Universidade Federal de São Carlos– UFSCar
Centro de Ciências Exatas e de Tecnologia– CCET
Departamento de Computação– DC
Programa de Pós-Graduação em Ciência da Computação– PPGCC

Ricardo Alexandre Neves

Sistema de Visão e Inteligência Computacional em Ambiente de Nuvem para Gestão de Risco da Ferrugem Asiática na Cultura da Soja

São Carlos
2024

Ricardo Alexandre Neves

Sistema de Visão e Inteligência Computacional em Ambiente de Nuvem para Gestão de Risco da Ferrugem Asiática na Cultura da Soja

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências Exatas e de Tecnologia da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Doutor em Ciência da Computação.

Área de concentração: Visão Computacional

Orientador: Prof.Dr.Paulo Estevão Cruvinel

São Carlos
2024

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Defesa de Tese de Doutorado do candidato Ricardo Alexandre Neves, realizada em 19/02/2024.

Comissão Julgadora:

Prof. Dr. Paulo Estevão Cruvinel (EMBRAPA)

Prof. Dr. Alexandre Luis Magalhães Levada (UFSCar)

Profa. Dra. Agma Juci Machado Traina (USP)

Profa. Dra. Kalinka Regina Lucas Jaquie Castelo Branco (USP)

Prof. Dr. José Mario de Martino (UNICAMP)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Ciência da Computação.

Este trabalho é dedicado a Deus e à Família, especialmente à minha esposa Geise e ao meu filho Felipe José, pelo amor incondicional e apoio nesta caminhada.

Agradecimentos

Ao Instituto Federal de São Paulo (IFSP), pela concessão de licença remunerada para qualificação stricto sensu, portaria Nº 1.644, de 09 de maio de 2019;

Aos Professores e Coordenação do Programa de Pós-Graduação (PPGCC), do Departamento de Computação (DC) da UFSCar;

À Embrapa Instrumentação, pelo acolhimento e suporte aos trabalhos desenvolvidos;

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Processo 17/19350-2, via IBM Brasil e Embrapa Instrumentação, pelo suporte financeiro à infraestrutura Oracle *Cloud* utilizada nesta pesquisa;

Aos amigos desta trajetória: Alex, Wilbur, Prof. Dr. Maurício Fernando Lima Pereira (UFMT) e, com destaque, ao Prof. Dr. Gabriel Marcelino Alves, do IFSP;

Aos pesquisadores da Embrapa que colaboraram na identificação dos especialistas para a validação do sistema, Dr. Augusto Guerreiro Fontoura Costa (Agrônomo, Embrapa Algodão), Dr. José Francisco da Silva Martins (Entomologista, Embrapa Clima Temperado) e Dr. Robson Rolland Monticelli Barizon (Agrônomo, Embrapa Meio Ambiente). E também, aos pesquisadores especialistas que colaboraram com as análises sobre a verdade de campo, utilizadas como referência na etapa de validação do sistema, Dr. Bernardo de Almeida Halfeld Vieira (Fitopatologista, Embrapa Meio Ambiente), Dr. Cley Donizeti Martins Nunes (Fitopatologista, Embrapa Clima Temperado), Dr. Dartanha José Soares (Fitopatologista, Embrapa Algodão); Dr. José Marcos Garrido Beraldo (Agrônomo, IFSP Matão), Dra. Katia de Lima Nechet (Fitopatologista, Embrapa Meio Ambiente) e Dr. Rafael Moreira Soares (Fitopatologista, Embrapa Soja) e ao pesquisador que colaborou com informações sobre a base de dados de folhas de soja, Mestre em Engenharia Elétrica, Luciano Vieira Koenigkan (Processamento de Imagens, Embrapa Agricultura Digital);

Deixo meus agradecimentos especiais ao Orientador, Professor Dr.Paulo Estevão Cruvinel, Pesquisador da Embrapa Instrumentação e Professor do quadro permanente do Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos - UFSCar, por sua sabedoria, conhecimento, paciência e primor na orientação.

"A persistência é o caminho do êxito."
(Charles Chaplin)

Resumo

O controle da Ferrugem Asiática (*Phakopsora pachyrhizi*) da soja (*Glycine max* (L.) Merril) requer alto uso de fungicidas, o que pode gerar resistência. Logo, novas soluções de controle têm sido requeridas para sua mitigação. Este trabalho apresenta um sistema de inteligência e visão computacional para a avaliação da presença ou não desta doença em área de cultura e estágio de severidade. Envolve o uso de técnicas de reconhecimento de padrões e aprendizado de máquina, viabilizando ações de diagnóstico para prognósticos e controle. Considera um modelo de suporte à decisão que utiliza variáveis aleatórias relacionadas ao clima, às plantas e às características de padrões reconhecidos em imagens digitais de folhas da soja sob monitoramento. Para a extração de características utiliza as técnicas de transformada de características invariantes à escala, histograma de gradientes orientados e momentos invariantes de Hu. Faz uso de infraestrutura computacional em ambiente de nuvem e processamento baseado em redes inteligentes, bem como técnica de análise de componentes principais na etapa de redução da dimensionalidade das características que são classificadas por máquina de vetor de suporte. Adicionalmente, considera o modelo baseado no uso de cadeias ocultas de Markov, que é dedicado à fusão do conjunto das variáveis aleatórias, oferecendo robustez, eficácia e eficiência, conforme resultados validados por correlação cruzada com respostas de especialistas. Para a avaliação da qualidade dos dados, nas diversas etapas do sistema, são ainda considerados conjuntos de métricas, tais como razão sinal-ruído de pico, erro médio quadrático, índice de similaridade estrutural, valores ausentes, acurácia, precisão, *F1-score* e revocação. A solução desenvolvida oferece prevenção e minimização do uso de fungicidas, agregando valor ao processo produtivo e orientando o futuro monitoramento espaço-temporal precoce da doença em escala agrícola.

Palavras-chave: processamento de imagens, aprendizado de máquina, suporte à decisão, ferrugem asiática da soja.

Abstract

Controlling Asian Soybean Rust (*Phakopsora pachyrhizi*) in soybeans (*Glycine max* (L.) Merril) often requires high fungicide use, which can lead to resistance. Thus, new control solutions are needed for mitigation. This work presents an intelligent computer vision system for assessing the presence and severity of this disease in crop areas. It involves pattern recognition and machine learning techniques, enabling diagnostic actions for prognosis and control. It considers a decision-support model using random variables related to climate, plants, and characteristics recognized in digital images of monitored soybean leaves. For feature extraction, it uses scale-invariant feature transform, histogram of oriented gradients, and Hu's invariant moments techniques. It uses cloud-based computational infrastructure and intelligent network processing, as well as principal component analysis for dimensionality reduction of features classified by support vector machines. Additionally, a hidden Markov model is used to fuse random variables, offering robustness, effectiveness, and efficiency, as confirmed by expert cross-correlation. To evaluate data quality at various system stages, metric sets such as peak signal-to-noise ratio, mean squared error, structural similarity index, missing values, accuracy, precision, F1-score, and recall are considered. This solution prevents and reduces fungicide use, enhancing production and guiding future early spatio-temporal monitoring of the disease on an agricultural scale.

Keywords: image processing, machine learning, decision support, Asian soybean rust.

Lista de ilustrações

Figura 1 – Diagrama de Blocos do Sistema de Inteligência e Visão Computacional 34

Figura 2 – Visão Geral da Revisão Sistemática 38
Figura 3 – Critérios de Pesquisa da Revisão Sistemática 39
Figura 4 – Estrutura de Busca Sistemática 39
Figura 5 – Produção de Soja no Brasil 41
Figura 6 – Ocorrências de FAS no Brasil Vs. Estádios Fenológicos - Safra 2022/23 42
Figura 7 – Ciclo de Evolução da Ferrugem da Soja 44
Figura 8 – Escala Diagramática da Ferrugem Asiática da Soja 46
Figura 9 – Tipos de Dados 54
Figura 10 – Arquitetura em Alto Nível do *Data Warehouse* 54
Figura 11 – Processo *MapReduce* 56
Figura 12 – Algoritmo SVM 77
Figura 13 – Algoritmo Naïve Bayes 78

Figura 14 – Diagrama Conceitual 93
Figura 15 – Diagrama Estruturação das Bases de Dados 95
Figura 16 – Visão Geral de Extração de Características com a Técnica SIFT 103
Figura 17 – Visão Geral de Extração de Características com a Técnica HOG 106
Figura 18 – SVM Linear: Margens Rígidas e Margens Suaves 112
Figura 19 – SVM Não Linear 112
Figura 20 – Interpretação geométrica 2D 116
Figura 21 – Variáveis Projetadas na Janela Temporal 120
Figura 22 – Exemplos de Figuras de Mérito no Círculo Trigonométrico 125
Figura 23 – Função de Pertinência Triangular 128
Figura 24 – Esquema do Modelo Difuso de Mamdani 129
Figura 25 – Estados, Transições e Matriz - Cadeia de Markov 135
Figura 26 – Grafo de um Modelo Oculto de Markov 137

Figura 27 – Configuração das Tabelas do *Data Warehouse* 139
Figura 28 – Framework Qualidade de Dados 141

Figura 29 – Arquitetura Oracle *Cloud* - Cenário 1 149
Figura 30 – Arquitetura Oracle *Cloud* - Cenário 2 150
Figura 31 – Arquitetura Oracle *Cloud* - Cenário 3 151
Figura 32 – Oracle *Cloud* - *Tela Inicial* 153
Figura 33 – Oracle *Cloud* Acesso Remoto Instância Computação 154
Figura 34 – Oracle *Cloud* - *Buckets* 155
Figura 35 – Oracle *Cloud* - *Autonomous Database* 157
Figura 36 – Oracle *Cloud* - *Data Science* 158
Figura 37 – Arranjo para a Série Temporal de Dados das Variáveis Utilizadas no Suporte à Decisão 160
Figura 38 – Avaliação de Diferentes Modelos para Completar as Séries Temporais de Dados 161
Figura 39 – Preparação dos Dados Ferramenta Oracle *Cloud* 163
Figura 40 – Imagens de Folhas de Soja com Fundo Complexo 165
Figura 41 – Pré-Processamento Equalização de Histograma 165
Figura 42 – Pré-Processamento Suavização 166
Figura 43 – Escolha dos Limiares para a Segmentação 167
Figura 44 – Exemplo Segmentação Etapas I e II 169
Figura 45 – Boxplot Métricas Segmentação 170
Figura 46 – Pontos-Chave por Padrão de Cor 172
Figura 47 – Descritores Hu, Hog e SIFT Não Normalizados 173
Figura 48 – Boxplot Comparativo dos Classificadores 176
Figura 49 – Classificador SVM - Kernel Polinomial 179
Figura 50 – Classificador SVM - *Kernel* RBF 179
Figura 51 – Classificador SVM - *Kernel* Linear 180
Figura 52 – Classificador SVM Kernel: Linear, Polinomial e RBF 181
Figura 53 – Resultado Figura de Mérito 182
Figura 54 – Funções de Pertinência 184
Figura 55 – Lógica Difusa: Saída de Defuzzificação 185
Figura 56 – Modelo de Cadeias Ocultas de Markov para Qualificação da Favorabilidade de Ocorrência da FAS 186
Figura 57 – Gráfico de Regras Contabilizadas 189
Figura 58 – Análise Dados Relatório Analítico - *Assunto 1* 193
Figura 59 – Relatório Analítico Favorabilidade Média - *Assunto 3* 195
Figura 60 – Análise Dados Relatório Analítico - *Assunto 2* 196
Figura 61 – Relatório Analítico Favorabilidade Alta - *Assunto 3* 197
Figura 62 – Relatório Analítico Favorabilidade Alta Detalhado - *Assunto 3* 198

Figura 63 – Interface Sobre o Projeto 200
Figura 64 – Interface do Processamento Imagens 201
Figura 65 – Interface do Relatório Técnico 202
Figura 66 – Interface de Qualidade de Dados 204
Figura 67 – Interface de Qualidade de Dados: Segmentação 205
Figura 68 – Interface de Qualidade de Dados: Segmentação - Continuação 206
Figura 69 – Interface de Qualidade de Dados: Extração de Características 208
Figura 70 – Interface de Qualidade de Dados: Aprendizado de Máquinas 209
Figura 71 – Interface de Qualidade de Dados: Aprendizado de Máquinas - Continuação 210
Figura 72 – Interface de Qualidade de Dados: Fusão de Dados 211
Figura 73 – Qualidade Dados Markoviano 213
Figura 74 – Qualidade Dados Banco Oracle 214
Figura 75 – Interface de Qualidade de Dados: *Data Warehouse* 215
Figura 76 – Interface de Relatório de Recomendações 216
Figura 77 – Interface do *Dashboard* 217
Figura 78 – Interface do *Dashboard - Assunto 1* 218
Figura 79 – Interface do *Dashboard - Assunto 2* 219
Figura 80 – Interface do *Dashboard - Assunto 3* 220
Figura 81 – Validação Quanto à Presença ou Não de Ferrugem Asiática 221
Figura 82 – Validação Quanto ao Nível de Severidade da Ferrugem Asiática 222

Figura 83 – Dicionário de Dados *Data Warehouse* 251
Figura 84 – Modelo Estrela - *Data Warehouse* 252

Figura 85 – Dicionário de Dados - Banco de Dados Relacional 253
Figura 86 – Modelo - Banco de Dados Relacional 254

Lista de tabelas

Tabela 1 – Estudos dos Classificadores 73

Tabela 2 – Requisitos de Projeto *Data Warehouse* 96
Tabela 3 – Kernels - Classificador SVM 113
Tabela 4 – Base de Regras - Ferrugem Asiática da Soja 119
Tabela 5 – Vetor de Dados - Entrada Fusão de Dados 123
Tabela 6 – Variáveis e Grandezas Físicas - Fusão de Dados 124
Tabela 7 – Configuração das Variáveis - Antecedentes e Consequentes 130
Tabela 8 – Inferências Difusas 131

Tabela 9 – Descrição dos *Buckets* Oracle *Cloud* 156
Tabela 10 – Exemplos de Janelamento nas Séries Temporais de Dados 161
Tabela 11 – Valores dos Coeficientes de Correlação para os Vários Modelos Analisados no Exemplo Considerado 162
Tabela 12 – Dados Métricas Segmentação 170
Tabela 13 – Comparação Redução de Dimensionalidade 174
Tabela 14 – Testes Classificador SVM Binário e Multiclasse 176
Tabela 15 – Classificadores - 50% Treino e 50% Teste 177
Tabela 16 – Classificadores - 70% Treino e 30% Teste 177
Tabela 17 – Classificadores - 80% Treino e 20% Teste 177
Tabela 18 – Configurações de Hiperparâmetros - *Grid Search* 178
Tabela 19 – Dados do Relatório do Classificador SVM - *Kernel* Polinomial 178
Tabela 20 – Dados do Relatório do Classificador SVM - *Kernel* RBF 179
Tabela 21 – Dados do Relatório do Classificador SVM - *Kernel* Linear 180
Tabela 22 – Hiperparâmetros Selecionados - *kernel* Polinomial 180
Tabela 23 – Estatística Descritiva - *kernel* Liner, Polinomial e RBF 181
Tabela 24 – Configuração das Funções de Pertinência 183
Tabela 25 – Dados Cadeia Oculta de Markov 187

Tabela 26 – Vetor Entrada de Dados 188
Tabela 27 – Testes de Comparação das Abordagens 190
Tabela 28 – Resultado Cadeia de Markov 191
Tabela 29 – Resultados Relatório Analítico - *Assunto 1* 194
Tabela 30 – Resultados Relatório Analítico - *Assunto 2* 195
Tabela 31 – Resultados Relatório Analítico - *Assunto 3* 199
Tabela 32 – Análise de Qualidade da Segmentação - Métricas e *Outliers* 207
Tabela 33 – Dados de Qualidade Markoviana 212

Tabela 34 – Tabela de Fungicidas 255
Tabela 35 – Tabela de Boas Práticas de Manejo da Soja 256

Lista de siglas

ADW	*Autonomous Data Warehouse*
APEX	*Application Express*
API	*Application Programming Interface*
AWS	*Amazon Web Services*
BDPA	Bases de Dados de Pesquisa Agropecuária
BIC	*Border/Interior Classification*
BLOB	*Binary Large Object*
BPNN	*Backpropagation Neural Network*
BSON	*Binary-encoded Serialization of JSON*
CCF	*Comprehensive Color Feature*
CCM	*Color Co-ocorrence Method*
CONAB	Companhia Nacional de Abastecimento
CPTEC	Centro de Previsão de Tempo e Estudos Climáticos
CSS	*Curvature Scale Space*
DCNN	Rede Neural Convolucional Profunda
DML	*Data Manipulation Language*
DN	*Digital Numbers*
DQD	Dimensão da Qualidade de Dados

DSIFT	*Dense Scale-Invariant Feature Transform*
DW	*Data Warehouse*
EC2	*Amazon Elastic Compute Cloud*
ELM	*Extreme Learning Machine*
EMBRAPA	Empresa Brasileira de Pesquisa Agropecuária
EMR	*Amazon Elastic MapReduce*
EROI	*Extended Region Of Interest*
ETL	*Extract, Transform and Load*
FAPESP	Fundação de Amparo à Pesquisa do Estado de São Paulo
FAS	Ferrugem Asiática da Soja
FCM	*Fuzzy C-Means*
GB	*Giga Bytes*
GLCM	*Gray Level Co-occurrence Matrix*
GPS	*Global Positioning System*
HDFS	*Hadoop Distributed File System*
HD	*High Definition*
HIST	Características do histograma de cores
HOG	*Histogram of Gradients*
HSI	*Hue, Saturation, Intensity*
IEEE	Instituto de Engenheiros Elétricos e Eletrônicos
IEFD	*Invariant Elliptic Fourier Descriptor*
INMET	Instituto Nacional de Meteorologia
IoT	*Internet of Things*
JPEG	*Join Photographic Experts Group*
JSON	*JavaScript Object Notation*
KNN	*K-Nearest Neighbour*

LBP	*Local Binary Pattern*
LDA	*Linear Discriminant Analysis*
LiDAR	*Light Detection and Ranging* ou *Laser Imaging Detection and Ranging*
LoRa	*Long Range*
MAPA	Ministério da Agricultura, Pecuária e Abastecimento
MPP	*Massively Parallel Processing*
MSE	*Mean Squared Error*
NIR	*Banda Infravermelho Próximo*
NRMSE	*Normalized Root Mean Square Error*
NoSQL	*Not Only SQL*
OCI	*Oracle Cloud Infrastructure*
OCPU	*Oracle Compute Unit*
OLAP	*Online Analytical Processing*
OLTP	*Online Transaction Processing*
ONGs	Organizações Não Governamentais
PCA	*Principal Component Analysis*
PHOW	*Pyramid Histograms of Visual Words*
PIS	Processamento de Imagens e Sinais
PITE	Parceria para Inovação Tecnológica
PLSR	*Partial Least Squares Regression*
PNN	Rede Neural Probabilística
PSNR	*Peak Signal-to-Noise Ratio*
PSO	*Particle Swarm Optimization*
RBF	*Radial Basis Function*
RCI	Índice de Cor de Ferrugem
RGB	Modelo de Cores RGB - *Red, Green, Blue*

RIA	Proporção de Área Infectada
RIWD	*Rotation Invariant Wavelet Descriptor*
ROI	*Region of Interest*
SGBD	Sistema de Gerenciamento de Banco de Dados
SGDM	*Gray-Level Dependence Matrices*
SIFT	*Scale Invariant Feature Transform*
SLIC	*Simple Linear Iterative Clustering*
SNR	*Signal-to-Noise Ratio*
SOSD	Sistema Online de Suporte à Decisão
SPAD	*Soil Plant Analysis Development*
SQL	*Structured Query Language*
SRG	*Seeded Region Growing*
SSH	*Secure Socket Shell*
SSIM	*Structural Similarity Index*
SSLBP	Padrão Binário Local Simétrico Quadrado
SURF	*Speeded-Up Robust Features*
SVM	*Support Vector Machine*
StAr	*State of the art through Systematic Review*
TIFF	*Tagged Image File Format*
UDB	*Unstructured Data Base*
VANT	Veículo Aéreo Não Tripulado
WDH	Histograma de cores decompostas *wavelet*
XML	*Extensible Markup Language*

Lista de símbolos

A	matriz que contém os dados originais antes da transformação PCA e deve ser assimétrica, positiva e semidefinida
AR	área total da figura de mérito
AUC	cálculo da área da curva ROC do classificador
C	*cluster*
$F1$	cálculo do *f1-score* do classificador
H_1	hiperplano classificador SVM
$I(cx, cy)$	imagem na posição cx e cy
IA	imagem A para o cálculo de Erro Quadrático Médio entre duas imagens
IB	imagem B para o cálculo de Erro Quadrático Médio entre duas imagens
LM	limiar
M_{ij}	magnitude do gradiente na posição (i, j) da imagem
N	medições para o processo de Markov
OF	ocorrência de favorabilidade resultante da intersecção entre as possibilidades normalizadas de todas as variáveis da figura de mérito
$P_{ij}^{(n)}$	probabilidade de transição de n passos
Q	matriz dos autovetores de colunas (PCA)
R_{ij}	orientação do gradiente na posição (i, j) da imagem

S	subespaço a ser eliminado durante o processo de redução da dimensionalidade (PCA)
$SC_3(xi)$	*spline* cúbica interpolante de $f(xi)$
T	novo subespaço no processo de redução de dimensionalidade (PCA)
TFP	taxa de falsos positivos do classificador
TMh	limiar superior
TVP	taxa de verdadeiros positivos do classificador
$Temp$	temperatura atual de bulbo seco dada em (°C)
$Temp_{po}$	ponto de orvalho com a temperatura dada em (°C)
UR	umidade relativa dada em porcentagem (%)
$\Delta\bar{\jmath}$	erro padrão que mede a incerteza associada à estimativa da média amostral $\bar{\jmath}$
$\Gamma(xi)$	*b-spline* cúbica: particularidade das *splines*, cujo significado se traduz em *basis spline*
$\Im$	espaço de características para transformação no classificador SVM
Λ	matriz diagonal dos autovalores (PCA)
Ψ	representa a função de pertinência difusa
Θ	ângulo formado entre os vetores do triângulo da figura de mérito
β_0	variável de saída para a qual calcula-se a pertinência no conjunto difuso B'
$\bigwedge$	operador de interseção para combinar as condições do antecedente da regra difusa
$\mathbf{\Phi}(\mathbf{xo})$	transformação aplicada às componentes xo_1 e xo_2 de **xo**
χ	objeto que pode ou não ser membro de um conjunto na lógica difusa
$\delta^{(i)}$	resultado da i-ésima regra difusa
γ_j^2	quadrado da variância da j-ésima componente principal (PCA)
$\hat{C}(0)$	valor da autocorrelação em $t = 0$ para uma observação
$\hat{C}(t)$	valor da autocorrelação para uma observação em um dado momento t

ι_i	ponto na clusterização *fuzzy*
ι_i	ponto na imagem
κ	ponto na imagem
λ_j	penalização aplicada às componentes principais para controlar sua contribuição com o objetivo de equilibrar a maximização da variância
$\mathbf{X}$	conjunto de treinamento composto de n pares $(\mathbf{xo}_i, yr_i)$ para geração um classificador SVM particular $\widehat{h} \in H$
$\mathbf{\Sigma_x}$	matriz de covariância dos dados originais **xo** simétrica e definida como positiva (PCA)
$\mathbf{v}_j$	componentes principais (PCA)
$\mu_\alpha(\chi)$	função de pertinência triangular
$\mu_{B'}(\beta)$	pertinência de β a um conjunto fuzzy B'
μ_i	centroide do *cluster* i
μ_{pq}	momentos centrais de ordem $(p+q)$ e invariantes às transformações de translação e rotação
$\overline{cx}$	coordenada do centro de massa momentos de HU
$\overline{cy}$	coordenada do centro de massa momentos de HU
π_i	distribuição inicial de estados
σ	cálculo da variância
σ^2	cálculo do desvio padrão
$\star_S$	operador *s-norm* que representa a operação de máximo (ou *OR* máximo)
$\star_T$	operador *t-norm* que representa a operação de mínimo (ou *AND* mínimo)
τ_{int}	tempo de integração ou o tempo de correlação integrada
$\underline{\mathbb{X}}$	conjunto de valores que considera a variabilidade *fuzziness*, ou seja, falta de precisão ou à incerteza associada
$\varnothing$	invariante ortogonal dos momentos de HU

ξ_i	variável de folga aplicada ao classificador SVM Linear com Margem Suave
ac	cálculo de acurácia do classificador
$argmax$	argumento que maximiza
c	coeficientes para configuração particularidade das *splines*, cujo significado se traduz em *basis spline*
cx	*pixel* na posição x
cy	*pixel* na posição y
dp_{ij}	probabilidade de transição de um estado so_i para um estado so_j em uma cadeia de Markov
e_i	configurações geradas para a série temporal e i é a ordem temporal medida entre medições N
g	ordem da *b-spline*
la	lado "a" conhecido do triângulo da figura de mérito
lb	lado "b" conhecido do triângulo da figura de mérito
mb_{pq}	momentos bidimensionais de HU
$prec$	cálculo de precisão do classificador
rev	cálculo de revocação do classificador
t	representam os nós da *b-spline* cúbica
t	valor do limiar
u_{ij}	grau de pertinência do ponto ι_i na clusterização *fuzzy*
v_j	centroide na clusterização *fuzzy*
w	vetor de pesos do classificador SVM após a transformação de características
xi	ponto nas funções de interpolação polinomial e *spline* cúbica
PI(xi)	forma geral da função interpoladora polinomial

Sumário

1 INTRODUÇÃO 31
1.1 Contextualização 31
1.2 Hipótese 32
1.3 Motivação 32
1.4 Objetivo Geral 33
1.5 Objetivos Específicos 33
1.6 Visão Geral do Sistema 33
1.7 Principais Contribuições 34
1.8 Organização do Documento 35

2 REVISÃO DA LITERATURA EM ESTADO DA ARTE . . . 37
2.1 Abordagem Utilizada para Revisão Sistemática da Literatura . 37
2.2 Principais Aspectos da Revisão Bibliográfica 39
2.2.1 Desenvolvimento da Soja no Brasil 40
2.2.2 Ferrugem Asiática da Soja 41
2.2.3 Doenças na Cultura da Soja e Diagnósticos Baseados em Uso de Visão Computacional 48
2.2.4 Estruturação de Bases de Dados 51
2.2.4.1 Soluções Aplicadas às Fontes de Dados *Big Data* 53
2.2.4.2 Soluções de Processamento Paralelo Aplicadas à Dados *Big Data* 55
2.2.4.3 Infraestrutura para Armazenamento e Gerenciamento de Dados *Big Data* em Ambiente de Nuvem 58
2.2.4.4 Processo de Preparação de Dados *Big Data* via Aprendizado de Máquina . . 60
2.2.4.5 A Qualidade dos Dados na Obtenção de Resultados Confiáveis 62
2.2.5 Pré-processamento e Processamento de Imagens Digitais 66
2.2.6 Extração de Características e Aprendizado de Máquina (Classificação) . 71
2.2.6.1 Reconhecimento de Padrões 78

2.2.7 Fusão de Dados de Sensores para o Controle de Doenças de Plantas . . 84
2.3 Destaques e Contextualização 88
2.4 Considerações Finais 90

3 MATERIAIS E MÉTODOS 91
3.1 Materiais 91
3.2 Métodos 93
3.2.1 Métodos para avaliações da Qualidade de Dados 140
3.2.2 Considerações Finais 145

4 RESULTADOS E DISCUSSÕES 147
4.1 Configuração, Integração e Estruturação dos Dados 147
4.1.1 Estruturação de Dados das Séries Temporais 159
4.2 Pré-Processamento e Processamento Digital das Imagens 164
4.3 Redução de Dimensionalidade do Vetor de Características . . . 173
4.4 Reconhecimento de Padrões, Classificação e Aprendizado de Máquina 175
4.5 Fusão de Variáveis e Auxílio à Tomada de Decisão 182
4.6 Relatórios Analíticos, Recomendações e *Dashboard* 191
4.7 Validação do Sistema a partir de Visão Especialista 221
4.8 Limitações e Dificuldades Encontradas no Desenvolvimento do Trabalho 222
4.9 Considerações Finais 223

Conclusões 231

REFERÊNCIAS 233

APÊNDICES 249

APÊNDICE A – MODELO MULTIDIMENSIONAL (DW) 251

APÊNDICE B – MODELO BANCO DADOS RELACIONAL . . 253

APÊNDICE C – RECOMENDAÇÕES AGRÍCOLAS 255

APÊNDICE D – QUESTIONÁRIO ELABORADO PARA A ETAPA DE VALIDAÇÃO, RESPONDIDO POR ESPECIALISTAS 257

Capítulo 1

Introdução

Este Capítulo apresenta a motivação, o problema a ser tratado, hipótese e os objetivos e uma visão geral do sistema. Também, detalha a organização deste documento.

1.1 Contextualização

O advento da agricultura digital trouxe oportunidades para o desenvolvimento de trabalhos na linha das ferramentas de gestão de risco, tanto no contexto de melhorias, quanto para novas abordagens. Nesse sentido, a Embrapa, ao longo dos anos, vem desenvolvendo projetos nesta frente de pesquisa, com apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). Inicialmente, oportunizaram-se iniciativas sobre a análise de risco de doenças da bananicultura atuando, especificamente, sobre a Sigatoka Negra (FERNANDES, 2005; CRUVINEL et al., 2011).

Atualmente, outros trabalhos em gestão de risco estão em desenvolvimento. O primeiro deles trata o risco agrícola de uma maneira mais abrangente e conta com fomento da FAPESP, Auxílio à Pesquisa - Parceria para Inovação Tecnológica - PITE, Processo 17/19350-2, Convênio/Apoio IBM Brasil. Tal projeto trata o risco sob três aspectos: risco de perda, risco agrícola do processo de produção e o risco logístico.

Os tipos de dados do ambiente agrícola são variados e os modelos de decisão que consideram o manejo baseado na agricultura de precisão geram um grande volume de dados, o que necessita de sistemas robustos e conhecimentos para tomada de decisão qualificada. As ações baseadas em conhecimento permitem que os produtores agrícolas possam se antecipar, evitar e reagir aos acontecimentos decorrentes de externalidades negativas que possam ocorrer em áreas de cultivo. Os serviços de processamento que compõem a arquitetura, para a gestão de riscos agrícolas, consideram uma infraestrutura de computação

em nuvem, sendo específicos para aplicações agrícolas que envolva a geoespacialização de grandes quantidades de dados e análises, desde o suporte à decisão (SIMIONATO et al., 2021).

Essa massa de dados, após sua coleta, processamento e transformação, produz conhecimento, por meio de modelos e algoritmos, tais que os resultados possam ser disponibilizados em forma de relatórios analíticos e recomendações para tomada de decisão, frente à necessidade da diminuição das perdas e dos riscos agrícolas decorrentes da ferrugem asiática da soja.

O aprimoramento de modelos da visão computacional, que processam variáveis do universo agrícola, vêm propiciando oportunidades para estudos sobre incidência das doenças em plantas. Abordagens baseadas em inteligência computacional podem agregar valor à gestão de riscos e aos prognósticos que visem soluções sustentáveis.

1.2 Hipótese

A fusão de variáveis climáticas, imagens e dados de planta da soja, com inteligência e visão computacional, viabiliza o diagnóstico preciso sobre a ocorrência da ferrugem asiática.

1.3 Motivação

A motivação do trabalho está focada na utilização dos conceitos fundamentados em Ciência da Computação, na área de Visão Computacional, por meio do Processamento Digital de Imagens e Sinais e Inteligência Computacional, para atuar na solução de problemas relacionados a produtos, alimentos e energia de biomassa, sabendo dos desafios que indicam a necessidade do aumento de produção de alimentos em 70%, até 2050. Neste contexto, é buscado contribuir com a redução de perdas ocasionadas pela Ferrugem Asiática na produção de soja, por meio do desenvolvimento de uma nova abordagem para avaliação da favorabilidade da doença, fornecendo subsídios para a tomada de decisão agrícola. Dentre as motivações, destaca-se o desenvolvimento de tal abordagem em ambiente de nuvem, a partir de imagens digitais no formato RGB e dados de séries temporais climáticas relacionadas à área de cultivo.

Adicionalmente, prover um ambiente com facilidades e serviços de cunho científico e tecnológico, via web, para atender às necessidades do produtor de soja, quanto ao estabelecimento de um diagnóstico preciso e dinâmico de uma análise real relacionada a ferrugem asiática.

1.4 Objetivo Geral

Desenvolver um sistema de inteligência e visão computacional, em ambiente de nuvem, para monitorar estágios e avaliar a dinâmica de ocorrência da ferrugem asiática, em cultura da soja.

1.5 Objetivos Específicos

1. Organizar estrutura de dados de séries temporais, de forma a viabilizar o uso de fontes estruturadas, semiestruturadas e não estruturadas;

2. Selecionar, a partir do estado da arte da literatura, métodos para o processamento de imagens digitais de folhas da soja e aprendizado de máquina que viabilizem a organização de informações a serem completadas com dados climáticos para a definição dos estágios de severidade da ferrugem asiática;

3. Estabelecer método, para a fusão de variáveis aleatórias, contidas em um vetor com informações climáticas e de classificação para o suporte à decisão e controle;

4. Elaborar *dashboard*, para a apresentação de resultados, incluindo relatórios analíticos e recomendações.

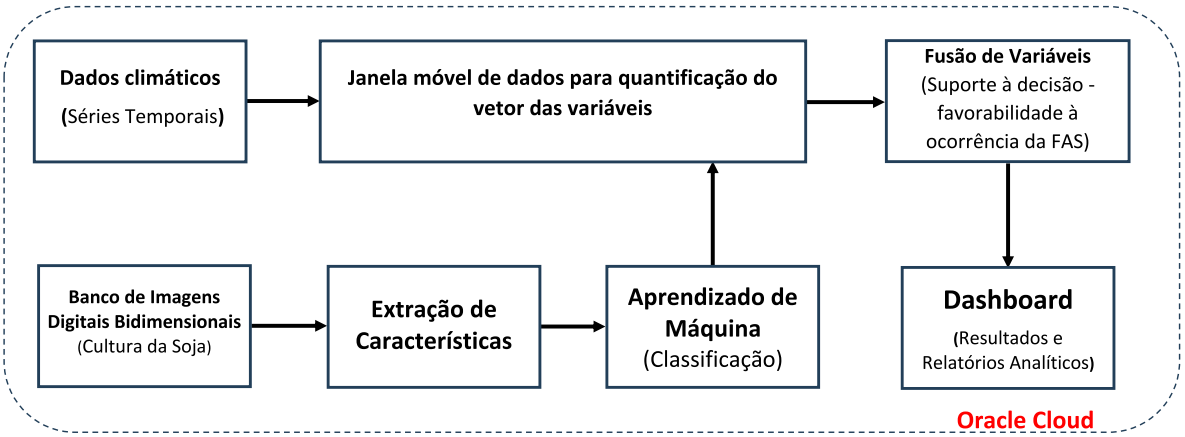
1.6 Visão Geral do Sistema

A Figura 1 apresenta uma visão macro, em diagrama de blocos, para o sistema de visão e inteligência computacional em ambiente de nuvem para a gestão de risco da ferrugem asiática na cultura da soja. O bloco (1), relacionado a dados climáticos, tem o propósito de consolidar as séries temporais de dados, originadas das estações climáticas, referentes às variáveis: precipitação, temperaturas mínima e máxima, umidade relativa do ar, ponto de orvalho e temperatura média compensada; O bloco (2) tem o propósito de quantificar o vetor das variáveis, recebendo como entrada as séries temporais consolidadas para compor o vetor das variáveis, bem como estabelecer as janelas temporais para a análise do conjunto das variáveis, conforme acima mencionado (períodos de 10 dias); O bloco (3) tem o propósito de extrair características que, como entrada, recebem as imagens originadas do banco de imagens digitais bidimensionais da cultura da soja (bloco 4), para efetuar a extração das características das imagens; O bloco (5) tem o propósito de executar o modelo de aprendizado de máquina, o qual recebe como entradas as características dos padrões identificados nas imagens digitais. Este resultado é, em seguida, adicionado no vetor das variáveis; O bloco (6) tem o propósito de executar o modelo para a fusão das variáveis. Recebe como entrada os dados das janelas móveis pré-estabelecidas para a análise

da favorabilidade à doença, bem como sua severidade, quando pertinente; Os resultados para suporte à decisão são apresentados em um *dashboard*.

Os desdobramentos do diagrama de blocos do sistema estão descritos no Capítulo 3.2.

Figura 1 – Diagrama de Blocos do Sistema de Inteligência e Visão Computacional



Fonte: Próprio Autor

A favorabilidade é definida como inferência estatística em função do comportamento do conjunto das variáveis consideradas e relacionadas à ocorrência da FAS. Neste contexto, os conceitos relacionados à favorabilidade, risco e ocorrência da FAS são entendidos como:

1. **Favorabilidade e risco:** O risco trata sobre a possibilidade de ocorrer a doença em decorrência de sua favorabilidade, frente ao comportamento do conjunto das variáveis. Assim, a favorabilidade pode ser reconhecida como baixa, média ou alta.

2. **Ocorrência da FAS:** É definida quando cada variável envolvida, no conjunto, apresentar condição propícia à doença, de acordo com os dados climáticos, de classificação da imagem da folha da soja. Neste caso, tem-se, para a favorabilidade baixa, o número de variáveis com ocorrência à FAS igual a 0 (zero) ou menor igual a 2. Para a favorabilidade média, são consideradas de 3 a 4 variáveis com ocorrência e, para a favorabilidade alta, considera-se de 5 a 7 variáveis com ocorrência à FAS.

1.7 Principais Contribuições

1. A elaboração de um modelo de decisão para a avaliação do estágio de favorabilidade da ferrugem asiática provocada pelo patógeno (*Phakopsora pachyrhizi*), baseado em visão e inteligência computacional, em ambiente de nuvem;

2. A concepção de uma estruturação das bases de dados que atende às necessidades do projeto da ferrugem asiática da soja, a partir de diferentes fontes de dados heterogêneas;

3. A elaboração de uma arquitetura em nuvem, via Oracle *Cloud*, que permite integrar suas tecnologias e fornecer um resultado em interfaces web amigáveis, considerando os conceitos da agricultura 4.0 e suas arquiteturas associadas à Internet das Coisas (IoT);

4. O desenvolvimento de uma técnica para fusão de dados fundamentada na integração das variáveis originadas de diferentes fontes e grandezas físicas e de interesse agrícola;

5. A elaboração de um *framework* de qualidade de dados para avaliação das diferentes etapas envolvidas no desenvolvimento do sistema.

1.8 Organização do Documento

Este documento está organizado em cinco Capítulos. O primeiro Capítulo retrata a introdução que apresenta o trabalho quanto à definição do problema, a motivação, os objetivos e uma visão geral do sistema. O segundo Capítulo é dedicado ao estado da arte, que inclui o processo de revisão sistemática, os referenciais bibliográficos, tais como: as doenças na cultura de soja com ênfase nos aspectos da ferrugem asiática, a estruturação das bases de dados, o processamento e pré-processamento de imagens digitais, a extração de características e aprendizado de máquina (classificação) e a fusão de dados de sensores para o ambiente agrícola, com vista à doenças de plantas. No terceiro Capítulo são apresentados os materiais e métodos para o desenvolvimento do sistema para avaliação da favorabilidade da ferrugem asiática em cultura de soja. O quarto Capítulo apresenta os resultados e discussões referentes aos experimentos realizados. Finalmente, o quinto Capítulo apresenta as conclusões deste trabalho.

Capítulo 2

Revisão da Literatura em Estado da Arte

Este Capítulo apresenta a revisão do estado da arte, visando a avaliação de técnicas já publicadas, como subsídio à composição de parte do sistema para gestão de risco da ferrugem asiática na cultura da soja.

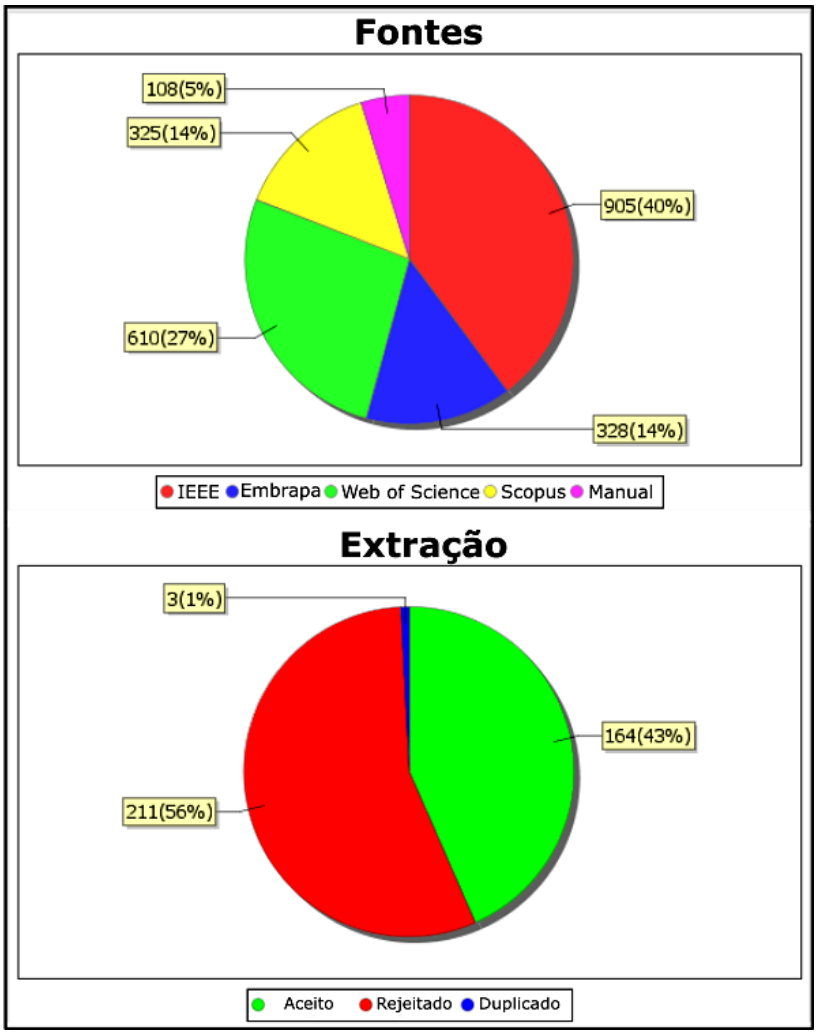
2.1 Abordagem Utilizada para Revisão Sistemática da Literatura

A revisão sistemática compreendeu o processo utilizado para a construção do estado da arte, o qual foi dividido nas etapas de Planejamento e Condução. A primeira delas é relacionada às tarefas de elaboração do protocolo de pesquisa: definição das questões de pesquisa; definição da estratégia de busca; definição das fontes de pesquisa; definição das *strings* de busca; definição dos critérios de seleção (inclusão e exclusão) e definição dos critérios de qualidade. A segunda etapa, de condução, consistiu em identificar os estudos primários, via estratégia de busca; selecionar os estudos primários pelos critérios de seleção e qualidade; extração dos dados e, por fim, o processo de sintetização dos mesmos. Essas etapas de revisão sistemática foram apoiadas pela ferramenta StArt (*State of the art through Systematic Review*), desenvolvida pelo LaPES (Laboratório de Pesquisa em Engenharia de Software) do Departamento de Computação da UFSCar, São Carlos.

Na Figura 2, foram representadas as fontes de busca utilizadas, tais como o Instituto de Engenheiros Elétricos e Eletrônicos (IEEE), as Bases de Dados de Pesquisa Agropecuária (BDPA) da Embrapa, *Web of Science*, *Scopus*, busca manual, e o processo de extração,

ilustrados por meio dos gráficos de pizza, os quais mostraram as porcentagens sob os critérios de inclusão e exclusão.

Figura 2 – Visão Geral da Revisão Sistemática



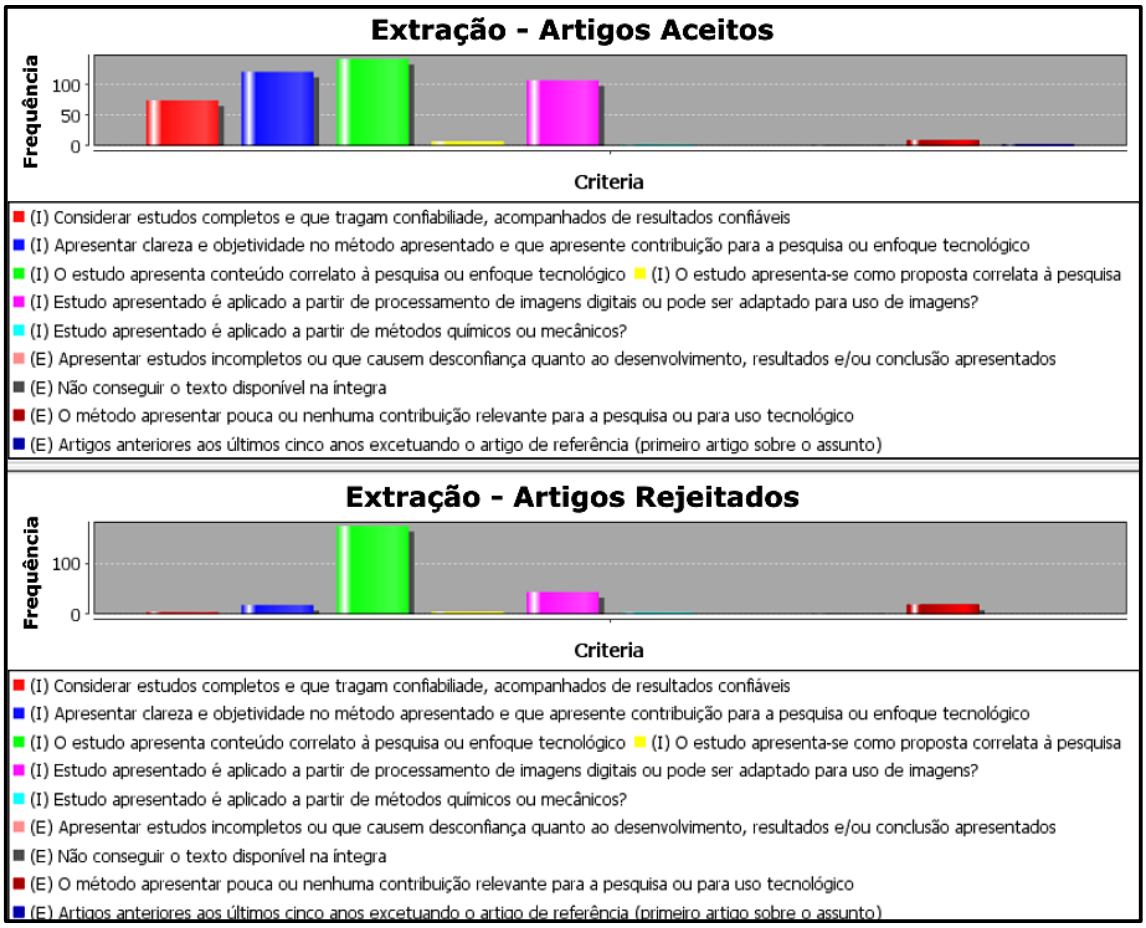
Fonte: Próprio Autor

A Figura 3 ilustra a etapa de extração de informações de bases bibliográficas, diante dos critérios de pesquisa, conforme os quesitos de aceite e rejeição de documentos.

As palavras-chave, em ordem alfabética, utilizadas para a definição das *strings* nas línguas Inglesa e Portuguesa foram as seguintes: *agriculture*; agricultura; *classification algorithms*; classificação de algoritmos; *crops*; campos; cultivo; *digital images*; imagens digitais; *image processing*; processamento de imagens; *method*; método; *pattern recognition*; reconhecimento de padrões; *remote sensing*; sensoriamento remoto; *risk managment*; gerenciamento de risco; *soybean*; soja.

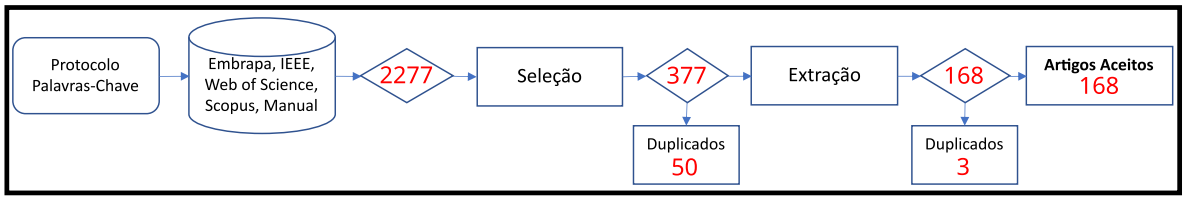
Na Figura 4, se encontra representada a estrutura de busca sistemática, considerando todas as etapas do processo.

Figura 3 – Critérios de Pesquisa da Revisão Sistemática



Fonte: Próprio Autor

Figura 4 – Estrutura de Busca Sistemática



Fonte: Próprio Autor

2.2 Principais Aspectos da Revisão Bibliográfica

Esta seção apresenta os referenciais bibliográficos dos assuntos relativos à pesquisa na seguinte sequência: (1) Desenvolvimento da soja no Brasil; (2) Ferrugem asiática da soja; (3) Doenças na cultura de soja com ênfase nos aspectos da ferrugem asiática; (4) Estruturação das bases de dados: refere-se à organização dos dados pautada por suas tecnologias,

infraestrutura, preparação e qualidade dos dados; (5) Processamento e Pré-processamento de Imagens Digitais; (6) Extração de características e aprendizado de máquina (classificação); e (7) Fusão de dados de sensores para o ambiente agrícola com vista às doenças de plantas.

2.2.1 Desenvolvimento da Soja no Brasil

A soja (*Glycine max* (L.) Merril) é considerada uma das leguminosas mais importantes do mundo, pois pode ser utilizada tanto para o consumo humano, quanto para consumo animal. Também é utilizada para fabricação de produtos industrializados, tais como óleo, adesivos não tóxicos, velas e tintas. A soja tem alto teor de proteína e inúmeros nutrientes benéficos à vida humana, assim como fatores bioativos. Estas características qualificam a soja como um produto adequado para a melhoria da dieta e o combate à desnutrição de milhões de pessoas nos países em desenvolvimento, quando incorporada com outros alimentos. Adicionalmente, a soja tem um papel importante na melhoria da fertilidade do solo, pois consegue contribuir com a diminuição da aplicação de fertilizantes à base de nitrogênio, dada sua característica de conseguir fixar, aproximadamente, de 43 a 103 kg de nitrogênio atmosférico por hectare ao ano. Tal fator proporciona ganhos econômicos e ambientais importantes (MURITHI et al., 2016).

De acordo com Godoy e colaboradores, no Brasil, a partir da década de 1970, a soja passou a ser considerada um produto economicamente importante e, desde então, sua relevância no mercado agrícola mundial aumentou (GODOY et al., 2016).

Frente aos dados do CONAB (2023), referentes à produção de soja na safra de 2021/22, foram observados destaques dos índices de produtividade das regiões Centro-Oeste e Sul do Brasil, evidenciando-se os melhores resultados para os estados do Paraná, sendo o Mato Grosso o maior produtor.

A Figura 5 ilustra a evolução da produção da soja no Brasil, levando em consideração as safras de 1990/91 até 2021/22 e a previsão para 2022/23.

Destacou-se, na safra de 2022/23, segundo os dados econômicos da Embrapa Soja (2023), a produção de 369.029 milhões de toneladas de soja no mundo, sendo o Brasil, nesta safra, o maior produtor mundial de soja, responsável pela produção de 154.566,3 milhões de toneladas e os Estados Unidos (EUA), o segundo maior produtor, com 116.377 milhões de toneladas.

Ainda de acordo com Godoy e colaboradores, no Brasil as perdas de produtividade da soja variaram entre as safras e se mostraram significativas ao longo do período considerado. Nesse contexto, as perdas ocorrem devido às doenças que acometem a cultura da soja, em particular a ferrugem asiática como a doença mais ameaçadora observada no período por esses autores. Ademais, sem que as medidas adequadas de controle tivessem sido consideradas, provocou perdas de até 90% na produtividade (GODOY et al., 2016).

Figura 5 – Produção de Soja no Brasil

Fonte: Adaptado de (CONAB, 2023)

Ao evidenciar a importância da soja para Brasil e o mundo, também se faz necessário estabelecer métodos que possam garantir a compreensão sobre seu manejo cultural, o que envolve uma especial atenção ao controle de pragas e doenças, principalmente no que tange a ferrugem asiática, ou seja, quanto a sua presença ou não, bem como sobre os estágios de severidade ou favorabilidade para a sua presença na área de cultura.

2.2.2 Ferrugem Asiática da Soja

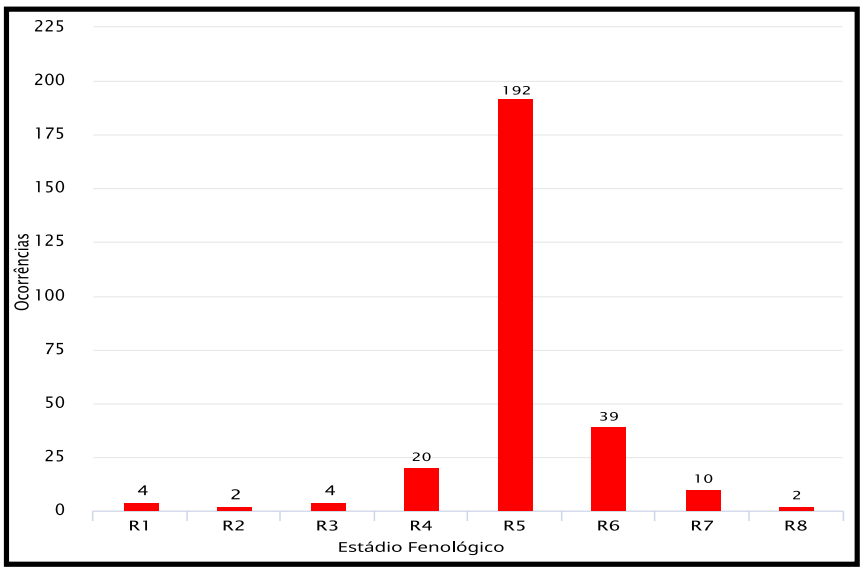
A Ferrugem Asiática da Soja (FAS) é uma doença que ameaça a sustentabilidade da cultura da soja e que, segundo Yorinori e colaboradores, espalhou-se pelo Brasil em 2001, a partir da primeira constatação da doença no Paraguai e no estado do Paraná (YORINORI et al., 2003).

Segundo os dados do Consórcio Antiferrugem (2023), na safra de 2022/23, todas as regiões do Brasil apresentaram pelo menos uma ocorrência da FAS nas áreas produtivas. Entretanto, as ocorrências da FAS foram observadas em diferentes estádios fenológicos da planta (NUNES; MARTINS; PONTE, 2018). Ademais, esta doença pode se expressar em todas as fases fenológicas da cultura, quando não tratada adequadamente.

Os estádios fenológicos são relacionados ao desenvolvimento das culturas, sendo que para a soja estas fases são compostas pelas vegetativas (VC,V1,V2,V3,VN) e reprodutivas (R1,R2,R3,R4,R5,R6,R7,R8 e R9), conforme descrito por Fehr e Caviness (1977).

Na Figura 6 estão apresentados os dados das ocorrências totais da FAS, frente aos estádios fenológicos da soja no Brasil, para a safra de 2022/23.

Figura 6 – Ocorrências de FAS no Brasil Vs. Estádios Fenológicos - Safra 2022/23



Fonte: Adaptado de (CONSÓRCIO ANTIFERRUGEM, 2023)

Ao considerar os pontos abordados sobre a FAS, passou a ser fundamentalmente importante o conhecimento sobre o comportamento do patógeno, assim como sua atuação na área de cultura. Neste contexto, fatores como a forma de disseminação da doença, as condições climáticas, a influência de outras variáveis do ambiente e as modificações e o surgimento de outros padrões nas folhas da soja passaram a ser relevantes para o entendimento e o controle da doença, visando a minimização das perdas.

Quanto ao desenvolvimento da doença, o fungo *Phakopsora Pachyrhizi* é o patógeno causador da FAS, doença altamente agressiva, que atua diretamente nas folhas da planta, causando a desfolha precoce. O patógeno somente sobrevive e se multiplica em plantas vivas. Com o aparecimento do fungo na planta, a doença se desenvolve em três estágios: inicial, intermediário e avançado. No estágio inicial, a doença se manifesta como pequenas manchas de cor amarelada ou alaranjada na superfície das folhas. No estágio intermediário, caracterizado pelo aumento das lesões nas folhas de soja, as manchas de ferrugem se expandem e se fundem, formando áreas maiores com uma coloração mais avermelhada. Entretanto, no estágio avançado, as folhas infectadas apresentam um aspecto amarelado ou bronzeado, devido à intensa colonização do fungo, cobrindo grandes áreas das folhas, levando à desfolha precoce da planta. A doença impede a completa formação dos grãos da soja e, como consequência, diminui consideravelmente a produtividade (YORINORI et al., 2003; GODOY et al., 2009; LELIS et al., 2009; GOULART; FURLAN; FUJINO, 2011; TANIMOTO et al., 2011; GODOY et al., 2016).

A disseminação deste patógeno é feita pelo vento, dificultando seu controle. Assim, o vento pode levar o fungo tanto na própria lavoura, quanto para lavouras vizinhas próximas ou distantes.

No Brasil, a disseminação e o estabelecimento do patógeno é agravada pela existência de plantas hospedeiras que, mesmo no período de vazio sanitário[1], nas entressafras, mantém o fungo em seu ciclo de reprodução (YORINORI et al., 2003; GODOY et al., 2016).

A calendarização[2] tem sido utilizada com a finalidade de atrasar as primeiras ocorrências da doença nos estádios iniciais do desenvolvimento da soja, viabilizando um melhor controle (GODOY et al., 2017).

Analisando de uma forma mais profunda a evolução desta doença, na área de cultura, é possível considerar que:

1. Aparecimento de pontos mais escuros que o tecido sadio da folha, de no máximo 1mm de diâmetro, com coloração esverdeada a cinza-esverdeada;

2. Uma protuberância semelhante a uma bolha é observada, o que caracteriza o início da formação da estrutura de reprodução do fungo, denominado urédia;

3. Urédias adquirem a coloração de castanho claro a castanho escuro e se abrem em um minúsculo poro, expelindo os uredósporos, os quais são carregados pelo vento, de forma a provocar a disseminação do fungo para outras folhas ou plantas;

4. O tecido em volta das urédias, a partir da esporulação, é lesionado e sua coloração alterada para castanho claro, denominada lesão do tipo TAN ou para a coloração castanho avermelhado, caracterizada por lesão do tipo RB;

5. Lesões são facilmente visíveis a olho nu, entretanto apresentam dificuldade de classificação.

A Figura 7 ilustra estes ciclos da evolução da doença na planta de soja.

Segundo a CONAB, a FAS é considerada uma doença que demanda muitos cuidados, tanto na monitoramento, quanto em seus processos de controle, para que as perdas na produtividade possam ser minimizadas (CONAB, 2023). Nesse sentido, as atenções estão voltadas, principalmente, no combate à doença por meio dos fungicidas químicos (ZUNTINI et al., 2019; DORIGHELLO et al., 2020). No entanto, percebe-se que o patógeno da FAS está cada vez mais resistente às diversas classes de fungicidas (JULIATTI et al., 2017). Adicionalmente, o excesso de aplicações de fungicidas, fora da calendarização, implica em prejuízos ao meio ambiente e também aos produtores, pois acarreta no aumento dos custos de produção (NUNES, 2014).

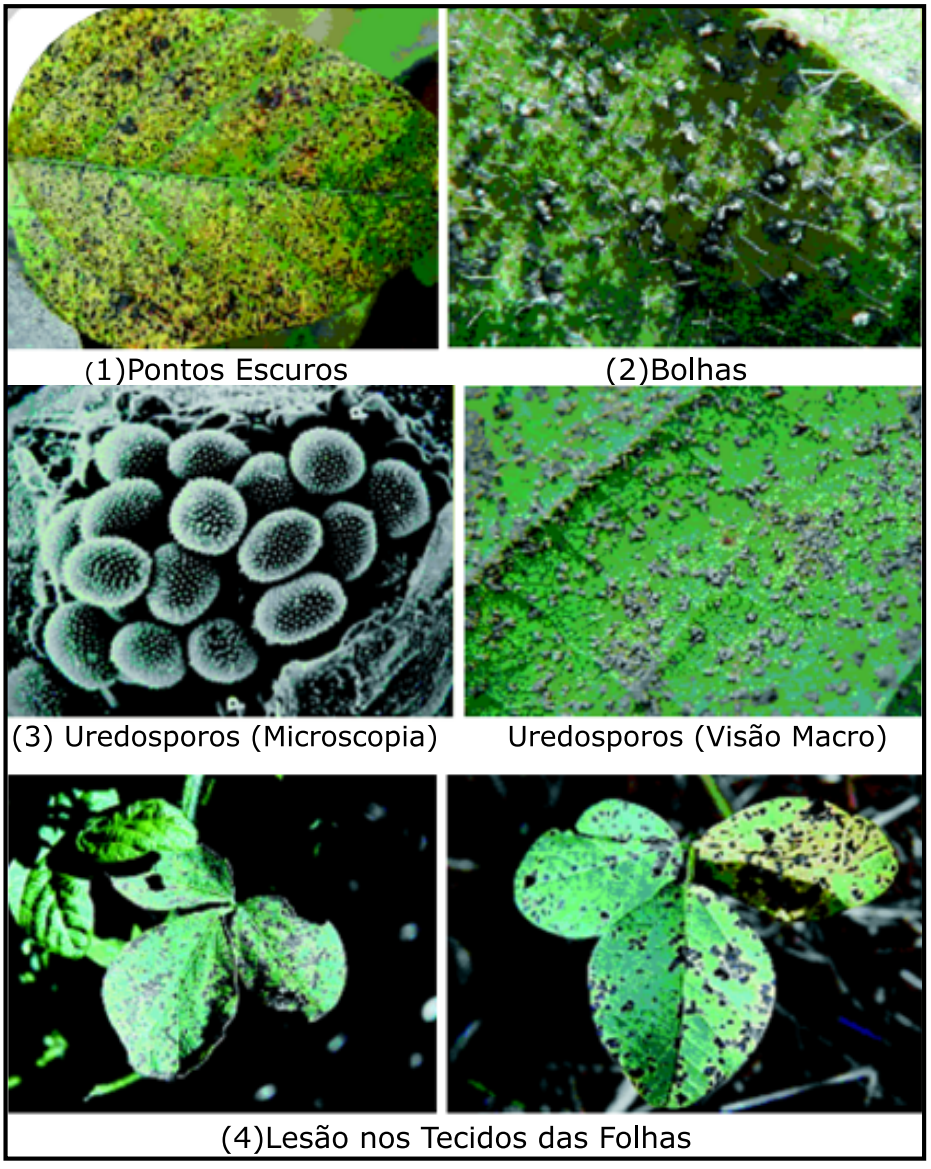
[1] Vazio sanitário é um período estabelecido, de no mínimo 60 dias, em que treze estados e o Distrito Federal adotaram para a redução da sobrevivência do fungo causador da FAS durante a entressafra. A medida tem por objetivo atrasar a doença na safra com a diminuição dos esporos presentes no ambiente(EMBRAPA FERRUGEM SOJA, 2020).

[2] Calendarização da semeadura da soja: trata-se de uma medida estabelecida para determinar uma data limite para que se possa plantar a soja, a fim de reduzir a aplicação de fungicidas ao longo da safra, o custo com os insumos e a resistência do fungo aos fungicidas (EMBRAPA FERRUGEM SOJA, 2020).

A favorabilidade para a ocorrência da FAS depende deste conhecimento sobre o ciclo de evolução da doença. Tal aspecto, envolve ações relacionadas ao manejo da cultura da soja, de forma a se poder minimizar perdas de produtividade, bem como a redução do uso de fungicidas.

Quanto à favorabilidade da FAS, Lelis e colaboradores destacaram que, para a identificação da doença é necessário analisar um conjunto de fatores, de maneira que cada um deles esteja representado por uma variável em seus diferentes aspectos, agregados à localização geográfica, bem como ao ciclo de desenvolvimento da planta (LELIS et al., 2009). Cada uma dessas variáveis traz a representação de um pequeno domínio do problema e o seu conjunto permite o entendimento da doença com maior detalhe desse domínio, o que possibilita a qualificação e o controle da doença (BAHRY et al., 2020).

Figura 7 – Ciclo de Evolução da Ferrugem da Soja



Fonte: Adaptado de (YORINORI et al., 2003)

Ao entender que o Brasil possui uma diversidade de regiões de cultivo da soja e que, para cada região, há condições climáticas diferentes, fator que influencia diretamente na severidade da FAS, Yorinori, Júnior e Lazzarotto (2004) relataram não ser possível fazer uma recomendação generalizada sobre o controle da doença que atenda a todas regiões.

Segundo Godoy e colaboradores, o processo de infecção da FAS depende basicamente da disponibilidade de água livre na superfície da planta, a duração do molhamento foliar de seis a doze horas, além de temperaturas de 15°C a 28°C. Assim, esses autores perceberam que, com a observação de tais variáveis climáticas junto às chuvas que ocorrem na safra, foi possível a avaliação da favorabilidade da ocorrência da doença nas áreas de cultura (GODOY et al., 2017).

De acordo com Nunes, Martins e Ponte (2018), as temperaturas extremas, tais como 20°C, considerada baixa, e 30°C, considerada alta, provocam a ausência ou atraso do desenvolvimento de epidemia da FAS, pois nestes extremos há perspectiva de redução em até 80%, quanto à produção de esporos do patógeno. Outro aspecto importante relatado por esse autor foi a fraca relação entre a severidade da doença da FAS e a temperatura para um período de trinta dias, ao sugerir que os valores máximos e mínimos atingidos dentro da faixa de variação, na maioria das regiões brasileiras, não foram fatores limitantes para o desenvolvimento epidêmico da doença. Por outro lado, a chuva, juntamente com a umidade decorrente do orvalho, afeta diretamente a infecção e a esporulação do fungo da FAS, provocando aceleração de epidemias, bem como sua disseminação regional.

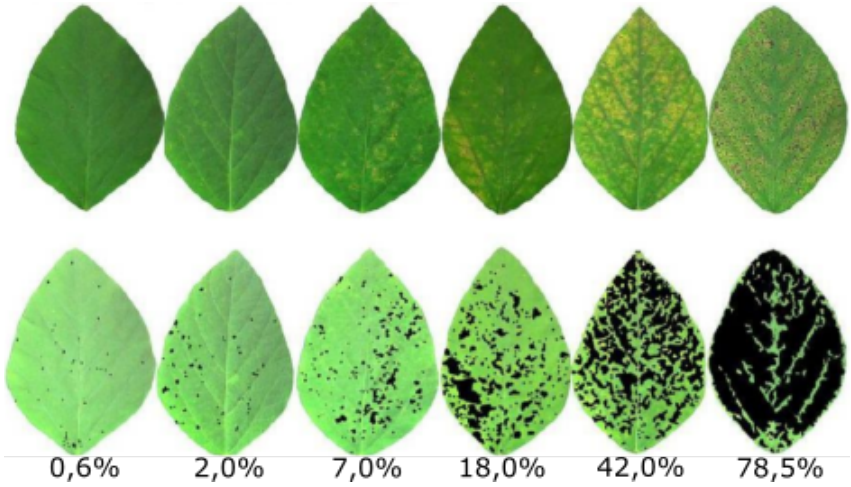
Conforme Beruski e colaboradores, as variáveis meteorológicas atuam como entrada para sistemas de alerta à FAS. Assim, esses autores consideraram as variáveis de duração do molhamento foliar e a temperatura do ar noturno. Tais variáveis têm efeito direto na ocorrência e disseminação da doença, com influência na taxa de infecção, o que compreende a germinação, a penetração do fungo e também o processo de esporulação. Nessa abordagem, os autores entenderam, como duração do molhamento foliar, a presença de água livre nas superfícies das plantas que, em condições de campo, deu-se pela chuva, neblina, irrigação, orvalho da atmosfera ou ainda evapotranspiração do solo. No trabalho, esses autores trata a duração do molhamento foliar como uma variável que não foi medida, em todas as estações meteorológicas convencionais ou automáticas, devido a ser de difícil medição, situação que ocorre devido à falta de padronização nos processos de medição. Na ausência dessa informação, optaram pela forma alternativa para sua obtenção, considerando métodos para a estimação do período de molhamento foliar. Nesse contexto, consideraram a variável umidade relativa (UR), que foi estimada, sendo em seguida realizada calibração ajustada às condições climáticas locais (BERUSKI et al., 2019).

Para reduzir erros na identificação visual sobre a severidade da FAS foi desenvolvida por Godoy e colaboradores uma escala diagramática, que permite avaliar diferentes níveis, a partir das percentagens preestabelecidas a seguir: 0,6; 2,0; 7,0; 18,0; 42,0 e 78,5%. Esta escala diagramática tem sido frequentemente utilizada na literatura (GODOY; KOGA;

CANTERI, 2006). Bahry e colaboradores utilizaram essa escala diagramática na análise de estratégias combinadas para o manejo da ferrugem asiática da soja (BAHRY et al., 2020). Nascimento e colaboradores também utilizaram em seus experimentos tal escala para avaliação semanal da severidade da FAS, fazendo uso de um método para captura de uredósporos, a partir de um coletor de esporos tipo cata-vento, inserido no campo na época de semeadura, permanecendo durante as safras e também nas entressafras (NASCIMENTO et al., 2012). A Figura 8 ilustra a escala diagramática.

De acordo com Nascimento e colaboradores, a precipitação foi apontada como principal causa de variação na severidade das epidemias de FAS. O estudo apresentado por esses autores mostrou uma alta correlação entre a precipitação e a severidade da doença. A abordagem apresentou que a precipitação provoca a liberação dos uredósporos do fungo *Phakopsora Pachyrhizi*, por meio de gotas de chuva, ou seja, pelo impacto dos respingos nas folhas. Foram apresentados também, nesse estudo, relatos de experimentos que indicaram uma relação direta entre a duração do período de molhamento foliar, o aparecimento dos primeiros sintomas da FAS, assim como a relação de dias chuvosos e a propagação da doença (NASCIMENTO et al., 2012).

Figura 8 – Escala Diagramática da Ferrugem Asiática da Soja



Fonte: Adaptado de (GODOY; KOGA; CANTERI, 2006)

Como abordagem em formato de estudo de caso, relatado no estado de Minas Gerais pelo Centro de Previsão de Tempo e Estudos Climáticos (CPTEC), a partir dos dados de quatorze estações meteorológicas da plataforma de coleta entre 2004 a 2006, foram identificadas condições favoráveis ao desenvolvimento da FAS em dois diferentes modelos (LELIS et al., 2009). O primeiro modelo evidenciou o número de horas com umidade relativa maior ou igual a 90% e, no segundo, foi destacada a depressão do ponto de orvalho menor que 2°C. Para os dois modelos, a faixa de temperatura de trabalho foi de 18°C a 25°C, considerada como temperatura ideal para o desenvolvimento do fungo

causador da ferrugem asiática. Dessa forma, foi identificado que, nos meses de julho e agosto, houve as piores condições de desenvolvimento do fungo e que, em contrapartida, o período de outubro a abril apresentou as melhores condições para o desenvolvimento do mesmo.

De acordo com Bedin e colaboradores, as plantas com deficiência nutricional ficam mais suscetíveis ao ataque de patógenos do que aquelas que estão nutridas adequadamente. Essa abordagem tratou da importância da nutrição da soja com o cobre (Cu), associado a tratamentos fitossanitários. O cobre é um micronutriente essencial que, além de outras funções importantes na planta, atua na resistência à doença da FAS. As deficiências de cobre ocorrem em solos orgânicos alcalinos e em solos de textura arenosa. A reposição das quantidades de cobre, necessárias ao solo, é realizada por meio da aplicação desse nutriente. Outro fato também mencionado pelos autores foi a importância da aplicação do Silício (Si). Esse nutriente contribui com o melhoramento da estrutura da planta, aumentando a rigidez na parede celular, provocando folhas mais eretas, o que limita a incidência de patógenos na planta. Os resultados sugerem que a nutrição foliar com cobre, em conjunto com tratamentos fitossanitários com fungicidas, reduz a severidade da FAS e melhora o desempenho da planta (BEDIN et al., 2018).

De acordo com Nunes, Martins e Ponte (2018), no que tange ao uso racional de fungicidas, há necessidade da adoção de programas supervisionados de controle e monitoramento da FAS, sob as orientações de especialistas, os quais devem ser norteados por informações obtidas por dados preditivos da doença, bem como por dados obtidos também por observações, a fim de oportunizar melhores práticas de manejo. Adicionalmente, esse autor mencionou que modelos específicos podem ser aplicados para o monitoramento, de forma a relacionar informações meteorológicas, dados da cultura e da doença, uma vez conhecido o deslocamento, a deposição e a infecção provocada pelos esporos. Nunes e colaboradores também destacaram que outros fatores podem ser inclusos nos modelos, tais como as fontes de inóculo (e.g.contágio ou difusão), direção e velocidade do vento, temperatura, umidade relativa do ar, molhamento das folhas, intensidade da radiação solar e o estágio de desenvolvimento da cultura.

Ponte e colaboradores consideraram que ao se entender a importância do controle da FAS, nas diversas áreas produtivas espalhadas pelo mundo, passou a ser necessário o uso de modelos matemáticos em diferentes escalas, tanto espacial quanto temporal, para suporte à decisão visando a minimização dos riscos de epidemias (PONTE et al., 2006). Nesse sentido, o monitoramento das safras é um recurso importante para esta avaliação sobre a severidade da doença, a fim de se efetivar o seu controle. Esta avaliação depende não só das variáveis climáticas, mas também da avaliação de índices de reflectância das folhas sadias ou não, o que pode ser obtido tanto com o uso de câmeras com sensores adequados embarcadas ou não em drones ou satélites (ZAMBENEDETTI et al., 2007).

2.2.3 Doenças na Cultura da Soja e Diagnósticos Baseados em Uso de Visão Computacional

Cui e colaboradores apresentaram duas abordagens acerca da detecção da ferrugem da soja, considerando a severidade da doença. A primeira abordagem trabalhou com o ajuste de limiar manual e, a segunda, sem o ajuste manual. Entretanto, as imagens de entrada foram multiespectrais e capturadas, inicialmente, em RGB para posteriormente serem convertidas para HSI, de forma que fossem obtidas as coordenadas dos centroides de distribuição de cores da folha. Assim, a partir dos índices definidos, tanto para a Proporção de Área Infectada (RIA), quanto para o Índice de Cor de Ferrugem (RCI). Os métodos envolvidos no processamento das abordagens foram o filtro de mediana para a melhoria visual e também para redução de ruídos aleatórios. Esses autores também utilizaram técnicas de segmentação na expectativa de avaliarem a severidade da ferrugem, considerando representação com coordenada polar, de acordo com a distribuição de cores em diferentes situações relacionadas a gravidade da ferrugem pré-identificada por especialistas (CUI et al., 2010).

Uma abordagem para reconhecimento da doença da soja, de acordo com Pires e colaboradores, adotou técnicas com base no uso de descritores locais e *Bag of Visual Words*. Esses autores utilizaram os descritores *Speed Up Robust Features* (SURF), *Histogram of Oriented Gradients* (HOG), *Dense Scale-Invariant Feature Transform* (DSIFT), *Scale Invariant Feature Transform* (SIFT) e *Pyramid Histograms of Visual Words (PHOW)*. Os experimentos foram conduzidos sob um grande conjunto de imagens digitais, tanto em escala de cinza, quanto coloridas, no espectro visível, adquiridas em campos de soja localizados no Brasil. Os autores relataram que, para os experimentos, não houve necessidade de trabalhar com imagens hiperspectrais e, por essa razão, as imagens puderam ser adquiridas por hardwares tradicionais, como *smartphones*. Portanto, as imagens consideradas foram de folhas saudáveis e infectadas pelas doenças da soja: míldio, ferrugem TAN e ferrugem RB. Adicionalmente, o vetor de características processado pelos experimentos foi submetido à redução de dimensionalidade, via algoritmo de Análise de Componentes Principais (PCA), no espaço de duas dimensões (PIRES et al., 2016).

Conforme Shrivastava, Singh e Hooda (2017), no cenário de classificação de doenças de plantas da soja, houve a exploração de descritores baseados em textura e cor, para testagem de algoritmos, com o objetivo de resolverem o problema de falhas de detecção de doenças, quando utilizado métodos de recuperação das imagens. Os descritores avaliados por esses autores envolveram: (1) histograma de cores (HIST) e as características do histograma de cores decompostas em *Wavelet* (WDH) que, por definição, codificou as informações de cor da parte doente das folhas, cujo procedimento consistiu no cálculo para todas as imagens, sendo em seguida avaliada uma informação sobre média de valores; (2) classificação de fronteira / interior (BIC): Tratou-se da caracterização dos *pixels* de fronteira ou interior, cujo método viabilizou trabalhar com as cores quantizadas, dividindo as

mesmas no espaço RGB em quatro cores (4x4x4=64 cores), para a obtenção do descritor de dimensão mais baixa. Desse modo, cada *pixel* foi reconhecido como borda ou interior e, em seguida foi possível a geração de um histograma correspondente; (3) o uso do Vetor de Coerência de Cor (CCV) viabilizou comparar as imagens, a partir da densidade de cor. Foi efetuada a quantização e calculou-se os componentes conectados, classificando-os em coerentes, quando os *pixels* foram pertencentes a essa região contígua de cor e, quando não atendiam a este quesito, foram classificados como incoerentes; (4) o histograma de diferença de cores (CDH) viabilizou a codificação da aparência visual da imagem, permitindo sua recuperação. Como primeiro passo, a imagem RGB foi convertida no espaço de cores L*a*b e calculadas as duas características de orientação de gradientes, nas direções xx, xy e xy e suas derivadas parciais em x, y e xy. Então, realizou-se o processo de discretização da orientação em 18 direções, enquanto a quantização de cores foi feita nos canais L*a*b em 10x3x3 = 90 cores. Assim, os dois vetores de características foram calculados de acordo com as diferenças de cores locais, em uma vizinhança, sendo um vetor para uma orientação discretizada e o outro para cor quantizada; (5) o padrão binário local (LBP), conceitualmente, viabilizou calcular as informações de textura, localmente, para classificação, encontrando as diferenças locais e convertendo-as em binário, considerando o sinal da diferença. O LBP foi calculado para o canal da imagem, separadamente, concatenando-se todos os canais para a obtenção de um único vetor de características de dimensão; (6) o conceito de Padrão Binário Local Simétrico Quadrado (SSLBP) foi considerado um padrão único e a dimensão do vetor de características reduzida para 49 bins; (7) a fase angular localizada (LAP) descreveu as características de textura robusta ao embaçamento, dimensionamento e iluminação da imagem, sendo que em seguida esses autores calcularam, por meio do LAP, a magnitude e a fase da transformada de *Fourier* da vizinhança local de cada *pixel*, convertidas em formato binário e combinadas para a obtenção de um único padrão binário para cada *pixel*; (8) o histograma do elemento de estrutura (SEH) foi introduzido para aplicação de recuperação de imagem, codificando a cor e a textura da imagem, onde a cor foi representada pela quantização HSV em (8x3x3=72 bins) e a textura codificada para as estruturas básicas: horizontal, vertical, diagonal e quadrada para cada cor quantizada da imagem. A estrutura proposta para detecção de doenças da soja, pelo método de recuperação, consistiu na remoção do fundo, segmentação, extração dos vetores de características da região infectada pelas doenças, em cada imagem do banco de dados e também na imagem de consulta. Assim, uma pontuação de similaridade foi gerada entre a imagem de consulta e as imagens do banco de dados e que, de acordo com tal pontuação, as melhores imagens foram recuperadas, encontrando o tipo de doença presente na maioria das folhas, ao considerar uma votação por maioria.

Uma revisão abrangente da literatura, realizada por Dhingra e colaboradores, apresentou detalhamentos de métodos de processamento de imagens para detecção de doenças de plantas foliares, considerando uma grande diversidade de culturas. Entre tais abordagens,

houve destaques para as que trataram da cultura da soja. A primeira discorreu sobre um método que identificou doenças da folha da soja, por meio de imagens coloridas com fundo complexo, a partir de regiões salientes. O processo envolveu o uso do algoritmo *K-Means*, combinado ao uso do limiar do componente R do espaço RGB e o algoritmo de morfologia, para preenchimento de regiões e correções dos segmentos doentes da folha. A segunda abordagem discutiu uma técnica que identificou e estimou, de forma automática, o nível de severidade da doença em folhas de plantas. O detalhamento do método consistiu na conversão da imagem de RGB para os canais Y, C_b,C_r e, em seguida, em segmentos que utilizam o método de limiar. Foram destacados os parâmetros: nível de doença, índice de gravidade, índice de gravidade da doença e área infectada, sendo os três últimos responsáveis pela medição automatizada da gravidade e do nível da doença (DHINGRA; KUMAR; JOSHI, 2018).

Uma combinação de técnicas de processamento digital de imagens, apresentadas por Araujo e Peixoto (2019), também destacou os momentos de cor, Padrões Binários Locais (LBP) e modelo *Bag of Words* (BoVW), com o objetivo de detectar, de forma automática, doenças da soja, por meio de sintomas apresentados nas folhas. Os descritores de características considerados foram a cor, textura e características locais de manchas, cujas características extraídas compuseram o vetor de entrada para um classificador do tipo *Support Vector Machine* (SVM). O algoritmo de segmentação adotado foi o *K-Means*, em dois *clusters*, onde um *cluster* correspondeu à região da folha e, o outro, ao fundo. Na sequência ao processo de segmentação, as características foram extraídas e aplicadas ao algoritmo SURF, com a finalidade da obtenção das características locais para a construção do modelo BoVW.

Tetila (2019) apresentou um sistema de visão computacional que identificou as doenças foliares e insetos-praga, na cultura da soja, a partir do uso de imagens coletadas com VANT. A metodologia do sistema foi formada por cinco etapas: (1) aquisição de imagens, onde a altura para a coleta das imagens pelo VANT foi igual a 2 metros, a partir da plantação, por apresentar melhores resultados; (2) a segmentação fez uso do algoritmo SLIC *superpixels*, onde cada segmento de *superpixel* foi classificado visualmente, em uma classe específica para as doenças: ferrugem asiática, mancha-alvo, míldio, oídio, solo ou amostras de folhas saudáveis; (3) conjunto de imagens para treino e teste selecionados por especialistas; (4) extração de atributos das imagens de *superpixels* baseados nas características de cor, textura e forma, incluindo a última etapa (5) classificação. Um dos objetivos do experimento realizado, na abordagem desse autor, foi a análise de desempenho de cada característica visual, responsável por descrever as propriedades físicas da folha. Entretanto, a cor foi a característica mais significativa diante das demais, haja vista que as manchas das doenças provocaram diferentes cores na folha, como por exemplo: a ferrugem asiática foi caracterizada por pequenos pontos de coloração esverdeada a cinza-esverdeada, mais escuros que o tecido sadio da folha.

Segundo Manavalan (2020), o algoritmo *K-means* de aprendizado de máquina não supervisionado foi utilizado para segmentar a região de interesse das imagens de folhas de plantas para particionar objetos em *k clusters*, com base na similaridade, a fim de extrair características locais e globais para identificação da severidade da ferrugem da soja. O método SIFT foi aplicado, a partir de cores, para atuação em folhas de soja.

No estudo apresentado por Zagui e colaboradores, foi desenvolvida uma modelagem espaço-temporal, para simulação da ferrugem asiática da soja, baseada em um sistema difuso. A proposta combinou variáveis de entradas para o modelo de decisão, envolvendo presença do patógeno, a planta suscetível e condições ambientais favoráveis, tendo, como saída, a vulnerabilidade da região da doença. A presença do patógeno foi quantificada, usando uma equação de difusão-aderência, apropriada para o problema. As condições ambientais propícias foram determinadas, usando um sistema difuso com entradas de temperatura e umidade foliar. As análises foram desenvolvidas no ambiente Matlab® (ZAGUI et al., 2022).

Yu, Ma e Guan (2023) propuseram um método de reconhecimento de doenças foliares da soja, usando modelos tradicionais de aprendizado profundo (AlexNet, ResNet18, ResNet50 e TRNet50). O modelo desenvolvido por esses autores possibilitou o reconhecimento de doenças foliares de soja, com base no algoritmo aprimorado de aprendizado profundo (TRNet18), o qual apresentou o melhor resultado. Os autores utilizaram uma câmera digital para capturar imagens digitais de folhas de soja doentes, previamente identificadas, viabilizando análises individuais de folhas em estádios fenológicos de crescimento de R2 até R5.

Estes trabalhos revisados abriram caminho ao aprofundamento dos estudos complementares, como desenvolvido nesta tese, entretanto incluindo no modelo de decisão não somente as informações de imagens digitais relacionadas às folhas da soja, inclusive com diagnóstico precoce da presença da ferrugem asiática já estabelecidos, mas também com as informações climáticas para uma avaliação precisa sobre a presença da doença ou não e seu estágio de severidade na cultura.

2.2.4 Estruturação de Bases de Dados

No contexto agrícola, a organização e arquitetura dos dados, bem como seu armazenamento, suas diferenças estruturais e formato estão entre os principais desafios para a construção de sistemas computacionais. Tais desafios implicam na avaliação das complexidades, para atender à interoperabilidade, quando o objetivo é a elaboração de sistemas que possam operar, de modo eficiente e confiável, em rede Internet. A infraestrutura da Internet é, naturalmente, composta por uma arquitetura distribuída e em ascensão, a qual depende de premissas, como escalabilidade, heterogeneidade, transparência, segurança e tolerância à falhas (LU; HOLUBOVÁ, 2019).

Frente as tecnologias aplicadas às fontes de dados, de acordo com Alekseev e colaboradores, o banco de dados relacional permite a coleta rápida e ótima alocação de dados para assegurar a sua exaustividade, pertinência e coerência, durante operações DML (*Data Manipulation Language*). No entanto, os sistemas SQL (*Structured Query Language*) não foram projetados para efetuar análise eficiente e rápida em modelos multidimensionais, inclusive, quando se referiram às grandes quantidades de dados *Big Data*. Nesse sentido, os bancos de dados relacionais, assim como dados de *Bussiness Inteligence*, de sensores de campo, de séries temporais, de imagens de satélites, de Sistemas de Informações Geográficas (GIS), entre outras diversas fontes de dados, com suas diversificadas origens, alimentaram uma arquitetura de *Big Data*, ora com dados históricos, ora com dados em tempo real via web, de modo que estes dados pudessem ser analisados sob o contexto desejado, em busca de conhecimento (ALEKSEEV et al., 2016).

Com base na literatura, foram citados trabalhos com foco em agricultura, os quais envolveram soluções de tecnologia que trabalharam com fontes de dados heterogêneas, considerando as variadas opções de gerenciadores de bancos de dados relacionais (WANG; WU; LI, 2017) e não relacionais (SHAH; HIREMATH; CHAUDHARY, 2017), assim como o uso de plataformas de nuvem (RAJESWARI; SUTHENDRAN; RAJAKUMAR, 2017; NEVES; CRUVINEL, 2020; NEVES; CRUVINEL, 2021), com a possibilidade de reunir soluções híbridas e integradas, a fim de tratar problemas de pesquisa.

Em contrapartida, Ghiwari e colaboradores apresentaram uma abordagem para armazenamento de dados agrícolas, pré-reunidos, usando tecnologia IoT (Internet das Coisas), por meio do banco de dados de documentos *Couchbase NoSQL*. A solução foi projetada e implementada por esses autores, considerando a estrutura hierárquica da agricultura Indiana. Em seu levantamento bibliográfico, o autor apontou que 95% do universo digital continha dados não estruturados e que as técnicas de coleta de dados expandiram seu armazenamento por várias fontes, devido ao aumento diário e exponencial na geração dos dados (GHIWARI; SAMBREKAR; RAJPUROHIT, 2017).

Nesse contexto, foram considerados, na área agrícola, como dados não estruturados, os dados originados de dispositivos IoT, sensores (sonar, sísmico e magnético), sites, imagens, áudio, vídeo e textuais (periódicos, metadados, registros de saúde, documentos, etc). Esses dados não estruturados deveriam ser convertidos para dados semiestruturados ou estruturados para sua utilização. Então, Ghiwari, Sambrekar e Rajpurohit (2017) propuseram o uso de banco de dados NoSQL (*Not Only SQL*) para trabalhar o processamento dos dados, com seus vários formatos, a partir do uso da técnica *MapReduce*, de maneira que o algoritmo *MapReduce* fosse aplicado nos dados não estruturados, para convertê-los em formato de dados semiestruturados e estruturados. A fim de lidar com uma grande quantidade de informações, foi utilizado o *Couchbase Server*[3]. Adicionalmente, esses autores

[3] Couchbase: banco de dados NoSQL aplicado em nuvens distribuídas ou híbridas. Oferece recursos de versatilidade, desempenho e escalabilidade. Couchbase Cloud é um banco de dados como serviço (DBaaS) utilizado para aplicativos de missão crítica (COUCHBASE, INC., 2021).

utilizaram os formatos CSV, XML e Texto.

2.2.4.1 Soluções Aplicadas às Fontes de Dados *Big Data*

No contexto de utilização de dados *Big Data*, com a finalidade de atender às demandas agrícolas, o rápido desenvolvimento das tecnologias, impulsionado pela Internet das Coisas, tem proporcionado, à agricultura, uma expressiva geração de dados originados de sensores, máquinas, instrumentos de medição, satélites, veículos aéreos não tripulados (drones), robôs autônomos e softwares em geral, justificados pela automação dos processos agrícolas, devido às necessidades de aumento da produção de alimentos.

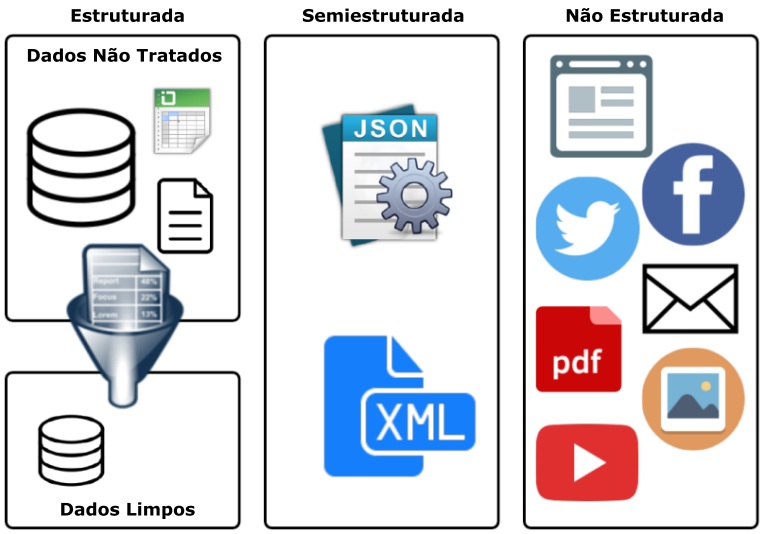
Com esse massivo volume de dados gerados, foram previstas, no setor, infraestruturas de grande porte, para garantir o armazenamento e processamento desses dados. Tais recursos puderam ser providos pelos centros de pesquisa ou por meios externos, ou seja, em *datacenters* de terceiros, acessados via rede Internet.

A partir do entendimento de que dados de origem *Big Data* são dados heterogêneos e complexos e, também, de que a análise desses dados demanda a aplicação de métodos interdisciplinares, Begoli (2012) recomendou considerar, nesse caso, a aplicação de métodos de aprendizado de máquina, mineração de dados e análise estatística para a obtenção de bons resultados, assim como o uso de tecnologias que pudessem suportar o volume de dados a ser trabalhado.

Conforme Emmanuel e Stanier (2016), a aplicação do *Big Data*, na área agrícola, apresentou uma variedade de formatos de dados, devido à diversidade desse setor, de proporcionar informações de formas extremamente diversas e heterogêneas. Em outras palavras, os dados provenientes de bancos de dados relacionais ou *Data Warehouses* puderam ser estruturados, não estruturados como texto, vídeo, áudio ou semiestruturados, contendo informações de *web-logs*, *e-mails* ou *tweets*, conforme ilustrado na Figura 9.

