UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

CARLOS RENATO BUENO

MONITORAMENTO ESTATÍSTICO APLICADO À CIÊNCIA DOS DADOS: UMA ABORDAGEM PARA VALIDAÇÃO CONTÍNUA DE MODELOS PREDITIVOS CLASSIFICATÓRIOS

SÃO CARLOS - SP
Maio, 2022

CARLOS RENATO BUENO

MONITORAMENTO ESTATÍSTICO APLICADO À CIÊNCIA DOS DADOS: UMA ABORDAGEM PARA VALIDAÇÃO CONTÍNUA DE MODELOS PREDITIVOS CLASSIFICATÓRIOS

Tese apresentada à Universidade Federal de São Carlos, como parte dos requisitos para obtenção do título de Doutor em Engenharia de Produção.

Orientador: Prof. Dr. Pedro Carlos Oprime

SÃO CARLOS - SP
Maio, 2022

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Engenharia de Produção

Folha de Aprovação

Defesa de Tese de Doutorado do candidato Carlos Renato Bueno, realizada em 27/05/2022.

Comissão Julgadora:

Prof. Dr. Pedro Carlos Oprime UFSCar)

Prof. Dr. Felipe Schoemer Jardim (UFF)

Profa. Dra. Fabiane Letícia Lizarelli (UFSCar)

Prof. Dr. Roberto Fernandes Tavares Neto (UFSCar)

Prof. Dr. Subhabrata Chakraborti (UA)

Profa. Dra. Marcela Aparecida Guerreiro Machado de Freitas (UNESP)

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

Dedico este trabalho a minha mãe e meu pai, que tanto se privaram para que os filhos pudessem se dedicar aos estudos, e com certeza, ainda nos protegem fora dos planos físicos.

Dedico este trabalho à minha esposa, que cuidou sozinha de nossa pequena, sempre que eu precisava de longas horas de concentração.

Dedico às pessoas anônimas, que estudam e trabalham para que nossa sociedade evolua, e, na maioria das vezes, não recebem o reconhecimento devido, enquanto outros acumulam fortunas em virtude da superficialidade da sociedade de nosso tempo.

AGRADECIMENTOS

Em primeiro lugar agradeço a Deus por manter meu ânimo perante todos os desafios que surgiram. Novamente meu muito obrigado aos meus pais José Carlos Bueno e Guiomar Sampaio Bueno, que, em sua simplicidade, ensinaram as maiores lições de caráter. Meu muito obrigado à minha esposa Rejane Patricia Pioto e aos meus filhos Eduardo Benedini Bueno e Laura Pioto Bueno, por compreender minha constante ausência.

Agradeço ao meu orientador Prof. Dr. Pedro Carlos Oprime por realmente ter me guiado nesta jornada e espero fazer jus a todo seu tempo dispendido, algo sem preço.

Agradeço ao Prof. Dr. Shuba Chakraborty pelo inestimável conselho de simplificar a matemática e ser profundo nos conceitos, e, como ele diz, quando não souber...pergunte.

Agradeço aos meus amigos de jornada em especial ao Juliano Endrigo Sordan, ao Daily Morales, e ao Giovanni Condé pelas inúmeras contribuições de conhecimento. Agradeço ao Erick Silva pelas incontáveis aulas de programação. Também não posso deixar de lembrar do Robson e do Lucas, com seu trabalho de secretaria rápido e preciso no suporte prestado aos alunos.

Agradeço a Profa. Dra. Juliana Keiko Sagawa pela inestimável contribuição no processo de qualificação e aos membros da banca: Profa. Dra. Fabiane Leticia Lizarelli; Prof. Dr. Roberto Fernandes Tavares Neto; Profa. Dra. Marcela Aparecida Guerreiro Machado de Freitas; e Prof. Dr. Felipe Schoemer Jardim, por terem aceitado participar deste momento tão importante da minha vida pessoal e acadêmica. Suas contribuições serão de extrema relevância para a finalização deste trabalho.

Agradeço a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior pelo apoio que presta à ciência neste país, tarefa não menos que árdua.

Deus não joga dados

(Albert Einstein)

RESUMO

Modelos preditivos são aqueles que se aplicam a dados de variáveis observáveis, denominadas independentes, inferindo-se sobre o comportamento de outra variável, observável ou não, denominada dependente. Caso particular, e muito utilizado, são os modelos classificatórios binários, em que a variável dependente pode receber dois valores: Sim ou não (positivo/negativo, sucesso/falha). A presente tese demonstra que ambientes operacionais cada vez mais digitalizados possibilitam aplicações mais complexas destes modelos classificatórios. Junte-se a isso a necessidade de aumento da competitividade nos negócios, através da busca por informações que reduzam custos ou aumentem a lucratividade das empresas: é a "Tempestade Perfeita", que aumenta a importância, o escopo, os impactos financeiros e o horizonte de tempo de uso desses modelos. Esse fenômeno ocorre tanto dentro da indústria com a Análise de Big Data (em inglês Big Data Analytics – BDA), quanto em outros setores, com o desenvolvimento da Ciência dos Dados (em inglês Data Science – DS). Entretanto, as condições de contorno, ou condições operacionais existentes, quando da criação do modelo, podem sofrer variações significativas, devido problemas técnicos na geração, captura, fluxo das informações ou mesmo nas relações entre as variáveis estudadas, que podem reduzir a qualidade da previsão do modelo criado. A revisão da literatura demonstrou que vários pesquisadores afirmam ser importante verificar periodicamente o desempenho de acertos e erros destes modelos, no entanto faltam critérios e métodos mais específicos que definam a frequência de checagem e tamanhos de amostra adequados a este monitoramento. Com o objetivo de preencher essa lacuna, utilizou-se os conceitos de Pesquisa em Ciência de Projeto para integrar os conceitos do Monitoramento Estatístico do Processo (em inglês: Statistical Process Monitoring – SPM) à métodos de elaboração de modelos aplicados no campo da DS. Na construção dessa integração relacionou-se as Fases I e II do SPM com um processo estruturado de análise de dados e geração de modelos, criando-se uma abordagem para sua validação contínua. Esta foi validada com o uso de técnicas analíticas e de simulação aplicadas ao índice Kappa de Cohen, resultando em critérios prescritivos para seu uso, apoiado em comparações baseadas no coeficiente de correlação de Matthews (em inglês: Matthews Correlation coefficient – MCC) e no índice de Youden. Verificou-se que gráficos de controle, baseados em Kappa, tem desempenho adequado para quantidades de amostra m=5 e tamanhos de amostra n=500, desde que o valor de Pe seja menor que 0,8. As simulações também demonstraram que para o monitoramento através de *Kappa* são necessárias menos amostras que para os outros índices estudados.