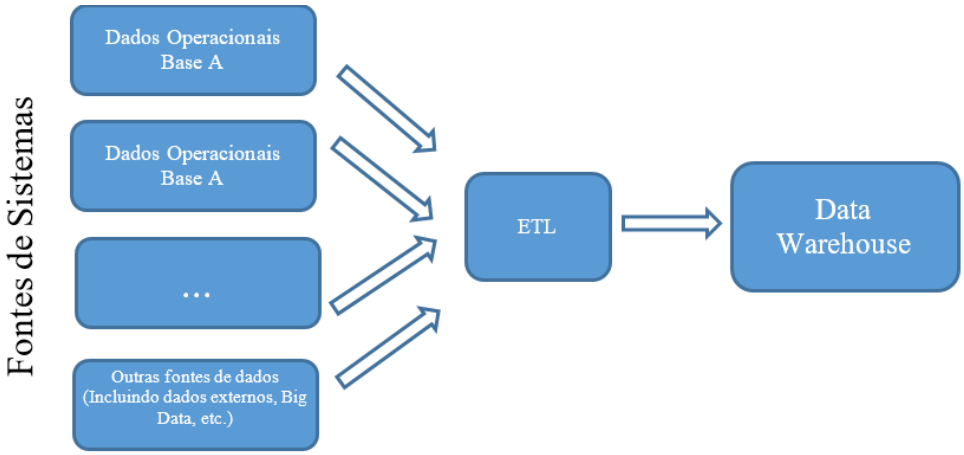
A abordagem de Jukić e colaboradores apresentou o *Hadoop* como solução para administrar a explosão de grandes volumes de dados, por meio de práticas existentes no *Data Warehouse* corporativo. Tais práticas fizeram parte do arsenal dos desenvolvedores de ETL (*Extract, Transform and Load*), a partir de uma variedade de fontes de dados. Com esses avanços tecnológicos, o *Big Data*, apesar de seus desafios, tornou-se outro tipo de conjunto de dados de origem para o *Data Warehouse*, em sistemas de auxílio à tomada de decisão. A Figura 10 apresenta uma arquitetura de um *Data Warehouse* em alto nível, onde é possível observar as várias fontes de dados de entrada, incluindo também os dados *Big Data*. A origem desses dados são decorrentes de múltiplas fontes, haja vista que o conceito de IoT esteve diretamente ligado aos setores de produção e logística agrícola. Também, os subsídios gerados pelas tecnologias de agricultura de precisão proporcionaram características comuns, ou seja, grande quantidade de informação, podendo ser unidimensional, multidimensional e de alta precisão, estruturadas, não estruturadas ou semiestruturadas. (JUKIĆ et al., 2015).

Figura 9 – Tipos de Dados



Fonte: Adaptado de (JOHN; MISRA, 2017)

Figura 10 – Arquitetura em Alto Nível do *Data Warehouse*



Fonte: Adaptado de (JUKIĆ et al., 2015)

Em contrapartida, Alekseev e colaboradores, e também, Amghar e colaboradores destacaram a tecnologia NoSQL, frente ao aumento de tarefas para processamento de *Big Data*, as quais, diante da necessidade do modelo de dados, tiveram que considerar a utilização de três classes de sistemas NoSQL: (1) sistemas orientados à coluna ou colunar; (2) sistemas chave-valor; (3) sistemas orientados à documento; e (4) sistemas orientados a grafo. Esses autores definiram as classes de sistemas NoSQL da seguinte forma: (1) colunar, onde se consideram sistemas de bancos de dados que armazenam dados em colunas, diferente dos bancos de dados tradicionais que armazenam em linhas. Neste caso, cada coluna é associada a uma chave, porém colunas semelhantes são armazenadas em

conjunto, ou seja, em famílias de colunas; (2) sistemas de chave-valor, os quais são sistemas NoSQL que implementaram um modelo de dados semelhante a um dicionário, no qual os dados são armazenados em pares de chaves e valores. Dependendo do sistema, os valores podem ser qualquer tipo de dados e são abordados por uma chave única; (3) na classe orientada a documento, onde os dados em formato de documentos podem estar na forma de JSON (*JavaScript Object Notation*), BSON (*Binary-encoded Serialization of JSON*) ou XML (*Extensible Markup Language*); e (4) em sistemas Orientados a Grafo, os quais consideram o armazenamento dos dados realizado em nós e arestas, cada um com as suas propriedades particulares. Esta última estrutura permite representar, em formato de grafo, as relações entre os dados, o que levou Alekseev et al. (2016), Amghar, Cherdal e Mouline (2019) a considerar esta categoria de sistemas mais adequada para o armazenamento de dados ligados.

Porém, diante do cenário de identificação de doenças de plantas, a partir da similaridade de sintomas, Kaur e colaboradores recomendaram um *framework* com suporte à *Big Data*, cuja solução se deu com base em evidências de dados históricos. O *framework* apresentado utilizou, como tecnologias, o sistema de arquivos distribuídos e escalável HDFS (*Hadoop Distributed File System*), assim como *Apache Hadoop*, ou seja, uma estrutura desenvolvida para armazenamento distribuído e processamento de grandes quantidades de dados em uma única plataforma, além do *Apache Hive*, em se tratando de um software de *Data Warehouse* que serviu ao propósito de facilitar as análises, as consultas e sumarização de dados de *Big Data*. Outras ferramentas, baseadas nas soluções de tecnologia da Apache Software Foundation (ASF) (2021), puderam ser utilizadas para estudos nessa linha de trabalho, a partir de dados *Big Data*. Segundo Kaur, com destaque para *MapReduce* (paradigma para escrever aplicações com recursos para processar grandes quantidades de dados em paralelo), *Apache Mahout* (paradigma livre para produção de algoritmos de Aprendizado de Máquina), *Apache Pig* (plataforma destinada para a execução paralela de fluxos de dados *Hadoop*), *Zookeeper* (interface simples para serviços operacionais para atuar em *clusters Hadoop* com as características de ser simples, confiável e rápida acessibilidade) (KAUR; GARG; AGGARWAL, 2016).

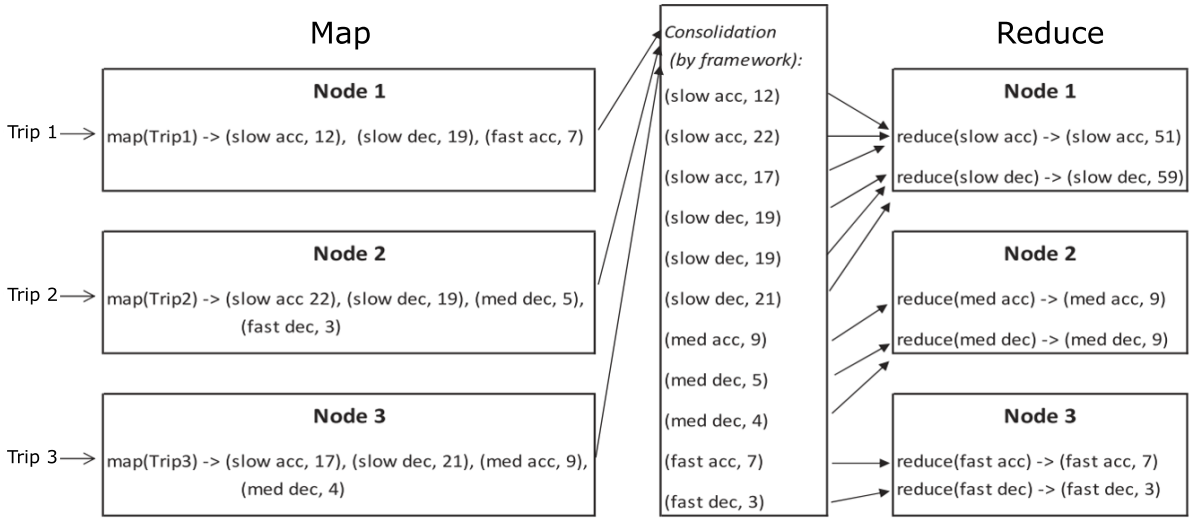
2.2.4.2 Soluções de Processamento Paralelo Aplicadas à Dados *Big Data*

Quanto ao processamento de dados *Big Data*, em arquitetura paralela, foram destacadas as abordagens consideradas mais relevantes.

Em Begoli (2012) foram apresentadas algumas dessas arquiteturas e estruturas de dados, que foram adotadas com frequência, tais como: (1) *Data Warehouse* e (2) NoSQL; (3) Processamento Massivamente Paralelo (*Massively Parallel Processing* (MPP) *Databases*): definido como uma técnica que emprega uma grande quantidade de unidades computacionais, sendo os núcleos, de diversas máquinas separadas, preparados para processar, de forma coordenada, a computação em paralelo. Ao considerar a aplicação do MPP, em

bancos de dados, as arquiteturas MPP podem fazer o armazenamento dos dados de forma distribuída e processá-los em paralelo; (4) Arquitetura de dados sem compartilhamento: termo aplicado à gestão de dados em arquitetura distribuída, em que cada servidor de gerenciamento de dados armazena e gerencia sua própria cópia de dados, em armazenamento local, aumentando a velocidade de acesso aos dados e evitando a sobrecarga associada aos dispositivos de armazenamento compartilhado em ambiente de rede; (5) sistemas de arquivos distribuídos e paralelos: trata-se de mecanismos fundamentais para o processamento em larga escala. Tais sistemas de arquivos distribuem os dados por vários servidores, para a obtenção de alto desempenho, acesso e processamento de dados paralelos; (6) *Hadoop*: consiste em uma estrutura de programação de código aberto, baseada na tecnologia Java, que suporta o processamento de dados *Big Data*, em ambiente distribuído. Também, utiliza um sistema de arquivo conhecido como HDFS, projetado para operar na maioria dos hardwares, incluindo o armazenamento de dados. Adicionalmente, possui gerenciamento tolerante a falhas, implementado em Java; (7) *MapReduce*: trata-se de um algoritmo do tipo "dividir para conquistar", para processamento paralelo de dados. O primeiro passo desse algoritmo consiste no mapeamento de uma parte dos dados, os quais são indexados por uma chave. Após o mapeamento, os resultados são reduzidos, utilizando-se do mesmo esquema de indexação por chave, para um conjunto de resultados, conforme ilustra a Figura 11; e, por fim, (8) armazenamento de dados orientados à coluna e compressão de colunas, onde além do armazenamento de dados também é considerada a compressão por colunas, que significa que os bancos de dados que trabalham com orientação à coluna podem armazenar os dados de forma compactada, ou seja, armazena a contagem da quantidade de dígitos repetidos, ao invés de armazenar o mesmo dígito de forma redundante.

Figura 11 – Processo *MapReduce*



Fonte: Adaptado de (JUKIĆ et al., 2015)

Liu e colaboradores propuseram um algoritmo de segmentação *fuzzy c-means* (FCMs) paralelo, baseado na plataforma de computação de memória distribuída *Apache Spark*, para *Big Data* de imagens agrícolas. Esse algoritmo foi adotado com frequência para segmentação de imagens agrícolas. A partir de uma imagem de entrada, nuvem de pontos foram geradas pelos autores, que foram particionados e armazenados em nós de computação diferentes, utilizando-se graus de associação de pontos de *pixels*. Assim, foram calculados os diferentes centros de *cluster*, de forma que a atualização desses centros fosse interativa e paralela, até que a condição de parada fosse satisfeita. Então ocorreu a restauração dos dados da nuvem de pontos, após o agrupamento, para que a imagem segmentada fosse reconstruída. A avaliação do algoritmo ocorreu com base na comparação da implementação paralela em *Hadoop*, em que a implementação em *Apache Spark* se mostrou com um aumento significativo de desempenho na ordem de 128%, frente à implementação em *Hadoop* (LIU et al., 2019).

Pereira e colaboradores apresentaram um modelo paralelo, baseado em semântica e no *Apache Spark*, relacionado à qualidade do solo, de maneira que a relação do manejo e qualidade pudesse ser avaliada por indicadores, classificados em quatro grupos: visual, físico, químico e biológico. A abordagem foi estruturada, em gestão de risco agrícola, e também para o suporte às tomadas de decisão. Os autores propuseram um algoritmo paralelo para distribuir imagens tomográficas de solo, em diversos nós, a fim de trabalhar, tanto o reconhecimento de porosidade do solo, quanto para as medições de densidade e compactação do solo, a partir dessas imagens. As tecnologias LoRa (*Long Range*) foram empregadas para operacionalizar os sensores de campo, enquanto o *MapReduce* foi adotado nas ferramentas digitais de suporte à redução de riscos agrícolas. Essa escolha ocorreu devido à crescente demanda de processamento de vastos conjuntos de dados e ao custo computacional dos algoritmos envolvidos. Adicionalmente, adotou-se o *framework Apache Spark* e as linguagens *Scala* e *Python*, nas implementações do modelo e o algoritmo. Com a utilização do algoritmo desenvolvido, foi possível obter ganhos de desempenho e redução do tempo para gerar os mapas de qualidade do solo agrícola (PEREIRA et al., 2020).

A abordagem de Alves (2020), Alves e Cruvinel (2023), na linha de trabalhos com aplicação de *Big Data*, com abordagem paralela no âmbito agrícola, apresentou um método de reconstrução tomográfica, aplicado às amostras agrícolas em alta resolução. O método fez uso da densidade espectral das projeções tomográficas de raios-x, como critério para minimizar o tempo de processamento e a obtenção de imagens digitais de boa qualidade, permitindo a escalabilidade. Assim, algumas etapas do processo foram paralelizadas, considerando a tecnologia *Apache Spark*. Para a execução do método desenvolvido, foi organizado um ambiente *Big Data* que contou com um *cluster*, instalado na plataforma *Amazon Web Services* (AWS), e uma pilha de tecnologias, além do *Apache Spark*: (1) *Amazon S3* (armazenamento dos dados); (2) *Amazon Elastic MapReduce* (EMR) (estruturação dos clusters); (3) *Amazon Elastic Compute Cloud (EC2)* (instâncias dos Nós do cluster); (4)

Hadoop YARN (gerência dos clusters); (5) Biblioteca *MRJob* (integração de aplicações *Python* com serviços de computação em nuvem / configuração dos *clusters*); (6) Linguagem *Python* (desenvolvimento do método de reconstrução de imagens tomográficas 2D e 3D). Os autores trabalharam com imagens 2D de sementes agrícolas, considerando um grande volume de imagens, na ordem de 242GB. Na etapa de avaliação, foram utilizadas as métricas SSIM (*Structural Similarity Index*), NRMSE (*Normalized Root Mean Square Error*) e PSNR (*Peak Signal-to-Noise Ratio*). Os resultados mostraram que o método possibilitou a redução, entre 28% e 38%, do número de projeções tomográficas, em cada amostra analisada, sem comprometer a qualidade das imagens reconstruídas, mostrando ser viável para a análise de grandes quantidades de amostras e também relevante para o processo de tomada de decisão.

Acrescentou-se ainda, de acordo com Tang, Aridas e Talip (2023), um método inteligente de processamento de dados que foi desenvolvido a partir do uso de tecnologia em nuvem, para monitoramento ambiental de áreas agrícolas. Esse método foi baseado no uso de componentes aprimorados do *Apache Spark* e utilizou o algoritmo denominado *FAST-Join* para otimização. Os resultados demonstraram que o método processou, com eficiência, dados de monitoramento ambiental de áreas agrícolas, apresentando uma redução significativa do volume de dados para consultas, bem como maior velocidade de cálculos, devido à computação em *cluster*. Destaca-se o potencial dessa abordagem para outros tipos de dados agrícolas, como imagens de pragas e doenças.

2.2.4.3 Infraestrutura para Armazenamento e Gerenciamento de Dados *Big Data* em Ambiente de Nuvem

Os serviços de nuvem têm sido considerados uma alternativa viável para os casos onde a opção por infraestrutura externa, tem justificado uma redução de custo, ou acesso às novas tecnologias, dada à facilidade de instalação de softwares e sua integração, bem como de soluções para tratamento dos dados, a partir de ferramentas pré-configuradas. Tais iniciativas, para uso da nuvem, contribuem para o desenvolvimento de novas soluções computacionais, viabilizando o rápido avanço das pesquisas na área agrícola e nas áreas computacionais, no que tange às etapas práticas de aplicação de metodologias, em especial, no gerenciamento de culturas, de produção, logística, previsão e tratamento de riscos agrícolas.

Nesse sentido, há bons produtos comerciais que oferecem recursos para se trabalhar com *Big Data*, no ambiente de nuvem. Foram destacados os serviços elásticos, ou seja, por demanda, para projetos com diversidade de dados caracterizados por diferentes tipos, formatos, fontes, assim como volume. No ambiente em nuvem, foram oferecidas arquiteturas distribuídas, paralelas, ou seu uso integrado, cujas características principais foram o alto desempenho, a alta disponibilidade, a redundância dos dados, a variedade de ferramentas automatizadas e integradas, com a possibilidade de configuração customizada e, princi-

palmente, com custo acessível para a maioria dos projetos (SOSINSKY, 2010; BASSOI et al., 2019; REHMAN et al., 2019).

Por meio da utilização de tecnologias de *Business Intelligence* (BI), providas por processos sistematizados e automatizados, para apoio à tomada de decisão, notou-se que a quantidade, a escalabilidade e a heterogeneidade são características que, minimamente, compõem a realidade das bases de dados, em nível mundial, nos mais variados segmentos, sejam esses governamentais, da produção industrial, ou agrícola. No entanto, é de suma importância considerar um projeto de *Big Data* cuja arquitetura possa atender requisitos, como processamento, performance e escalabilidade. Conceitualmente, uma arquitetura é escalável quando, ao aumentar seus recursos, como CPU, memória RAM e disco, observa-se um aumento de desempenho proporcional aos recursos adicionados (FLOREA et al., 2015).

Segundo Pokornỳ (2015), a adoção de uma abordagem que considere uma arquitetura de bancos de dados distribuídos, tais como uma arquitetura sem compartilhamento e particionamento horizontal, viabiliza uma menor latência, o que implica em qualidade no processamento dos dados. Nesse contexto, os bancos de dados NoSQL ofereceram maior desempenho, ao trabalhar em escala horizontal, em contraste com os bancos de dados relacionais, os quais foram concebidos para trabalhar em escala vertical.

Quando se faz opção pelo uso de infraestrutura de nuvem, o armazenamento de dados heterogêneos, via *Data Lake*, torna-se viável. O conceito de *Data Lake*, descrito por Miloslavskaya e Tolstoy, é um repositório para armazenamento massivo de arquitetura escalável e plana, onde cada elemento, o dado, tenha um identificador único, bem como um conjunto de etiquetas de metadados. Tal estrutura suporta dados em seu formato nativo. Logo, os sistemas de processamento podem receber dados em grandes quantidades, sem comprometer a estrutura. Um *Data Lake* é construído para trabalhar com dados estruturados, não estruturados e semiestruturados, com entradas rápidas, quase em tempo real, sendo acessível após sua concepção, o que difere de uma estrutura de um *Data Warehouse*. Adicionalmente, um modelo conceitual e semântico pode ser agregado ao *Data Lake*, para que o mesmo possa considerar a adição relacionada a contexto sobre os dados e, também, trabalhar suas inter-relações com outros dados. O uso de estratégias para uso do *Data Lake* pode combinar abordagens de bancos de dados, tanto com tecnologia SQL, quanto NoSQL, assim como para análise de dados online. Nesse caso, recomenda-se a análise com OLAP (*Online Analytical Processing*) e OLTP (*Online Transaction Processing*), para processamento de transações online (MILOSLAVSKAYA; TOLSTOY, 2016).

Um ponto importante a ser considerado sobre *Data Lake*, segundo Mehmood e colaboradores, é a integração de várias fontes de dados, por apresentar a necessidade do gerenciamento de metadados. No entanto, a falta de metadados, em dados brutos disponíveis no repositório, pode tornar complexos os processos de consulta e integração dos dados. Esse autor entendeu que, para resolver esses problemas, o *Data Lake* deve ser de-

senvolvido de modo a gerenciar de forma inteligente os metadados das fontes de dados heterogêneas, pois esse procedimento facilita a integração de dados heterogêneos. Ainda nesse trabalho, ocorreu a apresentação de uma discussão relacionada a evolução do modelo de dados, a partir do uso de inteligência artificial, a fim de tratar dados e metadados de diferentes fontes (MEHMOOD et al., 2019).

2.2.4.4 Processo de Preparação de Dados *Big Data* via Aprendizado de Máquina

A mineração de dados assumiu uma natureza multidisciplinar, recebendo diversas contribuições das comunidades de banco de dados, aprendizado de máquina, estatística, recuperação da informação, visualização de dados, computação distribuída e paralela. Notou-se que as principais contribuições foram provenientes das três primeiras comunidades. No contexto da mineração de dados, foram empregadas técnicas de banco de dados, aprendizado de máquina e estatística, cada uma apresentando considerações cruciais, como eficiência, eficácia e validade, respectivamente. Faceli e colaboradores consideraram o processo de mineração de dados como bem-sucedido, quando todos estes aspectos foram cuidadosamente abordados. Os autores também destacaram que a área de aprendizado de máquina experimentou um notável crescimento, resultando frequentemente no surgimento de diversas formas de utilização e aprimoramento dos algoritmos existentes, para atender às demandas específicas de diferentes áreas de aplicação. (ZHOU, 2003; FACELI et al., 2011).

Por outro lado, Wu e colaboradores afirmaram que, quando há disponibilidade de dados de diferentes naturezas, necessita-se de uma etapa de integração, para que os mesmos possam ser utilizados conjuntamente. Esses autores consideraram que processo de integração se resumiu na fusão de dados de fontes heterogêneas, a partir de diferentes representações conceituais, contextuais e tipográficas. Então, o conjunto de dados fundidos é diferente de um grande conjunto combinado. Também consideraram destaque para o conjunto de dados fundidos, em que há atributos e metadados que não podem ser incluídos, ao considerar os conjuntos dos dados originais. A integração foi utilizada, tanto na mineração de dados, quanto na consolidação de dados de recursos não estruturados ou semiestruturados, o que se referiu a representações textuais de conhecimento e que também puderam ser aplicadas ao conteúdo *rich media* (HAO et al., 2013).

Garcia e colaboradores destacaram aspectos na etapa de pré-processamento, quanto à importância e aplicação em técnicas de mineração de dados. O primeiro aspecto abordado relatou que o desempenho dos métodos aplicados, para a descoberta do conhecimento, dependia da qualidade e adequação dos dados analisados. O segundo aspecto tratou da apresentação de fatores negativos, associados aos dados, tais como o ruído, os valores ausentes, dados inconsistentes e supérfluos, bem como tamanhos muito longos em exem-

plos pré-estabelecidos em formatos específicos[4]. Tais fatores influenciaram fortemente os dados, tanto para a aprendizagem, quanto para a extração do conhecimento. O terceiro aspecto relevante comentado tratou do pré-processamento, etapa importante que viabiliza o processamento dos dados pelos algoritmos, ao entender que houve imposições para adaptações nos dados nos mesmos. Como quarto ponto observado, esses autores consideraram destaque sobre a etapa do pré-processamento, a qual apesar de sua importância, deveria prever que poderia consumir grandes quantidades de tempo de processamento, tomada por tarefas de limpeza, normalização, transformação, imputação de dados ausentes, integração dos dados, identificação de ruídos, seleção de características, seleção de instâncias e discretização. Os autores pontuaram também um problema recorrente no domínio da classificação, conhecido como classes desbalanceadas. O desbalanceamento ocorreu quando houve classes com uma percentagem muito pequena de dados, em comparação com as demais classes. Esse comportamento foi relatado por ser comum em ambientes *Big Data*, uma vez que contém milhões de instâncias (GARCÍA et al., 2016).

A abordagem de Sharma, Rathee e Saini (2018) atuou no cenário agrícola, a partir de dados *Big Data*, a qual aplicou técnicas de aprendizado de máquina para previsão de safra, em uma configuração híbrida. Assim, combinou o algoritmo SVM ao recurso de Otmização de Grew Wolf (GWO). A otmização GWO atuou na seleção do subconjunto ideal para a classificação que, nessa abordagem, encontrou a melhor relação das características, ou seja, uma relação semântica que fosse possível obter os melhores hiperparâmetros. Portanto, atuou na diminuição da dimensionalidade dos dados a serem aplicados ao algoritmo SVM, de forma a reduzir o erro na etapa de treinamento do classificador e aumentar a precisão da classificação. A etapa de pré-processamento, além de proporcionar a escolha dos melhores hiperparâmetros, resolveu o problema quanto à retirada de dados redundantes e ruidosos.

Em trabalhos que demandaram o uso de grandes quantidades de dados, especialmente com fontes de dados *Big Data*, despertou-se o interesse para o uso de técnicas de aprendizado de máquina, diante da gama de opções disponíveis (ABDUL-RAHMAN et al., 2021; LIAKOS et al., 2018; ANG; SENG, 2021). Em uma linha conceitual, Faceli et al. (2011) descreveram que as técnicas de aprendizado de máquina e seus algoritmos são diretamente afetados pelo estado dos dados, o que influencia diretamente na extração do conhecimento. Os conjuntos de dados podem apresentar problemas, por conterem dados com ruído ou incorretos, duplicados, inconsistentes ou ausentes. Nesse caso, foram utilizadas técnicas de pré-processamento para melhorar a qualidade dos dados, a fim de eliminar ou minimizar os problemas relacionados. A etapa de pré-processamento, quando bem executada, leva à construção de modelos mais fiéis à distribuição real dos dados, proporcionando a redução da complexidade computacional e, adicionalmente, facilita a configuração dos hiperparâmetros dos algoritmos, tornando-os adequados para a obtenção

[4] O dado pode também ser chamado de objeto, exemplo, padrão ou registro que corresponde a uma tupla formada pelos valores de características ou atributos (FACELI et al., 2011).

de melhores resultados.

2.2.4.5 A Qualidade dos Dados na Obtenção de Resultados Confiáveis

A qualidade dos dados é um fator muito importante para que se possa ter uma análise de dados com resultados confiáveis e evitar erros em tomadas de decisão (ALAOUI; GAHI; MESSOUSSI, 2019), tanto em bases de dados relacionais como em *Data Warehouses*, onde há modelos de dados bem definidos, a partir da necessidade de constantes tratamentos nos dados para eliminar incertezas, ruídos, dados nulos, dados incompletos, dados sem precisão, redundância desnecessária ou até mesmo para corrigir problemas que o modelo ou a aplicação pudesse gerar na base de dados. O problema que pode ocorrer relacionado à má qualidade dos dados gerados, quando utilizando grandes massas de dados *Big Data*, tomou uma dimensão significativa, requerendo cuidados dado à diversidade e à quantidade dos mesmo que normalmente é considerada e, principalmente, pelo fator da veracidade, um dos "Vs" considerados como relevantes por Jukić e colaboradores e também por Pipino e colaboradores (PIPINO; LEE; WANG, 2002; JUKIĆ et al., 2015).

Em Pipino, Lee e Wang (2002), destacou-se que a qualidade de dados deve ser considerada como um conceito multidimensional. Portanto, o autor apresentou uma lista de dimensões de qualidade, sendo que a mesma tem sido muito utilizada, segundo o autor, em diversas aplicações ou casos. As dimensões listadas foram: (1) acessibilidade; (2) quantidade apropriada de dados; (3) acreditabilidade; (4) completude; (5) representação concisa; (6) representação consistente; (7) facilidade de manipulação; (8) livre-de-erro; (9) objectividade; (10) relevância; (11) reputação; (12) segurança; (13) oportunidade; (14) compreensibilidade e (15) valor agregado. Esses autores concluíram que a utilização de um conjunto único de métricas não foi eficiente para avaliar a qualidade dos dados, pois a avaliação foi fruto de um esforço que deve ser realizado de forma contínua, a qual requer conhecimento dos princípios fundamentais das métricas, tanto subjetivas, expressas pela experiência dos especialistas (*stakeholders*), quanto das objetivas, obtidas por medições baseadas em um conjunto de dados.

Na abordagem de Alaoui e colaboradores, foi apresentado um estudo que atuou na análise de dados *Big Data*, em redes sociais, e propuseram a classificação de métricas de qualidade sobre os seguintes grupos: confiabilidade, usabilidade e pertinência, de forma que, para cada grupo, foram selecionadas um conjunto de métricas. Para o grupo de confiabilidade, as métricas utilizadas foram: (1) precisão; (2) completude e (3) singularidade. Já no grupo de usabilidade, as métricas consideradas foram: (1) transformação; (2) conformidade; (3) penalidade de armazenamento; (4) normalização e (5) referencial de integridade. No entanto, no grupo de pertinência, as métricas foram: (1) consistência; (2) credibilidade; e (3) atualidade. O autor colocou como destaque as muitas abordagens na área de qualidade de dados em *Big Data* que consideraram apenas duas características de *Big Data*: volume e variedade. Ressaltou que as demais características, conhecidas como

os “V” de *Big Data*, também deveriam ser levadas em consideração, quando a discussão pautasse a qualidade de dados (ALAOUI; GAHI; MESSOUSSI, 2019).

As dimensões da qualidade dos dados, abordado por Ramasamy e Chowdhury (2020), tem foco em *Big Data*. Nesse estudo realizado foi considerado um método de revisão sistemática, para o período de 2013 a 2019. Foram trabalhadas as dimensões da qualidade dos dados, bem como sua importância para a avaliação em sistemas que utilizam dados *Big Data*. Quando se afirmou que a qualidade não pôde ser avaliada, os dados tornaram-se duvidosos e, por consequência, as tomadas de decisão foram consideradas imprecisas. As três perguntas chave, realizadas para esta pesquisa, foram: (1) "Quais são as várias dimensões de qualidade de dados usadas na pesquisa para definir uma estrutura de avaliação de qualidade de dados em sistemas de *Big Data*?"; (2) "As dimensões convencionais ainda são aplicáveis aos sistemas de *Big Data*?" e (3) "Existem novas dimensões que emergem e são aplicáveis ao *Big Data* em específico?". A partir dessas premissas, os autores apresentaram um levantamento das dimensões-chave da qualidade dos dados relacionadas ao contexto *Big Data*, relacionando então seus respectivos autores, de acordo com o Quadro 1.

Quadro 1 – Dimensões-chave da Qualidade dos Dados em *Big Data*

Dimensão de Qualidade	**Definições**
Acessibilidade	"Acessibilidade e disponibilidade estavam relacionadas à capacidade do usuário em acessar dados de sua cultura, estado físico, funções e tecnologias disponíveis" (BATINI et al., 2015 apud RAMASAMY; CHOWDHURY, 2020).
Coesão	"Consistência, coesão e coerência referiram-se à capacidade dos dados cumprirem, sem contradições, com todas as propriedades da realidade de interesse, conforme especificado em termos de restrições de integridade, edições de dados, regras de negócio e outros formalismos" (BATINI et al., 2015 apud RAMASAMY; CHOWDHURY, 2020).
Confidencialidade	"Essa dimensão de qualidade determina se os dados certos estavam nas mãos certas. Os dados estão seguros?" (CATARCI et al., 2017 apud RAMASAMY; CHOWDHURY, 2020).
Credibilidade	"Os dados vieram de organizações especializadas de um país, área ou indústria. Experts ou especialistas auditaram regularmente e verificaram a exatidão do conteúdo dos dados. Os dados apareceram na faixa de valores conhecidos ou aceitáveis"(CAI; ZHU, 2015 apud RAMASAMY; CHOWDHURY, 2020).
Pedigree / Linhagem	"Esta dimensão ajudou a conhecer a fonte dos dados para que qualquer inconsistência fosse corrigida na fonte e não em quaisquer outras instâncias."
	continua

Quadro 1 – continuação

Dimensão de Qualidade	Definições
Legibilidade	"Também representada como clareza, simplicidade, facilidade de compreensão, interpretabilidade, compreensibilidade. Essa dimensão referiu-se à facilidade de compreensão dos dados pelos usuários"(BATINI et al., 2015 apud RAMASAMY; CHOWDHURY, 2020).
Analisabilidade em tempo real	"Às vezes, os dados precisaram ser analisados em tempo real. O tempo gasto para armazenamento pôde impactar a qualidade dos resultados" (ALAOUI; GAHI; MESSOUSSI, 2019 apud RAMASAMY; CHOWDHURY, 2020).
Redundância	"Redundância, minimalidade, compactação e concisão referiram-se à capacidade de representar a realidade de interesse com o uso mínimo de recursos informativos" (BATINI et al., 2015 apud RAMASAMY; CHOWDHURY, 2020).
Confiabilidade	"Confiança, incluindo credibilidade e reputação, captando quantos dados derivaram de uma fonte autorizada" (BATINI et al., 2015 apud RAMASAMY; CHOWDHURY, 2020).
Volume	"Esta dimensão da qualidade forneceu a porcentagem dos valores contidos no objeto de dados analisado em relação à fonte autorizada" (ARDAGNA et al., 2018 apud RAMASAMY; CHOWDHURY, 2020).

fim do Quadro 1

Fonte: Adaptado de (RAMASAMY; CHOWDHURY, 2020)

Adicionalmente, Taleb e colaboradores afirmaram que a avaliação da qualidade dos dados de fontes *Big Data* é uma fase importante e integrada à etapa de pré-processamento dos dados. Também, que a preparação dos dados está relacionada aos requisitos dos usuários e das aplicações. No entanto, a avaliação da qualidade é facilitada quando os dados são estruturados com seus formatos e esquemas bem definidos, cuja descrição colabora com o mapeamento dos atributos para as dimensões de qualidade. Estes atributos foram considerados pelos autores como essenciais para a escolha das métricas. Porém, quando não há dados estruturados, faz-se necessária uma fase intermediária, cujas tarefas podem ser de análise, mineração, detecção de um esquema, ou identificação de sua origem. Para estas tarefas intermediárias, esses pesquisadores lançaram mão de técnicas, tais como classificação, agrupamento, pesquisa, mineração ou filtros, disponíveis por meio dos algoritmos de aprendizado de máquina.

Por se tratar de uma tarefa de alto custo computacional, o processo de avaliação da qualidade dos dados não estruturados, originados de fontes *Big Data*, são elencadas algumas características que foram consideradas como fatores complicadores do processo, tais como o tamanho dos dados, a heterogeneidade, os vários tipos e formatos de dados, as fontes múltiplas e a escolha de Dimensão da Qualidade de Dados (DQD). A atuação na

escolha das DQD está diretamente relacionada à qualidade de sua fonte de dados. Nesse caso são sugeridas as tarefas: (1) definição de um projeto de qualidade, considerando dados *Big Data* não estruturados (*Unstructured Data Base* - UDB); (2) um conjunto de requisitos com DQDs padrão para começar a tarefa; (3) uma estratégia de amostragem para UDB, como extração de recursos, variáveis, atributos que caracterizam UDB e avaliação da qualidade e, finalmente, (4) a seleção das melhores técnicas, métodos e estratégias que extraem informações úteis do UDB ou as convertam em dados baseados em um esquema conhecido, conforme Taleb, Serhani e Dssouli (2018).

O desenvolvimento de um *framework*, para avaliar qualidade de dados aplicados à agricultura, foi um assunto pouco discutido pela comunidade científica, embora cada vez mais indispensável, haja vista as diversas aplicações de novos conhecimentos e tecnologias para este setor (BASSOI et al., 2019; MASSRUHÁ; LEITE, 2017; CRUVINEL, 2018; GASPAR, 2020; SILVEIRA; LERMEN; AMARAL, 2021).

Apesar da qualidade de dados ser um tema amplamente discutido e estudado, desde a década de 1950, o mesmo não ocorreu com a estruturação de *frameworks*, sendo entretanto baseado nos mesmos, que o assunto qualidade de dados ganhou maior expressão e sua evolução ocorreu (CAI; ZHU, 2015). Porém, o mesmo tem sido verificado em relação a área agrícola, onde não somente a qualidade dos dados, mas também a estruturação de *frameworks* qualificados são de grande importância, sendo portanto estratégico para o segmento da pesquisa e para o setor produtivo. Em meio a tais evoluções, notou-se também que a definição do termo qualidade de dados vem sofrendo mudanças, assim como os descritores de qualidade envolvidas para as diversas áreas de pesquisa, incluindo estatística, ciência da computação, geo-computação, rede de sensores e o uso da IoT nos processos de tomada de decisão, com destaque para aplicações *Big Data*.

O termo qualidade de dados se refere a um conjunto de características que os dados devem possuir, quanto à precisão, atualidade e aos desafios como definir, medir e melhorar a qualidade dos dados eletrônicos, armazenados em bancos de dados, *Data Warehouses* e sistemas legados (HASSENSTEIN; VANELLA, 2022).

Neste contexto, o nível de qualidade adequado para apoiar corretamente o processo de tomada de decisão na cadeia de valor agrícola passou a ser entendido como de fundamental importância, quer relacionado ao fator produtividade, quer relacionado ao fator sustentabilidade ambiental. De fato, a qualidade dos dados também se relaciona com a metrologia, conceitos físicos dos fenômenos, aspectos práticos de medições e análises, bem como conhecimentos. Portanto é possível considerar: (1) unidades de medida, padrões e métodos; (2) calibração, erros de medição, avaliação periódica; (3) infraestrutura: sensores, instrumentos de medição, redes e topologias, conectividade, arquiteturas de computadores e seus algoritmos; (4) precisão e completude nos modelos de decisão; (5) inspeção e técnicas, ou seja, verdade de campo e (6) aspectos de projeto, fabricação (medidores de todos os tipos e fusão de dados).

2.2.5 Pré-processamento e Processamento de Imagens Digitais

Quanto às abordagens que tratam o pré-processamento e processamento digital de imagens, evidenciou-se as principais técnicas que colaboram para estudos de doenças de plantas e, em especial, à ferrugem asiática da soja.

De acordo com Filho e Neto (1999), a etapa de pré-processamento pode ser entendida como o processo de aprimoramento da imagem coletada na etapa de aquisição, com o objetivo de melhorar a qualidade da mesma, preparando para as etapas subsequentes. As operações a serem executadas no pré-processamento da imagem, devem ser condizentes com os objetivos a serem atingidos, o que envolve uma melhor definição e em contraste, bem como uma melhor relação Sinal/Ruído (SNR), entre outros. As etapas de pré-processamento de imagens de sensoriamento remoto podem ser feitas trabalhando os aspectos de calibração radiométrica da imagem, ou efetuando correções de distorções geométricas provocadas por satélites, câmeras ou inerentes ao modelo da Terra, ou remoção de ruídos, fatores esses que ocorrem em processos de extração de informações, ou pela delimitação da área de estudo.

Ao considerar as aplicações de pré-processamento, em diferentes áreas de atuação do Processamento de Imagens e Sinais (PIS), notou-se que algumas operações foram comuns na maioria dos casos e que houve uma variação de entendimento de quais operações ou processos podem ser executados. Nesse contexto, de acordo com Filho e Neto (1999), ficou evidenciado que os processos de pré-processamento podem ocorrer, fazendo uso de processos de filtragem, realce, suavização de imagens, morfologia matemática e compressão de imagens.

Sonka e colaboradores expuseram que as etapas de pré-processamento podem ocorrer por etapas de transformações de brilho de *pixel*, transformações geométricas, pré-processamento local ou restauração de imagem (SONKA; HLAVAC; BOYLE, 2014).

Em contrapartida, Krig (2014) explorou o pré-processamento, com o foco em visão computacional e apresentou ajustes na taxonomia relacionada às operações de ponto, linha, área, algoritmos e conversão de dados. Inserido nesse contexto se encontram os algoritmos de métodos de pré-processamento de imagens que envolvem códigos puramente seriais. Como exemplo esses autores consideraram dados de sensores digitalizados e em formato de variáveis declaradas como tipo inteiro, os quais puderam ser convertidos para o tipo ponto flutuante ou para cálculos geométricos, ou ainda para conversões nos espaços de cores. Neste cenário de realidades, as conversões de dados são partes significativas do pré-processamento, em muitos casos, de imagens digitais.

As aplicações que envolvem o uso de imagens RGB, bem como seu pré-processamento, são muito diversificadas e podem ser apresentadas de diversas formas, conforme as necessidades. Na abordagem de Anagha e Baskar (2021), discutiu-se um método para a detecção automática de histogramas e extração de informações. A metodologia apresentou que a extração da informação foi feita via histograma, utilizando os dados dos eixos horizontal

e vertical, assim como o padrão do gráfico e o título. A extração das informações ocorreu por meio de um detector de linha, via Transformada de *Hough*, e pela aplicação de operadores morfológicos para identificar, de forma automática, a frequência dos dados presentes no histograma. O pré-processamento faz a conversão da imagem RGB em tons de cinza e, em seguida, permite converter o resultado em uma imagem binária e, com o objetivo de tornar o primeiro plano mais claro, considera o negativo da binarização. Como etapa final do pré-processamento, é realizada a detecção de bordas do histograma.

Na abordagem de Minz e colaboradores, no que tange aos processos de colorimetria de alimentos, evidenciou-se, na etapa de pré-processamento, o uso de imagens RGB e apresentou o desenvolvimento de um algoritmo de recorte de imagens para fins de análise em alta definição (HD). Esse algoritmo pode extrair alguma parte específica da imagem ou alterar sua proporção. Este recurso foi usado, principalmente, para remover detalhes indesejados ou considerados não essenciais na imagem. A implementação da solução ocorreu em duas etapas, sendo, a primeira, uma matriz gerada para funcionar como região de interesse e, a segunda etapa, uma matriz de amostra para medição de cor. O algoritmo apresentado por esses autores permitiu a análise rápida de imagens em HD, reduzindo o custo computacional, de forma a eliminar a necessidade de um sistema de processamento mais elaborado e de maior custo (MINZ; SAWHNEY; SAINI, 2020).

Ainda na análise de imagens RGB, a pesquisa de Monteiro e colaboradores identificou, por meio do processamento de imagens, os danos causados às sementes de soja, tais como sementes esverdeadas e enrugadas devido às variações de umidade e temperatura. No entanto, a etapa de pré-processamento dessa abordagem consistiu, primeiramente, em verificar qual das faixas de cores permitiram a separação das sementes de forma mais fácil, para uso em equipamentos de seleção por cor, nas unidades de processamento. Para tanto, esses autores desenvolveram um algoritmo que executou a verificação da faixa de cores. Adicionalmente, transformaram-se as imagens RGB, selecionadas em imagens de oito bits em escala de cinza contendo 256 tons possíveis, cuja escala considerou 0 como tom de cinza preto e 255 como tom de cinza branco. Para que as sementes possam ser identificadas de forma individualizada, foi aplicada a técnica de binarização, de modo a auxiliar na obtenção das características, tais como a área projetada no plano, perímetro e contagem de *pixels* em cada uma das regiões identificadas, conforme o interesse de análise. Após a etapa de pré-processamento, foram aplicadas as técnicas de segmentação *threshold* e detecção de bordas. Para a técnica de *threshold*, cada imagem foi dividida em duas ou mais classes de *pixels*, a partir da imagem binária. Assim, as imagens foram analisadas *pixel* a *pixel*, para a obtenção do valor total de *pixels* em cada semente, separando o fundo e classificando como branco, de acordo com o objeto de interesse. Na segunda etapa, aplicou-se o detector de bordas para que fosse possível verificar as rugas de um tegumento[5]

[5] Tegumento é o revestimento externo das sementes. Fonte: https://michaelis.uol.com.br/. Acesso em 01/04/2021.

causadas pelas variações de umidade e temperatura (MONTEIRO et al., 2021).

Além do uso de imagens RGB, foram também utilizadas imagens multiespectrais que, segundo Sonka e colaboradores, geradas de forma que cada banda espectral fosse digitalizada de forma independente, representadas por uma função de imagem digital individualizada. Os comprimentos de onda foram utilizados em aplicações de sensoriamento remoto, em sensores aerotransportados, via VANT, e também com o uso do satélite *Landsat* 4, entre outros, que transmite imagens digitalizadas em bandas (SONKA; HLAVAC; BOYLE, 2014).

No contexto de aplicação agrícola das imagens multiespectrais, de acordo com Stempliuk e Menotti (2020), seu registro, ao longo do tempo, considerou-se a etapa de pré-processamento como importante para extração de informações relevantes nas imagens estudadas. Portanto, os autores propuseram estudar o registro, que consistiu no alinhamento de duas imagens de uma cena comum, deslocadas por *viewports*.

Adicionalmente, esses autores consideraram também que os sensores para a captação da imagem multiespectral foram montados em uma mesma estrutura de um VANT, dispostos um ao lado do outro e que, mesmo assim, puderam apresentar respostas diferentes. Antes do processo de registro, as imagens brutas extraídas pelos sensores foram pré-processadas, para calibrar a refletância, normalizar a radiação considerada no momento da aquisição da imagem e correção da lente.

Ainda no cenário agrícola, com aquisição de imagens via VANT, Shin e colaboradores trabalharam com imagens multiespectrais, nas bandas azul, verde, vermelho, borda vermelha e infravermelho próximo e, além disso, fizeram uso de um sensor de irradiância, ou seja, um sensor de luz de fundo (DLS), para calibração radiométrica. A calibração radiométrica, incluindo correção atmosférica, foi uma importante etapa de pré-processamento para obter informações biofísicas da refletância espectral.

A etapa de pré-processamento, que envolveu o procedimento de calibração radiométrica, foi descrita em sete partes. A primeira compreendeu a calibração radiométrica vicária[6] de uma imagem de referência, a partir de painéis de referência terrestres. Na segunda parte do procedimento, os valores de Números Digitais (DN) dos painéis de referência foram medidos e os coeficientes para calibração vicária foram estimados por meio de análise de regressão. A terceira parte consistiu na formação de uma rede de nós, onde cada imagem foi definida como um nó. A quarta parte do procedimento consistiu em encontrar o caminho ideal de uma imagem para a outra, por meio do algoritmo de *Dijkstra*. Em seguida, como quinta parte, os pontos de ligação entre as imagens, a partir do caminho ideal, foram processados para estimar os coeficientes para a calibração geométrica rela-

[6] Trata-se da caracterização espectral de objetos de referência localizados na superfície terrestre. Deve ser concomitante à passagem do sensor sobre o objeto para que possam ser comparados os dados oriundos do objeto caracterizado espectralmente em campo com aqueles coletados pelo sensor a ser calibrado. Neste caso vale-se do uso de um espectrorradiômetro portátil atuando nas mesmas faixas espectrais do sensor que se pretende calibrar (PONZONI et al., 2015).

tiva. Como sexta parte, os DNs das imagens foram convertidos em DNs equivalentes à imagem de referência, usando os coeficientes de calibração relativos e, então, convertidos em reflectância, usando os coeficientes de calibração vicários. Como última parte do procedimento, realizou-se um processo de mosaico geométrico nas imagens de reflectância, para gerar o mapa de reflectância em formato de mosaico (SHIN et al., 2020).

Segundo Ding e Zhang (2020), a tecnologia de imagem multiespectral pôde, efetivamente, estimar o conteúdo de nutrientes da cultura. O estudo apontou que a qualidade da imagem foi reduzida por fatores, como as condições de iluminação do campo e ondulação das folhas e sobreposição, o que impactou negativamente no modelo SPAD (*Soil Plant Analysis Development*) para a cultura do tomate. Os fatores de enfraquecimento foram atribuídos à três métodos de pré-processamento que influenciaram na iluminação, foram eles: a Correção Gama Auto-Adaptativa, *Retinex* Multiescala e Reconstrução de Reflectância. Dessa forma, o modelo SPAD foi construído a partir de 16 parâmetros de entrada, onde foram comparados e analisados os três métodos, a fim de melhorar a precisão do teor de clorofila do tomate, com base em imagens multiespectrais.

Um estudo apresentado por Moghimi e colaboradores teve, como objetivo, o desenvolvimento de modelos preditivos robustos para atuar em estimativas de nitrogênio, em cultura de videiras. Os modelos preditivos foram desenvolvidos por meio de técnicas avançadas de aprendizado de máquina, a partir de imagens aéreas multiespectrais de alta resolução, cujas características foram compostas por cinco bandas espectrais discretas, incluindo azul, verde, vermelho, borda vermelha e infravermelho próximo. A ferramenta desenvolvida, chamada de *Micasense Preprocessing* (MP), automatizou as etapas do processo de pré-processamento elencadas, pela calibração radiométrica e remoção da distorção de lente, para melhorar o alinhamento da imagem (*Unwarping*) e alinhamento de bandas. Na etapa de calibração radiométrica, esses autores converteram as imagens brutas (DN) em radiância, considerando fatores dependentes do sensor, tais como ganho, configuração de exposição e efeitos de vinheta[7]. A MP fez uso de informações embutidas no arquivo de cabeçalho das imagens para conversão de radiância, de modo que as imagens de radiância fossem convertidas para refletância e contabilizado o fator dependente do tempo, ou seja, a variação na intensidade da luz incidente. Assim, a MP converteu as imagens de radiância em refletância, usando irradiância de entrada, medida por um sensor de luz incidente de cinco bandas integrado à câmera. Além disso, a MP removeu as distorções das imagens, alinhou e cortou cada imagem no quadro comum, de acordo com as cinco bandas (MOGHIMI et al., 2020).

A imagem hiperespectral é uma combinação de duas tecnologias consolidadas que incluem os aspectos de espectroscopia e imagem, em que uma imagem é adquirida ao longo dos comprimentos de onda na região do NIR, para especificar o espectro completo

[7] O efeito de vinheta é apresentado como a redução de brilho, do centro para as bordas, causado por anteparo físico aos feixes de raios que entram obliquamente em um sistema óptico (SILVA; CANDEIAS, 2009).

de comprimento de onda de cada *pixel* do plano de imagem. A vantagem oferecida pela imagem hiperespectral é sua capacidade de caracterizar as propriedades químicas inerentes de uma amostra, observando sua distribuição no espaço (TIBOLA et al., 2018).

Segundo Nagasubramanian e colaboradores, a imagem hiperespectral pode capturar informações espectrais e espaciais em uma faixa mais ampla do espectro eletromagnético, incluindo as regiões visível e infravermelho próximo. Dessa forma, esses autores aplicaram a imagem hiperespectral para detecção precoce de sintomas de doenças. Diante do cenário apresentado pelos métodos atuais de fenotipagem e doenças de plantas, apresentaram aspectos predominantemente visuais, com características de lentidão, variações e ocorrências de erros humanos. O objetivo principal da abordagem desses autores foi atuar na identificação precoce da doença da podridão de carvão na soja. As imagens foram obtidas por uma câmera hiperespectral, instalada em um drone, e organizadas em cubos de dados hiperespectrais de caules de soja saudáveis e infectados com podridão de carvão, coletadas em diferentes pontos no tempo (NAGASUBRAMANIAN et al., 2018).

Ainda de acordo com Nagasubramanian e colaboradores, a extração dos espectros de refletância de cada *pixel* permitiu relacionar as mudanças nos valores de refletância aos sintomas da doença. Por se tratar de dados com alta dimensionalidade, aplicou-se um pipeline de análise para reduzir a dimensionalidade dos dados e a seleção de comprimentos de onda ideais, para a identificação da doença, em se tratando de um Algoritmo Genético (GA), como um otimizador, em conjunto com o algoritmo SVM, como um classificador para a seleção efetiva da bandas relacionadas à faixas de comprimento de onda. O modelo baseado em GA-SVM apresentou sucesso na seleção de bandas para classificação de imagens hiperespectrais detectadas remotamente.

Lu e colaboradores descreveram os avanços e aplicações de tecnologias em imagens hiperespectrais em um documento abrangente, cuja revisão se deu para o período de 1990 até maio de 2020, com foco em agricultura. O documento, em um primeiro momento, destacou as aquisições das imagens hiperespectrais, com origem satelital e também de VANTs. Pelo fato dos sensores hiperspectrais se configurarem como minoria, diante dos sensores multiespectrais nos satélites, o estudo apontou que o sensor *EO-1 Hyperion* foi o mais utilizado em aplicações agrícolas, com destaque no monitoramento de diferentes culturas, propriedades de solo e detecção de doenças. Esse sensor atuou nas faixas de infravermelho visível, infravermelho próximo e de ondas curtas, com uma resolução espectral de 10nm e uma resolução espacial de 30m. No aspecto da aquisição das imagens, os VANTs foram muito utilizados, considerando os modelos multi-rotores, os quais foram caracterizados por serem mais competitivos que os modelos de asa fixa. Ao considerar a aquisição das imagens por diferentes plataformas e sensores e que foram normalmente fornecidas em um formato bruto, comumente em números digitais, houve necessidade da realização da etapa de pré-processamento para correções atmosféricas, radiométricas e espectrais para recuperar informações espectrais precisas (LU et al., 2020; VERAMENDI,

2022).

Ainda de acordo com Lu e colaboradores, também foram descritas as etapas de pré-processamento e processamento das informações de imagens hiperspectrais. A etapa de pré-processamento revelou-se crucial diante da necessidade de correções radiométricas para recuperação de informações espectrais precisas, a fim de investigar características agrícolas de interesse, como propriedades da cultura e do solo. Os autores destacaram a importância da redução de dimensionalidade como processo essencial nessa etapa. Na etapa de processamento das informações espectrais, foram citados como os métodos mais utilizados: regressão linear, regressão avançada (Regressão por Mínimos Quadrados Parciais - PLSR), aprendizado de máquina e aprendizado profundo. Adicionalmente, foram empregadas técnicas avançadas de modelagem de transferência radiativa, exemplificadas pelos modelos PROSPECT e PROSAIL (LU et al., 2020).

2.2.6 Extração de Características e Aprendizado de Máquina (Classificação)

A etapa de classificação, presente em várias metodologias, tem por objetivo a divisão dos objetos em classes, podendo ser caracterizadas em abordagens tanto binárias, quando envolveram duas classes, ou multiclasses, quando a classificação considera a utilização de mais de duas classes.

Para a agricultura, os estudos dos classificadores foram iniciados com a análise da abordagem de Pires e colaboradores, que se propuseram a detectar doenças da soja. Consideraram, na etapa de classificação, a aplicação do classificador SVM, acompanhado do método de avaliação de validação cruzada estratificada em dez vezes, de maneira que as imagens do conjunto de dados foram particionadas em dez vezes e cada dobra deveria conter a mesma proporção de dados de cada classe. Dentre as dez dobras, uma dobra foi usada para testes e, as demais, para treinamento do classificador. Esse processo foi repetido dez vezes, sendo que, cada dobra foi usada somente uma vez. A taxa de classificação correta foi medida, considerando a média da execução das dez rodadas (PIRES et al., 2016).

No sentido de avaliar o melhor classificador para o problema de detecção foliar de plantas de soja, considerando o uso de visão computacional e processamento de imagens, diante dos três descritores de características utilizados, Shrivastava e colaboradores observaram que os classificadores *K-Nearest Neighbour* (KNN) e SVM obtiveram os melhores resultados de separatividade de classes, quanto ao uso do descritor BIC (Classificação de fronteira / interior) (SHRIVASTAVA; SINGH; HOODA, 2014; SHRIVASTAVA; SINGH; HOODA, 2015; SHRIVASTAVA; SINGH; HOODA, 2017).

Os estudos, para o classificador Rede Neural Probabilística (PNN), apresentaram os melhores resultados com o uso dos descritores HIST (Características do histograma de

cores) e WDH (Histograma de cores decompostas *wavelet*). Cabe ressaltar que cada experimento, nessa abordagem, foram executados dez vezes e a precisão final foi calculada por meio da média de todas as rodadas. Assim, os melhores resultados para os descritores avaliados, frente aos experimentos e considerando o número de exemplos de imagens para treinamento igual a 38 foram listados, conforme segue: (1) classificador SVM: o descritor BIC indicou, em média para o conjunto analisado, 90% de precisão; (2) classificador KNN: o descritor BIC também indicou o melhor resultado frente aos demais descritores, ou seja, 95%; (3) classificador (PNN): os descritores HIST e WDH indicaram igualmente o percentual aproximado de 92,5% (SHRIVASTAVA; SINGH; HOODA, 2017).

O estudo realizado por Dhingra e colaboradores se referiu a uma revisão abrangente quanto às contribuições e métodos de processamento de imagens na detecção de doenças foliares de plantas. As técnicas de classificação exploradas nas abordagens por esses autores foram mostradas na Tabela 1, organizadas de acordo com a técnica, cultura aplicada, doença, precisão e características (DHINGRA; KUMAR; JOSHI, 2018).

Na abordagem de Devaraj e colaboradores, foi utilizado o classificador *Random Forest* para a identificação das doenças *Alternaria Alternata, Antracnose, Bacterial Blight e Cercospora Leaf Spot*. O algoritmo de *Random Forest* também foi utilizado na abordagem de Ma e colaboradores. Entretanto, foram também utilizados os algoritmos SVM, Rede Neural Convolucional Profunda (DCNN) e *AlexNet* para a cultura de pepino para reconhecimento das doenças: Antracnose, Míldio, Oídio e Manchas alvo na folha (MA et al., 2018; DEVARAJ et al., 2019).

Araujo e Peixoto (2019) propuseram, como objetivo, a identificação de oito doenças da soja, foram elas: crestamento bacteriano, ferrugem da soja, fitotoxicidade do cobre, mosaico da soja, mancha-alvo, oídio e mancha marrom septoria. A metodologia apresentada identificou somente uma doença por folha. A etapa de classificação das doenças fez uso do classificador SVM, com função *Kernel* polinomial. Em se tratando de uma tarefa multiclasse, foi considerada também a aplicação do método "um contra um", o qual consistiu na classificação por pares para cada classe, em relação a todas as demais classes, sendo que cada par representou um voto para uma doença. Assim, o método de identificação levou em consideração a doença que obtivesse mais votos para a folha selecionada.

Ainda de acordo com Araujo e Peixoto (2019), a precisão da abordagem foi de aproximadamente de 73,1% a 77,5% com intervalo de confiança de 5%. A maior taxa de identificação observada foi para a doença da ferrugem asiática da soja, com valor de 88,4% a 92% e a menor taxa observada foi atribuída para a doença da mancha marrom septória, com valor de 23% a 42%. Essa abordagem obteve um resultado melhor em 17%, em relação a um estudo semelhante, realizado por Barbedo (2016), que também realizou os testes com a mesma base de dados de imagens, obtida por meio da EMBRAPA (Empresa Brasileira de Pesquisa Agropecuária), disponível em: <https://www.digipathos-rep.cnptia.embrapa.br/>.

Tabela 1 – Estudos dos Classificadores

Técnica	**Tipo de Folha**	**Doença**	**Precisão**
Back-Propagation Neural Network	*Phalaenopsis* (Orquídea)	Bacteriana (podridão mole bacteriana, mancha marrom bacteriana, *Phytophthora* e podridão negra)	89,6% com doença e 97,25% sem doença
Multilayer Perceptron	Folha de seringueira	*Corynespora*, olho de rã e *Collectotrichum*	95.50%
Self Organizing MAP (SOM)	Arroz	Mancha marrom da folha, crosta de arroz, podridão da bainha e mancha marrom	94.21%
Redes Neurais Artificiais	Doenças da folha e do caule	Queima precoce, leve algodão, bolor acinzentado, brancura minúscula e queima tardia	93%
Redes Neurais Artificiais	Plantas e folhas de frutas	Queimadura precoce, bolor algodoado, bolor acinzentado, queima tardia e alvura minúscula.	94.67%
Redes Neurais Artificiais	Mandioca	Mancha marrom	79,23% doente e 89,92% saudável
Redes Neurais Artificiais	Milho	*Holcus spot*	98%
Rede Neural Probabilística	Tabaco	Antracnose e olho de sapo	88.59%
Redes Neurais Artificiais	Jujuba	Ferrugem da jujuba, antracnose da jujuba, podridão branca da jujuba, doença da ferrugem da fruta jujuba, mancha de *Ascochita* de jujuba e vassoura de bruxa da jujuba	Ferrugem da jujuba (91%) *Antracnose* jujuba (89%) podridão branca da jujuba (94%) doença da ferrugem da fruta jujuba (84%) *Ascochita* mancha da jujuba (73%) jujuba bruxas vassoura (81%)
Back-Propagation Neural Network	Uva	Oídio e *Antracnose*	100% (usando apenas recursos de matiz)
Toolbox Matlab - Reconhecimento de Padrões e Redes Neurais	Melancia	Míldio e *Antracnose*	75% (míldio) 76,9% (antracnose)
Rede Neural Probabilística	Mandioca	Doença do Mosaico da Mandioca	91.46%
Forward Neural Network	Algodão	Míldio cinzento e crestamento bacteriano	94%
Back-Propagation Neural Network	Arroz	Praga bacteriana	100%
Adaboost	Cítrica	Cancro cítrico	88%
Adaboost	Arroz	Doença da mancha marrom, brusone e doença bacteriana	identificação (83,3%) e Reconhecimento (91,1% e 93,3%)
PSO melhorado baseado no método de aprendizagem por oposição	Quiabo	Doença do vírus do mosaico das veias amarelas	93.30%
Naïve Bayes	Quiabo	Doenças do vírus do mosaico da veia amarela	87%
SVM	Uva	Sarna e Ferrugem	97.80%
SVM	Cana	Anel de Cana, Ferrugem e Manchas Amarelas	80%
SVM	Alfafa	Mancha foliar e Ferrugem	80%
SVM (3 Combinações)	Trigo	Oídio, ferrugem, ferrugem e *Puccinia Striiformis* (Ferrugem Amarela do Trigo)	95.16%

Fonte: Adaptado de Dhingra, Kumar e Joshi (2018)

A tese apresentada por Tetila (2019) discutiu a identificação de insetos-praga e de doenças foliares na cultura da soja a partir de imagens originadas de VANTs (Veículos Aéreos Não Tripulados), capturadas de várias alturas de voo, sendo a altura de dois metros estabelecida por apresentar melhores resultados. Os experimentos dessa abordagem foram realizados com a aquisição das imagens de diferentes alturas, considerando o dossel da planta e condições reais de campo para a soja. A etapa de classificação das imagens deu-se por meio da comparação de abordagens tradicionais com os métodos de aprendizagem profunda, de ajuste fino e transferência de aprendizagem. Os resultados obtidos pelos experimentos foram: (1) O classificador SVM combinado com o método de segmentação SLIC (*Simple Linear Iterative Clustering*) *Superpixels* obteve 98,34% na identificação da doença foliar; (2) Arquiteturas de aprendizagem profunda, tais como a *Inception-V3, Resnet-50, VGG-19 e Xception* foram testadas, considerando também o mesmo método de segmentação aplicado ao experimento anterior, e os resultados mostraram que a acurácia obtida foi de até 99,04%. Além disso, as estratégias de ajuste fino, que também foram submetidas aos testes, resultaram na acurácia de 100% e 75% e alcançaram as maiores taxas de classificação, em comparação a outras estratégias. Cabe ressaltar que as estratégias de treinamento, no caso da utilização de ajuste fino, necessitaram de um tempo maior de treinamento, cuja justificativa apresentada é a de que, além de substituir e treinar o conjunto de dados, houve também a necessidade de ajuste dos pesos da rede neural pré-treinada, com o algoritmo de retropropagação. Assim, esses autores concluíram que os modelos de aprendizagem profunda com pesos computados de ajuste fino generalizaram de forma adequada suas operações para conjuntos de dados com imagens de doenças da soja.

Segundo Kuchipudi e Babu (2019), na etapa de classificação, as redes neurais de retropropagação foram aplicadas em imagens de folhas que envolveram diversas culturas, tais como feijão, algodão e flores (rosas), para a identificação de doenças de plantas que, como resultado, trouxeram a melhoria na atualização dos mapas de características auto-organizáveis (SOFM), utilizados para identificação das cores da doença na folha.

Nessa mesma linha, um estudo de revisão, desenvolvido por Sudhesh e colaboradores, também foi identificado, haja vista que as redes neurais de retropropagação foram utilizadas com sucesso no processo de classificação, quanto ao reconhecimento de doenças, com destaque na cultura de romã, com 97% de acurácia. Ainda nesse estudo, Sudhesh citou o uso de outros classificadores, tais como: *Five-Fold Cross Validation And Interaction Procedure, Fuzzy Real Time Technique, Fuzzy* KNN, Redes Neurais, SVM e que, de acordo com os resultados dos classificadores estudados, foi mostrado outro destaque para a cultura de arroz, com a aplicação do algoritmo SVM, com acurácia de 97,2% (SUDHESH; NAGALAKSHMI; AMIRTHASARAVANAN, 2019).

Ainda, segundo a revisão da literatura, Nisar e colaboradores apresentaram diferentes tipos de culturas e de doenças que foram discutidas, a partir das seguintes técnicas de

classificação: SVM, Redes Neurais Artificiais, *Naïve Bayes*, Algoritmo *AdaBoost*, Rede Neural Probabilística (PNN), Redes Neurais Convolucionais e Lógica *Fuzzy*. As maiores acurácia, de acordo com os experimentos, foram 95% *k-means clustering*, 93% LDA (*Linear Discriminant Analysis*), para a cultura de folhas de trevo, 93,33% - KNN, 91,10% - SVM e 79,5% - Classificador *Naïve Bayes*, para a cultura de Arroz, respectivamente (NISAR et al., 2020).

Silva e colaboradores propuseram identificar as doenças míldio, macha alvo, ferrugem asiática em plantações de soja, a partir de imagens capturadas com o uso de Veículo Aéreo Não Tripulado (VANT). A abordagem utilizou recursos de visão computacional e aprendizado de máquina, tais como a segmentação das imagens em textitsuperpixels, por meio do algoritmo *Simple Linear Iterative Clustering* (SLIC) e as técnicas SVM, J48, *Random Forest* e KNN. Os resultados mostraram que o algoritmo SVM e *Random Forest* se destacaram em desempenho e que se apresentou eficaz na diferenciação das doenças analisadas (SILVA et al., 2020).

Outro artigo de revisão abrangente, de acordo com Manavalan (2020), apresentou um estudo de técnicas de classificação adaptadas para identificação de doenças de grãos, tais como *Partial Least Squares Regression* (PLSR), KNN, *Principal Component Analysis* (PCA), SVM, *Ensemble Learning Classifiers*, Redes Neurais Artificiais e suas variações. Dada a diversidade e relevância desse estudo, reuniram-se os principais conceitos e detalhamentos das técnicas apresentadas, bem como os melhores resultados dos principais classificadores, conforme pode ser observado no Quadro 2.

Quadro 2 – Classificadores - Abordagem de Manavalan (2020)

Classificadores - Definições
***Partial Least Squares Regression* (PLSR):** Método de construção de modelos preditivos aplicado quando as variáveis são altamente colineares. Técnica que generaliza os recursos da PCA e de Regressão Múltipla. Utilizado também para prever o conjunto de variáveis dependentes a partir de variáveis independentes. **Vantangens:** (1) Os mínimos quadrados permitem que os resíduos sejam tratados continuamente e (2) Fornecer uma solução de máxima probabilidade; **Limitações:** Ser sensível a *outliers* e à tendência a dados sobreajustados.
***K-Nearest Neighbour* (KNN):** O algoritmo trabalha com semelhanças de características para determinação de valores de novas amostras. Desta forma, o valor recebido pela nova amostra depende da semelhança com o conjunto de treinamento. Observa-se a aplicação do classificador KNN na identificação de doenças em folhas de milho, assim como para as doenças do arroz com precisão de classificação de 89,23%. O método de avaliação de generalização do modelo foi utilizada por meio da aplicação do método de validação cruzada (*three-fold-cross-validation*). **Vantagens:** (1) Abordagem de simples aplicação e não paramétrica; e (2) Não demanda processo de treinamento, o que implica que novos dados podem ser adicionados sem falhas; **Limitações:** (1) Dificuldade em trabalhar com dados de alta dimensionalidade; (2) Sensibilidade a outliers, os quais degradam o desempenho; (3) Possui convergência lenta; e (4) Necessidade em fixar o número de vizinhos.
***Principal Component Analysis* (PCA):** Aplicou-se a técnica de PCA, como método de classificação, na identificação da brusone do arroz, a partir de imagens hiperspectrais com comprimento de onda de infravermelho próximo de 900 a 1700 nm. **Vantagens:** (1) Elimina recursos correlatos, melhorando a eficiência; e (2) Reduz o sobreajuste; *continua*

Quadro 2 – continuação

Classificadores

Limitações: (1) Requer escalonamento de recursos como padronização e normalização de dados e (2) O uso de variáveis independentes causa a diminuição do desempenho.

***Support Vector Machine* (SVM):** Técnica projetada para decidir o melhor limite de decisão para divisão do espaço N-Dimensional em classes para projeção de um novo ponto de dados na classe correta. O limite ideal de decisão é um hiperplano formado por uma seleção de pontos ou vetores extremos chamados de vetores de suporte.

Na identificação de doenças da folha do trigo foram aplicadas estratégias de combinação de classificação múltipla de empilhamento baseadas em SVM com precisão de 95,16%, a partir das características de cor, textura e forma.

Fez-se experimentos do SVM com funções de *kernel* linear, polinomial e *Radial Basis Function* (RBF) para identificação de doenças foliares como *septoria leaf blinght, mildew e frogeye*.

Vantagens: (1) Alto rendimento; (2) Elimina problemas de sobreajuste; e (3) Suporta transformações lineares e não lineares;

Limitações: (1) Dificuldade em fixar os parâmetros ideais; e (2) Para o grande conjunto de dados o processo de treinamento é lento.

***Ensemble Learning Classifiers*:** Os algoritmos de ensemble podem combinar algoritmos com as mesmas características ou de tipos diferentes para o reconhecimento de objetos. Observa-se a utilização dos algoritmos *Random Forest* e CART (Árvores de Regressão) para classificação de doenças de plantas, cujas características dos modelos preditivos são de encontrar uma variável com base em outras variáveis rotuladas. Nota-se também o uso de Regressão Logística e modelos CART para a previsão da severidade da doença, de acordo com dados como data do plantio, por exemplo. Os dados dos experimentos mostram que os modelos de regressão logística classificaram corretamente de 60 a 70% dos casos, enquanto os modelos CART classificaram corretamente de 57 a 77% dos casos.

Vantagens: (1) Eficiente para interpretar e compreender; (2) Estratégia de votação aumenta o desempenho; e (3) Robusto para *outliers*.

Limitações: (1) Enorme conjunto de dados aumenta a complexidade; (2) Previsões mais lentas e (3) Uso de muita memória.

Redes Neurais Artificiais (RNA) e suas variações: Entende-se que as redes neurais são sistemas paralelos e hierárquicos e sua criação influenciada por modelos biológicos de cognição, assim como o cérebro humano. No entanto, sua arquitetura é implementada como uma coleção de estruturas denominadas neurônios com características paralelas e não lineares. As RNAs são aplicadas na identificação de doenças de folhas de arroz e recomendada-se sua combinação com algoritmo genético para a identificação de mancha marrom da soja.

Para o diagnóstico de doenças em folhas de grãos utilizou-se a técnica *Perceptron* Multicamadas (MLP) que consiste em uma camada de entrada para interpretação do sinal e uma camada de saída para a interpretação dos dados, bem como as camadas ocultas que determinam a capacidade computacional do MLP. Esta técnica foi aplicada na identificação da ferrugem amarela do trigo com desempenho maior que 99% e para doenças de folhas de arroz o resultado foi de 88,56% de classificações corretas.

Adicionalmente, a técnica MLP para a identificação da doença de brusone e mancha marrom do arroz comprovou 89% de acerto na classificação. Em contrapartida, a técnica de retropropagagação (BPNN) é definida pelo processamento em rede neural que faz um ajuste fino dos pesos baseada na taxa de erro estimada da época anterior. Assim, o ajuste de peso torna o modelo robusto por meio da minimização das taxas de erro, levando ao aumento da generalização do modelo. O uso da técnica de BPNN para identificação da mancha marrom do arroz alcançou mais de 90% de taxa de reconhecimento da doença.

A técnica de rede neural com base radial (RBP) foi também explorada para a classificação de doenças de grãos. É composta por uma camada de entrada, uma camada oculta formada por neurônios cuja função de ativação é uma função gaussiana e uma camada de saída. O ponto central é representado pelos pesos que conectam o vetor de entrada na arquitetura RBF. A rede neural RBF, por meio de resultados experimentais, apresentou o resultado de 91,32%. Por fim, a rede neural convolucional (CNN) é composta, em sua arquitetura, de camadas convolucionais iniciais, camadas ReLU (Unidade Linear Retificada), camadas *Max Pooling* e camadas totalmente conectadas, sendo estas últimas camadas responsáveis por formar o módulo de classificação.

continua

Quadro 2 – continuação

Classificadores

Com base no conceito de aprendizagem por transferência, uma otimização utilizada em máquinas de aprendizagem profunda, o conhecimento é adquirido por um modelo pré-treinado por meio de um grande conjunto de dados para resolução de um problema. E, desta forma, o conhecimento da arquitetura pré-treinada é transferido para a resolução de problemas relacionados. Para o contexto das CNN's podem ser citados os modelos de arquitetura GoogleNet e Cifar10 utilizados para estudo de doenças de folhas de milho com precisão média de reconhecimento de 98,9% e 98,8% respectivamente.
=> **MLP - Vantagens:** (1) Usado como uma metáfora para rede neural biológica; (2) Possui mapeamento não linear de sinal de entrada e saída; e (3) Convergência em mínimos locais.
Limitações: (1) Muitos parâmetros; e (2) Descarte de informações espaciais.
=> **RBF - Vantagens:** (1) Alta confiabilidade; e (2) Convergência mais rápida.
Limitação: O processo de classificação leva mais tempo, pois o cálculo do RBF é feito em cada nó da camada oculta.
=> **BPNN - Vantagens:** (1) O ajuste dos parâmetros não é necessário, exceto o número de entradas; e (2) Ajuste fino dos pesos da rede neural devido a taxa de erro correspondente obtida na época anterior.
Limitações: (1) Propenso a dados ruidosos; e (2) Convergência lenta.
=> **CNN - Vantagens:** (1) O ajuste do parâmetro inicial evita mínimos locais; (2) Extração de características, de forma automática, sem intervenção humana; e (3) Permitir o compartilhamento de parâmetros, uma vez que um único filtro é usado por várias seções de uma entrada para construção do mapa de características.
Limitações: (1) Requer muitos dados de treinamento; (2) Não tem a capacidade de discriminar dados espaciais invariáveis; e (3) Requer sistema de GPU para um rendimento mais rápido.

fim do Quadro 2

Fonte: Adaptado de Manavalan (2020)

A dissertação apresentada por Brito (2020) considerou um método de classificação de sementes oleaginosas, a partir de imagens de alta resolução, originadas de reconstruções tomográficas. Os métodos utilizados para estabelecimento da classificação foram o SVM e *Naïve Bayes*, conforme a Figura 12 e Figura 13, respectivamente.

Figura 12 – Algoritmo SVM

Algoritmo para treinamento do classificador SVM

Input: Id, Vetor de característica, tamanho
Output: Vetor de característica
Criar o classificador
vetor_características_treinamento, vetor_características_teste, id_treinamento, id_teste ← divide(Vetor de característica, Id, tamanho)
Treinar o classificador SVM(vetor_características_treinamento, id_treinamento)
Testar o classificador SVM(vetor_características_teste)

Fonte: Adaptado de (BRITO, 2020)

Adicionalmente, Brito (2020) apresentou um estudo com resultados decorrentes da avaliação de classificadores aplicados para classificar sementes de girassol, pinhão-manso e soja, tanto sadias como com defeitos, presentes em imagens digitais, incluindo situações compostas pelo conjunto das mesmas. Os resultados obtidos mostraram a acurácia de 0,82, para o algoritmo de *Naïve Bayes* e, de 0,94, para o algoritmo SVM. Na execução dos experimentos, esse autor adotou a proporção de 80% de imagens submetidas para treinamento e 20% de imagens para testes para os dois algoritmos. Entretanto, a partir

dos resultados, o autor atestou que os dois classificadores apresentaram alto índice de acurácia, porém com variações quanto à precisão na classificação das diferentes sementes submetidas. O algoritmo *Naïve Bayes*, frente ao SVM, trouxe resultados melhores para as sementes de girassol sem defeitos, com precisão de 100%. Em contrapartida, o resultado do processamento do algoritmo SVM, quanto à precisão, para as sementes de pinhão-manso sem defeitos, também foi de 100%.

Figura 13 – Algoritmo Naïve Bayes

Algoritmo para treinamento do classificador Naïve Bayes
Input: Id, Vetor de característica, tamanho
Output: Vetor de característica
Criar o classificador
vetor_características_treinamento, vetor_características_teste, id_treinamento, id_teste ← divide(Vetor de característica, Id, tamanho)
Treinar o classificador Naïve Bayes(vetor_características_treinamento, id_treinamento)
Testar o classificador Naïve Bayes(vetor_características_teste)

Fonte: Adaptado de (BRITO, 2020)

2.2.6.1 Reconhecimento de Padrões

Com o objetivo de pautar tecnologias e métodos do reconhecimento de padrões na classificação de doenças de plantas em diferentes culturas, bem como seus desdobramentos e refinamentos para a cultura da soja, foram apresentadas as principais abordagens que trataram os conceitos de identificação e classificação das doenças foliares, sob a ótica do aprendizado de máquinas, acerca dos métodos mais recentes e utilizados, bem como suas inovações.

Inicialmente, a abordagem de Chaki e Parekh (2012) propôs o reconhecimento de padrões com base nas características de forma das folhas. A técnica consistiu na conversão de duas imagens para a forma binária, dado um valor de limiar apropriado, de maneira que estas seriam sobrepostas e, a imagem resultante, calculada pela operação lógica *AND*. Todavia, considerou-se que os *pixels* diferentes de "0" foram somados, e o resultado utilizado como característica. Portanto, ao analisar a soma da resultante, foi percebida a baixa similaridade entre as imagens, quando o valor da soma foi baixo e vice-versa para a alta similaridade. No processo de classificação, a imagem de teste foi submetida à comparação em relação a todo o conjunto de imagens de treinamento para cada classe. Na sequência, foi calculada a média resultante para cada classe. No entanto, a classe que apresentasse a maior média seria a responsável por classificar a imagem de teste submetida. A técnica, apesar de simples, mostrou-se promissora diante da comparação com as técnicas mais complexas de reconhecimento de forma, tais como Momentos Invariantes e *Centroid-Radii*.

Wang e colaboradores, descreveram um método para o reconhecimento de folhas de plantas, elaborado pela decomposição da imagem, em escala dupla, e pelo uso de descritores binários locais. O método foi desenvolvido por meio da técnica *Wavelet*, de levantamento adaptativo, associada ao uso de um grupo de filtros gaussianos, de escala variável, permitindo, desta forma, a decomposição da imagem em escala dupla. O processo consistiu em decompor a imagem da folha em várias sub-bandas, utilizando a técnica *Wavelet* de levantamento adaptativo. Cada sub-banda foi convoluída com o grupo de filtros gaussianos de escala variável (WANG; LIANG; GUO, 2014).

Um novo descritor de forma foi apresentado por Yousefi e colaboradores, definido como *Rotation Invariant Wavelet Descriptor* (RIWD), o que tratou de uma versão melhorada de *Invariant Elliptic Fourier Descriptor* (IEFD), desenvolvido para classificação de grãos de cereais, que atuou independente do tamanho e rotação do contorno da forma. No RIWD foram utilizados coeficientes *Wavelet*, enquanto no método IEFD foram utilizados os coeficientes de *Fourier*. Na abordagem, foi prevista uma etapa de pré-processamento, por meio do uso de operações morfológicas para eliminação de erros de borda, de maneira que o contorno da folha pudesse ser extraído após essa etapa. Na etapa de extração de recursos, outras características morfológicas e de textura foram concatenadas ao vetor de características, com o objetivo de melhoria de desempenho do processo de reconhecimento (YOUSEFI; BALEGHI; SAKHAEI, 2017).

Turkoglu e Hanbay (2019) propuseram os métodos *Region Mean-LBP, Overall Mean-LBP e ROM-LBP* tratando-se de versões melhoradas do método LBP (*Local Binary Pattern*), para o reconhecimento de padrões de folhas de plantas. Para o funcionamento desses métodos os autores consideraram a região e a média geral, em vez do *pixel* central para a codificação, fazendo o uso de filtragem. Assim, enquanto o método LBP converteu as imagens em tons de cinza, os métodos *Region Mean-LBP, Overall Mean-LBP* utilizaram os canais R e G das imagens. Porém, o método *ROM-LBP* se baseou na combinação de parâmetros obtidos em ambos os métodos, ou seja, de *Region Mean-LBP, Overall Mean-LBP*. Os métodos foram avaliados quanto à robustez contra ruídos sal e pimenta e gaussiano e seu desempenho foi medido pelo classificador ELM(*Extreme Learning Machine*).

Quanto à identificação automática de plantas, a partir da imagens de folhas, conforme o trabalho de Sachar e Kumar (2020), um procedimento bem estabelecido foi adotado para o reconhecimento de folhas, envolvendo a aquisição de imagem, pré-processamento, segmentação, extração de características e classificação. Nesse trabalho, as imagens RGB utilizadas como entrada na etapa de pré-processamento, foram convertidas em tons de cinza e em seguida binarizadas. Os ruídos foram removidos por filtros de média, mediana, *Laplaciano, Wiener, Wavelet Log Gabor, Sobel, Canny, Prewitt e Roberts*. Outro recurso aplicado, como pré-processamento, tratou da correção da faixa estreita de níveis de cinza, por meio da normalização, o que significou aplicar métodos de prolongamento de contraste para esticar a faixa disponível de nível de cinza.

Ainda de acordo com Sachar e Kumar (2020), a nitidez foi uma característica desejável nas imagens e foi tratada por meio de operações morfológicas, tais como erosão e pre-enchimento. A etapa de segmentação levou em consideração a região de interesse (ROI) nas imagens da folha, desejável quando houve a necessidade de separar a imagem em segmentos ou grupo de *pixels*, tais como separar objetos indesejáveis e remoção de fundo, por exemplo. Neste sentido, a extração de características das folhas foram baseadas na forma, cor, textura e venação.

No Quadro 3 foram descritos os diferentes métodos presentes na literatura, segundo a revisão de Sachar e Kumar (2020), para a extração de características. Os métodos abordados foram divididos em: forma (contorno, região), cor, textura (espacial, espectral, veia) e a combinação de forma e textura.

Quadro 3 – Extração de Características

Métodos
Forma: 1. **Contorno**: Morfologia básica (diâmetro, perímetro e excentricidade), *Curvature Scale Space* - CSS, *Fourier*, Fractral e Ângulos de Contorno Integral (ICA); 2. **Região**: Morfologia (área, casco convexo e fator de forma), Momentos de imagem (HU e Zernike), características locais (*Scale Invariant Feature Transform* - SIFT, *Histogram of Gradients* - HOG, *Local Binary Pattern* - LBP e *Speeded-Up Robust Features* - SURF).
Cor: 1. Momentos de Cor e histogramas de cor.
Textura: 1. **Espacial** (*Gray Level Co-occurrence Matrix* - GLCM, Fractal, LBP e SURF); 2. **Espectral** (Filtro Gabor, *Fourier, Wavelet e Curvelet*); 3. **Veia** (Morfologia, Representação em Grafo, Fractal e SIFT, Análise de Componentes Principais - PCA para extração de características de cor).
Forma e Textura: 1. *Fast Discreet Curvelet Transform* (FDCT).
fim do Quadro 3

Fonte: Adaptado de (SACHAR; KUMAR, 2020)

As doenças de plantas em diferentes culturas agrícolas, cujos diagnósticos puderam contar, ainda que parcialmente, com a utilização de técnicas do reconhecimento de padrões, consideraram características que foram decorrentes de reconhecimentos de padrões e suas propriedades, incluindo detalhes quantificáveis de regiões alvo da doença.

Baseado nessa afirmação, Abdu e colaboradores atribuíram, como prática comum, o uso da segmentação para a remoção de fundo da imagem e dos *pixels* verdes saudáveis das folhas, com o objetivo da obtenção da região de interesse (ROI). Nesse contexto,

foi apresentado o uso de algoritmos *soft computing* que se basearam na manipulação aritmética, junto ao uso do limiar dos canais de cores, combinando o conceito de sistema especialista humano, com a visão computacional e, também, técnicas de aprendizado de máquina para identificar, segmentar e quantificar, de forma automatizada, as colônias de lesões de doenças locais. Foi destacado, nesse estudo, a segmentação da região de interesse estendida automaticamente (EROI), de maneira que permitisse a incorporação de informações da progressão dos sintomas da doença. Entretanto, a região de borda foi estendida para cobrir as zonas difusas da imagem, referentes aos tecidos borrados, por meio do uso do limiar de homogeneidade de cor (ABDU; MOKJI; SHEIKH, 2019).

O reconhecimento de padrões em doenças da cultura do pepino foi estudado por Ma e colaboradores (2018), mediante o uso de uma rede neural convolucional profunda (DCNN), cujo processo de segmentação ocorreu com a combinação das características de cor e de crescimento da região. A característica de cor abrangente foi formada por três componentes: Índice de Excesso de Vermelho (ExR), componente H (do espaço de Cor HSV) e componente b* (do espaço de cor L* a* b*). Tal combinação foi motivada por eliminar a influência da iluminação desigual, dado que os *pixels* das regiões, quando influenciados pela luz, foram classificados, de forma equivocada, como manchas da doença. O método de segmentação utilizado na abordagem foi desenvolvido por Ma e colaboradores (2017) e dividido em dois estágios: o primeiro estágio envolveu um recurso abrangente de cores (*Comprehensive Color Feature* - CCF) e seu método de detecção. No segundo estágio foi aplicada a segmentação crescente de região interativa, para que a segmentação de manchas da doença pudesse ser obtida, a partir da seleção de sementes em crescimento no mapa das características de cores CCF (MA et al., 2017; MA et al., 2018).

Adicionalmente, Kuchipudi e Babu (2019) trouxe a definição de um conjunto de passos considerados básicos para a detecção de doenças de plantas, considerando os recursos do PIS, foram eles: (1) Aquisição de imagem: passo em que a imagem digital é capturada em RGB, com previsão de mudança deste espaço de cor; (2) Pré-processamento: aplicado para a suavização da imagem, bem como para a expansão do contraste. Nessa etapa as imagens foram convertidas em tons de cinza e, posteriormente, aplicadas à equalização do histograma para alocação das intensidades, com foco no realce das imagens das doenças de plantas; (3) Segmentação: aplicada por meio das técnicas de *Otsu, K-Means, Fuzzy C-Means* (FCM), *Kernel Based FCM*, seguida da mudança de imagens RGB para o modelo HSI, dividindo a imagem em vários pedaços com características similares; (4) Extração de características: consolidada pelo uso de cor, morfologia, textura e bordas. Quanto ao uso de cores, utilizaram os processos de Co-ocorrência de cor e de extração da cor da folha por meio do "H"(*Hue* - Matiz) do espaço de cores HSI, junto aos segmentos "B"(do espaço de cor L* a* b*); (5) Classificação: como última etapa do processo, foi aplicada a rede neural artificial via algoritmo de rede neural de retropropagação (BPNN).

Um estudo de revisão foi apresentado por Sudhesh e colaboradores, no qual foram

relatadas abordagens diversificadas que trataram as doenças de plantas, assim como suas deficiências nutricionais em diferentes culturas, fazendo uso de técnicas e suas combinações para identificação, detecção, reconhecimento, segmentação e classificação. Em destaque, sob a ótica da doenças de folhas de plantas, puderam ser elencadas as seguintes discussões sobre as culturas de alfafa, maçã e pepino, onde: (1) para diagnóstico de doenças de folhas de alfafa aplicou-se, como método de segmentação, o uso de redes neurais e, na etapa de classificação, o método *Scale Invariant Feature Transform* (SIFT). A avaliação foi realizada em cinco etapas repetidas, em cinquenta instâncias; (2) Para a abordagem que tratou da cultura de maçãs esses autores utilizaram o método de identificação de doenças, com a aplicação da técnica de *Wavelet*, para diagnosticar as particularidades das doenças quanto às características de cor, textura e forma, de maneira que características consistentes das folhas fossem adquiridas e efetuados os diagnósticos corretos. Além disso, no trabalho desses autores foi observado o uso do filtro de Sobel e da técnica de Contagem Homogênea de *Pixels* para classificação das doenças (SUDHESH; NAGALAKSHMI; AMIRTHASARAVANAN, 2019).

Outro estudo de revisão foi apresentado por Manavalan (2020), para a distinção de doenças detectadas nas folhas dos grãos de arroz, trigo, milho e soja. As técnicas de pré-processamento foram focadas para a eliminação de ruídos, ou remoção de objetos, por meio do uso do algoritmo morfológico e do filtro de mediana. Na etapa de segmentação, utilizou-se o algoritmo *K-Means*, baseado em limiar, isolando a região discrepante em diferentes agrupamentos, de forma que cada intersecção do histograma de grupo pudesse representar a relação de cores, a fim de extrair o local afetado pela doença. Esses autores ainda utilizaram outras técnicas para a segmentação, a saber: (1) limiarização; (2) *Clustering Fuzzy C-Means*; (3) algoritmos baseados em crescimento regional (*Seeded Region Growing* - SRG), e para em sua implementação original foi utilizada a distância euclidiana para identificar as áreas adjacentes. Porém, como aprimoramento da SRG, passou a ser empregada uma tabela de pesquisa bidimensional para rotulagem dos vizinhos, na realização da fusão das regiões. Quanto à extração de características, a revisão abordou os descritores de cor, textura e forma, extraídos para reconhecimento das folhas saudáveis ou afetadas pela doença, compondo assim o vetor de características que atendeu como entrada para a etapa de classificação. Ainda, esse autor considerou como técnicas para extração de características: (1) técnicas de extração de características morfológicas, onde a dilatação e a erosão foram utilizadas como estratégia para calcular a diferença entre uma imagem original e a respectiva imagem processada (BAI, 2015); (2) função de transformação SIFT, onde foi tratada a operação de um algoritmo de identificação de características locais em imagens de folhas de plantas, cujo funcionamento básico consistiu na localização de pontos-chave e, assim, permitiu a geração de informações quantitativas sobre as características(LOWE, 2004); (3) características orientadas por histogramas, onde foi utilizado o descritor criado por Dalal e Triggs (2005), baseado na avaliação de histogramas locais,

sendo que a característica estatística do histograma de gradiente foi obtida pela decomposição de imagens em uma matriz densa de regiões. Em seguida, calculou-se o histograma de gradientes orientados para cada região, normalizado pela sobreposição do contraste local das mesmas; (4) sumarização de descritores locais, tais como SURF, HOG e PHOW para identificação de doenças da soja, como o Mofo e Ferrugem (Tan e RB)(PIRES et al., 2016); (5) técnica com Matriz de Co-ocorrência de Nível de Cinza (GLCM), onde os autores sugeriram para extração de características e fornecimento de dados de entrada, tanto o uso de rede neural, quanto o uso do algoritmo SVM (CHANDRAPRABHA; BHARATHI, 2019); (6) Filtro Gabor 2D, onde se pode promover a extração de características Gabor (frequência e orientação) conjuntamente com características de textura e gradiente, aplicados para reconhecimento da severidade de doenças (HAN; HALEEM; TAYLOR, 2016); e (7) aprendizado profundo (*Deep Learning*)(LIANG et al., 2017).

Diante destes principais recursos apresentados é possível observar também o uso de métodos para a redução do conjunto de características, tais como: (1) Análise de Componentes Principais (PCA); (2) seleção de recursos baseados em grafos; (3) técnicas de otimização de características; (4) técnica baseada na teoria dos conjuntos brutos; e (5) fusão de características.

Na linha de identificação de doenças, Devaraj e colaboradores desenvolveram uma metodologia conduzida pelas etapas de aquisição da imagem no espaço de cor RGB; pré-processamento contemplando os processos de eliminação de ruídos, conversão de imagem em tons de cinza, execução de operações morfológicas e alteração do tamanho da imagem para fins de aumento do contraste; segmentação com adoção do método *K-Means* para particionamento das imagens em agrupamentos; extração de características com a utilização do método *Gray Level Co-occurrence Matrix* - GLCM e, por fim, a etapa de classificação (DEVARAJ et al., 2019).

A revisão da literatura proposta por Nisar e colaboradores, destacou o reconhecimento de doenças em folhas de plantas sob os aspectos do uso de filtros do domínio espacial, tais como o *Gabor* e *Wavelet*, para a etapa de pré-processamento das imagens; os métodos de *Otsu* e *K-Means Clustering*, para a segmentação e os métodos estatísticos, para extração de características: *Color Co-ocorrence Method - CCM, GLCM, Spacial Gray-Level Dependence Matrices - SGDM*, a partir de descritores de textura e cor aplicados (NISAR et al., 2020).

Neves e Cruvinel (2022), envolvendo inteligência computacional, desenvolveram um primeiro estudo para a identificação da ferrugem asiática em plantas de soja. Esses autores, utilizaram o processamento de imagens da cultura, infraestrutura em nuvem e recursos do aprendizado de máquina. O resultado alcançado mostrou ser possível agregar ao conhecimento especialista o suporte da inteligência computacional para o auxílio à decisão sobre a incidência ou não da ferrugem asiática em áreas de plantio da soja.

2.2.7 Fusão de Dados de Sensores para o Controle de Doenças de Plantas

O primeiro trabalho sobre fusão de dados de múltiplos sensores versou sobre uma revisão que foi realizada por Khaleghi, Khamis e Karray (2016). Nesse trabalho os autores definiram sob vários pontos de vista e aplicações a fusão de múltiplos sensores, como visão de integração de dados. Tal definição foi considerada com foco no processo de integração, utilizando a combinação real de diferentes fontes de dados coletados de sensores em um formato representacional único. Entretanto, outras definições foram também mencionadas por esses autores, entre elas: (1) para melhorar o funcionamento do sistema de fusão de dados, incluindo o planejamento e arquitetura de sensores; (2) para um processo que lida com a combinação de dados em uma representação interna, ou sobre ação coerente e consistente; (3) sobre processo multifacetado em vários níveis que lida com a detecção automática, associação, correlação, estimativa, combinação automática de dados e informações de várias fontes; (4) sobre dados fornecidos por uma única fonte, ou por várias fontes que podem ser aplicados em diferentes campos de atuação, incluindo sensoriamento remoto; (5) sobre fusão de dados e/ou informações (considerando métodos eficientes para transformar, de forma automática ou semi-automática, as informações de diferentes fontes e momentos, em uma representação que ofereça um suporte eficaz para a tomada de decisões humanas ou automatizadas, incluindo combinações decorrentes de processamento de sinais, teoria da inferência da informação estatística e inteligência artificial).

Conforme Khaleghi, Khamis e Karray (2016), o processo da fusão de sensores foi também utilizado em aplicações militares, baseado nos dados de entradas e saídas, produzidas de acordo com modelos que consideram o método em quatro níveis: (1) abstração; (2) situação; (3) impacto; e (4) refinamento dos processos. Do ponto de vista da engenharia de software, o método de fusão é visto como um fluxo de dados caracterizado por entrada, saída e funcionalidades. Outro conceito também considerado, baseia-se na organização de conjuntos aleatórios, em se tratando de um *framework*, que combina incertezas de decisões com as decisões em si, por meio de um esquema genérico de incerteza. Outro *framework* de fusão de dados foi citado nessa revisão dos autores, baseado na característica abstrata, bem como na teoria das categorias que atuam de forma generalizada, incluindo fusão de dados, de características, de decisões e de informações relacionais.

LI e colaboradores propuseram a técnica de fusão de sensores no nível de características que abordam o problema da detecção de alvos em ambientes dinâmicos, em um cenário de aprendizado de dados semissupervisionado, com sensores passivos de baixo custo. Essa técnica apresentou, simultaneamente altas probabilidades de detecção correta dos alvos e baixas probabilidades de alarme falso, sob as restrições de recursos limitados de computação e comunicação. Os autores ainda apresentaram um algoritmo de teste de hipóteses binárias, com base no agrupamento de características extraídas de múltiplos sensores utilizados para observação de alvos. As características foram extraídas indivi-

dualmente, a partir de sinais de séries temporais de diferentes sensores e agrupadas em *clusters*, para avaliação da homogeneidade das respostas dos sensores. A decisão para a detecção de alvos foi tomada com base nas medições de distância entre pares de *clusters* de sensores (LI et al., 2016).

A teoria da evidência de *Dempster-Shafer*, proposta por Boudaren e Pieczynski (2016), considera o uso de um modelo oculto de Markov, visando raciocinar sobre incertezas ou fusão de informações que oferecem possibilidades para a modelagem e processamento em relação à imprecisão das informações, sobre a confiabilidade dos sensores, assim como para a fusão de dados.

Ainda, de acordo com LI e colaboradores, há uma abordagem que analisa o desempenho teórico da informação de redes de sensores passivos, para detecção de alvos móveis que se enquadram na categoria de fusão de informações, em nível de dados em redes de sensores. Essa pesquisa foi motivada na perspectiva da fusão de informações, em uma rede local de sensores passivos, onde o ambiente de fundo interferiu nas medições dos sensores, o que causou limitações do espaço de decisão estatística, durante o processo de treinamento. O estado das informações da rede foi representado pelo maior componente principal da série temporal coletada. Esses autores quantificaram a contribuição de cada sensor para a geração do conteúdo da informação, por meio dos modelos de máquinas de Markov e também pelo uso dos modelos de máquinas x-Markov, denominado (cross-Markov), que consistiu em representar um modelo de Markov que incorporou o comportamento de um processo estocástico simbólico baseado nas observações de outro processo estocástico. Para tanto, o método apresentado utilizou a estrutura algébrica de um autômato de estados finitos. A diferença entre as entropias condicionais dessas máquinas foi tratada como medida aproximada da contribuição de informação dos respectivos sensores. Os modelos x-Markov representaram as estatísticas temporais condicionais, de acordo com o estado da informação da rede. A validação foi realizada com dados experimentais coletados de uma rede local de sensores passivos, para detecção de alvos, cujas características estatísticas dos distúrbios ambientais foram semelhantes às do sinal alvo em escala de tempo e textura (LI et al., 2017).

Outros autores, como por exemplo, Bhatnagar e Liu (2017) propuseram uma técnica de fusão de múltiplos sensores, realizada no domínio espacial, com o objetivo de obter as características de atividades locais, ao considerarem que todas as imagens de entrada eram campos aleatórios. As características de atividades locais foram organizadas em uma matriz e então utilizada para construção da imagem fusionada. A homogeneidade da imagem fusionada final foi avaliada antes do término do processo para verificação de consistência. O desempenho da técnica proposta por esses autores foi validado subjetivamente e objetivamente, por meio de experimentos abrangentes em diferentes imagens de múltiplos sensores.

A fusão de imagens de múltiplos sensores, de acordo com Li e colaboradores, foi ba-

seada em uma proposta de uma nova camada de aplicação da Internet das Coisas (IoT) denominada (IFIOT), para preservar as informações espectrais das imagens multiespectrais. Nesse método, os autores identificaram as áreas locais homogêneas, por meio de segmentação *superpixel*, sendo que as áreas homogêneas apresentaram-se uniformes e contiveram apenas um tipo de objeto. Em seguida, esses autores estimaram os pesos espectrais para diferentes bandas na área homogênea e calcularam os coeficientes de ganho de forma adaptativa, minimizando o erro entre as imagens multiespectrais degradadas espectralmente e a imagem pancromática (PAN). As imagens fusionadas foram produzidas após a injeção de detalhes espaciais obtidos da imagem PAN. Os experimentos realizados mostraram que o método IFIOT forneceu bons resultados de fusão, preservando as informações espectrais (LI et al., 2018).

Em contrapartida, Xia e colaboradores utilizaram a fusão de informações de múltiplas fontes aplicada a um modelo que se baseou na rede bayesiana, chamado aliança de inovação. O processo de fusão de informações de várias fontes proposto foi classificado em três camadas: (1) camada de percepção de informações; (2) camada de agrupamento de características; e (3) camada de fusão de decisões. As informações recebidas dos sensores envolvidos, assim como as características, foram agrupadas e fundidas pela aliança de inovação, com base no algoritmo de fusão para obtenção de informações completas e abrangentes. O modelo foi aplicado a um estudo sobre a previsão de informações econômicas, onde a precisão dos resultados de fusão foi maior do que a de uma única fonte e os erros foram menores, com o MPE inferior a 3% (XIA et al., 2018).

Um método de fusão multiespectral-hiperespectral, que explorou as correlações espaciais e espectrais, foi proposto por Yi, Zhao e Chan (2018), para melhorar a resolução espacial de uma imagem hiperespectral. A alta correlação espacial entre a imagem multiespectal e a imagem desejada hiperespectral de alta resolução foi preservada por meio de um dicionário sobrecompleto[8] e a degradação espectral entre elas, projetada no espaço de esparsidade, sendo aplicada como uma restrição espectral. A alta correlação espectral entre a imagem hiperespectral de alta resolução e a de baixa resolução espacial foi preservada por meio da explicitando espectral linear.

Ghamisi e colaboradores reuniram os avanços das abordagens de fusão de dados multissensoriais e multitemporais, fornecendo uma visão abrangente de contribuições dedicadas, especificamente, aos tópicos de fusão para aplicações de imagens satelitais, tais como espaçoespectral ou *pansharpening* e aprimoramento de resolução, fusão de dados de nuvem de pontos, fusão de dados hiperespectrais e LiDAR[9], fusão de dados multitemporais e *Big Data* e mídias sociais (GHAMISI et al., 2019).

[8] Refere-se a um conjunto de funções ou elementos que é maior do que o estritamente necessário para representar um sinal. Essas funções são frequentemente usadas em técnicas de decomposição ou representação de sinais, como a decomposição em componentes de base ou a análise espectral.

[9] *Light Detection and Ranging* ou *Laser Imaging Detection and Ranging*. É uma tecnologia de sensoriamento remoto que utiliza pulsos de laser para medir distâncias precisas e obter informações em 3D sobre a superfície da Terra.

Também, Miyagusuku, Yamashita e Asama (2018) propuseram uma abordagem para fusionar informações de múltiplos pontos de acesso, a fim de aprimorar a auto-localização baseada na tecnologia *WiFi*[10]. Tal abordagem para a fusão foi baseada na teoria da informação e produziu distribuições conjuntas, aplicadas para aprender mapeamentos de localização, de acordo com a intensidade do sinal de cada ponto de acesso. Cada mapeamento foi usado para calcular a probabilidade da localização condicionada aos dados de intensidade do sinal detectado, resultando em funções de probabilidade, frente aos mapeamentos disponíveis.

O problema de fusão de informações de conjuntos foi estudado para sistemas dinâmicos multissensores gerais, conforme Shen e colaboradores, baseado na teoria de conjuntos. Nesse contexto, foram construídos três algoritmos de fusão, centralizados para sistemas multissensores com distúrbios limitados, baseados na estimativa de limite elipsoidal, dentre eles: algoritmo aumentado, algoritmo de filtragem pseudo-sequencial e algoritmo de filtragem de medição combinada. Também, esses autores realizaram uma análise sobre as propriedades dos algoritmos propostos, incluindo estabilidade, convergência, complexidade computacional, a permutabilidade da ordem de atualização das medições e a equivalência entre diferentes algoritmos. Como resultados esses autores mostraram que esses algoritmos apresentaram funcionalidades equivalentes e precisão para as estimativas (SHEN et al., 2019).

Ao considerar a utilização de sensores na agricultura, em suas diversas aplicações, a abordagem de Aygün e colaboradores considerou a implementação de um método de baixo custo, com diferentes dispositivos de entrada, sincronizados com microcontroladores. Os dados obtidos pelos sensores foram enviados por tecnologia sem fio, por intermédio de um dispositivo IoT para a nuvem, registrados e monitorados, em tempo real, via web para submissão à algoritmos de mineração de dados. As variáveis consideradas para análise e estudo foram luz, temperatura, umidade, chuva, umidade do solo, pressão atmosférica, qualidade do ar e ponto de orvalho. As relações entre os dados dos sensores foi obtida com o uso do algoritmo de árvore de regressão, com o objetivo de aplicar a fusão de dados para a redução do número de sensores do conjunto e custo, sem prejudicar o monitoramento. Dessa forma, após os testes de fusão para todos os casos das combinações possíveis do conjunto de variáveis, bons resultados relacionados à otimização do número de sensores foram apresentados. Adicionalmente, a temperatura e o ponto de orvalho puderam ser obtidos usando outros sensores, ou seja, fundindo os dados de treinamento na árvore de regressão com precisão de 92% e 84%, respectivamente, com uma margem de erro de 5% nos nós folha da árvore de regressão (AYGÜN et al., 2019).

O problema de fusão foi investigado por Wang e Liang (2019), para aproveitar ao máximo as informações de dados multimodais. Esses autores propuseram um método de

[10] *WiFi* ou *Wireless Fidelity* é uma tecnologia de comunicação sem fio que permite que dispositivos se conectem a uma rede local (LAN) ou à Internet sem a necessidade de fios físicos.

fusão em duas etapas, para resolverem o problema de fusão de dados heterogêneos, considerando a não compatibilidade e também a não disponibilidade da função de densidade de probabilidade conjunta dos sensores. A primeira etapa transformou os dados multimodais em uma mesma forma de representação, usando uma transformação linear ou não linear específica. Esses autores ainda treinaram cada modalidade de sensor, em modelos separados, considerando as disparidades entre eles o que manteve a informação de cada modalidade preservada. Cada representação foi utilizada como entrada no *framework* de fusão probabilística, o que permitiu o processamento de dados em diferentes modalidades, em um espaço unificado de fusão de informações. A relação intrínseca entre os sensores foi explorada para codificar os dados originais do sensor em um grafo. A relação entre dois sensores foi caracterizada pelo fator de correlação. Resultados numéricos foram fornecidos para validar a eficácia do método proposto na fusão de redes de sensores heterogêneos.

A técnica de fusão de informações, conforme Zhou e colaboradores, teve o objetivo de aplicar a fusão em imagens multifoco e integrar as informações de foco das imagens de origem no resultado fundido. Como principais contribuições, esses autores consideraram: (1) desenvolvimento de um esquema de detecção de *pixels* de foco multiescala para geração de mapas de decisão de foco; (2) verificação de consistência de bloco e também a elaboração da técnica de filtragem rápida e orientada, para a remoção de *pixels* com foco inapropriados e geração de mapas refinados. Para a fase da fusão, os autores utilizaram um filtro de distância entre vizinhos, para extrair *pixels* relacionados aos detalhes das imagens originais (mantendo as informações de alta frequência ou bordas, bem como as de baixa frequência ou áreas uniformes). Também, os resultados da aplicação do método, apresentados por esses autores, retrataram a oportunidade de se trabalhar com mapas de decisão obtidos calculando a distância entre *pixels* vizinhos e correspondentes (ZHOU et al., 2019).

2.3 Destaques e Contextualização

A ferrugem asiática tem impacto significativo no Brasil e em outros países, tanto do ponto de vista financeiro, afetando diretamente o índice do Produto Interno Bruto (PIB), quanto ambiental, devido ao uso intensivo de fungicidas para o controle da doença.

A revisão da literatura possibilitou identificar oportunidades para o desenvolvimento de pesquisa que auxilie na construção de soluções efetivas e completas para o controle desta doença. Adicionalmente, a revisão da literatura apresenta, como estado da arte, soluções que partem do pressuposto de que esta doença já se encontrava instalada nas áreas de cultura analisadas sem propriamente avaliar as evidências relacionadas aos fatores que de fato favoreceram ou favorecem a presença do patógeno e sua evolução. Segundo a revisão analisada, foi possível sistematizar situações decorrentes de variáveis climáticas, quando associadas às informações sobre padrões que ocorrem nas folhas da soja em virtude de

modificações de uma situação normal para outras decorrentes da presença da doença e seu estágio de evolução. Uma vez integradas essas informações, em uma base de regras robusta, passou a ser possível considerar a estruturação de um sistema de visão e inteligência computacional para a gestão de risco da ferrugem asiática em cultura de soja. Assim, buscou-se considerar as principais técnicas e métodos identificados a partir dos melhores resultados apresentados, no referido estado da arte, bem como, a partir de fundamentação teórica bem estabelecida, elencar e avaliar modelos matemáticos e computacionais de forma a completar e viabilizar a estruturação do referido sistema. Neste contexto, são considerados:

1. Quanto à estruturação e qualidade dos dados, foi identificada a possibilidade de trabalhar em contexto *Big Data* a partir do uso de diferentes tipos de dados, dentre eles estruturados, semiestruturados e não estruturados. Também, houve a identificação sobre a viabilidade de uso de soluções para armazenamento de dados via *Data Warehouses* e também por *Data Lake*. Identificou-se também que o ambiente de nuvem é uma opção adequada para a organização de um arcabouço de tecnologias que permitam integrar e desenvolver diferentes cenários para aplicações nas áreas de processamento de séries temporais de dados climáticos, processamento de imagens e sinais, aprendizado de máquina e trabalhos com dados *Big Data*;

2. Diante das abordagens que trataram do processamento de imagens, foram identificadas oportunidades no uso das técnicas de segmentação baseada no algoritmo *k-means*, técnica de limiarização com dois limiares e técnica de extração de fundo complexo em imagens RGB. Adicionalmente, observou-se a possibilidade da análise de classes de *pixels*, considerando faixas de cores e a utilização de transformações por meio de técnicas de conversão para escala de cinza e binarização para a identificação de características do problema a ser investigado;

3. Para os processos de extração de características, foram identificadas oportunidades relacionadas ao uso das técnicas SIFT, HOG e Momentos Invariantes de HU para extração de características de cor, textura e forma relacionadas aos padrões que podem ocorrer nas folhas de soja e identificados com o processamento das imagens digitais. Percebeu-se também o maior uso da técnica de Análise de Componentes Principais (PCA) em relação a outras técnicas de vetores de características, vindo a mesma a ser também considerada uma oportunidade para uso no sistema, vez que colabora, principalmente, no desempenho computacional;

4. Quanto ao aprendizado de máquina, o uso de SVM ficou evidenciado na revisão da literatura, o que o levou a ser considerado uma oportunidade na classificação de padrões relacionados às doenças da soja, incluindo a ferrugem asiática. Também, por apresentar alto desempenho para configurações binárias e multiclasses.

Adicionalmente, os classificadores do tipo árvore de decisão, KNN e *Naïve Bayes* se mostraram promissores para avaliações relacionadas a parâmetros envolvidos em doenças da soja, abrindo assim oportunidades adicionais para uso no sistema e serão também avaliados no âmbito deste trabalho.

5. No que se refere à fusão de dados de sensores, no que tange aos assuntos relacionados às doenças de plantas, o mesmo se mostrou aberto às oportunidades de pesquisa, vez que são poucos os trabalhos disponíveis na literatura. Dentre os modelos que apresentaram oportunidade de ser avaliados, encontraram destaque a integração com base em figura de mérito e integração baseada em lógica difusa. Também, é considerada uma oportunidade a avaliação de modelos baseados em x-Markov e outros a eles relacionados.

Isto posto, no desenvolvimento do sistema de inteligência e visão computacional para a gestão de risco, além das técnicas apontadas acima, serão também consideradas outras inclusões, de forma a se obter uma solução robusta para auxílio à tomada de decisão para o controle desta doença.

2.4 Considerações Finais

As abordagens apresentadas neste Capítulo correspondem, prioritariamente, aos últimos cinco anos de pesquisa nas áreas de agricultura e controle de doenças, infraestrutura de nuvem e *Big Data*, processamento de imagens e sinais, inteligência computacional e modelagem para fusão de dados de sensores. As técnicas apresentadas acima, e outras que se façam necessárias, serão incluídas e avaliadas para o estabelecimento do sistema de inteligência e visão computacional para gestão de risco da ferrugem asiática na cultura de soja.

O próximo capítulo trata sobre os materiais e métodos, com um aprofundamento nas bases matemáticas necessárias para se entender e construir o sistema de visão e inteligência computacional em ambiente de nuvem para gestão de risco da ferrugem asiática na cultura da soja.

Capítulo 3

Materiais e Métodos

Este Capítulo trata sobre materiais, recursos e técnicas utilizadas para o desenvolvimento do sistema de inteligência e visão computacional, em ambiente de nuvem, para gestão de risco da ferrugem asiática na cultura da soja.

3.1 Materiais

Os materiais utilizados envolvem um *dataset* de imagens de folhas de soja, coletadas em condição de campo e durante o cultivo (EMBRAPA, 2021; BARBEDO et al., 2018); um *dataset* de dados climáticos (INMET, 2019) e um *dataset* de dados de planta da soja (SOJA, 2020). Também é utilizada uma estrutura laboratorial na Embrapa Instrumentação de São Carlos, em especial, o Laboratório de Pesquisa e Desenvolvimento em Aplicação Agrícola de Precisão, incluindo uma *workstation* local e o acesso à infraestrutura em nuvem da Oracle *Cloud* (mediante recurso financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), de acordo com Processo 17/19350-2, via convenio IBM Brasil e Embrapa Instrumentação).

As características dos *datasets* utilizados envolvem:

1. Imagens: *Dataset* de imagens sRGB de folhas de soja com diferentes sintomas da FAS, contendo fundo complexo; dimensões: 4128x3096 *pixels*; resolução: 12780288 *pixels*;

2. Dados Climáticos: Vinte anos de dados, dos quais foram considerados os anos de 2015 a 2017, designados como ciclos de produção da soja. Para cada ano, adotou-se 1 ciclo de 120 dias, ocorrendo durante os meses de setembro a dezembro. As características da estação climatológica utilizada para a coleta dos dados: nome da

estação/localização; código da estação: 83358; município de Poxoréo, MT; latitude: -15,82749999, longitude: -54,39555555; data inicial dos dados: 01/01/2000; data final dos dados: 03/07/2019; periodicidade de medição: diária; e

3. Dados de Planta: os dados da planta de soja disponibilizados contém informações sobre a variedade da cultura, a distância entre linhas da cultura, a distância entre plantas, a altura da planta e a quantidade de plantas por metro linear. A variedade da cultura foi a BRS-536, por ser suscetível à FAS. O número de amostras está associado ao número de plantas na cultura, pois as imagens foram tomadas de forma aleatória (amostragens realizadas na área da cultura), porém distribuída de modo a contemplar a área cultivada.

A infraestrutura computacional conta com uma *Workstation*, envolvendo programação, leitura e escrita, conforme as configurações:

1. Tipo Sistema: x64-based PC; Processador AMD64 - 3893 Mhz; Memória física de 64 GB;

2. Sistemas operacionais: *Microsoft Windows* 10, Sistema Operacional *Linux Mint* 20 (Virtualizado);

3. Principais *softwares* instalados: *Python* 3, Editor de código fonte: *Visual Studio Code*, *Microsoft Office* 2019, *Software Putty(SSH)*, *Oracle SQL Developer, Oracle SQL Developer Data Modeler*.

A infraestrutura computacional da *Oracle Cloud* contratada, está planejada para o processamento das etapas estabelecidas no desenvolvimento da pesquisa. A configuração do ambiente de nuvem, conforme as diferentes etapas e necessidades do trabalho, envolve o uso dos seguintes recursos:

1. Acesso à rede publica (Internet);

2. Acesso à rede privada via protocolo *SSH (Secure Socket Shell)* para acesso seguro em instâncias de computação via *Software Putty*; armazenamento de objetos;

3. Armazenamento em banco de dados transacional;

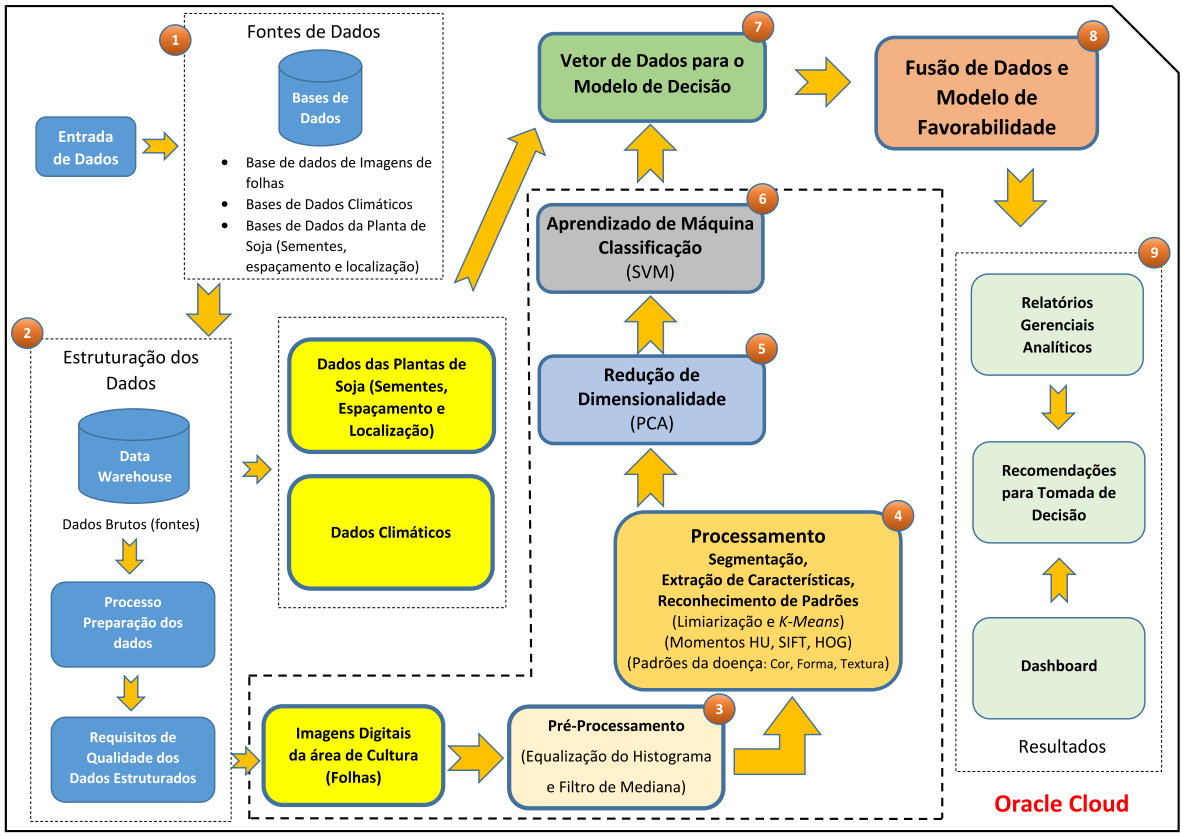
4. Acesso ao ambiente de desenvolvimento integrado às tecnologias emergentes (*Python* e aprendizado de máquinas);

5. Ambiente para construção de *Data Warehouse*, ambiente para análise de dados e ambiente para geração de relatórios analíticos (*Analytics* e *Insights*) auxílio à tomada de decisão.

3.2 Métodos

O Diagrama Conceitual (Figura 14) ilustra detalhadamente o conceito do sistema desenvolvido. Sua estrutura é dividida em blocos interligados, desde o processo inicial até o processo final. É possível observar que há, entre os blocos, uma integração tanto na aplicação conceitual, quanto nas tecnologias de nuvem abordadas.

Figura 14 – Diagrama Conceitual



Fonte: Próprio Autor

A fonte de dados é caracterizada pela entrada de dados de fontes públicas diversas, tais como: (1) base de dados de imagens de folhas de soja; (2) bases de dados climáticas; e (3) bases de dados da planta de soja (sementes, localização e espaçamento). Essas bases são obtidas em órgãos de controle do Governo Federal, ou outros órgãos do terceiro setor: Organizações Não Governamentais (ONGs) que atuam em território nacional. Cada fonte considerada representa um universo de informações, cujas características são essenciais para o escopo desta pesquisa, tais como: base de dados de imagens de folhas; bases de dados de estações climáticas; e bases de dados de planta de soja.

Os dados de imagens são providos de satélites ou coletados in loco, na propriedade rural. A coleta é feita com câmeras portáteis ou por serviços de imageamento, via drones. Ao considerar que as imagens são ricas em informações de diversas naturezas do domínio do problema, seu processamento contribui significativamente para a análise do problema

tratado nesta pesquisa, haja vista que os resultados deste processamento compõem parte relevante do vetor de dados de entrada para o modelo de decisão.

As bases de dados de estações climáticas são acessadas via banco de dados meteorológicos, do Instituto Nacional de Meteorologia (INMET), órgão do Ministério da Agricultura, Pecuária e Abastecimento (INMET, 2019).

Das variáveis climáticas, na configuração de dados históricos diários, disponíveis para acesso público, são consideradas para o trabalho: precipitação total, dada em milímetros (mm); temperatura máxima, dada em graus celsius (°C); temperatura mínima, dada em graus celsius (°C); temperatura média compensada, dada em graus celsius (°C); e umidade relativa do ar, dada em percentagem (%). Considera-se também, para cada registro coletado, o cálculo da variável ponto de orvalho, dado pela Equação 1.

$$Temp_{po} \approx \ Temp - \ \frac{100 - UR}{5} \tag{1}$$

onde $Temp_{po}$ é o ponto de orvalho com a temperatura em (°C); $Temp$ é a temperatura atual de bulbo seco dada em (°C) e UR é a umidade relativa dada em porcentagem (%).

Além disso, as variáveis Região e Nome da Estação Climática também são consideradas como dados para identificação.

A estruturação de dados determina a estrutura dos dados, originados da etapa de fonte de dados, de modo que sua organização facilita o entendimento do domínio trabalhado e subsidia os demais blocos no aproveitamento do potencial das informações coletadas. Para a estruturação elaborada, (Figura 15), se faz uso dos seguintes componentes: (1) Diferentes Fontes de Dados; (2) *Data Lake*; (3) *Data Marts*; (4) *Data Warehouse*; (5) Banco de Dados Relacional; (6) Preparação de Dados; (7) Requisitos de Qualidade e (8) Vetor de Dados.

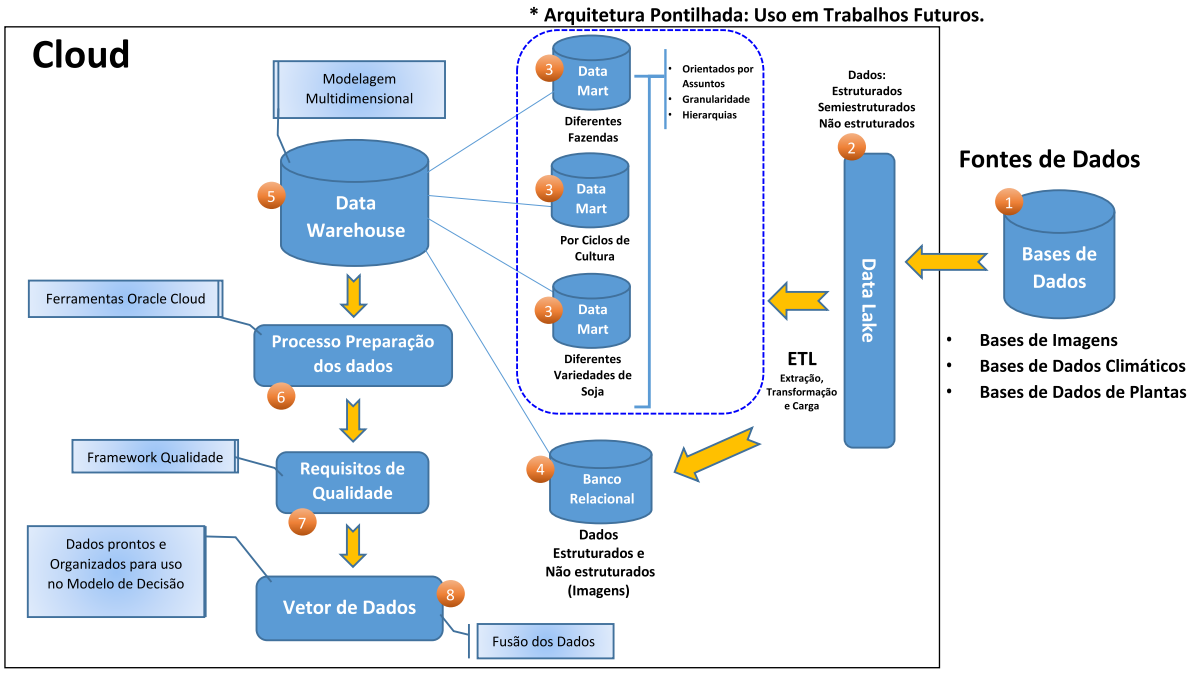
Para a estruturação dos dados, é utilizada a plataforma em nuvem da *Oracle Cloud*. O *Data Lake* compõe a entrada do bloco de estruturação dos dados, recebendo objetos de dados estruturados, semiestruturados e não estruturados, tanto por sistemas legados, quanto por processos individuais de exportação de dados sujeitos a grandes fluxos. O *Data Lake* é um componente fundamental na estrutura que permite suportar dados sob demanda e o tratamento de diferentes tipos de dados, para que possam ser redirecionados, de maneira supervisionada ou semisupervisionada, para outras estruturas internas da nuvem (JOHN; MISRA, 2017; MEHMOOD et al., 2019).

O redirecionamento dos objetos recebidos do *Data Lake*, do diagrama elaborado, na Figura 15, é efetivado, armazenado e preparado de forma supervisionada, compondo os pré-requisitos para o processamento do vetor de dados, ou seja, a última etapa do fluxo do bloco de estruturação de dados.

A infraestrutura de estruturação de dados está planejada para atender a quatro cenários de projeto, sendo, o primeiro, a entrada de dados exportados via *Data Marts* de sistemas legados; o segundo, a entrada de dados semiestruturados e não estruturados via

Data Lake; o terceiro cenário, a entrada de dados somente estruturados com entrada via *Data Lake* e armazenamento no banco de dados relacional e, por fim, o quarto e último cenário, o qual combina os três cenários anteriores, ou seja, o uso de dados de entrada via *Data Marts*, dados semiestruturados e não estruturados e dados estruturados.

Figura 15 – Diagrama Estruturação das Bases de Dados



Fonte: (NEVES; CRUVINEL, 2020)

Os dados de sistemas legados são, inicialmente, carregados no *Data Lake* e posteriormente transferidos para um dos três *Data Marts* previstos na estrutura, são eles: *Data Mart* de Dados de diferentes fazendas; *Data Mart* de diferentes ciclos de cultura ou *Data Mart* de diferentes variedades de soja.

Em contrapartida, os dados semiestruturados e não estruturados são carregados e permanecem armazenados, por padrão, na estrutura do *Data Lake*. Sabendo que o sistema apresentado prevê somente o uso de imagens, como dados não estruturados, opcionalmente, as imagens, após serem armazenadas no *Data Lake*, podem ser transferidas para um banco de dados relacional, que ofereça suporte ao formato *BLOB (Binary Large Object)*. O Sistema Gerenciador de Banco de Dados (SGBD) *Oracle*, presente na nuvem, suporta o tipo de dado *BLOB* para o armazenamento de imagens, vídeos ou áudios, com tamanho de arquivo de 4 *Giga Bytes (GB)*.

A infraestrutura prevista para o *Data Warehouse* tem por objetivo o armazenamento de dados históricos, originados tanto dos *Data Marts*, quanto da carga de dados estruturados do banco de dados relacional, via processo de ETL. O projeto do modelo multidimensional do *DW* está dimensionado de acordo com o documento de requisitos, dividido por assuntos de interesse e elaborado para atender à descoberta de conhecimento dos

processos realizados, no sistema, com foco no monitoramento e identificação da FAS. O documento de requisitos do *DW* pode ser observado na Tabela 2.

Adicionalmente, o modelo multidimensional elaborado constitui-se do modelo estrela com tabela de fatos: FATO_Favorabilidades_FAS e 6 tabelas de Dimensão: DIM_Tempo; DIM_Banco_Imagens; DIM_Dados_Planta_Soja ; DIM_Dados_Climáticos; DIM_Imagem_Clima_Favorabilidade e DIM_Classificações. Este conjunto de tabelas reunidas no modelo estrela se traduz em um cubo de dados que, por sua vez, é elaborado com base em um conjunto de requisitos supracitados.

Tabela 2 – Requisitos de Projeto *Data Warehouse*

***Assunto 1*: Influência das Variáveis Climáticas na Favorabilidade da Ferrugem Asiática da Soja:**
a.Qual o período do ano que a Temperatura (mínima, máxima) pode favorecer o aparecimento da FAS?
b.Qual o período do ano que a Temperatura (Faixa de Temperatura) pode favorecer o aparecimento da FAS?
e.Qual o período do Ciclo de Cultura que a Umidade Relativa contribuiu para o aparecimento da doença da FAS?
f.Qual o período do Ciclo de Cultura que o Ponto de Orvalho contribuiu para o aparecimento da FAS?
g.Qual o período do Ciclo de Cultura que a Precipitação contribui para o aparecimento da FAS?
h.Qual o período do Ciclo de Cultura que a Período Mínimo de Molhamento Foliar contribui para o aparecimento da FAS?
i.Qual o período do Ciclo de Cultura que a Período de Molhamento Foliar contribui para o aparecimento da FAS?
***Assunto 2*: Contabilização da Favorabilidade Baixa, Média e Alta por Ano:**
j.Qual o período do Ciclo de Cultura, que compreende a etapa de plantio e colheita por Região, mostra a favorabilidade Baixa da FAS?
k.Qual o período do Ciclo de Cultura, que compreende a etapa de plantio e colheita por Região, mostra a favorabilidade Média da FAS?
l.Qual o período do Ciclo de Cultura, que compreende a etapa de plantio e colheita por Região, mostra a favorabilidade Alta da FAS?
***Assunto 3*: Influência da Imagem da Folha de Soja na Favorabilidade da Ferrugem Asiática da Soja:**
m.Qual o período do ano, que compreende a etapa de plantio e colheita (R5 e R6) por Região, no qual a informação da Imagem da Folha de Soja contribuiu para o aparecimento da doença da FAS?

Fonte: Próprio Autor

Dessa forma, o detalhamento do modelo multidimensional estrela está desenvolvido na ferramenta *Oracle SQL Developer*, o qual pode ser visualizado no Apêndice A.

O banco de dados relacional é uma estrutura indicada para compor o bloco de estruturação de dados para armazenar e gerenciar dados estruturados, sob o aspecto transacional. Os dados estruturados, no sistema, são inicialmente originados via *Data Lake* e, ao passo que as fases do sistema passam a produzir resultados, estes também são armazenados no banco relacional. A partir dos dados populados neste SGBD, é possível atender às requisições para o processamento da fases de fusão de dados do sistema e também alimentar o *Data Warehouse* com cargas via ETL, com a finalidade de realizar análises de dados para tomadas de decisão, por meio de ferramentas de *Business Intelligence* do ambiente Oracle *Cloud*.

O modelo de concepção da estrutura do banco de dados relacional é desenvolvido de acordo com a necessidade de armazenamento dos dados de entrada e saída para o sistema. A ferramenta utilizada para o desenvolvimento da estrutura do banco de dados relacional é a *Oracle SQL Developer Data Modeler*. O detalhamento do Modelo Relacional pode ser visualizado no Apêndice B.

Após o armazenamento dos dados, tanto no *Data Warehouse*, quanto no banco de dados relacional, o próximo passo do bloco de estruturação trata do processo de preparação dos dados, aplicando técnicas de pré-processamento para minimizar problemas com dados incorretos, inconsistentes, duplicados e ausentes. O processo de preparação dos dados também ocorre, se necessário, na etapa de migração de diferentes bases de dados para o *Data Lake*. Um conjunto de dados preparados, segundo Facelli e colaboradores, leva à construção de modelos mais adequados e também à diminuição da complexidade do processamento dos algoritmos que produzem, por consequência, melhores resultados, assim como na construção do conhecimento (FACELI et al., 2011).

As funcionalidades disponíveis nas principais operações que envolvem a preparação de dados consiste na análise de diferentes fontes de dados para tratar a integridade, consistência e completude dos mesmos. O ambiente da Oracle *Cloud* disponibiliza ferramentas que facilitam a execução do processo de preparação dos dados de forma semi-automatizada. Entre as ferramentas disponíveis, podem ser citadas algumas opções, sendo a primeira, a Transformação de Dados, recurso disponível no *Data Studio* e agregado aos serviços do *Autonomous Database*. Como segunda ferramenta, é possível elencar a Integração de Dados, disponível para o recurso de *Data Lake*. Como terceira opção, entre as principais, menciona-se o Fluxo de Dados, disponível como ferramenta de análise do *Analytics Cloud*.

No pré-processamento das imagens das folhas de soja, o processo de remoção do fundo complexo das imagens e a investigação da caracterização das cores da doença são realizados por meio da segmentação.

A segmentação é definida, de acordo com Gonzalez e Woods (2010), como processo que particiona a imagem em regiões de interesse ou objetos que a compõem. A segmentação é trabalhada por diferentes métodos, sendo os principais: por limiar (RIDLER; CALVARD et al., 1978), baseada em regiões (ZUCKER, 1976), bordas (CANNY, 1986), agrupamento (CELEBI; KINGRAVI; VELA, 2013) e detecção de contornos (LEUNG; MALIK, 1998).

Segundo Gonzalez e Woods (2010) as Equações 2 e 3 se referem, respectivamente, às técnicas de segmentação por limiar global e limiarização por histerese.

$$f(cx, cy) = \begin{cases} 1, & \text{se } I(cx, cy) > LM \\ 0, & \text{caso contrário} \end{cases} \tag{2}$$

onde $I(cx, cy)$ é o valor do pixel na posição (cx, cy) da imagem e LM é o valor de limiar.

$$\mathrm{M}(cx, cy) = \begin{cases} 1, & \text{se I(cx, cy)} > \text{TMh e conectado a um pixel já segmentado} \\ 0, & \text{caso contrário} \end{cases} \tag{3}$$

onde $M(cx, cy)$ é a máscara resultante da limiarização por histerese, onde cada *pixel* da imagem de saída pode ser 1 ou 0, dependendo das condições especificadas; $I(cx, cy)$ é o valor do *pixel* na posição (cx, cy) da imagem, TMh é o limiar superior e a conexão se refere à ligação de *pixels* vizinhos.

Em contrapartida, a técnica de clusterização K-means tem como conceito a atribuição de *pixels* a um número K de *clusters*, definido previamente, dado que K define os pontos centrais como centroides, aleatoriamente, em seguida, cada *pixel* é atribuído a um determinado *cluster*, de acordo com a menor distância euclidiana calculada entre os centroides. Assim, os centroides são atualizados interativamente, movendo-se para o centro de massa de seus *pixels* atribuídos. O processo continua até atingir a convergência ou o número máximo de iterações. A técnica *K-means* está apresentada pela Equação 4.

$$\arg\min_{C} \sum_{i=1}^{k} \sum_{\iota \in C_i} \|\iota - \mu_i\|^2 \tag{4}$$

onde C é o *cluster* que possui dados semelhantes; C_i é o *cluster* específico; μ_i é o centroide do *cluster* i; k é o número de *clusters* com valor igual a 6 e $\|\iota - \mu_i\|^2$ é a distância euclidiana ao quadrado entre um ponto ι e o centroide μ_i do cluster C_i.

A segmentação de imagens por cores (GARCIA-LAMONT et al., 2018) utiliza o método de agrupamento, a partir da técnica *K-means*, de limiarização por cor que, segundo Gonzalez e Woods (2010), se baseia na definição de limiares de cor, considerando os principais espaços de cores RGB, HSI, LAB, com aplicações em detecção de objetos, reconhecimento de padrões e diferentes análises ou processamentos de imagens médicas ou de satélites.

Diante das técnicas de segmentação discorridas, adota-se para o sistema a segmentação por agrupamento pela técnica *K-means*, apoiada pela segmentação via técnica de limiarização, a partir do uso de dois limiares.

O processo de segmentação envolve duas etapas, sendo a primeira (etapa I) responsável pela retirada do fundo complexo (Algoritmo 1) e a segunda (etapa II) o processamento do conhecimento customizado (Algoritmo 2) com foco na fenomenologia do processo agrícola relacionado à FAS.

Algoritmo 1 Pseudocódigo para Segmentação Etapa I

Entrada: Imagem RGB com fundo complexo
Saída: Imagem Preparada para Segmentação Etapa II

```
1  início
2      k ← 6;                                          ▷ número de clusters: K-means
3      imagem ← folha_fundo_RGB;
4      Dividir_canais_RGB(imagem):
5      retorna imagens canais R,G,B;
6      Equalizar_histograma(canal_G):
7      retorna canal_G_equalizado;
8      Limiarizar_imagem(canal_G_equalizado, lim_1, lim_2);   ▷ lim_1, lim_2: limiares
9      retorna imagem_limiarizada;
10     Segmentar_k_means(k, imagem_limiarizada):
11     retorna rotulos_imagens;
12     Escolha_rotulo_significativo(rotulos_imagens):
13     retorna rotulo_significativo;
14     matting_imagens(rotulo_significativo, imagem):
15     retorna imagem_segmentada;
16     suavizar_imagem(imagem_segmentada):
17     retorna imagem_segmentada_suavizada;
18 fim
19 retorna imagem segmentada suavizada (sem fundo);
```

De acordo com o Algoritmo 1, são descritas abaixo as suas respectivas funções:

1. *Dividir_canais_RGB*: a partir de uma imagem RGB de entrada, é feita a separação dos canais vermelho (R), verde (G) e azul (B). Considera-se o canal G por ser mais adequado para o contexto;

2. *Equalizar_histograma*: sob a imagem do canal G, aplica-se a técnica de equalização de histograma para facilitar o processo de segmentação do fundo;

3. *Limiarizar_imagem*: dada a imagem do canal G equalizada, aplica-se a técnica de limiarização global, utilizando-se de dois limiares para obtenção de um intervalo entre *lim_1* e *lim_2*, com o objetivo de retirada parcial do fundo da imagem;

4. *Segmentar_k_means*: a partir de uma imagem limiarizada como entrada e da definição do número de *clusters*, é utilizada a técnica *k-means* para segmentar a imagem do fundo, possibilitando a geração de rótulos da imagem segmentada;

5. *Escolha_rotulo_significativo*: ao considerar como entrada os rótulos segmentados, escolhe-se, de modo supervisionado, o rótulo que melhor segmentou a imagem;

6. *Matting_imagens*: dado o rótulo escolhido, faz-se o *matting* com a imagem original para recuperação dos *pixels* com as cores originais;

7. *Suavizar_imagem*: a partir da imagem resultante do *matting*, suaviza-se a imagem para minimização de ruídos para colaboração às próximas etapas de processamento.

A etapa II da segmentação (Algoritmo 2) se caracteriza como uma etapa semiautomatizada, pois, em meio às rotinas automatizadas, há também rotinas que necessitam de supervisão. O algoritmo da etapa II da segmentação tem, como principal objetivo, trabalhar as cores obtidas pelo conhecimento da FAS, expressas por suas referências marrom e amarela, e representar o comportamento na forma em que essas se configuram na folha. A descrição das funções, estabelecidas no Algoritmo 2, estão relacionadas abaixo.

1. *Coletar_melhor_semente*: ao considerar a imagem do canal G, o tamanho da janela fixa e a definição da região de interesse, são calculadas as coordenadas do *pixel* central baseado na vizinhança para cada *pixel* da imagem. O procedimento é executado para cada cor envolvida (verde, amarela e marrom). Como retorno, são coletadas as coordenadas do *pixel* central a partir do menor erro associado, bem como o conjunto de sementes baseadas na janela fixa, de acordo com o *pixel* central;

2. *Calcular_janela_pixels*: é calculada uma janela de tamanho variável baseada em cada cor considerada. Retorna-se as coordenadas do ponto central e as sementes calculadas a partir da vizinhança dessa janela, baseada no cálculo da população de *pixels*. O procedimento é feito de acordo com o número de amostras estatisticamente suficientes para o cálculo da faixa de limiares;

3. *Calcular_limiares*: tendo como entrada o conjunto de sementes calculadas pela janela, de acordo com cada cor, são obtidos, como retorno, dois limiares e o erro associado do cálculo realizado. O cálculo dos limiares envolve a definição de uma faixa de limiares, levando em consideração os valores de variância, desvio padrão e a mediana do pixel, de forma que os limites a serem calculados se aproximem o máximo possível do valor da cor do pixel de referência. Assim, para a faixa dos limiares deve ser considerado o tamanho de $\pm$ 1 sigma, a partir do valor da mediana. Além disso, a faixa de limiares é estabelecida quando o valor do erro associado, neste cálculo, fica menor ou igual a 5%. Quando o erro associado não atinge este percentual, as sementes são reavaliadas e retirados, deste conjunto, os *outliers* correspondentes. Neste conjunto sem *outliers*, o cálculo é refeito até que a percentagem do erro atinja o valor esperado;

4. *Calcular_boxplot*: avalia-se a qualidade dos dados que se referem aos conjuntos de *pixels* sementes, antes e depois do processo de cálculo dos limiares.

Algoritmo 2 Pseudocódigo para Segmentação Etapa II

Entrada: imagem segmentada Etapa I
Saída: Imagem Preparada para Segmentação Etapa II

```
início
    k ← 6;
    imagem ← folha_fundo_RGB;
    vet_cor_ref ← Pixels(verde : 104, amarelo : 137, marrom : 75);      ▷ vetor de cores de referência
    tam_janela_fixa ← 5x5;                                  ▷ janela fixa para busca da melhor semente
    tam_janela_cor ← num_pixels;  ▷ num_pixels= 144 (cores marrom e amarela); 196 (cor verde)
    Dividir_canais_RGB(imagem):
    retorna imagens canais R,G,B;
    Coletar_melhor_semente(img_canal_G, vetor_cor_ref, tam_janela_fixa, ROI):
    início
        para (pixel_img_canal_G == vetor_cor_ref) faça
            Calcular_vizinhança(tam_janela_fixa, ROI, cor_ref):
            retorna coordenadas_pixel_central, sementes[], erro;
        fim
    fim
    retorna pixel_semente_menor_erro;
    Calcular_janela_pixels(vetor_cores_ref, pixel_semente_menor_erro, sementes[], tam_janela_cor):
    retorna coordenadas_janela, sementes_janela[];
    Calcular_boxplot(sementes_janela[]):
    retorna boxplot;
    Calcular_limiares(sementes_janela[]):
    início
        Calcular_erro_associado(sementes_janela[]):
        erro_associado ← desvio_padrão/mediana;
        retorna erro_associado;
        se erro_associado ≥ 5% então
            repita
                Retirar_outliers(sementes_janela[]):
                retorna sementes_janela[] - outliers;
                Calcular_limiares(sementes[]):
                desvio_padrão ← sementes[];
                mediana ← sementes[];
                erro ← desvio_padrão/mediana;
                lim_1 ← mediana − desvio_padrão;
                lim_2 ← mediana + desvio_padrão;
                retorna lim_1, lim_2, desvio_padrão, mediana, erro_associado;
            até erro_associado ≤ 5%;
            retorna lim_1, lim_2, desvio_padrão, mediana, erro_associado (valores finais);
        fim
    fim
    retorna lim_1, lim_2 (calculados), sementes_janela_sem_outliers[];
    Calcular_boxplot(sementes_janela_sem_outliers[]):
    retorna boxplot;
    Limiarizar_imagem(img_canal_G, lim_1, lim_2):
    retorna imagem_limiarizada;
    Segmentar_k_means(k, imagem_limiarizada):
    retorna labels_imagens;
    Escolha_rotulo_significativo(rotulos_imagens):
    retorna rotulo_significativo;
    Matting_imagens(rotulo_significativo, imagem):
    retorna imagem_segmentada;
    Suavizar_imagem(imagem_segmentada):
    retorna imagem_segmentada_suavizada;
fim
retorna imagem segmentada suavizada (por cor);
```

Destaca-se, como ponto de automatização do processo, a identificação do *pixel* semente, de acordo com as cores de referência da doença e o processo de definição dos limiares, frente ao uso de técnicas estatísticas como cálculo de mediana, desvio padrão e retirada de *outliers*, considerando o erro máximo associado $\leq 5\%$.

A segunda etapa da segmentação leva à definição das classes da doença, onde a cor amarela, em uma área pequena da folha, mostra o estágio inicial ou, em áreas maiores, o estágio intermediário da doença. Porém, quando a classe amarela está associada à cor marrom, a doença pode se tornar dominante, de acordo com a ocupação da área da folha, fato este que pode levar à desfolha precoce da planta. Por este motivo, considera-se importante a busca pelo diagnóstico no estágio inicial da doença, um ponto importante a ser considerado no sistema apresentado.

A etapa de pré-processamento, realizada por meio da segmentação e dos filtros de cores RGB e também da mediana, proporciona que as imagens digitais da cultura de soja possam ser investigadas quanto à caracterização da FAS. Dadas as cores de referência, tais caracterizações podem ser aprofundadas no processo de reconhecimento de padrões da doença.

O reconhecimento de padrões tem por objetivo a extração das características da FAS, por meio da imagem da folha da soja. Os descritores adotados para trabalhar no sistema são de cor, implementado pela técnica *Scale-Invariant Feature Transform* (SIFT); de forma, implementado pela técnica de Momentos Invariantes de HU; e descritor de textura, implementado via técnica *Histogram of Oriented Gradients* (HOG).

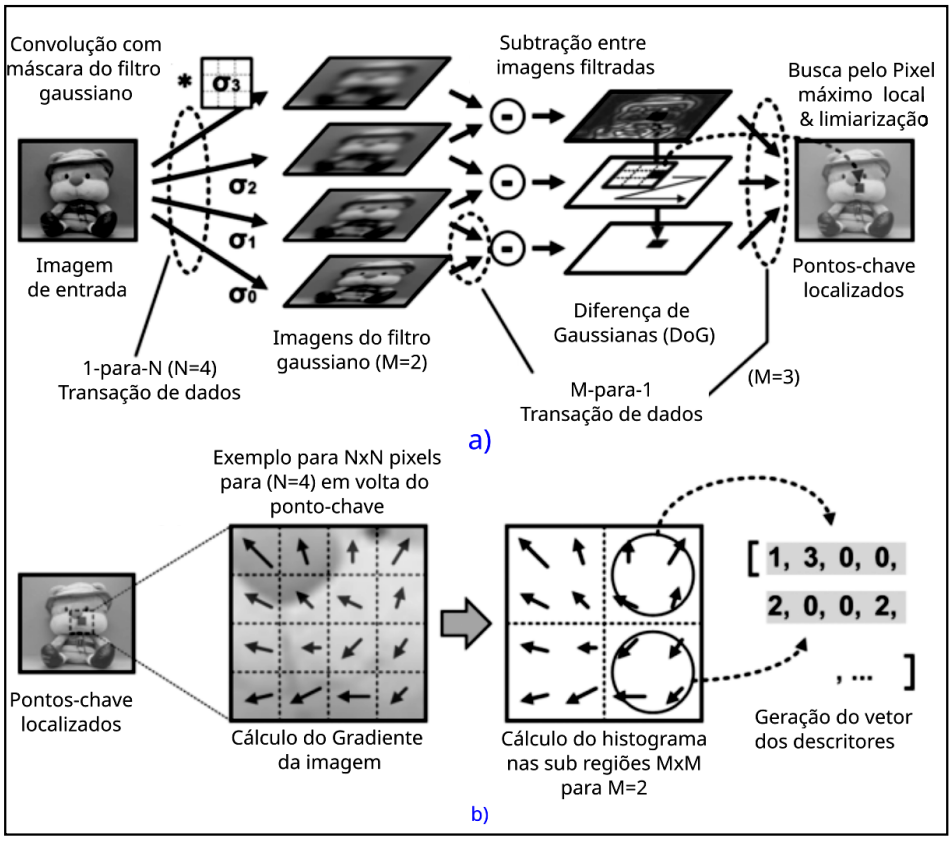
Entende-se por SIFT, segundo Lowe (1999), uma abordagem que transforma uma imagem em uma grande coleção de vetores de características locais invariantes à imagem, translação, dimensionamento e rotação, sendo parcialmente invariante às mudanças de iluminação e projeção. Os recursos invariantes de escala, nesta abordagem, é eficientemente identificada por meio de filtragens em diversos estágios, sendo o primeiro responsável por procurar locais-chave no espaço da escala cujos locais sejam valores máximos ou mínimos, de acordo com a função de Gauss. Assim, cada local é um ponto usado para gerar um vetor de recursos que descreve a região da imagem local amostrada, referente às coordenadas de escala-espaço. Cada recurso atinge uma invariância parcial de acordo com as variações locais, produzindo o desfoque dos locais de gradiente da imagem. Então, os vetores que são resultantes desse processo são definidos como Chaves SIFT que, por sua vez, suas localizações 2D representam modelos de objetos atuais afins.

A visão geral das etapas do processo de extração de características com a técnica SIFT está ilustrada na Figura 16, a qual contém: em (a) pontos-chave e em (b) ilustração sobre a geração do vetor de descritores SIFT.

Quanto ao descritor de forma geométricas, é considerada a utilização da técnica de Momentos Invariantes de HU que, de acordo com Hu (1962), trata-se do teorema fundamental que relaciona invariantes de momentos bidimensionais para figuras planares, com

invariantes algébricos e suas aplicações para processamento de informações visuais. Sistemas completos de momentos invariantes são estabelecidos sob os aspectos de translação, similitude e transformações ortogonais, os quais demonstram que o reconhecimento de padrões geométricos e caracteres alfabéticos, independentemente da posição, tamanho e orientação flexível, sendo assim suficientes para o aprendizado de qualquer conjunto de padrões.

Figura 16 – Visão Geral de Extração de Características com a Técnica SIFT



Fonte: Adaptada de (KIM et al., 2009)

De acordo com Zhao e Wang (2010), são necessários os cálculos dos momentos bidimensionais, centrais e centrais normalizados, para que os sete momentos invariantes de HU possam ser calculados.

As equações de 5 a 10 representam os seis momentos invariantes ortogonais absolutos de segunda e terceira ordem. A Equação 11 apresenta um momento invariante ortogonal assimétrico utilizado na distribuição de espelhamento de imagens.

$$\varnothing_1 = \eta_{20} + \eta_{02} \tag{5}$$

onde $\varnothing_1$ é a invariante ortogonal que se refere ao primeiro momento invariante; η_{20} é o momento central de segunda ordem, que é calculado a partir da imagem, ou da região de interesse (ROI) e representa a dispersão da distribuição de *pixels* ou *voxels* ao longo do

eixo X; η_{02} é o momento central de segunda ordem, calculado a partir da imagem ou da ROI, e representa a dispersão da distribuição de *pixels* ou *voxels* ao longo do eixo Y.

$$\varnothing_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \tag{6}$$

onde $\varnothing_2$ é a segunda invariante ortogonal à rotação que se refere a uma medida das características geométricas da imagem ou da ROI; η_{11} é o momento central de segunda ordem entre os eixos X e Y que representa a covariância entre os eixos da distribuição de *pixels* ou *voxels*.

$$\varnothing_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \tag{7}$$

onde $\varnothing_3$ é a terceira invariante ortogonal de rotação que se refere a medida das características geométricas da imagem ou da ROI; η_{30} é o momento central de terceira ordem ao longo do eixo principal X que representa a dispersão da distribuição *pixels* ou *voxels* ao longo desse eixo; η_{12} e η_{21} são momentos centrais de terceira ordem que envolvem misturas de deslocamento, ao longo dos eixos principais X e Y, que representam a dispersão da distribuição de *pixels* ou *voxels* devido às interações entre os eixos; η_{03} é o momento central de terceira ordem, ao longo do eixo Y, que representa a dispersão da distribuição de *pixels* ou *voxels* ao longo desse eixo.

$$\varnothing_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} - \eta_{03})^2 \tag{8}$$

onde $\varnothing_4$ é a quarta invariante ortogonal à rotação que se refere à medida das características geométricas da imagem ou da ROI.

$$\begin{aligned}\varnothing_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + \\ (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]\end{aligned} \tag{9}$$

onde $\varnothing_5$ é a quinta invariante ortogonal à rotação que se refere à medida das características geométricas da imagem ou da ROI.

$$\begin{aligned}\varnothing_6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\ + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})\end{aligned} \tag{10}$$

onde $\varnothing_6$ é a sexta invariante ortogonal à rotação que descreve as características geométricas de uma imagem ou de uma ROI.

$$\begin{aligned}\varnothing_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + \\ (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]\end{aligned} \tag{11}$$

onde $\varnothing_7$ é a sétima invariante ortogonal à rotação usada para descrever as características geométricas de uma imagem ou de uma ROI.

Adicionalmente, a técnica de Histogramas de Gradientes Orientados (HOGs), segundo Dalal e Triggs (2005), propõe um método baseado na avaliação de histogramas locais normalizados quanto às orientações de gradientes da imagem, de acordo com as Equações 12 e 13. Essa abordagem utiliza uma grade densa que consiste na extração de características e detecção de objetos. A partir da imagem de entrada, divide-se a janela de imagem em pequenas regiões chamadas de células, onde acumulam-se histogramas de informações locais 1D de direções de gradiente ou orientações de borda sobre os *pixels* desta célula. As entradas do histograma são combinadas para formar a representação. Realiza-se a normalização sobre regiões espaciais maiores denominadas Blocos, a partir dos dados de contraste das respostas locais, acumulando medidas de energia dos histogramas de cada bloco. Os blocos de descritores normalizados são denominados como descritores de histogramas de gradientes orientados. Após coletados os HOGs sobre os blocos dos dados, estes são submetidos ao classificador SVM com *Kernel* linear para obtenção da resposta, ou seja, se é ou não é o objeto que se busca. Nessa técnica, os objetos experimentais tratam-se de imagens para detecção humana. A Figura 17 ilustra a extração de características da técnica HOG.

$$Hist(\theta) = \sum_{pixels \text{ na célula}} w(\theta - \Theta) \tag{12}$$

onde $Hist(\theta)$ representa o valor acumulado do histograma de orientação para uma célula em um determinado ângulo; θ[1] indica a orientação do gradiente em um determinado *pixel* dentro da célula, representado pelo valor de 22,5°; $w(\theta - \Theta)$ função de ponderação que determina como a contribuição de um gradiente específico para o histograma, ponderada com base na diferença angular; Θ[2] indica o ângulo do "bin" ou (barra) dos gradientes na célula, representado pelo valor de 2,22°; *pixels* na célula indica que a soma é realizada sobre todos os *pixels* na célula.

$$v = \frac{v}{\sqrt{\|v\|_2^2 + \epsilon^2}} \tag{13}$$

onde ϑ representa o vetor concatenado de histogramas de orientação em um bloco; $||\vartheta||_2$ representa a norma euclidiana (ou comprimento) do vetor ϑ, calculada como a raiz quadrada da soma dos quadrados dos elementos do vetor; ϵ é uma pequena constante adicionada na raiz quadrada para evitar possíveis divisões por zero, representada pelo valor $0,01$.

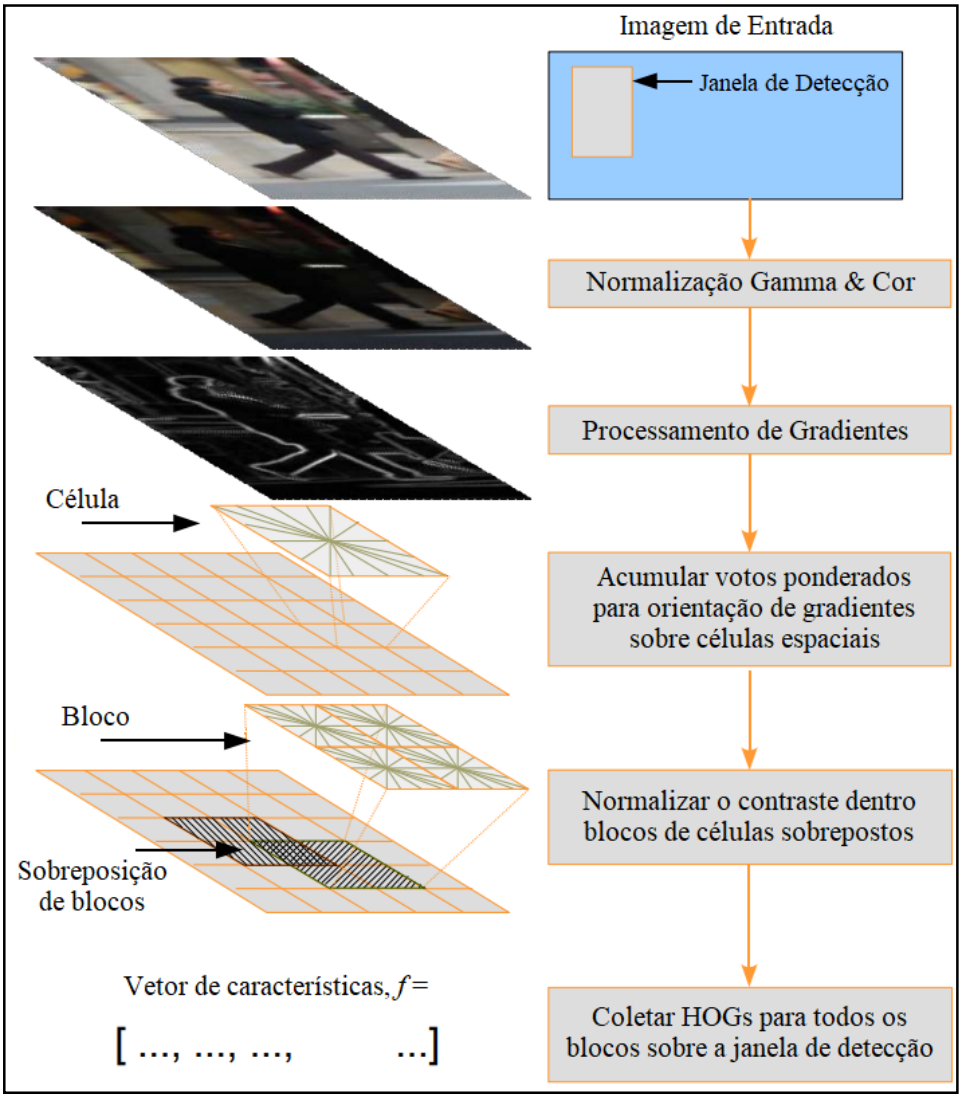
Para o início do processamento do Bloco de Reconhecimento de Padrões do sistema, são definidos os padrões de cor verde, amarela e marrom, conforme referências de cores

[1] É definido considerando a divisão do ângulo de 180° pelo número de bins definido, igual a 8.

[2] É calculado a partir de θ, considerando: o número de bins e o uso da função gaussiana para ponderar a contribuição de cada *pixel* para o histograma.

observadas nas amostras das imagens do banco de imagens para a FAS, considerando desde os primeiros sinais de aparecimento da doença, até o último estágio.

Figura 17 – Visão Geral de Extração de Características com a Técnica HOG



Fonte: Adaptada de (DALAL; TRIGGS, 2005)

Em um segundo momento, submete-se as imagens segmentadas, sem o fundo complexo, aos métodos de reconhecimento de padrões. Os algoritmos SIFT, HOG e Momentos de HU, respectivamente, extraem as características de cor, textura e forma nas imagens das folhas segmentadas de soja e as armazenam, inicialmente, em vetores de características distintos.

Na sequência, os vetores de características são organizados e preparados por meio de processos de tratamento, como a verificação de dados faltantes, a normalização dos dados de características e a redução da alta dimensionalidade de 130 para 19 dimensões.

Ao fechar o processo, a etapa de reconhecimento de padrões entrega para cada imagem processada um vetor de características por cor, ou seja, um vetor verde, um amarelo e

outro marrom.

O aprendizado de máquina é responsável por classificar os padrões reconhecidos para cada imagem de folha da soja.

Neste trabalho, os classificadores árvore de decisão (BREIMAN et al., 1984), *K-Nearest Neighbor* (KNN) (COVER; HART, 1967), *Naïve Bayes* (CESTNIK; KONONENKO; BRATKO, 1987) e Máquinas de Vetores de Suporte (SVM) (VAPNIK, 1995) são considerados para avaliação, como alternativas para o sistema de visão e inteligência computacional em ambiente de nuvem para gestão de risco da ferrugem asiática na cultura da soja.

Árvores de decisão (BISHOP; NASRABADI, 2006) são modelos utilizados para tarefas de classificação e regressão, e seu desenvolvimento envolve a escolha cuidadosa das divisões para minimizar o erro de predição. A predição ótima para uma região R_τ em uma árvore de decisão é dada pela média dos valores dos dados pertencentes a essa região, sendo a predição expressa pela Equação 14.

$$y_\tau = \frac{1}{N_\tau} \sum_{x_n \in R_\tau} t_n \tag{14}$$

onde y_τ é a predição para a região R_τ, N_τ é o número total de exemplos na região R_τ, e t_n é o valor da resposta para o exemplo n. Esta média, y_τ, é usada como a predição para todos os dados que caem na região R_τ.

A eficácia de uma divisão é medida utilizando a soma dos quadrados dos resíduos, que é calculada pela Equação 15.

$$Q_\tau(T) = \sum_{x_n \in R_\tau} \{t_n - y_\tau\}^2 \tag{15}$$

Na Equação 15, $Q_\tau(T)$ quantifica a variabilidade dos valores de resposta em relação à predição média y_τ, com o objetivo de minimizar a soma total dos resíduos em toda a árvore, o que é alcançado somando as contribuições de todas as regiões.

Para evitar o sobreajuste e balancear a complexidade do modelo com o erro de predição é utilizado um critério de poda, dado pela Equação 16.

$$C(T) = \sum_{\tau=1}^{|T|} Q_\tau(T) + \lambda|T| \tag{16}$$

onde $C(T)$ é o critério de poda para a árvore T, $|T|$ é o número de folhas na árvore T, e λ é o parâmetro de regularização. O termo $\lambda|T|$ penaliza a complexidade do modelo, incentivando árvores menores e mais simples. O parâmetro λ é ajustado via validação cruzada para encontrar o equilíbrio ideal entre o erro de predição e a complexidade do modelo.

Para problemas de classificação, em vez da soma dos quadrados dos resíduos, são usadas medidas de desempenho apropriadas para avaliar a qualidade das divisões. A

Entropia Cruzada é uma medida considerada, dada pela Equação 17, assim como o Índice de Gini, expressa pela Equação 18.

$$Q_\tau(T) = \sum_{k=1}^{K} p_{\tau k} \ln p_{\tau k} \tag{17}$$

onde $p_{\tau k}$ é a proporção de pontos de dados na região R_τ atribuídos à classe k. A entropia cruzada mede a incerteza das previsões de classe e penaliza regiões com distribuições de classes mais uniformes.

$$Q_\tau(T) = \sum_{k=1}^{K} p_{\tau k}(1 - p_{\tau k}) \tag{18}$$

onde $p_{\tau k}$ representa a proporção de pontos de dados na região R_τ atribuídos à classe k.

O Índice de Gini avalia a impureza da região, com valores mais altos, indicando uma mistura desigual das classes. Este índice ou ainda a Entropia Cruzada podem ser selecionados como uma escolha para a formação de regiões, onde a maioria dos dados pertence a uma única classe específica. Essas medidas são preferidas em comparação com a taxa de erro de classificação devido à sua sensibilidade às probabilidades de classe e à sua capacidade de serem diferenciáveis, o que é útil para métodos de otimização baseados em gradiente. Além disso, a estrutura de uma árvore é sensível aos dados de treinamento, o que pode resultar em divisões muito diferentes com pequenas alterações no conjunto de dados.

O classificador *K-Nearest Neighbors* (KNN) (BISHOP; NASRABADI, 2006) é um método de classificação baseado na proximidade das características das amostras de treinamento. Dado um conjunto de dados de treinamento $\{(x_i, y_i)\}_{i=1}^{N}$, onde $x_i \in \mathbb{R}^d$ é composto por vetores de características e $y_i \in \{1, 2, \ldots, C\}$ é composto por rótulos de classes, onde a tarefa deste classificador é prever o rótulo de classe y para uma nova amostra x.

No classificador KNN, a densidade é estimada localmente usando uma esfera centrada no ponto x e ajustando seu raio até que ele contenha precisamente K pontos de dados, sendo a estimativa da densidade $p(x)$, dada pela Equação 19.

$$p(x) = \frac{K}{NV} \tag{19}$$

onde K é o número de pontos de dados dentro da esfera, N é o número total de pontos de dados no conjunto de treinamento, e V é o volume da esfera. Para cada classe C_k, a densidade condicional $p(x \mid C_k)$ é estimada pela Equação 20.

$$p(x \mid C_k) = \frac{K_k}{N_k V} \tag{20}$$

onde K_k é o número de pontos da classe C_k dentro da esfera, N_k é o número total de pontos da classe C_k, e V é o volume da esfera.

Aplicando o teorema de Bayes, a probabilidade posterior da classe dada a amostra x é dada pela Equação 21.

$$p(C_k \mid x) = \frac{p(x \mid C_k) \cdot p(C_k)}{p(x)} = \frac{K_k}{K} \tag{21}$$

onde $p(C_k)$ é a probabilidade a priori da classe C_k, e $p(x)$ é a densidade não condicionada. A classe predita $\hat{y}$ é aquela com a maior probabilidade posterior, conforme descrito na Equação 22.

$$\hat{y} = \arg \max_{c \in \{1,2,\ldots,C\}} \frac{K_c}{K} \tag{22}$$

onde K_c é o número de pontos da classe c entre os K vizinhos mais próximos. O caso particular de $K = 1$ é conhecido como a regra do vizinho mais próximo, onde o ponto de teste é simplesmente atribuído à mesma classe do ponto de treinamento mais próximo.

O classificador *Naïve Bayes* (BISHOP; NASRABADI, 2006) é um método de classificação baseado na suposição de independência condicional das características, dado a classe. Conforme um conjunto de dados de treinamento $\{(x_i, y_i)\}_{i=1}^{N}$, onde $x_i \in \mathbb{R}^d$ é composto por vetores de características e $y_i \in \{1, 2, \ldots, C\}$ é composto por rótulos de classe, o objetivo é prever o rótulo de classe y para uma nova amostra x. Adicionalmente, no modelo *Naïve Bayes*, as características são consideradas independentes condicionalmente à classe. Portanto, a probabilidade condicional $p(x \mid C_k)$ pode ser expressa pela Equação 23.

$$p(x \mid C_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i} (1 - \mu_{ki})^{1 - x_i} \tag{23}$$

onde D é o número de características, x_i é o valor da i-ésima característica da amostra, e μ_{ki} é a probabilidade da característica x_i ser 1, dado que a amostra pertence à classe C_k.

A função de decisão do classificador *Naïve Bayes* é baseada no cálculo do logaritmo da probabilidade posterior para cada classe, descrita conforme a Equação 24.

$$a_k(x) = \sum_{i=1}^{D} [x_i \ln \mu_{ki} + (1 - x_i) \ln(1 - \mu_{ki})] + \ln p(C_k) \tag{24}$$

onde $a_k(x)$ é uma função linear das características x_i. A classe predita $\hat{y}$ é aquela que maximiza a função de decisão, conforme a Equação 25.

$$\hat{y} = \arg \max_{c \in \{1,2,\ldots,C\}} a_c(x) \tag{25}$$

Para o caso de $K = 2$ classes, a Equação 25 pode ser interpretada também na forma de uma função sigmoide logística. Para variáveis discretas com $M > 2$ possíveis valores, são obtidos resultados análogos.

O classificador SVM é baseado na Teoria de Aprendizado Estatístico (TAE), proposta por Vapnik (1995) e proporciona, matematicamente, auxílio na escolha do classificador, a partir do conjunto de dados de treinamento.

A teoria estabelece que h é um classificador e H um conjunto de todos os classificadores que um algoritmo de aprendizado de máquina pode gerar, dado que $\mathbf{X}$ trata do conjunto de treinamento composto de n pares $(\mathbf{xo}_i, yr_i)$ para geração de um classificador particular $\widehat{h} \in H$ (CORTES; VAPNIK, 1995; FACELI et al., 2011).

A dimensão de Vapnik-Chervonenkis (VC) é uma medida da capacidade do conjunto de hipóteses para ajustes de diferentes conjuntos de dados. Sendo assim, um hiperplano é considerado ótimo, na geração de um classificador linear, quando a busca de um hiperplano possui a margem ρ elevada e poucos erros marginais, o que possibilita a minimização do erro sobre os dados, tanto de treinamento, quanto de novos dados. Entende-se por classificadores lineares os algoritmos que categorizam os dados em diferentes classes por meio de uma função linear, cujo hiperplano (ou uma linha, no caso de duas dimensões) melhor separa as diferentes classes.

As SVMs lineares possuem duas definições, sendo a primeira com margens rígidas e a segunda com margens suaves. SVM linear de margem rígida define-se por fronteiras lineares a partir de dados linearmente separáveis.

A partir da Equação 26 do hiperplano, apresenta-se o produto escalar entre os vetores $\mathbf{w}$ e $\mathbf{xo}, \mathbf{w} \in \mathbf{X}$, descrito por $\mathbf{xo}.\mathbf{w}$, onde $\mathbf{X}$ é o vetor normal ao hiperplano descrito e $\frac{b}{||\mathbf{w}||}$ é a distância em relação à origem, dado $b \in \Re$, sendo b o termo conhecido como o viés (ou interceptação) do hiperplano que representa a distância do hiperplano à origem em direção perpendicular ao vetor de peso w. Então, dado o espaço xo, pela Equação 26, têm-se duas regiões, onde $\mathbf{w}.\mathbf{xo} + b > 0$ e $\mathbf{w}.\mathbf{xo} + b < 0$, dando origem a uma função sinal $g(xo) = sgn(f(xo))$ a ser utilizada nas classificações, de acordo com a Equação 27.

$$f(\mathbf{xo}) = \mathbf{w}.\mathbf{xo} + b = 0 \tag{26}$$

$$g(\mathbf{xo}) = sgn(f(\mathbf{xo})) = \begin{cases} +1 \text{ se } \mathbf{w}.\mathbf{xo} + b > 0 \\ \text{-1 se } \mathbf{w}.\mathbf{xo} + b < 0 \end{cases} \tag{27}$$

Na Figura 18 em (a) está representada a SVM linear com margem rígida, sendo o hiperplano $H_1 : \mathbf{w}.\mathbf{xo} + b = +1$ e $\mathbf{xo}_1$ como seu respectivo ponto, assim como $\mathbf{xo_2}$ um ponto no hiperplano $H_2 : \mathbf{w}.\mathbf{xo} + b = -1$. A Equação 28 define a projeção de $\mathbf{xo_1} - \mathbf{xo_2}$ na direção de $\mathbf{w}$ e perpendicular ao hiperplano separador, definido por $\mathbf{w}.\mathbf{xo} + b = 0$, possibilitando a obtenção da distância entre H_1 e H_2. Adicionalmente, a Equação 29 define as restrições para que não haja dados de treinamento entre as margens de separação das classes para a SVM de margens rígidas.

Em (b) está representada a SVM Linear com Margem Suave, que se caracteriza por suavizar as margens do classificador linear, permitindo que dados possam permanecer

entre os hiperplanos H_1 e H_2, bem como a ocorrência de erros de classificação. Para permitir tal suavização, há violação das restrições definidas na Equação 29. Neste caso, é inserida uma variável de folga ξ_i para todo $i = 1, ..., n$, conforme a Equação 30. Ainda na Figura 18 (b), são ilustrados, por pontos coloridos, os possíveis Vetores de Suporte (VSs). Os pontos cinza representam os VSs livres, os pretos indicam os VSs limitados, os pretos com borda representam os VSs limitados que são erros de treinamento, e os brancos representam classificações corretas.

$$(\mathbf{xo_1} - \mathbf{xo_2}) \left(\frac{\mathbf{w}}{||\mathbf{w}||} . \frac{(\mathbf{xo_1} - \mathbf{xo_2})}{(||\mathbf{xo_1} - \mathbf{xo_2}||)} \right) \tag{28}$$

$$y_i(\mathbf{w}.\mathbf{xo}_i + b) - 1 \geq 0, \forall i = 1, ..., n \tag{29}$$

$$y_i(\mathbf{w}.\mathbf{xo}_i + b) \geqslant 1 - \xi_i, \quad \xi_i \geqslant 0, \forall i = 1, ..., n \tag{30}$$

O SVM não linear (Figura 19) pode lidar com situações em que os dados de treinamento não podem ser divididos de forma satisfatória por um hiperplano.

O hiperplano é capaz de separar objetos e possibilitar o calculo de acordo com a Equação 31. Em seguida, utilizando o SVM Linear com Margens Suaves é possível trabalhar com *outliers* e ruídos presentes nos dados.

$$h(x) = \boldsymbol{w}.\boldsymbol{\Phi}(\boldsymbol{xo}) + b = w_1 xo_1^2 + w_2\sqrt{2}xo_1xo_2 + w_3 xo_2^2 + b = 0 \tag{31}$$

onde $\boldsymbol{\Phi}(\mathbf{xo})$ é a transformação, aplicada aos componentes xo_1 e xo_2 de $\mathbf{xo}$, que realiza um mapeamento não linear para um espaço de características tridimensional $\mathbf{w}$: é o vetor de pesos do classificador SVM, após a transformação de características; xo_1^2 são os componentes transformados: xo_1^2, $\sqrt{2}xo_1xo_2$, e xo_2^2 são características transformadas e combinadas linearmente com os pesos correspondentes w_1, w_2, e w_3, e o termo de polarização b para calcular o valor da função discriminante.

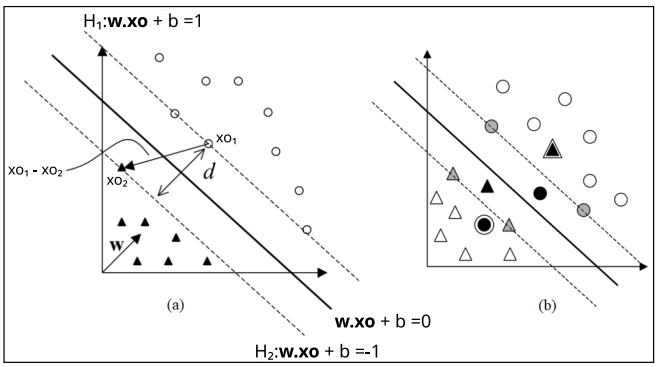
O SVM utiliza funções chamadas *Kernels*, conforme a Equação 32, que possuem a capacidade de representar espaços abstratos recebendo, portanto, dois objetos $\mathbf{xo}_i$ e $\mathbf{xo}_j$, no espaço de entrada, para o cálculo do produto escalar de tais objetos, no espaço de características $\Im$, haja vista que $\Im$ pode alcançar dimensões muito altas e que a computação de Φ pode ser muito custosa.

$$K(xo_i, xo_j) = \boldsymbol{\Phi}(xo_i).\boldsymbol{\Phi}(xo_j) \tag{32}$$

Para que o *kernel* possa representar mapeamentos em que seja possível o cálculo de produtos escalares, conforme função definida na Equação 32, são estabelecidas as condições previstas pelo Teorema de Mercer, que se caracteriza por dar origens a matrizes semidefinidas $\mathbf{k}$, onde cada elemento K_{ij} seja definido por $K_{ij} = K(xo_i, xo_j)$, para todo

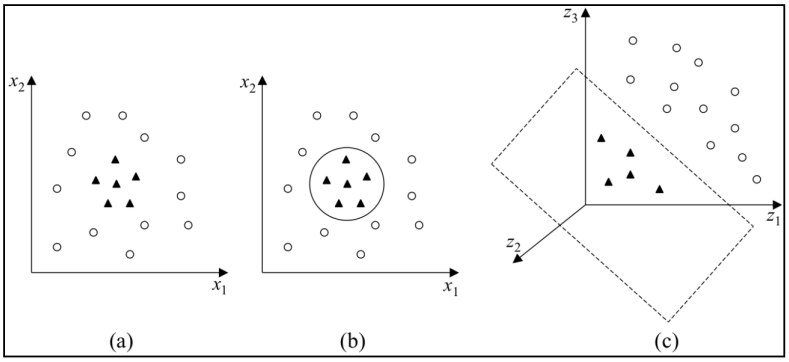
$i, j = 1, ..., n$, sendo que $\mathbf{\Phi}(\mathbf{xo}_i)$ e $\mathbf{\Phi}(\mathbf{xo}_j)$ são as representações de $\mathbf{xo}_i$ e $\mathbf{x}_j$, após a aplicação da função de mapeamento de características $\mathbf{\Phi}(\mathbf{x})$.

Figura 18 – SVM Linear: Margens Rígidas e Margens Suaves



Fonte: Adaptada de (FACELI et al., 2011)

Figura 19 – SVM Não Linear



Fonte: Adaptada de (FACELI et al., 2011)

Na Tabela 3 podem ser observadas as Funções *Kernels* mais comuns, tais como a Função Polinomial, a Função de Base Radial (*Radial Basis Function* - RBF) e a Função Sigmoidal.

Destaca-se que, para o uso de um classificador SVM, a escolha adequada da função *kernel*, assim como a melhor configuração para seus parâmetros e constante de regulação C, são pré-requisitos para a definição de uma melhor fronteira de decisão induzida, o que afeta diretamente o resultado do classificador a ser estabelecido.

A busca por melhores hiperparâmetros, também conhecida pelo processo de *Tunning* dos hiperparâmetros, é uma etapa que antecede o ajuste do modelo de aprendizado de máquina, quanto ao treinamento e teste que confere, ao processo de aprendizado, um melhor resultado de classificação, diante do uso do SVM. Uma maneira eficiente para executar o processo de *Tunning*, evitando experimentações manuais, é o uso do método *Grid Search* por meio da biblioteca *Scikit-Learning* (BUITINCK et al., 2013).

Tabela 3 – Kernels - Classificador SVM

Tipo Kernel	**Função** $k(xo_i, yo_j)$	**Parâmetros**
Linear	$(\delta(\mathbf{xo}_i.\mathbf{yo}_j) + k)^d$	$d = 1$, $\delta = 1$, $k = 0$
Polinomial	$(\delta(\mathbf{xo}_i.\mathbf{yo}_j) + k)^d$	d, δ, k
RBF	$exp(-\theta\|\|\mathbf{xo}_i - \mathbf{yo}_j\|\|^2)$	θ

Fonte: Adaptada de (FACELI et al., 2011)

Para o processamento ser realizado no aprendizado de máquina, são geradas métricas para avaliação do modelo, tais como: Variância (Equação 33), Desvio Padrão (Equação 34), Precisão (Equação 35), Acurácia (Equação 36), Suporte e Revocação (Equação 37), *F1-Score* (Equação 38), Área sob a Curva ROC (Equação 39) composta pelas medidas: TVP - Taxa de Verdadeiro Positivo (Equação 40), TFP - Taxa de Verdadeiro Negativo (Equação 41) e Matriz de Confusão composta pelas medidas: VP - Verdadeiro Positivo, VN -Verdadeiro Negativo, FP - Falso Positivo e FN - Falso Negativo.

$$\sigma^2 = \frac{\sum(xo_i - \bar{xo})^2}{n-1} \tag{33}$$

$$\sigma = \sqrt{(\sigma^2)} \tag{34}$$

$$PREC(f) = \frac{VP}{VP + FP} \tag{35}$$

$$AC(f) = \frac{VP + VN}{n} \tag{36}$$

$$REV(f) = \frac{VP}{VP + FN} \tag{37}$$

$$F1 - Score(f) = \frac{2.prec(f).rev(f)}{prec(f) + rev(f)} \tag{38}$$

$$\begin{cases} ROC(a) = TVP\left\{TFP^{-1}(a)\right\}, a \in (0,1) \\ \\ AUC = \int_0^1 ROC(a)da \end{cases} \tag{39}$$

$$TVP(f) = \frac{VP}{VP + FN} \tag{40}$$

$$TFP(f) = \frac{VN}{VN + FP} \tag{41}$$

Após a definição do classificador, análise e preparação dos dados para a submissão ao mesmo, torna-se tarefa essencial a busca de um resultado satisfatório de classificação.

Neste sentido, os dados de características são analisados no âmbito de sua distribuição para se definir, por meio de estatística descritiva, possíveis transformações dos mesmos, caso sejam necessárias. Ao final desse processo, é possível validar a escolha do *Kernel*, por meio da melhor configuração obtida pelos testes realizados, e aplicar os dados preparados junto ao classificador para treinamento e teste do modelo.

A fim de reduzir a dimensionalidade do vetor de características, é utilizada a técnica de Análise de Componentes Principais (PCA), etapa que também antecede o uso do classificadores árvore de decisão, KNN, *Naïve Bayes* e SVM.

PCA pode ser definida como uma técnica não supervisionada, por tratar dados de alta dimensionalidade. É também conhecida por Transformação de Karhunen-Loève (KARHUNEN, 1947), Transformação de Hotelling (HOTELLING, 1933) ou Decomposição de Valores Singulares (KLEMA; LAUB, 1980).

Essa técnica é caracterizada pelo interesse em encontrar um mapeamento de entradas no espaço d-dimensional, original para um novo espaço $K < d$-dimensional com a perda mínima de informações (JOLLIFFE, 2002).

Por definição, conforme a Equação 42, $Z = [T^T, S^T]$, constitui uma base ortonormal para R^d, onde d representa os componentes retidos, dado $T^T = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_k]$ e $S^T = [\mathbf{v}_{k+1}, \mathbf{v}_{k+2}, ..., \mathbf{v}_d]$, onde T representa o novo subespaço PCA e S o subespaço a ser eliminado, durante o processo de redução da dimensionalidade (LEVADA, 2022).

$$\mathbf{xo} \in R^d = \begin{bmatrix} xo_1 \\ xo_2 \\ ... \\ xo_d \end{bmatrix} \quad \underset{\mathbf{v}_{pca}}{\Rightarrow} \quad \mathbf{yo} \in R^k = \begin{bmatrix} yo_1 \\ ... \\ yo_k \end{bmatrix}, k \ll d \tag{42}$$

Os objetivos do PCA podem ser elencados como a extração das informações mais importantes da tabela de dados; comprimir o tamanho do conjunto de dados, mantendo apenas as informações relevantes; simplificar a descrição do conjunto de dados; e analisar a estrutura das observações e as variáveis (ABDI; WILLIAMS, 2010; LEVADA, 2022).

O PCA efetua o cálculo de novas variáveis chamadas de componentes principais, obtidas como combinações lineares das variáveis originais. A primeira componente principal deve ter a maior variância, ou seja, inércia denotada por I e definida pela soma do quadrado de todos γ_j^2 dos elementos $xo_{i,j}$ da coluna, conforme descrito pela Equação 43. A segunda componente é calculada sob a restrição de ser ortogonal à primeira componente e também deve ter a maior inércia possível, desconsiderando a componente principal calculada. As demais componentes também são calculadas da mesma forma, sendo os valores das novas variáveis, chamados de escores de fator, interpretados como projeções sobre as componentes principais.

$$\gamma_j^2 = \sum_i^I xo_{i,j}^2 \tag{43}$$

Ao considerar a PCA pela maximização da variância, entende-se que, dado um espaço de entrada, é pretendido encontrar as direções $\mathbf{v}_i$ que, na projeção dos dados, possibilita maximizar a variância retida na nova representação.

A Equação 44 representa o conjunto de dados $\mathbf{xo} \in R^d$, como expansão da base ortonormal, onde os c_j são coeficientes da expansão:

$$\mathbf{xo} = \sum_{j=1}^{d} (\mathbf{xo}^T \mathbf{v}_j)\mathbf{v}_j = \sum_{j=1}^{d} c_j \mathbf{v}_j \tag{44}$$

onde $\mathbf{xo}$ representa um vetor de dados originais de alta dimensão; $\mathbf{xo}^T$ é a transposta do vetor coluna de dados originais $\mathbf{xo}$; $\mathbf{v}_j$ é uma componente principal (autovetor); c_j é um coeficiente que indica a projeção do vetor de dados originais $\mathbf{xo}$ na direção da j-ésima componente principal $\mathbf{v}_j$.

A Equação 45 representa a transformação linear T que maximiza a variância retida nos dados, dado $\|\mathbf{v}_j\| = 1$ em que $\mathbf{E}[\mathbf{xoxo}^T] = \mathbf{\Sigma}_{xo}$ e denota a matriz de variância dos dados observados, ou seja:

$$J^{PCA}(T) = \sum_{j=1}^{k} \mathbf{E}[\mathbf{v}_j^T \mathbf{xoxo}^T \mathbf{v}_j] = \sum_{j=1}^{k} \mathbf{v}_j^T \mathbf{E}[\mathbf{xoxo}^T]\mathbf{v}_j = \sum_{j=1}^{k} \mathbf{v}_j^T \mathbf{\Sigma}_{\mathbf{xo}} \mathbf{v}_j \tag{45}$$

onde $\mathbf{E}[\mathbf{v}_j^T \mathbf{xoxo}^T \mathbf{v}_j]$ é o valor esperado de média do produto escalar entre o j-ésimo autovetor e o vetor de dados $\mathbf{x}$. Isso representa a variância explicada pela j-ésima componente principal; $\mathbf{E}[\mathbf{xoxo}^T]$ é a matriz de covariância dos dados originais $\mathbf{xo}$; $\mathbf{\Sigma}_{\mathbf{x}}$ representa a matriz de covariância dos dados originais $\mathbf{xo}$, simétrica e definida como positiva.

A Equação 46 é utilizada para calcular as k direções ortogonais $\mathbf{v}_j$ dado $\|\mathbf{v}_j\| = \mathbf{1}$, para $j = 1, 2, ...k$:

$$\underset{\mathbf{v}_j}{argmax} \sum_{j=1}^{k} \mathbf{v}_j^T \mathbf{\Sigma}_{\mathbf{xo}} \mathbf{v}_j \tag{46}$$

A Equação 47 trata o problema de otimização com restrições de igualdade, onde são aplicados os multiplicadores de Lagrange (λ), incorporada diretamente na função objetivo do PCA:

$$J^{PCA}(T, \lambda_j) = \sum_{j=1}^{k} \mathbf{v}_j^T \mathbf{\Sigma}_{\mathbf{xo}} \mathbf{v}_j - \sum_{j=1}^{k} \lambda_j (\mathbf{v}_j^T \mathbf{v}_j - 1) \tag{47}$$

onde $J^{PCA}(T, \lambda_j)$ representa a função objetivo da PCA, que é uma combinação da variância explicada pelas componentes principais ($\mathbf{v}_j$) e um termo de penalização (λ_j); $lambda_j$ é uma penalização aplicada às componentes principais para controlar sua contribuição, com o objetivo de equilibrar a maximização da variância para evitar que as componentes principais tenham pesos muito altos.