Palavras-chave: Indústria 4.0, Big Data, Mineração de Dados, Regressão Logística, Validação temporal, Modelos Preditivos, Validação Contínua.

ABSTRACT

Predictive models are those that apply to data of observable variables, called independent, inferring on the behavior of another variable, observable or not, called dependent. A particular case, and widely used, are the binary classificatory models, in which the dependent variable can receive two values: Yes or no (positive/negative, success/failure). The present thesis demonstrates that operational environments increasingly digitized allow more complex applications of these classification models. Add to this the need to increase business competitiveness, through the search for information that reduces costs or increases the profitability of companies: it is the "Perfect Storm", which increases the importance, scope, financial impacts, and horizon time of use of these models. This phenomenon occurs both within the industry with Big Data Analytics (BDA), and in other sectors, with the development of Data Science (DS). However, the boundary conditions, or existing operational conditions, when creating the model, can undergo significant variations, due to technical problems in the generation, capture, flow of information, or even in the relationships between the variables studied, which can reduce the quality of the forecast. of the created model. The literature review showed that several researchers claim that it is important to periodically check the performance of hits and misses of these models, however, there is a lack of more specific criteria and methods that define the checking frequency and sample sizes suitable for this monitoring. To fill this gap, the concepts of Project Science Research were used to integrate the concepts of Statistical Process Monitoring (SPM) with the methods of elaboration of models applied in the field of DS. In the construction of this integration, Phases I and II of the SPM were related to a structured process of data analysis and model generation, creating an approach for its continuous validation. This was validated using analytical and simulation techniques applied to the Cohen's Kappa index, resulting in prescriptive criteria for its use, supported by comparisons based on the Matthews correlation coefficient (MCC) and the index from *Youden*. Control charts, based on *Kappa*, were found to perform well for sample amounts of m=5 and sample sizes of n=500, provided the Pe value is less than 0.8. The simulations also showed that for monitoring through *Kappa* fewer samples are needed than for the other studied indices.

Keywords: Industry 4.0, Big Data, Data Mining, Logistic Regression, Temporal Validation, Predictive Models, Continuous Validation.

LISTA DE FIGURAS

Figura 1 - Mapa mental: o que fazer 13
Figura 2 - Mapa mental: Por que fazer 14
Figura 3 - Procedimento de Análise 16
Figura 4 - Aplicações da Regressão Logística no site Kaggle.com 17
Figura 5 - Etapas de desenvolvimento da Indústria 19
Figura 6 - Desenvolvimento da Probabilidade Estatística 20
Figura 7 - Evolução das Publicações sobre Indústria 4.0 22
Figura 8 - Classificação e Seleção de Tecnologias para a Indústria 4.0 26
Figura 9 - Framework Ambiente Sustentável Indústria 4.0 28
Figura 10 - Classificação e Seleção de Tecnologias para a Manufatura Inteligente 30
Figura 11 - Integração entre Mundo Físico e Virtual 32
Figura 12 - 4V's do Big Data 33
Figura 13 - Classificação dos problemas e técnicas referentes a análise do Big Data 34
Figura 14 - Espaço de Oportunidades de Produtividade 35
Figura 15 - Integração entre Mundo Físico, Virtual e o Big Data 36
Figura 16 - Arquitetura 5C para CPS 41
Figura 17 - Estrutura da Fábrica Inteligente 43
Figura 18 - Elementos integrantes de um CPS 44
Figura 19 - Roteiro para Manufatura Inteligente 45
Figura 20 - Evolução da Estrutura de Decisão na Manufatura 4.0 46
Figura 21 - Evolução dos Sistemas de Suporte à Decisão 50
Figura 22 - Maturidade de uso das ferramentas de análise no processo decisório 50
Figura 23 - Relacionamento entre DS, BI e BA 51
Figura 24 - O surgimento da Ciência dos Dados 55
Figura 25 - Exemplos de técnicas para modelos baseados em conhecimento 62
Figura 26 - Exemplos de técnicas para modelos baseados em dados 63
Figura 27 - Demonstração de pontos de convergência entre DS e o RDMP 67
Figura 28 - Validação de Modelos Preditivos 71
Figura 29 - Generalização como objetivo de validação de modelos estatísticos 71
Figura 30 - Validação Contínua 72
Figura 31 - Resultados da pesquisa 73
Figura 32 - Erros de classificação e o Limiar 78
Figura 33 - Matriz de Confusão ou Concordância 79
Figura 34 - Revocação e Precisão 80
Figura 35 - Curva ROC 86
Figura 36 - Pontos Particulares da curva ROC 87
Figura 37 - *AUC* da Curva ROC 88
Figura 38 - Exemplo de Carta de Controle 95
Figura 39 - Etapas Fase I SPM 97
Figura 40 - Evolução do Monitoramento Estatístico de Processo 100
Figura 41 - Abordagem de Integração do SPM a DS 105
Figura 42 - Diagrama de Ishikawa adaptado a análise de modelos 108
Figura 43 - Etapas para Análise de Desempenho dos Gráficos de *Kappa* 111
Figura 44 - Estrutura geral das Simulações 122
Figura 45 - Algoritmo geral utilizado na programação em Phyton 124
Figura 46 - Análise de normalidade de *Kappa* em função de m e n 127