Ao derivar a função em relação a $\mathbf{v}_j$ e igualando o resultado a zero, há que se considerar as questões relacionadas aos autovalores e autovetores que, por sua vez, os vetores de $\mathbf{v}_j$

devem ser autovetores da matriz de covariância $\mathbf{\Sigma_{xo}}$ e os vetores da base PCA representam autovetores de $\mathbf{\Sigma_{xo}}$, de acordo com a Equação 48:

$$\mathbf{\Sigma_{xo}}\mathbf{v}_j = \lambda_j \mathbf{v}_j \tag{48}$$

A Equação 49 trata a otimização do critério definido anteriormente, na Equação 46, maximiza-se a soma dos k autovetores da matriz de covariância, associados aos k maiores autovalores.

$$\underset{\mathbf{v}_j}{argmax} \sum_{j=1}^{k} \mathbf{v}_j^T \mathbf{\Sigma_{xo}} \mathbf{v}_j = \underset{\mathbf{v}_j}{argmax} \sum_{j=1}^{k} \mathbf{v}_j^T \lambda_j \mathbf{v}_j = \underset{\mathbf{v}_j}{argmax} \sum_{j=1}^{k} \lambda_j \left\|\mathbf{v}_j\right\|^2 = \underset{\mathbf{v}_j}{argmax} \sum_{j=1}^{k} \lambda_j \tag{49}$$

onde Q é a matriz dos autovetores de colunas e Λ a matriz diagonal dos autovalores. A matriz A contém os dados originais antes da transformação PCA e deve ser assimétrica, positiva e semidefinida com composição $A = Q\Lambda Q^T$:

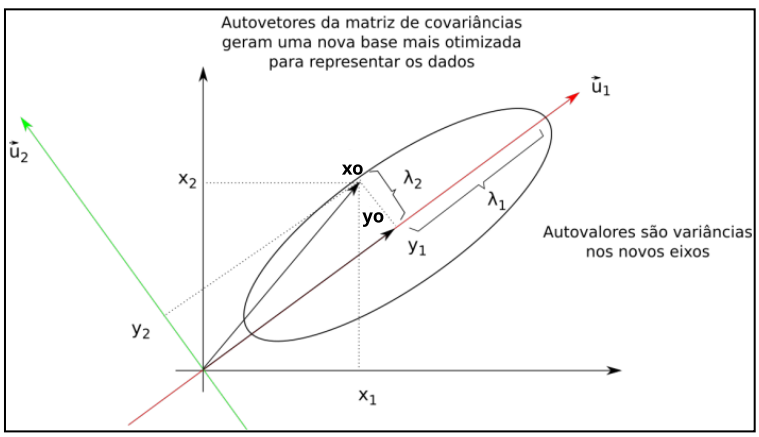
$$\Sigma_{yo} = Q^T Q \Lambda Q^T \ A \tag{50}$$

Após a transformação PCA, os dados encontram-se descorrelacionados, ou seja, a matriz de covariâncias é diagonal, não havendo correlação entre os novos atributos gerados pelo PCA.

A Figura 20 ilustra a interpretação geométrica ao final do processo da transformação PCA para dados em $\Re^2$.

Neste trabalho, as características extraídas das imagens de folhas de soja integram um único vetor de dados que, após ter sua dimensionalidade reduzida com o uso de PCA, é utilizado como entrada para o conjunto de classificadores, para em seguida compor um segundo vetor de decisão que também contempla variáveis climáticas.

Figura 20 – Interpretação geométrica 2D



Fonte: Adaptada de (LEVADA, 2022)

O vetor de dados para o modelo de favorabilidade consiste na construção do vetor de dados para submissão às abordagens de fusão de dados a serem processadas no bloco de fusão de dados e no modelo de favorabilidade.

O processo para a construção do vetor de dados consiste na conexão com o banco de dados relacional Oracle, conforme descrito no bloco de estruturação de dados, para acesso à série temporal de dados climáticos e demais dados das imagens segmentadas. Após o acesso, faz-se a consulta dos dados climáticos para uma janela de tempo. É viável o monitoramento da doença durante o cultivo, no ciclo de cultura, com a utilização das séries temporais de dados e uso de janelamento, para amostragens consecutivas, em subperíodos de 10 dias. O primeiro período de janelamento deve ser informado, considerando a data inicial da ocorrência.

Em uma janela temporal estabelecida são considerados, simultaneamente, além da informação de classificação, decorrente do processamento das imagens, os seguintes dados de um conjunto de 6 variáveis ($V's$) climáticas: $V1$ - Umidade Relativa; $V2$ - Precipitação; $V3$ - Temperatura Máxima; $V4$ - Temperatura Mínima; $V5$ - Ponto de Orvalho e $V6$ - Temperatura Média Compensada. Os dados das 6 variáveis, a partir da janela de tempo definida, são armazenados em um vetor inicial de dados. Consequentemente, faz-se uma verificação se o vetor inicial possui os dados referentes aos 10 dias, pois podem ocorrer dados faltantes na série temporal de dados climáticos, devido à indisponibilidade ou erro de medições pela estação climática que forneceu os dados. Quando a janela temporal apresentar incompletude nos dados, utiliza-se o conceito de interpolação de dados para suprir os dados faltantes. Para tanto, neste trabalho são também consideradas avaliações sobre diferentes métodos de interpolação, ou seja, por polinômios, conforme Davis (1975), de grau 1 até grau 5 e também por *spline* cúbica (GREVILLE, 1969).

Conceitualmente, segundo Ruggiero e Lopes (1997), ao considerar $(np+1)$ pontos distintos $xi_0, xi_1, ...xi_n$, chamados de nós de interpolação e os valores da função $f(xi)$ nesses pontos são considerados $f(xi_0)$, $f(xi_1), ...f(xi_n)$. Assim, a interpolação consiste em determinar a função $g(xi)$, tal que $g(xi_0) = f(xi_0)$; $g(xi_1) = f(xi_1)$; $g(xi_2) = f(xi_2)...g(xi_n) = f(xi_n)$.

Considera-se na interpolação polinomial (Equação 51) a forma geral da função interpoladora $PI(xi)$ e, na Equação 52, a forma de Newton para o polinômio $PI_n(xi)$ que interpola $f(xi)$ nos pontos distintos $xi_0, xi_1, ..., xi_n$.

$$PI(xi) = ca_0 + ca_1 xi + ca_2 xi^2 + ca_3 xi^3 + ... + ca_n xi^n \tag{51}$$

onde np é o número de pontos; xi um ponto arbitrário; $n-1$ o grau do polinômio e $(ca_0, ca_1, ..., ca_n)$ são os coeficientes.

$$\begin{aligned} PI_n(xi) = f(xi_0) + d_1(xi - xi_0) + d_2(xi - xi_0)(xi - xi_1) \\ + ... + d_{np}(xi - xi_0)(xi - xi_1)...(xi - xi_{np-1}) \end{aligned} \tag{52}$$

onde $d_k = f[xi_0, xi_1, ..., xi_k]$; *para* $0 \leq k \leq np$; $f(xi_0)$ é o valor da função $f(xi)$ no ponto inicial x_0, quando o polinômio começa a interpolar os dados; $d_1, d_2, \ldots, d_n$ são as diferenças divididas de Newton, ou seja, coeficientes que são calculados para determinar o comportamento do polinômio interpolador. Cada d_i é calculado com base nas diferenças divididas anteriormente e nos pontos de dados: $xi, xi_0, xi_1, \ldots, xi_{np-1}$ são os valores de xi que correspondem aos pontos de dados nos quais desejamos realizar a interpolação; xi_0 é o ponto inicial a partir do qual a interpolação começa.

Em contrapartida, a *spline* cúbica é uma função polinomial constituída por partes contínuas. Logo, cada parte é composta por um polinômio de grau 3 no intervalo $[xi_{k-1}, xi_k], k = 1, 2, 3, ..n$. Além disso, é caracterizada também por se obter uma fórmula de interpolação que seja suave na primeira derivada e contínua na segunda, tanto dentro de um intervalo, quanto em seus limites (GREVILLE, 1969).

A *spline* cúbica se mostrou mais utilizada frente à *spline* quadrática, haja vista que a quadrática possui derivadas contínuas somente de ordem 1 que, por consequência, a curvatura pode tocar os nós. Por outro lado, a *spline* cúbica tem a primeira e segunda derivadas contínuas, fazendo com que a curva não tenha picos e nem toque de forma abrupta a curvatura dos nós.

Formalmente, ao saber que a função $f(xi)$ está tabelada nos pontos $xi_k, k = 0, 1, 2, ..., np$ a função $SC_3(xi)$ é chamada de *spline* cúbica interpolante de $f(xi)$, dado os nós $xi_k, k = 0, ..., np$ e se existem np polinômios de grau 3, onde $SC_k(xi), k = 1, ..., np$. Na Equação 53 é exibida a notação simplificada da *spline* cúbica.

$$SC_k(xi) = ca_k(xi - xi_K)^3 + cb_k(xi - xi_K)^2 + cc_k(xi - xi_K) + cd_k, \ onde\ k = 1, 2, ..., np \quad (53)$$

Entretanto, para o cálculo de $SC_3(xi)$ são necessários 4 coeficientes para cada k, totalizando $4n$ coeficientes: $ca_1, cb_1, cc_1, cd_1, ca_2, cb_2, ..., ca_{np}, cb_{np}, cc_{np}, cd_{np}$.

Adicionalmente, apresenta-se o conceito e representação formal da *b-spline* que, também de acordo com Greville (1969) e expressa pela Equação 54, trata-se de uma particularidade das *splines*, cujo significado se traduz em *basis spline*.

$$\Gamma(xi) = \sum_{i=0}^{n-1} c_i B_{i,g;t}(\iota) \quad (54)$$

onde c_i são os coeficientes, g a ordem da *b-spline*, t são os nós e $B_{i,g}(\iota)$ definida pelas Equações 55 e 56.

$$B_{i,0}(xi) = \begin{cases} 1, & se \quad t_i \leq \iota < t_{i+1} \\ 0, & se \quad caso\ contrário \end{cases} \quad (55)$$

$$B_{i,k}(xi) = \frac{x - t_i}{t_{i+k} - t_i} B_{i,k-1}(\iota) + \frac{t_{i+k+1} - xi}{t_{i+k+1} - t_{i+1}} B_{i+1,k-1}(xi) \quad (56)$$

Após a apresentação dos principais conceitos e equações para a interpolação polinomial, *spline* cúbica e sua particularidade, *b-spline* cúbica, destaca-se que no desenvolvimento do processo de interpolação para o sistema, estão sendo aplicados tais conceitos. Isso possibilita a análise e a decisão do melhor método de interpolação para o ajuste mais adequado da curva, considerando os dados faltantes das variáveis climáticas supracitadas.

Após a obtenção do conjunto completo de dados, a partir do uso da interpolação quando necessário, o próximo passo é a submissão dos mesmos à base de regras elaborada, conforme a Tabela 4, a qual descreve as condições climáticas para a favorabilidade da FAS (conforme visão especialista e dados avaliados da literatura).

Tabela 4 – Base de Regras - Ferrugem Asiática da Soja

Condições Climáticas para Favorabilidade da Ferrugem Asiática da Soja

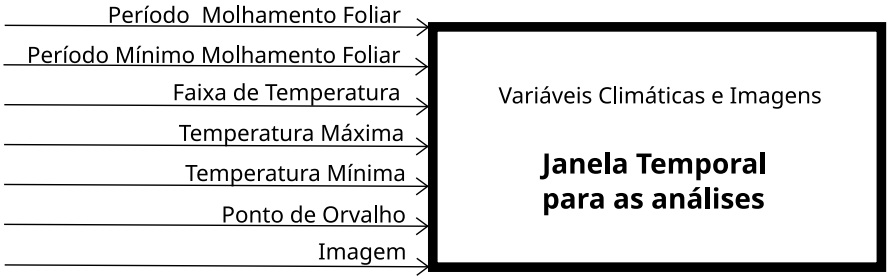
Descrição	**Variável**	**Valor estimado**
Dados Climatológicos Conhecidos		
Período de molhamento foliar	Quantidade de horas	Umidade relativa maior ou igual a 90%
Ponto de orvalho	Temperatura	Diferença menor que $2°C$
Faixa de temperatura favorável ao desenvolvimento do fungo	Temperatura	Faixa entre 18 a $25°C$
Temperatura Mínima e Máxima no período de molhamento foliar	Faixa de temperatura	Faixa entre 18 a $26{,}5°C$
Período mínimo de molhamento foliar	Tempo	6 horas
Novos Dados Apresentados		
Dados do Cultivar: Folha da Soja	Classificação	Análise dos pixels
Fenomenologia do Problema da Ferrugem Asiática da Soja	Descoberta da Classes de Cores	Análise dos pixels verde, amarelo e marrom
Identificação do estágio da doença	Percentual de ocorrência das classes	Quantidade de pixels de cada classe
Probabilidade de Favorabiliade	Conjunto das variáveis dos indicadores	Baixo, Média e Alto

Fonte: Próprio Autor

O processamento dos dados na base de regras considera a fenomenologia do problema da doença, sendo iniciado com a seleção da janela temporal disponível na série histórica de dados climáticos. Neste intervalo temporal, tem-se, como entrada, as variáveis de V1 a V6 e os dados de classificação do processamento da imagem da folha de soja. Essa imagem foi processada anteriormente e baseia-se na ocorrência da favorabilidade, de acordo com a análise dos *pixels* verdes, amarelos e marrons, conforme representado na Figura 21.

Para o modelo de decisão considerado, é estabelecida a base de regras que envolve as variáveis climáticas de V1 a V6 e uma variável adicional V7, que corresponde aos dados de classificação de padrões de imagens, obtidas de folhas da soja. O detalhamento das regras constam nos Algoritmos 3, 4, 5, 6, 7, 8, 9, onde estão incluídas as faixas que correspondem às diferentes situações que orientam a favorabilidade de ocorrência da FAS.

Figura 21 – Variáveis Projetadas na Janela Temporal



Fonte: Próprio Autor

Algoritmo 3 Pseudocódigo para Elaboração da Regra 1

Entrada: Dados Umidade Relativa (Janela Temporal)
Saída: Período de Molhamento Foliar

início
  se *dad_sel* ← *umidade relativa* ≤ 90% **então**
    dad_sel ← *dados_regra_1*;
    favorabilidade ← alta;
  fim
  se (*dad_sel* ← ∅) *E (umidade relativa* ≤ *80% E* ≥ *90%)* **então**
    dad_sel ← *dados_regra_1*;
    favorabilidade ← média;
  fim
  se (*dad_sel* ← ∅) *E (umidade relativa* < *80% E* > *0%)* **então**
    dad_sel ← *dados_regra_1*;
    favorabilidade ← baixa;
  fim
fim
retorna m_d*(período de molhamento foliar), favorabilidade: baixa, média ou alta, dados_regra_1*;

Algoritmo 4 Pseudocódigo para Elaboração da Regra 2

Entrada: Dados Precipitação (Janela Temporal)
Saída: Período Mínimo de Molhamento Foliar

início
  se *dad_sel* ← *precipitação* ≤ 25% **então**
    dados_regra_2 ← *dad_sel*;
    favorabilidade ← alta;
  fim
  se (*dad_sel* ← ∅) *E (umidade relativa* < *25% E* ≥ *20%)* **então**
    dados_regra_2 ← *dad_sel*;
    favorabilidade ← média;
  fim
  se (*dad_sel* ← ∅) *E (precipitação* < *20% E* > *0%)* **então**
    dados_regra_2 ← *dad_sel*;
    favorabilidade ← baixa;
  fim
fim
retorna m_d*(período mínimo de molhamento foliar), favorabilidade: baixa média ou alta, dados_regra_2*;

Algoritmo 5 Pseudocódigo para Elaboração da Regra 3

Entrada: Dados Temperatura Mínima, Temperatura Máxima (Janela Temporal)
Saída: Faixa de Temperatura

início
  se *(dad_sel ←temperatura mínima* $\leq 18\%$*)* E *(temperatura máxima* $\geq 26,5\%$*)* **então**
    dados_regra_3 ← *dad_sel*;
    m_d faixa inicial ← dados regra 3 [temperatura mínima];
    m_d faixa final ← dados regra 3 [temperatura máxima];
    favorabilidade ← alta;
  fim
  se *dados_regra_3* ← ∅ **então**
    se (*dad_sel ←temperatura mínima* $\leq 15,1\%$*)* E *(temperatura máxima* $\geq 17,9\%$*)* **então**
      dados_regra_3 ← *dad_sel*;
      m_d faixa inicial ← dados regra 3 [temperatura mínima];
      m_d faixa final ← dados regra 3 [temperatura máxima];
      favorabilidade ← média;
    fim
  fim
  se *dados_regra_3* ← ∅ **então**
    se (*dad_sel ←temperatura mínima* $\leq 0\%$*)* E *(temperatura máxima* $\geq 15\%$*)* **então**
      dados_regra_3 ← *dad_sel*;
      m_d faixa inicial ← dados regra 3 [temperatura mínima];
      m_d faixa final ← dados regra 3 [temperatura máxima];
      favorabilidade ← baixa;
    fim
  fim
  se *dados_regra_3* ← ∅ **então**
    dados_regra_3 ← 0;
    m_d faixa inicial ← 0;
    m_d faixa final ← 0;
    favorabilidade ← baixa;
  fim
fim
retorna m_d*(faixa inicial e faixa final), favorabilidade: baixa média ou alta, dados_regra_3*;

Algoritmo 6 Pseudocódigo para Elaboração da Regra 4

Entrada: Dados Temperatura Máxima (Janela Temporal)
Saída: Temperatura Máxima - Regra 4

início
  v_temp_max ← max(temperatura máxima);
  se $v\_temp\_max \leq 18$ *&* $\geq 26,5$ **então**
    favorabilidade ← alta;
    valor_regra_4 ← 1;
  fim
  senão se $(v\_temp\_max \leq 15,1$ *&* $\geq 17,9)|(v\_temp\_max \leq 26,5$ *&* $\geq 30)$ **então**
    favorabilidade ← média;
    valor_regra_4 ← 1;
  fim
  senão se $(v\_temp\_max > 0$ *&* $\geq 15)|(v\_temp\_max > 30$ *&* $\geq 42)$ **então**
    favorabilidade ← baixa;
    valor_regra_4 ← 1;
  fim
  senão se $(v\_temp\_max == 0))|(v\_temp\_max > 42)$ **então**
    favorabilidade ← baixa;
    valor_regra_4 ← 0;
  fim
fim
retorna *favorabilidade: baixa média ou alta, valor_regra_4*;

Algoritmo 7 Pseudocódigo para Elaboração da Regra 5

Entrada: Dados Temperatura Mínima (Janela Temporal)
Saída: Temperatura Mínima - Regra 5

início
  $v\_temp\_min \leftarrow$ min(temperatura mínima);
  se $v\_temp\_min \leq 18$ *&* $\geq 26,5$ **então**
    favorabilidade $\leftarrow$ alta;
    v_regra_5 $\leftarrow 1$;
  fim
  senão se $(v\_temp\_min \leq 15,1$ *&* $\geq 17,9) | (v\_temp\_max \leq 26,5$ *&* $\geq 30)$ **então**
    favorabilidade $\leftarrow$ média;
    v_regra_5 $\leftarrow 1$;
  fim
  senão se $(v\_temp\_max \leq 0$ *&* $\geq 15) | (v\_Temp\_max > 30)$ **então**
    favorabilidade $\leftarrow$ baixa;
    valor_regra_5 $\leftarrow 1$;
  fim
  senão se $(v\_Temp\_Max \geq 0)$ **então**
    favorabilidade $\leftarrow$ baixa;
    valor_regra_5 $\leftarrow 0$;
  fim
fim
retorna *favorabilidade: baixa média ou alta, valor_regra_5*;

Algoritmo 8 Pseudocódigo para Elaboração da Regra 6

Entrada: Temperatura Média Compensada e Ponto de Orvalho (Janela Temporal)
Saída: Diferença Temperatura Média Compensada e Ponto de Orvalho - Regra 6

início
  $temp\_med\_comp \leftarrow$ Dados Temperatura Média Compensada;
  $p\_orvalho \leftarrow$ Dados Ponto de Orvalho;
  $dad\_sel \leftarrow (temp\_med\_comp - p\_orvalho \geq 2)$
  se $dad\_sel == \varnothing$ **então**
    favorabilidade $\leftarrow$ baixa;
    $m_d(temp\_med\_comp - p\_orvalho \leftarrow 0)$;
    $dados\_regra\_6 \leftarrow dad\_sel$;
  senão
    favorabilidade $\leftarrow$ alta;
    $m_d(temp\_med\_comp)$;
    $m_d(p\_orvalho)$;
    $m_d(temp\_med\_comp - p\_orvalho)$;
    $dados\_regra\_6 \leftarrow dad\_sel$;
  fim
fim
retorna $m_d(temp\_med\_comp - p\_orvalho)$, *favorabilidade: baixa média ou alta,* $dados\_regra\_6$;

Algoritmo 9 Pseudocódigo para Elaboração da Regra 7

Entrada: Resultado Classificação, Imagens Segmentadas (Classificação)
Saída: Resultado da Classificação - Regra 7

```
1  início
2      result_classif ← (int) Resultado Classificação;
3      img_seg_class[] ← Imagens Segmentadas;
4      se result_classif ≠ ∅ então
5          se result_classif == 1 então
6              favorabilidade ← alta;
7              valor_regra_7 ← 1;
8          fim
9          senão se result_classif == 0 então
10             favorabilidade ← baixa;
11             valor_regra_7 ← 0;
12         fim
13         para i = 0 até n == 2 faça
14             Mostrar_imagens ← img_seg_class[i];
15         fim
16     senão
17         Mostrar_falha_classificação;
18     fim
19 fim
20 retorna favorabilidade: baixa, média ou alta; valor_regra_7;
```

Diante da finalização do processamento das regras, o vetor de dados é carregado com os dados prontos para serem submetidos à entrada dos processos da etapa de fusão de dados, conforme apresenta a Tabela 5.

Tabela 5 – Vetor de Dados - Entrada Fusão de Dados

Dados Processados das Regras de 1 a 7			
Regra	**Descrição**	**Valor**	**Favorabilidade**
1	Período de Molhamento Foliar	90,00	Alta
2	Período Mínimo de Molhamento Foliar	2,40	Baixa
3	Faixa Temperatura	20,00 a 25,80	Alta
4	Temperatura Máxima	34,70	Baixa
5	Temperatura Mínima	17,60	Baixa
6	Ponto de Orvalho	1,80	Alta
7	Imagem	0	Baixa

Fonte: Próprio Autor

A Fusão de Dados e o modelo de favorabilidades são baseados na concepção dos dados originados do aprendizado de máquina e dos dados das séries temporais climáticas.

O processo de fusão é fundamentado na integração das variáveis originadas destas diferentes fontes e grandezas físicas normalizadas, conforme apresentado na Tabela 6.

Tabela 6 – Variáveis e Grandezas Físicas - Fusão de Dados

Id.	**Descrição das Variáveis**	**Grandeza Física**
V1	Umidade Relativa	Porcentagem (%)
V2	Precipitação	Milímetros (mm)
V3	Temperatura Máxima	Graus Celsius (°C)
V4	Temperatura Mínima	Graus Celsius (°C)
V5	Ponto de Orvalho	Graus Celsius (°C)
V6	Temperatura Média Compensada	Graus Celsius (°C)
V7	Classificação de Padrões (Imagem Folha de Soja)	Admensional (Classificação 0 ou 1)

Fonte: Próprio Autor

Neste trabalho são avaliadas três deferentes técnicas para a fusão de dados, sendo elas Figura de Mérito (CRUVINEL et al., 2011; BENDINI et al., 2013; CRUVINEL, 2022), uso de Lógica Difusa (BRESSAN et al., 2006; LIU et al., 2019; ZAGUI et al., 2022) e Cadeias Ocultas de Markov (BAUM, 1972; BOUDAREN; PIECZYNSKI, 2016; LI et al., 2017).

A técnica que utiliza o conceito de figura de mérito (CRUVINEL et al., 2011) prevê, que as probabilidades normalizadas de cada variável sejam plotadas nos eixos de uma circunferência de raio unitário. Assim, o cálculo da favorabilidade total corresponde à área da figura poligonal formada pela união dos vértices, determinada pelo somatório de n variáveis das áreas dos triângulos formados dentro da figura. Uma vez conhecido um dos ângulos e os dois lados de cada triângulo, a área total da figura pode ser determinada pela Equação 57:

$$AR = \sum_{i=1}^{n} \frac{la_i lb_i sen\Theta_i}{2} \tag{57}$$

onde AR é área total da figura de mérito, la e lb são os lados conhecidos dos triângulos e Θ é o ângulo formado entre os vetores.

A partir da área total da figura de mérito, define-se a favorabilidade total, como a relação da intersecção entre as possibilidades normalizadas de ocorrência da favorabilidade entre as variáveis, conforme descreve a Equação 58.

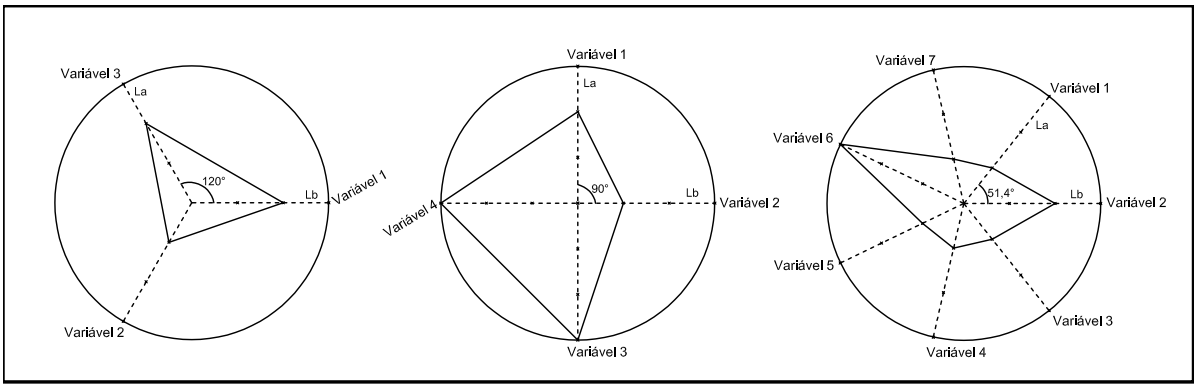
$$OF = O(var_1) \cap O(var_2) \cap O(var_3)...O(var_n) \tag{58}$$

onde OF é a ocorrência de favorabilidade resultante da intersecção entre as possibilidades normalizadas de todas as variáveis; $O(var_1) \cap O(var_2) \cap O(var_3)...O(var_n)$ representam a intersecção das ocorrências das favorabilidades das variáveis.

Adicionalmente, o círculo trigonométrico de raio unitário é dividido em partes equivalentes pela quantidade n de variáveis. Tal divisão determina o ângulo Θ para cada triângulo, cujo cálculo consiste em dividir o ângulo total da circunferência (360°) pela quantidade de variáveis n.

Em cada eixo traçado, do centro do círculo trigonométrico até a borda, marca-se um ponto de acordo com o valor correspondente à variável, na escala normalizada de 0 a 1. Cada eixo está dividido em três marcações para representar a magnitude da variável, se baixa, média e alta com os valores normalizados de 0,333 para 33,3%, 0,666 para 66,6% e 1 para 100%, respectivamente. A figura de mérito que forma no conjunto dos eixos traduz a fusão dessas variáveis, ou seja, quando todos os eixos estiverem marcados, com seus respectivos pontos. Consequentemente, o estágio de favorabilidade à ocorrência da doença será baixa, média ou alta, conforme ilustra a Figura 22, sob diferentes exemplos, assim como o Algoritmo 10 que corresponde à lógica da abordagem.

Figura 22 – Exemplos de Figuras de Mérito no Círculo Trigonométrico



Fonte: Adaptado de (CRUVINEL et al., 2011)

Por outro lado, a lógica difusa definida por Zadeh (1965), que também pode ser utilizada para fusão de dados de sensores, é definida como uma teoria clássica de conjuntos. Dado um objeto χ, este pode pertencer ou ser membro de um conjunto α, ou não pertencer a esse conjunto, ou seja, não ser membro. Estas duas opções são denotadas por $\chi \in \alpha$ ou $\chi \notin \alpha$.

Em continuidade às definições da lógica difusa, um conjunto clássico pode ser descrito por uma função característica X_α que assume dois valores, de acordo com a Equação 59, onde o valor "1" significa que o objeto pertence ao conjunto A e o valor "0" significa que o objeto não pertence ao conjunto α.

$$X_\alpha(\chi) = \begin{cases} 1, \chi \in \alpha, \\ 0, \chi \notin \alpha. \end{cases} \tag{59}$$

Algoritmo 10 Pseudocódigo Abordagem Figura de Mérito no Circulo Trigonométrico

Entrada: Regras de 1 a 7
Saída: Favorabilidade

início
  $dados\_regras[] \leftarrow flags\_favorabilidade$;
  $angulo \leftarrow 360/num\_variaveis$;
  $angulos[] \leftarrow 0$;
  $divisao \leftarrow num\_variaveis$;
  $raio \leftarrow 1$;
  $quadrantes \leftarrow 0, 90, 180, 270$;
  $cores[vermelho, amarelo, verde, azul, roxo, ciano, laranja]$;
  para $num \leftarrow 0$ *até* $num\_variaveis$ **faça**
    $angulos[num] \leftarrow 360 - (angulo \times num)$;
  fim
  Desenhar_circulo(raio):
  retorna *circulo*;
  Desenhar_quadrantes(circulo, quadrantes):
  retorna *circulo_quad*;
  Desenhar_linhas(circulo_quad, angulos[], divisao, cores[]):
  retorna *circulo_div_color, pontos_divisao[]*;
  Marcar_pontos(circulo_div_color, pontos_divisao[], dados_regras[]):
  para cada (*linha in circulo_div_color*) **faça**
    $circulo\_dividido \leftarrow dados\_regras[linha]$;
  fim
  retorna *circ_div_color_marca*;
  Desenhar_figura_merito(circ_div_color_marca):
  para cada $ponto \leftarrow 0$ *in* $marcas[]$ **faça**
    $desenhar\_figura \leftarrow marcas[ponto]$;
    $calcular\_triangulos \leftarrow marcas[ponto]$;
  fim
  retorna *circulo_figura_merito, triangulos[]*;
  Calcular_area_figura(triangulos[]):
  para cada *triangulo in triangulos*[] **faça**
    $calcular\_area\_triangulos \leftarrow triangulos[triangulo]$;
  fim
  retorna *area_total_calculada*;
  Calcular_favorabilidade(area_total_calculada):
  se $area\_total\_calculada \leq 33,3\%$ **então**
    $favorabilidade \leftarrow baixa$;
  fim
  senão se $(area\_total\_calculada \geq 33,4\%)\ E\ (\leq 66,6\%)$ **então**
    $favorabilidade \leftarrow média$;
  fim
  senão se $(area\_total\_calculada \geq 66,6\%)\ E\ (\leq 100\%)$ **então**
    $favorabilidade \leftarrow alta$;
  fim
fim
retorna *favorabilidade*;

Assim, conforme as definições apresentadas por Prokopowicz e colaboradores, as operações básicas de Produto (intersecção e conjunção), Soma (união e disjunção) e Negação (complemento) são aplicadas na abordagem de lógica difusa para o desenvolvimento do sistema. Formalmente, tais operações são definidas, consecutivamente, nas Equações 60, 61 e 62 (PROKOPOWICZ et al., 2017).

$$\alpha \cap \beta = \{\chi \in \mathbb{X} | \, \chi \in \alpha \; and \; \chi \in \beta\} \tag{60}$$

onde o objeto χ pertence ao universo $\mathbb{X}$, tal que o objeto χ pertence ao conjunto α e β.

$$\alpha \cup \beta = \{\chi \in \mathbb{X} | \, \chi \in \alpha \; or \; \chi \in \beta\} \tag{61}$$

onde o objeto χ pertence ao universo $\mathbb{X}$, tal que o objeto χ pertence ao conjunto α ou β.

$$\bar{\alpha} = \{\chi \in \mathbb{X} | \, \chi \notin \alpha\} \tag{62}$$

onde o objeto χ pertence ao universo $\mathbb{X}$, tal que o objeto χ não pertence ao conjunto α.

Os conjuntos difusos também são descritos usando funções de pertinência $\mu\alpha : \mathbb{X} \rightarrow [0, 1]$ que, ao contrário de um conjunto clássico α, um objeto χ pode pertencer a esse conjunto, a partir de graus de pertinência no intervalo do valor "0", com significado de sem pertinência, até "1" com pertinência plena (PROKOPOWICZ et al., 2017).

Consideram-se como as quatro principais funções de pertinências: (1) Função de Pertinência Gaussiana; (2) Função de Pertinência Trapezoidal; (3) Função de Pertinência Triangular; e (4) Função de Pertinência *Singleton*. Neste trabalho, a função de pertinência utilizada no sistema, está descrita na Equação 63, e representada na Figura 23, dado que $L(\chi)$ é uma função crescente estritamente contínua com $L(a) = 0$, $L(b) = 1$ e $R(\chi)$ uma função contínua estritamente decrescente com $R(b) = 1$, $R(c) = 0$ (PEDRYCZ, 1994).

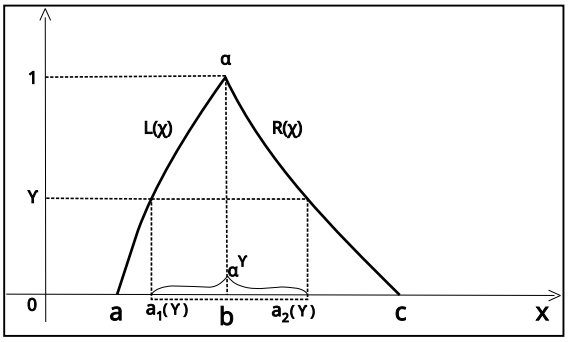
$$\mu_\alpha(\chi) = \begin{cases} 0, & if \; \chi < a \\ L(\chi), & if \; a \leq \chi \leq b \\ R(\chi), & if \; b \leq \chi \leq c \\ 0, & if \; \chi > c \end{cases} \tag{63}$$

Adicionalmente, um universo discreto $\mathbb{X}$ é definido como objetos ordenados ou não ordenados, de acordo com a Equação 64, segundo Prokopowicz e colaboradores. Esses autores também relatam que uma regra difusa condicional pode ser definida como uma relação difusa, assim como conjuntos difusos são as declarações antecedentes e consequentes (PROKOPOWICZ et al., 2017).

$$\alpha = \sum_{\chi \in \mathbb{X}} \mu_\alpha(\chi)/\chi \tag{64}$$

onde $\mu_\alpha(\chi)$ e χ representam o grau de pertinência do par de objetos χ, sendo que o simbolo "/" denota o separador do par e $\sum$ a sumarização idempotente.

Figura 23 – Função de Pertinência Triangular



Fonte: Adaptada de (PEDRYCZ, 1994)

Afirma-se, ainda segundo Prokopowicz e colaboradores, que uma regra difusa condicional pode ser definida como uma relação difusa, assim como as declarações dos antecedentes e consequentes como conjuntos difusos. Uma declaração X é um fato $L_{\alpha'}$, onde $L_{\alpha'}$ denota o rótulo de uma variável linguística X, dado um conjunto difuso α' em o universo $\mathbb{X}$. Sabe-se também que o conhecimento é representado por uma regra condicional difusa se X for L_α então Y é L_β, onde que $L\alpha$ e $L\beta$ são os valores linguísticos de variáveis linguísticas X e Y, definidos pelos conjuntos difusos α e β, nos universos $\mathbb{X}$ e $\mathbb{Y}$, respectivamente.

No contexto da construção do sistema, as entradas são definidas como dados *crisp* ou precisos, representados pelos dados climáticos e pelas imagens das folhas de soja. Porém, para compor um raciocínio aproximado, são utilizadas tais entradas para representar conjuntos difusos.

Por consequência, o processo de mapeamento de valores *crisp* reais, formalmente representado por $\alpha_0 = [\alpha_{01}, \alpha_{02}, \ldots, \alpha_{0n}]^T \in \underline{\mathbb{X}} \subset \mathbb{R}^N$ para um conjunto difuso, também formalmente representado por α' $N-$dimensional em $\underline{\mathbb{X}}$, é conhecido por fuzzificação.

Baseada na abordagem *Tsukamoto Fuzzy System* (TSUKAMOTO, 1979), a base de conhecimento é uma coleção de declarações condicionais difusas, dado que $f_i(y)$ é uma função monotônica no i-ésimo consequente, conforme a Equação 65.

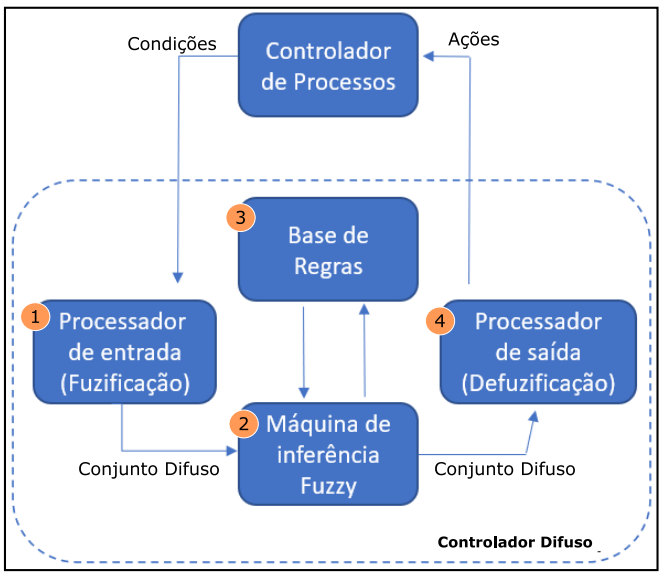
$$\delta^{(i)} = if \bigwedge_{n=1}^{m} \left(\mathbf{xo}_{0n} \text{ is } L_{\alpha_n}^{(i)}\right), \quad then \quad \zeta = f_i^{-1}\left(F^{(i)}\left(\mathbf{xo}_0\right)\right) \tag{65}$$

onde $\delta^{(i)}$ este é o resultado da i-ésima regra difusa, que representa a conclusão ou consequente da regra; $\bigwedge_{n=1}^{m}$ é o operador de interseção $\bigwedge$, usado para combinar as condições do antecedente da regra; x_{0n} são as variáveis de entrada para a regra e $\mathbf{xo}_0$ são as variáveis de entrada para as quais a regra difusa está sendo avaliada; $L_{\alpha_n}^{(i)}$ é o conjunto difuso correspondente à variável de entrada xo_{0n} na i-ésima regra, representando a pertinência da entrada ao conjunto difuso; ζ é o resultado da regra difusa que representa a saída ou

ação a ser tomada; f_i^{-1} é a função inversa da i-ésima regra difusa usada para calcular a saída ζ com base na função de pertinência $F^{(i)}$ da variável de entrada $\mathbf{xo}_0$; $F^{(i)}$ é a função de pertinência difusa da variável de entrada $\mathbf{xo}_0$ na i-ésima regra. $F^{(i)}$ atribui graus de pertinência aos valores de entrada em relação ao conjunto difuso $L_{\alpha_n}^{(i)}$.

Na Figura 24 é mostrado o esquema do modelo difuso, baseado na abordagem de Mamdani e Assilian (1999), apresentado em duas partes. A primeira é o controlador de processos, e a segunda parte o controlador difuso composto por quatro blocos: (1) processador de entrada que recebe as condições do controlador de processos e responsável pelo processo de fuzzificação dos dados; (2) máquina de inferência que recebe o conjunto de dados difusos e interage com a base de regras, processando as inferências; (3) base de regras que armazena as inferências geradas pelo especialista da fenomenologia do problema e (4) processador de saída que recebe o conjunto de dados difusos processados pela máquina de inferências e responsável pelo processo de defuzzificação dos dados, resultado enviado para o controlador de processos.

Figura 24 – Esquema do Modelo Difuso de Mamdani



Fonte: Adaptada de (MAMDANI; ASSILIAN, 1999)

Os conceitos de lógica difusa estão organizados em um modelo difuso, dado pela configuração das variáveis antecedentes e consequentes, exibidos na Tabela 7.

Neste trabalho é considerado uma base de regras para o modelo difuso de favorabilidade, elaborado com base nas configurações das variáveis antecedentes e consequentes, bem como na configuração das funções de pertinência. Também, é Utilizada a formação (**se** $<$ *antecedente* $>$ **então** $<$ *consequente* $>$), obedecendo as condições quando totalmente satisfeitas ou parcialmente satisfeitas, conforme o mecanismo de inferência difusa, o que define o disparo da regra. Destaca-se também que as regras utilizadas foram cons-

truídas, de acordo com o modelo de inferência de Mamdani (MAMDANI; ASSILIAN, 1999).

Tabela 7 – Configuração das Variáveis - Antecedentes e Consequentes

(Antecedente) V1 - Período Molhamento Foliar:
Se período de Molhamento Foliar (umidade relativa) maior ou igual a 90%
Conjunto Universo (intervalo crisp): 0 a 100%
Conjunto Difuso: umidade baixa, umidade média, umidade favorabilidade
(Antecedente) V2 - Período Mínimo de Molhamento Foliar:
Se período de Molhamento Foliar maior que 6h - equivalente a 1/4 de 100%
Precipitação de 25% em diante (limite 70%)
Conjunto Universo (intervalo crisp): 0 a 100%
Conjunto Difuso: pouco tempo, tempo favorabilidade, muito tempo
(Antecedente) V3 - Dados do Cultivar - Imagem Segmentada da Folha de Soja
Conjunto Universo (intervalo crisp): 0 ou 1
Conjunto Difuso: com favorabilidade, sem favorabilidade
(Antecedente) V4 - Ponto Orvalho Temp. Média $< 2°C$ que o Ponto de Orvalho
Universo (intervalo crisp): -1 a $3°C$
Conjunto Difuso: temperatura baixa, temperatura favorabilidade, temperatura alta
(Antecedente) V5 - Faixa Temperatura (Temperatura Inicial e Temperatura Final)
Conjunto Universo (intervalo crisp): (Inicial: 0 a $27°C$ e Final: 13 a $44°C$)
Conjunto Difuso: temperatura baixa, temperatura favorabilidade, temperatura alta
(Antecedente) V6 - Temperatura Mínima
Conjunto Universo (intervalo crisp): 0 a $27°C$
Conjunto Difuso: temp. mínima baixa, temp. miníma favorabilidade, temp. miníma alta
(Antecedente) V7 - Temperatura Máxima
Conjunto Universo (intervalo crisp): 13 a $44°C$
Conjunto Difuso: temp. máxima baixa, temp. máxima favorabilidade, temp. máxima alta
(Consequente): Favorabilidade
Conjunto Universo (valores crisp): 0 a 100%
Conjunto Difuso: baixa, média, alta

Fonte: Próprio Autor

Semanticamente, as regras de inferências Min-Max aplicam os operadores de união e intersecção entre os conjuntos, conforme a Equação 66. Porém, o mapeamento completo de entrada e saída do modelo difuso está representado na Equação 67, dada a coleção de regras difusas condicionais Se-Então, que compõem a base de regras difusas.

$$\Psi(\alpha_1, \beta_1) = max(1 - \alpha_1, min(\alpha_1, \beta_1)) \tag{66}$$

onde $\Psi(\alpha_1, \beta_1)$ representa a função de pertinência difusa usada para calcular o grau de pertinência com base nos valores de α e β_1, que são as variáveis de entrada da função de pertinência; max função de máximo que retorna o maior valor entre os valores fornecidos como argumentos, usada para calcular o grau de pertinência difusa; min função de mínimo que retorna o menor valor entre os valores fornecidos como argumentos.

$$\mathfrak{R} = \left\{\delta^{(i)}\right\}_{i=1}^{n} = \left\{se \underset{n=1}{\overset{m}{and}} \left(X_n \text{ é } L_{\alpha}^{(i)}\right), então\ Y \text{ é } L_{\beta}^{(i)}\right\}_{i=1}^{n} \tag{67}$$

onde $\mathfrak{R}$ representa um conjunto de regras de inferência difusa, em que cada regra é identificada por $\delta^{(i)}$, na qual i é um índice que varia de 1 a n, indicando o número total de regras no conjunto; Cada $\delta^{(i)}$ representa uma regra de inferência individual no conjunto

de regras, composta por uma condição e uma conclusão; "se" palavra-chave que indica o início da condição da regra e "então" é a palavra-chave que indica o início da conclusão da regra. A conclusão é a parte da regra que especifica o que acontece, quando a condição é verdadeira; $\overset{m}{\underset{n=1}{and}}$ representa uma conjunção lógica (operador "E") entre m condições, o que significa que todas as m condições dentro da conjunção precisam ser verdadeiras para que a regra seja ativada; X_n variáveis de entrada da regra associada a uma função de pertinência $L_{\alpha}^{(i)}$, que descreve a pertinência da variável de entrada X_n ao conjunto $A\alpha_n$ no contexto da regra i; Y representa a variável de saída da regra que está sendo determinada, com base nas variáveis de entrada e nas funções de pertinência associadas; $L_{\beta}^{(i)}$ função de pertinência associada à variável de saída Y no contexto da regra i, que descreve como o grau de pertinência de Y varia em relação aos valores da variável de entrada X_n que satisfaçam a condição da regra.

Na Tabela 8 constam as descrições das inferências construídas, levando em consideração as favorabilidades baixa, média e alta e, adicionalmente, também é representada a quantidade de combinações de regras geradas para cada inferência.

As combinações surgem diante das variações, traduzidas pelo conhecimento da fenomenologia do problema da FAS, expressa pelas sete variáveis de V1 a V7 (antecedentes) que alimentam o modelo difuso e que, necessariamente, são compostas por conjunções "OU" e "E" formando regras únicas. Portanto, ao somar todas as combinações para as três possibilidades de favorabilidade, têm-se 120 regras construídas, as quais compõem a base de regras a serem submetidas à máquina de inferência difusa.

Tabela 8 – Inferências Difusas

Se	**Favorabilidade**	**Combinações**
Favorabilidade for TRUE para até duas variáveis ENTÃO 1 opção: V1 ou V2 ou V3 ou V4 ou V5 ou V6 ou V7	Baixa	1
Favorabilidade for TRUE para até duas variáveis ENTÃO 2 opções: V1 ou grupo(V2 ou V3 ou V4 ou V5 ou V6 ou V7)	Baixa	8
Favorabilidade for TRUE para até quatro variáveis ENTÃO 3 opções: V1 E V2 E grupo(V3 ou V4 ou V5 ou V6 ou V7)	Média	21
Favorabilidade for TRUE para até quatro variáveis ENTÃO 4 opções: V1 E V2 E V3 E grupo(V4 ou V5 ou V6 ou V7)	Média	35
Favorabilidade for TRUE para acima de quatro variáveis ENTÃO 5 opções: V1 E V2 E V3 E V4 E grupo(V5 ou V6 ou V7)	Alta	35
Favorabilidade for TRUE para acima de quatro variáveis ENTÃO 6 opções: V1 E V2 E V3 E V4 E V5 E grupo(V6 ou V7)	Alta	20

Fonte: Próprio Autor

Ainda na abordagem de Mamdani e Assilian (1999), a interpretação conjuntiva das regras difusas condicionais fazem uso das funções *s-norm* máxima ($\vee$), conforme a Equação 68, e da função *t-norm* mínima ($\wedge$), de acordo com a Equação 69.

$$\alpha \star_S, \beta = max(\alpha, \beta) = \beta \vee y \tag{68}$$

onde α e β são variáveis difusas ou conjuntos difusos que estão sendo combinados, usando a função *s-norm* máxima, podendo ser um valor escalar ou um conjunto difuso; $\star_S$ é o operador *s-norm* que representa a operação de máximo (ou *OR* máximo), usado para combinar os conjuntos difusos α e β; $\alpha \vee \beta$ significa que a saída da função é o máximo entre os valores de pertinência de α e β para um dado elemento do universo do discurso.

$$\alpha \star_T, \beta = min(\alpha, \beta) = \beta \wedge y \tag{69}$$

onde α e β são variáveis difusas ou conjuntos difusos que estão sendo combinados usando a função *t-norm* mínima, podendo ser um valor escalar ou um conjunto difuso; $\star_T$ é o operador *t-norm* que representa a operação de mínimo (ou *AND* mínimo), usado para combinar os conjuntos difusos α e β; $x \wedge y$ significa que a saída da função é o mínimo valor de pertinência entre os dois conjuntos para um dado elemento do universo do discurso.

Em contrapartida, o processo de defuzzificação consiste em calcular uma saída numérica representativa, em que $\beta_0 \in \mathbb{Y}$, a partir do conjunto difuso resultante $B'(\beta)$ em $\mathbb{Y}$. Portanto, consiste no mapeamento de conjuntos difusos do espaço $\mathbb{Y}$ para um único valor numérico de $\mathbb{Y}$, onde $\mathfrak{F}(\mathbb{Y}) \rightarrow \mathbb{Y}$. Assim, o resultado numérico é calculado usando o método *Center of Gravity* (COG), utilizando as Equações 70 e 71 (PROKOPOWICZ et al., 2017).

$$\beta_0 = \frac{\int_{\mathbb{Y}} \beta \mu_{B'}(\beta) d\beta}{\int_{\mathbb{Y}} \mu_{B'}(\beta) d\beta} \tag{70}$$

$$\mu_{B'}(\beta) = \bigvee_{i=1}^{m} \left[F^{(i)}(\chi_0) \wedge \mu_{B^{(i)}}(\beta) \right] \tag{71}$$

onde $\mu_{B'}(\beta)$ representa a pertinência de β a um conjunto fuzzy B'; β variável de saída para a qual calcula-se a pertinência no conjunto difuso B'; $\bigvee_{i=1}^{m}$ representa a operação de "supremo" ou operação de máximo para calcular a suprema pertinência entre os m conjuntos difusos resultantes da inferência difusa; i índice usado para iterar de 1 a m através dos conjuntos difusos que participam da inferência; $F^{(i)}(x_0)$ função de pertinência do conjunto difuso $\alpha^{(i)}$ em relação à variável de entrada χ_0, que representa a pertinência de χ_0 ao conjunto difuso $\alpha^{(i)}$; $\mu_{B^{(i)}}(\beta)$ função de pertinência do conjunto difuso $B^{(i)}$ em relação à variável de saída β, que representa a pertinência de β ao conjunto difuso $B^{(i)}$.

Após o cálculo efetuado pelo processo de defuzzificação, o valor numérico resultante é computado a uma taxa de erro de 5%, deste valor, para que a favorabilidade possa ser conhecida. O valor do consequente "favorabilidade" tem valor de 0 a 100%, o que mantém o padrão utilizado na abordagem de figura de mérito. Então, o resultado da favorabilidade, dado o valor numérico de defuzzificação, é para favorabilidade baixa de 0 a 33,3%, para favorabilidade média de 33,4 a 66,6%, e para favorabilidade alta de 66,7 até 100%. O Algoritmo 11, abaixo apresentado, utiliza métodos da biblioteca *Scikit-Fuzzy* (WARNER et al., 2019).

Algoritmo 11 Pseudocódigo Abordagem Lógica Difusa

Entrada: Regras de 1 a 7
Saída: Favorabilidade

```
início
    regras[] ← dados Regras de 1 a 7
    Definir_universo(Regras[]):
    para cada regra in regras[] faça
        define_antecedentes ← regras[regra];
        define_conjunto_difuso ← regras[regra];
    fim
    retorna antecedentes[], consequentes[];
    Definir_fuzzificação(antecedentes[], consequentes[]):
    para cada antecedente in antecedentes[] faça
        definir_grau_pertinência ← antecedentes[antecedente];
    fim
    para cada consequente in consequentes[] faça
        definir_grau_pertinência ← consequentes[consequente];
    fim
    retorna grau_pertinências[];
    Definir_base_regras_difusas():
    se período de molhamento foliar ≥ 90% então
        favorabilidade ← TRUE;
    fim
    senão se (período mínimo molhamento foliar ≥ 6h) E (precipitação ≤ 25%) então
        favorabilidade ← TRUE;
    fim
    senão se (faixa temperatura ≥ 18 ≤ 25) então
        favorabilidade ← TRUE;
    fim
    senão se (Faixa Temperatura ≥ 18 ≤ 26, 5) então
        favorabilidade ← TRUE;
    fim
    senão se (temperatura média − ponto orvalho) < 2 então
        favorabilidade ← TRUE;
    fim
    senão se (temperatura máxima ≥ 18 ≤ 26) então
        favorabilidade ← TRUE;
    fim
    senão se (temperatura mínima ≥ 18 ≤ 26) então
        favorabilidade ← TRUE;
    fim
    senão se (classificação imagem ← 1) então
        favorabilidade ← TRUE;
    fim
    retorna base_regras[];
    Avaliar regras(base_regras[]):
    para regra in base_regras[] faça
        calcular_grau_ativação ← base_regras[regra];
    fim
    retorna regras_ativadas[];
    Combinar regras(regras_ativadas[]):
    para regra in regras_ativadas[] faça
        combinar_regras_função_max_min ← regras_ativadas[regra];
    fim
    retorna regras_função_max_min[];
    Defuzzificar resultado(regras_função_max_min[]):
    para defuzzifica in combinar_função_max_min[] faça
        resultado_defuzzificado[] ← combinar_função_max_min[defuzzifica];
    fim
    retorna resultado_defuzzificado[];
    Calcular_favorabilidade(resultado_defuzzificado[]):
    se resultado_defuzzificado[] ≤ 33, 3% então
        favorabilidade ← baixa;
    fim
    senão se (resultado_defuzzificado[] ≥ 33, 4%) E (≤ 66, 6%) então
        favorabilidade ← média;
    fim
    senão se (resultado_defuzzificado[] ≥ 66, 6%) E (≤ 100%) então
        favorabilidade ← alta;
    fim
fim
retorna favorabilidade;
```

Outra abordagem considerada e avaliada no âmbito deste trabalho, é fundamentada em conceitos do método baseado em Cadeias Ocultas de Markov (BAUM; PETRIE, 1966; RABINER, 1989; KARBOWSKI et al., 2019).

Segundo Baum e Petrie (1966), a cadeia oculta de markov é um processo estocástico duplo que possui um componente observável e outro componente não observável. As cadeias ocultas de Markov são uma extensão do conceito das cadeias de Markov (BAUM; EAGON, 1967; MARKOV, 1971), definidas como modelo estocástico $\{X_n, n \in \mathbb{N}\}$, que descreve uma sequência de eventos, em que a probabilidade de um evento futuro depende apenas do estado atual e não dos estados anteriores. Esta propriedade Markoviana é expressa formalmente na Equação 72.

$$Pr(\zeta_n = \xi_n | \zeta_{n-1} = \xi_{n-1}, \ldots, \zeta_0 = i_0) = Pr(\zeta_n = \xi_n | \zeta_{n-1} = \xi_{n-1}) \tag{72}$$

onde $P = p_{(ij)}$ é a matriz de transição que controla a cadeia de Markov, se ζ_n denota o estado da cadeia de Markov, na época n, então $p_{i,j} = Pr(\zeta_n = j | \zeta_{n-1} = \xi)$ que, por consequência, toda entrada de P satisfaz $p_{ij} \geq 0$ e cada linha de P satisfaz $\sum_j p_{ij} = 1$.

As cadeias de Markov possuem os estados, que podem ser discretos ou contínuos, e que representam as condições ou configurações possíveis em um sistema que está sendo modelado. Cada estado na cadeia de Markov em tempo discreto é uma representação discreta de uma situação em que o sistema se encontra em um determinado momento. Quando ocorrem as mudanças de um estado para outro, seguindo o modelo de probabilidades, tais mudanças são chamadas de transições. As probabilidades de transição descrevem as possibilidades ou as chances de um sistema fazer a transição de um estado para outro, em um determinado período de tempo ou etapa, o que depende de cada contexto de aplicação. A Figura 25 ilustra a representação dos estados discretos, as transições entre os estados e as probabilidades, que descrevem as chances das transições (RABINER, 1989), bem como a matriz de transição. Consequentemente, uma matriz de transição em que cada elemento na posição i, j da matriz representa a probabilidade de transição do estado i para o estado j. As probabilidades de transição devem satisfazer às condições de não serem negativas e também a soma destas probabilidades de transição ser igual a 1 (LANGE; CHAMBERS; EDDY, 2010).

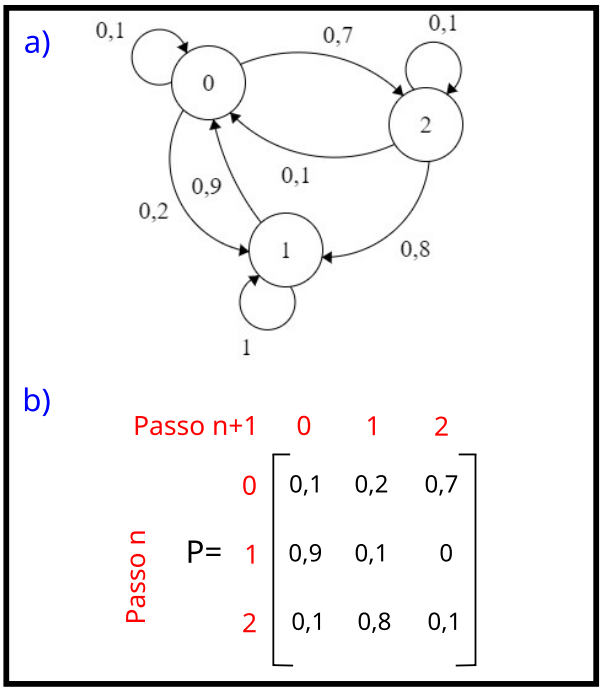
A probabilidade de transição de n passos $p_{ij}^{(n)} = \Pr(Z_n = j \,|\, Z_0 = i)$ é dada pelo elemento na linha i e coluna j da matriz P^n, expressa pela decomposição na Equação 73 que, ao longo de todos os caminhos $i \to i_1 \to \cdots \to i_{n-1} \to j$, ocorre via multiplicação de matrizes (LANGE; CHAMBERS; EDDY, 2010).

$$P_{ij}^{(n)} = \sum_{i_1}, \ldots, \sum_{i_{n-1}} P i i_1 \ldots P i_{n-1} j \tag{73}$$

As cadeias ocultas de Markov, de acordo com Lange, Chambers e Eddy (2010), incorporam tanto dados observáveis, quanto dados não observáveis ou ausentes. Entendem-se

por dados ausentes a sequência de estados visitados por uma cadeia de Markov e os dados observáveis, responsáveis por fornecerem informações parciais sobre essa sequência de estados, o que denota a sequência de estados visitados como $Z_1, \ldots, Z_n$, e a observação realizada na época i quando a cadeia está no estado Z_i por $Y_i = y_i$.

Figura 25 – Estados, Transições e Matriz - Cadeia de Markov



Fonte: Adaptada de (RABINER, 1989)

Conforme Ching e Ng (2006), as cadeias ocultas de Markov são caracterizadas pelos elementos: (1) N : O número de estados ocultos no modelo, denotados pelos estados individuais conforme a Equação 74:

$$SO = \{so_1, so_2, \ldots, so_N\} \tag{74}$$

onde SO representa o conjunto de estados individuais em uma cadeia de Markov, que são os possíveis estados que a cadeia de Markov pode assumir; $s_1, s_2, \ldots, s_N$ são as variáveis de estado individuais que compõem o conjunto S, em que cada s_i representa um estado específico na cadeia de Markov. O subscrito i varia de 1 a N, em que N é o número total de estados na cadeia de Markov.

(2) M : O número de símbolos de observações distintas por estado oculto, denotado pela Equação 75, bem como as observações no instante t como O_t.

$$VO = \{vo_1, vo_2, \ldots, vo_M\} \tag{75}$$

onde VO é o conjunto de todas as observações distintas em uma cadeia de Markov, representado como um conjunto que contém os elementos $vo_1, vo_2, \ldots, vo_M$; $vo_1, vo_2, \ldots, vo_M$

são as observações distintas que podem ocorrer em um estado oculto da cadeia de Markov, onde vo_1 é a primeira observação, vo_2 é a segunda observação e assim por diante, até vo_M, que é a última observação.

(3) A distribuição de probabilidade de transição de estados $[DP]_{ij} = dp_{ij}$ é denotada pela Equação 76:

$$dp_{ij} = DP(H_{t+1} = so_j | H_t = so_i), \quad 1 \leq i, j \leq N; \tag{76}$$

onde dp_{ij} denota a probabilidade de transição de um estado so_i para um estado so_j em uma cadeia de Markov. Trata-se de uma matriz de probabilidades de transição, em que i e j são índices que variam de 1 a N, representando os estados possíveis na cadeia de Markov; $DP(H_{t+1} = so_j | H_t = so_i)$ é a notação condicional que indica a probabilidade de que o próximo estado H_{t+1} seja igual a so_j, dado que o estado atual H_t é igual a so_i. Em outras palavras, é a probabilidade de fazer a transição do estado so_i para o estado so_j em um passo de tempo na cadeia de Markov.

(4) A distribuição de probabilidade dos símbolos de observação no estado oculto j, $[DP]_{jk} = \{dp_j(vo_k)\}$ é denotada pela Equação 77:

$$dp_j(vo_k) = DP(O_t = vo_k | H_t = so_j), \quad 1 \leq j \leq N, 1 \leq k \leq M \tag{77}$$

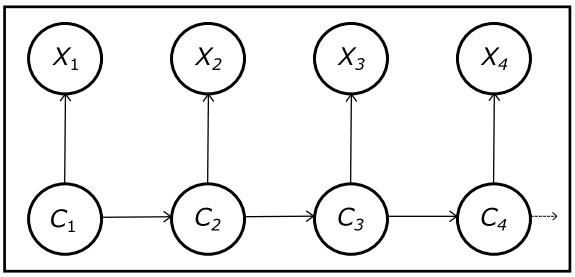
onde $dp_j(vo_k)$ denota a probabilidade de observar o símbolo vo_k, quando a cadeia de Markov está no estado oculto so_j, ou seja, $dp_j(vo_k)$ é a probabilidade condicional de observar o símbolo vo_k no momento t, dado que a cadeia de Markov está no estado oculto so_j no mesmo momento t; $DP(H_t = vo_k | H_t = so_j)$ é a notação condicional que indica a probabilidade de que a observação H_t seja igual a vo_k, dado que o estado oculto H_t seja igual a so_j no momento t; $1 \leq j \leq N, 1 \leq k \leq M$ são as condições que limitam os valores possíveis de j e k, de maneira que j varia de 1 a N, representando os estados ocultos possíveis na cadeia de Markov. k varia de 1 a M, representando os símbolos de observação possíveis. Tais condições garantem que estão sendo consideradas todas as combinações possíveis de estados ocultos e observações.

(5) A distribuição inicial de estados $\prod = \{\pi_i\}$ é denotada pela Equação 78:

$$\pi_i = DP(H_1 = so_i), \quad 1 \leq i \leq N \tag{78}$$

A Figura 26 ilustra o grafo de um modelo baseado no uso de cadeias ocultas de Markov, assim como o Algoritmo 12, desenvolvido para a abordagem de Cadeias Ocultas de Markov.

Figura 26 – Grafo de um Modelo Oculto de Markov



Fonte: Adaptada de (ZUCCHINI; MACDONALD; LANGROCK, 2009)

Algoritmo 12 Pseudocódigo Abordagem Cadeias Ocultas de Markov

Entrada: Vetor de Ocorrências, Cadeia Oculta Markov
Saída: Resultados Favorabilidade, Combinação Markov Selecionada e Probabilidades

```
1  início
2      vo[] ← Vetor de Ocorrências;
3      c[] ← Cadeia Oculta Markov;                    ▷ Vetor Cadeia Oculta Markov
4      vn[] ← 0 ;                                     ▷ Vetor Ocorrências Normalizado
5      result[] ← 0;                                  ▷ Vetor Resultado
6      para (ele in vo) faça
7          se (vo[ele] >0) então
8              vn[ele] ← 1;
9          senão
10             vn[ele] ← 0;
11         fim
12     fim
13     para cada ele in c faça
14         se c[ele] == vn[0] então
15             result[0] ← c[ele];                    ▷ Buscar Combinação Cadeia Oculta de Markov
16         fim
17     fim
18     para cada ele in Result faça
19         contab ← 0;
20         se ele ==1 então
21             contab ← + = 1;
22         fim
23     fim
24     porcentagem ← 1/contab;
25     se porcentagem ≥ 0 ≤ 0,333 então
26         estado ← baixa;                            ▷ Favorabilidade
27     fim
28     senão se porcentagem ≥ 0,334 ≤ 0,666 então
29         estado ← média;                            ▷ Favorabilidade
30     fim
31     senão se porcentagem ≥ 0,667 ≤ 1 então
32         estado ← alta;                             ▷ Favorabilidade
33     fim
34 fim
35 retorna estado, result;          ▷ Combinação Markov Selecionada e Probabilidades
```

Neste trabalho são avaliadas a eficácia e a eficiência do uso dos três métodos descritos acima, de forma a selecionar àquele que venha a apresentar o melhor resultado para compor o sistema de visão e inteligência computacional em ambiente de nuvem para gestão de risco da ferrugem asiática na cultura da soja.

As entregas do sistema estão organizadas de forma a refletirem a construção de recomendações para tomada de decisão, a partir de informações originadas no processamento do modelo da fusão e apresentadas em relatórios gerenciais, disponíveis em uma interface de usuário, em formato *Dashboard*.

Para os relatórios gerenciais, são considerados assuntos e questões previstas no levantamento de requisitos, supracitado na Tabela 2, cuja estrutura está materializada no modelo dimensional e implementado no ambiente da Oracle *Cloud*.

Adicionalmente, os relatórios gerenciais têm, por objetivo, dar apoio às tomadas de decisão, tanto ao usuário, quanto às áreas de plantio, em conformidade com os assuntos: (1) Existência da ferrugem asiática nas áreas de cultura (influenciado pelas das variáveis climáticas e características identificadas em imagens digitais, variedade da planta na área de cultura); (2) estágio de severidade da doença (contabilização da favorabilidade baixa, média e alta no ciclo de cultura); e (3) recomendações agronômicas baseadas no diagnóstico para o controle da doença.

O ambiente Oracle *Cloud* é configurado para atender à elaboração destes Relatórios, levando em consideração a estruturação da Base de Dados históricos, ou seja, o *Data Warehouse*. Este é elaborado para que as junções criadas entre as tabelas de Dimensão e Fatos possam responder às consultas previstas nos requisitos planejados, com exceção do terceiro reaquisto (3), o qual é somente considerado em operação via *dashboard*.

A Figura 27 apresenta a configuração das junções e das Tabelas de Dimensão e Fatos frente às junções definidas, disponíveis na *View* Analítica do Modelo Dimensional do ambiente.

A partir de operações *OLAP*, é possível trabalhar os dados para serem visualizados no relatório, de acordo com o nível de detalhes e informações pertinentes. As ferramentas consideradas para a operacionalização dos dados são: (1) *drill across*: permite acessar dados em uma dimensão diferente, muitas vezes de outras tabelas ou fontes; (2) *drill down*: utilizada quando há necessidade de aumentar o nível de detalhe, ou seja, ela aumenta a granularidade; e (3) *drill up*: utilizada quando há necessidade do nível de detalhe ser mais amplo ou ainda resumido, ou seja, há diminuição da granularidade.

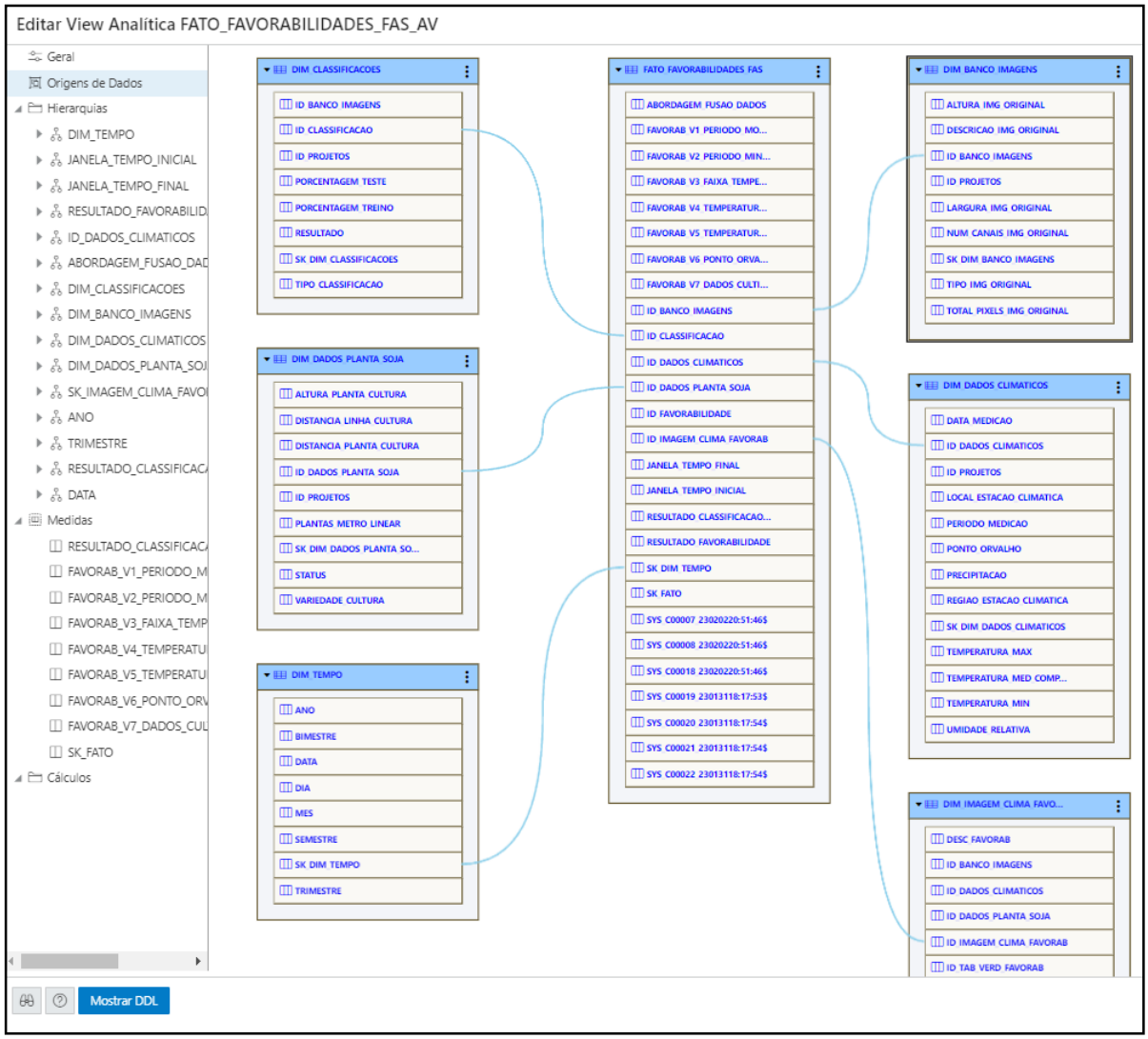
No contexto do bloco de resultados, está previsto que o usuário pode usufruir das informações produzidas pelo sistema, haja vista que as informações dos relatórios gerenciais orientam as decisões frente ao cenário da FAS.

Outro aspecto para a interação dos usuários é o uso de uma interface que possibilita, por meio da web, processar os dados no sistema de acordo com os séries históricas do banco de dados. Os dados armazenados são originados das séries temporais climáticas e também

dos processamentos das diferentes etapas do sistema, ou seja, das séries de imagens que envolvem a segmentação, o reconhecimento de padrões e o aprendizado de máquina. O conjunto desses dados processados, de acordo com o vetor de decisão, passam em seguida pela fusão, gerando elementos para o suporte à decisão e dados para os relatórios gerenciais analíticos.

O bloco resultados consolida os processos apresentados nos outros blocos do sistema, o que inclui as interfaces com os usuários, bem como a apreciação dos resultados históricos expressos pelos relatórios do *Data Warehouse*. Este compilado de ações visa prover recursos integrados para tomada de decisão, a partir do processamento das imagens de laboratório das plantas de soja em verdade de campo, ou seja, considerando imagens reais.

Figura 27 – Configuração das Tabelas do *Data Warehouse*



Fonte: Próprio Autor

3.2.1 Métodos para avaliações da Qualidade de Dados

Outro aspecto metodológico considerado é quanto à avaliação da qualidade dos dados, nas diferentes fases dos processos que compõem o sistema de visão e inteligência computacional em ambiente de nuvem para gestão de risco da ferrugem asiática na cultura da soja.

Para tanto, neste trabalho o seguinte conjunto de descritores de qualidade é utilizado para a descrição do *framework* de funcionalidade de dados:

1. **Validade**: é uma característica decorrente da medição, indicando o quanto os instrumentos realmente medem em relação ao que se pretendia medir;
2. **Confiabilidade**: é a característica de medição preocupada com a consistência;
3. **Menor Janela Temporal de Coleta**: é a relação entre o momento da coleta dos dados, sua compilação, interpretação e relevância para os processos de tomada de decisão;
4. **Exatidão**: é a probabilidade de que os dados reflitam a verdade, ou seja, medida que representa fielmente a realidade. O termo está relacionado à precisão e à ausência de erros ou distorções nas informações coletadas e armazenadas;
5. **Integridade**: está relacionada com a veracidade dos dados, frente ao fenômeno observado.

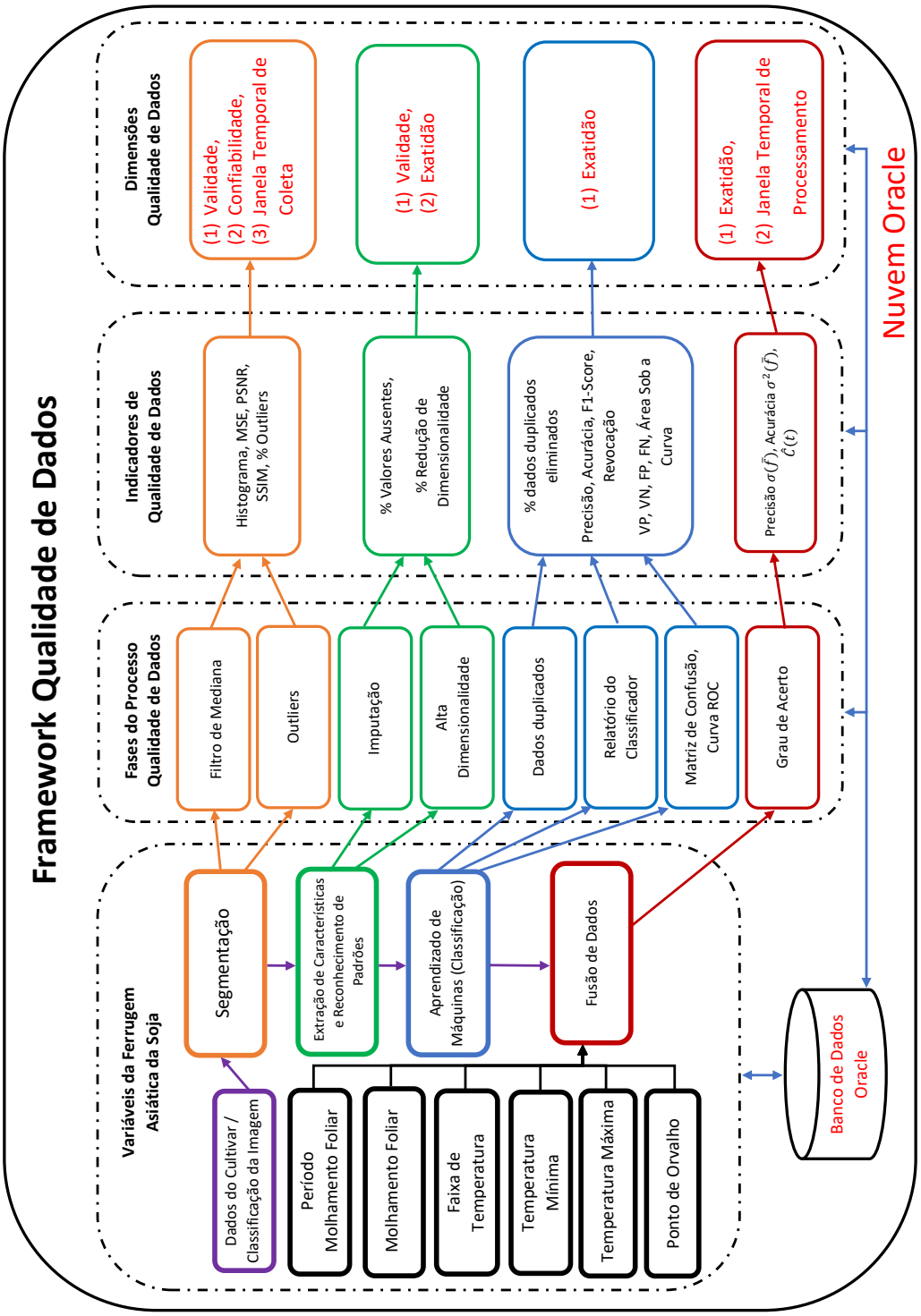
A Figura 28 ilustra o fluxo das operações e o conjunto de indicadores e as dimensões de qualidade que podem ser observados. Nota-se também que há uma ligação de todos os fluxos de qualidade com o banco de dados Oracle, responsável por suportar a infraestrutura dos dados que participam do processo.

A aplicação do *framework* de qualidade está inserida diretamente nos algoritmos de cada etapa do sistema, haja vista que, para facilitar a visualização dos indicadores, as informações são reunidas em apenas uma tela da interface web, respeitando cada etapa do processamento, diante das janelas das séries temporais de dados.

A fase do *framework*, que se refere à segmentação das imagens, avalia as dimensões: validade, confiabilidade e janela temporal de coleta e também os indicadores: histograma das imagens, as métricas MSE, PSNR, SSIM e *outliers*.

O Erro Quadrático Médio (MSE - *Mean Square Error*), a Relação Sinal-Ruído de Pico (PSNR - *Peak Signal-to-Noise Ratio*) e o Método de Índice de Similaridade de Estrutura (SSIM - *Structure Similarity Index Method*) são métricas frequentemente aplicadas para avaliar a qualidade, frente às diversas tarefas de processamento de imagens, estabelecendo comparações entre imagens, de modo a atender a diferentes sensibilidades e contextos de degradação, considerando ou não a percepção humana (HORé; ZIOU, 2010; PREEDANAN et al., 2018; SARA; AKTER; UDDIN, 2019).

Figura 28 – Framework Qualidade de Dados



Fonte: Próprio Autor

Ao saber que o Erro Quadrático Médio (MSE - *Mean Square Error*), conforme descrita na Equação 79, é uma métrica muito utilizada para a medição da qualidade de dados que, no contexto da avaliação da qualidade de imagens, quantifica a diferença entre a imagem original de entrada e uma imagem comprimida ou processada.

$$\text{MSE} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(IA[i,j] - IB[i,j])^2 \tag{79}$$

onde m e n são as dimensões de largura e altura das imagens; $IA[i,j]$ e $IB[i,j]$ são os valores dos *pixels* na posição i,j nas imagens IA e IB, respectivamente.

A Relação Sinal-Ruído de Pico (PSNR - *Peak Signal-to-Noise Ratio*), descrita na Equação 80, é uma métrica também utilizada para avaliar a qualidade de uma imagem, que compara o nível máximo do sinal com o nível do ruído submetido à imagem. Quanto maior for a relação sinal-ruído de pico, melhor será a qualidade da imagem ou do sinal. As vantagens do uso da métrica PSNR, em relação à MSE, são evidenciadas pelas melhorias de sensibilidade quanto à detecção de pequenas diferenças entre a imagem original e a imagem processada, o que facilita o reconhecimento de pequenas degradações em imagens, fatos estes que podem passar despercebidos com a métrica MSE. O uso do PSNR também permite a comparação da qualidade em diferentes escalas de intensidade de *pixels*.

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{MVP^2}{\text{MSE}}\right) \tag{80}$$

onde MVP representa o valor máximo que um *pixel* pode ter em uma imagem. Em imagens com 8 bits por canal, como é o caso das imagens coloridas RGB, o valor máximo do pixel é 255; MSE é o erro médio quadrático Médio encontrado entre a imagem de referência e a imagem processada.

Além das métricas MSE e PSNR, há também as de Índice de Similaridade de Estrutura (SSIM - *Structure Similarity Index Method*), conforme disposto na Equação 81, que viabiliza avaliar a qualidade das imagens, com destaque na percepção visual humana e seus aspectos estruturais. Fatores importantes, baseados na percepção, como mascaramento de luminância e mascaramento de contraste, são considerados para o cálculo SSIM. Adicionalmente, o cálculo da semelhança estrutural, entre as duas imagens, leva em consideração como os valores dos *pixels* estão distribuídos.

$$\text{SSIM}(IA, IB) = \frac{(2 \cdot \mu_{IA} \cdot \mu_{IB} + c_1) \cdot (2 \cdot \sigma_{IAIB} + c_2)}{(\mu_{IA}^2 + \mu_{IB}^2 + c_1) \cdot (\sigma_{IA}^2 + \sigma_{IB}^2 + c_2)} \tag{81}$$

onde IA e IB são as duas imagens de referência e a imagem processada, respectivamente; μ_{IA} e μ_{IB} são as médias dos valores dos *pixels* nas imagens IA e IB, respectivamente; σ_{IA} e σ_{IB} são os desvios padrão dos valores dos *pixels* nas imagens IA e IB, respectivamente; σ_{IAIB} é a covariância entre os valores dos *pixels* nas imagens IA e IB; c_1 e c_2 são constantes pequenas com o valor de 0,01 que são adicionadas para evitar a divisão por zero e estabilizar o cálculo, tal que $c_1 = (k_1 \cdot L)^2$ e $c_2 = (k_2 \cdot L)^2$, onde L é o intervalo dinâmico dos valores de pixel (por exemplo, 255 para imagens de 8 bits por canal) e k_1 e k_2 são constantes predefinidas.

Na etapa do sistema que processa a extração de características e reconhecimento de padrões, deve ser avaliada, no âmbito da qualidade dos dados, as dimensões: validade e exatidão e também os indicadores: valores ausentes e redução de dimensionalidade.

Quanto aos valores ausentes, os mesmos são avaliados no momento do processo em que ocorre a junção dos vetores de dados de características processadas com os algoritmos SIFT, HOG e Momentos de HU. Esta junção pode retornar diferentes tamanhos para o vetor resultante de características dos padrões para as diferentes imagens processadas. Este arranjo é feito de forma horizontal, sendo entretanto necessário completar com zeros, as posições das linhas que, eventualmente, não se encontram preenchidas.

Ainda, na etapa de extração de características e reconhecimento de padrões, o indicador de alta dimensionalidade é tratado com a redução da dimensionalidade do vetor de características, que apresenta tamanho máximo de 130 colunas, resultantes da união dos vetores de características, podendo entretanto ter sua dimensionalidade final reduzida com um número menor de colunas, o que é obtido com a aplicação da técnica PCA.

Na etapa de aprendizado de máquina, a avaliação da qualidade remete ao estudo das dimensões relacionadas à exatidão, bem como os indicadores relacionados às instâncias duplicadas.

A necessidade de avaliar os dados duplicados, de acordo com Chitralekha e Roogi (2021), ocorre por ocasionar a falsa impressão de que o dado duplicado é mais importante que os demais. Também, devido ao fato de provocar o aumento do tempo de indução do modelo de aprendizado de máquina, que é decorrente dos conjuntos de dados duplicados.

Quanto ao procedimento para o tratamento dos dados redundantes no modelo de favorabilidade, o mesmo consiste na identificação e depois na remoção, aplicado no conjunto avaliado para cada imagem de características repetidas. Tais eliminações colaboram com a simplificação do conjunto de dados, com o objetivo de melhorar o desempenho do classificador ou modelo de aprendizado de máquina.

Por outro lado, para as avaliações da qualidade do processamento dos dados relacionados na etapa de fusão, são consideradas as dimensões exatidão e janela temporal de processamento, bem como os indicadores de precisão e acurácia.

Os indicadores de qualidade, quando utilizados nas cadeias de Markov, são baseados na teoria de autocorrelação, de acordo com Berg (2004), onde se encontram estimados os valores esperados a partir das observações utilizadas.

Para tanto, a partir de uma série temporal de N medições para o processo de Markov (Equação 82), onde e_i representa as configurações geradas para a série temporal, i representa a ordem temporal entre as medições. Assim, o estimador[3] do valor esperado para a observação $\hat{j}$ é demonstrado na Equação 83, onde o símbolo $\bar{}$ (traço) representa a média amostral.

[3] O símbolo $\hat{}$ (chapéu) representa a estimativa.

$$\jmath_i = \jmath_i(e_i),\ \ i = 1, ..., N \tag{82}$$

$$\bar{\jmath} = \frac{1}{N}\sum \jmath_i \tag{83}$$

No entanto, a função de autocorrelação para uma observação $\hat{\jmath}$ está definida na Equação 84, considerando a invariância de translação no tempo para equilíbrio do conjunto de dados estabelecidos no processo. Adicionalmente, a Equação 85 demonstra que a variância de $\jmath$ é um caso especial de autocorrelação.

$$\hat{C}(t) = \hat{C}_{ij} = \langle(\jmath_i - \langle\jmath_i\rangle)(\jmath_j - \langle\jmath_j\rangle)\rangle = \langle\jmath_i\jmath_j\rangle - \langle\jmath_i\rangle\langle\jmath_j\rangle = \langle\jmath_0\jmath_t\rangle - \hat{\jmath}^2 \tag{84}$$

onde $\hat{C}(t)$ é o valor da autocorrelação para uma observação em um dado momento t; $\hat{C}_{ij}$ é o valor da autocorrelação entre duas variáveis $\jmath_i$ e $\jmath_j$; $\jmath_i$ e $\jmath_j$ são as observações das variáveis. Cada uma delas pode ser vista como uma série temporal de dados; $\langle\jmath_i\rangle$ e $\langle\jmath_j\rangle$ são as médias das séries temporais $\jmath_i$ e $\jmath_j$, respectivamente; $\langle(\jmath_i - \langle\jmath_i\rangle)(\jmath_j - \langle\jmath_j\rangle)\rangle$ é uma medida da covariância entre as séries temporais; $\langle\jmath_i\jmath_j\rangle$ é a média do produto das séries temporais $\jmath_i$ e $\jmath_j$, representando a covariância bruta entre as duas variáveis; $\hat{\jmath}^2$ é a média quadrada da variável $\jmath$, ou seja, $\langle\jmath\rangle^2$ que representa a variância de $\jmath$; $\langle\jmath_0\jmath_t\rangle$ é a média do produto das variáveis $\jmath_0$ e $\jmath_t$, onde $\jmath_0$ é uma observação em um determinado momento e $\jmath_t$ é uma observação em um momento posterior t na série temporal. Ela representa a covariância entre $\jmath_0$ e $\jmath_t$.

$$\hat{C}(0) = \sigma^2(\jmath) \tag{85}$$

onde $\hat{C}(0)$ é o valor da autocorrelação em $t = 0$ para uma observação, ou seja, a covariância da variável $\jmath$ consigo mesma no mesmo instante de tempo; $\sigma^2(\jmath)$ é o símbolo para a variância da variável $\jmath$. A variância mede a dispersão dos valores da variável $\jmath$ em relação à sua média, calculada como a média dos quadrados das diferenças entre cada valor de $\jmath$ e a média de $\jmath$.

Outro ponto a considerar na teoria apresentada por Berg (2004) é quanto à análise da autoconsistência versus o erro razoável, o que implica analisar os aspectos de equilíbrio do sistema, considerando as séries temporais sob o contexto da cadeia de Markov e o controle dos tempos de autocorrelação integrados, tomados por diferentes medições de $\bar{\jmath}$.

As Equações 86, 87 e 88 definem o cálculo do erro $\Delta\bar{\jmath}$, a variância do estimador $\bar{\jmath}$ e o tempo de correlação integrada τ_{int}, respectivamente.

$$\Delta\bar{\jmath} = \sqrt{\sigma^2(\bar{\jmath})}\ com\ \sigma^2(\bar{\jmath}) = \tau_{int}\frac{\sigma^2(\bar{\jmath})}{N} \tag{86}$$

onde $\Delta\bar{\jmath}$ é o erro padrão que mede a incerteza associada à estimativa da média amostral $\bar{\jmath}$; $\sigma^2(\bar{\jmath})$ é a variância que indica o quão espalhados estão os valores das médias amostrais em relação à verdadeira média populacional; τ_{int} representa o tempo de integração ou o

tempo de correlação integrada. Trata-se de um parâmetro que descreve a autocorrelação dos dados; $\frac{\sigma^2(\jmath)}{N}$ é a estimativa da variância da média $\bar{\jmath}$ com base no tamanho da amostra N.

$$\sigma^2(\bar{\jmath}) = \sigma^2\frac{(\jmath)}{N}\left[1 + 2\sum_{t=1}^{N-1}\left(1 - \frac{t}{N}\right)\hat{\epsilon}(t)\right] com\ \hat{\epsilon}(t) = \frac{\hat{C}(t)}{\hat{C}(0)} \tag{87}$$

onde $\hat{\epsilon}(t)$ é a função de autocorrelação normalizada em $t = 0$, ou seja, $\hat{C}(0) = 1$ que mede a autocorrelação da variável $\jmath$ em diferentes momentos de tempo t, normalizada em relação à autocorrelação.

$$\tau_{int} = \left[1 + 2\sum_{t=1}^{N-1}\left(1 - \frac{t}{N}\right)\hat{\epsilon}(t)\right] \tag{88}$$

onde τ_{int} é o tempo de integração da correlação.

3.2.2 Considerações Finais

Neste Capítulo, além de materiais, a partir da revisão bibliográfica da literatura em estado da arte, foram considerados os principais métodos selecionados para serem, após avaliados, utilizados no sistema de visão e inteligência computacional em ambiente de nuvem para gestão de risco da ferrugem asiática na cultura da soja.

O próximo Capítulo trata sobre resultados e discussões, incluindo a validação do sistema, frente à visão de especialistas.

Capítulo 4

Resultados e Discussões

Neste Capítulo são apresentados os resultados, discussões e validação do sistema baseado em visão computacional para avaliação da favorabilidade da ocorrência da ferrugem asiática em soja. Neste contexto, os tópicos se encontram organizados na forma: (1) configuração e integração de tecnologias da arquitetura em nuvem e aplicação de técnicas para estruturação dos dados; (2) pré-processamento e processamento digital das imagens, envolvendo o uso de técnicas para extração de características; (3) redução de dimensionalidade do vetor de características; (4) reconhecimento de padrões e classificação de características com aprendizado de máquina; (5) organização do vetor de dados estruturados para o modelo de decisão; (6) fusão de dados estruturados; e (7) relatórios analíticos, recomendações e *dashboard*. O processamento e a visualização dos resultados ocorrem via interfaces web.

4.1 Configuração, Integração e Estruturação dos Dados

A Plataforma Oracle *Cloud* foi configurada de acordo com a melhor opção de arquitetura, diante do estudo realizado com três possíveis cenários que atenderam à pesquisa. Os principais elementos observados, que compuseram as arquiteturas estudadas, foram: o acesso às redes privadas e públicas, interligação dos elementos de armazenamento de objetos, infraestrutura para instância de computação, ambiente de *Data Science* ou infraestrutura para ciência de dados, serviços para análises de dados analíticos, bancos de dados tanto transacionais, quanto os multidimensionais.

Nos cenários avaliados, foram observados pontos em comum, tais como a origem dos

dados de fontes públicas e privadas. Inicialmente, o armazenamento foi feito na estrutura *Object Storage*, caracterizada como um repositório temporário *Data Lake*, para que esses objetos pudessem ser separados em dados de imagens e dados estruturados, inclusive os das séries climáticas, de forma supervisionada, utilizando-se *buckets* distintos. Os três cenários avaliados estudaram diferentes possibilidades com foco no armazenamento dos dados estruturados, não estruturados, e nas possibilidades de integração entre os elementos da arquitetura.

No primeiro cenário, foi representado o fluxo dos dados estruturados, cujo armazenamento contemplou somente o uso do SGBD MySQL. Outras duas opções para armazenamento dos dados estruturados foram propostas nos demais cenários 2 e 3.

No cenário 2, foi considerado o foco no armazenamento caracterizado pelo uso do *Autonomous Data Warehouse* (ADW) e concebido a partir do banco de dados Mysql, ao invés do uso da opção com o Oracle 19c. O uso do Mysql, no cenário 2, foi pensado para estruturar o *Data Warehouse*, sem o uso de licenças de softwares com alto custo do Oracle 19c. Por não encontrar essa escolha, na *Oracle Cloud Infrastructure* (OCI), foi avaliada a viabilidade dessa implementação junto à equipe de suporte da Oracle.