Figura 47 - Análise da distribuição dos RL´s - *Kappa* 128
Figura 48 - Análise da distribuição do ARL´s - *Kappa* 128
Figura 49 - AARL para n fixo e m e η variáveis 131
Figura 50 - AARL para η fixo e m e n variáveis 132
Figura 51 - AARL para m fixo e η e n variáveis 133
Figura 52 - Demonstração dos pontos interpolados 135
Figura 53 - Análise de normalidade de *Kappa* OOC 136
Figura 54 - Análise de normalidade de *MCC* em função de m e n 137
Figura 55 - Análise da distribuição dos RL´s - MCC 138
Figura 56 - AARL para n fixo e m e η variáveis 140
Figura 57 - AARL para η fixo e m e n variável 141
Figura 58 - AARL para m fixo e η e n variável 142
Figura 59 - Análise de normalidade de *Youden* em função de m e n 144
Figura 60 - Análise da distribuição dos RL´s - Youden 144
Figura 61 - AARL para n fixo e m e η variáveis - *Youden* 147
Figura 62 - AARL para η fixo e m e n variável - *Youden* 148
Figura 63 - AARL para m fixo e η e n variável - *Youden* 149
Figura 64 - Análise da distribuição do ARL´s - *MCC* 153
Figura 65 - Análise da distribuição do ARL´s - *Youden* 153

LISTA DE QUADROS

Quadro 1 - Objetivos Específicos X Capítulos 7
Quadro 2 - Indicação das escolhas metodológicas do Trabalho 11
Quadro 3 - Iniciativas Governamentais de Suporte ao Desenvolvimento da Indústria 4.0 21
Quadro 4 - Termos correlatos utilizados para denominar a Indústria 4.0 22
Quadro 5 - Dimensão: Objetivos de Desempenho Indústria 4.0 24
Quadro 6 - Características técnicas de uma Fábrica Inteligente comparadas com a Fábrica Tradicional 27
Quadro 7 - Características da Qualidade dos Dados 29
Quadro 8 - Habilidades necessárias para adaptação do ser Humano à Indústria 4.0 47
Quadro 9 - Termos e definições comuns no contexto da DS 49
Quadro 10 - Conjunto de habilidades para as famílias de conteúdo de cargo 52
Quadro 11 - Detalhamento das habilidades e competências por categoria 53
Quadro 12 - Habilidades mais relevantes para o Cientista de Dados no GCS 54
Quadro 13 - Oportunidades de aplicação da Ciência dos Dados nas Indústrias 57
Quadro 14 - CRISP-DM vs. demais abordagens e Processo Refinado de Mineração de Dados 65
Quadro 15 - Descrição das Etapas do RDMP 66
Quadro 16 - Artigos que abordam o tema Validação Contínua 74
Quadro 17 - Matriz de Confusão para Cálculo de *Po* 82
Quadro 18 - Proporções baseadas na Matriz de Confusão para Cálculo de *Po* 83
Quadro 19 - Proporções baseadas na Matriz de Confusão para Cálculo de *Pe* 83
Quadro 20 - Interpretação do Índice Kappa 85
Quadro 21 - Critérios de análise para Cartas de Controle *X* (X Barra) 95
Quadro 22 - Implementação Clássica do SPM versus Oportunidades do Big Data 101
Quadro 23 - Dimensões de Qualidade dos Dados 108
Quadro 24 - Variáveis utilizadas no cálculo analítico *de Kappa* 111
Quadro 25 - Cenários analisados 115
Quadro 26 - Exemplo para a expansão de um registro 125
Quadro 27 - Cenários de Simulação para *Kappa* 126
Quadro 28 - Descrição dos Campos da Tabela 11 134
Quadro 29 - Cenários de Simulação para *MCC* 138
Quadro 30 - Cenários de Simulação para *Youden* 145

LISTA DE TABELAS

Tabela 1 - Categorias de pesquisa sobre Indústria 4.0 22
Tabela 2 - Principais disciplinas em programas de bacharelado 54
Tabela 3 - Trabalhos por área do conhecimento 73
Tabela 4 - Resultados para n e m fixos, variando-se δ e *Pe* 117
Tabela 5 - Resultados para n e δ fixos, variando-se m e *Pe* 118
Tabela 6 - Resultados para δ e m fixos, variando-se n e *Pe* 119
Tabela 7 - AARL para maiores valores de n com δ = 0,05 120
Tabela 8 - AARL para maiores valores de n com δ = 0,1 120
Tabela 9 - Informações dos parâmetros gerados 123
Tabela 10 - Resultados de AARL nos cenários de simulação de *Kappa* 130
Tabela 11- Análise de AARL em função de $\boldsymbol{\delta}$ 135
Tabela 12 - Resultados de AARL nos cenários de simulação de *MCC* 139
Tabela 13 - Resultados de AARL nos cenários de simulação de *Youden* 146
Tabela 14 - Comparações da Distribuição de ARL entre os indicadores (m=10, n=300, $\boldsymbol{\eta}$ = 1) 152

LISTA DE ABREVIATURAS E SIGLAS

ANN - Artificial Neural Network

AARL - Média do comprimento médio de frequência

ARL - Comprimento médio de frequência

AUC - Área sob a Curva da Curva Característica de Operação do Receptor

BD - Big Data

BDA - Análise do Big Data

BI - Inteligência de Negócios

CBM - Manutenção Baseada em Condição

CDF - Função de Densidade Acumulada

CL - Limites de Controle

CRISP-DM - Processo padrão entre setores para mineração de dados

CPS - Sistemas Cibernético Físicos

CPPS - Sistemas de Produção Cibernéticos Físicos

CUSUM - Soma Cumulativa

DL - Aprendizado profundo de Máquina (em inglês – Deep Learning)

DM - Mineração de Dados

DM & KD - Mineração de Dados e descoberta de Conhecimento

DS - Ciência dos Dados

e.g - Por Exemplo

EWMA - Média Móvel Exponencialmente Ponderada

FDD - Detecção e Diagnóstico de Falha

FAR - Taxa de Falso Alarme

FN - Falso Negativo

FNR - Taxa de falsos negativos

FP - Falso Positivo

FPR - Taxa de falsos positivos

FTC - Sistema de recuperação automática de Falha

GE - General Eletric

GCS - Gerenciamento da Cadeia de Suprimento

IC - Sob Controle

i.i.d - Independentes e identicamente distribuídas

IoS - Internet dos Serviços

IoT - Internet das Coisas

IoTS - Internet das Coisas e Serviços

IPM - Monitoramento de Processos Industriais

KDD - Descoberta de Conhecimento em Bancos de Dados

LC - Limite de Controle

LCL - Limite Inferior de Controle

MCC - Coeficiente de Correlação de Matthews

MDT - Critério para o Limiar que minimiza a diferença entre sensibilidade e especificidade

ML - Aprendizado de Máquina (em inglês – Machine Learning)

MSPM - Monitoramento Estatístico do Processo Multivariado

MST - Critério para o Limiar que considera a maximização da soma entre sensibilidade e especificidade

NPV - Valor Preditivo Negativo

OEE - Eficiência Global de Equipamentos

OOC - Fora de Controle

PDF - Função Densidade de Probabilidade

PdM - Manutenção baseada em estatística

PHM - Gestão de Saúde e Prognóstico de equipamentos

PM - Monitoramento de Processos

PPV - Valor Preditivo Positivo

RBS - Revisão Bibliográfica Sistemática

RDMP - Processo Refinado de Data Mining

RF - Florestas aleatórias

RFID - Dispositivo de identificação por rádio frequência

ROC - Característica de Operação do Receptor

RL - Comprimento de Sequência

RSP - Probabilidade Real de Sucesso

SPC - Controle Estatístico do Processo

SPM - Monitoramento Estatístico do Processo

SVR - Regressão por Vetores de Suporte

TI – Tecnologia de Informação

TN - Verdadeiros Negativo

TNR - Taxa de Verdadeiros Negativos

TP - Verdadeiro Positivo

TPR - Taxa de Verdadeiros Positivos

TIC - Tecnologias de Informação e Comunicação

UCL - Limite Superior de Controle

WoS - Web of Science

SUMÁRIO

1 INTRODUÇÃO 1
1.1 PROBLEMA DE PESQUISA 3
1.2 OBJETIVO GERAL DO TRABALHO 7
1.3 OBJETIVOS ESPECÍFICOS 7
1.4 JUSTIFICATIVA DO TRABALHO 8
1.5 ORGANIZAÇÃO DO TEXTO 9
2 MÉTODO 10
2.1 CARACTERIZAÇÃO E NATUREZA DA PESQUISA 10
2.2 PROCEDIMENTO DE PESQUISA 11
2.3 PROCEDIMENTO DE ANÁLISE 15
3 INDÚSTRIA 4.0, TECNOLOGIA E MODELOS ESTATÍSTICOS 19
3.1 INDÚSTRIA 4.0: UM NOVO MODELO INDUSTRIAL? 20
3.2 DEMANDAS COMPETITIVAS E OS OBJETIVOS DA INDÚSTRIA 4.0 23
3.3 ELEMENTOS DA INDÚSTRIA 4.0 25
3.4 IoT e BIG DATA: DESAFIOS PARA UMA NOVA INDÚSTRIA 29
3.5 AMBIENTES CIBERNÉTICO – FÍSICOS, PESSOAS E MODELOS PREDITIVOS . 38
4 CONHECIMENTO, HABILIDADES E INTEGRAÇAO NO AMBIENTE DO DATA SCIENCE 48
4.1 O AMPLO ESPECTRO DA CIÊNCIA DOS DADOS 48
4.2 EXEMPLOS DE APLICAÇÕES DA DS NA INDÚSTRIA 56
4.3 PROCESSOS DE MODELAGEM 61
5 VALIDAÇÃO DE MODELOS DE CONHECIMENTO 69
5.1 VALIDAÇÃO CONTÍNUA DE MODELOS PREDITIVOS CLASSIFICATÓRIOS ... 69
5.2 ERROS DE CLASSIFICAÇÃO EM MODELOS BINÁRIOS 77
5.3 MÉTRICAS PARA AVALIAÇÃO DE MODELOS CLASSIFICATÓRIOS BINÁRIOS 81
5.3.1 Índice *Kappa* de Cohen 81
5.3.2 Área sob a Curva da Curva Característica de Operação do Receptor 86
5.3.3 Coeficiente de Correlação de Matthews (*MCC*) 89

5.3.4 Índice de Youden **91**
6 MONITORAMENTO ESTATÍSTICO DE PROCESSO, SPM 93
6.1 CONCEITOS BÁSICOS DE SPM 93
6.2 NOVAS APLICAÇÕES DO SPM 98
7 INTEGRAÇÃO DO SPM AOS MODELOS DE GERAÇÃO DE CONHECIMENTO 104
7.1 PROPOSTA DE ABORDAGEM DE INTEGRAÇÃO ENTRE O SPM E A CIÊNCIA DOS DADOS 104
7.2 MODELO ANALÍTICO DE DESEMPENHO PARA *KAPPA* 110
7.2.1 Procedimento de Análise **110**
7.2.2 Desempenho Teórico dos gráficos de controle de *Kappa* com $P0$ conhecido **111**
7.2.3 Desempenho Teórico dos gráficos de controle de *Kappa* com $P0$ desconhecido **113**
7.2.4 Análise de desempenho dos Gráficos de Controle de *Kappa* **114**
7.2.5 Considerações sobre os resultados **121**
7.3 SIMULAÇÕES ESTOCÁSTICAS DE DESEMPENHO 121
7.3.1 Caso 1: Índice *Kappa* de Cohen **127**
7.3.2 Caso 2: Coeficiente de Correlação de Matthews **137**
7.3.3 Caso 3: Índice de Youden **143**
8 ANÁLISE E DISCUSSÃO DOS RESULTADOS 151
9 CONCLUSÕES 154
10 BIBLIOGRAFIA 158
APÊNDICE A 175
APÊNDICE B 180
APÊNDICE C 181
APÊNDICE D 182
APÊNDICE E 188
APÊNDICE F 194
APÊNDICE G 200

1 INTRODUÇÃO

Desde o surgimento da Internet, bem como os avanços nas tecnologias relacionadas a capacidade de transmissão e armazenamento dos dados, que teve seu início aproximadamente na última década do século XX, a quantidade de dados tem crescido a uma velocidade exponencial. Estimou-se que no final de 2020 o volume de dados na Internet ultrapassou a 40 trilhões de gigabytes (EXAME, 2021). Diante disso, um dos grandes desafios é absorver, processar e analisar essa quantidade imensa de dados, que são, em muitos casos, não estruturados, dificultando ainda mais o seu processamento e análise.