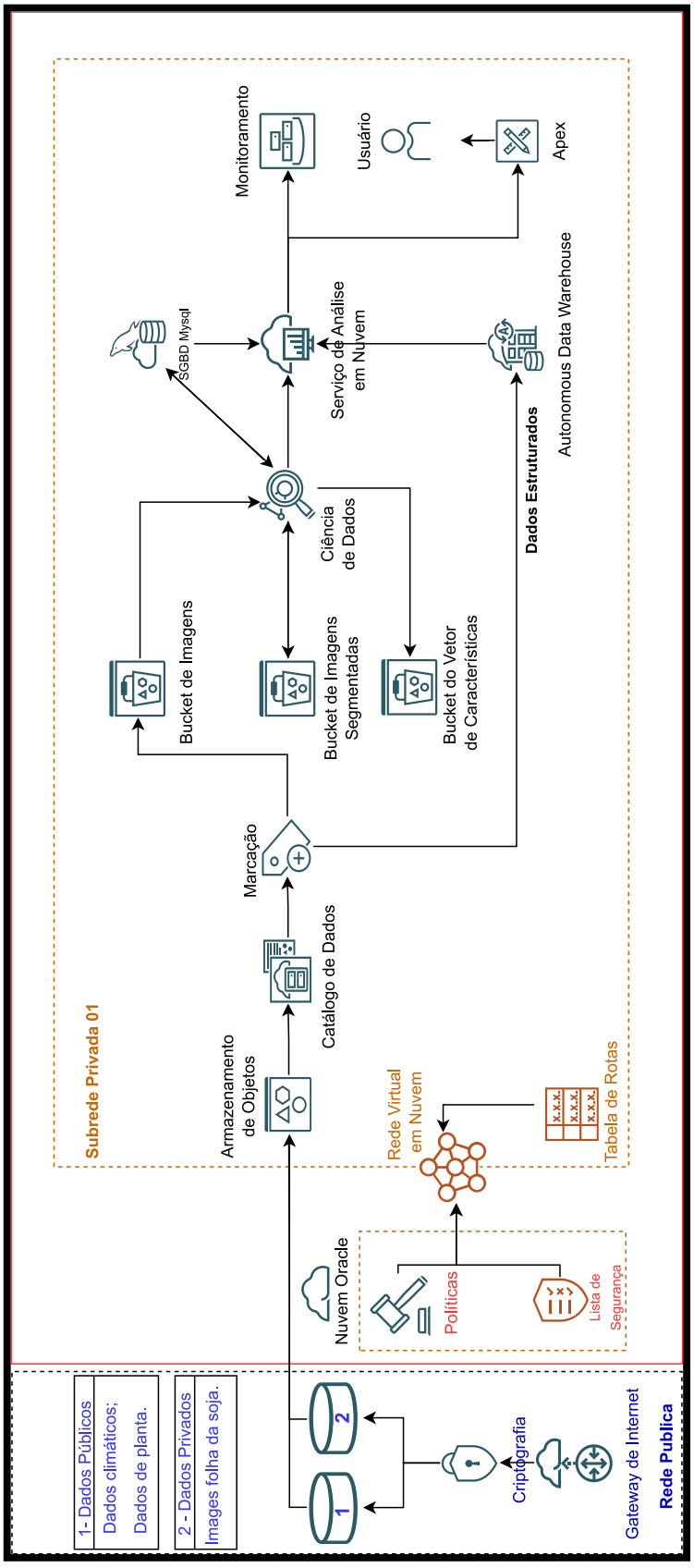
Como terceiro cenário, o armazenamento dos dados fez uso do elemento ADW, conforme o padrão estabelecido na OCI, ou seja, criado com o uso da opção do Oracle 19c, sendo assim viabilizada sua implementação. Nesse caso, foi considerada também a adição da licença de software do Oracle 19c, configurada no ato da criação da instância do ADW.

Para esses três cenários considerados foram utilizados algoritmos de aprendizado de máquina para processar os dados relacionados as etapas envolvidas na visão computacional. As Figuras 29, 30 e 31 ilustraram, respectivamente, as arquiteturas desenvolvidas para cada cenário.

Após avaliação, o cenário 3 foi escolhido devido à sua configuração, possibilitar a completa integração dos serviços envolvidos na infraestrutura. Tal integração permitiu que os dados pudessem ser armazenados, inicialmente, nos *buckets* quando não estruturados ou semiestruturados ou o fluxo destes dirigidos, tanto ao ambiente de desenvolvimento e processamento (Ciência de Dados), quanto para armazenamento no *Autonomous Database* para alimentar o sistema de apoio à decisão. Adicionalmente, a construção do *Data Warehouse* e sua integração com o serviço de *Analytics Cloud Service* foi somente compatível quando utilizado o *Autonomous Database*.

Além disso, o cenário 3 se mostrou robusto também por contemplar a infraestrutura de hospedagem web para monitoramento do usuário, por meio da instância de computação linux. Nessa implementação, as etapas de catalogação e marcação dos objetos não foram consideradas, por não fazerem parte do escopo do desenvolvimento do sistema de visão e inteligência computacional em ambiente de nuvem para gestão de risco da ferrugem asiática na cultura da soja.

Figura 29 – Arquitetura Oracle *Cloud* - Cenário 1



Fonte: Próprio Autor

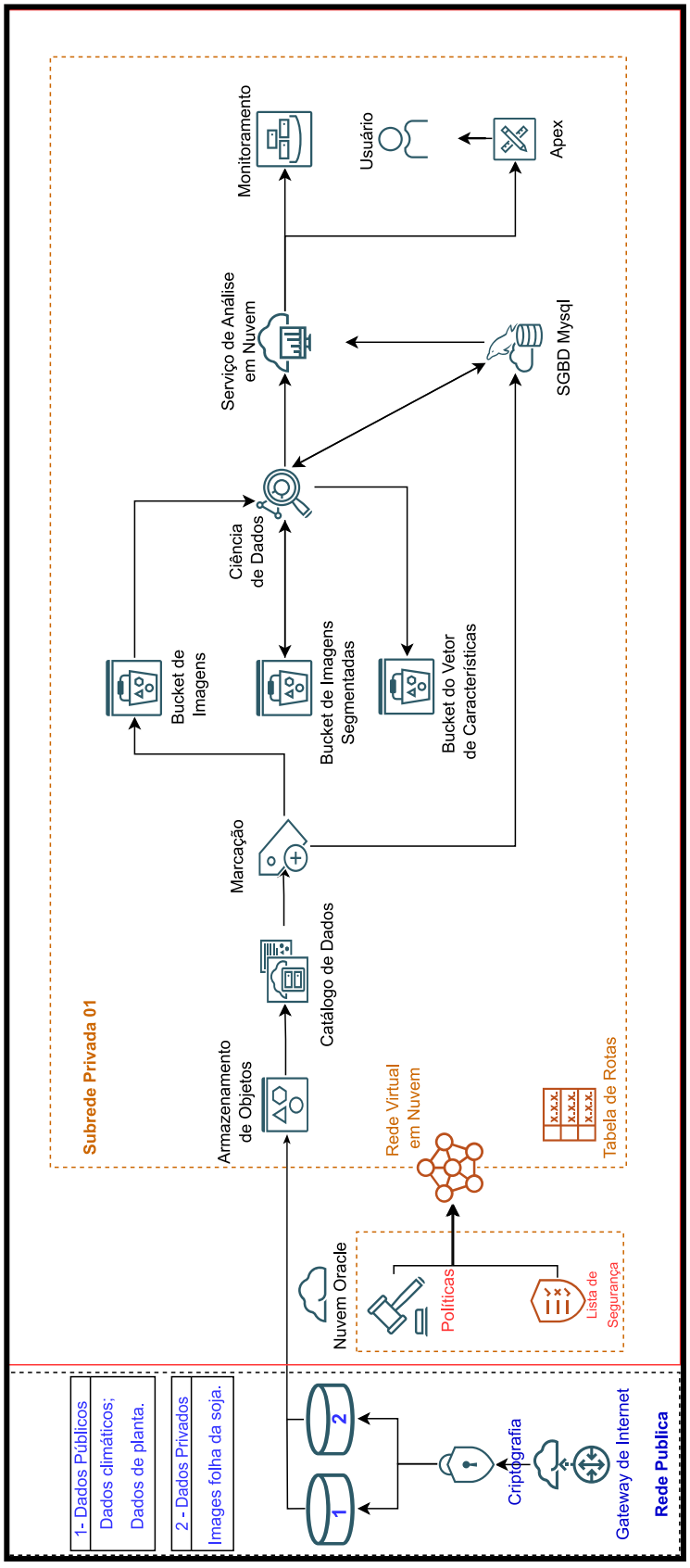


Figura 30 – Arquitetura Oracle *Cloud* - Cenário 2

Fonte: Próprio Autor

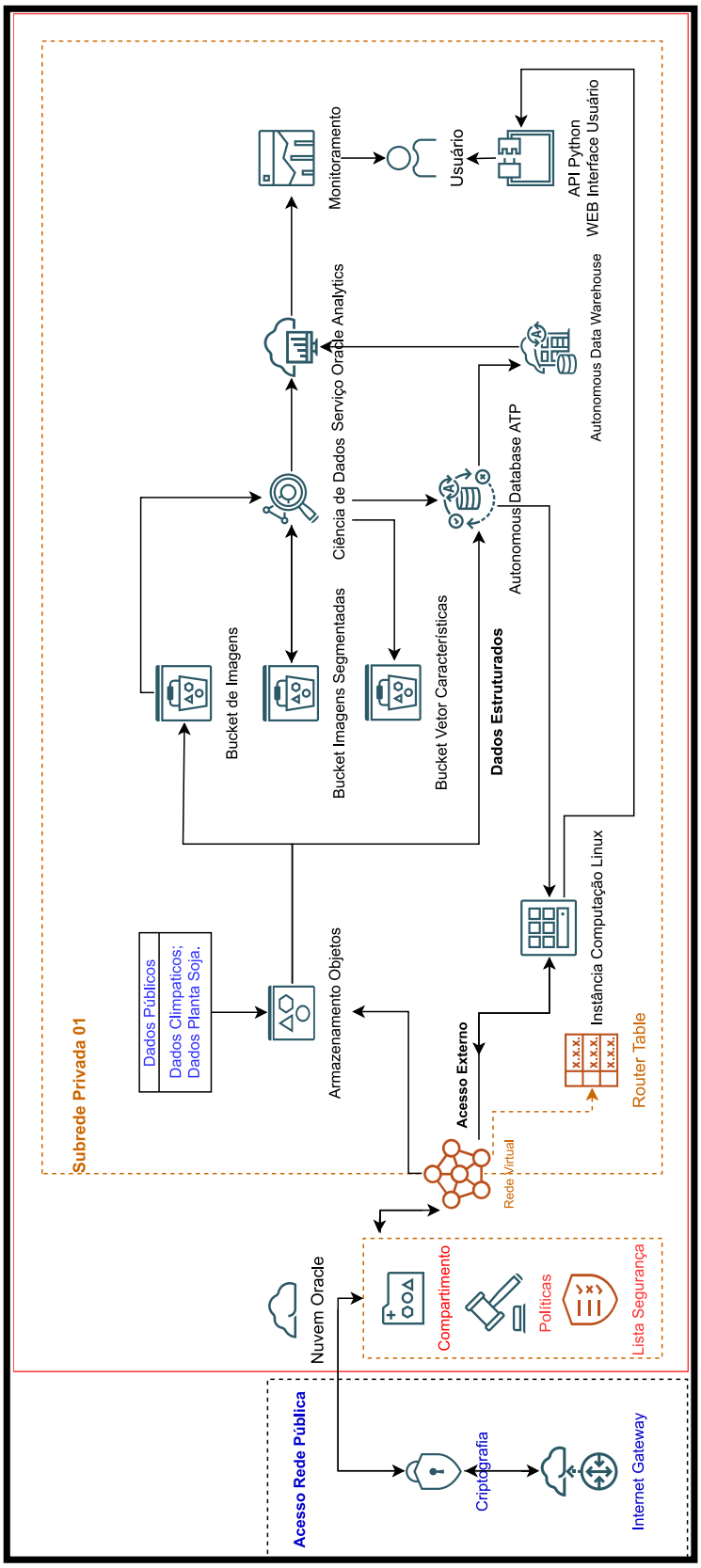


Figura 31 – Arquitetura Oracle *Cloud* - Cenário 3

Fonte: Próprio Autor

Os cenários 1 (Figura 29) e 2 (Figura 30) não atenderam aos critérios de integração das tecnologias envolvidas, vez que foram desenhados visando o menor custo, logo com menor disponibilidade de tecnologias agregadas. As incompatibilidades decorrentes desse fato, ocorreram tanto pela opção de uso do SGBD MySQL, que não se comunica com o *Data Warehouse* do *Autonomous Database* e também com o *Analytics*, quanto pelo uso da tecnologia APEX (*Application Express*) para a implementação web, que não permite aproveitar o código desenvolvido na etapa de ciência de dados.

A Figura 32 ilustra resultados para a tela inicial do ambiente da Plataforma Oracle, após a autenticação. O menu, uma vez aberto, exibiu conforme as marcações em cor vermelha as opções dos recursos utilizados para o desenvolvimento do sistema, sem considerar o detalhamento das configurações de rede, usuários e permissões de acesso, que também foram efetuadas.

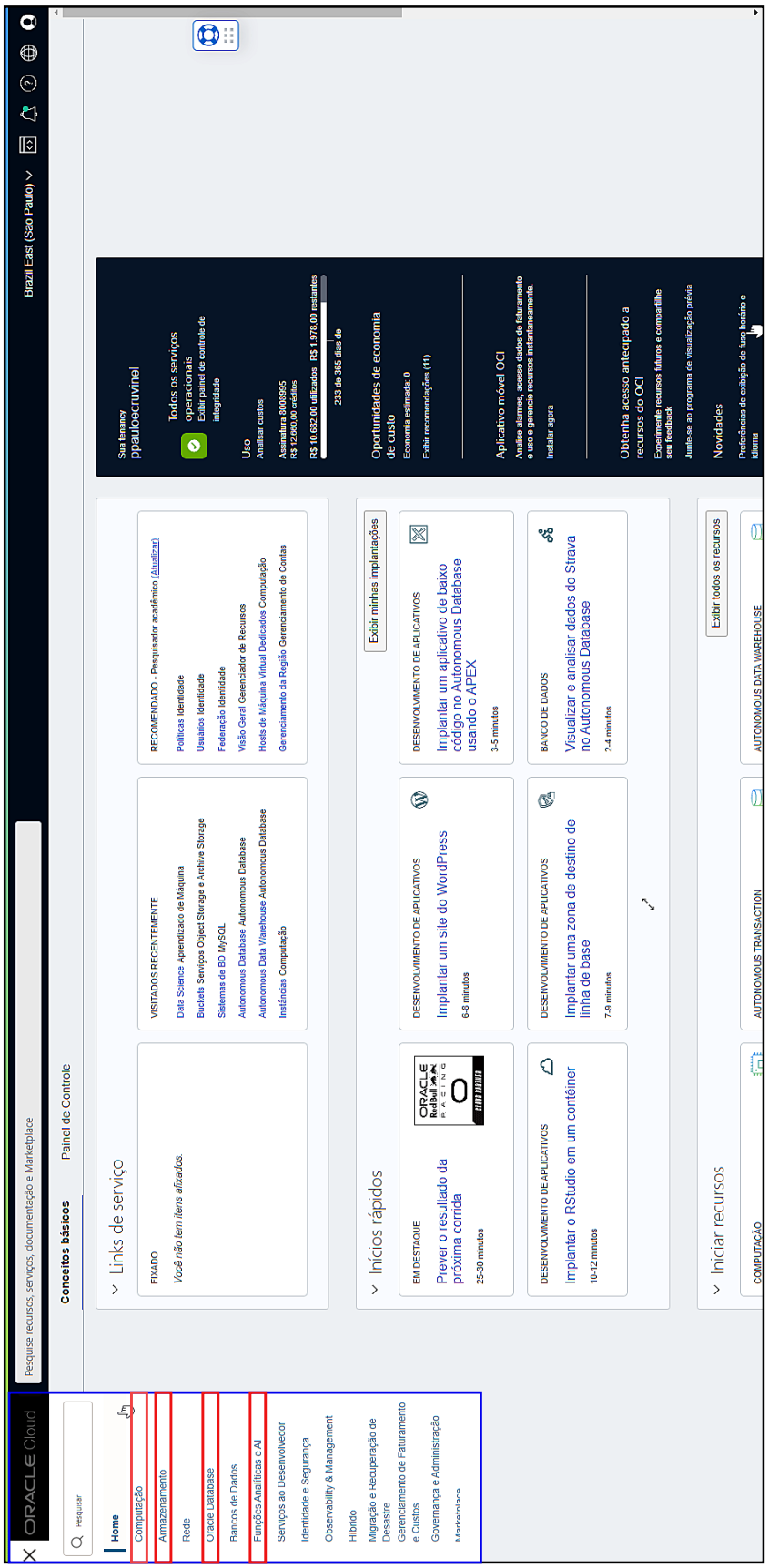
Como primeiro item destacado no menu da tela principal da Figura 32, identificou-se a instância de computação que foi dimensionada para hospedagem do código desenvolvido em linguagem de programação *Python*, a qual proporcionou acesso externo, via IP público, para o processamento do sistema na fase de fusão de dados, via *framework* web. A instância de computação, conforme a Figura 33, foi configurada para uso do sistema operacional Oracle Linux 8.0, com a configuração de 1 OCPU[1] em arquitetura AMD e 16GB de memória RAM. O acesso à instância foi realizado de forma remota, por meio do protocolo SSH, via aplicativo *Putty*, com criptografia de chaves pública e privada de 2048 bits, a qual também foi visualizada.

No segundo item do menu, foi destacado o armazenamento de objetos denominado *buckets*, dividido de acordo com a estruturação do processamento para prover recursos de armazenamento, tanto de entrada das fontes de dados, quanto para a saída do processamento. Na Tabela 9, se encontra apresentada uma visão geral sobre o armazenamento de objetos com a identificação, descrição e devida funcionalidade na infraestrutura, na sequência dos processamentos. A Figura 34 ilustra o ambiente de armazenamento de objetos, que foi configurado de acordo com as respectivas identificações e funcionalidades na infraestrutura *cloud*.

O terceiro item do menu (Figura 32), retratado em detalhes na Figura 35, se refere ao item Oracle *Database*, cuja configuração é relacionada com os bancos de dados transacional (relacional) e multidimensional (DW) respectivamente. As configurações em comum das arquiteturas, que envolveram tais bancos de dados, foram as seguintes: 1 OCPU, 20GB de espaço de armazenamento, versão 19C do banco de dados Oracle e acesso à conexão externa via credenciais *Wallet*.

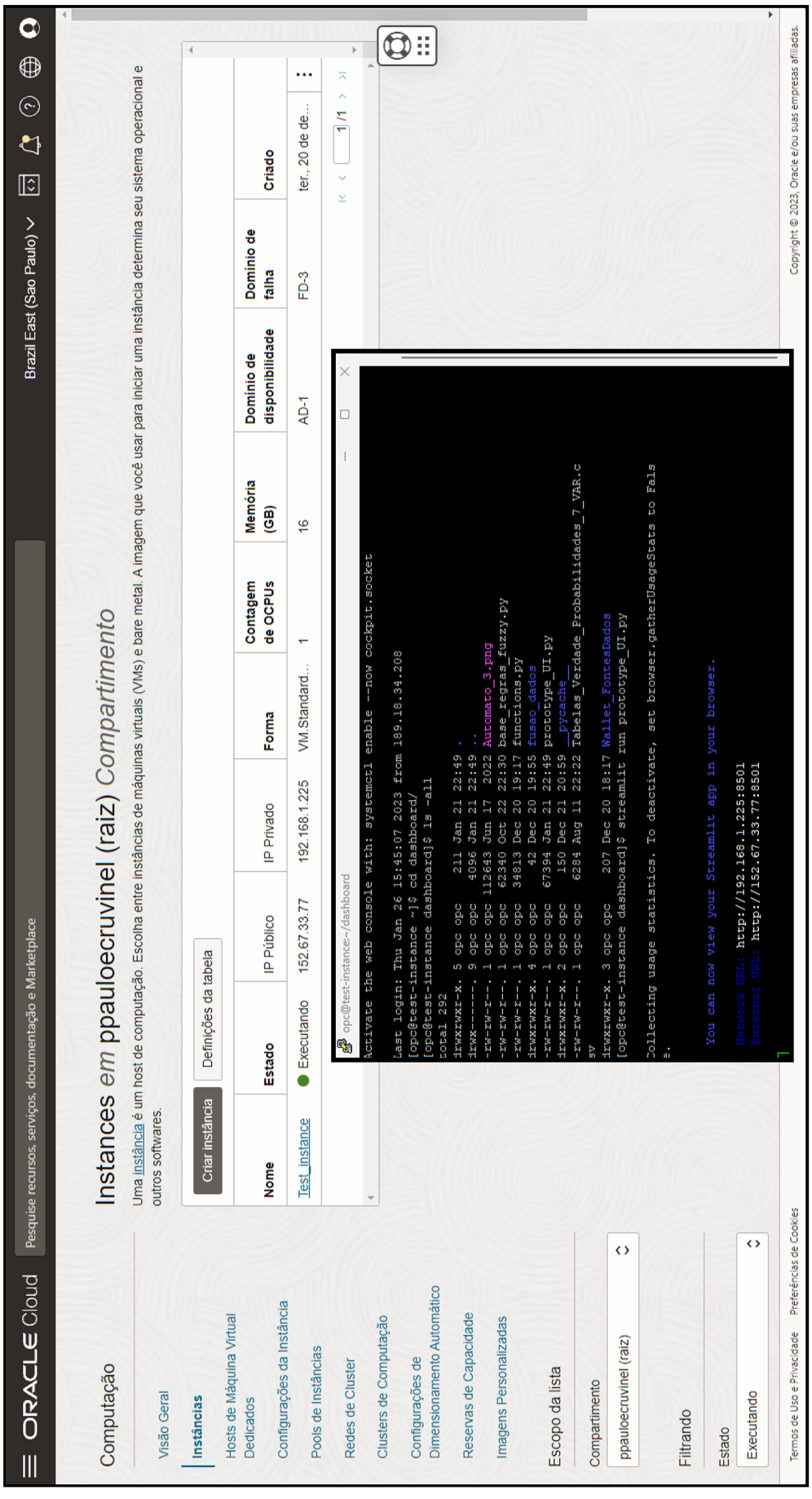
[1] Uma OCPU (*Oracle Compute Unit*) é a unidade de processamento que a Oracle utiliza para criação do serviço. Quanto maior o valor de OCPUs, maior a potência de processamento.

Figura 32 – Oracle *Cloud* - *Tela Inicial*



Fonte: Próprio Autor

Figura 33 – Oracle *Cloud* Acesso Remoto Instância Computação



Fonte: Próprio Autor

Figura 34 – Oracle *Cloud* - *Buckets*

Criar Bucket

Nome	Camada de Armazenamento Padrão	Visibilidade
BK-classification	Padrão	Privado
BK-feature-selected	Padrão	Privado
BK-feature-vector	Padrão	Privado
BK-images-1	Padrão	Privado
BK-input-data-classification	Padrão	Privado
BK-segmented-images-1	Padrão	Privado
BK-segmented-images-2	Padrão	Privado
BK-segmented-images-selected	Padrão	Privado

Fonte: Próprio Autor

Tabela 9 – Descrição dos *Buckets* Oracle *Cloud*

Armazenamento de Objetos (*Buckets*)		
Identificação	**Função**	**Descrição**
BK-images-1	Entrada	Cópia manual de imagens da fonte de dados da Embrapa.
BK-segmented-images-1	Entrada	Cópia automática, via script python, do processamento da etapa 1 de segmentação.
	Saída	Cópia manual das imagens selecionadas da etapa 1 para o Bucket BK-segmented-images-selected (processar etapa 2 da segmentação).
BK-segmented-images-selected	Entrada	Cópia manual das imagens selecionadas dos clusters da etapa 1 de segmentação, para processar etapa 2 (segmentação).
BK-segmented-images-2	Entrada	Cópia automática, via script python, do processamento da etapa 2 de segmentação.
	Saída	Cópia manual das imagens, selecionadas da etapa 2 (segmentação), para o Bucket BK-feature-selected (processar extração de características).
BK-feature-selected	Entrada	Cópia manual das imagens selecionadas, processadas na etapa 2, para as cores verde, amarela e marrom.
BK-feature-vector	Entrada	Cópia automática, via script python, do resultado do processamento das características SIFT, HOG e Momentos de HU para as cores verde, amarela e marrom.
BK-input-data-classification	Entrada	Cópia manual dos dados de características selecionadas, por imagem, para o processamento do algoritmo de classificação SVM.
BK-classification	Entrada	Cópia automática, via script python, do resultado com todos os dados do processamento da classificação SVM.

Fonte: Próprio Autor

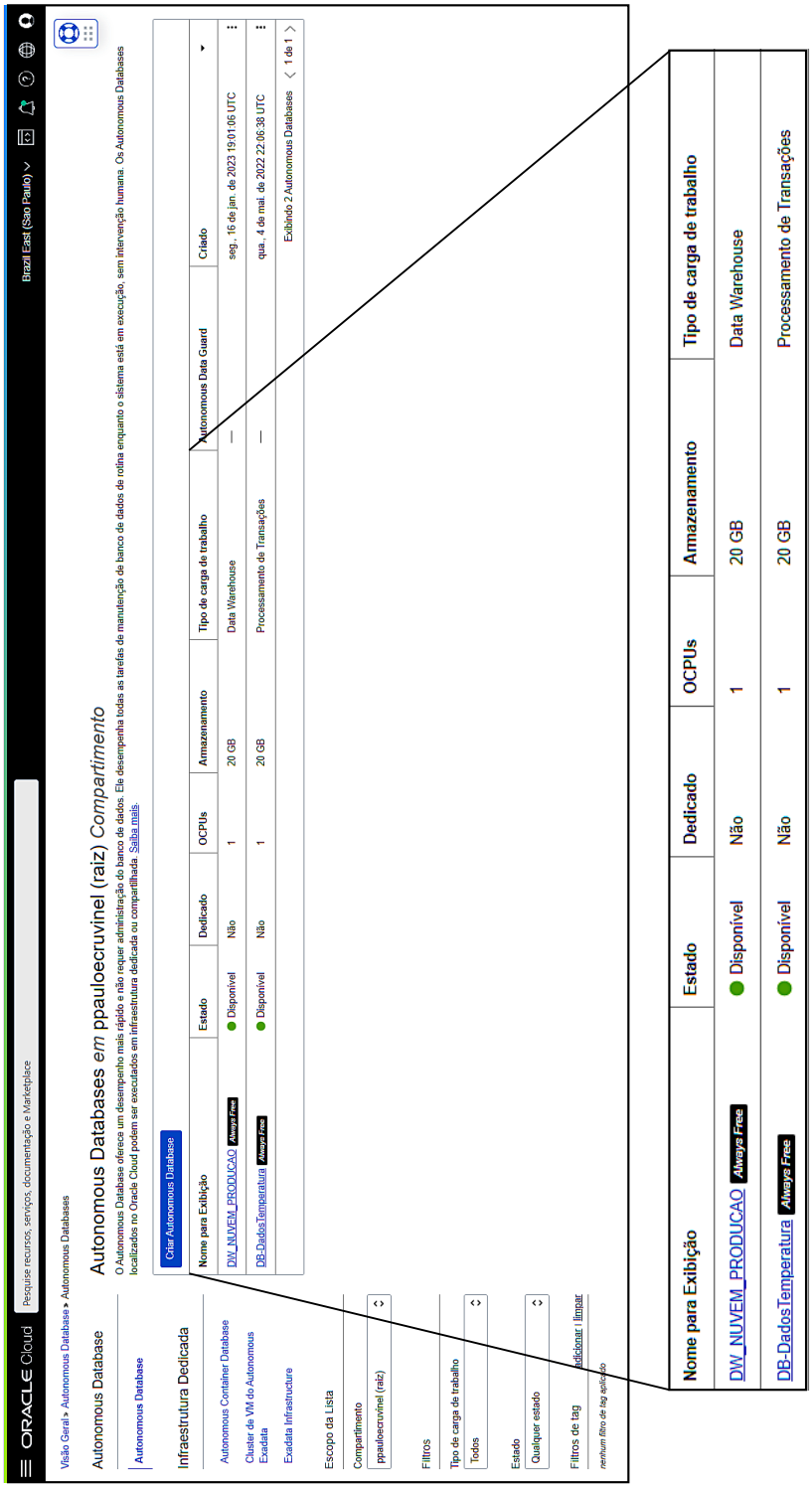
O último item elencado na tela inicial (Figura 32) incluiu o ambiente de desenvolvimento para ciência de dados. Para uso dos recursos, foi necessário ativar uma sessão de *notebook*, a qual incluiu uma instância de computação configurada no momento da criação do projeto *Data Science*. Na Figura 36, é possível observar a infraestrutura de criação desta instância e também a sessão de *notebook* utilizada.

A instância configurada para o processamento dos códigos *Python*, no ambiente de ciência de dados, foi preparada com a arquitetura do processador AMD, com 4 OCPU's de processamento, 64 GB de memória RAM e um disco de 250GB para armazenamento dos resultados do processamento, na forma de computação ou *shape* VM.Standard.E3.Flex, que permitiu de 1 até no máximo de 64 OCPU's. A visualização da sessão de *notebook* foi dividida em três partes, sendo a primeira ficou reservada ao menu, localizado ao lado esquerdo da tela, onde ao clicar sobre os itens, foi possível exibir as pastas, arquivos e detalhamentos correspondentes. Para a exibição ficou reservado a segunda parte da tela. Quanto à terceira divisão da sessão de *notebook*, a mesma referiu-se ao ambiente *JupyterLab*, disponível em abas para o desenvolvimento de código *Python*, bem como a integração deste com os demais ambientes da Oracle *Cloud*. Para este arranjo, encontrou destaque o console de linha de comando do linux, como principal elemento integrador, o qual permitiu o gerenciamento da instância do *notebook* em termos gerais.

Para os projetos relacionados ao desenvolvidos do sistema, neste ambiente *Data Science*, foi viabilizada a criação de pastas no disco local da instância com a mesma identificação dos *buckets* para a transferência de todas as imagens digitais de folhas da soja, necessárias aos processamentos. Como exemplo deste arranjo, é possível considerar a pasta "segmented-images-1", conforme apresentada na Figura 36. Essas pastas foram sincronizadas com os *buckets*, a cada início de processamento. O sincronismo dos arquivos foi

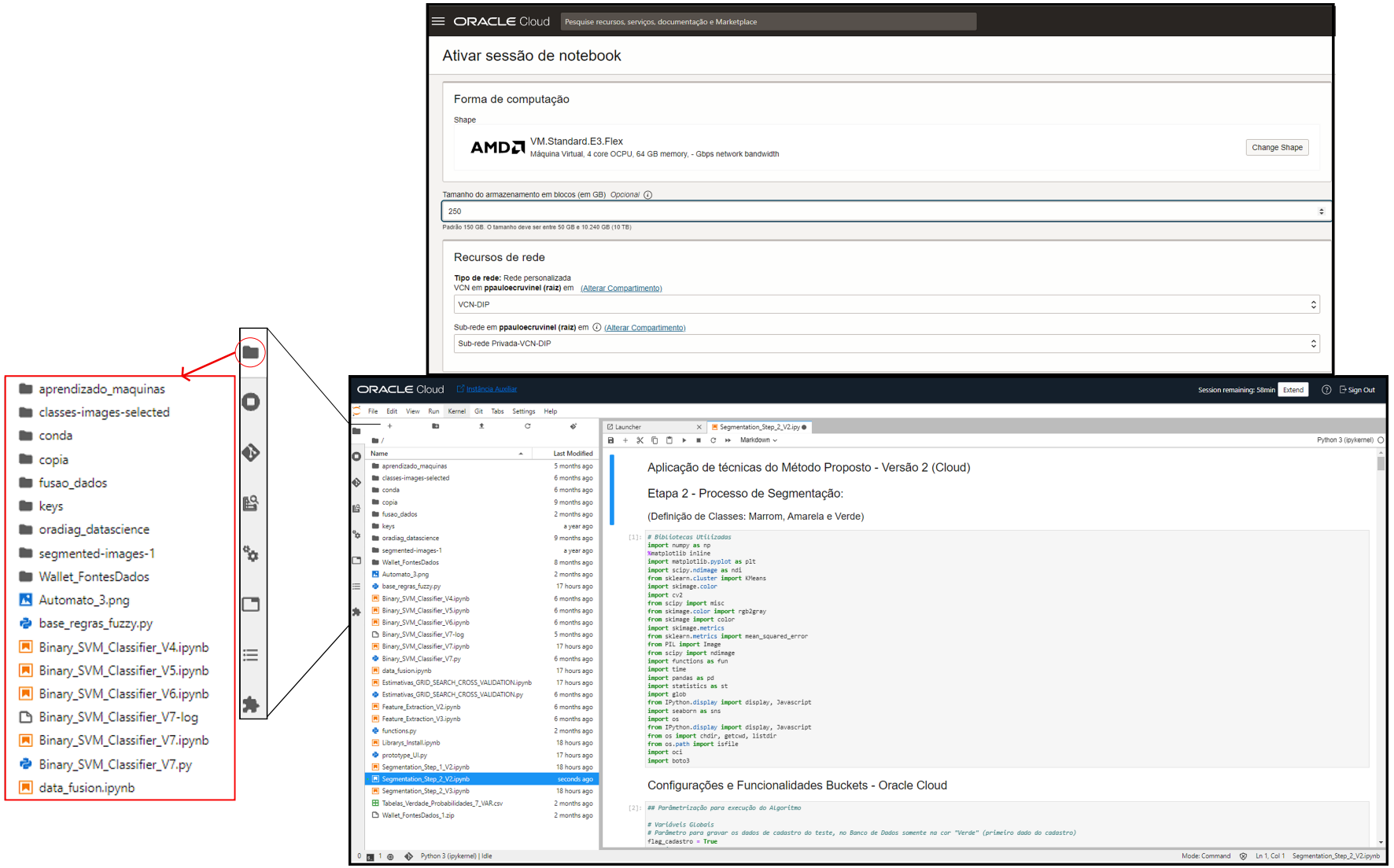
realizado entre as pastas locais e os *buckets*, via comandos em linha com o uso do software *rclone* para a versão linux. No entanto, para a gravação de arquivos dos resultados, durante os processamentos nos *buckets*, foram utilizadas as *Applications Programming Interfaces* (APIs) OCI e Boto3 para a linguagem *Python*.

Figura 35 – Oracle *Cloud* - *Autonomous Database*



Fonte: Próprio Autor

Figura 36 – Oracle *Cloud - Data Science*



Fonte: Próprio Autor

4.1.1 Estruturação de Dados das Séries Temporais

As fontes de dados, obtidas com apoio do INMET (dados climáticos) e da Embrapa (imagens de plantas da soja), deram suporte ao desenvolvimento da pesquisa. Essas fontes de dados, assim como os dados processados, foram armazenadas em banco de dados da plataforma da Oracle *Cloud* e tratadas como séries temporais de dados. Tal ação possibilitou a integração com as tecnologias da nuvem, tais como ciência de dados e instância de computação linux, favorecendo o desenvolvimento dos algoritmos em *Python* e a completa implementação web do sistema.

Neste contexto, os dados climáticos das séries temporais, correspondentes à região, localização, ano, ciclos de culturas e datas, possibilitaram estabelecer diferentes configurações em auxílio à avaliação da etapa de fusão das variáveis, incluindo a classificação decorrente da análise das imagens de plantas da soja.

As imagens das folhas de plantas de soja foram coletadas em campos experimentais e adicionadas a um fundo complexo, em laboratório, pela equipe da Embrapa e então disponibilizadas em repositório público. Este *dataset* de imagens foi escolhido para a pesquisa, por apresentar as condições adequadas à reprodução de experimentos, em ambiente controlado, inferindo desafios computacionais, principalmente para etapas de segmentação das folhas, visando o estabelecimento de parâmetros de referência para serem aplicados não somente em imagens coletadas com o uso de câmeras digitais, como também em imagens coletadas com câmeras multiespectrais, embarcadas em veículos aéreos não tripulados (VANTS). Ainda nesse contexto, fez-se importante observar que, na organização do *dataset* para as imagens digitais, foram considerados a variedade da cultura, a distância entre linhas de plantio, a altura da planta de soja e a quantidade de plantas por metro linear.

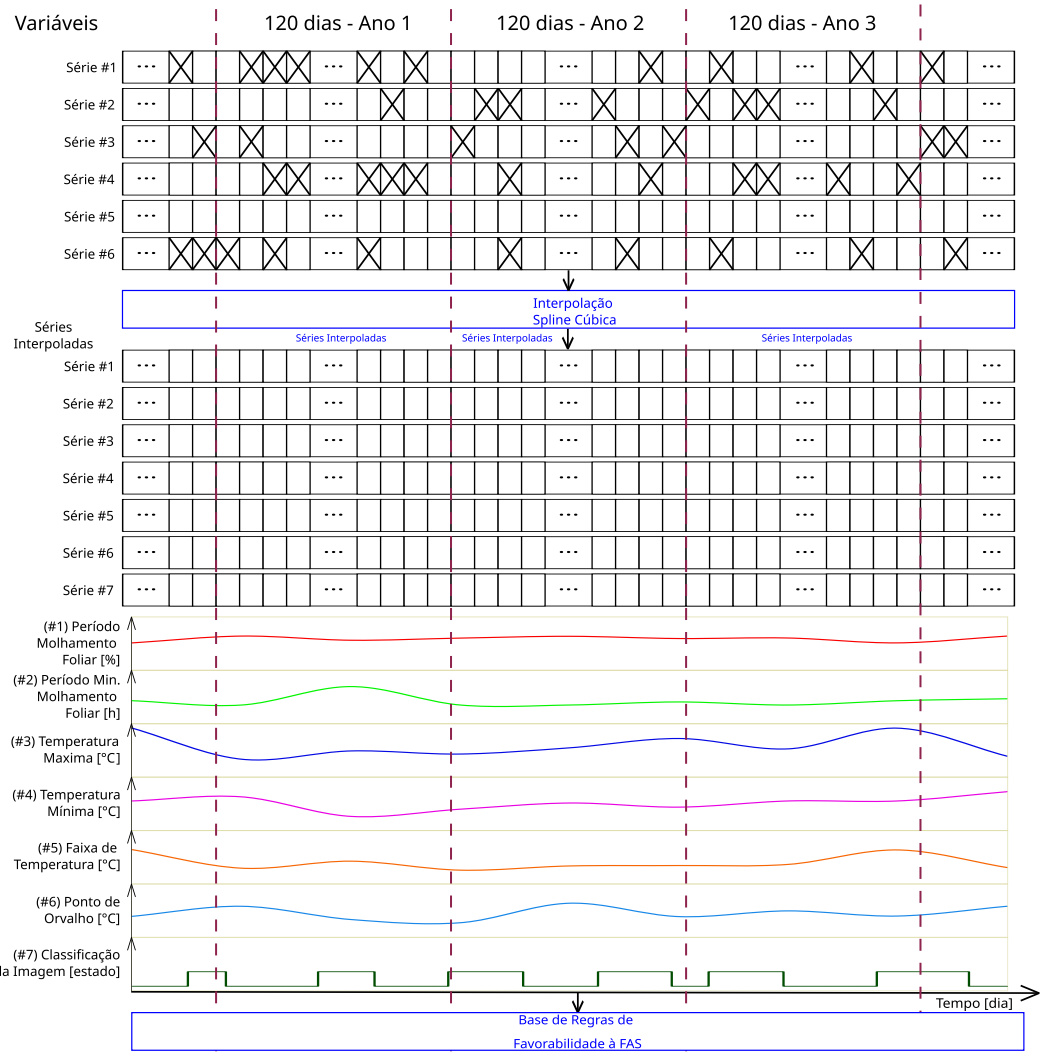
Adicionalmente, as séries temporais de dados climáticos, constituídas com histórico de vinte anos, possibilitaram avaliar um período de três ciclos da cultura, um em cada ano, o que correspondeu, para cada um deles, períodos de cento e vinte dias. Também, as séries temporais de dados das imagens disponíveis foram divididas em três grupos, sendo dois deles contendo vinte e uma imagens, utilizadas com dois primeiros ciclos da cultura considerados, bem como um terceiro grupo com vinte e duas imagens, para uso nas análises correspondentes ao terceiro período do ciclo da cultura. Para fins de avaliação do sistema, essas imagens foram distribuídas, ao longo das séries temporais de dados climáticos, de forma aleatória e sem repetição, visando as avaliações sobre a ocorrência da ferrugem asiática, nos períodos considerados.

A Figura 37 ilustra o arranjo utilizado na organização e uso das séries temporais de dados das variáveis consideradas para o suporte à decisão. Para as localizações específicas nas séries temporais, onde ocorreu a falta de dados, as mesmas foram preenchidas lançando mão de interpoladores. Para tanto, foram avaliados a aplicação de modelos interpoladores polinomiais de grau 1 até grau 5, incluindo também a avaliação de interpolação baseada

no uso de *b-spline* cúbica.

A partir daí, as séries completas de dados foram utilizadas como entradas para o modelo de fusão, em janelas de tempos específicas, de forma a viabilizar a análise sobre a ocorrência ou não da doença e seus estados de severidade, ou seja, considerando o conjunto de sete variáveis para o modelo de decisão.

Figura 37 – Arranjo para a Série Temporal de Dados das Variáveis Utilizadas no Suporte à Decisão



Fonte: Próprio Autor

Para o conjunto das séries temporais analisadas, tomando como exemplo uma dessas avaliações, a Tabela 10 ilustra registros incompletos decorrentes de um janelamento, durante o período de 30/10/2017 a 08/11/2017. Para esse caso, os registros de três posições da janela foram completados com o uso da interpolação, ou seja, posições de 8 a 10, conforme destacado na cor amarela.

Quanto à escolha do modelo de interpolação, para o exemplo considerado, a Figura 38

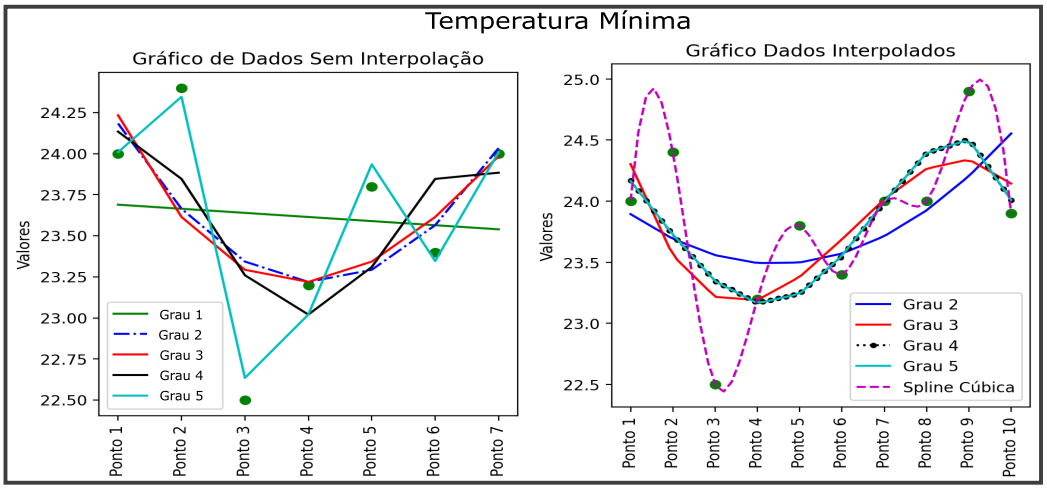
ilustra duas análises para a variável temperatura mínima. A primeira plotagem retrata, para a janela considerada, os dados originais, onde apareceram exclusivamente os sete pontos de medição da variável. A segunda plotagem ilustra os resultados obtidos com os diferentes interpoladores considerados, para se completar a série temporal de dados.

Tabela 10 – Exemplos de Janelamento nas Séries Temporais de Dados

N.	Data de Medição	Precip.	Temp. Max.	Temp. Mín.	Umidade Relativa	Ponto Orvalho	Temp. Méd. Compensada	Status
1	30/10/2017	4,20	35,50	24,00	72,75	23,08	28,44	Original
2	31/10/2017	0,00	32,50	24,40	88,75	23,80	25,80	Original
3	01/11/2017	18,00	33,30	22,50	79,00	22,85	26,80	Original
4	05/11/2017	0,00	33,00	23,20	84,00	22,62	25,52	Original
5	06/11/2017	0,00	33,60	23,80	88,25	24,02	26,12	Original
6	07/11/2017	3,00	34,50	23,40	83,00	23,06	26,18	Original
7	08/11/2017	0,00	33,50	24,00	84,25	23,47	26,34	Original
8	08/11/2017	4,20	35,50	24,00	72,80	23,10	28,40	Interpolado
9	08/11/2017	6,10	32,80	24,90	88,70	23,80	25,90	Interpolado
10	08/11/2017	5,40	32,60	23,90	86,80	23,70	26,00	Interpolado

Fonte: Próprio Autor

Figura 38 – Avaliação de Diferentes Modelos para Completar as Séries Temporais de Dados



Fonte: Próprio Autor

Nos resultados, o melhor ajuste da curva observado, inclusive para o exemplo considerado, ocorreu para a interpolação com *b-spline* cúbica (destacados em cor amarela). A Tabela 11 ilustra, para o exemplo, os valores dos coeficiente de correlação para os vários modelos analisados.

As diferentes fontes de dados, identificadas como *dataset* de imagens, dados climáticos e de plantas, assim como as fontes de dados armazenadas nos *buckets*, foram preparadas de acordo com as dimensões de qualidade, integridade e completude. Tal preparação ocorreu na execução da Extração, Transformação e Carga (ETL), a fim de carregar os dados no banco de dados transacional. Também, a dimensão de consistência foi verificada no

momento da carga de dados, pois as chaves primárias e estrangeiras, utilizadas para o relacionamento das tabelas do modelo transacional, garantiram a integridade referencial.

Tabela 11 – Valores dos Coeficientes de Correlação para os Vários Modelos Analisados no Exemplo Considerado

Sem Interpolação	R^2				
	Reta	**Ordem 2**	**Ordem 3**	**Ordem 4**	**Ordem 5**
Precipitação	0,083	0,133	0,208	0,210	0,727
Temperatura Máxima	0,017	0,316	0,828	0,837	0,980
Temperatura Mínima	0,007	0,348	0,354	0,424	0,969
Umidade Relativa	0,202	0,382	0,440	0,567	0,988
Ponto Orvalho	0,017	0,038	0,038	0,177	0,822
Temperatura Med.Compensada	0,248	0,613	0,710	0,760	0,865

Com Interpolação	R^2					
	Reta	**Ordem 2**	**Ordem 3**	**Ordem 4**	**Ordem 5**	**Spline Cúbica**
Precipitação	0,002	0,038	0,112	0,218	0,244	0,657
Temperatura Máxima	0,030	0,032	0,575	0,638	0,646	0,773
Temperatura Mínima	0,113	0,274	0,483	0,525	0,525	0,823
Umidade Relativa	0,074	0,094	0,353	0,354	0,363	0,633
Ponto Orvalho	0,108	0,150	0,151	0,151	0,187	0,823
Temperatura Med.Compensada	0,036	0,114	0,527	0,527	0,530	0,718

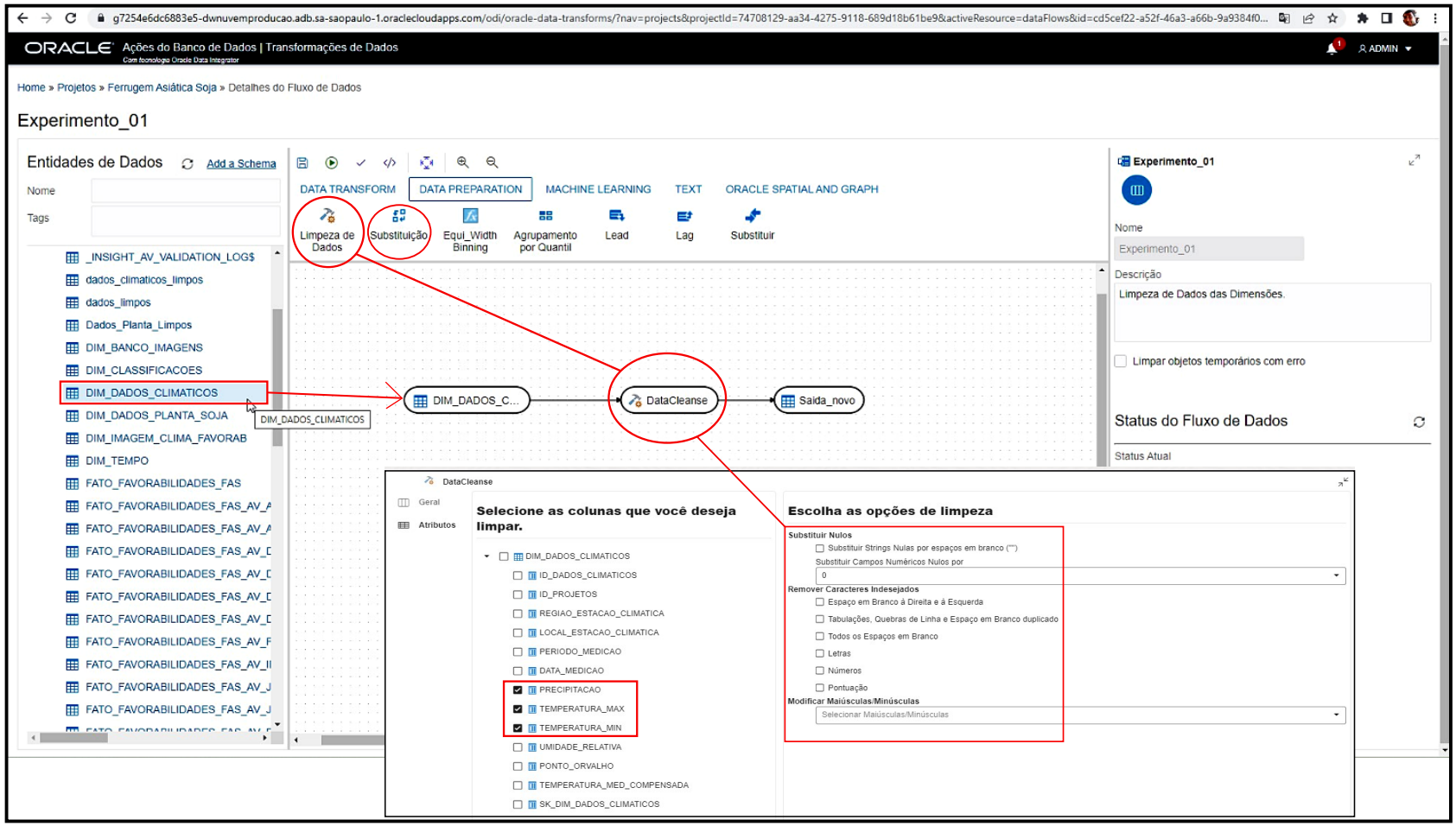
Fonte: Próprio Autor

O modelo de estruturação das bases de dados (Figura 15) contemplou, via DW, o armazenamento das diversas fontes de dados, frente ao cubo de dados dimensionado pelo modelo multidimensional, conforme modelo descrito no Apêndice A, seguido pela carga dos dados no SGBD transacional. Nesse contexto dada à necessidade da preparação dos dados, como pré-requisito de qualidade, foi utilizada a ferramenta Transformação de Dados do *Data Studio*, disponível nos serviços do *Autonomous Database* da Oracle *Cloud*. Tal ferramenta (Figura 39) viabilizou o uso de um conjunto de recursos, tais como a "limpeza de dados" e a "substituição", que possibilitou a preparação dos dados para atender às dimensões de integridade e completude, de forma semi-automatizada.

Adicionalmente, os critérios de preparação de dados, definidos como padrão, para a limpeza de dados, foram: (1) registros nulos; e (2) verificação de caracteres indesejados (caracteres numéricos em dados não numéricos e *strings* em dados numéricos). No âmbito da operação do sistema, em particular no que tange essa estruturação e organização dos dados, não houve necessidade de se promover a operação de limpeza, vez que durante a verificação não foram identificados casos que demandassem esse nível de tratamento. Entretanto, a disponibilidade dessas funcionalidades se fazem necessárias, dado a variabilidade que as séries temporais podem encontrar em diferente área de cultura da soja existentes.

A etapa de preparação dos dados, via Ferramenta de Transformação *Cloud*, para o conjunto de dados analisados, em relação aos *datasets* que foram considerados, ainda demandou um complemento na preparação da etapa ETL dos *buckets* executada, no que tange ao banco de dados transacional. Nesse caso, o número de tratamentos foi menor, porém continuou indispensável a execução desta etapa, de forma a se poder certificar que o vetor de dados estivesse nos padrões estabelecidos para o nível de qualidade desejado para o processo relacionado à fusão dos dados.

Figura 39 – Preparação dos Dados Ferramenta Oracle *Cloud*



Fonte: Próprio Autor

Ainda em relação a Figura 39, é possível também observar no seu lado direito a estrutura do DW, como Entidade de Dados. No centro desta figura se encontram ilustrados os recursos de limpeza de dados e demais detalhamentos das telas para as configurações. Destacado em cor vermelha, o quadro "DIM_DADOS_CLIMATICOS" se refere às colunas de precipitação, temperatura mínima e temperatura máxima, as quais eventualmente em função dos dados disponíveis, podem requer recursos para limpeza, o que implica na melhoria da qualidade de dados.

4.2 Pré-Processamento e Processamento Digital das Imagens

O pré-processamento das imagens foi realizado considerando as operações de equalização do histograma da imagem, incluindo a aplicação do filtro de mediana.

Após a divisão dos canais RGB, foi aplicada, ao canal verde, na etapa I da segmentação, a equalização do histograma da imagem, o que permitiu a redução das diferenças de brilho e contraste, contribuindo para a retirada do fundo (Figura 40). O trabalho sobre o canal verde, com prioridade, se mostrou mais adequado em relação aos demais canais vermelho e azul, de acordo com o contexto estudado, vez que o mesmo apresenta espectro de frequência predominante frente a outras componentes de frequências presentes em áreas de culturas agrícolas.

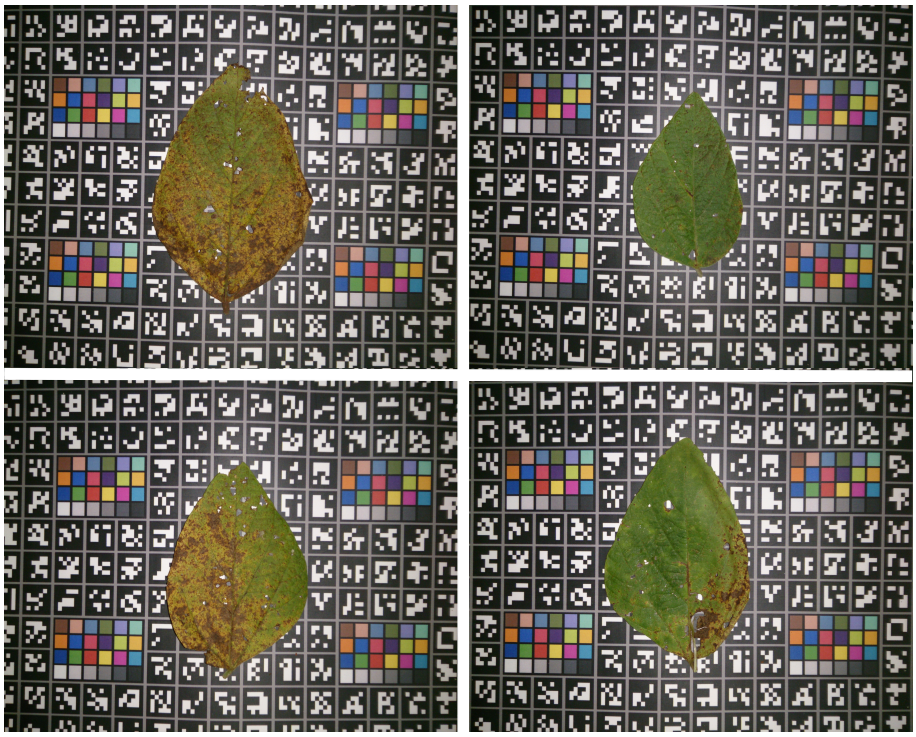
Assim, ao observar o histograma do canal verde equalizado de segmentação para a retirada do fundo, etapa I, (Figura 41), parte dos quadrados coloridos, do fundo da imagem, foram modificados em relação ao objeto de interesse, onde diante da equalização, refletiu o aumento da contagem dos valores de intensidades dos *pixels*.

Adicionalmente, ao final da etapa da segmentação a utilização do filtro de mediana, com janela de filtragem de 3x3, viabilizou suavizar a imagem (melhoria do SRN), o que auxiliou na extração de características para o reconhecimento dos padrões. Nesse sentido, conforme ilustra a Figura 42, puderam ser melhor observados os diferentes valores de intensidades, incluindo a minimização de valores de desvio padrão.

Em relação ao processamento das imagens, as análises envolveram a aplicação das técnicas de segmentação e extração de características para o reconhecimento de padrões.

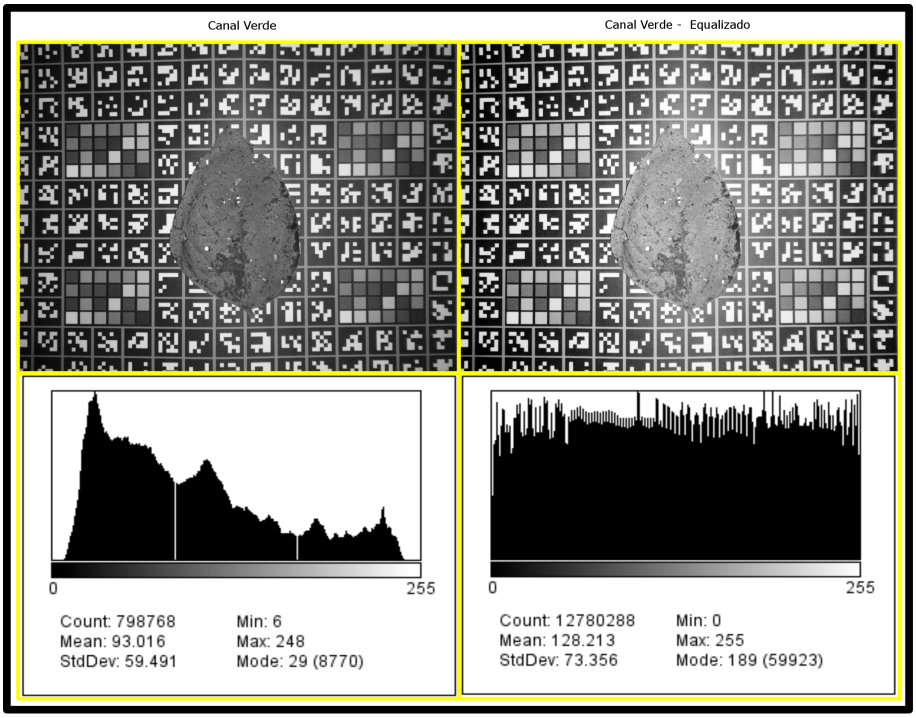
Em busca da robustez da segmentação, foram combinadas as técnicas de limiarização (Equação 2), orientada pela busca dos limiares (Figura 43), por meio de análise supervisionada do histograma das imagens, e a técnica de clusterização (Equação 4) via técnica *K-Means*. Tal associação de técnicas de segmentação se mostrou relevante na obtenção dos resultados, pois somente o uso da técnica *k-means* não foi suficiente para eliminar o fundo complexo das imagens, pela diversidade de cores envolvidas e tonalidades muito próximas, considerando a folha e o fundo.

Figura 40 – Imagens de Folhas de Soja com Fundo Complexo



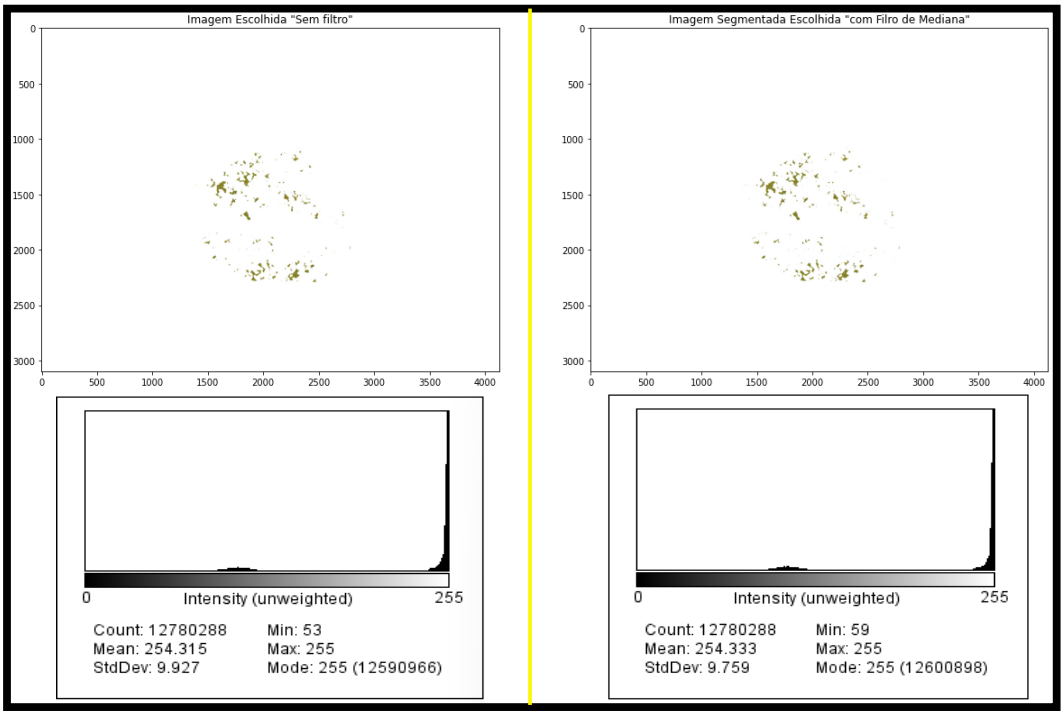
Fonte: Próprio Autor

Figura 41 – Pré-Processamento Equalização de Histograma



Fonte: Próprio Autor

Figura 42 – Pré-Processamento Suavização



Fonte: Próprio Autor

A segmentação foi adaptativa para a necessidade de novos conjuntos de imagens, a partir de diferentes fontes de coleta, seja por meio terrestre, VANTS ou outras. Na validação do sistema desenvolvido houve a utilização das séries temporais de dados de imagens coletadas diretamente em área de cultura da soja com câmera digital de alta resolução (12,78 Mega *pixels*), entretanto considerando níveis de adaptabilidade, de forma a se poder trabalhar com outras bases de imagens, adquiridas com outros tipos de sensores.

A seleção dos limiares envolveu a análise dos histogramas das imagens e a avaliação das regiões para segmentar os objetos de interesse, considerando que o fundo apresentava uma variedade significativa de tonalidades. Vale ressaltar que o fundo aplicado às imagens do *dataset* foi inserido pela equipe da Embrapa como uma camada de complexidade computacional ao contexto, com a finalidade de incentivar o uso de métodos eficientes para segmentação.

O procedimento adotado para avaliação dos histogramas foi supervisionado, considerando dois limiares para segmentar o fundo da imagem, sem prejudicar a região reservada à área da folha de soja, dada à diversidade de tonalidades de cores, relacionadas ao problema a ser abordado, ou seja, verde, amarelo e marrom, possivelmente associadas à FAS. Contudo, o estudo dos histogramas pautou 6 diferentes faixas, que puderam ser visualizadas na Figura 43, de acordo com os itens: (a) 0 a 85, (b) 31 a 165, (c) 70 a 159, (d) 83 a 159, (e) 100 a 130 e (f) de 18 a 200. Os testes foram realizados com as faixas (b) e (f), pois as demais apresentaram uma perda de *pixels* considerável no objeto de interesse.

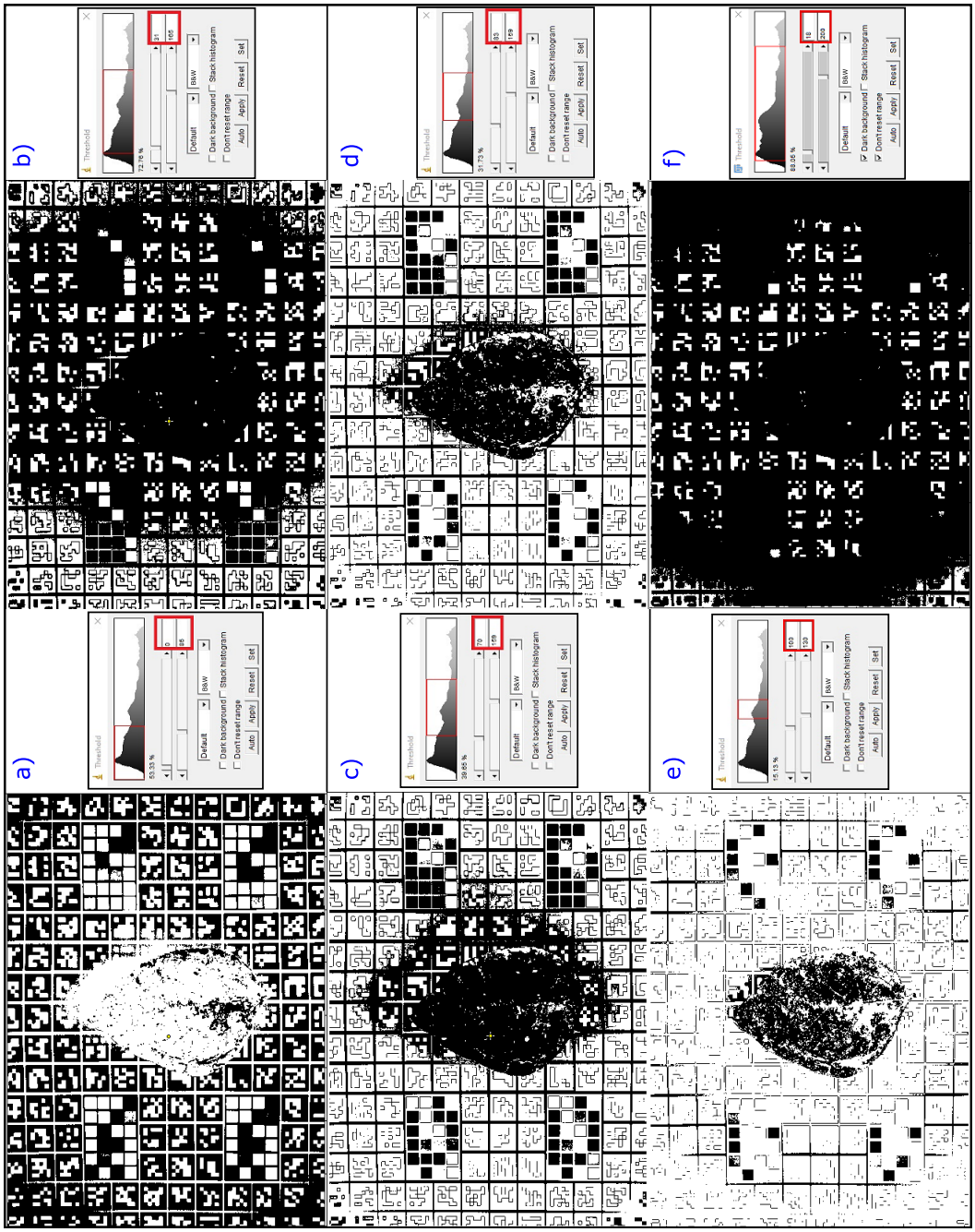


Figura 43 – Escolha dos Limiares para a Segmentação

Fonte: Próprio Autor

Ainda, conforme observado na Figura 43, a faixa de limiares (b), de 31 a 165, trouxe bons resultados e foi adotada, como padrão, para processamento do *dataset* de imagens. No entanto, após a aplicação da técnica *k-means*, percebeu-se que, com a retirada do fundo complexo, a faixa (f) de 18 a 200, também apresentou resultados muito próximos.

Porém, ao final da etapa I da segmentação, foi notada a presença de pequenos fragmentos do fundo presentes na imagem, em alguns casos. A justificativa para a presença de tais fragmentos se deu pelo dimensionamento fixo e supervisionado do tamanho da janela de interesse (ROI), para diferentes tamanhos de folha. Em contrapartida, em 95% dos casos, os fragmentos da etapa I foram eliminados ao final da etapa II.

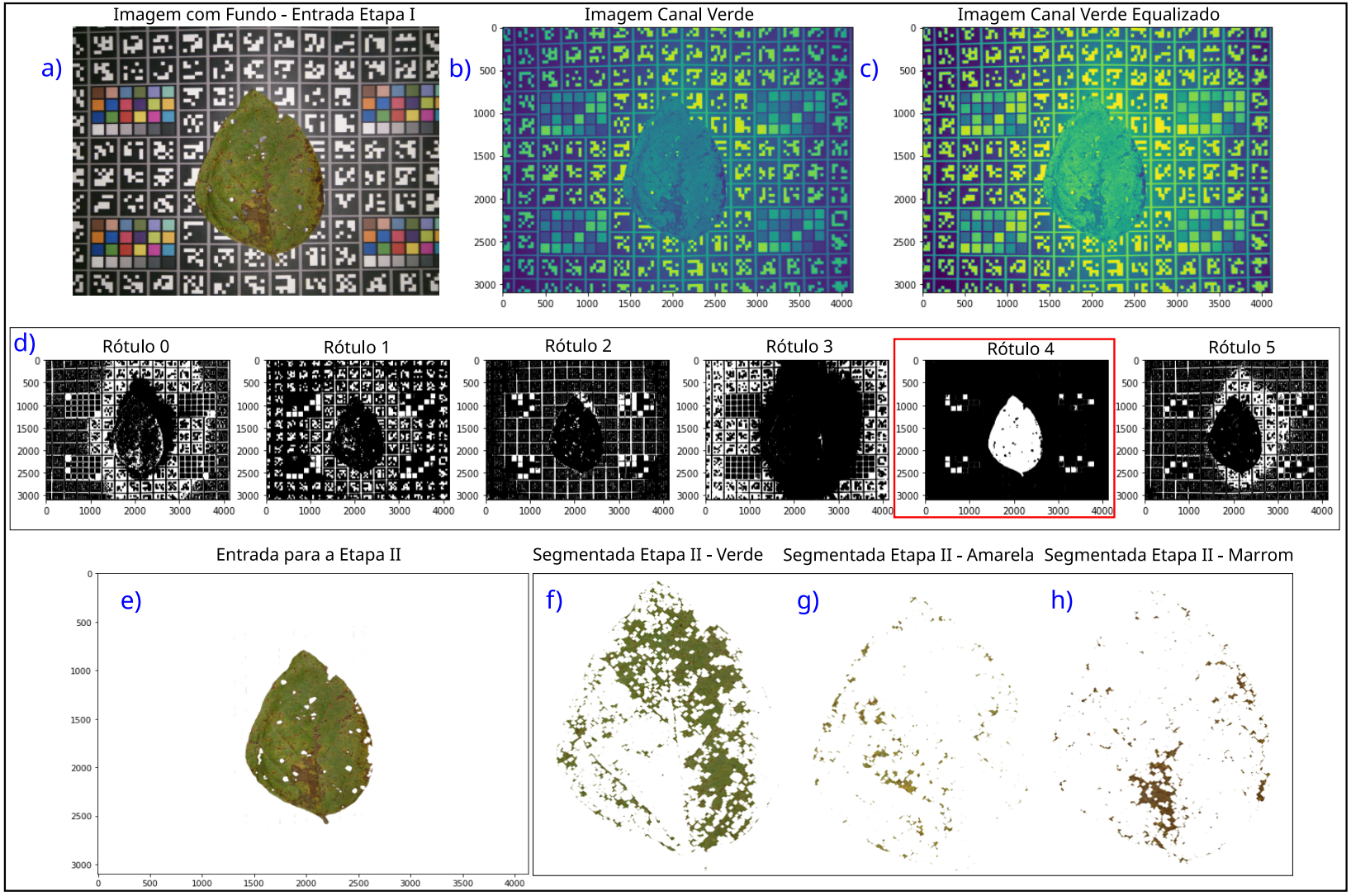
Outro ponto importante, observado na segmentação, foi quanto a definição do número de *clusters* utilizados pelo algoritmo. Para tal, utilizou-se a técnica *Elbow*[2] como referência. O valor calculado pelo algoritmo *Elbow*, diante dos dados submetidos para cálculo de referência, foi para $k = 4$, ou seja, 4 *clusters*. Com esse valor aplicado ao algoritmo *k-means*, o resultado da segmentação, nas etapas I e II, não foram satisfatórios. No entanto, foi obtido o valor de $k = 6$, considerado ideal, conseguido por testes supervisionados.

Ainda, na Figura 44, também é possível observar detalhes sobre as imagens segmentadas, conforme método apresentado na etapa I, ou seja, onde a imagem digital "a" representa a folha da soja com fundo, a imagem digital "b" representa o canal verde e a imagem "c" representa o canal verde, após aplicação da técnica de equalização do histograma. Também, nesse mesmo arranjo, é possível observar, no centro, a imagem digital "d" que representa a rotulagem (seis rótulos), promovida pelo algoritmo *K-Means*, onde o rótulo com a marcação em cor vermelha, foi definido de forma supervisionada. Adicionalmente, se encontram ilustradas as imagens digitais finalizadas na etapa II do processamento, ou seja, as imagens digitais "e" que se referiram ao processo de *matting*, após a escolha do rótulo e a aplicação do filtro de mediana e as imagens digitais "f", "g" e "h" como exemplos de soluções obtidas com a aplicação da técnica de segmentação, considerando os limiares nos intervalos de cada cor de referência (verde, amarela e marrom) e de interesse para a composição do vetor de decisão para a identificação da presença ou não da ferrugem asiática e seus estágios de severidade.

O Algoritmo 2 detalha as rotinas da segmentação consideradas para o sistema com ênfase na segunda etapa (II), cujas as entradas são originadas como saídas da primeira etapa (I) do processo. Para este algoritmo encontram destaques funções que foram estabelecidas para automatização da busca do melhor *pixel* semente, para a coleta dos *pixels* baseado no ponto central, ou seja, na semente encontrada e para o cálculo automático dos limiares a partir dos *pixels* coletados, considerando erro $\leq 5\%$.

[2] A técnica *elbow* ou cotovelo é utilizada para determinar o número ideal de *clusters* em uma análise de clusterização de dados, comumente usada em algoritmos como o *k-means* (CUI et al., 2020).

Figura 44 – Exemplo Segmentação Etapas I e II

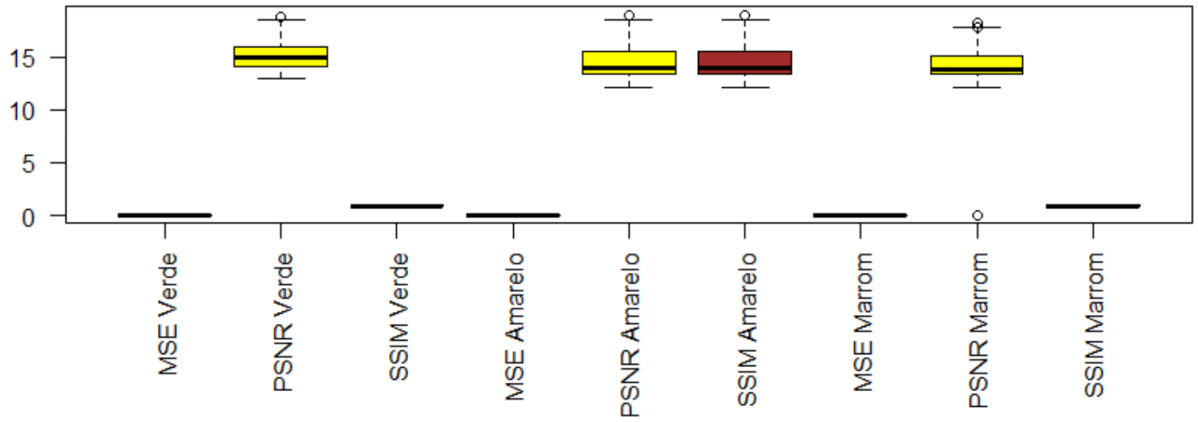


Fonte: Próprio Autor

Durante a operação do sistema, após cada realização de uma ação de segmentação são calculadas as métricas PSNR (Equação 79), MSE (Equação 80) e SSIM (Equação 81) para inferência da qualidade, tanto dos dados de entrada, quanto dos dados processados. A Figura 45 ilustra um exemplo de resultados, a qual exibe graficamente em formato *boxplot* a avaliação dessas métricas, assim como a Tabela 12, o seu detalhamento numérico. Para o exemplo considerado (linha mediana) é possível observar para a métrica MSE os valores de 0,03; 0,04 e 0,04 para as cores verde, amarela e marrom, respectivamente observadas.

Ainda para o exemplo considerado a avaliação da métrica PSNR apresentou valores próximos, indicando que a SNR se mostrou equivalente entre a imagem original e as imagens segmentadas, ou seja, em torno de 14,92. Quanto à aplicação da métrica SSIM, a mesma ficou em torno de 0,94, indicando alta similaridade entre as regiões da imagem original e as regiões segmentadas.

Figura 45 – Boxplot Métricas Segmentação



Fonte: Próprio Autor

Tabela 12 – Dados Métricas Segmentação

	MSE VD	PSNR VD	SSIM VD	MSE AM	PSNR AM	SSIM AM	MSE MR	PSNR MR	SSIM MR
Min.:	0,01	12,90	0,89	0,01	12,07	0,89	0,01	0,04	0,89
1st Qu.:	0,03	14,08	0,93	0,03	13,32	0,92	0,03	13,31	0,92
Median.:	0,03	14,92	0,94	0,04	14,01	0,93	0,04	13,77	0,93
Mean.:	0,03	15,17	0,94	0,04	14,46	0,94	0,04	14,00	0,94
3rd Qu.:	0,04	15,87	0,95	0,05	15,44	0,95	0,04	15,07	0,95
Max.:	0,05	18,74	0,97	0,06	18,98	0,97	0,06	18,23	0,97

Fonte: Próprio Autor

A extração de características foi iniciada considerando a identificação dos padrões de cores textura e forma pelos algoritmos *Scale-Invariant Feature Transform* (SIFT), *Histogram of Oriented Gradients* (HOG) e Momentos Invariantes de HU. Logo, a partir

dos padrões de cores da doença, segundo a literatura, a cor amarela representou o estágio intermediário da doença e, a cor marrom, o estágio avançado da doença.

Para a evolução da doença, em relação aos descritores, tanto para as cores verde, amarelo e marrom, como de textura e forma, todos contribuíram, juntamente com as demais variáveis consideradas para a avaliação da evolução da favorabilidade, quanto ao desenvolvimento do patógeno, na área de cultura. Quando presente exclusivamente a cor verde na imagem da folha de soja, a mesma foi tomada como um indicador de possível ausência da FAS.

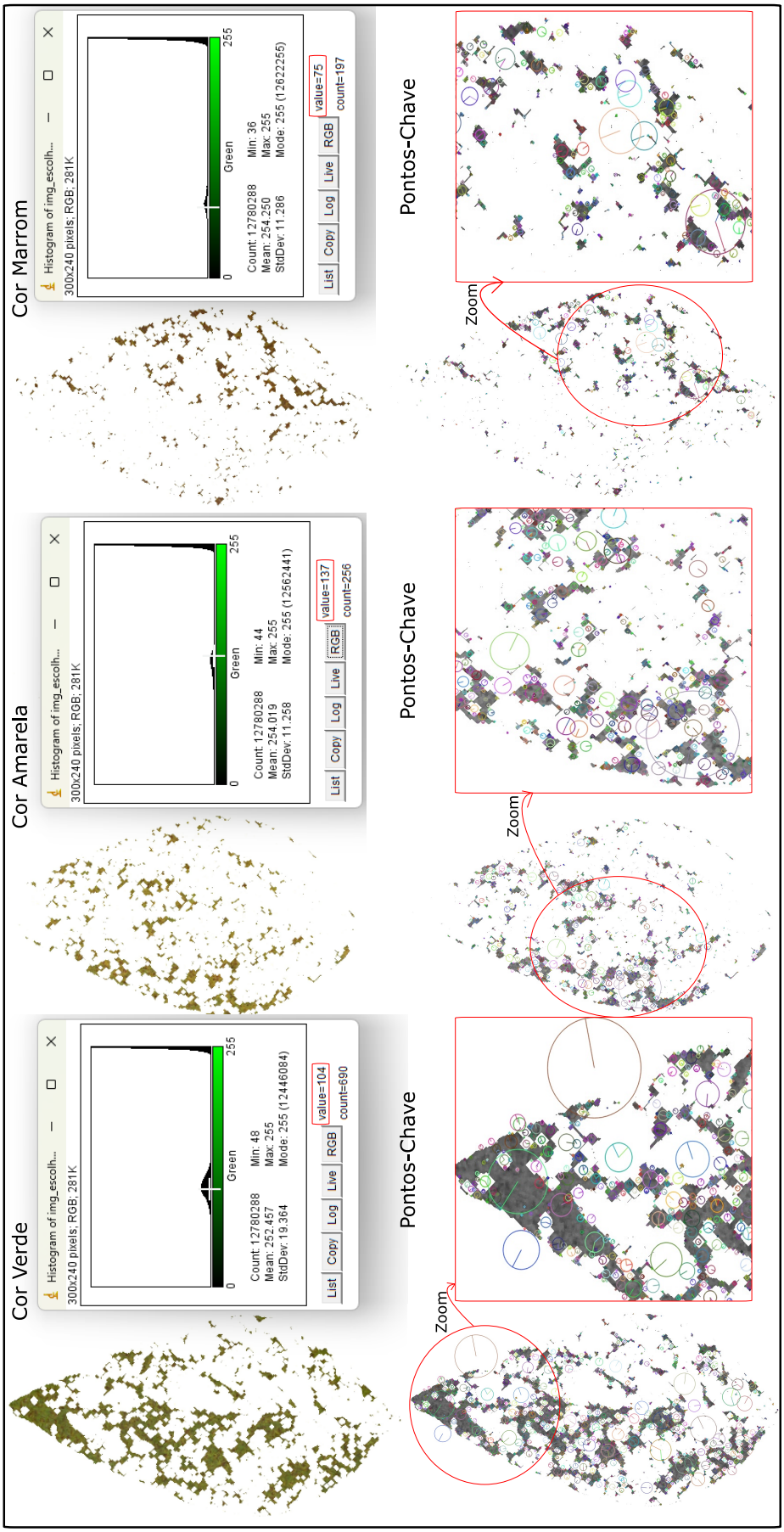
Para a aplicação dos algoritmos de extração de características foram considerados os seguintes valores dos *pixels* de referência: verde (75,**104**,29); amarela (153,**137**,50) e marrom (109,**75**,38). Adicionalmente, a questão associada à presença de ruídos nas imagens foi solucionada, considerando o uso de filtragem digital com filtro de mediana. As questões relacionadas as diferenças de iluminação na aquisição de imagens digitais no ambiente de campo foram contornadas com a associação dos dois outros descritores relacionados à textura (HOG) e as características geométricas (HU) que surgem na folha da planta, de forma aleatória, em função da dinâmica da infestação.

Na Figura 46 podem ser observadas as imagens segmentadas e seu respectivos histogramas no canal verde, destacado, em vermelho, o valor de referência nos padrões de cores verde, amarela e marrom. Nessa mesma Figura 46, foram mostrados os pontos chave da técnica SIFT para cada cor, com destaque ampliado, em cada caso, para melhor visualização.

Os algoritmos SIFT, Momentos de HU (Equações 5 a 11) e HOG (Equações 12 a 13) foram processados para todas as imagens do *dataset*, utilizando as bibliotecas *OpenCV* e *Skimage*, por meio da linguagem de programação *Python*, a partir dos parâmetros *default*, na maioria dos casos, das bibliotecas supracitadas. Cada processamento gerou um arquivo com as características para cada cor e seu armazenamento ocorreu no *bucket* da Oracle *Cloud*.

A Figura 47 exibiu parte de um dos processamentos realizados, de maneira que pudessem ser observadas as características geradas. Na primeira coluna foram relacionados os dados gerados pelo algoritmo HOG e, na última, os dados gerados pelo algoritmo dos Momentos de HU. As demais colunas, no total de 128, corresponderam às características geradas pelo algoritmo SIFT.

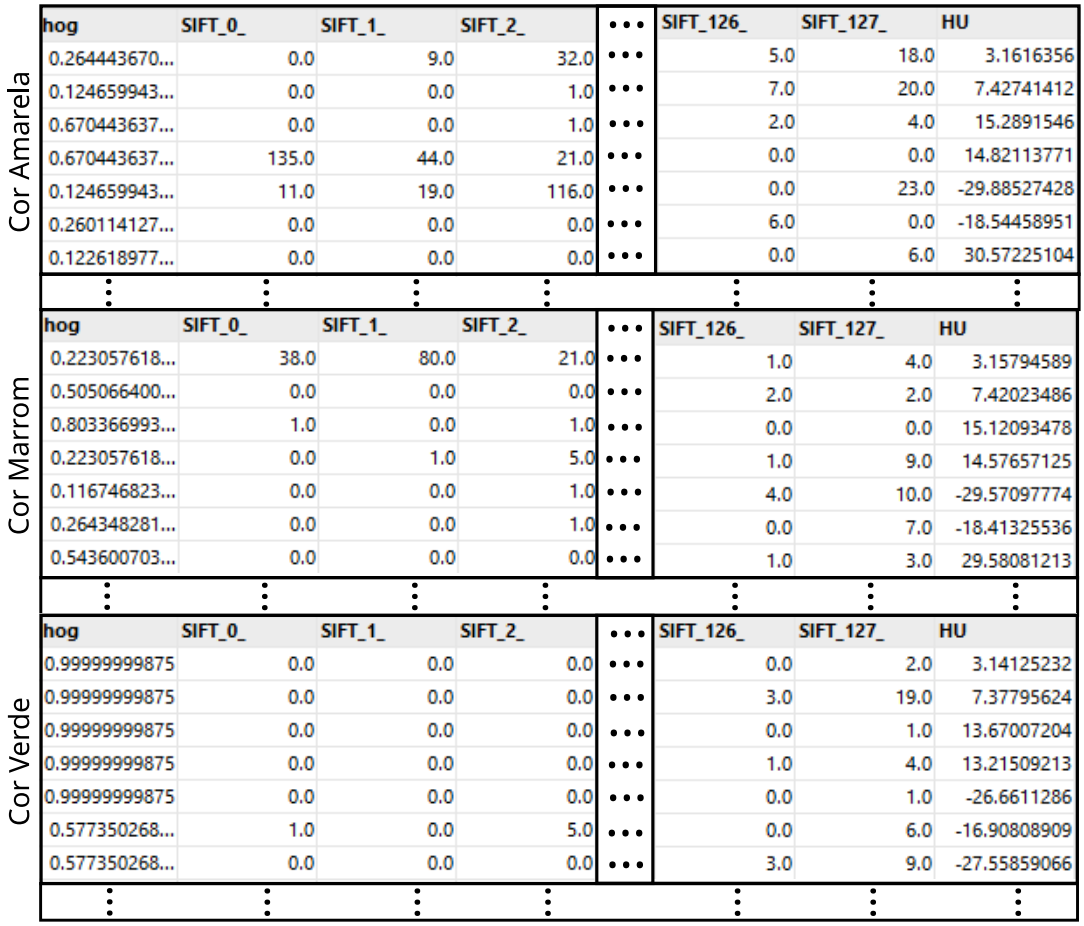
Para cada imagem processada foi considerado um conjunto de três subvetores de características obtidas com a aplicação das técnicas mencionadas, conforme ilustra a Figura 47. Entretanto, posteriormente estes subvetores foram integrados em um único vetor de características para, em seguida, receber a aplicação da técnica PCA para a redução de dimensionalidade.



Fonte: Próprio Autor

Figura 46 – Pontos-Chave por Padrão de Cor

Figura 47 – Descritores Hu, Hog e SIFT Não Normalizados



Cor Amarela

hog	SIFT_0_	SIFT_1_	SIFT_2_	•••	SIFT_126_	SIFT_127_	HU
0.264443670...	0.0	9.0	32.0	•••	5.0	18.0	3.1616356
0.124659943...	0.0	0.0	1.0	•••	7.0	20.0	7.42741412
0.670443637...	0.0	0.0	1.0	•••	2.0	4.0	15.2891546
0.670443637...	135.0	44.0	21.0	•••	0.0	0.0	14.82113771
0.124659943...	11.0	19.0	116.0	•••	0.0	23.0	-29.88527428
0.260114127...	0.0	0.0	0.0	•••	6.0	0.0	-18.54458951
0.122618977...	0.0	0.0	0.0	•••	0.0	6.0	30.57225104

Cor Marrom

hog	SIFT_0_	SIFT_1_	SIFT_2_	•••	SIFT_126_	SIFT_127_	HU
0.223057618...	38.0	80.0	21.0	•••	1.0	4.0	3.15794589
0.505066400...	0.0	0.0	0.0	•••	2.0	2.0	7.42023486
0.803366993...	1.0	0.0	1.0	•••	0.0	0.0	15.12093478
0.223057618...	0.0	1.0	5.0	•••	1.0	9.0	14.57657125
0.116746823...	0.0	0.0	1.0	•••	4.0	10.0	-29.57097774
0.264348281...	0.0	0.0	1.0	•••	0.0	7.0	-18.41325536
0.543600703...	0.0	0.0	0.0	•••	1.0	3.0	29.58081213

Cor Verde

hog	SIFT_0_	SIFT_1_	SIFT_2_	•••	SIFT_126_	SIFT_127_	HU
0.99999999875	0.0	0.0	0.0	•••	0.0	2.0	3.14125232
0.99999999875	0.0	0.0	0.0	•••	3.0	19.0	7.37795624
0.99999999875	0.0	0.0	0.0	•••	0.0	1.0	13.67007204
0.99999999875	0.0	0.0	0.0	•••	1.0	4.0	13.21509213
0.99999999875	0.0	0.0	0.0	•••	0.0	1.0	-26.6611286
0.577350268...	1.0	0.0	5.0	•••	0.0	6.0	-16.90808909
0.577350268...	0.0	0.0	0.0	•••	3.0	9.0	-27.55859066

Fonte: Próprio Autor

4.3 Redução de Dimensionalidade do Vetor de Características

Como técnica de redução de dimensionalidade, optou-se pela PCA, pois a mesma, após avaliações foi considerada adequada ao contexto trabalhado, em relação ao uso, por exemplo, da técnica *Linear Discriminant Analysis* - LDA, que demanda uso de rotulagem.

O vetor de características para o reconhecimento de padrões apresenta 130 posições. Entretanto, de forma a se buscar maior desempenho e melhor entendimento da representação destas características, a utilização da técnica PCA foi considerada.

Neste trabalho, dado os *datasets* utilizados para a validação, após a aplicação da técnica PCA, foi possível verificar a viabilidade de se trabalhar com uma redução de 130 para um total de 19 características.

A Tabela 13 apresentou resultados de um estudo elaborado que avaliou os dados obtidos da extração das características, em relação às possibilidades da redução de dimen-

sionalidade, de modo a ser estabelecido o entendimento das componentes principais no intervalo de 1 a 130, frente a cada componente principal (CP), o autovalor e sua variância estabelecidos.

Tabela 13 – Comparação Redução de Dimensionalidade

CP	Autovalor	% Variância	CP	Autovalor	% Variância	CP	Autovalor	% Variância
1	0,638181	12,298	45	0,0192137	0,37027	89	0,00558973	0,10772
2	0,571651	11,016	46	0,0189237	0,36468	90	0,00524686	0,10111
3	0,328703	6,3345	47	0,0181695	0,35015	91	0,00522982	0,10078
4	0,325943	6,2813	48	0,0169668	0,32697	92	0,00516094	0,099457
5	0,275776	5,3145	49	0,0166832	0,3215	93	0,00499258	0,096213
6	0,184056	3,547	50	0,0161889	0,31198	94	0,00480426	0,092583
7	0,175878	3,3894	51	0,0157513	0,30355	95	0,00470831	0,090734
8	0,144679	2,7881	52	0,0154428	0,2976	96	0,00467591	0,09011
9	0,13286	2,5604	53	0,0150127	0,28931	97	0,00448733	0,086476
10	0,125603	2,4205	54	0,0142839	0,27527	98	0,00430309	0,082925
11	0,122809	2,3667	55	0,013923	0,26831	99	0,00416152	0,080197
12	0,0967856	1,8652	56	0,013688	0,26378	100	0,00409866	0,078986
13	0,0949415	1,8296	57	0,0132774	0,25587	101	0,00397766	0,076654
14	0,0855482	1,6486	58	0,012608	0,24297	102	0,0039088	0,075327
15	0,081941	1,5791	59	0,0124476	0,23988	103	0,00380418	0,073311
16	0,0799518	1,5408	60	0,0122982	0,237	104	0,00348872	0,067232
17	0,0763869	1,4721	61	0,0120398	0,23202	105	0,00343486	0,066194
18	0,0678579	1,3077	62	0,0113486	0,2187	106	0,00331614	0,063906
19	0,0640759	1,2348	63	0,011029	0,21254	107	0,00318668	0,061411
20	0,0618595	1,1921	64	0,0109731	0,21146	108	0,00305105	0,058797
21	0,0573919	1,106	65	0,0107029	0,20626	109	0,00303647	0,058516
22	0,053187	1,025	66	0,0102752	0,19801	110	0,00289397	0,05577
23	0,0506325	0,97574	67	0,0099124	0,19102	111	0,00287129	0,055333
24	0,0478648	0,92241	68	0,00984752	0,18977	112	0,00275805	0,053151
25	0,0474375	0,91417	69	0,00948921	0,18287	113	0,00269183	0,051874
26	0,0466466	0,89893	70	0,00913085	0,17596	114	0,00259537	0,050016
27	0,0432379	0,83324	71	0,0087902	0,1694	115	0,0024602	0,047411
28	0,0400341	0,7715	72	0,00860113	0,16575	116	0,00244225	0,047065
29	0,0388116	0,74794	73	0,00819131	0,15786	117	0,0023265	0,044834
30	0,0373446	0,71967	74	0,00787662	0,15179	118	0,00226584	0,043665
31	0,0355076	0,68427	75	0,00784722	0,15122	119	0,00217171	0,041851
32	0,033911	0,6535	76	0,00764874	0,1474	120	0,00206117	0,039721
33	0,0323025	0,62251	77	0,00753469	0,1452	121	0,00197643	0,038088
34	0,0308543	0,5946	78	0,00741622	0,14292	122	0,00192024	0,037005
35	0,0286397	0,55192	79	0,00731357	0,14094	123	0,00184426	0,035541
36	0,0282012	0,54347	80	0,00697372	0,13439	124	0,00181284	0,034936
37	0,0274507	0,52901	81	0,00691314	0,13322	125	0,00174727	0,033672
38	0,0268133	0,51672	82	0,00661345	0,12745	126	0,00163573	0,031522
39	0,0247443	0,47685	83	0,00657032	0,12662	127	0,00153725	0,029625
40	0,0245677	0,47345	84	0,00625141	0,12047	128	0,00124301	0,023954
41	0,023151	0,44614	85	0,00612995	0,11813	129	0,00122559	0,023618
42	0,0230231	0,44368	86	0,00603637	0,11633	130	0,00025177	0,0048519
43	0,0218113	0,42033	87	0,00593295	0,11433			
44	0,0207261	0,39942	88	0,0056685	0,10924			

Fonte: Próprio Autor

A partir do *dataset* analisado e de dados de imagens processadas, a decisão sobre o número de componentes principais para a redução da dimensionalidade do vetor de características, baseou-se sobre o percentual de variância que possa descrever adequadamente o conjunto das variáveis, ou seja, considerando um valor $\geq 70\%$. Este procedimento é tomado como padrão, devendo ser considerado também quando da avaliação de outras bases de imagens, obtidas nas diferentes áreas de cultura. Observou-se que 19 componen-

tes principais alcançam 70,79% de variância, enquanto 18 componentes atingiram 69,56%, conforme ilustra a Tabela 13.