Para aprimorar a competitividade dos seus negócios, as empresas têm utilizado com frequência cada vez mais informações e dados de fontes primarias e secundarias. Os benefícios do uso desses dados estão em construir uma base do conhecimento dos clientes, melhorar a logística interna e externa, bem como melhorar os processos de decisões, dentre outros benefícios (SANTOS *et al*., 2018; JOPPEN *et al*., 2019).

Em relação a capacidade de coletar, armazenar, processar e analisar esses dados, Landset *et al* (2015) indicaram várias plataformas de domínio público (em inglês - open source), tais como Hadoop, Mahout, MLlib, SAMOA, H2O, Spark, Flink e Storm. Esses recursos propiciam o uso de técnicas estatísticas avançadas e complexas, bem como o uso de algoritmos de aprendizado de máquina (em inglês Machine Learning – ML e Deep Learning – DL). O desenvolvimento destas tecnologias foram determinantes para o surgimento da Ciência dos Dados (em inglês Data Science – DS) e a análise do Big Data (em inglês Big Data Analytics – BDA), cujos objetivos são extrair e entender as relações complexas entre diversas variáveis, suprindo os tomadores de decisão com informações em tempo real, precisas e acuradas (HE; WANG, 2018a; TAO *et al*., 2018; MAURO *et al*. 2017).

Mauro *et al*. (2017) analisaram as características das atividades do BDA e identificaram que essas se confundem com as características do DS; porém verifica-se que não há consenso sobre o significado desses dois elementos, o que abre espaço para especulações e novas teorias e conceitos sobre eles. O termo Mineração de dados (em inglês Data Mining - DM) é correlacionado ao DS; no entanto, alguns autores entendem que o DM é uma etapa do processo do DS, e não é compreendido como todo o processo (DEBUSE *et al*., 2001; MARISCAL, *et al*. 2010; FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996a). Outro termo comumente utilizado por estes autores, especialmente aplicado na área de sistemas computacionais, é o KDD - Knowledge Database Discovery, que é entendido como uma metodologia para criação

de modelos ou Descoberta de Conhecimento em Bancos de Dados (KDD), e que possui várias das etapas do processo de DS, sendo estes aspectos discutidos nos capítulos 3 e 4.

Os especialistas em DS afirmam a necessidade do uso de modelos estatísticos avançados, tais como os modelos de dependência ou modelos causais, para predizer resultados e riscos, a partir de variáveis explanatórias (PARA *et al*., 2019; SHMUEI; KOPPIUS, 2011).

O uso de tecnologias (BDA, DS e DM) tem catalisado mudanças na gestão (ZHONG *et al*., 2017; TAO; ZHANG, 2017; XU; DUAN, 2018;MAYNARD, 2015)

Toda essa gama de tecnologias, desenvolvida nas primeiras décadas do século XXI, constitui um grande guarda-chuva denominado de Indústria 4.0, considerada a quarta revolução industrial (WANG *et al*., 2016; DIAZ-ROZO; BIELZA; LARRAÑAGA, 2017; ZHONG *et al*., 2017; KAMBLE; GUNASEKARAN; GAWANKAR, 2018).

Embora tal denominação não seja consenso no meio acadêmico, utiliza-se o termo Indústria 4.0 quando se refere às transformações tecnológicas em diversos ambientes, não exclusivamente no empresarial (APREDA *et al*., 2016). No Brasil, indicam-se como tecnologias constituintes da Indústria 4.0 elementos como biologia sintética (Synbio), cidades inteligentes, construções inteligentes, mobilidade inteligente, redes inteligentes, logística inteligente, produtos inteligentes, dentre outros (MINISTÉRIO DA INDÚSTRIA E COMÉRCIO E SERVIÇOS, 2019). Incluem-se também nesse guarda-chuva, segundo Kang *et al*. (2016) e Lu (2017), a Internet das Coisas (em inglês -Internet of things - IoT) e a Internet dos Serviços (em inglês Internet of Services – IoS) como elementos da Indústria 4.0.

As tecnologias de Iot, agregadas a outras tecnologias, aumentam a produtividade dos processos industriais. Por sua vez, a aplicação de técnicas de BDA possibilita predizer resultados dos sistemas, e, consequentemente, melhorar sua eficiência (CHENG et al., 2017). Sobre isso, Reis e Gins (2017) identificaram três facilitadores principais que melhoram o desempenho das organizações: dados, tecnologia e capacidade analítica.

Sharma *et al*. (2014) constataram que a principal dificuldade para a indústria é o entendimento das relações entre dados de entrada e dados de saída dos processos; em outras palavras, compreender as relações de causalidade. Para isso, de acordo com O'Donovan *et al*.(2015b) há a necessidade de transformação do ambiente de manufatura de reativo para proativo. Porém, há desafios em estabelecer padrões de comunicação entre os equipamentos, pois as estruturas de dados e suas interfaces são heterogêneas. Para He e Wang (2018) a aplicação do BDA auxilia nessa transformação, mitigando os problemas de comunicação para

um ambiente proativo. O seu objetivo é entender as relações entre variáveis causais e as variáveis dependentes, e assim dar suporte às ações preventivas. Nesse aspecto, os modelos estatísticos têm papel central.

A literatura aponta gargalos, como a falta de mão de obra capacitada, para que as organizações usufruam dos benefícios do uso de modelos estatísticos, aplicados em condições de processamento online, com grande volume e variedade de dados de baixa qualidade (devido falhas de leitura, quebras de sensores, leituras incorretas). Essa capacitação é complexa pois engloba habilidade em programação, domínio de estatística e matemática (LANDSET *et al*., 2015).

Feitas a considerações acima, é possível extrair alguns problemas de pesquisa que contribuam para o campo dos estudos sobre modelos preditivos aplicados aos negócios e processos.

1.1 PROBLEMA DE PESQUISA

Percebe-se que o ambiente de negócios tem se tornado mais complexo no século XXI, em razão de diversos fatores, enumerando-se nesse trabalho os mais relevantes para o desenvolvimento de uma tese que buscará avaliar os métodos de predição baseados em modelos matemáticos e estatísticos. Os principais fatores deste século XXI, que tem transformado os negócios, são as inovações tecnológicas da informação, também chamada de transformação digital, os problemas de sustentabilidade e da preservação do meio ambiente.

A inovação tecnológica e as transformações digitais nos remetem ao estudo de modelos estatísticos e matemáticos, à inteligência artificial, à realidade aumentada, ao aprendizado profundo (em inglês *deep learning*), à algoritmos genéticos, dentre outros. Por consequência, ao que se denomina de I 4.0, como sendo um novo paradigma organizacional. Esse novo paradigma se define pelo uso de novos recursos tecnológicos já disponíveis, bem como das metodologias estatísticas, com a perspectiva que os gestores tomem decisões embasadas em dados (e ciência), de modo proativo, e com significativos ganhos financeiros e ambientais para a organização.

A partir do século XXI observa-se um aumento significativo do uso de modelos classificatórios preditivos. Este fenômeno está associado a vários elementos presentes nas transformações da I 4.0, como por exemplo o Big Data e o BDA. Modelos estatísticos como regressões logísticas, regressões lineares multivariadas, modelos probit, e modelos baseados

em dados como: redes neurais artificiais, árvores de decisão, dentre outros, aplicam-se no apoio a tomada de decisão em situações de maior impacto financeiro, bem como aumenta-se a abrangência das aplicações para outras áreas (ÇINAR *et al*., 2020; DIEZ-OLIVAN *et al*., 2019; SIRYANI; TANJU; EVELEIGH, 2017; NAMDARI; JAZAYERI-RAD, 2014; STETCO *et al.*, 2019). Essa crescente importância no uso dos modelos preditivos, sugere a necessidade da revisão dos atuais métodos de validação da sua eficácia.

O background do corpo teórico que sedimenta os caminhos para aplicação de modelos preditivos está presente em diversas áreas da ciência. A teoria corrente do processo de modelagem estatística, aplicado no escopo da DS, tem como uma de suas etapas, a validação destes modelos. A prática e a teoria recomendam a divisão dos dados disponíveis, para essa modelagem, em dois blocos. O primeiro bloco contendo 70% dos dados utiliza-se na construção do modelo (ou treinamento do modelo), e o segundo bloco de dados é aplicado para validar o modelo gerado (AUSTIN, 2008; HAIR *et al*., 2009; SPENCER, 2014; EVERITT *et al*., 2011; LI *et al.*, 2020; DU *et al*., 2021; ALTMAN *et al*., 2009; HUANG *et al*., 2018). Propõe-se várias métricas para tal finalidade, entretanto não há consenso sobre qual a melhor, ou mais adequada, para cada contexto de aplicação. Além disso, uma possível fragilidade destas propostas de validação é não considerar a necessidade de validação contínua dos modelos, uma vez que podem ser utilizados por longos períodos.

Dentre os diversos modelos preditivos, os classificatórios binários são bastante utilizados devido sua simplicidade e relevância. Um exemplo prático deste tipo de modelo são os testes de Covid-19, que classificam o indivíduo entre duas categorias: positivo para a doença ou negativo para a doença. Quanto maior o nível de acerto, ou seja, classificar quem é positivo como positivo e quem é negativo como negativo, maior a qualidade do modelo. Du *et al*. (2021) criaram um modelo onde resultados de exames de saúde de baixo grau de complexidade (variáveis de entrada) foram aplicados para prever se um paciente está ou não está com covid. Esse resultado é de relevância pois possibilita o acesso de comunidades a exames menos complexos e de menor custo que o RT-PCR. Além da área de saúde, estes tipos de modelos estão presentes nas mais diversas áreas, como a área financeira, manutenção de equipamentos, controle de processos, dentre outras (WESSLER *et al.*, 2019; DAINES *et al*., 2019; LI *et al*., 2020; CARVALHO *et al*., 2019).

Na literatura, o procedimento de validação mais usual para estes modelos utiliza uma matriz cruzada, também denominada de matriz de confusão ou de contingência. A validação do modelo baseia-se em seu desempenho com relação a índices, calculados a partir dessa matriz.

Dentre os índices mais utilizados estão a área sob a curva da curva característica de operação do receptor (em inglês Área Under Curve of Receiving Operator Characteristic Curve – AUC of ROC curve), precisão, revocação, especificidade, sensibilidade (CHICCO; WARRENS; JURMAN, 2021; BEN-DAVID, 2008). Entretanto, como será demonstrado no item 5.3, não há consenso sobre qual o melhor índice para cada situação problema.

A importância de processos eficazes de validação é discutida por Lang e Bolton (1991a) e Lang e Bolton (1991b) para métodos bioanalíticos. Apesar de seu objeto de estudo não ser um modelo classificatório, estes autores utilizam cartas de controle para validar continuamente a qualidade dos resultados gerados por análise química de precisão, demonstrando a importância do monitoramento de variações, mesmo em ambientes muito controlados, como o da química analítica. No campo da DS, Sornette *et al*. (2007) afirmam que um modelo é progressivamente validado através da confirmação de repetidas predições, desta forma sugerindo a necessidade de seu monitoramento estatístico. Kuncheva (2009) analisa questões referentes a detecção de mudanças de contexto (em inglês Concept Drift), no campo de aprendizado de máquinas. A mudança de contexto ocorre quando um determinado comportamento não é mais adequadamente previsto pelo algoritmo de predição. A autora analisa gráficos de controle padrão, e outros métodos, para detecção destas variações, entretanto não justifica o tamanho e quantidade de amostras escolhidas nas análises, aspecto fundamental do monitoramento estatístico do processo (em inglês Statistical Process Monitoring – SPM). Na área da saúde, Minne *et al*., 2012a, Minne *et al*., 2012b e Minne *et al.*, 2012c, utilizam gráficos de controle para comparar diferentes índices de validação, também sem justificar os tamanhos de amostra escolhidos ou discutir aspectos relevantes das Fases I e II do SPM, afirmando que estudos deveriam ser feitos nesse sentido.