4.4 Reconhecimento de Padrões, Classificação e Aprendizado de Máquina

A escolha do classificador para uso no modelo de aprendizado de máquina do sistema foi pautada pelo ensaio de diferentes classificadores, ou seja, Árvore de Decisão (AD), *K-Nearest Neighbors* (KNN), *Naïve Bayes* (NB) e *Support Vector Machine* (SVM).

Nesse contexto, a partir da organização dos vetores de características com dimensões reduzidas para 19 componentes, via algoritmo PCA, para os diferentes conjuntos de imagens analisadas, os mesmos foram submetidos ao processamento de cada um dos classificadores mencionados, sendo, após análise dos resultados, selecionado o que apresentou melhor eficácia e eficiência, frente ao problema a ser tratado. Para tal avaliação, foram considerados conjuntos de dados de características das imagens, em proporções de treinamento-teste, respectivamente como: 50-50, 70-30 e 80-20.

Os resultados dos classificadores, processados com parametrização padrão, foram demonstrados nas Tabelas 15, 16 e 17. Diante dos resultados apresentados pelo processamento dos classificadores supracitados, na Figura 48, pôde ser observado o *ranking* sob a proporção 80-20. Nesse contexto, observou-se que os melhores resultados foram proporcionados pelo classificador SVM, onde a partir das três configurações de treino e teste, a proporção 80-20 apresentou os melhores resultados com valores de 0,718 para acurácia (Acc) e 0,282 como erro quadrático médio (MSE). Entretanto, foi observado que o valor da área sob a curva ROC (AUC) (0,755 unidades de área) foi menor do que aquele apresentado pelo uso da proporção 70-30 (0,760 unidades de área).

Em continuidade aos testes para o classificador SVM, foram realizados experimentos, tanto para a versão multiclasse, quanto para a versão binária, utilizando os parâmetros *default*. No primeiro momento, realizou-se os testes da versão multiclasse, considerando os dados de características das três classes de cores que potencialmente estão associadas ao problema da doença: verde, amarela e marrom. Tais classes representaram a fenomenologia do problema da FAS, onde cor verde significou a ausência da doença e, as cores amarela e marrom, a presença da ferrugem asiática. Na sequência, efetuou-se também os testes da versão binária do classificador, com os mesmos dados e configurações, porém associando as cores amarela e marrom, em somente uma classe, a fim de representar a presença da doença e, a cor verde, para a sua ausência.

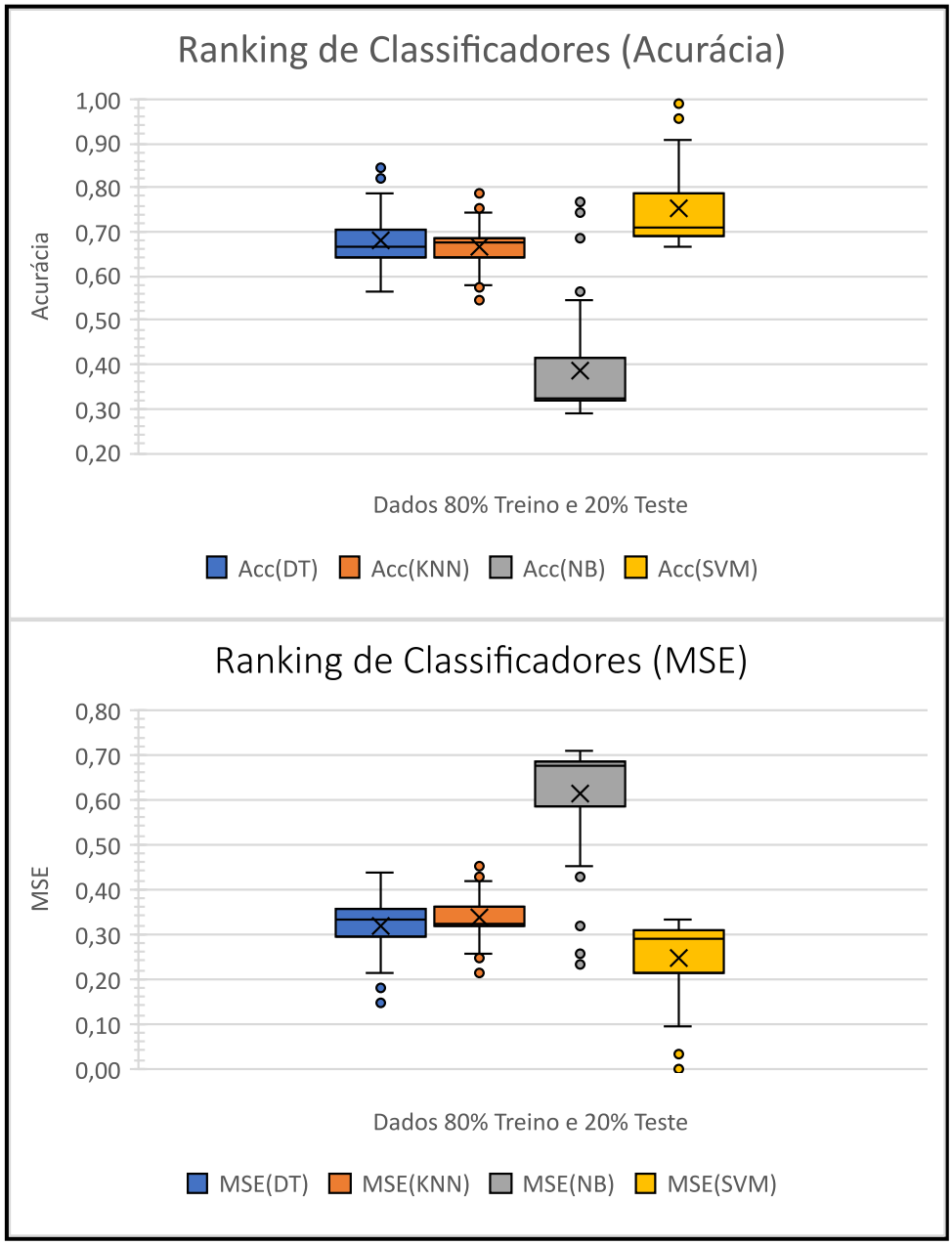
Os resultados dos testes supracitados mostraram que a versão binária foi mais eficiente se comparada à versão multiclasse do classificador SVM, quanto à acurácia, MSE e área sob a curva ROC (considerando o *kernel default*), de acordo com a Tabela 14.

Tabela 14 – Testes Classificador SVM Binário e Multiclasse

	50% Treino e 50% Teste			**70% Treino e 30% Teste**			**80% Treino e 20% Teste**		
	Acc	**MSE**	**AUC**	**Acc**	**MSE**	**AUC**	**Acc**	**MSE**	**AUC**
SVM Multiclasse	0,231	0,382	0,580	0,237	0,378	0,590	0,259	0,387	0,590
SVM Binário	0,708	0,292	0,705	0,709	0,291	0,760	0,718	0,282	0,755

Fonte: Próprio Autor

Figura 48 – Boxplot Comparativo dos Classificadores



Fonte: Próprio Autor

Tabela 15 – Classificadores - 50% Treino e 50% Teste

	Acc (DT)	MSE (DT)	Acc (KNN)	MSE (KNN)	Acc (NB)	MSE (NB)	Acc (SVM)	MSE (SVM)	AUC (SVM)	AUC (DT)	AUC (KNN)	AUC (NB)
Mínimo	0,472	0,164	0,503	0,297	0,267	0,282	0,667	0,005	0,560	0,410	0,180	0,330
Máximo	0,836	0,528	0,703	0,497	0,718	0,733	0,995	0,333	1,000	0,880	0,850	0,970
Soma	39,901	20,099	38,947	21,053	22,577	37,423	44,957	15,043	44,620	35,740	22,620	31,730
Média	0,665	0,335	0,649	0,351	0,376	0,624	0,749	0,251	0,744	0,596	0,377	0,529
Erro Padrão	0,009	0,009	0,006	0,006	0,015	0,015	0,011	0,011	0,016	0,017	0,022	0,023
Variância	0,005	0,005	0,002	0,002	0,013	0,013	0,008	0,008	0,016	0,017	0,030	0,033
Desvio Padrão	0,069	0,069	0,047	0,047	0,116	0,116	0,087	0,087	0,125	0,130	0,173	0,180
Mediana	0,667	0,333	0,669	0,331	0,318	0,682	0,708	0,292	0,705	0,540	0,320	0,465
25 porcentil	0,632	0,308	0,626	0,313	0,308	0,581	0,697	0,246	0,643	0,493	0,250	0,400
75 porcentil	0,692	0,368	0,687	0,374	0,419	0,692	0,754	0,303	0,825	0,718	0,448	0,658
Moda	0,667	0,333	0,687	0,313	0,308	0,692	0,697	0,303	NA	0,540	NA	0,430
Coef. Variação	10,336	20,519	7,214	13,345	30,731	18,540	11,677	34,900	16,817	21,898	45,910	34,121

Fonte: Próprio Autor

Tabela 16 – Classificadores - 70% Treino e 30% Teste

	Acc (DT)	MSE (DT)	Acc (KNN)	MSE (KNN)	Acc (NB)	MSE (NB)	Acc (SVM)	MSE (SVM)	AUC (SVM)	AUC (DT)	AUC (KNN)	AUC (NB)
Mínimo	0,564	0,145	0,547	0,214	0,291	0,231	0,667	0,000	0,570	0,370	0,180	0,240
Máximo	0,855	0,436	0,786	0,453	0,769	0,709	1,000	0,333	1,000	0,890	0,890	0,990
Soma	40,872	19,128	39,889	20,111	23,222	36,778	45,316	14,684	46,640	37,430	25,280	31,200
Média	0,681	0,319	0,665	0,335	0,387	0,613	0,755	0,245	0,777	0,624	0,421	0,520
Erro Padrão	0,008	0,008	0,005	0,005	0,017	0,017	0,012	0,012	0,016	0,017	0,023	0,027
Variância	0,004	0,004	0,002	0,002	0,017	0,017	0,009	0,009	0,015	0,017	0,032	0,045
Desvio Padrão	0,063	0,063	0,042	0,042	0,130	0,130	0,094	0,094	0,122	0,132	0,179	0,213
Mediana	0,667	0,333	0,675	0,325	0,325	0,675	0,709	0,291	0,760	0,600	0,390	0,430
25 porcentil	0,643	0,295	0,641	0,316	0,316	0,585	0,692	0,214	0,670	0,530	0,280	0,360
75 porcentil	0,705	0,357	0,684	0,359	0,415	0,684	0,786	0,308	0,878	0,720	0,540	0,693
Moda	0,667	0,333	0,684	0,316	0,316	0,684	0,692	0,308	1,000	NA	0,310	0,360
Coef. Variação	9,310	19,893	6,343	12,581	33,646	21,245	12,396	38,257	15,720	21,129	42,377	40,928

Fonte: Próprio Autor

Tabela 17 – Classificadores - 80% Treino e 20% Teste

	Acc (DT)	MSE (DT)	Acc (KNN)	MSE (KNN)	Acc (NB)	MSE (NB)	Acc (SVM)	MSE (SVM)	AUC (SVM)	AUC (DT)	AUC (KNN)	AUC (NB)
Mínimo	0,564	0,128	0,564	0,205	0,282	0,231	0,667	0,000	0,510	0,410	0,150	0,190
Máximo	0,872	0,436	0,795	0,436	0,769	0,718	1,000	0,333	1,000	0,910	0,890	0,980
Soma	41,885	18,115	40,744	19,256	22,590	37,410	45,513	14,487	46,290	38,590	27,380	30,610
Média	0,698	0,302	0,679	0,321	0,376	0,624	0,759	0,241	0,772	0,643	0,456	0,510
Erro Padrão	0,008	0,008	0,005	0,005	0,017	0,017	0,012	0,012	0,017	0,017	0,025	0,030
Variância	0,004	0,004	0,001	0,001	0,018	0,018	0,009	0,009	0,018	0,017	0,037	0,054
Desvio Padrão	0,065	0,065	0,038	0,038	0,134	0,134	0,092	0,092	0,134	0,131	0,193	0,233
Mediana	0,692	0,308	0,679	0,321	0,308	0,692	0,718	0,282	0,755	0,625	0,415	0,450
25 porcentil	0,654	0,269	0,667	0,308	0,308	0,590	0,696	0,212	0,670	0,533	0,303	0,320
75 porcentil	0,731	0,346	0,692	0,333	0,410	0,692	0,788	0,304	0,880	0,748	0,590	0,680
Moda	0,654	0,346	0,692	0,308	0,308	0,692	0,692	0,308	1,000	0,580	NA	NA
Coef. Variação	9,312	21,530	5,572	11,790	35,516	21,446	12,164	38,216	17,424	20,332	42,297	45,712

Fonte: Próprio Autor

Os hiperparâmetros para uso em cada *kernel* foram definidos pela técnica *Grid Search* via biblioteca *scikit-learn*, sob diferentes combinações, conforme retrata a Tabela 18.

Tabela 18 – Configurações de Hiperparâmetros - *Grid Search*

Configurações *Kernel* Linear
Parâmetros C: 1, 10, 100, Gama: 0,01; 0,1; 1, Peso: (0: 0,1 | 1: 0,9)
Configurações *Kernel* Polinomial
Grau: 3,5,7, **Parâmetros** C: 1, 10, 100, 1000, Gama: 0,001; 0,01; 0,1; 1, Peso: (balanceado, 0: 0,1 | 1: 0,9)
Configurações *Kernel* RBF
Grau: 3,5,7, **Parâmetros** C: 1, 10, 100, Gama: 0,001; 0,01; 0,1; 1, Peso: (0: 0,3 | 1: 0,7) (0: 0,1 | 1: 0,9)

Fonte: Próprio Autor

Para o exemplo considerado, a partir do *kernel* polinomial de terceira ordem que foi selecionado e de acordo com a análise da matriz de confusão (Figura 49) foi percebido conforme a diagonal principal da matriz que o modelo previu corretamente 346 casos que eram da classe "0", ou seja, não favorável à FAS. Em contrapartida, no quadrante superior direito, houve também 346 casos em que o modelo previu erroneamente que eram da classe "0", mas na realidade, eles eram da classe "1", ou seja, favorável à FAS. No segundo quadrante da diagonal principal da matriz foram registrados 1312 casos em que o modelo previu corretamente que eram da classe "1". Por outro lado, no quadrante inferior esquerdo, foram relatados 49 casos em que o modelo previu erroneamente que eram da classe "1" mas, na realidade, eles eram da classe "0". Frente a essa análise foi possível verificar que o classificador considerado operou adequadamente, entregando como resultado uma informação válida para compor o vetor das variáveis para o modelo de suporte à decisão.

Outro aspecto considerado, foi quanto à interpretação da curva ROC (Figura 49). Essa curva indicou que a resposta do classificador esteve distante da linha "laranja", que limita a região de operação do classificador para que o mesmo possa classificar adequadamente os dados, não operando de forma aleatória.

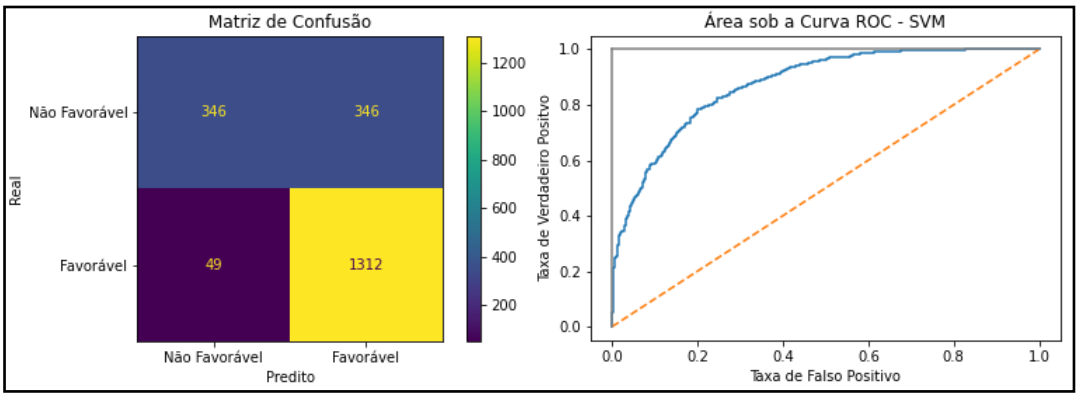
Também, como parte dos resultados foi observado que a eficácia do classificador com o *kernel* polinomial ficou na ordem de 0,87 (unidades de área) para a área sob a curva ROC; 0,81 para a acurácia e 0,19 para o MSE. Os demais resultados observados para o exemplo considerado, podem ser verificados na Tabela 19.

Tabela 19 – Dados do Relatório do Classificador SVM - *Kernel* Polinomial

	Precisão	Revocação	F1 score	Suporte
0	0,88	0,50	0,64	692
1	0,79	0,96	0,87	1361
Acurácia			0,81	2053
Média macro	0,83	0,73	0,75	2053
Média ponderada	0,82	0,81	0,79	2053

Fonte: Próprio Autor

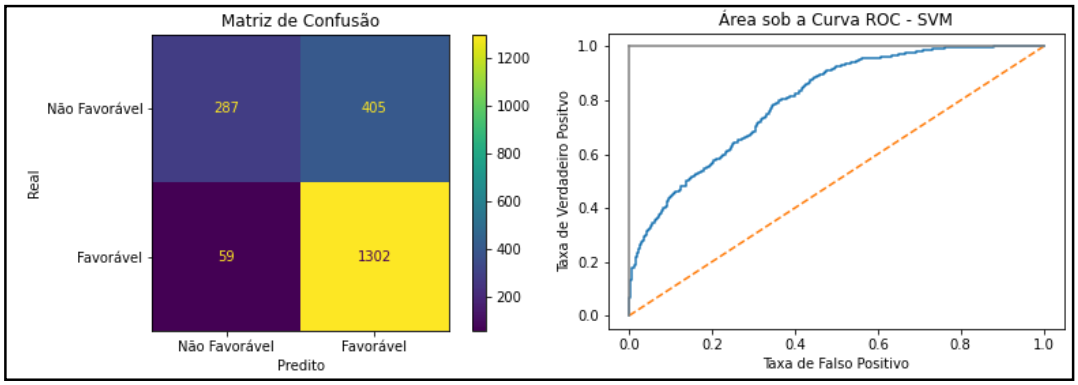
Figura 49 – Classificador SVM - Kernel Polinomial



Fonte: Próprio Autor

Como visto, as análises realizadas com o *kernel* RBF (Figura 50 e Tabela 20) e *kernel* linear (Figura 51 e Tabela 21) apresentaram resultados inferiores em relação ao uso do *kernel* polinomial de ordem três, o qual foi selecionado para integrar o sistema de visão e inteligência computacional em ambiente de nuvem para gestão de risco da ferrugem asiática na cultura da soja.

Figura 50 – Classificador SVM - *Kernel* RBF



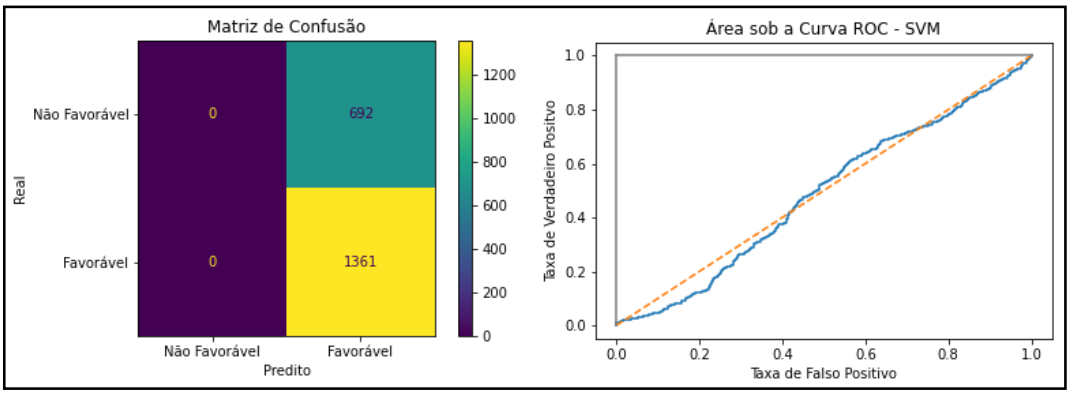
Fonte: Próprio Autor

Tabela 20 – Dados do Relatório do Classificador SVM - *Kernel* RBF

	Precisão	Revocação	F1 score	Suporte
0	0,83	0,41	0,55	692
1	0,76	0,96	0,85	1361
Acurácia			0,77	2053
Média macro	0,80	0,69	0,70	2053
Média ponderada	0,79	0,77	0,75	2053

Fonte: Próprio Autor

Figura 51 – Classificador SVM - *Kernel* Linear



Fonte: Próprio Autor

Tabela 21 – Dados do Relatório do Classificador SVM - *Kernel* Linear

	Precisão	Revocação	F1 score	Suporte
0	0,00	0,00	0,00	692
1	0,66	1,00	0,80	1361
Acurácia			0,66	2053
Média macro	0,33	0,50	0,40	2053
Média ponderada	0,44	0,66	0,53	2053

Fonte: Próprio Autor

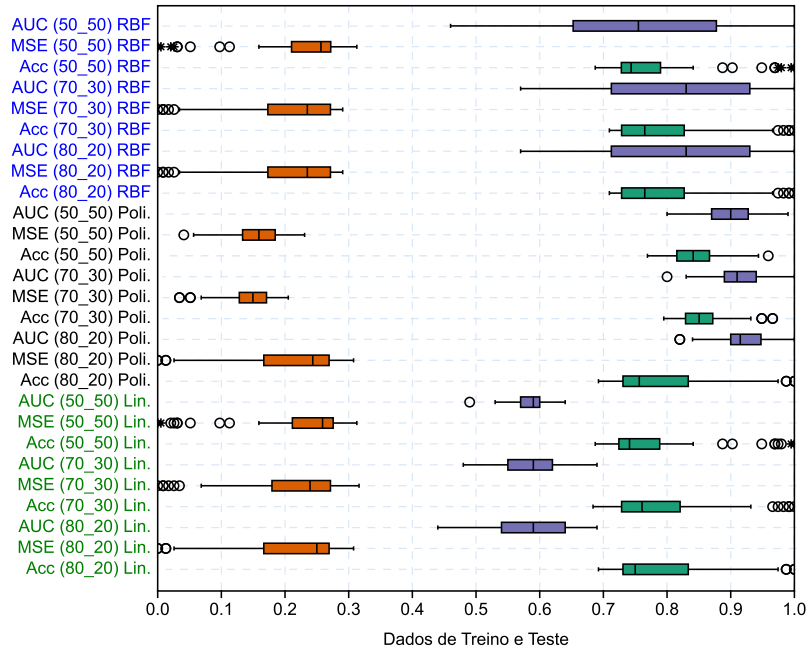
Uma vez selecionada a modalidade do classificador, para o sistema de visão e inteligência computacional em ambiente de nuvem para gestão de risco da ferrugem asiática na cultura da soja, importante se fez considerar uma abordagem sobre a escolha do *kernel* polinomial. Esse *kernel* se destacou em relação às avaliações realizadas com os *kernels* linear e RBF, uma vez que apresentou eficácia operacional conforme resultado da análise presente na Figura 52. O detalhamento numérico dessa análise pode ser visualizado na Tabela 23, onde vale destacar também, conforme apresentado na Tabela 22, os hiperparâmetros definidos para essa configuração. Nesse contexto, a melhor escolha considerada foi definida tomando em conta o classificador SVM binário com *kernel* polinomial de terceiro grau.

Tabela 22 – Hiperparâmetros Selecionados - *kernel* Polinomial

C	1
Peso (Classe 0)	0,3
Peso (Classe 1)	0,7
Grau	3
Gamma	1

Fonte: Próprio Autor

Figura 52 – Classificador SVM Kernel: Linear, Polinomial e RBF



Fonte: Próprio Autor

Tabela 23 – Estatística Descritiva - *kernel* Liner, Polinomial e RBF

Classificador SVM Kernel Linear									
Estatística Descritiva	**Configuração 80-20**			**Configuração 70-30**			**Configuração 50-50**		
	Acc	**MSE**	**AUC**	**Acc**	**MSE**	**AUC**	**Acc**	**MSE**	**AUC**
Mínimo	0,692	0,000	0,440	0,684	0,000	0,480	0,687	0,000	0,490
Máximo	1,000	0,308	0,690	1,000	0,316	0,690	1,000	0,313	0,640
Média	0,787	0,213	0,587	0,792	0,208	0,588	0,777	0,223	0,583
Erro Padrão	0,011	0,011	0,008	0,011	0,011	0,006	0,011	0,011	0,003
Variância	0,008	0,008	0,004	0,007	0,007	0,002	0,007	0,007	0,001
Desvio Padrão	0,089	0,089	0,062	0,085	0,085	0,050	0,085	0,085	0,025
Mediana	0,750	0,250	0,590	0,761	0,239	0,590	0,741	0,259	0,590
25 percentil	0,731	0,167	0,540	0,729	0,179	0,550	0,724	0,212	0,570
75 percentil	0,833	0,269	0,640	0,821	0,271	0,620	0,788	0,276	0,600
Classificador SVM Kernel Polinomial									
Estatística Descritiva	**Configuração 80-20**			**Configuração 70-30**			**Configuração 50-50**		
	Acc	**MSE**	**AUC**	**Acc**	**MSE**	**AUC**	**Acc**	**MSE**	**AUC**
Mínimo	0,692	0,000	0,820	0,795	0,034	0,800	0,769	0,041	0,800
Máximo	1,000	0,308	1,000	0,966	0,205	1,000	0,959	0,231	0,990
Média	0,790	0,210	0,917	0,860	0,140	0,916	0,844	0,156	0,900
Erro Padrão	0,011	0,011	0,006	0,005	0,005	0,005	0,005	0,005	0,005
Variância	0,008	0,008	0,002	0,002	0,002	0,001	0,001	0,001	0,001
Desvio Padrão	0,088	0,088	0,043	0,042	0,042	0,039	0,036	0,036	0,039
Mediana	0,756	0,244	0,915	**0,850**	**0,150**	**0,910**	0,841	0,159	0,900
25 percentil	0,731	0,167	0,900	0,829	0,128	0,890	0,815	0,133	0,870
75 percentil	0,833	0,269	0,948	0,872	0,171	0,940	0,867	0,185	0,928
Classificador SVM Kernel RBF									
Estatística Descritiva	**Configuração 80-20**			**Configuração 70-30**			**Configuração 50-50**		
	Acc	**MSE**	**AUC**	**Acc**	**MSE**	**AUC**	**Acc**	**MSE**	**AUC**
Mínimo	0,709	0,000	0,570	0,709	0,000	0,570	0,687	0,000	0,460
Máximo	1,000	0,291	1,000	1,000	0,291	1,000	1,000	0,313	1,000
Média	0,794	0,206	0,820	0,794	0,206	0,820	0,779	0,221	0,769
Erro Padrão	0,011	0,011	0,015	0,011	0,011	0,015	0,011	0,011	0,018
Variância	0,007	0,007	0,014	0,007	0,007	0,014	0,007	0,007	0,020
Desvio Padrão	0,084	0,084	0,119	0,084	0,084	0,119	0,084	0,084	0,143
Mediana	0,765	0,235	0,830	0,765	0,235	0,830	0,744	0,256	0,755
25 percentil	0,729	0,173	0,713	0,729	0,173	0,713	0,728	0,210	0,653
75 percentil	0,827	0,271	0,930	0,827	0,271	0,930	0,790	0,272	0,878

Fonte: Próprio Autor

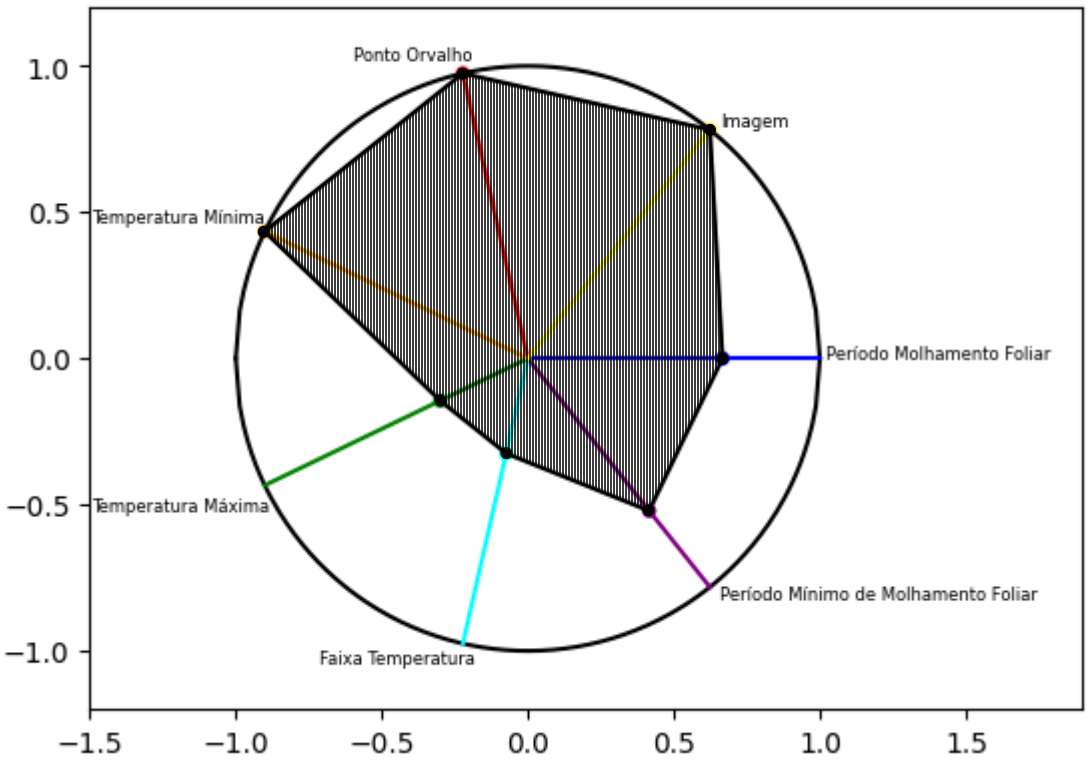
4.5 Fusão de Variáveis e Auxílio à Tomada de Decisão

O modelo de fusão de variáveis foi utilizado para verificar a presença ou não da ferrugem asiática em área de cultura e para o estabelecimento do estágio de favorabilidade. Nesse contexto, foram avaliadas três diferentes abordagens, as quais proporcionaram os seguintes resultados, conforme exemplo apresentado, a partir de todo o conjunto de dados do *dataset* das séries temporais consideradas. O exemplo considerado, para fins de elucidação, tomou em conta o processamento de dados compreendido ao período de 30/10/2017 a 08/11/2017.

Para a abordagem que considerou o método da figura de mérito, a mesma foi implementada por meio do Algoritmo 10 e pelas Equações 57 e 58. Nesse contexto, a figura de mérito final, foi originada pela união dos pontos relacionados aos valores das variáveis normalizadas (Figura 53).

A escala de referência utilizada para o cálculo da área da figura de mérito, diante da favorabilidade, foi definida na forma: "baixa", quando o valor encontrado ficou entre 0 e 0,90 [μA]; "média", quando o valor encontrado ficou entre 0,91 e 1,80 [μA]; e "alta", quando o valor encontrado ficou maior que 1,81 [μA]. Para o exemplo de resultado considerado, houve a confirmação da presença da ferrugem asiática e devido ao valor de 1,50 [μA] a caracterização como favorabilidade ou estágio de severidade "média".

Figura 53 – Resultado Figura de Mérito



Fonte: Próprio Autor

Para a abordagem que utilizou lógica difusa, foi considerado o uso de uma base de regras semânticas de inferência (Equação 66), construída por meio do conhecimento prévio sobre o comportamento das variáveis climáticas e classificação de padrões das imagens digitais das folhas de soja. Essas regras foram elaboradas por operações de união (Equação 61) e intersecção (Equação 60) entre conjuntos, o que descreveu o mapeamento completo (Equação 67) entre as entradas (antecedentes) e as saídas (consequentes).

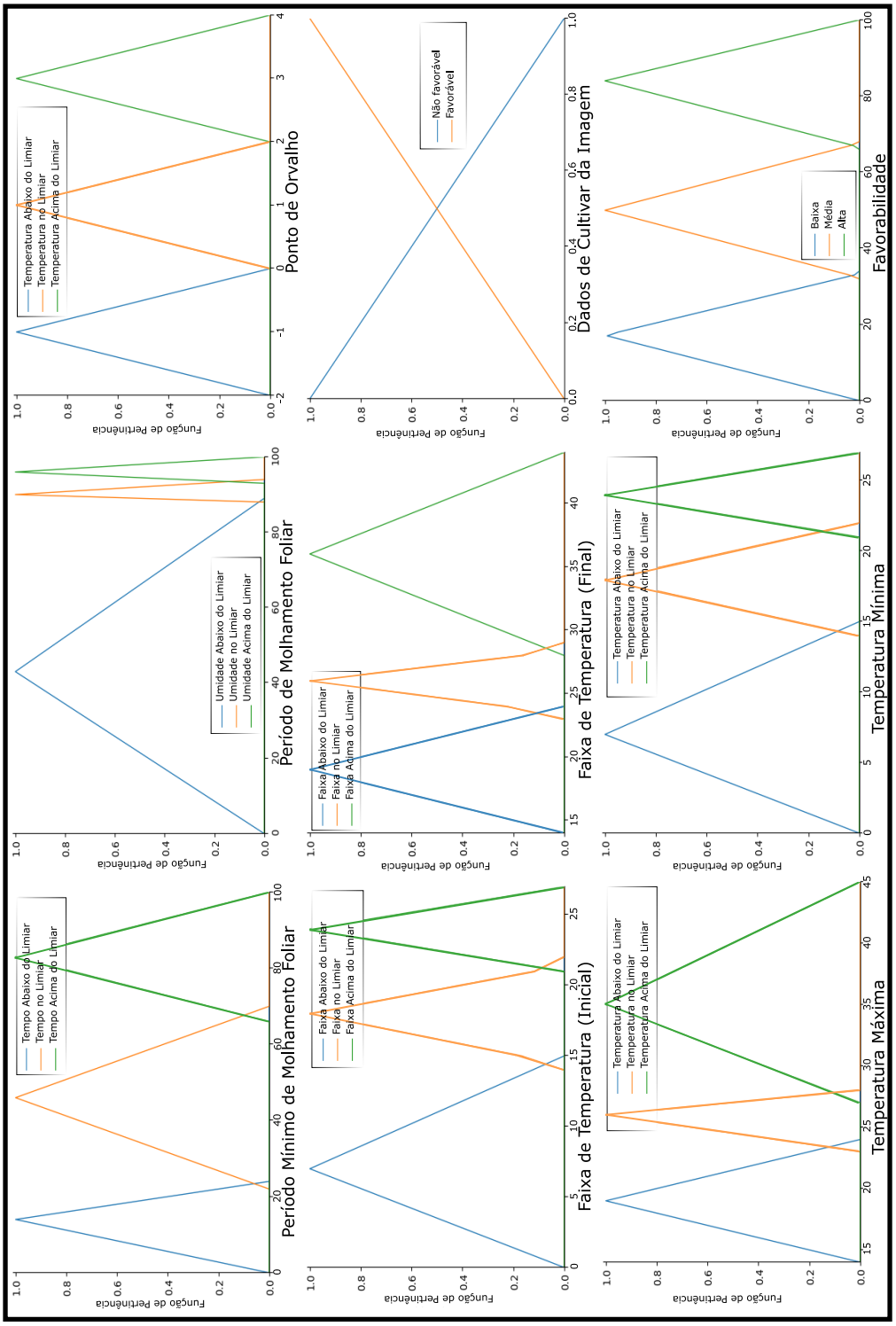
As configurações das variáveis difusas (Tabela 24) e as funções de pertinência correspondentes às 7 variáveis se encontram representadas nos gráficos da Figura 54. Foi possível destacar sobre esses resultados que para cada intervalo das funções de pertinência ocorreu um erro associado, da ordem de 5%, para mais e para menos, referente a cada faixa de transição entre as favorabilidades baixa, média e alta.

Tabela 24 – Configuração das Funções de Pertinência

Descrição	**Configuração**
Antecedente: Período Molhamento Foliar:	
Umidade abaixo do limiar	0, 43, 89
Umidade no limiar	88, 90, 94
Umidade acima do limiar	93, 96, 100
Antecedente: Período Mínimo de Molhamento Foliar:	
Tempo abaixo do limiar	0, 14, 24
Tempo no limiar	22, 46, 70
Tempo acima do limiar	66, 83, 100
Antecedente: Dados de Classificação da Imagem da Folha Soja:	
Não favorável	0, 0, 1
Favorável	1, 1, 1
Antecedente: Ponto de Orvalho:	
Temperatura abaixo do limiar	-2, -1, 0
Temperatura no limiar	0, 1, 2
Temperatura acima do limiar	2, 3, 4
Antecedente: Faixa Temperatura:	
Inicial: faixa abaixo do limiar	0, 7, 15
Inicial: faixa no limiar	14.4, 18, 21.4
Inicial: faixa acima do limiar	21, 24, 27
Final: faixa abaixo do limiar	14, 19, 24
Final: faixa no limiar	23.4, 26, 28.4
Final: faixa acima do limiar	28, 36, 44
Antecedente: Temperatura Mínima:	
Temperatura miníma abaixo do limiar	0, 7, 15
Temperatura miníma no limiar	14, 18, 22
Temperatura miníma acima do limiar	21, 24, 27
Antecedente: Temperatura Máxima:	
Temperatura máxima abaixo do limiar	14, 19, 24
Temperatura máxima no limiar	23, 26, 28
Temperatura máxima acima do limiar	27, 35, 43
Consequente: Favorabilidade:	
Baixa	0, 17.15, 33.3
Média	32.3, 50, 67.6
Alta	66.6, 84, 100

Fonte: Próprio Autor

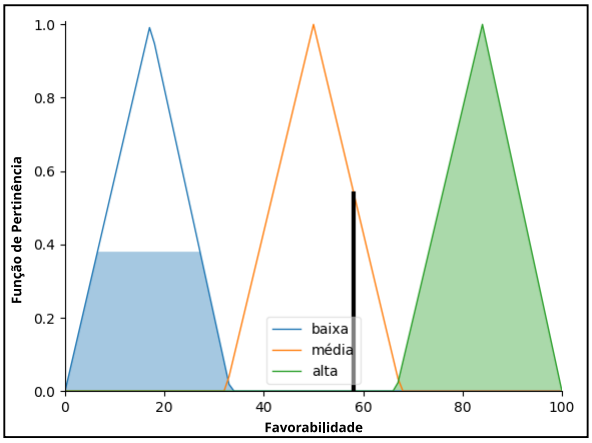
Figura 54 – Funções de Pertinência



Fonte: Próprio Autor

Na Figura 55 é possível observar um exemplo de saída em formato gráfico, entre os demais computados, diante da defuzzificação que registrou o valor numérico resultante, marcado por uma linha vertical na cor preta no eixo x, com valor de favorabilidade de aproximadamente 17%. Tal resultado, indicou favorabilidade baixa por se encontrar entre 0 e 33%.

Figura 55 – Lógica Difusa: Saída de Defuzzificação



Fonte: Próprio Autor

Por outro lado, para a aplicação de método baseado em x-Markov, foi considerado o udo de cadeias ocultas de Markov, com modelo adaptativo ao processo que trata sobre a favorabilidade da FAS. Assim, foi considerado um modelo que se encontra ilustrado na Figura 56. Suas características envolveram:

1. Estados de favorabilidade de ocorrência da doença da FAS caracterizados pelas favorabilidades baixa, média e alta;

2. Variáveis climáticas associadas a $R1$ - Período de Molhamento Foliar; $R2$ - Período Mínimo de Molhamento Foliar; $R3$ - Faixa de Temperatura; $R4$ - Temperatura Máxima; $R5$ - Temperatura Mínima; $R6$ - Ponto de Orvalho e $R7$ - Dados de Classificação de Imagens de Folhas da Soja;

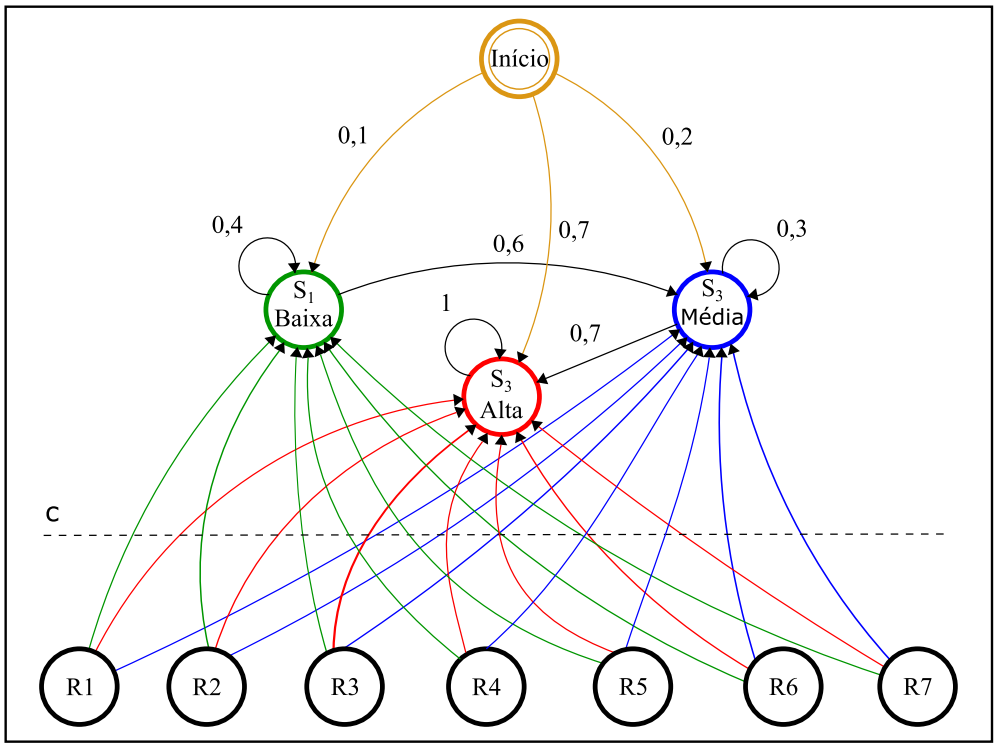
3. Probabilidade inicial, onde foram utilizados os valores de 0,1 para favorabilidade baixa; 0,3 para favorabilidade média e 0,7 para favorabilidade alta;

4. Combinações "C" que denotaram 2^7 possibilidades geradas pelas variáveis de R1 a R7, o que totalizou 128 possibilidades ou combinações. Os valores de preenchimento das cadeias ocultas de Markov, traduzidos em observações, foram definidos pelas combinações das sete variáveis e suas respectivas probabilidades, frente ao período de cada janela temporal de dados climáticos para predição da doença;

5. Probabilidades de transição de mudanças nos estados da favorabilidade da doença (observadas de acordo com cada variável identificada no processo, por meio das porcentagens indicadas em cada observação);

6. Probabilidades de emissão geradas pelas transições dos estados para as observações, em conformidade com a cadeia oculta de Markov. As combinações foram selecionadas pelo processo de coleta dos dados, via janela temporal, orientada pela base de regras de favorabilidade da FAS, em seus diferentes estágios, na forma: (1) transição para o estado de favorabilidade "Baixa": quando identificado que o conjunto das variáveis correspondeu à faixa de 0 a 33%, de acordo com as observações; (2) transição para o estado de favorabilidade "Média": quando as variáveis identificadas estiveram na faixa de 34 a 66%; ou (3) transição para o estado de favorabilidade "Alta": quando as variáveis identificadas se apresentaram com valores maiores que 66%. A cadeia oculta de Markov foi representada, parcialmente, para entendimento de acordo com a Tabela 25.

Figura 56 – Modelo de Cadeias Ocultas de Markov para Qualificação da Favorabilidade de Ocorrência da FAS



Fonte: Próprio Autor

Para a operação do modelo Markoviano utilizado, foi considerado o percentual de 10% para a primeira qualificação sobre a Favorabilidade Baixa (S_1), bem como 20% para Favorabilidade Média (S_2) e 70% para Favorabilidade Alta (S_3). No entanto, para

o Estado S_1, a probabilidade da doença foi definida em 40% de chances para permanecer em S_1 e 60% de chances para evoluir para o Estado S_2 de Favorabilidade Média. Todavia, para o Estado S_2, definiu-se a probabilidade de 30% para permanecer no Estado S_2 e a probabilidade de 70% para evoluir para o Estado S_3. Assim, quando o Estado S_3 foi atingido, a probabilidade de permanecer neste Estado foi de 100%, ou seja, não foi possível retornar para os estados S_1 e S_2.

Tabela 25 – Dados Cadeia Oculta de Markov

C	R1	R2	R3	R4	R5	R6	R7	S	P_R1	P_R2	P_R3	P_R4	P_R5	P_R6	P_R7
1	0	0	0	0	0	0	0	1	0,00	0,00	0,00	0,00	0,00	0,00	0,00
2	0	0	0	0	0	0	1	1	0,00	0,00	0,00	0,00	0,00	0,00	1,00
3	0	0	0	0	0	1	0	1	0,00	0,00	0,00	0,00	0,00	1,00	0,00
4	0	0	0	0	0	1	1	1	0,00	0,00	0,00	0,00	0,00	0,50	0,50
5	0	0	0	0	1	0	0	1	0,00	0,00	0,00	0,00	1,00	0,00	0,00
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9	0	0	0	1	0	0	0	1	0,00	0,00	0,00	1,00	0,00	0,00	0,00
10	0	0	0	1	0	0	1	1	0,00	0,00	0,00	0,50	0,00	0,00	0,50
11	0	0	0	1	0	1	0	1	0,00	0,00	0,00	0,50	0,00	0,50	0,00
12	0	0	0	1	0	1	1	2	0,00	0,00	0,00	0,33	0,00	0,33	0,33
13	0	0	0	1	1	0	0	1	0,00	0,00	0,00	0,50	0,50	0,00	0,00
14	0	0	0	1	1	0	1	2	0,00	0,00	0,00	0,33	0,33	0,00	0,33
15	0	0	0	1	1	1	0	2	0,00	0,00	0,00	0,33	0,33	0,33	0,00
16	0	0	0	1	1	1	1	2	0,00	0,00	0,00	0,25	0,25	0,25	0,25
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
128	1	1	1	1	1	1	1	3	0,14	0,14	0,14	0,14	0,14	0,14	0,14

Fonte: Próprio Autor

Parte do detalhamento da cadeia oculta de Markov pôde ser observado na Tabela 25, onde se encontram apresentados exemplos das 128 combinações possíveis para as 7 variáveis de observação e suas probabilidades correspondentes.

Há também, na Tabela 25, a representação dos Estados S_1 "Favorabilidade Baixa", S_2 "Favorabilidade Média" e S_3 "Favorabilidade Alta", (Figura 56), por meio da coluna S. Quando o valor foi igual a ”1”, em S, significou que as combinações estiveram na faixa de 0 até 33,3% e a quantidade de variáveis envolvidas foi de 0 até 2. Por outro lado, se o valor foi igual a ”2”, em S, significou que as combinações estiveram na faixa de $33,4$ até $66,6$% e a quantidade de variáveis envolvidas foi de 3 até 4. Finalmente, quando o valor, em S, foi igual a ”3”, significou que as combinações estiveram na faixa de $66,7$ até 100% e que a quantidade de variáveis evolvidas foi de 5 até 6. Para todos os casos considerados, para se identificar as variáveis envolvidas nas combinações, sejam elas de valores S iguais a "1", "2" ou ainda "3", bastou a verificação sobre aquelas que se encontraram com valor igual a "1", respectivamente, para as colunas de R1 a R7.

As probabilidades foram calculadas para cada combinação, de acordo com a seguinte regra: quando as variáveis de R1 a R7 apresentaram o valor igual a ”1”, ou seja, envolvidas na combinação, dividiu-se o valor total das probabilidades, 100% na escala de 0 a 1, pela

soma da quantidade de variáveis envolvidas. O resultado da probabilidade foi atribuído, igualmente, entre as variáveis participantes.

Para fins de entendimento, dada a combinação (16) da Tabela 25, tem-se os valores das variáveis de $R1$ a $R7$: "0; 0; 0; 1; 1; 1; 1". Desta forma, o cálculo foi a divisão do valor total das probabilidades, ou seja, "1" pelas 4 variáveis envolvidas. Assim, tem-se: $1/4 = 0,25$, de forma que o resultado "0, 25" foi atribuído, como probabilidade, para cada variável: "0; 0; 0; 0, 25; 0, 25; 0, 25; 0, 25".

Pôde ser observado também, na Tabela 25, que a soma das probabilidades, para cada combinação, foram iguais a "1". Quando ocorreram as dízimas periódicas, os valores de probabilidades foram aproximados.

Os dados de entrada para o algoritmo da abordagem das cadeias ocultas de Markov foram originados do vetor de ocorrências. Os dados que compuseram o vetor de ocorrências foram computados nas regras de 1 a 7, cujas variáveis participantes foram retornadas dos algoritmos de 3 a 9 a saber: (1) dados_Regra_1; (2) dados_Regra_2; (3) dados_Regra_3; (4) valor_Regra_4; (5) valor_Regra_5; (6) dados_Regra_6; e (7) valor_Regra_7.

Adicionalmente, a computação dos dados para alimentar o vetor de ocorrências foi realizada pela contabilização dos registros obtidos em razão da consulta à base de dados climáticos, de acordo com a composição de cada regra para cada variável, observando a janela da série temporal.

As faixas de porcentagens indicadas no algoritmo foram representadas pelos Estados S_1, S_2 e S_3 que, na Tabela 25, estiveram representadas na coluna S com os valores 1, 2 e 3, respectivamente.

A Figura 57 ilustra o gráfico, onde é possível observar que a contabilização das ocorrências para as variáveis, de acordo com a base de regras para a favorabilidade da FAS. Portanto, entendeu-se nesse gráfico que, das 7 variáveis envolvidas, a variável período de molhamento foliar registrou 7 ocorrências de favorabilidade para a FAS, a variável período mínimo de molhamento foliar registrou 6 ocorrências e as demais registraram 1 ocorrência, com exceção da faixa de temperatura que não registrou nenhuma ocorrência.

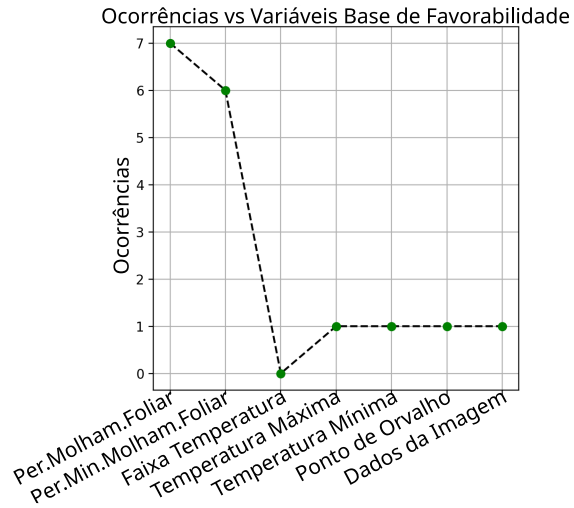
Em contrapartida, a Tabela 26 ilustra a composição do vetor de dados de entrada (Algoritmo 12), onde na primeira linha é apresentado o vetor de ocorrências e, na segunda linha, o vetor de ocorrências transformadas. Adicionalmente, a transformação do número de ocorrências versus o valor das variáveis "R" pôde ser estabelecido, considerando o número de ocorrências de $R \geq 1$ para $R = 1$, ou caso contrário, onde o número de ocorrências será igual a "0", ou seja, $R = 0$.

Tabela 26 – Vetor Entrada de Dados

Entrada (Algoritmo)	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**
Ocorrências:	7	6	0	1	1	1	1
Ocorrências Transformadas:	1	1	0	1	1	1	1

Fonte: Próprio Autor

Figura 57 – Gráfico de Regras Contabilizadas



Fonte: Próprio Autor

Para o exemplo ilustrativo considerado, a partir do gráfico de regras contabilizadas, apresentado na Figura 57, e do vetor de dados, conforme a Tabela 26, as ocorrências transformadas indicaram que 6 variáveis apresentaram valores iguais a "1", ou seja, a aplicação de cadeias ocultas de Markov indicou alta favorabilidade de ocorrência da FAS.

Diante das análises realizadas com os três diferentes modelos de fusão de dados, o uso de cadeia oculta de Markov foi o que apresentou o melhor resultado.

Como critério de avaliação, foram criados dois cenários para análise dos algoritmos na fusão de dados. Em cada cenário, avaliou-se a favorabilidade em suas categorias baixa, média e alta, tanto pela quantidade de pontos por acertos, quanto pela porcentagem desses acertos em cada abordagem avaliada.

Para a execução dos cenários de avaliação, foi elaborada uma base de dados com 128 combinações possíveis de favorabilidade (baixa, média e alta), a partir de dados das variáveis de R1 a R7, como entrada, e as saídas correspondentes, conhecidas à priori.

No contexto da base de testes, foi utilizado o seguinte conceito estabelecido de identificação das favorabilidades, frente à base de regras: favorabilidade baixa de 0 a 33,4%, representadas por 0 até 2 variáveis com valor igual a "1"; favorabilidade média de 33,4 a 66,6%, representadas por 3 até 4 variáveis com valor igual a "1" e para favorabilidade alta de 66,6 a 100%, representadas por 5 até 7 variáveis com valor igual a "1".

Após a contabilização das 128 combinações da base de dados de testes, organizadas utilizando os *datasets* de dados e imagens disponíveis, percebeu-se que as categorias de favorabilidade apresentaram-se desbalanceadas, oferecendo a seguinte situação: para as favorabilidades baixa e alta, identificou-se uma população de 29 combinações para cada categoria e, para a categoria da favorabilidade média, a população identificada foi de 70 combinações. Nesse sentido, organizou-se para o cenário 1 do teste uma população igual

a 29 combinações, para todas categorias de favorabilidade e, no cenário 2, as demais 41 combinações da categoria de favorabilidade média, não utilizadas no cenário 1.

A partir da base de dados construída, submeteu-se a mesma aos algoritmos das três abordagens. O resultado de cada abordagem foi comparado ao resultado já conhecido, a priori, para cada uma das combinações construídas. Dessa forma, para cada acerto foi atribuída uma nota com valor de 1 ponto e, ao final do processamento dos testes, foi apresentado o total da somatória de pontos, assim como a porcentagem de acertos para cada abordagem considerada. No entanto, a abordagem escolhida para aplicação na etapa de fusão de dados foi aquela que apresentou o maior número de pontos e a maior porcentagem de acertos.

Na Tabela 27 foi apresentada a comparação e os resultados dos testes nos cenários 1 e 2 para avaliação das abordagens Figura de Mérito, Lógica Difusa e Cadeia Oculta de Markov.

Tabela 27 – Testes de Comparação das Abordagens

Análise de Favorabilidade - Avaliação						
Figura de Mérito (Categoria)	**Cenário 1**			**Cenário 2**		
	População	**Pontos (Acertos)**	**% (Acertos)**	**População**	**Pontos (Acertos)**	**% (Acertos)**
Favorabilidade Baixa	29	29	100,00	0	———	———
Favorabilidade Média	29	14	48,28	41	29	70,74
Favorabilidade Alta	29	11	37,94	0	———	———
Lógica Difusa (Categoria)	**População**	**Pontos (Acertos)**	**% (Acertos)**	**População**	**Pontos (Acertos)**	**% (Acertos)**
Favorabilidade Baixa	29	8	27,59	0	———	———
Favorabilidade Média	29	12	41,38	41	25	60,98
Favorabilidade Alta	29	18	62,07	0	———	———
Cadeias Ocultas de Markov (Categoria)	**População**	**Pontos (Acertos)**	**% (Acertos)**	**População**	**Pontos (Acertos)**	**% (Acertos)**
Favorabilidade Baixa	29	29	100,00	0	———	———
Favorabilidade Média	29	29	100,00	41	41	100
Favorabilidade Alta	29	29	100,00	0	———	———

Fonte: Próprio Autor

A análise executada, de acordo com os resultados dos cenários 1 e 2, frente às abordagens para a fusão de dados apresentadas, indicaram que, para a figura de mérito, obteve 100% de acertos para a favorabilidade baixa, 48,28% para a favorabilidade média e 37,94% para a favorabilidade alta, para o cenário 1. Logo, para o cenário 2, foram obtidas 70,74% de acertos para a favorabilidade média.

Em contrapartida, a lógica difusa apresentou os resultados de 27,59% de acertos para a favorabilidade baixa, 41,38% para a favorabilidade média e 62,07% para a favorabilidade alta, no cenário 1. Porém, para o cenário 2, o resultado foi igual a 60,98% para a favorabilidade média.

No entanto, para as cadeias ocultas de Markov, foram obtidos os melhores resultados,

tanto no cenário 1 quanto no cenário 2, para as três categorias de favorabilidade com 100% de acertos. Por essa razão, a abordagem das cadeias ocultas de Markov foi escolhida para o processamento da fusão de dados.

Após a escolha da abordagem e também a partir do vetor de entrada, supracitado na Tabela 26, foi analisado o fluxo de processamento submetido à cadeia oculta de Markov, cujo entendimento do vetor de resultado foi exibido na Tabela 28: (1) a cadeia oculta de Markov; (2) as probabilidades selecionadas correspondentes ao cenário da base de regras de favorabilidade da FAS; (3) o estado traduzido por "S"; e (4) o resultado da ocorrência da favorabilidade na FAS.

Tabela 28 – Resultado Cadeia de Markov

Cadeia Oculta Selecionada:	1	1	0	1	1	1	1
Probabilidade Selecionada:	0,17	0,17	0,0	0,17	0,17	0,17	0,17
Estado (S):	3						
Favorabilidade:	Alta						

Fonte: Próprio Autor

O primeiro item da Tabela 28 mostra o arranjo da cadeia de Markov que é utilizado para a definição da favorabilidade de ocorrência, quando compatível com o estado das variáveis de entradas consideradas entre as 128 combinações possíveis, comparada ao vetor de dados transformado. O segundo item, da mesma tabela, mostra o resultado das probabilidades, associado à cadeia oculta de Markov customizada para a aplicação e selecionada em função do arranjo das sete variáveis. O terceiro item, da mesma tabela, apresenta o resultado do estado de favorabilidade, em decorrência dos itens 1 e 2 acima mencionados.

Também, ainda no exemplo considerado, tomando em conta a Tabela 28, a linha correspondente à cadeia oculta de Markov ora selecionada foi possível observar somente um valor igual a "0". Isso indica que apenas a faixa de temperatura não foi favorável à ocorrência da doença. Entretanto, como observado, o resultado final considerou a identificação de favorabilidade alta, vez que seis das sete variáveis contribuíram para que ocorresse tal definição.

4.6 Relatórios Analíticos, Recomendações e *Dashboard*

Como parte dos resultados, também foram consideradas as entregas de relatórios analíticos, também conhecidos como relatórios de apoio à decisão.

Esses relatórios possibilitam uma visão geral das séries temporais de dados climáticos e de imagens digitais das folhas de soja. Os relatórios foram originados de consultas a ferramentas OLAP, utilizando a base de dados histórica do DW, cujo modelo foi construído conforme os requisitos descritos na Tabela 2.

Para fins de exemplos sobre esses resultados, uma carga foi executada no DW, com dados coletados até 31/12/2017, por meio de um script escrito em SQL. Esse *script* foi elaborado a partir das junções das tabelas de dados do banco de dados transacional, atendendo às consultas dos requisitos elaborados.

Os requisitos foram divididos em três assuntos. O primeiro focando na influência das variáveis climáticas na favorabilidade da FAS (*Assunto 1*); o segundo tratou da contabilização da favorabilidade baixa, média e alta por ano (*Assunto 2*); e o terceiro possibilitando a investigação sobre a influência da imagem da folha da soja na favorabilidade da FAS, por ano, nas etapas de plantio e colheita, que correspondiam aos estádios reprodutivos, prioritariamente R5 e R6, pois estes são os mais atingidos pela doença (*Assunto 3*).

Para esse exemplo considerado, esse relatório analítico (*Assunto 1*), conforme a Figura 58, apresenta as variáveis consideradas na análise do *Assunto 1*, por meio de operações OLAP, e suas respectivas relações em uma escala anual na série temporal histórica de dados. Esse relatório mostrou que as maiores incidências da FAS ao longo do tempo foram provocadas pelas variáveis do período de molhamento foliar, com 142 ocorrências; temperatura máxima, com 137 ocorrências; temperatura mínima, com 142 ocorrências; período mínimo de molhamento foliar, com 97 ocorrências; e dados de classificação da imagem da folha da soja, com 8 ocorrências. Seguindo essa linha de raciocínio, observou-se que as demais variáveis mais significativas para a favorabilidade da doença foram o período mínimo de molhamento foliar, seguido pelo ponto de orvalho e pela faixa de temperatura.

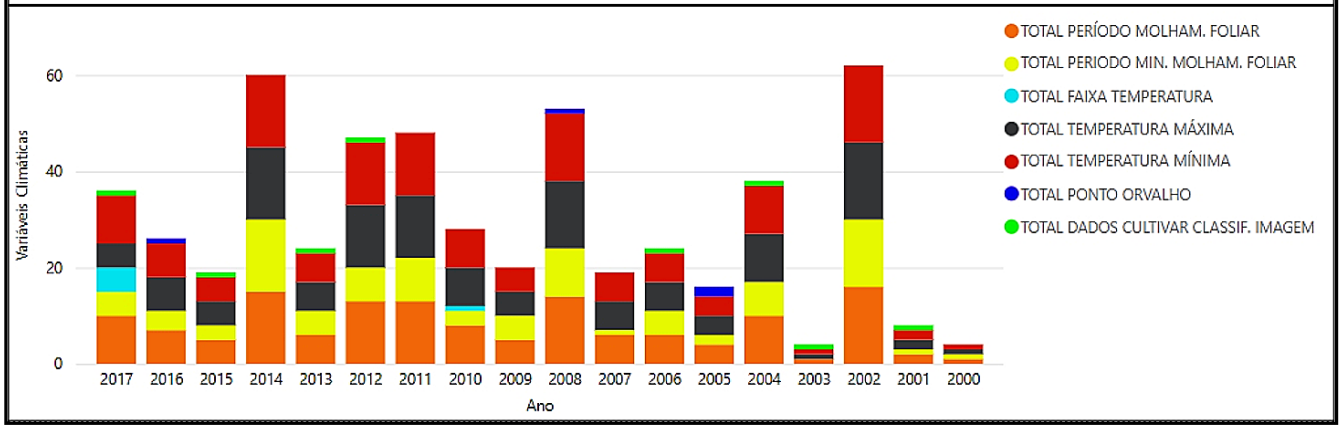
Adicionalmente, ao utilizar a operação OLAP *drill down*, foi possível considerar para esta modalidade de relatório, a verificação do detalhamento de ano para trimestre, equivalente ao período do segundo ciclo de cultura, ou seja, de setembro a dezembro. Portanto, observou-se ainda que o quarto trimestre foi o responsável pelos maiores índices de incidência da doença, abrangendo todas as variáveis analisadas, sabendo que o terceiro trimestre foi representado apenas pelo mês de setembro. Ainda, no (*Assunto 1*), outro detalhamento, via *drill down* para mês, foi possível identificar, na série temporal de dados, que os anos 2000, 2001, 2005, 2007, 2008, 2010, 2011, 2014 e 2016 não apresentaram índice de favorabilidade à FAS, em todos os meses do segundo ciclo de cultura. Em contrapartida, os demais anos apresentaram índices em todos os meses referentes ao ciclo avaliado.

No nível de detalhamento por mês, na série histórica (Figura 58), foram identificadas que as maiores incidências de favorabilidade ocorreram em 2002, nos meses de novembro e dezembro, em 2003, no mês de outubro e nos anos de 2013 e 2016, no mês de novembro. Outros picos de alta da favorabilidade foram identificados com menos intensidade que os anos citados, porém maiores que a média para os demais anos, para o mês de novembro, em 2015, 2016 e 2017.

Figura 58 – Análise Dados Relatório Analítico - *Assunto 1*

Gráfico de Favorabilidade por Variável por Ano

Ano	TOTAL PERÍODO MOLHAM. FOLIAR	TOTAL PERIODO MIN. MOLHAM. FOLIAR	TOTAL FAIXA TEMPERATURA	TOTAL TEMPERATURA MÁXIMA	TOTAL TEMPERATURA MÍNIMA	TOTAL PONTO ORVALHO	TOTAL DADOS CULTIVAR CLASSIF. IMAGEM
2014	15	15	0	15	15	0	0
2002	16	14	0	16	16	0	0
2008	14	10	0	14	14	1	0
2011	13	9	0	13	13	0	0
2004	10	7	0	10	10	0	1
2012	13	7	0	13	13	0	1
2006	6	5	0	6	6	0	1
2009	5	5	0	5	5	0	0
2013	6	5	0	6	6	0	1
2017	10	5	5	5	10	0	1
2016	7	4	0	7	7	1	0
2010	8	3	1	8	8	0	0
2015	5	3	0	5	5	0	1
2005	4	2	0	4	4	2	0
2000	1	1	0	1	1	0	0
2001	2	1	0	2	2	0	1
2007	6	1	0	6	6	0	0
2003	1	0	0	1	1	0	1
Total	**142**	**97**	**6**	**137**	**142**	**4**	**8**



Fonte: Próprio Autor

Na Tabela 29 relacionou-se os dados correspondentes ao ano de 2017, referentes ao trimestre que incluiu os exemplos de janelamento das séries temporais (Tabela 10) no ciclo 3, como parte da análise total e exibida na Figura 58. Tais análises relataram as influências de cada variável, de acordo com o mês de acontecimento e também o grau de influência de cada uma delas, avaliado do menor para o maior valor.

Tabela 29 – Resultados Relatório Analítico - *Assunto 1*

Ano 2017							
Mês	Período Molham. Foliar	Período Min. Molham. Foliar	Faixa Temperat.	Temperat. Max.	Temperat. Min.	Ponto Orvalho	Dados Classif. Imagem
Setembro	1	0	0	1	1	0	1
Outubro	1	1	0	1	1	1	1
Novembro	6	6	0	6	6	6	6
Dezembro	5	5	3	5	5	5	5

Fonte: Próprio Autor

Um outro aspecto que também foi observado na análise desses resultados, se encontra destacado na Tabela 29. Nesse caso, foi possível observar que nos meses de setembro, novembro e dezembro ocorreram maiores probabilidades para ocorrência da doença, conforme indicado pelos valores da favorabilidade que foram reportados para o referido ano. Também, na comparação dos dados, considerando o grau de influência crescente da FAS, por variável, os meses de novembro e dezembro registraram a maior incidência da doença para o ano de 2017, com exceção da variável "faixa de temperatura" para o mês de novembro, que apresentou menor grau de influência em relação as demais variáveis consideradas.

Os cenários para os meses de setembro e outubro apresentaram alterações na análise quanto ao grau de influência das variáveis. Logo, no mês de setembro, as variáveis de período mínimo de molhamento foliar, a faixa de temperatura e o ponto de orvalho apresentaram, igualmente, com menor influência que as variáveis do período de molhamento foliar, temperaturas mínima e máxima e dados de classificação da imagem. Entretanto, no mês de outubro, somente a variável faixa de temperatura se mostrou com menor influência em relação às demais variáveis, as quais apresentaram a mesma influência.

No relatório analítico (*Assunto 2*), conforme a Figura 60, a avaliação da contabilização das favorabilidades, dado o período de cultura de setembro a outubro, na região centro-oeste, indicou que não foram encontrados casos de favorabilidade baixa. Contudo, foram identificadas, para o mesmo período, ocorrências das favorabilidades média e alta.

Para a favorabilidade média, levando em consideração o ano de 2017, segundo a Tabela 30, para os meses de novembro e dezembro não foram identificadas ocorrências, porém, para os meses de setembro e outubro, houve resultados que puderam somar 3 e 7 para cada mês, respectivamente, com destaque para o mês de outubro, com registro de maior índice. Ainda para o *Assunto 2*, para o ano de 2017, com foco na favorabilidade alta e, de acordo com a Tabela 30, todos os meses apresentaram ocorrências para a favorabilidade alta, com maior índice para o mês de novembro.

Tabela 30 – Resultados Relatório Analítico - *Assunto 2*

Ano 2017				
Mês	Resultado Favorabilidade	Total Mês	Resultado Favorabilidade	Total Mês
Setembro	Favorabilidade Alta	2	Favorabilidade Média	3
Outubro	Favorabilidade Alta	3	Favorabilidade Média	7
Novembro	Favorabilidade Alta	6	———	———
Dezembro	Favorabilidade Alta	5	———	———

Fonte: Próprio Autor

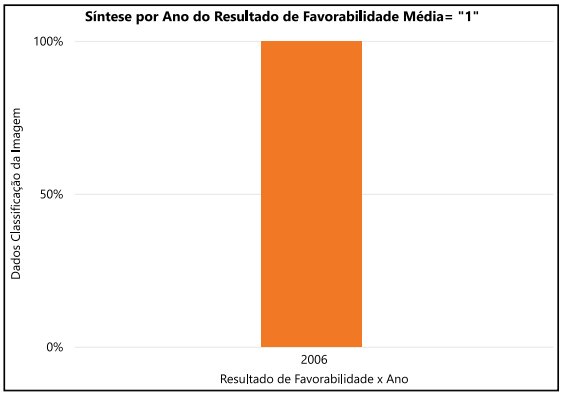
Os relatórios analíticos referentes ao *Assunto 3*, conforme ilustrado nas Figuras 59, 61 e 62, apresentaram os resultados das avaliações do período que compreendeu o estádio fenológico reprodutivo da soja R4, R5 e R6, com predominância em R5 e R6, sendo este o período de maior incidência da doença. O intervalo desse estádio fenológico ocorreu entre o octogésimo quinto e o nonagésimo quinto dia, abrangendo dezessete dias no total, o que correspondeu ao período de 17 de novembro a 4 de dezembro.

A Figura 61 apresentada síntese desses resultados discutidos, por favorabilidade e por ano, na série histórica de dados, em relação à variável de classificação da imagem. Observou-se que o valor de pico para favorabilidade alta foi registrado no ano de 2002, com 26 ocorrências, seguido pelos anos de 2004, com 11 ocorrências, e 2007, com 8 ocorrências. Os demais anos registraram valores abaixo desses índices.

Ainda sobre a favorabilidade alta, uma visão detalhada foi mostrada na Figura 62, sob a perspectiva dos dados de classificação da imagem versus o resultado de favorabilidade por mês. Foi possível observar que apenas nos anos 2000, 2003, 2006 e 2008 não houve registros de favorabilidade nos estágios de R5 e R6, nos meses de novembro e dezembro. Em contrapartida, nos demais anos da série temporal, foram apresentados registros em ambos os meses, confirmando a participação da variável analisada.

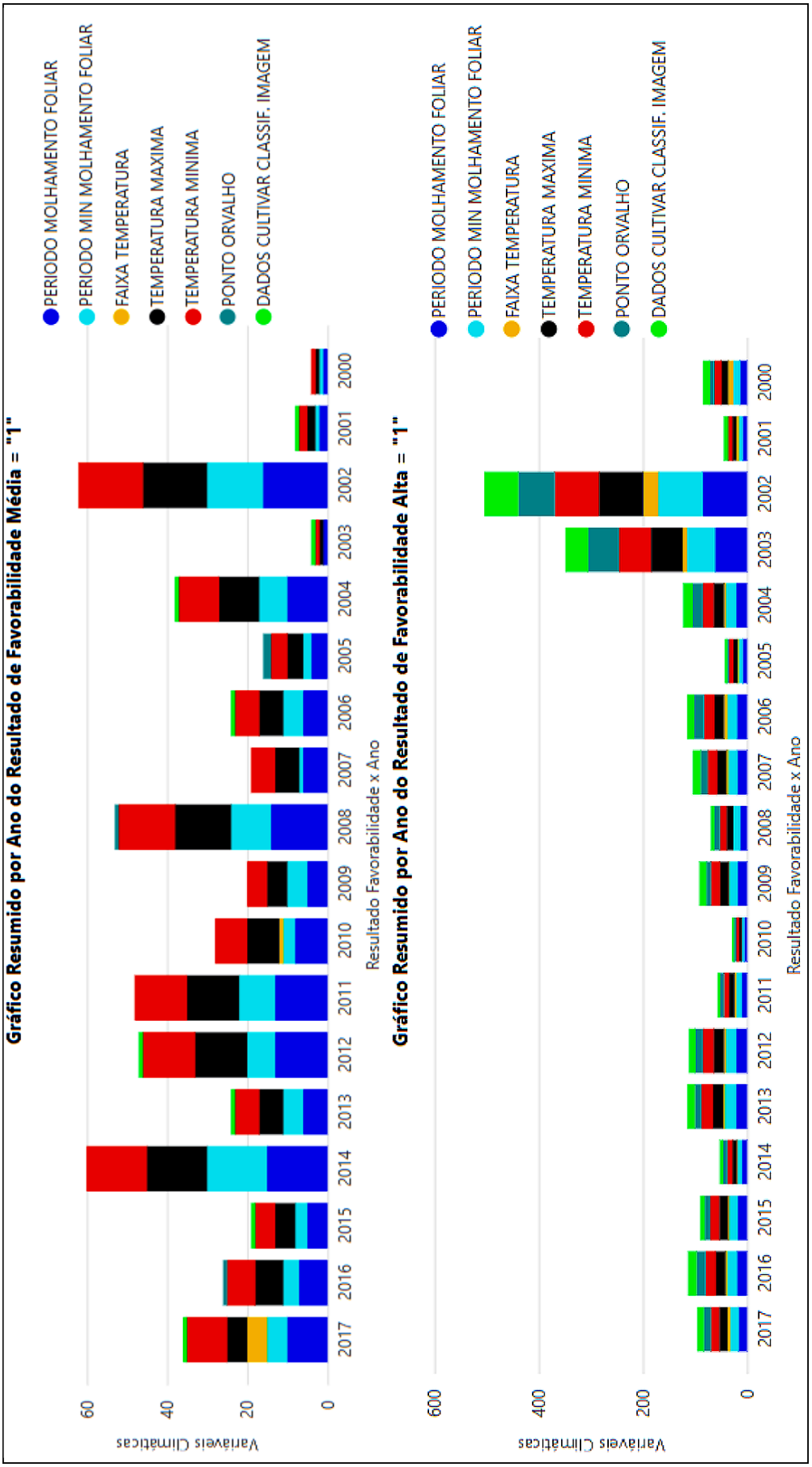
Por outro lado, durante o período considerado para R5 e R6, foi encontrado apenas um registro de favorabilidade média, ocorrido no mês de novembro de 2006, conforme mostrado na Figura 59.

Figura 59 – Relatório Analítico Favorabilidade Média - *Assunto 3*



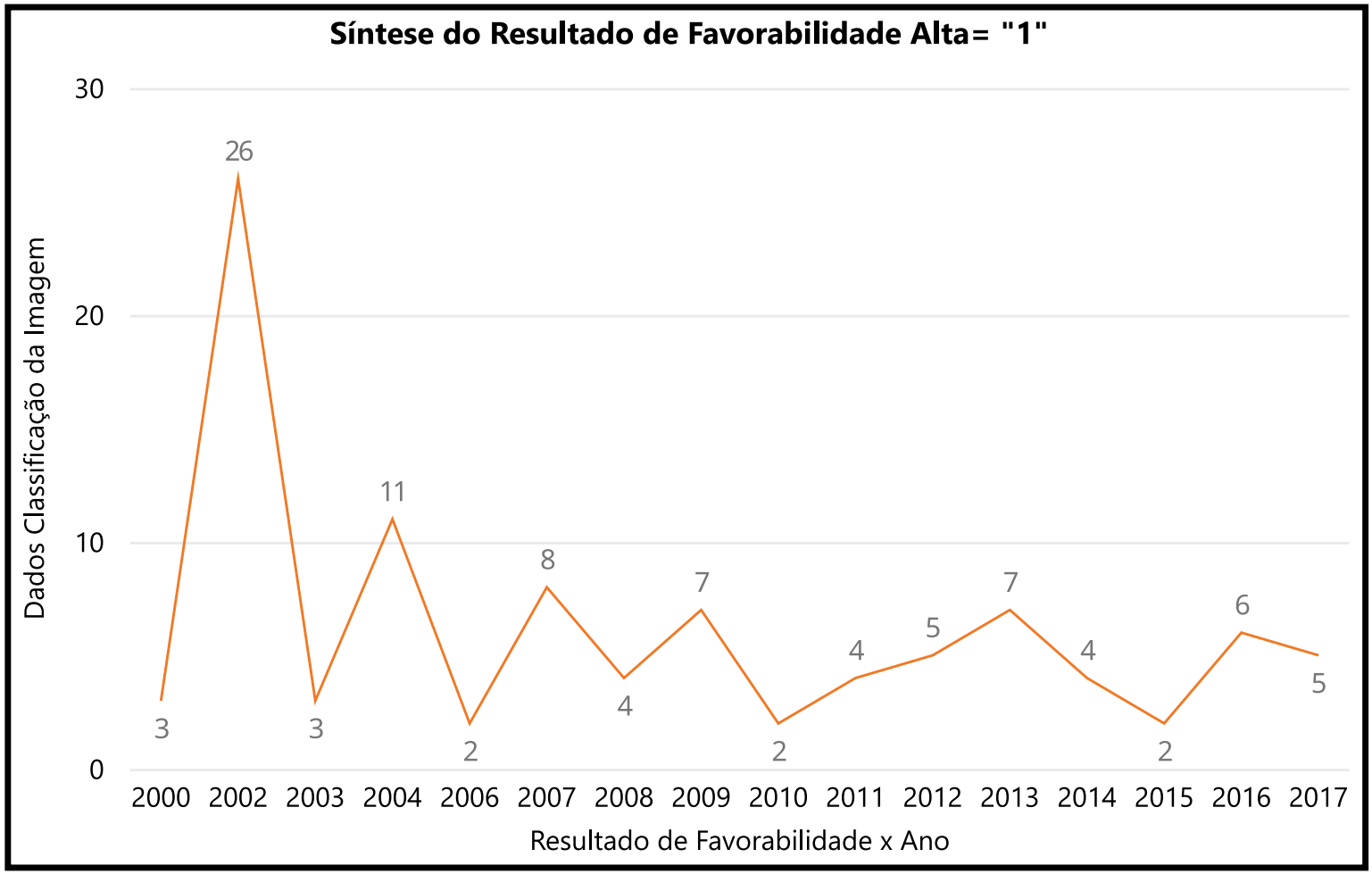
Fonte: Próprio Autor

Figura 60 – Análise Dados Relatório Analítico - *Assunto 2*



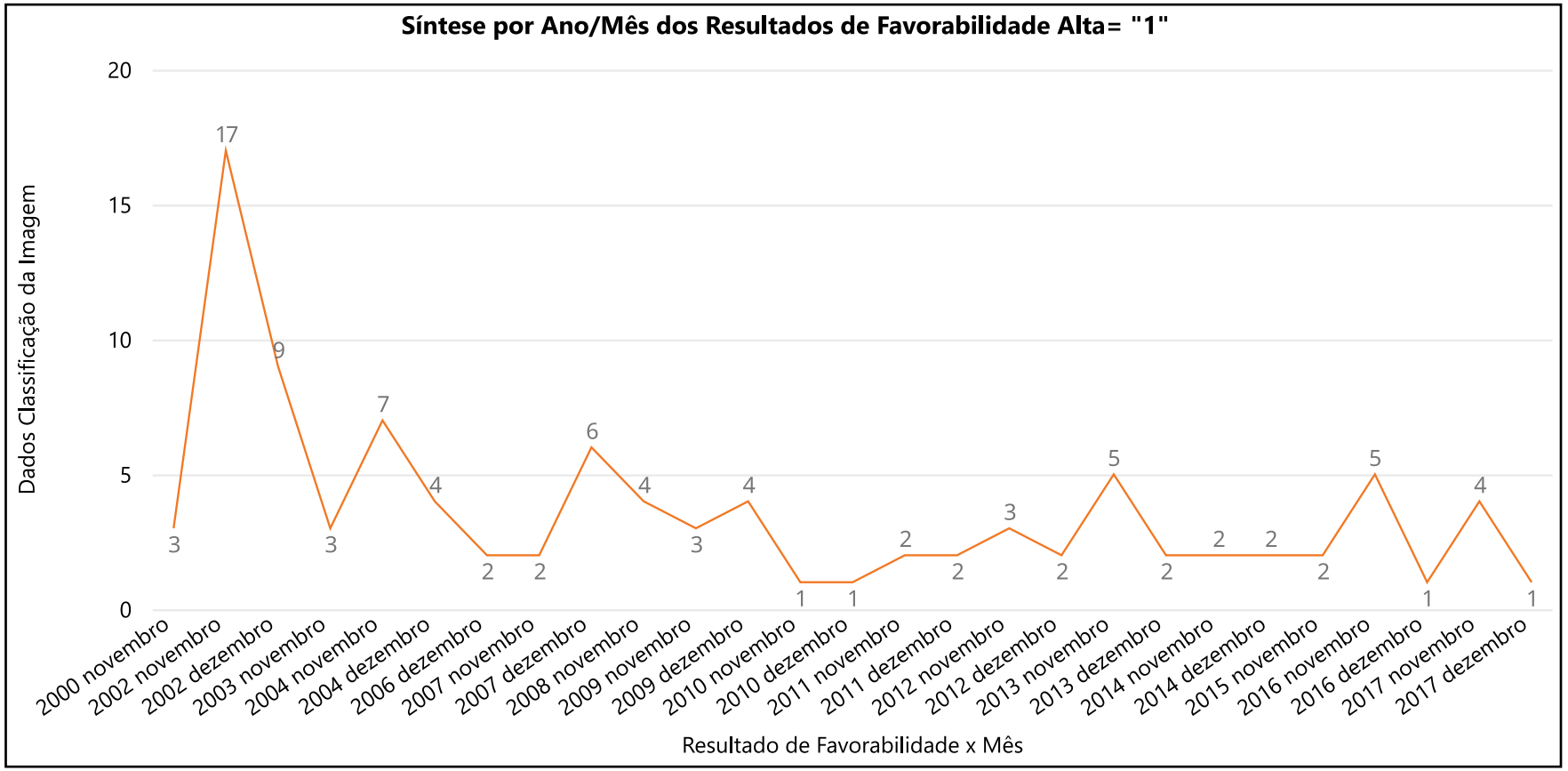
Fonte: Próprio Autor

Figura 61 – Relatório Analítico Favorabilidade Alta - *Assunto 3*



Fonte: Próprio Autor

Figura 62 – Relatório Analítico Favorabilidade Alta Detalhado - *Assunto 3*



Fonte: Próprio Autor

Com base nos dados de 2017, referentes ao trimestre que incluiu os exemplos de janelamento das séries temporais (Tabela 10) no ciclo 3, disponíveis na Tabela 31, a análise revelou a soma das ocorrências de favorabilidade média e alta nos meses de novembro e dezembro daquele período.

Tabela 31 – Resultados Relatório Analítico - *Assunto 3*

Ano 2017		
Mês	Resultado Favorabilidade	Dados Cultivar Classif. Imagem
Novembro	Favorabilidade Alta	4
Novembro	Favorabilidade Média	1
Dezembro	Favorabilidade Alta	1

Fonte: Próprio Autor

Para este relatório de recomendações também é ainda considerado, decorrente da parte analítica, conjunto de informações para o auxílio à decisão de agrônomos e produtores em formato *dashboard*. Sendo assim, este relatório de recomendações pode ser exibido em uma aba de "Recomendações Agrícolas", conforme retratado no Apêndice C. Além disso, foi incluído no mesmo o link do site Agrofit (2023) para consultas sobre atualizações de dosagens e novos fungicidas cadastrados pelo governo federal, divulgados pelo MAPA.

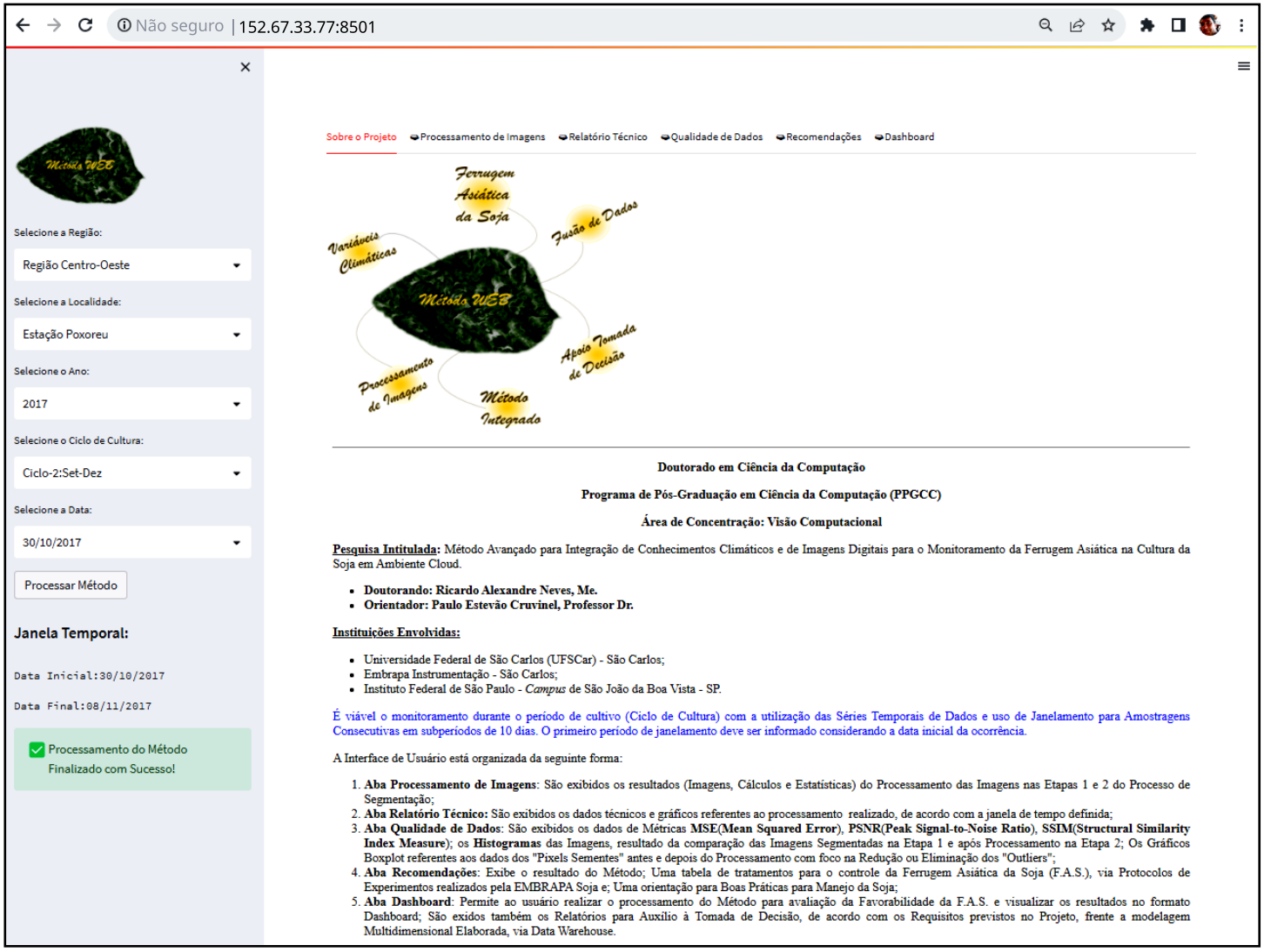
Ainda como parte dos resultados, as interfaces web, proporcionam uma navegação limpa e simples ao sistema de visão e inteligência computacional em ambiente de nuvem para gestão de risco da ferrugem asiática na cultura da soja, facilitando a operação para potenciais usuários. Neste sentido, as interfaces contribuem para a organização das informações, a qual divide a interface principal em cinco partes. Tal divisão organiza o ambiente e apresenta um texto com informações básicas sobre o sistema e o projeto de forma geral (Figura 64). Além disto, ainda na interface principal, é permitida a realização do processamento das séries temporais de dados, o qual é iniciado a partir do uso do botão "Processar sistema", possibilitando configurar os parâmetros para cada etapa e a leitura das informações para cada ciclo de análise relacionado aos ciclos da cultura.

A primeira aba, (Figura 64), apresenta os resultados do processamento dos cálculos envolvidos nas etapas de segmentação para as imagens digitais das folhas da soja.

A segunda aba, (Figura 65), apresenta os resultados do relatório técnico, originado a partir dos dados climáticos processados e respectivos gráficos de interpolação para cada variável climática considerado. Outros resultados relacionados às técnicas de interpolação podem ainda ser visualizados, no âmbito desta aba, a partir da operação de rolamento da página.

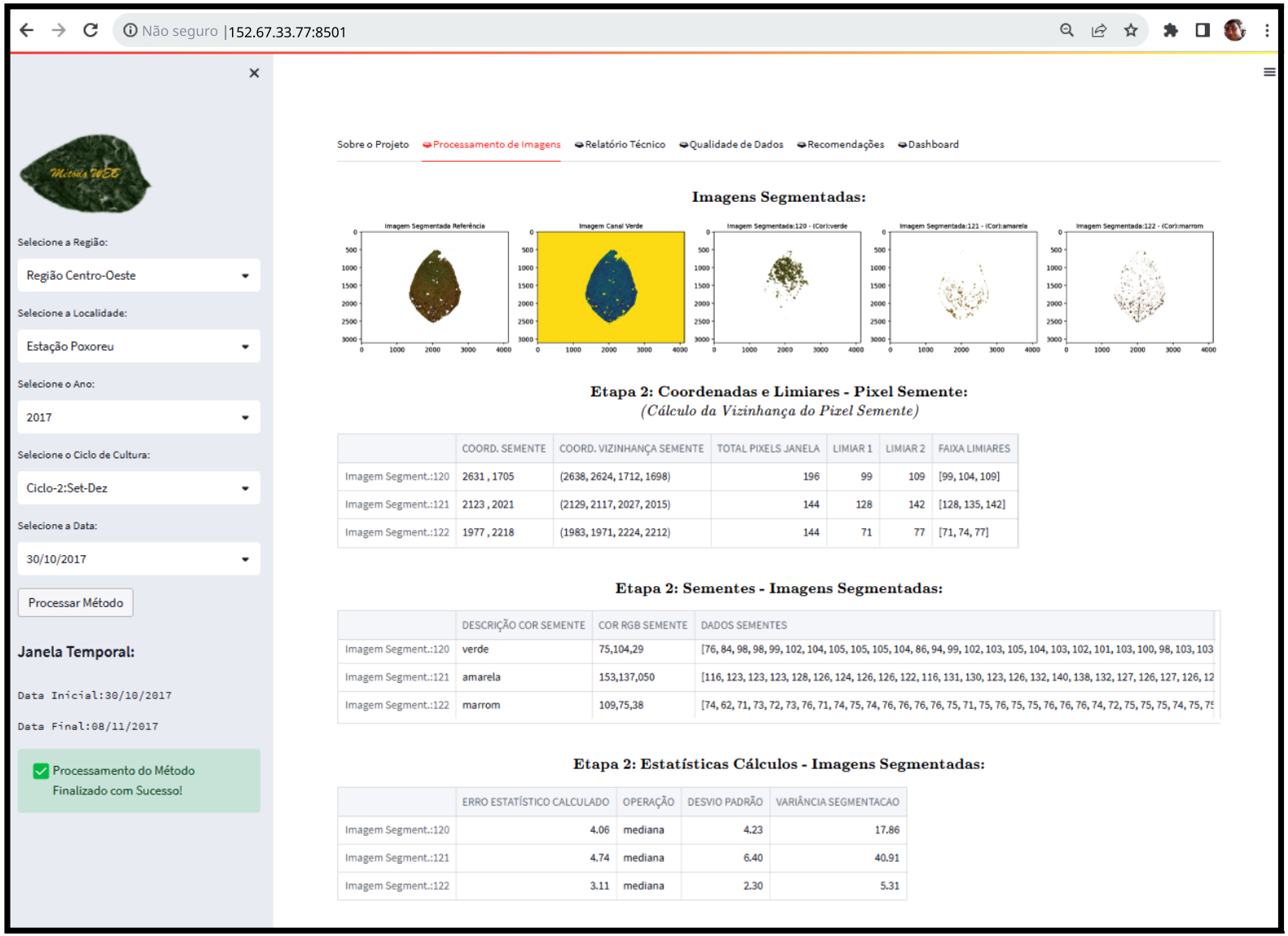
Quanto à terceira aba, (Figura 66), a mesma foi composta pelos relatórios de qualidade de dados, que apresentaram os dados de indicadores e suas respectivas dimensões, orientados pelo *framework* de qualidade.