Cintolo-gonzalez *et al*. (2017) e Dijkland *et al.* (2020) analisaram diversos modelos de predição e concluíram que a validação contínua é uma ferramenta importante para a implementação de melhorias dos algoritmos criados. Su *et al*. (2018) afirmam que o desempenho de modelos de predição pode se deteriorar com o tempo, sendo necessário o constante monitoramento dos seus resultados. Outra evidência sobre a importância da validação contínua encontra-se em Feidas *et al*. (2018), que avaliam modelos meteorológicos em vários momentos no tempo. Hoffmann *et al*. (2019) abordam a relevância destes processos contínuos de validação ao afirmarem que quaisquer índices, utilizados para caracterizar a qualidade de um modelo, são estimativas do real valor da população, uma vez que são calculados com base em amostras, e devem ser definidos em função de um intervalo de confiança. Por fim, Oduro *et*

al. (2016) e Myllyaho *et al*. (2021) afirmam que o dinamismo dos ambientes desatualizam os modelos, sugerindo que sejam validados continuamente.

A partir dessas informações percebe-se que o assunto ainda está aberto e há muito a ser pesquisado sobre a validação contínua de modelos preditivos classificatórios. Trabalhos recentes, como o de Eker *et al*. (2019), que faz uma revisão bibliográfica sistemática (RBS) sobre validação de modelos, não identificaram nas suas pesquisas o assunto "validação contínua", e o termo nem mesmo foi incluído no mapa de palavras chave.

Observa-se que, até o presente momento, a literatura indica a necessidade da validação contínua dos modelos preditivos, porém, os trabalhos sobre SPM nesse campo, que poderiam preencher essa lacuna, limitam-se aos seus aspectos mais elementares. Isso corrobora afirmações de Woodall e Montgomery (2014), Woodall (2016) e Jones-Farmer e Stevens (2016) sobre questões fundamentais como as Fases I e II do SPM e a influência do número e tamanho das amostras, para construção de gráficos de controle, não serem analisadas em maior profundidade.

Feitas estas considerações iniciais, algumas perguntas podem ser formuladas: *i)* é possível utilizar o SPM no contexto do BDA ou da DS, no processo de validação dos modelos, e também, como ferramenta de apoio a todo o processo decisório que o modelo auxilia? *ii)* os Índices *Kappa* de Cohen, *Youden* e Coeficiente de Correlação de Matthews (em inglês *Matthews Correlation Coefficient* – MCC), são adequados para a operacionalização do SPM no contexto estudado? *iii)* é possível melhorar os modelos preditivos inserindo o monitoramento estatístico do processo?

Com base nas perguntas acima, este trabalho propõe o uso do SPM na validação de modelos classificatórios, incorporando-o à procedimentos tradicionais, bem como à análise do seu desempenho, considerando a Fase I, em que são estimados os parâmetros estatísticos, e a Fase II de monitoramento (CHAKRABORTI; HUMAN; GRAHAM, 2009; JENSEN *et al.*, 2006; JARDIM; CHAKRABORTI; EPPRECHT, 2020; CHEN; SONG, 2012; JARDIM; CHAKRABORTI; EPPRECHT, 2019).

A justificativa para o uso do SPM como técnica de validação contínua deve-se ao fato de que, além de detectar variações significativas na variável de controle, seu uso auxilia na identificação das causas dessas variações e, assim, melhora-se o processo decisório, uma vez que decisões errôneas ocorrem tanto devido a imprecisão dos modelos estatísticos quanto a problemas no processo de coleta e análise de dados.

1.2 OBJETIVO GERAL DO TRABALHO

A partir das questões de pesquisa, entende-se que existem condições de contorno variáveis que afetam as respostas dos modelos, o que se configura em oportunidades para uso do SPM no monitoramento de modelos preditivos, uma vez que não há estudos completos que o integrem de forma estruturada com ferramentas da DS, no contexto do Big Data. O objetivo geral desta tese é estudar e verificar a adequação do índice *Kappa* de Cohen, doravante denominado como *Kappa*, na validação contínua de modelos classificatórios binários. Concomitantemente compará-lo com técnicas concorrentes: *Youden* e MCC.

1.3 OBJETIVOS ESPECÍFICOS

Como objetivos específicos propõem-se:

- Avaliar a aplicabilidade de modelos preditivos em projetos de Data Science e Big Data.
- Identificar as restrições e oportunidades no uso do *Kappa* como métrica na validação de modelos.
- Testar e validar o uso do SPM conjuntamente com o *Kappa*.
- Comparar o desempenho do SPM/Kappa com modelos concorrentes (SPM/*MCC* e SPM/*Youden*)
- Propor uma abordagem para validação contínua de modelos preditivos.

Quadro 1 - Objetivos Específicos X Capítulos

Avaliar a aplicabilidade de modelos preditivos em projetos de Data Science e Big Data	3.4; 3.5; 4; 5
Identificar as restrições e oportunidades no uso do índice Kappa como métrica na validação de modelos	5
Testar e validar o uso do SPM conjuntamente com o índice Kappa	5.3.1; 7.2; 7.3
Comparar o desempenho do SPM/Kappa com modelos concorrentes (SPM/MCC e SPM/Youden)	7.3
Propor um Framework para validação contínua de modelos preditivos	7.1

Fonte: Elaborado pelo Próprio Autor

1.4 JUSTIFICATIVA DO TRABALHO

A quantidade de dados disponíveis cresce a uma velocidade exponencial. As organizações, em sua busca por melhores posições competitivas, podem usá-los para criar valor nas suas operações, na construção de conhecimentos e no suporte às decisões. Entretanto a complexidade no processamento e análise desses dados também cresce proporcionalmente, e com isso a necessidade de verificar continuamente se o conhecimento gerado hoje, é aplicável amanhã.

O campo de pesquisa de técnicas para o monitoramento de processos tem tido, a partir dos trabalhos de Shewhart, significativo desenvolvimento (LIZARELLI *et.al*, 2016). Novas aplicações em diferentes áreas, tais como na medicina, na agricultura, em serviços e no campo das ciências sociais têm sido publicados em revistas científicas por diversos autores com destaque para Woodall que tem indicado temas de pesquisa sobre SPC; Castagliola, Celano, Chakraborti que avaliam os efeitos da estimativa de parâmetros sobre o desempenho de gráficos de controle, para citar alguns. Recentemente, na segunda década do Século XXI, pesquisas sobre aplicação do SPC e DS tem sido indicado por alguns autores. Sobre isso Woodall (2016) e Steinberg (2016) sugerem modificações nos métodos tradicionais de SPM, para adaptar ao ambiente de Big Data. Após revisão da literatura observou-se que há poucos trabalhos que abordam essa temática, ou seja, o uso de gráficos de controle no ambiente de Big Data ou DS.