Figura 63 – Interface Sobre o Projeto



Fonte: Próprio Autor

Figura 64 – Interface do Processamento Imagens



	COORD. SEMENTE	COORD. VIZINHANÇA SEMENTE	TOTAL PIXELS JANELA	LIMIAR 1	LIMIAR 2	FAIXA LIMIARES
Imagem Segment.:120	2631 , 1705	(2638, 2624, 1712, 1698)	196	99	109	[99, 104, 109]
Imagem Segment.:121	2123 , 2021	(2129, 2117, 2027, 2015)	144	128	142	[128, 135, 142]
Imagem Segment.:122	1977 , 2218	(1983, 1971, 2224, 2212)	144	71	77	[71, 74, 77]

	DESCRIÇÃO COR SEMENTE	COR RGB SEMENTE	DADOS SEMENTES
Imagem Segment.:120	verde	75,104,29	[76, 84, 98, 98, 99, 102, 104, 105, 105, 105, 104, 86, 94, 99, 102, 103, 105, 104, 103, 102, 101, 103, 100, 98, 103, 103
Imagem Segment.:121	amarela	153,137,050	[116, 123, 123, 123, 128, 126, 124, 126, 126, 122, 116, 131, 130, 123, 126, 132, 140, 138, 132, 127, 126, 127, 126, 12
Imagem Segment.:122	marrom	109,75,38	[74, 62, 71, 73, 72, 73, 76, 71, 74, 75, 74, 76, 76, 76, 76, 75, 71, 75, 76, 75, 75, 76, 76, 76, 74, 72, 75, 75, 75, 74, 75, 7

	ERRO ESTATÍSTICO CALCULADO	OPERAÇÃO	DESVIO PADRÃO	VARIÂNCIA SEGMENTACAO
Imagem Segment.:120	4.06	mediana	4.23	17.86
Imagem Segment.:121	4.74	mediana	6.40	40.91
Imagem Segment.:122	3.11	mediana	2.30	5.31

Fonte: Próprio Autor

Figura 65 – Interface do Relatório Técnico

	ID. Dados Clim.	ID.Projeto	Estação Climática	Período	Data Medição	Precipitação	Temp. Máxima	Temp. Mínima	Umidade Relativa	Ponto de Orvalho	Temp. Média Compensada	Região	Status
0	5998	1	Estação Poxoreu	safra	30/10/2017	4.20	35.50	24.00	72.75	23.08	28.44	Região Centro-Oeste	Original
1	5999	1	Estação Poxoreu	safra	31/10/2017	0.00	32.50	24.40	88.75	23.80	25.80	Região Centro-Oeste	Original
2	6000	1	Estação Poxoreu	safra	01/11/2017	18.00	33.30	22.50	79.00	22.85	26.80	Região Centro-Oeste	Original
3	6001	1	Estação Poxoreu	safra	05/11/2017	0.00	33.00	23.20	84.00	22.62	25.52	Região Centro-Oeste	Original
4	6002	1	Estação Poxoreu	safra	06/11/2017	0.00	33.60	23.80	88.25	24.02	26.12	Região Centro-Oeste	Original
5	6003	1	Estação Poxoreu	safra	07/11/2017	3.00	34.50	23.40	83.00	23.06	26.18	Região Centro-Oeste	Original
6	6004	1	Estação Poxoreu	safra	08/11/2017	0.00	33.50	24.00	84.25	23.47	26.34	Região Centro-Oeste	Original
7	6004	1	Estação Poxoreu	safra	08/11/2017	4.20	35.50	24.00	72.80	23.10	28.40	Região Centro-Oeste	Interpolado
8	6004	1	Estação Poxoreu	safra	08/11/2017	6.10	32.80	24.90	88.70	23.80	25.90	Região Centro-Oeste	Interpolado
9	6004	1	Estação Poxoreu	safra	08/11/2017	5.40	32.60	23.90	86.80	23.70	26.00	Região Centro-Oeste	Interpolado

Fonte: Próprio Autor

O *framework* desenvolvido possibilitou analisar, nas diferentes etapas do desenvolvimento do sistema, a perspectiva da qualidade dos dados, pois tratou de uma importante avaliação frente aos resultados, quanto ao grau de confiabilidade da informação trabalhada.

Tais informações foram organizadas de acordo com as descrições dos indicadores e dimensões que tratam sobre a qualidade dos dados, por meio de elementos textuais e tabelas para suas explicitações. Também, foi considerado explicitar as informações sobre as diferentes etapas envolvidas no sistema de visão e inteligência computacional em ambiente de nuvem para gestão de risco da ferrugem asiática na cultura da soja (Figura 66).

O primeiro *container* de expansão se refere à qualidade dos dados envolvidos na etapa II da segmentação e exibe os histogramas separados pelas cores verde, amarela e marrom.

Para o exemplo considerado, obtido como parte do conjunto das análises realizadas, a Figura 67, ilustra gráficos sobre a análise dos indicadores de qualidade desses dados. Nesse contexto, nos histogramas foi considerado a distribuição dos *pixels*, ao longo da imagem, em relação a sua intensidade decorrente do processo estabelecido na etapa I da segmentação, ou seja, considerando a retirada do fundo complexo e incluindo uma rotulagem como imagem "Original".

Também, os indicadores MSE, PSNR, SSIM e os gráficos dos *outliers* foram considerados para a avaliação da qualidade, ao longo da segmentação realizada. Os valores encontrados ficaram compreendidos nas faixas $0,01 \leq MSE \leq 0,06$, com mediana de 0,03; $0,04 \leq PSNR \leq 18,98$, com mediana de 14,29; $0,87 \leq SSIM \leq 0,97$ com mediana de 0,94.

A Tabela 32 apresenta dados com os valores das métricas MSE, PSNR e SSIM, calculadas para avaliar ruído e similaridade, frente às imagens segmentadas da etapa de extração do fundo (etapa I) e as imagens processadas nas três cores (etapa II).

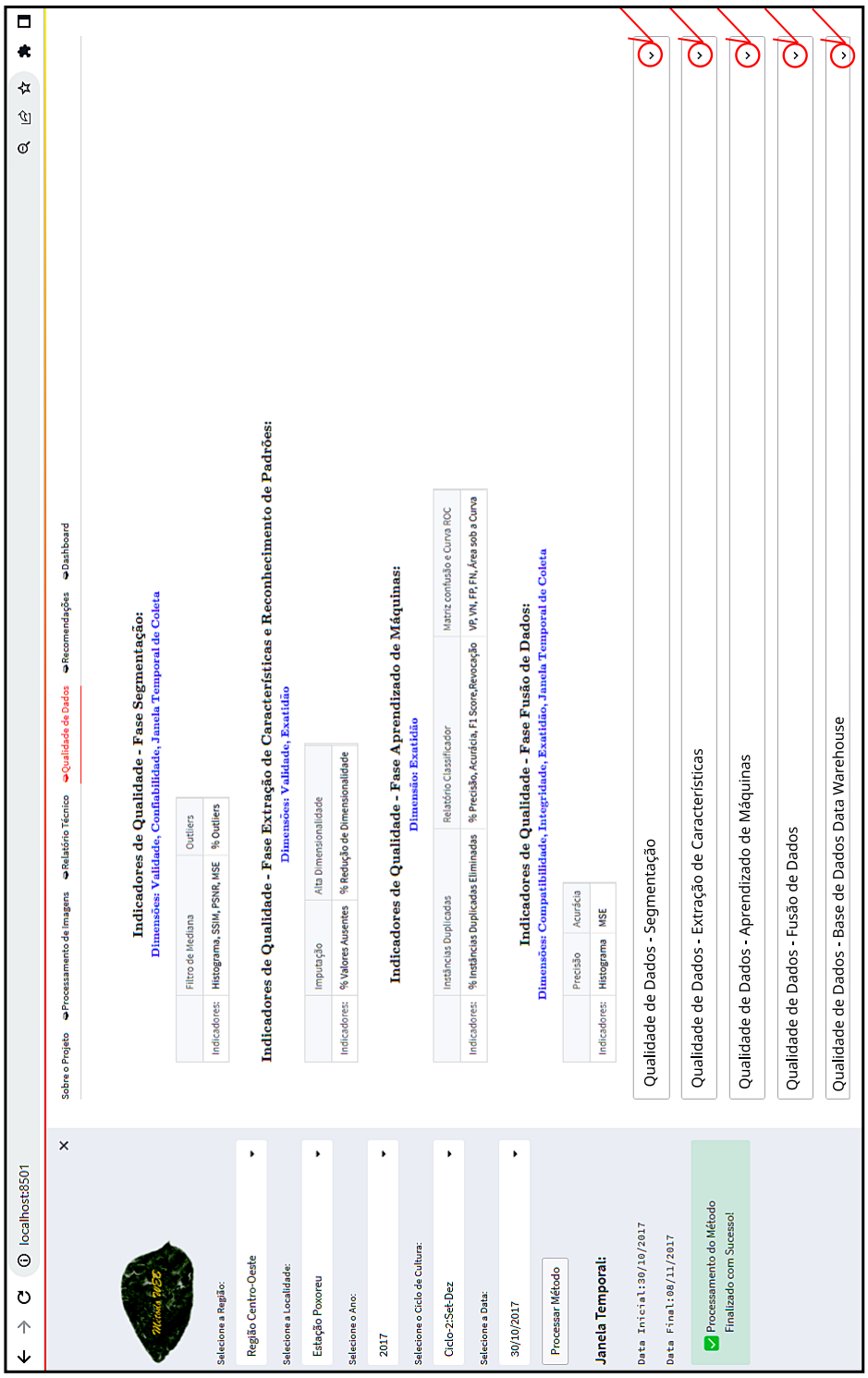
Para o exemplo, foram considerandos três histogramas (Figura 67). O primeiro histograma indica uma região cujas intensidades estão compreendidas na faixa de valores $50 \leq pixel \leq 210$, representados com linha de cor vermelha. Adicionalmente, nesses histogramas, também foram consideradas linhas de cor azul para representar dentro dessa faixa avaliada os picos relacionados as cores verde (104), amarela (137) e marrom (75).

Ainda nos histogramas, onde se utilizou a cor azul, foram identificadas diferenças de áreas em relação àqueles onde se utilizou a cor vermelha. Isso ocorreu decorrente da aplicação da operação de segmentação.

Adicionalmente, foi percebida, nos histogramas, uma perda de *pixels* na imagem de saída com relação à imagem de entrada, dada pela sobreposição dos picos, quando consideradas as três cores juntas. Essa perda ocorreu porque houve *pixels* na imagem de entrada que não corresponderam às cores de referência e também pelo fato da escolha da imagem segmentada final, que foi realizada na etapa de clusterização, ter sido executada de forma supervisionada. Nessa operação, optou-se por considerar os *pixels* cujos valores se encon-

travam mais próximos da cor de referência. Nesse contexto, em que pese alguma perda de *pixels*, foram considerados seis diferentes *clusters* sem perda significativa da qualidade dos dados.

Figura 66 – Interface de Qualidade de Dados

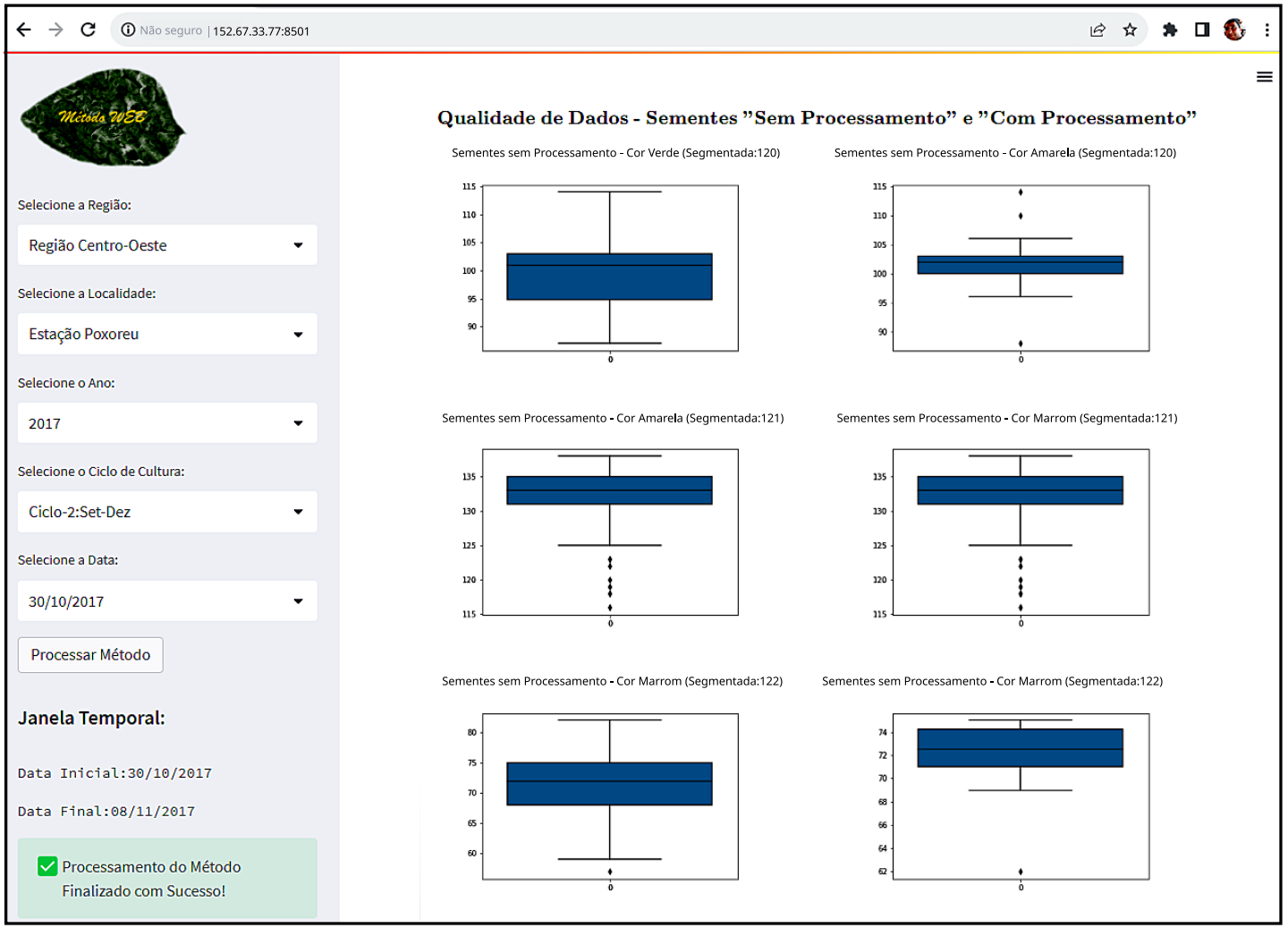


Fonte: Próprio Autor

Figura 67 – Interface de Qualidade de Dados: Segmentação

Fonte: Próprio Autor

Figura 68 – Interface de Qualidade de Dados: Segmentação - Continuação



Fonte: Próprio Autor

Tabela 32 – Análise de Qualidade da Segmentação - Métricas e *Outliers*

Imagens Segmentadas	Métricas			*Outliers*	
	MSE	PSNR (dB)	SSIM	Sementes	Cálculo
120 - Verde	0,05	13,35	0,91	0	3
121 - Amarela	0,06	12,59	0,91	14	14
122 - Marrom	0,05	12,94	0,91	1	1

Fonte: Próprio Autor

No caso da cor verde, na Tabela 32, para a coluna "Sementes", os dados não apresentaram nenhum valor de *outlier* para o conjunto de dados de sementes, pois a vizinhança do *pixel* semente apresentou cores com tonalidades iguais à referência, com limiar inicial calculado com o valor de 104, sendo este o valor referência para a cor verde. Em contrapartida, após o cálculo dos limiares, a variabilidade dos dados diminuiu, ou seja, os dados se aproximaram da mediana, o que ocasionou também a diminuição dos limites inferior e superior que, por essa razão, 3 *outliers* foram evidenciados, sendo dois deles do limite superior e o outro do limite inferior. No que tange à qualidade dos dados, para a cor verde, obteve-se 3% de erro esperado, após o processamento dos limiares, mesmo com o aumento dos *outliers*. Na Figura 68 foram representados os gráficos *boxplot* que se referiram aos dados dos *outliers* relatados na Tabela 32.

Ao continuar a análise, de acordo com a Tabela 32, os resultados de *outliers* na coluna "Sementes" para as cores amarela e marrom, frente aos valores de *outliers*, após o cálculo, foram mantidos os mesmos, em relação à coluna "Cálculo".

O fato do número de *outliers* ter se mantido em 14, para a cor amarela, significou que os *pixels* de vizinhança encontrados permaneceram com a mesma variabilidade, mesmo depois de calculados os limiares. Contudo, não houve o aparecimento de novos *outliers*.

O comportamento se repetiu para a análise da cor marrom, pois também foi mantido o mesmo número de *outliers*, ou seja, "1" para as colunas de "Sementes" e "Cálculo". Porém, foi percebida uma redução de variabilidade dos dados após o cálculo dos limiares, pois foram reduzidos os limites inferior e superior, o que ocasionou a diminuição da diferença de simetria dos dados.

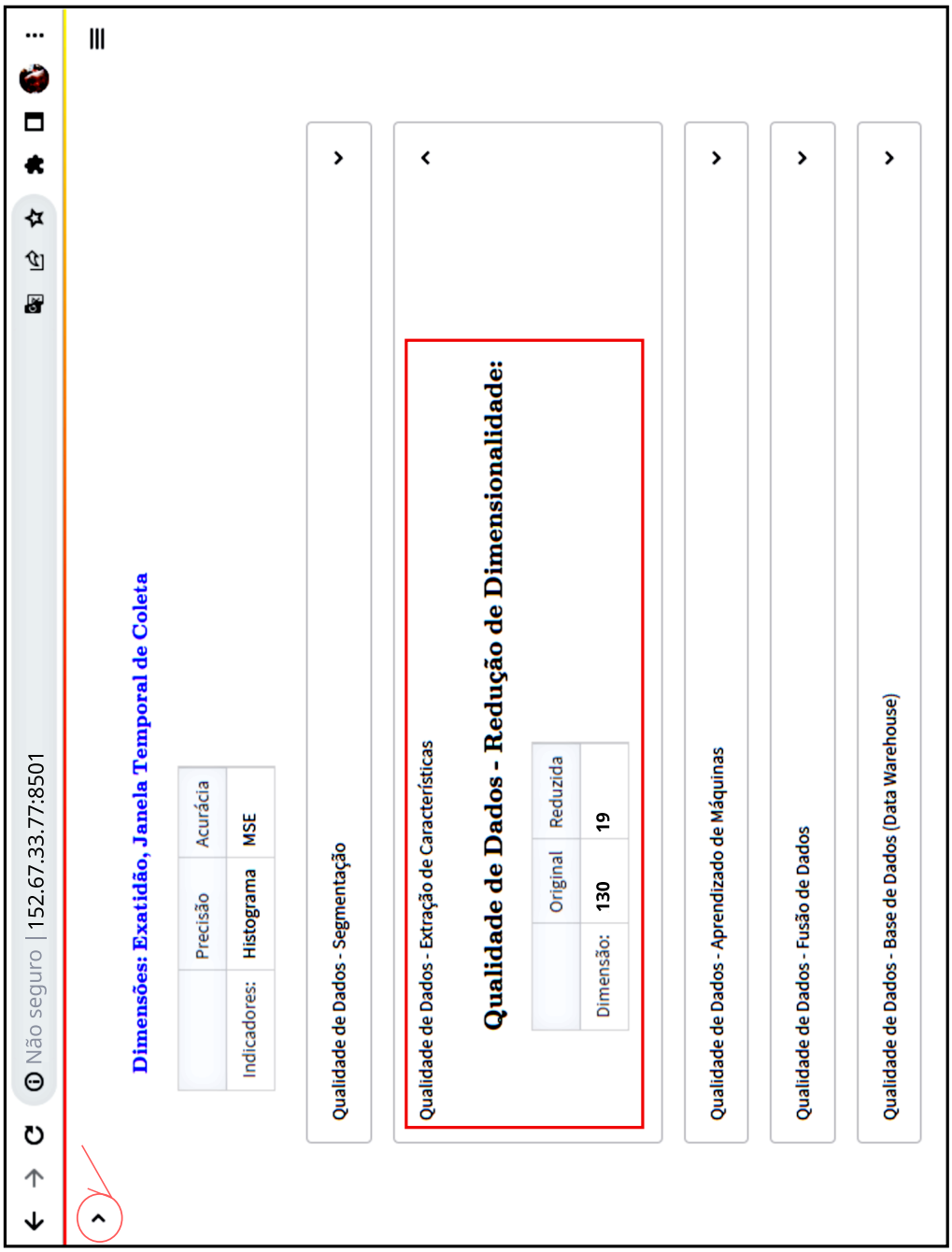
O *container* de extração de características, conforme destacado na Figura 69, exibiu as informações da dimensão dos dados de características processadas com 130 componentes e após de redução de dimensionalidade passou a ser igual a 19, para o exemplo considerado. Esse número de componentes respondeu por um valor de variância superior a 70%.

No *container* de aprendizado de máquina são visualizados os dados de qualidade que se referem aos indicadores de instâncias duplicadas. Adicionalmente, o relatório do classificador também é exibido, contendo os indicadores de acurácia, precisão, *F1-score* e revocação, caracterizadas como informações principais deste relatório. Além disso, outros indicadores de qualidade da etapa de classificação foram exibidos neste *container*, tais

como os gráficos matriz de confusão e área sob a curva ROC. As Figuras 70 e 71 ilustram o detalhamento do resultado obtido para o exemplo considerado.

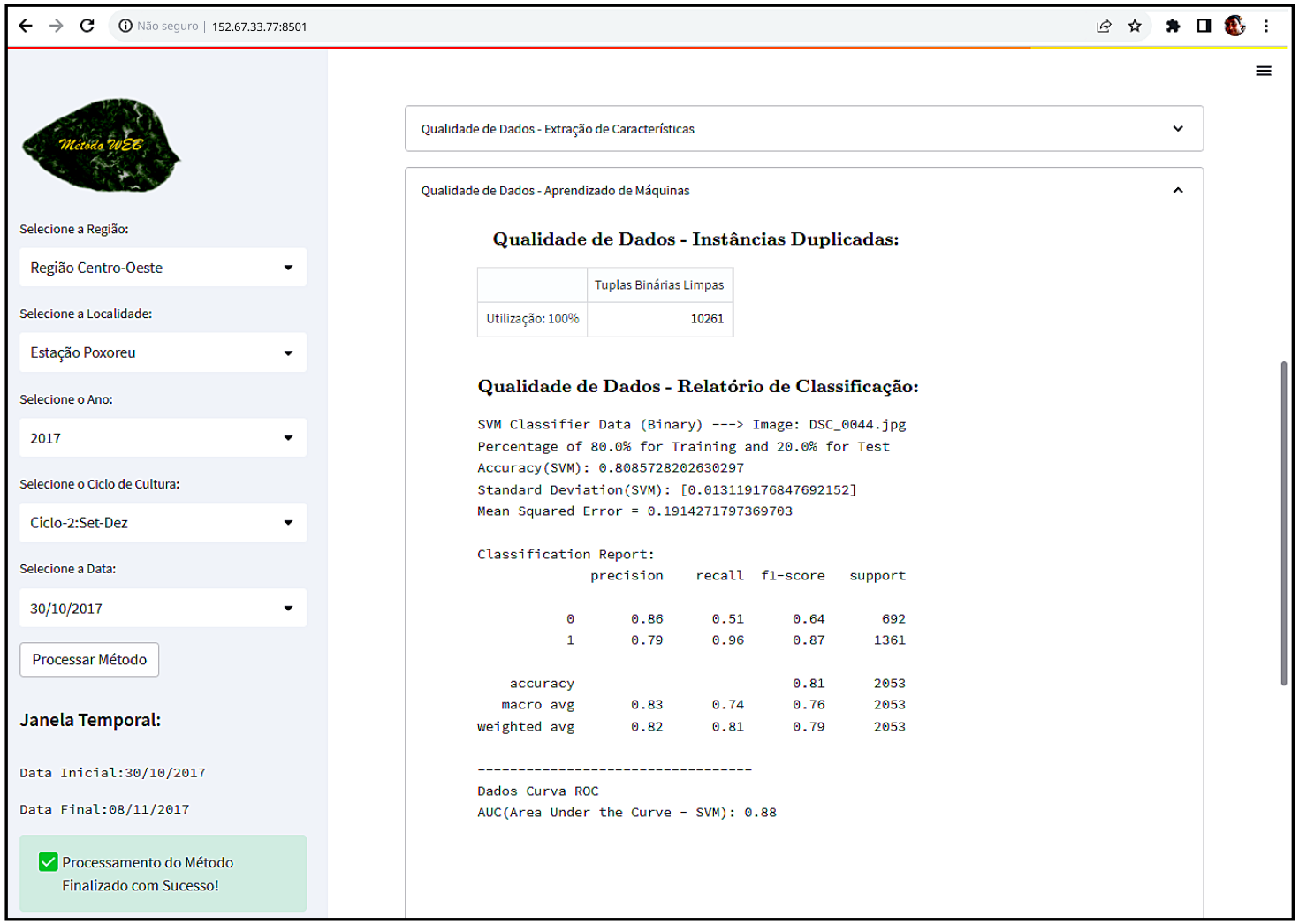
Resultados da análise da qualidade de dados decorrentes da etapa de classificação foram analisados considerando as informações sobre matriz de confusão, curva ROC, métricas sobre acurácia, precisão, relocação e *F1-score*. Esse conjunto de informações passou a compor um relatório customizado sobre o uso dos classificadores, consequentemente integrando o relatório geral do sistema sobre métricas e qualidade dos dados.

Figura 69 – Interface de Qualidade de Dados: Extração de Características



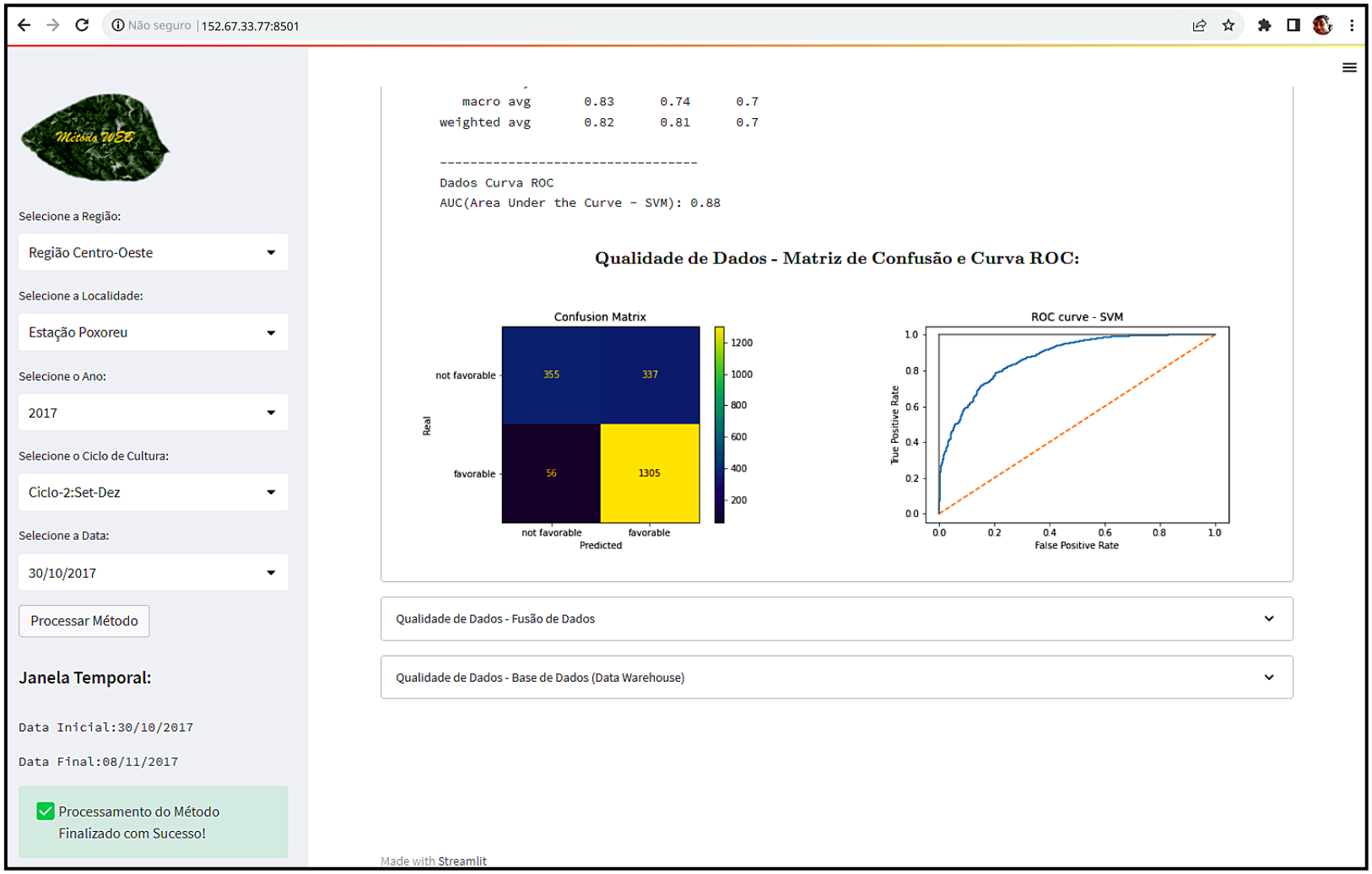
Fonte: Próprio Autor

Figura 70 – Interface de Qualidade de Dados: Aprendizado de Máquinas



Fonte: Próprio Autor

Figura 71 – Interface de Qualidade de Dados: Aprendizado de Máquinas - Continuação

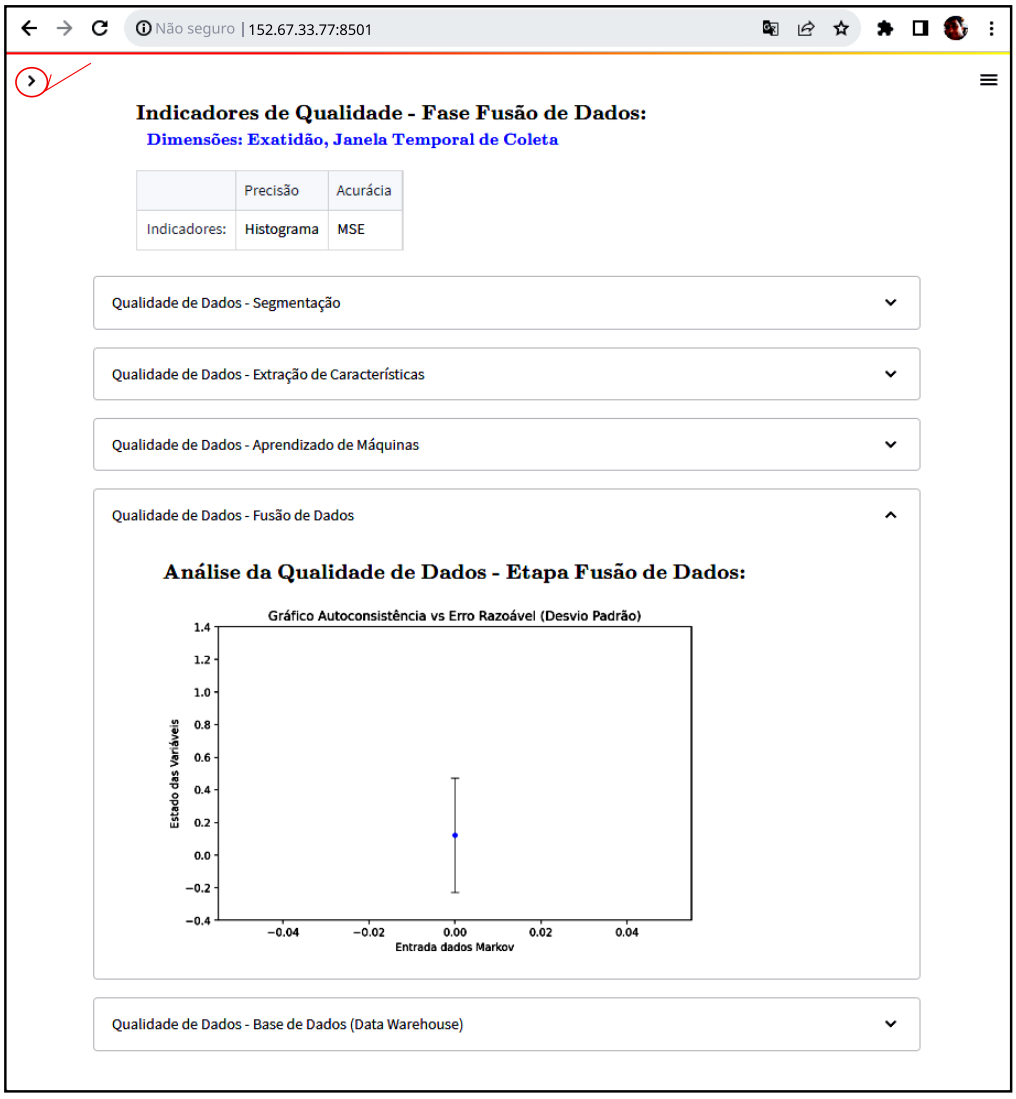


Fonte: Próprio Autor

De fato, destacou-se ainda que a curva ROC esteve mais próxima do canto superior esquerdo do gráfico, perto da coordenada (x=0, y=1), o que também indicou um bom desempenho do classificador, uma alta taxa de verdadeiros positivos e baixa taxa de falsos positivos.

Quanto a avaliação da qualidade dos dados envolvidos no processo de fusão, a Figura 72 ilustra o gráfico que corresponde a cada processamento realizado do modelo Markoviano, sob a ótica de uma janela temporal na série de dados climáticos e a informação da classificação dos padrões relacionados às imagens das folhas de soja que foram processadas.

Figura 72 – Interface de Qualidade de Dados: Fusão de Dados



Fonte: Próprio Autor

Quanto aos indicadores de qualidade considerados para este processo, os mesmos estão relacionados a precisão $\sigma(\bar{\jmath})$ (Equação 86), a acurácia $\sigma^2(\bar{\jmath})$ (Equação 87) e a estimativa do tempo de processamento Markoviano $\hat{\epsilon}(t)$ (Equação 88). Para essa avaliação foram considerados como entrada para tal processamento, os dados das medições de tempo e da

janela de dados da série temporal das variáveis. A qualidade dos dados no processo de Markov foi caracterizada pelo grau de acerto na fase em que os dados foram processados pelo algoritmo relacionado ao modelo baseado em cadeias ocultas. A Tabela 33 evidencia o resultado obtido diante do processamento de dados considerados em uma janela de tempo, tendo em conta, a título de exemplo, o ponto "15", conforme apresenta a Figura 73. Esse ponto traduziu a relação do estado das variáveis de entrada para a cadeia de Markov, frente aos valores de acurácia de $\sigma^2(\bar{\jmath})$.

A Figura 73 ilustra o resultado da avaliação da qualidade de dados para o processo de fusão, tomando em conta um conjunto de quinze janelas e seus respectivos tempos de processamento. Para esse ensaio foram considerados vetores de dados de entrada que traduziram situações relacionadas a favorabilidade baixa (os cinco primeiros resultados), favorabilidade média (os cinco próximos) e favorabilidade alta (cinco últimos casos).

Ao analisar os dados da Tabela 33, quando projetados no gráfico ilustrado na Figura 73, foi possível obter os erros para cada autocorrelação calculada, sendo que a faixa de valores para o desvio padrão foi observada na forma de $0,35 \leq \sigma \leq 0,49$. Esse resultado indicou um erro $< 1\%$, evidenciando um alto índice de qualidade para os dados processados na etapa da fusão de variáveis. Em seguida, para a favorabilidade baixa, foi notado que o primeiro ponto ficou destacado dos demais por apresentar erro igual a "0", acurácia e precisão iguais a "1". Essa situação ocorreu porque se tratou de uma particularidade dos dados de entrada, em que todas as variáveis possuíram valores iguais a "0".

Tabela 33 – Dados de Qualidade Markoviana

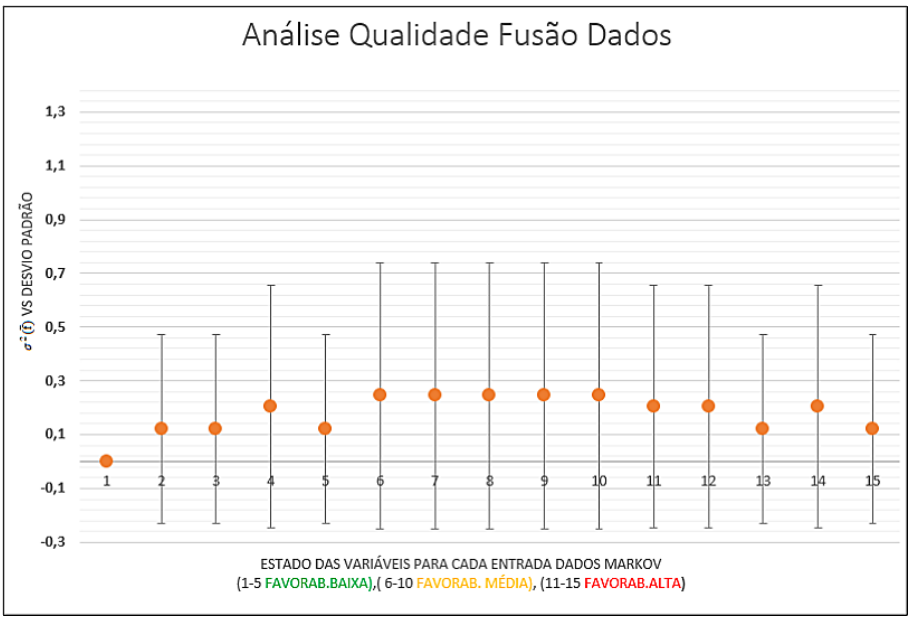
Favorab.	R1	R2	R3	R4	R5	R6	R7	$\hat{\epsilon}(t)$	$\sigma^2(\bar{\jmath})$	$\Delta\bar{\jmath}$	Acurácia	Precisão
Baixa	0	0	0	0	0	0	0	1,00	0	0	1	1
	0	0	0	0	0	0	1	1,00	0,12	0,35	0,88	0,65
	0	0	0	0	0	1	0	1,00	0,12	0,35	0,88	0,65
	0	0	0	0	0	1	1	1,00	0,20	0,45	0,80	0,55
	0	0	0	0	1	0	0	1,00	0,12	0,35	0,88	0,65
Média	0	0	0	0	1	1	1	1,00	0,24	0,49	0,76	0,51
	0	0	0	1	0	1	1	1,00	0,24	0,49	0,76	0,51
	0	0	0	1	1	0	1	1,00	0,24	0,49	0,76	0,51
	0	0	0	1	1	1	0	1,00	0,24	0,49	0,76	0,51
	0	0	0	1	1	1	1	1,00	0,24	0,49	0,76	0,51
Alta	0	1	1	1	1	0	1	1,00	0,20	0,45	0,80	0,55
	0	1	1	1	1	1	0	1,00	0,20	0,45	0,80	0,55
	0	1	1	1	1	1	1	1,00	0,12	0,35	0,88	0,65
	1	0	0	1	1	1	1	1,00	0,20	0,45	0,80	0,55
	1	1	0	1	1	1	1	1,00	0,12	0,35	0,88	0,65

Fonte: Próprio Autor

Em relação ao desvio padrão, as diferenças de erros observadas foram pequenas, sendo assim razoáveis os valores de autocorreção que foram encontrados. Para as autocorrelações calculadas identificou-se a dependência de dois fatores principais: as variáveis de entrada e o valor de $\hat{\epsilon}(t)$, ou seja, da dimensão janela temporal de processamento. A diferença dos valores de $\hat{\epsilon}(t)$ foram sutis, pois sofreu variação a partir da terceira ou quarta casa decimal, pois o processamento foi feito na mesma configuração de infraestrutura. Assim,

essa variável, nessa condição, contribuiu muito pouco para as diferenças dos valores de desvio padrão dos erros razoáveis para as autocorrelações calculadas.

Figura 73 – Qualidade Dados Markoviano



Fonte: Próprio Autor

No entanto, a variação de combinações das variáveis de entrada de R1 até R7 definem, com maior expressividade, as diferenças dos valores de desvio padrão dos erros razoáveis para as autocorrelações calculadas. As combinações das variáveis de R1 até R7, exibidas na Tabela 33, refletem que o aumento dos valores de desvio padrão que ocorreram a partir da presença de variáveis com valor igual "1". Um comportamento importante observado foi que o valor do desvio padrão atingiu o limite com até três variáveis com valor igual a "1", e estabilizou o valor na quarta variável. A partir da quinta variável com valor igual "1", o valor do desvio padrão dos erros razoáveis para as autocorrelações calculadas começou a decrescer, indicando maior consistência quanto ao processamento da informação.

Ao analisar a qualidade dos dados que a plataforma da Oracle *Cloud* ofereceu, como serviço para as funcionalidades dos bancos de dados *Autonomous*(Figura 74), foi notado que os dados de qualidade foram exibidos de modo individual, para a coluna do banco de dados, por meio de gráficos de barras, identificados na cor "verde". Para esse caso, uma visão macro sobre a qualidade foi exibida para todas as tabelas de um determinado banco de dados. Tal informação foi sinalizada na tabela na cor vermelha, quando ocorreu algum erro, e na cor verde, quando os dados se apresentaram consistentes. A parte superior central da Figura 74, ilustra para um dos bancos de dados uma situação favorável para a ausência de erros.

No *container* que se referiu à base de dados do *Data Warehouse*, foi apresentado um relatório da qualidade de dados, obtido na plataforma *Oracle Cloud*, que considerou os

recursos do *Autonomous Database* e *Analytics Cloud*. Este relatório, conforme ilustrado na Figura 75, oferece uma visão da qualidade dos dados das colunas das tabelas de dimensão e fatos do cubo de dados do *Data Warehouse*, abordando indicadores, como precisão, completude e integridade.

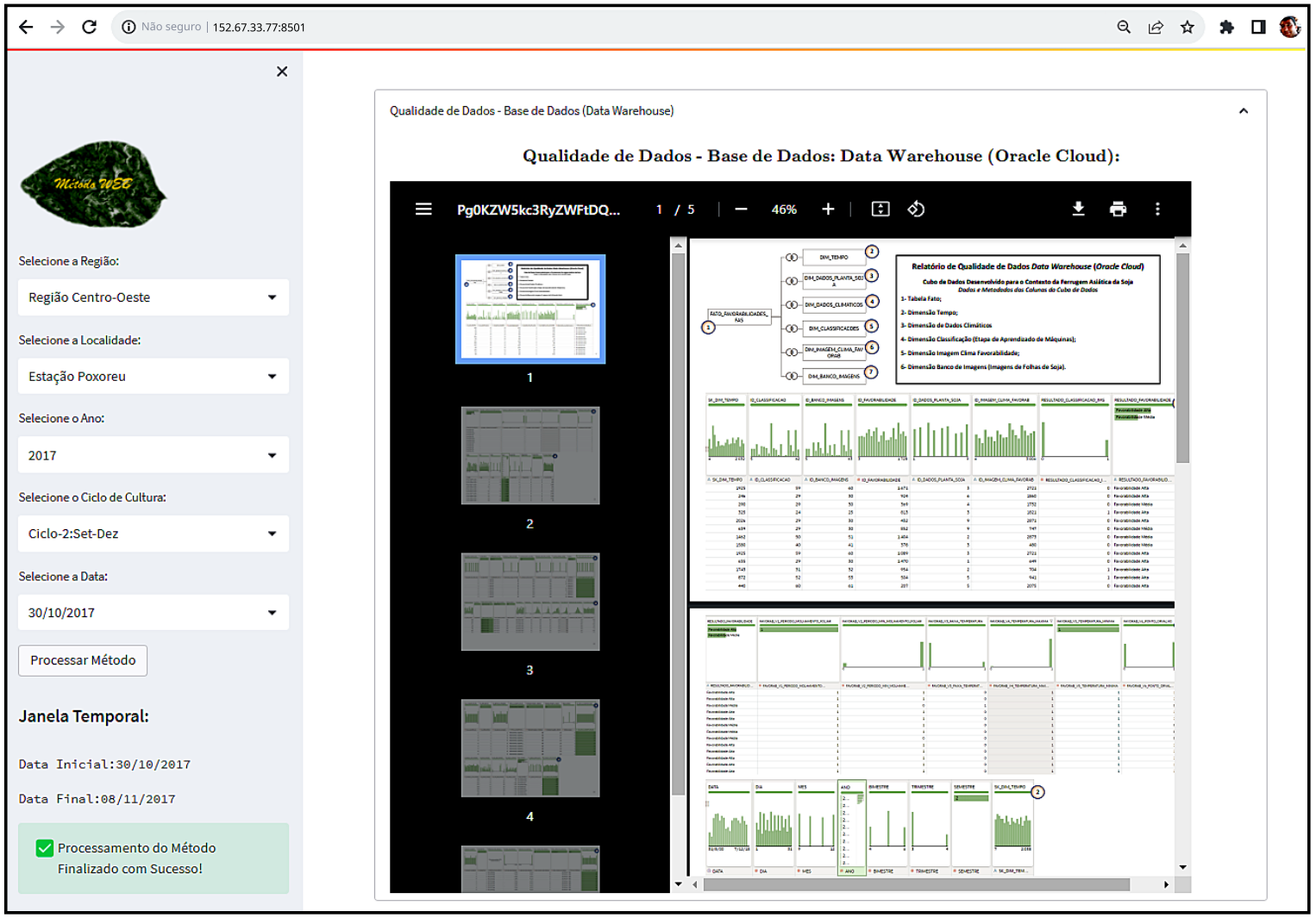
Figura 74 – Qualidade Dados Banco Oracle

Fonte: Próprio Autor

O *Framework* de Qualidade de Dados foi desenvolvido, tanto em seu modelo de concepção, quanto na aplicação prática, ou seja, utilizando a interface web. Além disso, mostrou-se também o modelo de qualidade de dados automatizado, via Plataforma Oracle *Cloud*, considerando o uso do Software *Analytics* para detecção de *insights* de qualidade de dados, o que permitiu a identificação e correção dos dados em colunas das tabelas, no banco de dados Oracle.

A quarta aba, (Figura 76), é considerada para as recomendações para tomada de decisão sobre aspectos relacionados à doença, considerando o resultado da favorabilidade obtida pelo processamento do sistema, as recomendações de fungicidas registrados no Sistema de Agrotóxicos Fitossanitários Agrofit (2023), e as orientações de boas práticas de manejo da soja, sob a perspectiva agronômica.

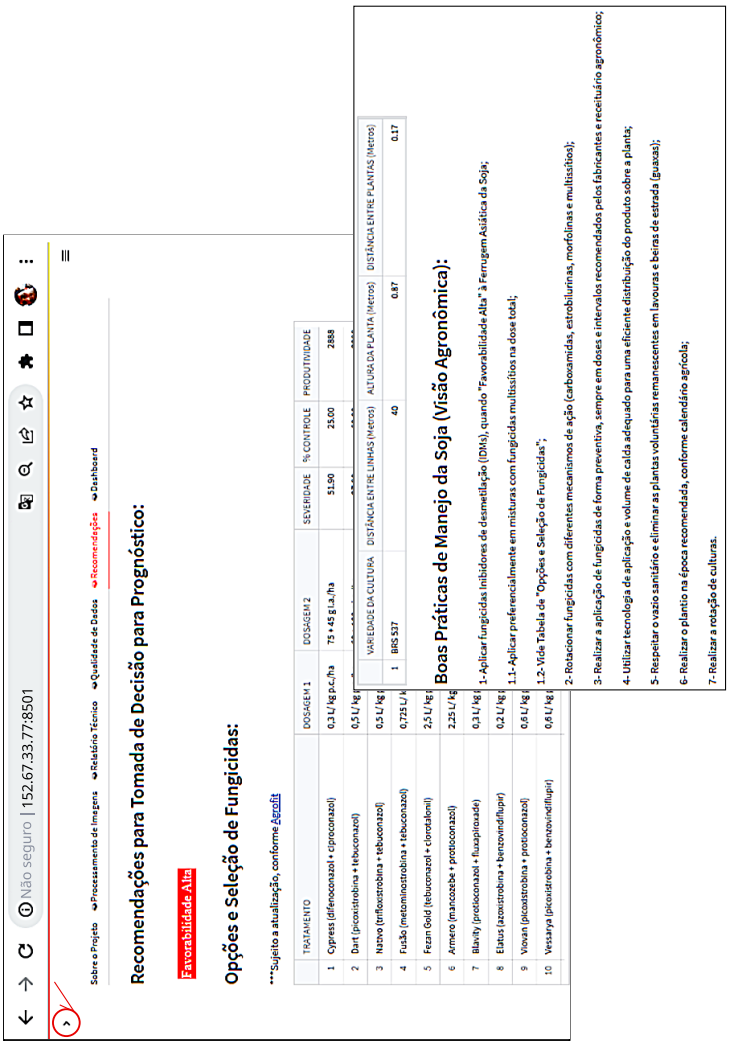
Figura 75 – Interface de Qualidade de Dados: *Data Warehouse*



Fonte: Próprio Autor

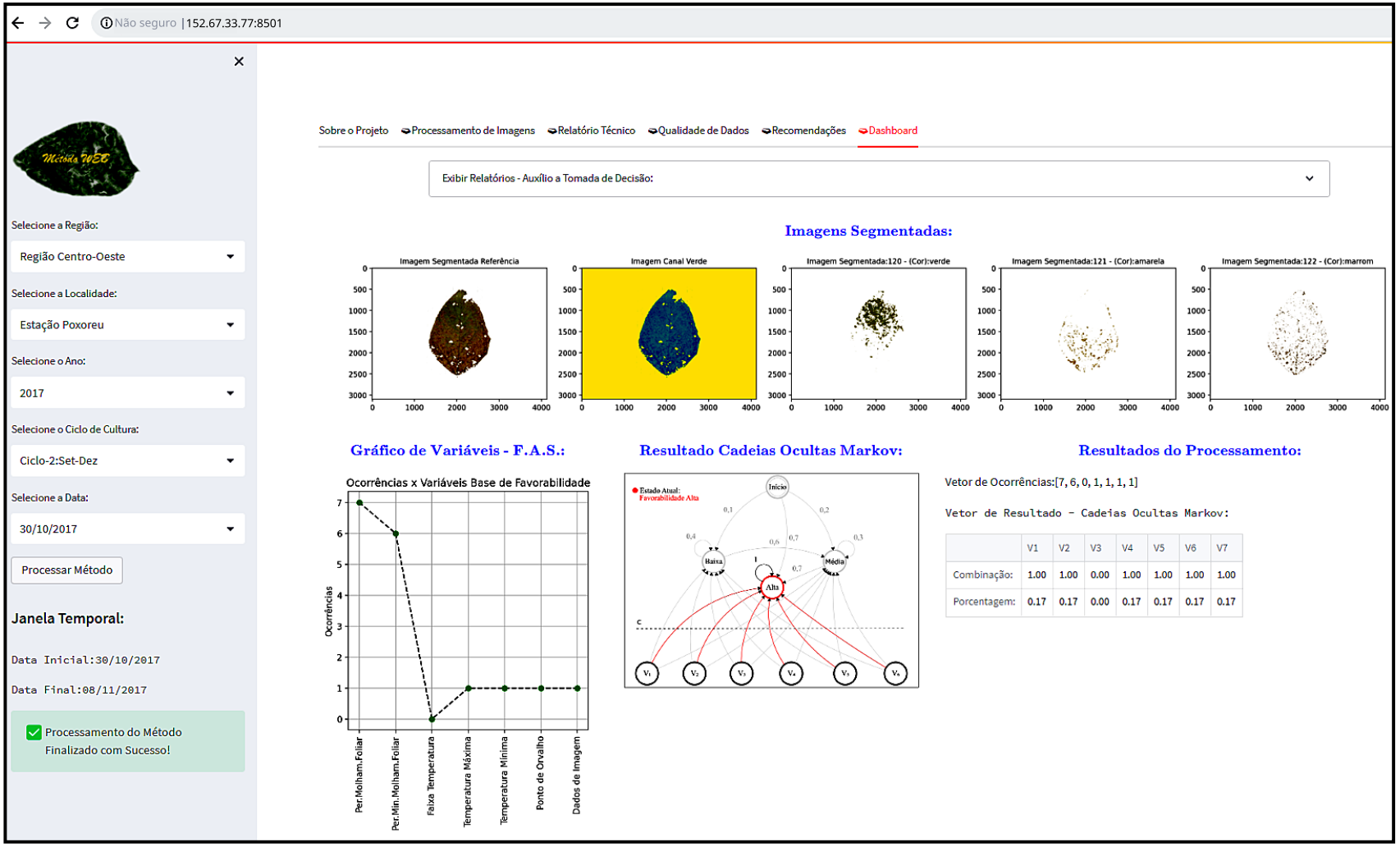
Completando as abas desta interface, uma quinta aba, (Figura 77), é também considerada para o painel *Dashboard*, que sintetiza as informações do processamento realizado. Isso incluiu as imagens segmentadas nas etapas I e II, o gráfico das variáveis climáticas, os resultados da fusão dos dados e da favorabilidade, além de um *container* contendo os relatórios para auxiliar na tomada de decisão, preparados com base no cubo de dados históricos do *Data Warehouse*. Os relatórios (*Assunto 1*), (*Assunto 2*) e (*Assunto 3*) são exibidos em diferentes *containers*, na parte superior do painel *dashboard*, o que agrega valor em análises a serem realizadas pelos usuários. No âmbito do exemplo considerado, as Figuras 78, 79 e 80 ilustram os relatórios que foram gerados no âmbito dessas funcionalidades. Esses dados exibidos ficam disponíveis para download e apoiam a tomada de decisão em relação aos aspectos de investigação da influência das variáveis climáticas na favorabilidade da ferrugem asiática da soja, da contabilização da favorabilidade baixa, média e alta por ano; e da influência da imagem da folha de soja na favorabilidade da FAS.

Figura 76 – Interface de Relatório de Recomendações



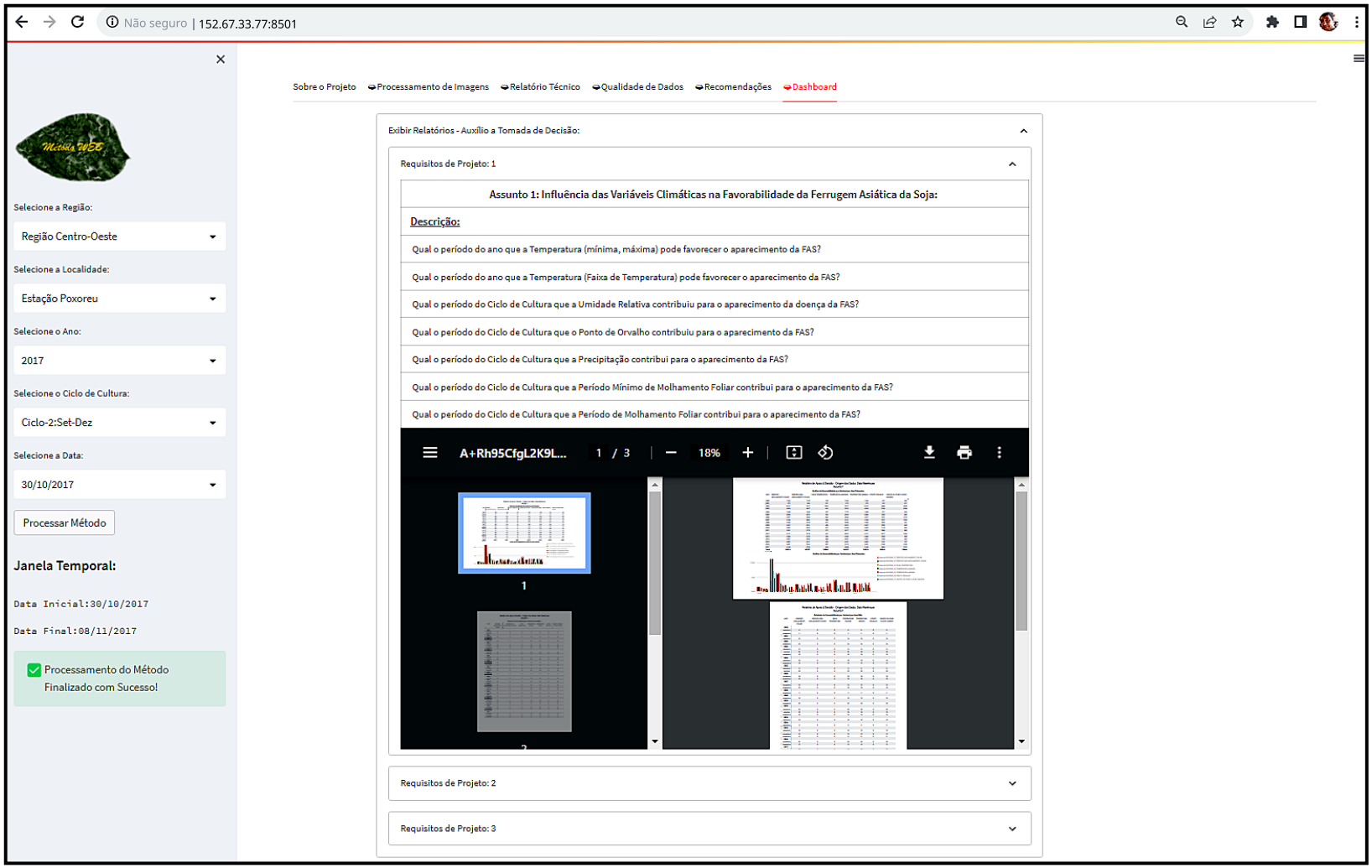
Fonte: Próprio Autor

Figura 77 – Interface do *Dashboard*



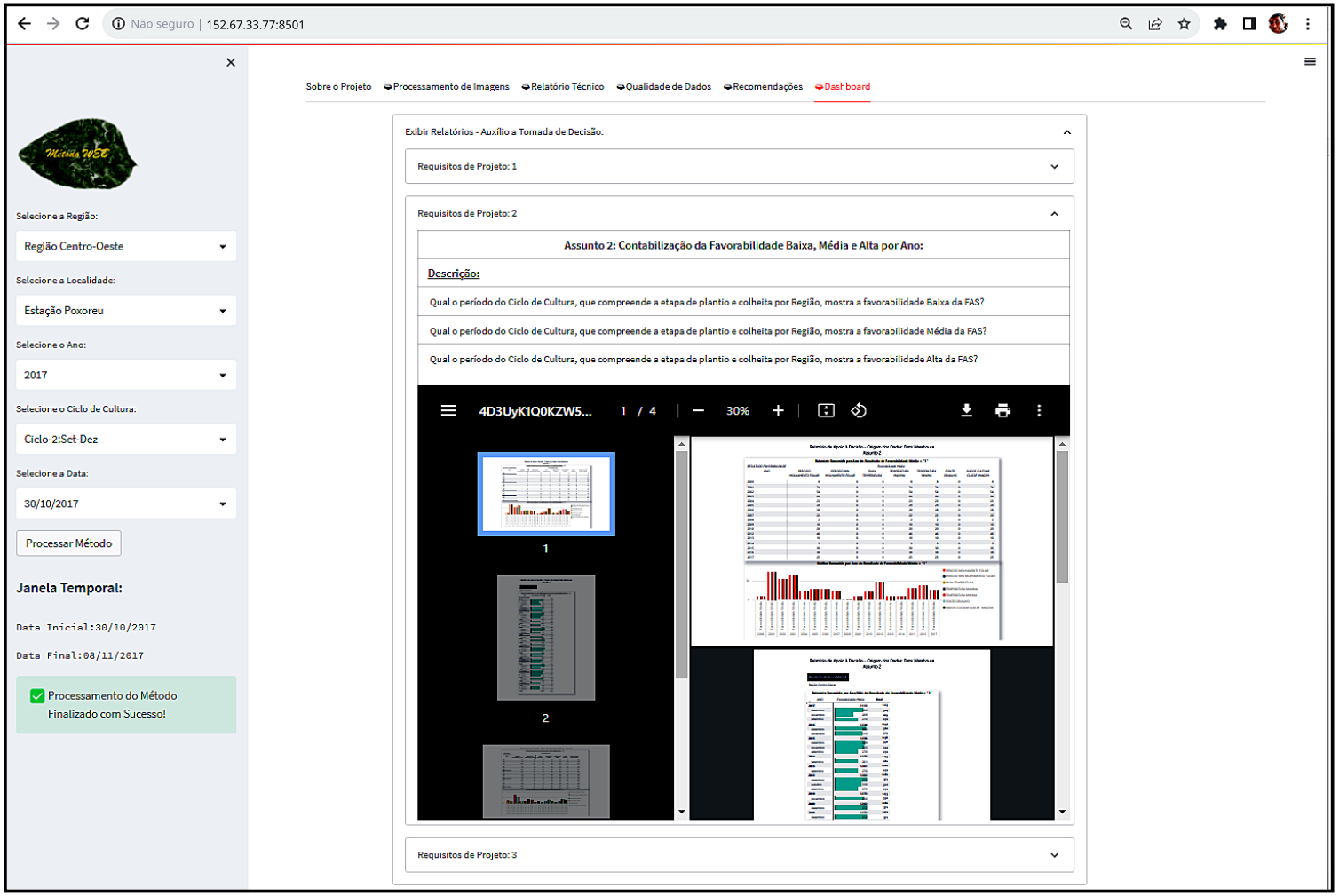
Fonte: Próprio Autor

Figura 78 – Interface do *Dashboard* - *Assunto 1*



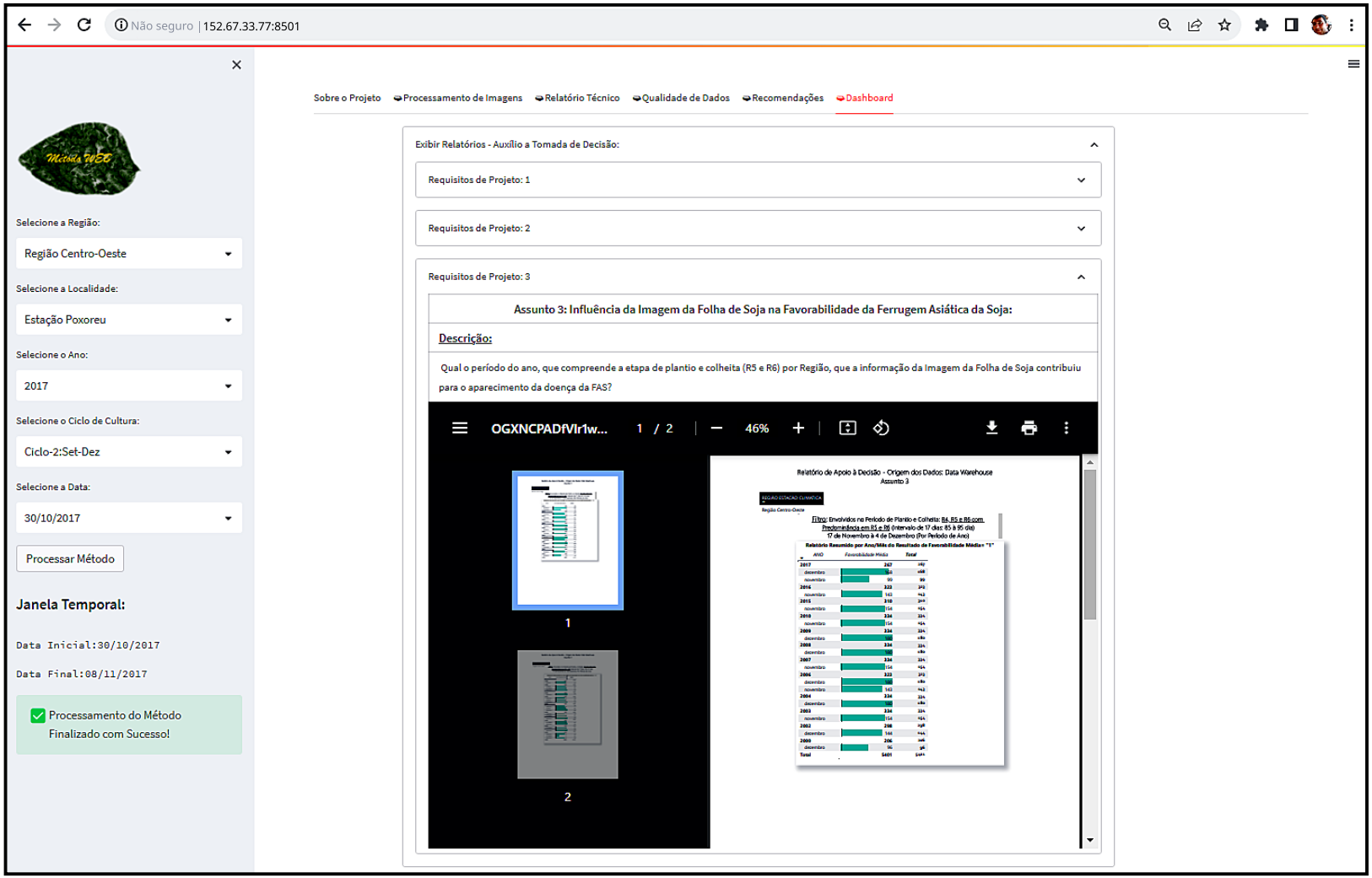
Fonte: Próprio Autor

Figura 79 – Interface do *Dashboard* - *Assunto 2*



Fonte: Próprio Autor

Figura 80 – Interface do *Dashboard* - *Assunto 3*



Fonte: Próprio Autor

4.7 Validação do Sistema a partir de Visão Especialista

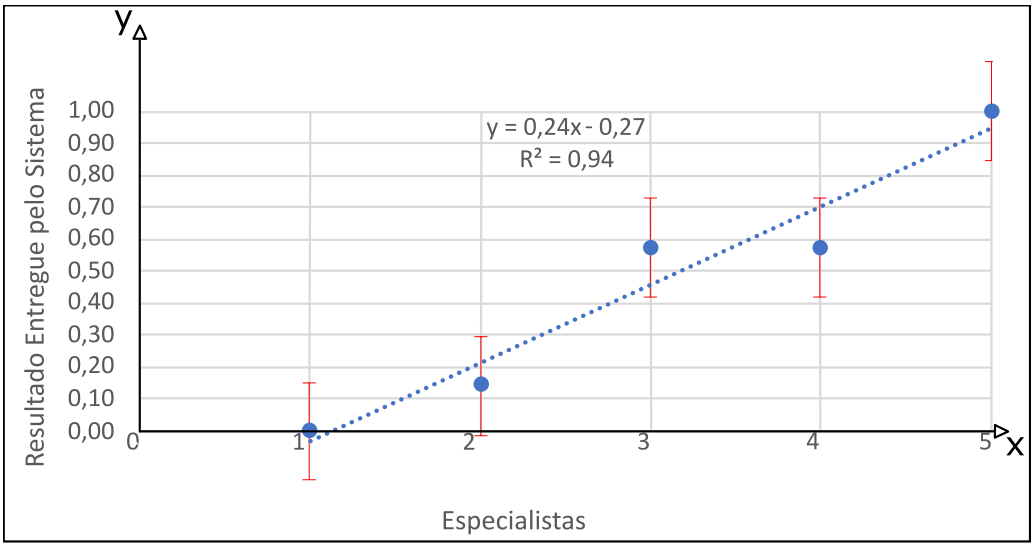
A validação do sistema foi realizada, tomando em conta a visão de cinco especialistas da área agronômica e da fitopatologia relacionada às doenças da soja, com ênfase na ferrugem asiática.

Para tanto, foi elaborado, para avaliação dos especialistas, um questionário contendo treze diferentes situações de ocorrências, observadas em área de cultura da soja, bem como quadros respectivos às informações climáticas e imagens digitais das folhas da cultura. Esse arranjo possibilitou, aos especialistas consultados, a verificação da presença ou não da ferrugem asiática, bem como sobre o seu estágio de severidade quando aplicável (conforme apresentado no Apêndice D).

Os testes submetidos aos especialistas também foram processados no sistema desenvolvido. Foram consideradas como referências as respostas dos especialistas e o respectivo percentual de acerto em relação às mesmas pelo sistema. Adicionalmente, de forma a sistematizar as respostas em um único conjunto de dados, foi realizada a normalização, considerando os valores máximos e mínimos do conjunto de respostas.

A Figura 81 apresenta o resultado em relação ao reconhecimento da presença ou a ausência da ferrugem asiática na soja, cujo R^2 calculado é igual a 0,94. A Figura 82 apresenta o resultado da validação sob o reconhecimento do índice de severidade da ferrugem asiática, nos testes avaliados ($R^2 = 0,88$).

Figura 81 – Validação Quanto à Presença ou Não de Ferrugem Asiática

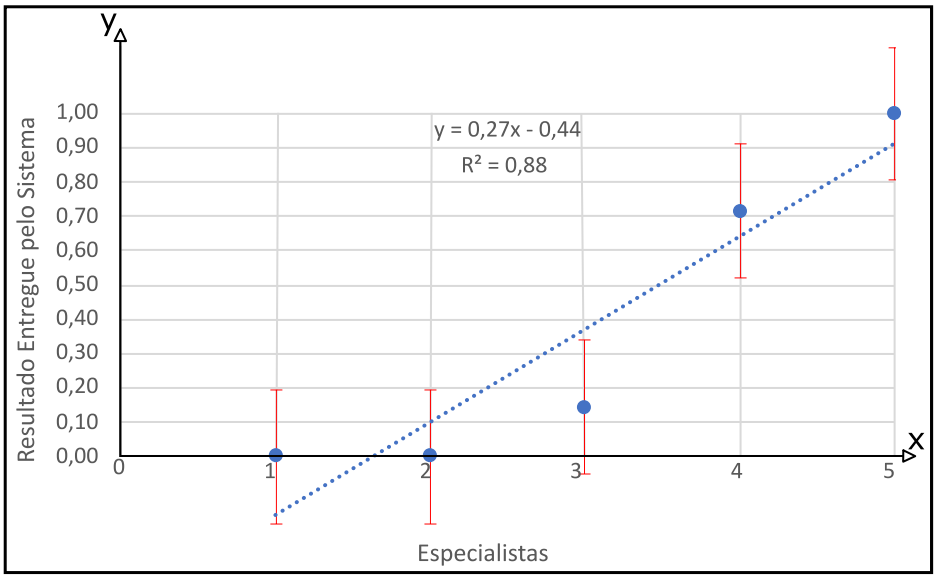


Fonte: Próprio Autor

Esses resultados expressam a quantidade da variância dos dados, os quais foram explicados utilizando o modelo linear. Logo, os altos valores de R^2 encontrados explicam

que o sistema desenvolvido atendeu satisfatoriamente em relação às respostas que foram atribuídas pelos especialistas consultados.

Figura 82 – Validação Quanto ao Nível de Severidade da Ferrugem Asiática



Fonte: Próprio Autor

4.8 Limitações e Dificuldades Encontradas no Desenvolvimento do Trabalho

A oportunidade de trabalhar na plataforma de nuvem da *Oracle* proporcionou novas experiências quanto ao uso de tecnologias computacionais e suas integrações no sistema construído. Junto a tais experiências, vale ressaltar as dificuldades associadas, as quais se referem, principalmente, ao período inicial do desenvolvimento do trabalho, quando a plataforma Oracle estava em fase de introdução no mercado. Diante desse fato, a plataforma passou por muitas melhorias implementadas durante o andamento dos trabalhos, o que ocasionou frequentes desatualizações na documentação de referência para consulta de cada tecnologia utilizada. Além disso, alguns recursos da nuvem ainda não estavam maduros, dentre eles podem ser citados bloqueios no ambiente de *Data Science*, os quais não permitiram acessos externos na instância configurada, o que não permitiu o uso da implementação do *framework* web. Nesse caso, foi necessário a criação de uma instância de computação utilizando *Oracle* Linux para implementar o ambiente computacional adequado, a fim de proporcionar a integração, via acesso externo, das tecnologias *Python*, do *framework* web e do *Autonomous Databases*.

Outra dificuldade constatada foi ainda em relação a plataforma *Oracle*, ou seja, na etapa de Transformação de Dados (*Autonomous Databases*), apresentando erros durante

o processo de preparação de dados, não permitindo a conclusão do mesmo. O mal funcionamento foi solucionado pela equipe de desenvolvimento da *Oracle*, por meio do relato do problema, via suporte técnico.

Adicionalmente, outras dificuldades podem ser relatadas, principalmente àquelas relacionadas ao desenvolvimento dos algoritmos. A primeira se referiu à retirada do fundo complexo das imagens, o que necessitou combinar diferentes abordagens de segmentação para que o conjunto de cores de fundo pudesse ser retirado com eficiência. Como segundo ponto de dificuldade nos algoritmos, a elaboração da base de regras da abordagem difusa provocou resultados não satisfatórios quanto à qualificação da ocorrência da FAS. Essa etapa de desenvolvimento demandou muitas horas de dedicação, apesar dos resultados não satisfatórios mencionados. Ainda, quanto às abordagens relacionadas à fusão de dados, é possível citar dificuldades encontradas para a implementação dos métodos de análise. Quanto à Figura de Mérito, a qual envolveu a utilização de bibliotecas gráficas associadas a conceitos matemáticos, houve dificuldades para o estabelecimento de métricas para normalização das variáveis envolvidas no processo de decisão. Quanto ao uso de lógica difusa, o estabelecimento da base de regras, dado a dinâmica do processo, demandou aprofundamento sobre os conceitos para a aplicação no problema customizado, ou seja, levando a necessidade do estabelecimento de um grande número de regras para o pleno atendimento do processo. Quanto ao uso de cadeias ocultas de Markov, as principais dificuldades envolveram o pleno entendimento do método e a sua customização para o atendimento e a avaliação da fenomenologia envolvida no entendimento do desenvolvimento da ferrugem asiática em cultura de soja. As dificuldades que envolveram as abordagens de fusão de dados foram resolvidas em uma série de reuniões técnicas desenvolvidas com o Orientador.

4.9 Considerações Finais

Este Capítulo apresentou os resultados e discussões sobre o desenvolvimento e a validação do sistema sob os aspectos desenvolvidos em sua totalidade, em ambiente de nuvem da Oracle, de forma integrada, a partir de uma nova abordagem desenvolvida para avaliação e qualificação da favorabilidade da ocorrência de ferrugem asiática da soja. Os resultados também foram apresentados considerando visualização em interfaces web, incluindo relatórios para o auxílio à tomada de decisão. Os resultados produzidos, tais como modelos de bancos de dados, *scripts* na linguagem SQL, imagens e vídeos referentes ao ambiente da Oracle *Cloud*, a tabela da cadeia oculta de Markov e os códigos-fonte do desenvolvimento do sistema estão disponíveis, na íntegra, no repositório online: <https://github.com/ricardo-a-neves/doutorado-UFSCar>.

O próximo Capítulo trata sobre as conclusões obtidas a partir dos resultados e discussões realizadas.

Conclusões

Neste capítulo, são apresentadas as conclusões, tendo em vista os resultados obtidos e as discussões realizadas, inclusive as relacionadas à validação com base em conhecimento especializado, também as referentes ao desenvolvimento do sistema de visão e inteligência computacional em ambiente de nuvem para gestão de risco da ferrugem asiática na cultura da soja.

Como visão geral das conclusões, foram priorizados quatro destaques, em que considerou-se os aspectos do funcionamento do sistema, etapas que compreendem o processamento das imagens digitais e o aprendizado de máquina, organização das séries temporais de dados, modelo de decisão para fusão de variáveis, *dashboard* e validação com base em visão especialista.

O primeiro destaque se refere ao funcionamento do sistema que pôde ser demonstrado, utilizando-se de um *framework* web para a linguagem *Python*, denominado *Streamlit*. Nesse sentido, a pilha de tecnologias disponíveis na nuvem permitiu integrar diferentes recursos, viabilizando conclusões sobre as múltiplas visualizações, ora via relatórios em tela ou analíticos, compilados em formato *Portable Document Format* (PDF), ora em painel *dashboard*.

O segundo destaque se refere a conclusões relacionadas às etapas que compreenderam o processamento das imagens, tais como a segmentação, extração de características e reconhecimento de padrões, assim como as etapas de redução de dimensionalidade e classificação, via aprendizado de máquina. Também, apresenta conclusões sobre os resultados armazenados em banco de dados Oracle, bem como sobre a utilização de um vetor característico das variáveis utilizadas no modelo de decisão, realizado via interface web para o processo de fusão das variáveis.

No terceiro destaque, são consideradas conclusões sobre os relatórios do *Data Warehouse* visualizados, estabelecidos previamente em formato PDF, em conjunto com os dados exibidos via painel *dashboard*. Esses relatórios retratam uma visão histórica das séries temporais de dados e os resultados produzidos pelo sistema, em consonância com

os requisitos utilizados na construção do *Data Warehouse*.

O quarto destaque envolve a validação do sistema desenvolvido, a partir da consulta a especialistas da agronomia, entomologia e fitopatologia, considerando a avaliação de correlação cruzada dessa visão especializada e resultados entregues pelo sistema desenvolvido.

Quanto ao primeiro destaque, a estruturação elaborada integrou tecnologias na obtenção de fontes de origem interna e externa, por meio da Internet, considerando a arquitetura implementada na Oracle *Cloud*, a partir do cenário 3, entre os demais avaliados. Vale ressaltar que a arquitetura em nuvem implementada permitiu a integração entre todos os serviços envolvidos. Nesse contexto, o uso de estruturas preparadas para absorver dados estruturados, semiestruturados e não estruturados, tais como o *Data Lake* e *Data Warehouse*, o que viabilizou a utilização de dados *Big Data* em um ambiente para tomada de decisão, destinado às análises de favorabilidade da ocorrência de ferrugem asiática da soja, em área real de cultivo. Adicionalmente, foi possível concluir que o uso de *b-spline* cúbica foi a melhor escolha para as etapas de interpolação e completude do conjunto de dados, conforme indicaram os resultados.

Quanto ao segundo destaque, em relação ao pré-processamento de imagens, as técnicas de equalização do histograma e do filtro de mediana foram aplicadas no canal verde (560 nanômetros ± 20 nanômetros), viabilizando, respectivamente, a equalização e a melhoria da relação Sinal/Ruído nas imagens de folhas da soja. Quanto aos limiares, uma vez selecionado o canal verde, foi possível estabelecer a seguinte região de interesse $31 \leq limiar \leq 165$, o que viabilizou a utilização de um valor médio de intensidade da ordem de 128, indicando uma melhor distribuição de valores para os *pixels* das imagens consideradas. Adicionalmente, essas operações viabilizaram a ampliação da faixa de valores de intensidade, ou seja, de 0 a 255, o que também aumentou o valor da medida modal da ordem de 29 para valores da ordem de 189, comprovando uma melhor redistribuição, ao longo das imagens, dos valores de intensidades. Também, a aplicação da filtragem mitigou interferências de ruído de baixa frequência, viabilizando melhorias na relação Sinal/Ruído (SNR ≥ 90), para o conjunto de imagens tratadas, preparando as mesmas para as etapas de processamento.

Quanto à etapa de segmentação, foi possível concluir que a combinação das técnicas de limiarização e *k-means* proporcionou a adequada separação dos objetos de interesse dos diferentes fundos, mesmo quando esses apresentaram cores próximas ao dos objetos de interesse nas imagens. O uso da técnica *k-means*, nas imagens que foram limiarizadas no pré-processamento, permitiu o agrupamento de pixels que apresentaram características semelhantes, o que definitivamente viabilizou a separação do fundo complexo dos objetos de interesse nas imagens. Nesse contexto a segmentação para as cores verde, amarela e marrom puderam ser realizadas e, para as mesmas, os seguintes valores de PSNR e MSE puderam ser observados: $14,01 \leq PSNR \leq 14,92$; $0,03 \leq MSE \leq 0,04$.

Adicionalmente, com a utilização combinada das técnicas de extração de características SIFT, HOG e momentos invariantes de Hu, foi possível ampliar a robustez para o reconhecimento de padrões nas imagens processadas. Nesse contexto, a extração das características de cor, textura e forma para os objetos de interesse nas imagens puderam ser analisados.

A técnica SIFT demonstrou ser eficaz na extração de informações sobre a distribuição das cores verde, amarelo e marrom, nas imagens digitais das folhas de soja. Quanto à sua aplicação, foi possível concluir que a suavização das imagens baseada na customização de filtros Gaussianos, com valores de desvio padrão no intervalo $1,6 \leq \sigma \leq 3,2$, viabilizou equilibrar as operações de subtração para as diversas oitavas consideradas. Adicionalmente, a configuração do contraste com sensibilidade ajustada para 4% auxiliou na detecção de extremos (máximo e mínimo locais), onde a comparação da intensidade de cada *pixel* foi realizada, considerando os oito vizinhos conectados. Assim, pontos-chave (invariantes a escala, rotação e mudanças de iluminação) sobre as características de cor puderam ser caracterizados, para em seguida viabilizar o estabelecimento dos gradientes e dos histogramas das regiões consideradas. Assim, foi possível gerar satisfatoriamente descritores de cor para as imagens processadas.

Quanto à aplicação da técnica HOG, foram considerados gradientes de intensidades para as imagens, os quais foram sensíveis às mudanças de intensidade, bem como a distribuição espacial das mesmas em relação às cores, onde as variáveis θ e Θ receberam os valores de 22,5° e 2,22°, respectivamente. O benefício dessa configuração foi o oferecimento de um equilíbrio entre o nível de detalhe e a eficiência na detecção e reconhecimento dos objetos nas regiões das imagens consideradas. Adicionalmente, essas escolhas garantiram a extração das características relevantes para o reconhecimento dos padrões, sem sobrecarregar ou demandar os recursos computacionais utilizados. Adicionalmente, de forma a se obter informações sobre a textura das regiões, foi utilizada uma granularidade com blocos de 9 *pixels*. Essa configuração proporcionou um adequado balanceamento entre a extração de características e a eficiência computacional. Portanto, foi possível concluir que a customização realizada influenciou na minimização da complexidade de extração das características de texturas, assim como na granularidade estabelecida para a descrição dos padrões, o que refletiu na eficácia e na robustez da etapa de reconhecimento dos padrões.

Quanto ao uso dos momentos invariantes de Hu, a técnica mostrou-se útil para a extração de características de padrões geométricos decorrentes de alterações no estágio normal das folhas de soja presentes nas imagens digitais processadas. De fato, a invariância apresentada pela técnica, quanto à translação, escala e rotação, trouxe robustez para a extração dessas características, vez que as mesmas ocorrem de forma aleatória nas folhas da soja, tanto para os estágios iniciais (geometrias circulares e elípticas), quanto para estágios mais avançados (geometrias irregulares e complexas), decorrentes de doenças que

podem ocorrer nesta cultura. Para a serie temporal de imagens digitais analisadas, foram observados, para o primeiro e segundo momento, valores medidos da ordem de $0,844 \leq$ médias $\leq 0,845$ e desvio padrão da ordem de $0,181 \leq \sigma \leq 0,176$, respectivamente. Para o terceiro, quarto, quinto, sexto e sétimo momento foram observados valores da ordem de $0,301 \leq$ médias $\leq 0,328$ e desvio padrão na ordem de $0,176 \leq \sigma \leq 0,190$, respectivamente. Nesse contexto, as análises possibilitaram concluir que o uso desses momentos permitiram avaliar com detalhe as formações geométricas que surgiram em folhas de soja e sua especificidade.

Estes descritores de HU, foram incluídos em um vetor de características, conjuntamente, com aqueles obtidos com as técnicas SIFT e HOG, permitindo caracterizar informações sobre padrões das imagens digitais de folhas de soja, de forma a contribuir com os demais dados das series temporais climáticas para a identificação ou não da presença da ferrugem asiática em cultura de soja e seu estágio de infestação.

Quanto à aplicação de PCA para a redução de dimensionalidade do vetor de características extraídas das imagens, com 128 componentes para 19 componentes, foi possível concluir que não houve perda significativa de informação, vez que a seleção de 19 componentes juntas explicou 70,79% da variância do conjunto dos dados obtidos com as técnicas SIFT, HOG e HU. De fato, a redução de dimensionalidade foi benéfica porque pôde simplificar os modelos de aprendizado de máquina que foram desenvolvidos, também reduzir o tempo de treinamento, diminuindo a necessidade adicional de recursos computacionais e aumentando o desempenho global dos algoritmos avaliados. Além disso, foi possível concluir que, utilizando o método de valores mínimos e máximos, para a normalização dos dados originados, no intervalo de 0 a 1, garantiu robustez com a devida proporcionalidade necessária para a classificação dos mesmos com SVM, vez que estes classificadores são sensíveis à escala dos dados.

Com relação à seleção do classificador, relacionado à etapa de aprendizado de máquina, foi possível concluir que o classificador SVM foi o que apresentou o melhor resultado, sendo, portanto, o selecionado. Esse classificador foi predominantemente melhor em relação aos demais classificadores avaliados, ou seja, quando comparado ao uso de classificadores baseados em Árvore de Decisão, *k-Nearest Neighbors* e o *Naïve Bayes*. De fato, para o classificador selecionado, foram avaliadas suas versões binário e multiclasses. A versão SVM binária apresentou melhor acurácia, com 45,9% melhor em relação ao multiclasse. Também, 10,5% menor em erro médio quadrático e 16,5% maior em área sob a curva ROC. Ao considerar a fenomenologia do problema tratado, foi possível concluir que o *kernel* polinomial de terceiro grau apresentou o melhor resultado quando considerada a relação 80-20%, respectivamente, para as etapas de treinamento e teste. As métricas para o conjunto de características do banco de imagens de folhas de soja ficaram em torno de 0,85 de acurácia, áreas sob a curva ROC com valores $\geq 0,91$ e erro médio quadrático com valores $\leq 0,15$. A seleção do classificador SVM foi adequada, visto que o mesmo

mostrou-se capaz de fornecer decisões de classificação robustas e resistentes a eventuais sobreajustes.

Outro aspecto a ser considerado é quanto ao uso de bancos de dados Oracle, os quais trouxeram flexibilidade ao ambiente de nuvem, permitindo a integração nos diversos níveis de processamento do sistema, que compreenderam desde a entrada de dados via rede pública, internet, até a entrada dos dados customizados como resultados das diversas etapas do sistema, via rede privada. Assim, foi possível concluir também que as principais estruturas de dados que se integraram ao Oracle foram os *buckets*, os quais desempenharam a funcionalidade do *Data Lake* de forma simplificada, mas eficiente. Esses *buckets* permitiram armazenar de forma customizada os dados não estruturados, ou seja, resultados dos processamentos das imagens das folhas de soja gerados pelos algoritmos no ambiente de *Data Science*, e também dados estruturados como carga no *Data Warehouse* para fins de geração dos dados históricos e relatórios analíticos, bem como permitiu uma dinâmica eficiente e rápida, via códigos SQL, na elaboração do vetor de dados de entrada para o algoritmo de fusão de variáveis com processamento em tempo real.

Quanto à fusão das variáveis, foi possível concluir que, comparativamente, o uso de Cadeias Ocultas de Markov, customizadas para a avaliação da ocorrência da ferrugem asiática em soja, foi melhor que o uso de Figura de Mérito, ou ainda em relação ao uso de técnica baseada em Lógica Difusa. De fato, para os diferentes cenários considerados, ou seja, relacionados às favorabilidades baixa, média ou alta, para a ocorrência da doença, esses métodos responderam em desempenhos com percentuais da ordem de 100%, 64% e 48% respectivamente. Ademais, o uso de Cadeias Ocultas de Markov, customizadas para fundir variáveis que definem a ocorrência ou não da ferrugem asiática e seu estágio de severidade em plantas de soja, mostrou-se adequado para integrar diferentes grandezas físicas ao longo de janelas em séries temporais de dados, permitindo a avaliação precisa da doença com confiabilidade e robustez.

Quanto ao terceiro destaque, foi possível concluir que as interfaces web contribuíram para a apresentação dos resultados do sistema, proporcionando organização e também uma visualização intuitiva dos relatórios analíticos e de recomendações via *dashboard*. Também, tendo em vista que a construção da interface de visualização que considerou processos distintos, relacionados ao fluxo de dados, levou a concluir que, em função de cada situação de favorabilidade sobre o estágio da ferrugem asiática em área de cultivo, foi possível viabilizar o planejamento de ações de controle. Outro aspecto a ser considerado nesse contexto foi que o desenvolvimento da interface, contemplando um *Framework* de Qualidade de Dados, viabilizou considerar um conjunto de indicadores de qualidade, bem como suas dimensões correspondentes sobre o fluxo do processamento, garantindo robustez na entrega de resultados de análises sobre os níveis de severidade da doença, nas áreas de cultivo.

Quanto ao quarto destaque, com base nos resultados da validação do sistema junto

aos cinco especialistas da área agronômica e da fitopatologia, concluiu-se que o sistema demonstrou uma performance robusta e confiável. As análises revelaram que os índices de acerto, medidos através do coeficiente de determinação R^2, foram significativos, alcançando $R^2 = 0,94$ para o reconhecimento da presença ou ausência da doença e $R^2 = 0,88$ para o reconhecimento do índice de severidade da ferrugem asiática. Portanto, é possível concluir que o sistema não apenas conseguiu replicar com precisão as avaliações dos especialistas, mas também demonstrou uma capacidade robusta de identificação e classificação da ferrugem asiática na soja. Esses resultados validam o sistema como uma ferramenta eficaz de suporte à decisão para o monitoramento e diagnóstico preciso da ferrugem asiática, contribuindo significativamente para o manejo integrado de doenças nas lavouras de soja.

Desta forma, com base nas conclusões parciais, apresentadas nos destaques considerados, é possível, conclusivamente, afirmar que o sistema de inteligência e visão computacional, em ambiente de nuvem viabiliza o monitoramento da presença ou não da ferrugem asiática da soja, bem como a avaliação da dinâmica de ocorrência da doença, em seus diferentes estágios de severidade e risco para o processo agrícola produtivo.

Trabalhos Futuros

1. Automatizar os processos de visão e inteligência computacional do sistema, os quais foram executados de forma supervisionada, fazendo uso de redes neurais com desdobramentos na utilização de redes convolucionais;

2. Automatizar o processo de recebimento dos dados a partir de novas conexões com a nuvem, via Internet, como servidores legados e Veículos Aéreos Não Tripulados (VANTs), possibilitando que o *Data Lake* possa direcionar, de forma automática, esses dados na infraestrutura de dados;

3. Ampliar o modelo desenvolvido para que possa operar com um número maior de variáveis de entrada, de forma a viabilizar predição e o uso de dados relacionados aos aspectos de fenótipo e genótipo das plantas de soja, a fim de possibilitar a customização da qualificação da favorabilidade relacionada a ocorrência da ferrugem;

4. Considerar, para novas versões do sistema, experimentações com a técnica *Manifold learning*, haja vista que trabalha com geometria não linear, e comparar aos resultados obtidos pelo PCA para a redução de dimensionalidade;

5. Buscar por novas abordagens de classificadores, na linha de aprendizado de máquinas, e efetuar comparações diante das abordagens utilizadas com o objetivo de melhoria na performance dos resultados de classificação;

6. Investigar se há abordagens padrão-ouro que podem ser utilizadas para comparação à abordagem desenvolvida, tanto na área computacional, quanto na área agrícola.

Referências

ABDI, H.; WILLIAMS, L. J. Principal component analysis. **WIREs Computational Statistics**, v. 2, n. 4, p. 433–459, 2010. Citado na página 114.

ABDU, A. M.; MOKJI, M. M.; SHEIKH, U. U. A pattern analysis-based segmentation to localize early and late blight disease lesions in digital images of plant leaves. In: IEEE. **2019 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)**. [S.l.], 2019. p. 116–121. Citado na página 81.

ABDUL-RAHMAN, M. et al. A framework to simplify pre-processing location-based social media big data for sustainable urban planning and management. **Cities**, Elsevier, v. 109, p. 102986, 2021. Citado na página 61.

AGROFIT. **Agrofit**. [S.l.]: Ministério da Agricultura, Pecuária e Abastecimento, 2023. Citado 2 vezes nas páginas 199 e 214.

ALAOUI, I. E.; GAHI, Y.; MESSOUSSI, R. Big data quality metrics for sentiment analysis approaches. In: **Proceedings of the 2019 International Conference on Big Data Engineering**. [S.l.: s.n.], 2019. p. 36–43. Citado 3 vezes nas páginas 62, 63 e 64.

ALEKSEEV, A. et al. Efficient data management tools for the heterogeneous big data warehouse. **Physics of Particles and Nuclei Letters**, Springer, v. 13, n. 5, p. 689–692, 2016. Citado 2 vezes nas páginas 52 e 55.

ALVES, G. M. Método de reconstrução tomográfica de amostras agrícolas com o emprego de técnicas big data. Universidade Federal de São Carlos, 2020. Citado na página 57.

ALVES, G. M.; CRUVINEL, P. E. Parallel and distributed processing for high resolution agricultural tomography based on big data. **Multimedia Tools and Applications**, Springer, p. 1–32, 2023. Citado na página 57.

AMGHAR, S.; CHERDAL, S.; MOULINE, S. Data integration and nosql systems: A state of the art. In: **Proceedings of the 4th International Conference on Big Data and Internet of Things**. [S.l.: s.n.], 2019. p. 1–6. Citado na página 55.

ANAGHA, P.; BASKAR, A. An automatic histogram detection and information extraction from document images. **International Journal of Speech Technology**, Springer, v. 24, n. 1, p. 77–85, 2021. Citado na página 66.

ANG, K. L.-m.; SENG, J. K. P. Big data and machine learning with hyperspectral information in agriculture. **IEEE Access**, IEEE, 2021. Citado na página 61.

APACHE SOFTWARE FOUNDATION (ASF). **Apache Project List**. 2021. Disponível em: <https://www.apache.org/index.html#projects-list>. Acesso em: 15 fev. 2021. Citado na página 55.

ARAUJO, J. M. M.; PEIXOTO, Z. M. A. A new proposal for automatic identification of multiple soybean diseases. **Computers and Electronics in Agriculture**, Elsevier, v. 167, p. 105060, 2019. Citado 2 vezes nas páginas 50 e 72.

ARDAGNA, D. et al. Context-aware data quality assessment for big data. **Future Generation Computer Systems**, Elsevier, v. 89, p. 548–562, 2018. Citado na página 64.

AYGÜN, S. et al. Sensor fusion for iot-based intelligent agriculture system. In: IEEE. **2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics)**. [S.l.], 2019. p. 1–5. Citado na página 87.

BAHRY, C. A. et al. Analysis of combined strategies for the management of asian soybean rust. **African Journal of Plant Science**, Academic Journals, v. 14, n. 8, p. 297–307, 2020. Citado 2 vezes nas páginas 44 e 46.

BAI, X. Morphological feature extraction for detail maintained image enhancement by using two types of alternating filters and threshold constrained strategy. **Optik**, Elsevier, v. 126, n. 24, p. 5038–5043, 2015. Citado na página 82.

BARBEDO, J. G. A. A review on the main challenges in automatic plant disease identification based on visible range images. **Biosystems engineering**, Elsevier, v. 144, p. 52–60, 2016. Citado na página 72.

BARBEDO, J. G. A. et al. Annotated plant pathology databases for image-based detection and recognition of diseases. **IEEE Latin America Transactions**, IEEE, v. 16, n. 6, p. 1749–1757, 2018. Citado na página 91.

BASSOI, L. H. et al. Agricultura de precisão e agricultura digital. **Embrapa Pecuária Sudeste-Artigo em periódico indexado (ALICE)**, TECCOGS, n. 20, jul./dez., 2019., 2019. Citado 2 vezes nas páginas 59 e 65.

BATINI, C. et al. From data quality to big data quality. In: **Big Data: Concepts, Methodologies, Tools, and Applications**. [S.l.]: IGI Global, 2015. p. 60–82. Citado 2 vezes nas páginas 63 e 64.

BAUM, L. E. An inequality and associated maximization technique in statistical estimation for probabilistic functions of a markov process. **Inequalities**, v. 3, p. 1–8, 1972. Citado na página 124.

BAUM, L. E.; EAGON, J. A. An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. 1967. Citado na página 134.

BAUM, L. E.; PETRIE, T. Statistical inference for probabilistic functions of finite state markov chains. **The annals of mathematical statistics**, JSTOR, v. 37, n. 6, p. 1554–1563, 1966. Citado na página 134.

BEDIN, E. et al. Aplicações foliares de cobre no manejo da ferrugem-asiática da soja. Universidade de Passo Fundo, 2018. Citado na página 47.

BEGOLI, E. A short survey on the state of the art in architectures and platforms for large scale data analysis and knowledge discovery from data. In: **Proceedings of the WICSA/ECSA 2012 Companion Volume**. [S.l.: s.n.], 2012. p. 177–183. Citado 2 vezes nas páginas 53 e 55.

BENDINI, H. d. N. et al. Risk analysis of black sigatoka occurrence based on polynomial models: a case study. **Tropical Plant Pathology**, SciELO Brasil, v. 38, p. 35–43, 2013. Citado na página 124.

BERG, B. A. **Markov chain Monte Carlo simulations and their statistical analysis: with web-based Fortran code**. [S.l.]: World Scientific Publishing Company, 2004. Citado 2 vezes nas páginas 143 e 144.

BERUSKI, G. C. et al. Leaf wetness duration estimation and its influence on a soybean rust warning system. **Australasian Plant Pathology**, Springer, v. 48, n. 4, p. 395–408, 2019. Citado na página 45.

BHATNAGAR, G.; LIU, Z. Multi-sensor fusion based on local activity measure. **IEEE Sensors Journal**, IEEE, v. 17, n. 22, p. 7487–7496, 2017. Citado na página 85.

BISHOP, C. M.; NASRABADI, N. M. **Pattern recognition and machine learning**. [S.l.]: Springer, 2006. v. 4. Citado 3 vezes nas páginas 107, 108 e 109.

BOUDAREN, M. E. Y.; PIECZYNSKI, W. Dempster–shafer fusion of evidential pairwise markov chains. **IEEE Transactions on Fuzzy Systems**, IEEE, v. 24, n. 6, p. 1598–1610, 2016. Citado 2 vezes nas páginas 85 e 124.

BREIMAN, L. et al. Classification and regression trees. wadsworth & brooks. **Cole Statistics/Probability Series**, 1984. Citado na página 107.

BRESSAN, G. M. et al. Sistema de classificação fuzzy para o risco de infestação por plantas daninhas considerando a sua variabilidade espacial. **Planta daninha**, SciELO Brasil, v. 24, p. 229–238, 2006. Citado na página 124.

BRITO, A. R. d. Método para classificação de sementes agrícolas em imagens obtidas por tomografia de raios-x em alta resolução. Universidade Federal de São Carlos, 2020. Citado 2 vezes nas páginas 77 e 78.

BUITINCK, L. et al. API design for machine learning software: experiences from the scikit-learn project. In: **ECML PKDD Workshop: Languages for Data Mining and Machine Learning**. [S.l.: s.n.], 2013. p. 108–122. Citado na página 112.

CAI, L.; ZHU, Y. The challenges of data quality and data quality assessment in the big data era. **Data science journal**, Ubiquity Press, v. 14, p. 1–10, 2015. Citado 2 vezes nas páginas 63 e 65.

CANNY, J. A computational approach to edge detection. **IEEE Transactions on pattern analysis and machine intelligence**, Ieee, n. 6, p. 679–698, 1986. Citado na página 97.

CATARCI, T. et al. My (fair) big data. In: IEEE. **2017 IEEE International Conference on Big Data (Big Data)**. [S.l.], 2017. p. 2974–2979. Citado na página 63.

CELEBI, M. E.; KINGRAVI, H. A.; VELA, P. A. A comparative study of efficient initialization methods for the k-means clustering algorithm. **Expert systems with applications**, Elsevier, v. 40, n. 1, p. 200–210, 2013. Citado na página 97.

CESTNIK, B.; KONONENKO, I.; BRATKO, I. A knowledge-elicitation tool for sophisticated users. In: **Proceedings of the 2nd European Conference on European Working Session on Learning EWSL**. [S.l.: s.n.], 1987. v. 87. Citado na página 107.

CHAKI, J.; PAREKH, R. Designing an automated system for plant leaf recognition. **International Journal of Advances in Engineering & Technology**, IAET Publishing Company, v. 2, n. 1, p. 149, 2012. Citado na página 78.

CHANDRAPRABHA, K.; BHARATHI, C. Texture analysis using glcm & glrlm feature extraction methods. **International Journal for Research in Applied Science &Engineering Technology (IJRASET)**, v. 7, n. 5, p. 2059–2064, 2019. Citado na página 83.

CHING, W.-K.; NG, M. K. Markov chains. **Models, algorithms and applications**, Springer, 2006. Citado na página 135.

CHITRALEKHA, G.; ROOGI, J. M. A quick review of ml algorithms. p. 1–5, 2021. Citado na página 143.

CONAB, C. **Acompanhamento da safra brasileira Soja**. 2023. Disponível em: <https://www.conab.gov.br/info-agro/safras/serie-historica-das-safras?start=30>. Acesso em: 13 jul. 2023. Citado 3 vezes nas páginas 40, 41 e 43.

CONSÓRCIO ANTIFERRUGEM. Consórcio antiferrugem: parceria público privada no combate à ferrugem asiática da soja. 2023. Disponível em: <http://www.consorcioantiferrugem.net/#/numeros>. Acesso em: 21 set. 2023. Citado 2 vezes nas páginas 41 e 42.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine learning**, Springer, v. 20, p. 273–297, 1995. Citado na página 110.

COUCHBASE, INC. **COUCHBASE Cloud - Database-as-a-Service**. 2021. Disponível em: <https://www.couchbase.com/>. Acesso em: 24 fev. 2021. Citado na página 52.

COVER, T.; HART, P. Nearest neighbor pattern classification. **IEEE transactions on information theory**, IEEE, v. 13, n. 1, p. 21–27, 1967. Citado na página 107.

CRUVINEL, P. et al. Avaliação de modelos para estabelecimento de figura de risco de ocorrência da sigatoka-negra em bananais. **Embrapa Instrumentação-Documentos (INFOTECA-E)**, São Carlos: Embrapa Instrumentação, 2011. 28 p., 2011. Citado 3 vezes nas páginas 31, 124 e 125.

CRUVINEL, P. E. **Agricultural data for geo-computing, sensors network, and advances in IoT: quality requested in the decision making processes**. 2018. In:Panel on Quality of Data and Services: The Essences From Geo-computing, Society/Crowd Sensing, and Mobility/Service-related Sensing. Available on <site>. Citado na página 65.

_____. Advanced digital platform for agricultural risk management. In: IEEE. **2022 IEEE 16th International Conference on Semantic Computing (ICSC)**. [S.l.], 2022. p. 299–306. Citado na página 124.

CUI, D. et al. Image processing methods for quantitatively detecting soybean rust from multispectral images. **Biosystems engineering**, Elsevier, v. 107, n. 3, p. 186–193, 2010. Citado na página 48.

CUI, M. et al. Introduction to the k-means clustering algorithm based on the elbow method. **Accounting, Auditing and Finance**, Clausius Scientific Press, v. 1, n. 1, p. 5–8, 2020. Citado na página 168.

DALAL, N.; TRIGGS, B. Histograms of oriented gradients for human detection. In: IEEE. **2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)**. [S.l.], 2005. v. 1, p. 886–893. Citado 3 vezes nas páginas 82, 105 e 106.

DAVIS, P. J. **Interpolation and approximation**. [S.l.]: Courier Corporation, 1975. Citado na página 117.

DEVARAJ, A. et al. Identification of plant disease using image processing technique. In: IEEE. **2019 International Conference on Communication and Signal Processing (ICCSP)**. [S.l.], 2019. p. 0749–0753. Citado 2 vezes nas páginas 72 e 83.

DHINGRA, G.; KUMAR, V.; JOSHI, H. D. Study of digital image processing techniques for leaf disease detection and classification. **Multimedia Tools and Applications**, Springer, v. 77, n. 15, p. 19951–20000, 2018. Citado 3 vezes nas páginas 50, 72 e 73.

DING, Y.; ZHANG, J. Estimation of spad value in tomato leaves by multispectral images. In: IOP PUBLISHING. **Journal of Physics: Conference Series**. [S.l.], 2020. v. 1634, n. 1, p. 012128. Citado na página 69.

DORIGHELLO, D. V. et al. Management of asian soybean rust with bacillus subtilis in sequential and alternating fungicide applications. **Australasian Plant Pathology**, Springer, v. 49, n. 1, p. 79–86, 2020. Citado na página 43.

EMBRAPA FERRUGEM SOJA. **Vazio sanitário e calendarização da semeadura da soja**. 2020. Disponível em: <https://www.embrapa.br/en/soja/ferrugem/vaziosanitariocalendarizacaosemeadura>. Acesso em: 18 nov. 2020. Citado na página 43.

EMBRAPA, S. **Repositório Digipathos - Embrapa Soja**. 2021. Disponível em: <https://www.digipathos-rep.cnptia.embrapa.br/>. Acesso em: 12 fev. 2021. Citado na página 91.

EMBRAPA SOJA. **Soja em números (safra 2019/20)**. 2023. Disponível em: <https://www.embrapa.br/web/portal/soja/cultivos/soja1/dados-economicos>. Acesso em: 21 set. 2023. Citado na página 40.

EMMANUEL, I.; STANIER, C. Defining big data. In: **Proceedings of the International Conference on Big Data and Advanced Wireless Technologies**. [S.l.: s.n.], 2016. p. 1–6. Citado na página 53.

FACELI, K. et al. Inteligência artificial: Uma abordagem de aprendizado de máquina. **Rio de Janeiro: LTC**, v. 2, p. 192, 2011. Citado 6 vezes nas páginas 60, 61, 97, 110, 112 e 113.

FEHR, W. R.; CAVINESS, C. E. Stages of soybean development. Iowa State University. Agricultural and Home Economics Experiment Station, 1977. Citado na página 41.

FERNANDES, C. d. F. Doenças da bananicultura: sigatoca-negra. **Embrapa Rondônia-Circular Técnica (INFOTECA-E)**, Porto Velho: Embrapa Rondônia, 2005., 2005. Citado na página 31.

FILHO, O. M.; NETO, H. V. **Processamento digital de imagens**. [S.l.]: Brasport, 1999. Citado na página 66.

FLOREA, A. M. I. et al. Data integration approaches using etl. **Database System Journal**, v. 6, n. 3, p. 19–27, 2015. Citado na página 59.

GARCIA-LAMONT, F. et al. Segmentation of images by color features: A survey. **Neurocomputing**, Elsevier, v. 292, p. 1–27, 2018. Citado na página 98.

GARCÍA, S. et al. Big data preprocessing: methods and prospects. **Big Data Analytics**, BioMed Central, v. 1, n. 1, p. 1–22, 2016. Citado na página 61.

GASPAR, P. D. Fruticultura 4.0: Novas tecnologias na fruticultura. **III AGROCIÊNCIA ABRIL 2020 VOZ DO CAMPO**, 2020. Citado na página 65.

GHAMISI, P. et al. Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art. **IEEE Geoscience and Remote Sensing Magazine**, IEEE, v. 7, n. 1, p. 6–39, 2019. Citado na página 86.

GHIWARI, S.; SAMBREKAR, K.; RAJPUROHIT, V. Hierarchical storage for agro informatics system using nosql technology. In: IEEE. **2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)**. [S.l.], 2017. p. 1–5. Citado na página 52.

GODOY, C. et al. Boas práticas para o enfrentamento da ferrugem-asiática da soja. **Embrapa Soja-Comunicado Técnico (INFOTECA-E)**, Londrina: Embrapa Soja, 2017., 2017. Citado 2 vezes nas páginas 43 e 45.

GODOY, C. V. et al. Eficiência do controle da ferrugem asiática da soja em função do momento de aplicação sob condições de epidemia em londrina, pr. **Tropical Plant Pathology**, SciELO Brasil, v. 34, n. 1, p. 56–61, 2009. Citado na página 42.

GODOY, C. V.; KOGA, L. J.; CANTERI, M. G. Diagrammatic scale for assessment of soybean rust severity. **Fitopatologia Brasileira**, SciELO Brasil, v. 31, n. 1, p. 63–68, 2006. Citado na página 46.

GODOY, C. V. et al. Asian soybean rust in brazil: past, present, and future. **Pesquisa Agropecuária Brasileira**, SciELO Brasil, v. 51, n. 5, p. 407–421, 2016. Citado 3 vezes nas páginas 40, 42 e 43.

GONZALEZ, R. C.; WOODS, R. E. Processamento digital de imagem. **Pearson, ISBN-10: 8576054019,**, v. 10, p. 11–27, 2010. Citado 2 vezes nas páginas 97 e 98.

GOULART, A. C. P.; FURLAN, S. H.; FUJINO, M. T. Controle integrado da ferrugem asiática da soja (phakopsora pachyrhizi) com o fungicida fluquinconazole aplicado nas sementes em associação com outros fungicidas pulverizados na parte aérea da cultura. **Summa phytopathologica**, v. 37, n. 2, p. 113–118, 2011. Citado na página 42.

GREVILLE, T. N. E. Theory and applications of spline functions. **Theory and applications of spline functions**, 1969. Citado 2 vezes nas páginas 117 e 118.

HAN, L.; HALEEM, M. S.; TAYLOR, M. Automatic detection and severity assessment of crop diseases using image pattern recognition. In: SPRINGER. **Emerging Trends and Advanced Technologies for Computational Intelligence: Extended and Selected Results from the Science and Information Conference 2015**. [S.l.], 2016. p. 283–300. Citado na página 83.

HAO, W. et al. Application of information fusion technologies for multi-source data. **Journal of Chemical and Pharmaceutical Research**, v. 5, n. 12, p. 560–564, 2013. Citado na página 60.

HASSENSTEIN, M. J.; VANELLA, P. Data quality—concepts and problems. **Encyclopedia**, MDPI, v. 2, n. 1, p. 498–510, 2022. Citado na página 65.

HORé, A.; ZIOU, D. Image quality metrics: Psnr vs. ssim. p. 2366–2369, 2010. Citado na página 140.

HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of educational psychology**, Warwick & York, v. 24, n. 6, p. 417, 1933. Citado na página 114.

HU, M.-K. Visual pattern recognition by moment invariants. **IRE transactions on information theory**, IEEE, v. 8, n. 2, p. 179–187, 1962. Citado na página 102.

INMET, D. **Banco de dados meteorológicos para ensino e pesquisa**. 2019. Disponível em: <https://portal.inmet.gov.br>. Acesso em: 03 jul. 2019. Citado 2 vezes nas páginas 91 e 94.

JOHN, T.; MISRA, P. **Data lake for enterprises**. [S.l.]: Packt Publishing Ltd, 2017. Citado 2 vezes nas páginas 54 e 94.

JOLLIFFE, I. T. **Principal Component Analysis**. Second edition. New York, NY: Springer, 2002. (Springer Series in Statistics). ISSN 0172-7397. ISBN 0387954422. Citado na página 114.

JUKIĆ, N. et al. Augmenting data warehouses with big data. **Information Systems Management**, Taylor & Francis, v. 32, n. 3, p. 200–209, 2015. Citado 4 vezes nas páginas 53, 54, 56 e 62.

JULIATTI, F. C. et al. Sensitivity of two isolates of phakopsora pachyrhizi to dithiocarmamate, chloronitril, triazoles, strobilurins, and carboxamides fungicides. **Bioscience Journal**, v. 33, n. 4, 2017. Citado na página 43.

KARBOWSKI, A. et al. Critical infrastructure risk assessment using markov chain model. **Journal of Telecommunications and Information Technology**, Instytut Łączności-Państwowy Instytut Badawczy, n. 2, p. 15–22, 2019. Citado na página 134.

KARHUNEN, K. **Ueber lineare Methoden in der Wahrscheinlichkeitsrechnung**. Soumalainen Tiedeakatemia, 1947. (Suomalaisen Tiedeakatemian toimituksia : Serja A : I). Disponível em: <https://books.google.com.br/books?id=OyT7xwEACAAJ>. Citado na página 114.

KAUR, R.; GARG, R.; AGGARWAL, H. Big data analytics framework to identify crop disease and recommendation a solution. In: IEEE. **2016 International Conference on Inventive Computation Technologies (ICICT)**. [S.l.], 2016. v. 2, p. 1–5. Citado na página 55.

KHALEGHI, B.; KHAMIS, A.; KARRAY, F. Multisensor data fusion: a data-centric review of the state of the art and overview of emerging trends. **Multisensor Data Fusion**, CRC Press, p. 15–33, 2016. Citado na página 84.

KIM, D. et al. 81.6 gops object recognition processor based on a memory-centric noc. **IEEE transactions on very large scale integration (VLSI) systems**, IEEE, v. 17, n. 3, p. 370–383, 2009. Citado na página 103.

KLEMA, V.; LAUB, A. The singular value decomposition: Its computation and some applications. **IEEE Transactions on automatic control**, IEEE, v. 25, n. 2, p. 164–176, 1980. Citado na página 114.

KRIG, S. **Computer vision metrics: Survey, taxonomy, and analysis**. [S.l.]: Springer Nature, 2014. Citado na página 66.

KUCHIPUDI, D. P.; BABU, T. R. A review on segmentation of plant maladies and pathological parts from the leaf images in agriculture crop. In: IEEE. **2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)**. [S.l.], 2019. v. 1, p. 927–934. Citado 2 vezes nas páginas 74 e 81.

LANGE, K.; CHAMBERS, J.; EDDY, W. **Numerical analysis for statisticians**. [S.l.]: Springer, 2010. v. 1. Citado na página 134.

LELIS, V. d. P. et al. Favorabilidade ao desenvolvimento da ferrugem asiática da soja no estado de minas gerais. Engenharia na Agricultura, 2009. Citado 3 vezes nas páginas 42, 44 e 46.

LEUNG, T.; MALIK, J. Contour continuity in region based image segmentation. In: SPRINGER. **Computer Vision—ECCV'98: 5th European Conference on Computer Vision Freiburg, Germany, June, 2–6, 1998 Proceedings, Volume I 5**. [S.l.], 1998. p. 544–559. Citado na página 97.

LEVADA, A. Uma breve introdução ao reconhecimento de padrões. 05 2022. Citado 2 vezes nas páginas 114 e 116.

LI, H. et al. Application of multi-sensor image fusion of internet of things in image processing. **Ieee Access**, IEEE, v. 6, p. 50776–50787, 2018. Citado na página 86.

LI, Y. et al. Information fusion of passive sensors for detection of moving targets in dynamic environments. **IEEE transactions on cybernetics**, IEEE, v. 47, n. 1, p. 93–104, 2016. Citado na página 85.

_____. Information-theoretic performance analysis of sensor networks via markov modeling of time series data. **IEEE transactions on cybernetics**, IEEE, v. 48, n. 6, p. 1898–1909, 2017. Citado 2 vezes nas páginas 85 e 124.

LIAKOS, K. G. et al. Machine learning in agriculture: A review. **Sensors**, Multidisciplinary Digital Publishing Institute, v. 18, n. 8, p. 2674, 2018. Citado na página 61.

LIANG, H. et al. Text feature extraction based on deep learning: a review. **EURASIP journal on wireless communications and networking**, Springer, v. 2017, p. 1–12, 2017. Citado na página 83.

LIU, B. et al. A spark-based parallel fuzzy *c*-means segmentation algorithm for agricultural image big data. **IEEE Access**, IEEE, v. 7, p. 42169–42180, 2019. Citado 2 vezes nas páginas 57 e 124.

LOWE, D. G. Object recognition from local scale-invariant features. In: IEEE. **Proceedings of the seventh IEEE international conference on computer vision**. [S.l.], 1999. v. 2, p. 1150–1157. Citado na página 102.

_____. Distinctive image features from scale-invariant keypoints. **International journal of computer vision**, Springer, v. 60, p. 91–110, 2004. Citado na página 82.

LU, B. et al. Recent advances of hyperspectral imaging technology and applications in agriculture. **Remote Sensing**, Multidisciplinary Digital Publishing Institute, v. 12, n. 16, p. 2659, 2020. Citado 2 vezes nas páginas 70 e 71.

LU, J.; HOLUBOVÁ, I. Multi-model databases: a new journey to handle the variety of data. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 52, n. 3, p. 1–38, 2019. Citado na página 51.

MA, J. et al. A segmentation method for greenhouse vegetable foliar disease spots images using color information and region growing. **Computers and Electronics in Agriculture**, Elsevier, v. 142, p. 110–117, 2017. Citado na página 81.

_____. A recognition method for cucumber diseases using leaf symptom images based on deep convolutional neural network. **Computers and electronics in agriculture**, Elsevier, v. 154, p. 18–24, 2018. Citado 2 vezes nas páginas 72 e 81.

MAMDANI, E.; ASSILIAN, S. An experiment in linguistic synthesis with a fuzzy logic controller. **International Journal of Human-Computer Studies**, v. 51, n. 2, p. 135–147, 1999. ISSN 1071-5819. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1071581973603035>. Citado 3 vezes nas páginas 129, 130 e 131.

MANAVALAN, R. Automatic identification of diseases in grains crops through computational approaches: A review. **Computers and Electronics in Agriculture**, Elsevier, v. 178, p. 105802, 2020. Citado 4 vezes nas páginas 51, 75, 77 e 82.

MARKOV, A. Extension of the limit theorems of probability theory to a sum of variables connected in a chain. **Dynam Probabilist Syst**, v. 1, p. 552, 1971. Citado na página 134.

MASSRUHÁ, S. M. F. S.; LEITE, M. d. A. Agro 4.0-rumo à agricultura digital. In: IN: MAGNONI JÚNIOR, L.; STEVENS, D.; SILVA, WTL DA; VALE, JMF DO; PURINI, SR .... **Embrapa Informática Agropecuária-Artigo em anais de congresso (ALICE)**. [S.l.], 2017. Citado na página 65.

MEHMOOD, H. et al. Implementing big data lake for heterogeneous data sources. In: IEEE. **2019 ieee 35th international conference on data engineering workshops (icdew)**. [S.l.], 2019. p. 37–44. Citado 2 vezes nas páginas 60 e 94.

MILOSLAVSKAYA, N.; TOLSTOY, A. Big data, fast data and data lake concepts. **Procedia Computer Science**, Elsevier, v. 88, p. 300–305, 2016. Citado na página 59.

MINZ, P.; SAWHNEY, I. K.; SAINI, C. S. Algorithm for processing high definition images for food colourimetry. **Measurement**, Elsevier, v. 158, p. 107670, 2020. Citado na página 67.

MIYAGUSUKU, R.; YAMASHITA, A.; ASAMA, H. Data information fusion from multiple access points for wifi-based self-localization. **IEEE Robotics and Automation Letters**, IEEE, v. 4, n. 2, p. 269–276, 2018. Citado na página 87.

MOGHIMI, A. et al. A novel machine learning approach to estimate grapevine leaf nitrogen concentration using aerial multispectral imagery. **Remote Sensing**, Multidisciplinary Digital Publishing Institute, v. 12, n. 21, p. 3515, 2020. Citado na página 69.

MONTEIRO, R. d. C. M. et al. Image processing to identify damage to soybean seeds. **Ciência Rural**, SciELO Brasil, v. 51, n. 2, 2021. Citado na página 68.

MURITHI, H. et al. Soybean production in eastern and southern africa and threat of yield loss due to soybean rust caused by phakopsora pachyrhizi. **Plant pathology**, Wiley Online Library, v. 65, n. 2, p. 176–188, 2016. Citado na página 40.

NAGASUBRAMANIAN, K. et al. Hyperspectral band selection using genetic algorithm and support vector machines for early identification of charcoal rot disease in soybean stems. **Plant methods**, Springer, v. 14, n. 1, p. 1–13, 2018. Citado na página 70.

NASCIMENTO, J. F. d. et al. Progress of Asian soybean rust and airborne urediniospores of Phakopsora pachyrhizi in southern Brazil. **Summa Phytopathologica**, scielo, v. 38, p. 280 – 287, 12 2012. ISSN 0100-5405. Disponível em: <https://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-54052012000400002&nrm=iso>. Citado na página 46.

NEVES, R. A.; CRUVINEL, P. E. Model for semantic base structuring of digital data to support agricultural management. In: IEEE. **2020 IEEE 14th International Conference on Semantic Computing (ICSC)**. [S.l.], 2020. p. 337–340. Citado 2 vezes nas páginas 52 e 95.

_____. Ontology for structuring a digital databases for decision making in grain production. In: IEEE. **2021 IEEE 15th International Conference on Semantic Computing (ICSC)**. [S.l.], 2021. p. 386–392. Citado na página 52.

_____. Application of image processing and advanced intelligent computing for determining stage of asian rust in soybean plants. In: IEEE. **2022 IEEE 16th International Conference on Semantic Computing (ICSC)**. [S.l.], 2022. p. 280–286. Citado na página 83.

NISAR, N. et al. Image based recognition of plant leaf diseases: A review. In: IEEE. **2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)**. [S.l.], 2020. p. 373–378. Citado 2 vezes nas páginas 75 e 83.

NUNES, C. D. M. Manejo da ferrugem asiática da soja por número de aplicação de fungicidas, safra 2012/2013. In: IN: REUNIÃO DE PESQUISA DA SOJA DA REGIÃO SUL, 40., PELOTAS, 2014. ATAS E .... **Embrapa Clima Temperado-Artigo em anais de congresso (ALICE)**. [S.l.], 2014. Citado na página 43.

NUNES, C. D. M.; MARTINS, J. F. da S.; PONTE, E. M. D. Validação de modelo de previsão de ocorrência da ferrugem asiática da soja com base em precipitação pluviométrica. **Embrapa Clima Temperado-Circular Técnica (INFOTECA-E)**, Pelotas: Embrapa Clima Temperado, 2018., p. 1–13, 2018. Citado 3 vezes nas páginas 41, 45 e 47.

PEDRYCZ, W. Why triangular membership functions? **Fuzzy sets and Systems**, Elsevier, v. 64, n. 1, p. 21–30, 1994. Citado 2 vezes nas páginas 127 e 128.

PEREIRA, M. F. L. et al. Parallel computational structure and semantics for soil quality analysis based on lora and apache spark. In: IEEE. **2020 IEEE 14th International Conference on Semantic Computing (ICSC)**. [S.l.], 2020. p. 332–336. Citado na página 57.

PIPINO, L. L.; LEE, Y. W.; WANG, R. Y. Data quality assessment. **Communications of the ACM**, ACM New York, NY, USA, v. 45, n. 4, p. 211–218, 2002. Citado na página 62.

PIRES, R. D. L. et al. Local descriptors for soybean disease recognition. **Computers and Electronics in Agriculture**, Elsevier, v. 125, p. 48–55, 2016. Citado 3 vezes nas páginas 48, 71 e 83.

POKORNỲ, J. Database technologies in the world of big data. In: **Proceedings of the 16th International Conference on Computer Systems and Technologies**. [S.l.: s.n.], 2015. p. 1–12. Citado na página 59.

PONTE, E. M. D. et al. Models and applications for risk assessment and prediction of asian soybean rust epidemics. **Fitopatologia Brasileira**, SciELO Brasil, v. 31, n. 6, p. 533–544, 2006. Citado na página 47.

PONZONI, F. J. et al. **Calibração de sensores orbitais**. [S.l.]: Oficina de Textos, 2015. Citado na página 68.

PREEDANAN, W. et al. A comparative study of image quality assessment. p. 1–4, 2018. Citado na página 140.

PROKOPOWICZ, P. et al. **Theory and applications of ordered fuzzy numbers: a tribute to Professor Witold Kosiński**. [S.l.]: Springer Nature, 2017. Citado 2 vezes nas páginas 127 e 132.

RABINER, L. R. A tutorial on hidden markov models and selected applications in speech recognition. **Proceedings of the IEEE**, Ieee, v. 77, n. 2, p. 257–286, 1989. Citado 2 vezes nas páginas 134 e 135.

RAJESWARI, S.; SUTHENDRAN, K.; RAJAKUMAR, K. A smart agricultural model by integrating iot, mobile and cloud-based big data analytics. In: IEEE. **2017 international conference on intelligent computing and control (I2C2)**. [S.l.], 2017. p. 1–5. Citado na página 52.

RAMASAMY, A.; CHOWDHURY, S. Big data quality dimensions: A systematic literature review. **JISTEM-Journal of Information Systems and Technology Management**, SciELO Brasil, v. 17, 2020. Citado 2 vezes nas páginas 63 e 64.

REHMAN, M. H. ur et al. The role of big data analytics in industrial internet of things. **Future Generation Computer Systems**, Elsevier, v. 99, p. 247–259, 2019. Citado na página 59.

RIDLER, T.; CALVARD, S. et al. Picture thresholding using an iterative selection method. **IEEE Trans. Syst. Man Cybern**, v. 8, n. 8, p. 630–632, 1978. Citado na página 97.

RUGGIERO, M. A. G.; LOPES, V. L. d. R. **Cálculo numérico: aspectos teóricos e computacionais**. [S.l.]: Makron Books do Brasil, 1997. Citado na página 117.

SACHAR, S.; KUMAR, A. Survey of feature extraction and classification techniques to identify plant through leaves. **Expert Systems with Applications**, Elsevier, p. 114181, 2020. Citado 2 vezes nas páginas 79 e 80.

SARA, U.; AKTER, M.; UDDIN, M. S. Image quality assessment through fsim, ssim, mse and psnr—a comparative study. **Journal of Computer and Communications**, Scientific Research Publishing, v. 7, n. 3, p. 8–18, 2019. Citado na página 140.

SHAH, P.; HIREMATH, D.; CHAUDHARY, S. Towards development of spark based agricultural information system including geo-spatial data. In: IEEE. **2017 IEEE International Conference on Big Data (Big Data)**. [S.l.], 2017. p. 3476–3481. Citado na página 52.

SHARMA, S.; RATHEE, G.; SAINI, H. Big data analytics for crop prediction mode using optimization technique. In: IEEE. **2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC)**. [S.l.], 2018. p. 760–764. Citado na página 61.

SHEN, Q. et al. Centralized fusion methods for multi-sensor system with bounded disturbances. **IEEE Access**, IEEE, v. 7, p. 141612–141626, 2019. Citado na página 87.

SHIN, J. et al. An optimal image selection method to improve quality of relative radiometric calibration for uav multispectral images. **The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences**, Copernicus GmbH, v. 43, p. 493–498, 2020. Citado na página 69.

SHRIVASTAVA, S.; SINGH, S. K.; HOODA, D. S. Statistical texture and normalized discrete cosine transform-based automatic soya plant foliar infection cataloguing. **Journal of Advances in Mathematics and Computer Science**, p. 2901–2916, 2014. Citado na página 71.

_____. Color sensing and image processing-based automatic soybean plant foliar disease severity detection and estimation. **Multimedia Tools and Applications**, Springer, v. 74, n. 24, p. 11467–11484, 2015. Citado na página 71.

_____. Soybean plant foliar disease detection using image retrieval approaches. **Multimedia Tools and Applications**, v. 76, n. 24, p. 26647–26674, 2017. ISSN 1573-7721. Disponível em: <https://doi.org/10.1007/s11042-016-4191-7>. Citado 3 vezes nas páginas 48, 71 e 72.

SILVA, D. C. da; CANDEIAS, A. L. B. Causas da iluminação não uniforme em fotografias aéreas coloridas. **Revista Brasileira de Cartografia**, v. 61, n. 2, 2009. Citado na página 69.

SILVA, G. da et al. Recognition of soybean diseases using machine learning techniques based on segmentation of images captured by uavs. In: SBC. **Anais do XVI Workshop de Visão Computacional**. [S.l.], 2020. p. 12–17. Citado na página 75.

SILVEIRA, F. da; LERMEN, F. H.; AMARAL, F. G. An overview of agriculture 4.0 development: Systematic review of descriptions, technologies, barriers, advantages, and disadvantages. **Computers and Electronics in Agriculture**, Elsevier, v. 189, p. 106405, 2021. Citado na página 65.

SIMIONATO, R. et al. Survey on connectivity and cloud computing technologies: State-of-the-art applied to agriculture 4.0. **Revista Ciência Agronômica**, SciELO Brasil, v. 51, 2021. Citado na página 32.

SOJA, E. **Sistemas de produção 17**. Tecnologias de produção de soja, 2020. Disponível em: <https://ainfo.cnptia.embrapa.br/digital/bitstream/item/223209/1/SP-17-2020-online-1.pdf>. Acesso em: 04 out. 2023. Citado na página 91.

SONKA, M.; HLAVAC, V.; BOYLE, R. **Image processing, analysis, and machine vision**. [S.l.]: Nelson Education, 2014. Citado 2 vezes nas páginas 66 e 68.

SOSINSKY, B. **Cloud computing bible**. [S.l.]: John Wiley & Sons, 2010. v. 762. Citado na página 59.

STEMPLIUK, S.; MENOTTI, D. Agriculture multispectral uav image registration using salient features and mutual information. In: IEEE. **IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium**. [S.l.], 2020. p. 4108–4111. Citado na página 68.

SUDHESH, R.; NAGALAKSHMI, V.; AMIRTHASARAVANAN, A. A systematic study on disease recognition, categorization, and quantification in agricultural plants using image processing. In: IEEE. **2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)**. [S.l.], 2019. p. 1–5. Citado 2 vezes nas páginas 74 e 82.

TALEB, I.; SERHANI, M. A.; DSSOULI, R. Big data quality assessment model for unstructured data. In: IEEE. **2018 International Conference on Innovations in Information Technology (IIT)**. [S.l.], 2018. p. 69–74. Citado na página 65.

TANG, R.; ARIDAS, N. K.; TALIP, M. S. A. Design of data processing methods for the farmland environmental monitoring based on improved spark components. **Frontiers in Big Data**, Frontiers, v. 6, 2023. ISSN 2624-909X. Disponível em: <https://www.frontiersin.org/articles/10.3389/fdata.2023.1282352>. Citado na página 58.

TANIMOTO, O. S. et al. Aproach prima no controle da ferrugem da soja, comparando-se diversos tipos de adjuvantes. **Nucleus**, Fundação Educacional Ituverava, v. 8, n. 1, p. 1–12, 2011. Citado na página 42.

TETILA, E. **Detecção e classificação de doenças e pragas da soja usando imagens de veículos aéreos não tripulados e técnicas de visão computacional. 2019. 103p**. Tese (Doutorado) — Tese (Doutorado em Desenvolvimento Local)- Universidade Católica Dom Bosco . . . , 2019. Citado 2 vezes nas páginas 50 e 74.

TIBOLA, C. S. et al. Espectroscopia no infravermelho próximo para avaliar indicadores de qualidade tecnológica e contaminantes em grãos. **Embrapa Trigo-Livro científico (ALICE)**, Brasília, DF: Embrapa, 2018., 2018. Citado na página 70.

TSUKAMOTO, Y. An approach to fuzzy reasoning method. **Advances in fuzzy set theory and applications**, North Holland, 1979. Citado na página 128.

TURKOGLU, M.; HANBAY, D. Leaf-based plant species recognition based on improved local binary pattern and extreme learning machine. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 527, p. 121297, 2019. Citado na página 79.

VAPNIK, V. **The nature of statistical learning theory**. [S.l.]: Springer Science+Business Media New York, 1995. Citado 2 vezes nas páginas 107 e 110.

VERAMENDI, W. N. C. Método para contagem de plantas de milho baseado no processamento digital de imagens multiespectrais utilizando drones em ambiente de campo. Universidade Federal de São Carlos, 2022. Citado 2 vezes nas páginas 70 e 71.

WANG, L.; LIANG, Q. Representation learning and nature encoded fusion for heterogeneous sensor networks. **IEEE Access**, IEEE, v. 7, p. 39227–39235, 2019. Citado na página 87.

WANG, X.; LIANG, J.; GUO, F. Feature extraction algorithm based on dual-scale decomposition and local binary descriptors for plant leaf recognition. **Digital Signal Processing**, Elsevier, v. 34, p. 101–107, 2014. Citado na página 79.

WANG, Y.-s.; WU, H.-r.; LI, Q.-x. Design and simulation of agricultural big data cloud storage system based on the relational database. In: **International Conference on Mathematics, Modelling and Simulation Technologies and Applications (MMSTA 2017)**. [S.l.: s.n.], 2017. Citado na página 52.

WARNER, J. et al. **JDWarner/scikit-fuzzy: Scikit-Fuzzy version 0.4.2**. Zenodo, 2019. Disponível em: <https://doi.org/10.5281/zenodo.3541386>. Citado na página 132.

XIA, J. et al. An information fusion model of innovation alliances based on the bayesian network. **Tsinghua Science and Technology**, TUP, v. 23, n. 3, p. 347–356, 2018. Citado na página 86.

YI, C.; ZHAO, Y.-Q.; CHAN, J. C.-W. Hyperspectral image super-resolution based on spatial and spectral correlation fusion. **IEEE Transactions on Geoscience and Remote Sensing**, IEEE, v. 56, n. 7, p. 4165–4177, 2018. Citado na página 86.

YORINORI, J. T.; JÚNIOR, J. N.; LAZZAROTTO, J. J. Ferrugem"asiática"da soja no brasil: evolução, importância econômica e controle. **Embrapa Soja-Fôlder/Folheto/Cartilha (INFOTECA-E)**, Londrina: Embrapa Soja, 2004., 2004. Citado na página 45.

YORINORI, J. T. et al. Ferrugem da soja (phakopsora pachyrhizi): identificação e controle. **Informações agronômicas**, v. 104, 2003. Citado 4 vezes nas páginas 41, 42, 43 e 44.

YOUSEFI, E.; BALEGHI, Y.; SAKHAEI, S. M. Rotation invariant wavelet descriptors, a new set of features to enhance plant leaves classification. **Computers and Electronics in Agriculture**, Elsevier, v. 140, p. 70–76, 2017. Citado na página 79.

YU, M.; MA, X.; GUAN, H. Recognition method of soybean leaf diseases using residual neural network based on transfer learning. **Ecological Informatics**, Elsevier, v. 76, p. 102096, 2023. Citado na página 51.

ZADEH, L. A. Fuzzy sets. **Information and control**, Elsevier, v. 8, n. 3, p. 338–353, 1965. Citado na página 125.

ZAGUI, N. L. S. et al. Spatio-temporal modeling and simulation of asian soybean rust based on fuzzy system. **Sensors**, MDPI, v. 22, n. 2, p. 668, 2022. Citado 2 vezes nas páginas 51 e 124.

ZAMBENEDETTI, E. et al. Evaluation of monocyclic parameters and intensity of the asian soybean rust (phakopsora pachyrhizi) in both several soybean genotypes and canopy position. **Summa Phytopathologica**, v. 33, p. 75–78, 2007. Citado na página 47.

ZHAO, W.; WANG, J. Study of feature extraction based visual invariance and species identification of weed seeds. In: IEEE. **2010 Sixth International Conference on Natural Computation**. [S.l.], 2010. v. 2, p. 631–635. Citado na página 103.

ZHOU, F. et al. Multifocus image fusion based on fast guided filter and focus pixels detection. **IEEE Access**, IEEE, v. 7, p. 50780–50796, 2019. Citado na página 88.

ZHOU, Z.-H. Three perspectives of data mining. **Artificial Intelligence**, Elsevier, v. 143, n. 1, p. 139–146, 2003. Citado na página 60.

ZUCCHINI, W.; MACDONALD, I. L.; LANGROCK, R. **Hidden Markov models for time series: an introduction using R**. [S.l.]: CRC press, 2009. Citado na página 137.

ZUCKER, S. W. Region growing: Childhood and adolescence. **Computer graphics and image processing**, Elsevier, v. 5, n. 3, p. 382–399, 1976. Citado na página 97.

ZUNTINI, B. et al. Effect of adding fungicide to mixtures of triazoles and strobilurins in the control of downy mildew and asian soybean rust. **Pesquisa Agropecuária Tropical**, SciELO Brasil, v. 49, 2019. Citado na página 43.

Apêndices

APÊNDICE A

Modelo Multidimensional (DW)

Figura 83 – Dicionário de Dados *Data Warehouse*

DIM_Banco_Imagens	Tipo_Dado
ID_BANCO_IMAGENS	NUMBER (*,0)
ID_PROJETOS	NUMBER (*,0)
DESCRICAO_IMG_ORIGINAL	VARCHAR2 (100)
ALTURA_IMG_ORIGINAL	NUMBER (*,0)
LARGURA_IMG_ORIGINAL	NUMBER (*,0)
NUM_CANAIS_IMG_ORIGINAL	NUMBER (*,0)
TOTAL_PIXELS_IMG_ORIGINAL	NUMBER (*,0)
TIPO_IMG_ORIGINAL	VARCHAR2 (45)
SK_DIM_BANCO_IMAGENS	NUMBER

DIM_Classificacoes	Tipo_Dado
ID_CLASSIFICACAO	NUMBER (*,0)
ID_PROJETOS	NUMBER (*,0)
ID_BANCO_IMAGENS	NUMBER (*,0)
TIPO_CLASSIFICACAO	VARCHAR2 (30)
PORCENTAGEM_TREINO	NUMBER (*,0)
PORCENTAGEM_TESTE	NUMBER (*,0)
RESULTADO	NUMBER (*,0)
SK_DIM_CLASSIFICACOES	NUMBER

DIM_Dados_Planta_Soja	Tipo_Dado
ID_DADOS_PLANTA_SOJA	NUMBER (*,0)
ID_PROJETOS	NUMBER (*,0)
VARIEDADE_CULTURA	VARCHAR2 (100)
DISTANCIA_LINHA_CULTURA	FLOAT (126)
ALTURA_PLANTA_CULTURA	FLOAT (126)
DISTANCIA_PLANTA_CULTURA	FLOAT (126)
PLANTAS_METRO_LINEAR	NUMBER
STATUS	VARCHAR2 (50)
SK_DIM_DADOS_PLANTA_SOJA	NUMBER

Fato_Favorabilidades_FAS	Tipo_Dado
SK_DIM_TEMPO	NUMBER (*,0)
ID_CLASSIFICACAO	NUMBER (*,0)
ID_BANCO_IMAGENS	NUMBER (*,0)
ID_FAVORABILIDADE	NUMBER (*,0)
ID_DADOS_PLANTA_SOJA	NUMBER (*,0)
ID_IMAGEM_CLIMA_FAVORAB	NUMBER (*,0)
RESULTADO_CLASSIFICACAO_IMG	NUMBER (*,0)
RESULTADO_FAVORABILIDADE	VARCHAR2 (500)
FAVORAB_V1_PERIODO_MOLHAMENTO_FOLIAR	NUMBER (*,0)
FAVORAB_V2_PERIODO_MIN_MOLHAMENTO_FOLIAR	NUMBER (*,0)
FAVORAB_V3_FAIXA_TEMPERATURA	NUMBER (*,0)
FAVORAB_V4_TEMPERATURA_MAXIMA	NUMBER (*,0)
FAVORAB_V5_TEMPERATURA_MINIMA	NUMBER (*,0)
FAVORAB_V6_PONTO_ORVALHO	NUMBER (*,0)
FAVORAB_V7_DADOS_CULTIVAR_CLASSIF_IMAGEM	NUMBER (*,0)
SK_FATO	NUMBER
ID_DADOS_CLIMATICOS	NUMBER
ABORDAGEM_FUSAO_DADOS	VARCHAR2 (100)
JANELA_TEMPO_INICIAL	DATE
JANELA_TEMPO_FINAL	DATE

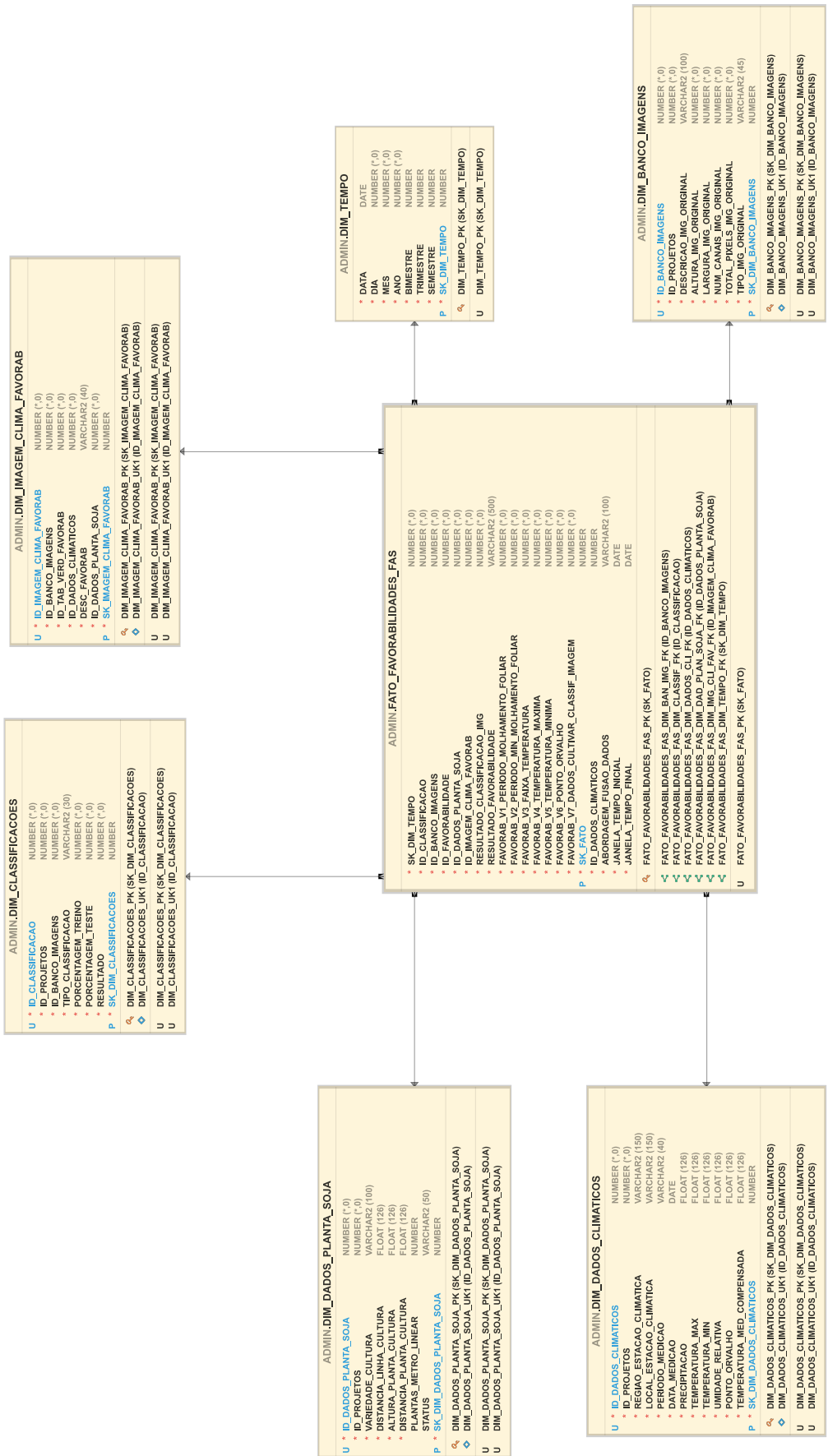
DIM_Dados_Climaticos	Tipo_Dado
ID_DADOS_CLIMATICOS	NUMBER (*,0)
ID_PROJETOS	NUMBER (*,0)
REGIAO_ESTACAO_CLIMATICA	VARCHAR2 (150)
LOCAL_ESTACAO_CLIMATICA	VARCHAR2 (150)
PERIODO_MEDICAO	VARCHAR2 (40)
DATA_MEDICAO	DATE
PRECIPITACAO	FLOAT (126)
TEMPERATURA_MAX	FLOAT (126)
TEMPERATURA_MIN	FLOAT (126)
UMIDADE_RELATIVA	FLOAT (126)
PONTO_ORVALHO	FLOAT (126)
TEMPERATURA_MED_COMPENSADA	FLOAT (126)
SK_DIM_DADOS_CLIMATICOS	NUMBER

DIM_Imagem_Clima_Favorab	Tipo_Dado
ID_IMAGEM_CLIMA_FAVORAB	NUMBER (*,0)
ID_BANCO_IMAGENS	NUMBER (*,0)
ID_TAB_VERD_FAVORAB	NUMBER (*,0)
ID_DADOS_CLIMATICOS	NUMBER (*,0)
DESC_FAVORAB	VARCHAR2 (40)
ID_DADOS_PLANTA_SOJA	NUMBER (*,0)
SK_IMAGEM_CLIMA_FAVORAB	NUMBER

DIM_Tempo	Tipo_Dado
DATA	DATE
DIA	NUMBER (*,0)
MES	NUMBER (*,0)
ANO	NUMBER (*,0)
BIMESTRE	NUMBER
TRIMESTRE	NUMBER
SEMESTRE	NUMBER
SK_DIM_TEMPO	NUMBER

Fonte: Próprio Autor

Figura 84 – Modelo Estrela - *Data Warehouse*

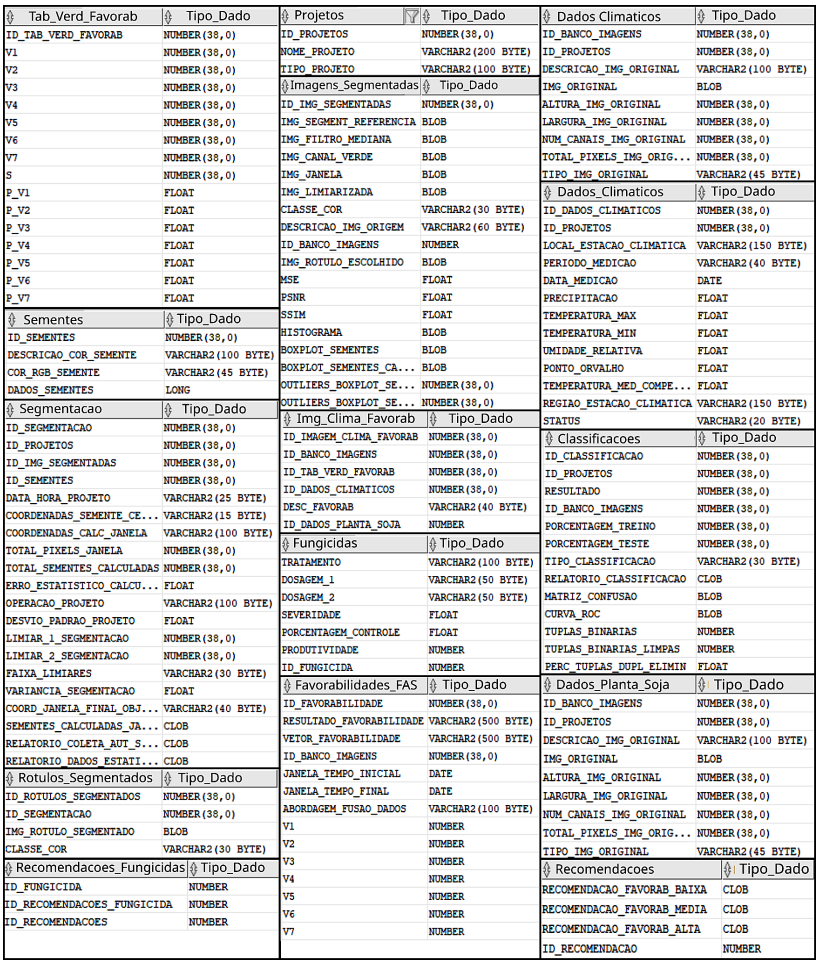


Fonte: Próprio Autor

APÊNDICE B

Modelo Banco Dados Relacional

Figura 85 – Dicionário de Dados - Banco de Dados Relacional



Tab_Verd_Favorab	Tipo_Dado
ID_TAB_VERD_FAVORAB	NUMBER(38,0)
V1	NUMBER(38,0)
V2	NUMBER(38,0)
V3	NUMBER(38,0)
V4	NUMBER(38,0)
V5	NUMBER(38,0)
V6	NUMBER(38,0)
V7	NUMBER(38,0)
S	NUMBER(38,0)
P_V1	FLOAT
P_V2	FLOAT
P_V3	FLOAT
P_V4	FLOAT
P_V5	FLOAT
P_V6	FLOAT
P_V7	FLOAT

Sementes	Tipo_Dado
ID_SEMENTES	NUMBER(38,0)
DESCRICAO_COR_SEMENTE	VARCHAR2(100 BYTE)
COR_RGB_SEMENTE	VARCHAR2(45 BYTE)
DADOS_SEMENTES	LONG

Segmentacao	Tipo_Dado
ID_SEGMENTACAO	NUMBER(38,0)
ID_PROJETOS	NUMBER(38,0)
ID_IMG_SEGMENTADAS	NUMBER(38,0)
ID_SEMENTES	NUMBER(38,0)
DATA_HORA_PROJETO	VARCHAR2(25 BYTE)
COORDENADAS_SEMENTE_CE...	VARCHAR2(15 BYTE)
COORDENADAS_CALC_JANELA	VARCHAR2(100 BYTE)
TOTAL_PIXELS_JANELA	NUMBER(38,0)
TOTAL_SEMENTES_CALCULADAS	NUMBER(38,0)
ERRO_ESTATISTICO_CALCU...	FLOAT
OPERACAO_PROJETO	VARCHAR2(100 BYTE)
DESVIO_PADRAO_PROJETO	FLOAT
LIMIAR_1_SEGMENTACAO	NUMBER(38,0)
LIMIAR_2_SEGMENTACAO	NUMBER(38,0)
FAIXA_LIMIARES	VARCHAR2(30 BYTE)
VARIANCIA_SEGMENTACAO	FLOAT
COORD_JANELA_FINAL_OBJ...	VARCHAR2(40 BYTE)
SEMENTES_CALCULADAS_JA...	CLOB
RELATORIO_COLETA_AUT_S...	CLOB
RELATORIO_DADOS_ESTATI...	CLOB

Rotulos_Segmentados	Tipo_Dado
ID_ROTULOS_SEGMENTADOS	NUMBER(38,0)
ID_SEGMENTACAO	NUMBER(38,0)
IMG_ROTULO_SEGMENTADO	BLOB
CLASSE_COR	VARCHAR2(30 BYTE)

Recomendacoes_Fungicidas	Tipo_Dado
ID_FUNGICIDA	NUMBER
ID_RECOMENDACOES_FUNGICIDA	NUMBER
ID_RECOMENDACOES	NUMBER

Projetos	Tipo_Dado
ID_PROJETOS	NUMBER(38,0)
NOME_PROJETO	VARCHAR2(200 BYTE)
TIPO_PROJETO	VARCHAR2(100 BYTE)

Imagens_Segmentadas	Tipo_Dado
ID_IMG_SEGMENTADAS	NUMBER(38,0)
IMG_SEGMENT_REFERENCIA	BLOB
IMG_FILTRO_MEDIANA	BLOB
IMG_CANAL_VERDE	BLOB
IMG_JANELA	BLOB
IMG_LIMIARIZADA	BLOB
CLASSE_COR	VARCHAR2(30 BYTE)
DESCRICAO_IMG_ORIGEM	VARCHAR2(60 BYTE)
ID_BANCO_IMAGENS	NUMBER
IMG_ROTULO_ESCOLHIDO	BLOB
MSE	FLOAT
PSNR	FLOAT
SSIM	FLOAT
HISTOGRAMA	BLOB
BOXPLOT_SEMENTES	BLOB
BOXPLOT_SEMENTES_CA...	BLOB
OUTLIERS_BOXPLOT_SE...	NUMBER(38,0)
OUTLIERS_BOXPLOT_SE...	NUMBER(38,0)

Img_Clima_Favorab	Tipo_Dado
ID_IMAGEM_CLIMA_FAVORAB	NUMBER(38,0)
ID_BANCO_IMAGENS	NUMBER(38,0)
ID_TAB_VERD_FAVORAB	NUMBER(38,0)
ID_DADOS_CLIMATICOS	NUMBER(38,0)
DESC_FAVORAB	VARCHAR2(40 BYTE)
ID_DADOS_PLANTA_SOJA	NUMBER

Fungicidas	Tipo_Dado
TRATAMENTO	VARCHAR2(100 BYTE)
DOSAGEM_1	VARCHAR2(50 BYTE)
DOSAGEM_2	VARCHAR2(50 BYTE)
SEVERIDADE	FLOAT
PORCENTAGEM_CONTROLE	FLOAT
PRODUTIVIDADE	NUMBER
ID_FUNGICIDA	NUMBER

Favorabilidades_FAS	Tipo_Dado
ID_FAVORABILIDADE	NUMBER(38,0)
RESULTADO_FAVORABILIDADE	VARCHAR2(500 BYTE)
VETOR_FAVORABILIDADE	VARCHAR2(500 BYTE)
ID_BANCO_IMAGENS	NUMBER(38,0)
JANELA_TEMPO_INICIAL	DATE
JANELA_TEMPO_FINAL	DATE
ABORDAGEM_FUSAO_DADOS	VARCHAR2(100 BYTE)
V1	NUMBER
V2	NUMBER
V3	NUMBER
V4	NUMBER
V5	NUMBER
V6	NUMBER
V7	NUMBER

Dados Climaticos	Tipo_Dado
ID_BANCO_IMAGENS	NUMBER(38,0)
ID_PROJETOS	NUMBER(38,0)
DESCRICAO_IMG_ORIGINAL	VARCHAR2(100 BYTE)
IMG_ORIGINAL	BLOB
ALTURA_IMG_ORIGINAL	NUMBER(38,0)
LARGURA_IMG_ORIGINAL	NUMBER(38,0)
NUM_CANAIS_IMG_ORIGINAL	NUMBER(38,0)
TOTAL_PIXELS_IMG_ORIG...	NUMBER(38,0)
TIPO_IMG_ORIGINAL	VARCHAR2(45 BYTE)

Dados_Climaticos	Tipo_Dado
ID_DADOS_CLIMATICOS	NUMBER(38,0)
ID_PROJETOS	NUMBER(38,0)
LOCAL_ESTACAO_CLIMATICA	VARCHAR2(150 BYTE)
PERIODO_MEDICAO	VARCHAR2(40 BYTE)
DATA_MEDICAO	DATE
PRECIPITACAO	FLOAT
TEMPERATURA_MAX	FLOAT
TEMPERATURA_MIN	FLOAT
UMIDADE_RELATIVA	FLOAT
PONTO_ORVALHO	FLOAT
TEMPERATURA_MED_COMPE...	FLOAT
REGIAO_ESTACAO_CLIMATICA	VARCHAR2(150 BYTE)
STATUS	VARCHAR2(20 BYTE)

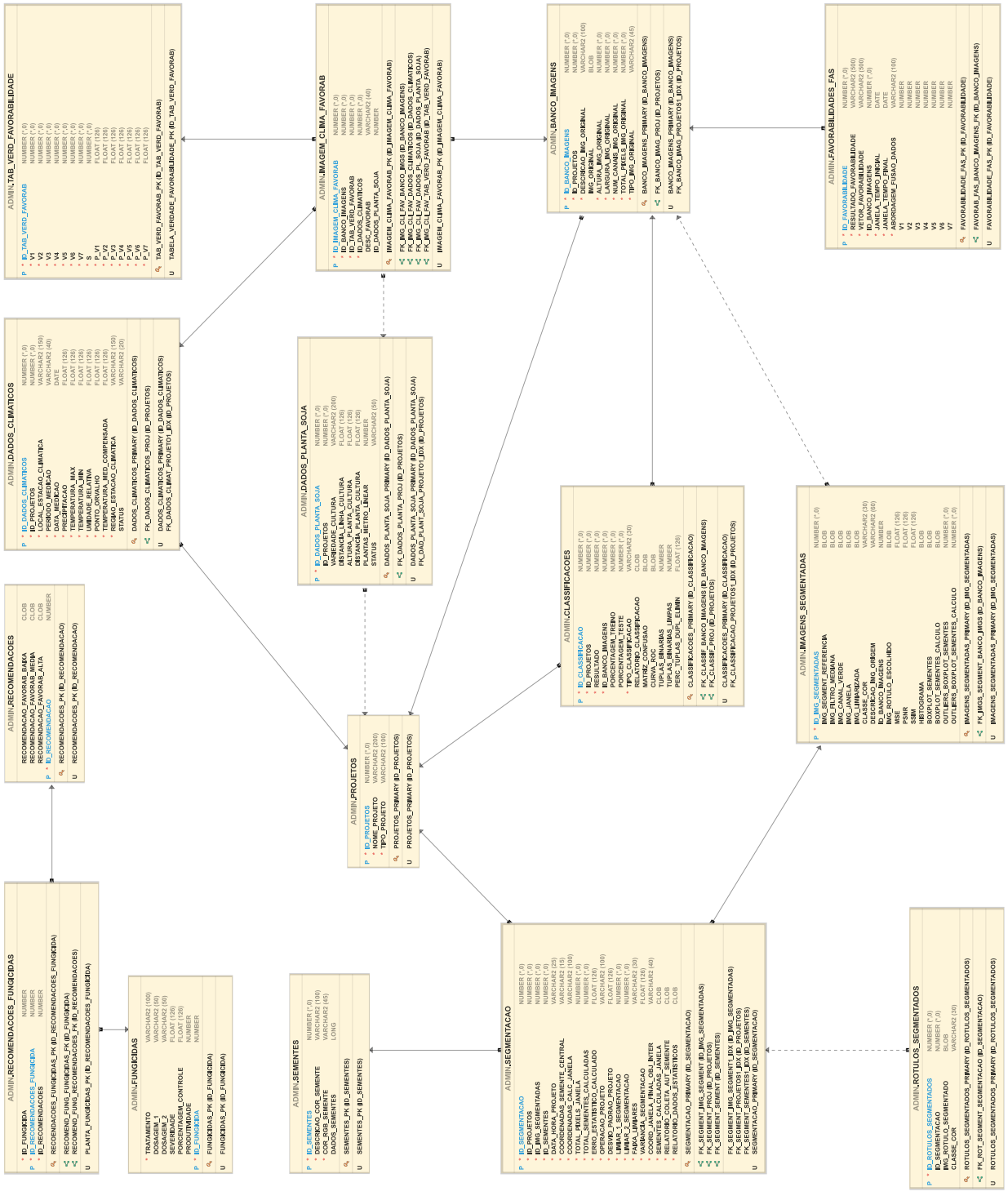
Classificacoes	Tipo_Dado
ID_CLASSIFICACAO	NUMBER(38,0)
ID_PROJETOS	NUMBER(38,0)
RESULTADO	NUMBER(38,0)
ID_BANCO_IMAGENS	NUMBER(38,0)
PORCENTAGEM_TREINO	NUMBER(38,0)
PORCENTAGEM_TESTE	NUMBER(38,0)
TIPO_CLASSIFICACAO	VARCHAR2(30 BYTE)
RELATORIO_CLASSIFICACAO	CLOB
MATRIZ_CONFUSAO	BLOB
CURVA_ROC	BLOB
TUPLAS_BINARIAS	NUMBER
TUPLAS_BINARIAS_LIMPAS	NUMBER
PERC_TUPLAS_DUPL_ELIMIN	FLOAT

Dados_Planta_Soja	Tipo_Dado
ID_BANCO_IMAGENS	NUMBER(38,0)
ID_PROJETOS	NUMBER(38,0)
DESCRICAO_IMG_ORIGINAL	VARCHAR2(100 BYTE)
IMG_ORIGINAL	BLOB
ALTURA_IMG_ORIGINAL	NUMBER(38,0)
LARGURA_IMG_ORIGINAL	NUMBER(38,0)
NUM_CANAIS_IMG_ORIGINAL	NUMBER(38,0)
TOTAL_PIXELS_IMG_ORIG...	NUMBER(38,0)
TIPO_IMG_ORIGINAL	VARCHAR2(45 BYTE)

Recomendacoes	Tipo_Dado
RECOMENDACAO_FAVORAB_BAIXA	CLOB
RECOMENDACAO_FAVORAB_MEDIA	CLOB
RECOMENDACAO_FAVORAB_ALTA	CLOB
ID_RECOMENDACAO	NUMBER

Fonte: Próprio Autor

Figura 86 – Modelo - Banco de Dados Relacional



Fonte: Próprio Autor

APÊNDICE C

Recomendações Agrícolas

As recomendações agrícolas são compostas pela Tabela 34 de fungicidas e também pela Tabela 35, de acordo com o resultado da favorabilidade do processamento.

Tabela 34 – Tabela de Fungicidas

Tratamento	Dosagem 1	Dosagem 2	Severid. %	Controle %	Produção kg/ha
Cypress (difenoconazol + ciproconazol)	0,3 L/ kg p.c./ha	75 + 45 g i.a./ha	51,90	25	2888
Dart (picoxistrobina + tebuconazol)	0,5 L/ kg p.c./ha	60 + 100 g i.a./ha	27,80	60	3288
Nativo (trifloxistrobina + tebuconazol)	0,5 L/ kg p.c./ha	50 + 100 g i.a./ha	29,60	57	3384
Fusão (metominostrobina + tebuconazol)	0,725 L/ kg p.c./ha	79,75 + 119,63 g i.a./ha	26,90	61	3345
Fezan Gold (tebuconazol + clorotalonil)	2,5 L/ kg p.c./ha	125 + 1125 g i.a./ha	25,40	63	3506
Armero (mancozebe + protioconazol)	2,25 L/ kg p.c./ha	1125 + 90 g i.a./ha	26,10	62	3494
Blavity (protioconazol + fluxapiroxade)	0,3 L/ kg p.c./ha	84 + 60 g i.a./ha	30,60	56	3484
Elatus (azoxistrobina + benzovindiflupir)	0,2 L/ kg p.c./ha	60 + 30 g i.a./ha	46,30	33	2917
Viovan (picoxistrobina + protioconazol)	0,6 L/ kg p.c./ha	60 + 70,02 g i.a./ha	31,00	55	3220
Vessarya (picoxistrobina + benzovindiflupir)	0,6 L/ kg p.c./ha	60 + 30 g i.a./ha	40,00	42	3150
Orkestra SC5 (piraclostrobina + fluxapiroxade)	0,35 L/ kg p.c./ha	116,55 + 58,45 g i.a./ha	41,70	40	3193
Alade (benzovindiflupir + ciproconazol + difenoconazol)	0,75 L/ kg p.c./ha	45 + 67,5 + 112,5 g i.a./ha	41,40	40	3125
Ativum (piraclostrobina + epoxiconazol + fluxapiroxade)	0,8 L/ kg p.c./ha	65 + 40 + 40 g i.a./ha	38,20	45	3046
Fox Xpro (bixafen + protioconazol + trifloxistrobina)	0,5 L/ kg p.c./ha	62,5 + 87,5 + 75 g i.a./ha	28,00	59	3395
Evolution (mancozebe + azoxistrobina + protioconazol)	2 L/ kg p.c./ha	1050 + 75 + 75 g i.a./ha	28,10	59	3483
Cronnos (mancozebe + picoxistrobina + tebuconazol)	2,5 L/ kg p.c./ha	1000 + 66,5 + 83,33 g i.a./ha	20,40	70	3620

Fonte: Próprio Autor

Tabela 35 – Tabela de Boas Práticas de Manejo da Soja

Boas Práticas de Manejo da Soja (Visão Agronômica)
1- Aplicar fungicidas Inibidores de desmetilação (IDMs), quando "Favorabilidade Alta" à FAS;
1.1- Aplicar preferencialmente em misturas com fungicidas multissítios na dose total;
1- Aplicar fungicidas Inibidores de desmetilação (IDMs), quando "Favorabilidade Média" à FAS;
1.1- Aplicar preferencialmente em misturas com fungicidas multissítios em 2/3 da dose total;
1- Não aplicar fungicidas Inibidores de desmetilação (IDMs), quando "Favorabilidade Baixa" à FAS;
1.1- Manter o monitoramento quando Favorabilidade Baixa à FAS;
1.2- Vide Tabela de "Opções e Seleção de Fungicidas";
2- Rotacionar fungicidas com diferentes mecanismos de ação (carboxamidas, estrobilurinas, morfolinas e multissítios);
3- Realizar a aplicação de fungicidas de forma preventiva, sempre em doses e intervalos recomendados pelos fabricantes e receituário agronômico;
4- Utilizar tecnologia de aplicação e volume de calda adequado para uma eficiente distribuição do produto sobre a planta;
5- Respeitar o vazio sanitário e eliminar as plantas voluntárias remanescentes em lavouras e beiras de estrada (guaxas);
6- Realizar o plantio na época recomendada, conforme calendário agrícola;
7- Realizar a rotação de culturas.

Fonte: Próprio Autor

A interpretação do relatório de recomendações consiste em observar as opções de fungicidas sugeridos, suas dosagens indicadas pelo fabricante, assim como os índices em percentagem da severidade da doença em que o fungicida foi aplicado, do controle da doença alcançado e a produção obtida após a aplicação, medida em kg/ha.

Além da observação das opções dos fungicidas sugeridos na Tabela 34 de boas práticas, outras iniciativas também são recomendadas para o manejo da soja na Tabela 35, que devem ser cuidadosamente observadas.

Esses fungicidas foram testados em campo pela Embrapa e os resultados divulgados em circular técnica que mostram a eficiência para o controle da ferrugem asiática da soja, por safra.

APÊNDICE D

Questionário Elaborado para a Etapa de Validação, Respondido por Especialistas

Teste de Validação

Estamos convidando Vossa Senhoria a participar da consulta sobre a ocorrência da Ferrugem Asiática da Soja, considerando três níveis de favorabilidade e a informação visual de folhas da respectiva cultura. Com a expectativa que possa aceitar participar seguem as considerações abaixo:

1. Deve-se ler as abordagens que versam sobre as condições de favorabilidade, quanto às variáveis climáticas, de acordo com a literatura, conforme sintetizado abaixo;
2. Deve-se assinalar as alternativas das questões 1 e 2 para cada teste e na questão 3, se considerar plausível, responder.

Síntese Teórica da Literatura

De acordo com Yorinori, Júnior e Lazzarotto (2004) *Phakopsora pachyrhizi*, o fungo causador da FAS, está adaptado a temperaturas que variam de 15°C até 30°C, onde ocorra molhamento de folha acima de seis horas. A doença ocorre com maior severidade sob condições de prolongado período de molhamento foliar e temperaturas médias abaixo de 28°C. Períodos prolongados com temperaturas acima de 28°C, segundo os autores, reduzem o desenvolvimento da ferrugem. Os uredosporos germinam em uma hora à temperatura ambiente de 25° à 27°C, porém a penetração no tecido da folha pode ocorrer à temperatura variando de 8°C a 28°C.

Del Ponte e colaboradores identificaram que períodos longos de orvalho e temperaturas variando de 15°C a 29°C parecem ser ótimos para o desenvolvimento da ferrugem da soja. A temperatura ótima para a germinação fúngica variou entre 15°C e 25°C, sendo necessário um período mínimo de umidade de 6 horas para a infecção, e uma eficiência de infecção crescente com uma duração de umidade de 8 a 12 horas (Del Ponte et al., 2006).

Lelis e colaboradores também identificaram que as condições favoráveis ao desenvolvimento da FAS em dois diferentes modelos. O primeiro modelo evidenciou o número de horas com umidade relativa maior ou igual a 90% e, no segundo, destacou a depressão do ponto de orvalho menor que 2°C. Para os dois modelos, a faixa de temperatura de trabalho foi de 18°C a 25°C, considerada como temperatura ideal para o desenvolvimento do fungo causador da ferrugem asiática (LELIS et al.,2009).

Segundo Godoy e colaboradores, o processo de infecção da FAS depende basicamente da disponibilidade de água livre na superfície da planta, a duração do molhamento foliar de seis a doze horas, além de temperaturas de 15°C a 28°C (GODOY et al., 2017).

De acordo com Nunes, Martins e Ponte (2018), as temperaturas extremas, tais como 20°C, considerada baixa, e 30°C, considerada alta, podem provocar a ausência ou atraso de desenvolvimento de epidemia da FAS, pois nesses extremos há perspectiva de redução em até 80% quanto a produção de esporos do patógeno.

Teste 1:

Imagem 1: Folha (DSC_0151)

Período Mínimo de Molhamento Foliar	Temperatura Máxima	Temperatura Mínima	Período de Molhamento Foliar	Ponto de Orvalho	Temperatura Média Compensada
0h00min	39,00°C	23,60°C	46,88%	16,96°C	28,89°C

Tabela 1: Dados Climáticos

<u>Responda as questões abaixo baseando-se pelas informações acima:</u>

1- Há presença da Ferrugem Asiática da Soja?

( ) Sim ( ) Não

2- Caso identifique a presença Ferrugem Asiática da Soja, selecione a severidade:

( ) baixa ( ) média ( ) alta

3- Se julgar necessário, justifique sua resposta:

_____

_____

_____

_____

_____

Teste 2:

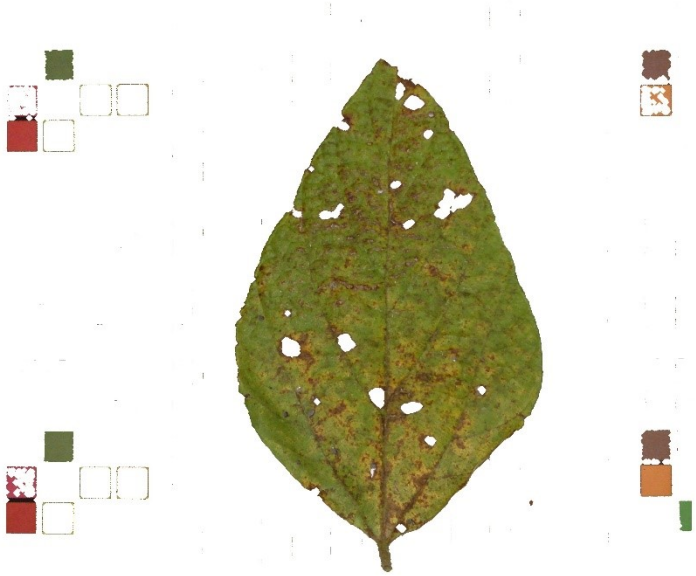


Imagem 2: Folha (DSC_0049)

Período Mínimo de Molhamento Foliar	Temperatura Máxima	Temperatura Mínima	Período de Molhamento Foliar	Ponto de Orvalho	Temperatura Média Compensada
1h37min	30,10°C	23,60°C	89,50%	23,32°C	25,30°C

Tabela 2: Dados Climáticos

Responda as questões abaixo baseando-se pelas informações acima:

1- Há presença da Ferrugem Asiática da Soja?

( ) Sim ( ) Não

2- Caso identifique a presença Ferrugem Asiática da Soja, selecione a severidade:

( ) baixa ( ) média ( ) alta

3- Se julgar necessário, justifique sua resposta:

_____

_____

_____

Teste 3:

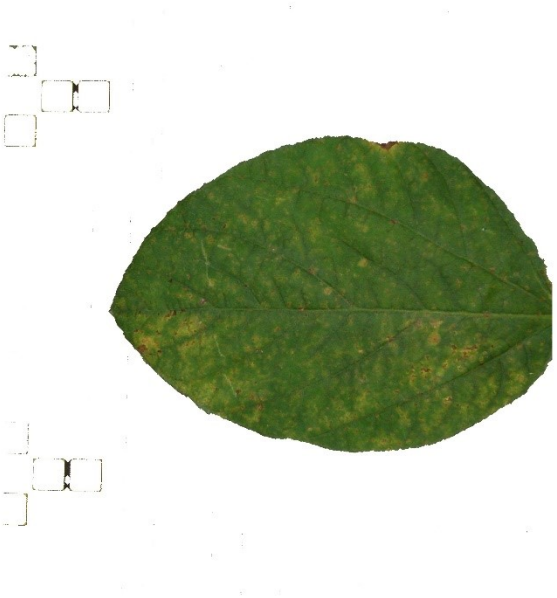


Imagem 3: Folha (DSC_0157)

Período Mínimo de Molhamento Foliar	Temperatura Máxima	Temperatura Mínima	Período de Molhamento Foliar	Ponto de Orvalho	Temperatura Média Compensada
0h00min	36,65°C	22,50°C	59,50%	18,33°C	28,14°C

Tabela 3: Dados Climáticos

Responda as questões abaixo baseando-se pelas informações acima:

1- Há presença da Ferrugem Asiática da Soja?

( ) Sim ( ) Não

2- Caso identifique a presença Ferrugem Asiática da Soja, selecione a severidade:

( ) baixa ( ) média ( ) alta

3- Se julgar necessário, justifique sua resposta:

_____

_____

_____

Teste 4:

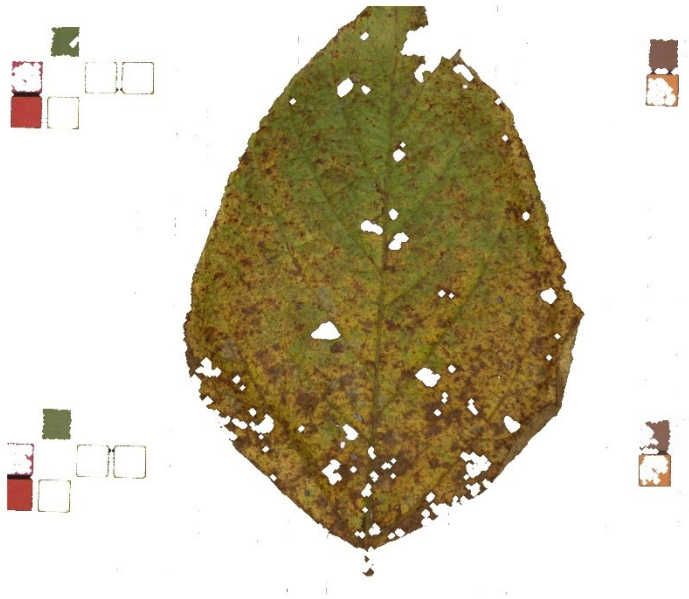


Imagem 4: Folha (DSC_0044)

Período Mínimo de Molhamento Foliar	Temperatura Máxima	Temperatura Mínima	Período de Molhamento Foliar	Ponto de Orvalho	Temperatura Média Compensada
0h52min	33,30°C	23,30°C	83,50%	22,96°C	26,15°C

Tabela 4: Dados Climáticos

Responda as questões abaixo baseando-se pelas informações acima:

1- Há presença da Ferrugem Asiática da Soja?

( ) Sim ( ) Não

2- Caso identifique a presença Ferrugem Asiática da Soja, selecione a severidade:

( ) baixa ( ) média ( ) alta

3- Se julgar necessário, justifique sua resposta:

_____

_____

_____

_____

Teste 5:

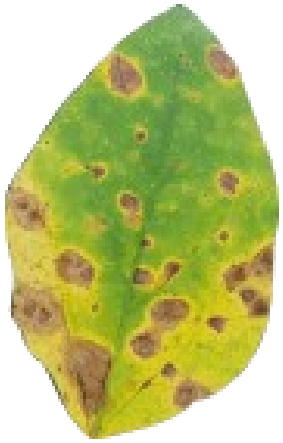


Imagem 5: Folha (DSC_0090)

Período Mínimo de Molhamento Foliar	Temperatura Máxima	Temperatura Mínima	Período de Molhamento Foliar	Ponto de Orvalho	Temperatura Média Compensada
0h00min	37,50°C	20,20°C	58,75%	19,16°C	27,08°C

Tabela 5: Dados Climáticos

Responda as questões abaixo baseando-se pelas informações acima:

1- Há presença da Ferrugem Asiática da Soja?

( ) Sim ( ) Não

2- Caso identifique a presença Ferrugem Asiática da Soja, selecione a severidade:

( ) baixa ( ) média ( ) alta

3- Se julgar necessário, justifique sua resposta:

_____

_____

_____

_____

_____

_____

Teste 6:

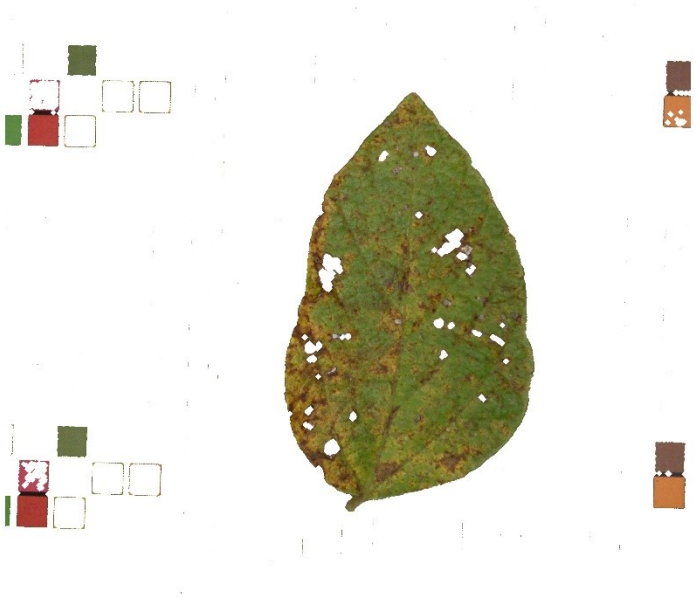


Imagem 6: Folha (DSC_0040)

Período Mínimo de Molhamento Foliar	Temperatura Máxima	Temperatura Mínima	Período de Molhamento Foliar	Ponto de Orvalho	Temperatura Média Compensada
0h00min	40,00°C	22,20°C	45,25%	15,84°C	28,56°C

Tabela 6: Dados Climáticos

Responda as questões abaixo baseando-se pelas informações acima:

1- Há presença da Ferrugem Asiática da Soja?

( ) Sim ( ) Não

2- Caso identifique a presença Ferrugem Asiática da Soja, selecione a severidade:

( ) baixa ( ) média ( ) alta

3- Se julgar necessário, justifique sua resposta:

_____

_____

_____

_____

Teste 7:

Imagem 7: Folha (DSC_0031)

Período Mínimo de Molhamento Foliar	Temperatura Máxima	Temperatura Mínima	Período de Molhamento Foliar	Ponto de Orvalho	Temperatura Média Compensada
0h00min	36,30°C	22,40°C	69,88%	21,23°C	27,10°C

Tabela 7: Dados Climáticos

Responda as questões abaixo baseando-se pelas informações acima:

1- Há presença da Ferrugem Asiática da Soja?

( ) Sim ( ) Não

2- Caso identifique a presença Ferrugem Asiática da Soja, selecione a severidade:

( ) baixa ( ) média ( ) alta

3- Se julgar necessário, justifique sua resposta:

_____

_____

_____

_____

Teste 8:

Imagem 8: Folha (DSC_0136)

Período Mínimo de Molhamento Foliar	Temperatura Máxima	Temperatura Mínima	Período de Molhamento Foliar	Ponto de Orvalho	Temperatura Média Compensada
1h00min	33,45°C	23,25°C	81,13%	22,76°C	26,28°C

Tabela 8: Dados Climáticos

Responda as questões abaixo baseando-se pelas informações acima:

1- Há presença da Ferrugem Asiática da Soja?

( ) Sim ( ) Não

2- Caso identifique a presença Ferrugem Asiática da Soja, selecione a severidade:

( ) baixa ( ) média ( ) alta

3- Se julgar necessário, justifique sua resposta:

_____

_____

_____

_____

Teste 9:

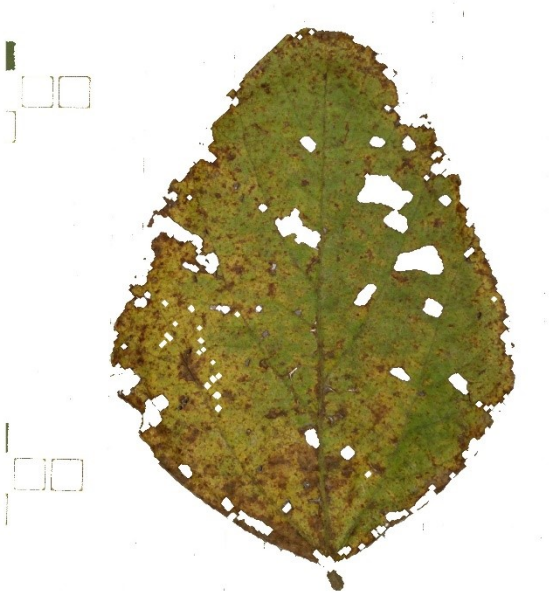


Imagem 9: Folha (DSC_0016)

Período Mínimo de Molhamento Foliar	Temperatura Máxima	Temperatura Mínima	Período de Molhamento Foliar	Ponto de Orvalho	Temperatura Média Compensada
1h19min	34,00°C	22,50°C	81,88%	23,39°C	26,57°C

Tabela 9: Dados Climáticos

Responda as questões abaixo baseando-se pelas informações acima:

1- Há presença da Ferrugem Asiática da Soja?

( ) Sim ( ) Não

2- Caso identifique a presença Ferrugem Asiática da Soja, selecione a severidade:

( ) baixa ( ) média ( ) alta

3- Se julgar necessário, justifique sua resposta:

_____

_____

_____

_____

Teste 10:

Imagem 10: Folha (DSC_0092)

Período Mínimo de Molhamento Foliar	Temperatura Máxima	Temperatura Mínima	Período de Molhamento Foliar	Ponto de Orvalho	Temperatura Média Compensada
0h00min	39,50°C	18,70°C	53,00%	17,80°C	28,30°C

Tabela 10: Dados Climáticos

<u>Responda as questões abaixo baseando-se pelas informações acima:</u>

1- Há presença da Ferrugem Asiática da Soja?

( ) Sim ( ) Não

2- Caso identifique a presença Ferrugem Asiática da Soja, selecione a severidade:

( ) baixa ( ) média ( ) alta

3- Se julgar necessário, justifique sua resposta:

_____

_____

_____

_____

_____

_____

_____

Teste 11:

Imagem 11: Folha (DSC_0017)

Período Mínimo de Molhamento Foliar	Temperatura Máxima	Temperatura Mínima	Período de Molhamento Foliar	Ponto de Orvalho	Temperatura Média Compensada
1h56min	34,00°C	22,00°C	80,63%	22,29°C	26,23°C

Tabela 11: Dados Climáticos

Responda as questões abaixo baseando-se pelas informações da Imagem 9 e da Tabela 9:

1- Há presença da Ferrugem Asiática da Soja?

( ) Sim ( ) Não

2- Caso identifique a presença Ferrugem Asiática da Soja, selecione a severidade:

( ) baixa ( ) média ( ) alta

3- Se julgar necessário, justifique sua resposta:

_____

_____

_____

_____

Teste 12:

Imagem 12: Folha (DSC_0042)

Período Mínimo de Molhamento Foliar	Temperatura Máxima	Temperatura Mínima	Período de Molhamento Foliar	Ponto de Orvalho	Temperatura Média Compensada
1h34min	34,75°C	23,30°C	76,50%	22,81°C	27,05°C

Tabela 12: Dados Climáticos

Responda as questões abaixo baseando-se pelas informações acima:

1- Há presença da Ferrugem Asiática da Soja?

( ) Sim ( ) Não

2- Caso identifique a presença Ferrugem Asiática da Soja, selecione a severidade:

( ) baixa ( ) média ( ) alta

3- Se julgar necessário, justifique sua resposta:

_____

_____

_____

_____

Teste 13:

Imagem 13: Folha (DSC_0091)

Período Mínimo de Molhamento Foliar	Temperatura Máxima	Temperatura Mínima	Período de Molhamento Foliar	Ponto de Orvalho	Temperatura Média Compensada
0h00min	37,50°C	19,60°C	59,25%	18,88°C	27,81°C

Tabela 13: Dados Climáticos

<u>Responda as questões abaixo baseando-se pelas informações acima:</u>

1- Há presença da Ferrugem Asiática da Soja?

( ) Sim ( ) Não

2- Caso identifique a presença Ferrugem Asiática da Soja, selecione a severidade:

( ) baixa ( ) média ( ) alta

3- Se julgar necessário, justifique sua resposta:

_____

_____

_____

_____

_____

_____

_____

_____

_____

Referências:

Yorinori, J. T., Nunes Junior, J., & Lazzarotto, J. J. (2004). Ferrugem" asiática" da soja no Brasil: evolução, importância econômica e controle. https://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/467712

Del Ponte, E. M., Godoy, C. V., Canteri, M. G., Reis, E. M., & Yang, X. B. (2006). Models and applications for risk assessment and prediction of Asian soybean rust epidemics. Fitopatologia Brasileira, 31, 533-544. https://doi.org/10.1590/S0100-41582006000600001

Lelis, V. D. P., Costa, L. C., Sediyama, G. C., & Vale, F. X. R. D. (2009). Favorabilidade ao desenvolvimento da ferrugem asiática da soja no estado de Minas Gerais. https://locus.ufv.br//handle/123456789/20313

Godoy, C. V., Seixas, C. D. S., Soares, R. M., Meyer, M. C., Costamilan, L. M., & Adegas, F. S. (2017). Boas práticas para o enfrentamento da ferrugem-asiática da soja. https://www.embrapa.br/busca-de-publicacoes/-/publicacao/1074899/boas-praticas-para-o-enfrentamento-da-ferrugem-asiatica-da-soja

Nunes, C. D. M., da Silva Martins, J. F., & Del Ponte, E. M. (2018). Validação de modelo de previsão de ocorrência da ferrugem asiática da soja com base em precipitação pluviométrica. *Embrapa Clima Temperado-Circular Técnica (INFOTECA-E)*, 1-13. https://ainfo.cnptia.embrapa.br/digital/bitstream/item/190270/1/CIRCULAR-199.pdf