Escobar *et al*. (2018) afirmam que o SPM deve ser reestudado para aplicações onde o tempo de maturação de conhecimento técnico em novos processos é muito reduzido. Segundo os autores, lançamentos de produtos em prazos cada vez menores pressionam para que testes finais de novos processos aconteçam com a fabricação de seus itens. Os autores também citam outros pontos a serem endereçados: *i)* quando a velocidade do processo é muito maior que o tempo necessário para o resultado da coleta e análise dos dados, possibilitando a não detecção de alterações entre estes momentos (por exemplo, casos onde o tempo para ensaios é longo, ou existe necessidade de baixa frequência de amostragem devido testes destrutivos), *ii)* quando o processo demanda o estabelecimento de buffers de itens semiprocessados que só podem ser avaliados após um determinado intervalo de tempo, dificultando a detecção e correção imediata das falhas desse processo *iii)* incerteza na definição das características da qualidade em processos que não atingiram sua maturidade. Percebe-se que o uso de modelos preditivos estáveis é fundamental na solução dessas questões, uma vez que é necessário conhecer o resultado do processo antes mesmo de sua finalização.

Acrescenta-se a importância econômica e social de previsões por meio de modelos

matemáticos e estatísticos. Por exemplo Özkundakci *et al*. (2018) apresenta aplicações do uso de modelos preditivos no campo jurídico relacionadas a questões ambientais; Abell *et al*. (2017), Carvajal Soto, Tavakolizadeh, Gyulai (2019) por sua vez mostram aplicações de modelos preditivos no ambiente de produção; na área de saúde Strobl *et al*. (2016) e Liu *et al*. (2020) afirmam que as condições de operação de um modelo variam ao longo do tempo; pois métodos diagnósticos podem mudar e equipamentos e protocolos de tratamento melhorarem; Li *et al*. (2020) descrevem aplicações de modelos na área financeira para análise de crédito.

A revisão da literatura demonstrou que os poucos estudos que utilizam o SPM, como apoio a processos de validação, pertencem, em sua maioria, a área de Saúde e não discutem todas as convergências entre o SPM e a DS, restringindo a discussão sobre variações aos aspectos pertinentes à qualidade da modelagem, (MINNE *et al*., 2012a; MINNE *et al*., 2012b; MINNE *et al.*, 2012c). Outra lacuna é a falta de discussão mais detalhada sobre as Fases I e II do SPM.

Considerando os avanços na área do SPM e pela carência de estudos relacionando essa técnica ao campo da DS e Big Data, justifica-se o desenvolvimento dessa tese.

1.5 ORGANIZAÇÃO DO TEXTO

Além da introdução, o trabalho possui outros 8 capítulos com os seguintes conteúdos:

O Capítulo 2 descreve o método aplicado no desenvolvimento da pesquisa. Através de mapas mentais resumem-se os conceitos abordados e as questões a se responder ao longo da trilha do trabalho, relacionando-se os capítulos do texto ao método proposto.

Os Capítulos 3 a 6 compõem o pilar de fundamentação teórica, utilizados tanto para a avaliação da lacuna de pesquisa proposta bem como para a elaboração de novas proposições.

No Capítulo 7 relaciona-se o SPM e a DS elaborando-se proposta para sua integração. Demonstra-se como o SPM pode se integrar e contribuir na avaliação do processo de geração do conhecimento e nos processos decisórios suportados pela aplicação das ferramentas de DS. Os tópicos 7.2 e 7.3 validam a integração proposta através de modelos teórico analíticos e de simulação.

A análise e discussão dos resultados, conclusões e referências bibliográficas estão descritas nos capítulos 8, 9 e 10 respectivamente.

2 MÉTODO

Neste capítulo são apresentados os métodos e procedimentos de pesquisa selecionados para que os objetivos especificados na introdução sejam alcançados. Inicialmente apresentam-se as características e natureza da pesquisa; em seguida os procedimentos de coleta e análise de dados.

2.1 CARACTERIZAÇÃO E NATUREZA DA PESQUISA

De acordo com Lakatos e Marconi (1995) e Gil (1999), os métodos de pesquisa são classificados em: Indutivo, Dedutivo, Dialético e Hipotético Dedutivo. O objetivo desta tese é propor, desenvolver e aplicar um método de validação periódica de modelos preditivos classificatórios pela abordagem clássica do SPM, utilizando diversos índices de concordância; assim, em síntese, a pesquisa utiliza o método Indutivo pois busca-se soluções particulares para se chegar a conclusões mais genéricas na forma de uma teoria.

Quanto a classificação em relação à natureza da pesquisa, segundo os conceitos e definições de Silva e Menezes (2000), esta é uma pesquisa aplicada, pois tem como foco a solução de problemas práticos específicos. No que diz respeito a sua abordagem, de acordo com Bryman (2005), a pesquisa classifica-se como quantitativa. Segundo esse autor a pesquisa qualitativa enfatiza a captação das perspectivas e interpretações dos indivíduos que estão sendo estudados, já a pesquisa quantitativa parte de uma teoria onde são geradas hipóteses explicadas por dados mensuráveis. Como detalhado, a presente pesquisa aborda o uso de modelos analíticos e de simulação, com base em modelos estatísticos do SPM, de maneira que se entende a pesquisa predominantemente quantitativa.

Do ponto de vista dos objetivos gerais, Gil (1991) propõe três tipos principais de pesquisa: a exploratória, a descritiva e a explicativa. No presente caso, como será proposto um Método, entende-se que a pesquisa é predominantemente descritiva – uma pesquisa descritiva descreve as características de determinado fenômeno, definindo e delimitando-o. Porém ela é também explicativa, pois pretende-se mostrar relações de causalidade entre o projeto do modelo e o seu desempenho.

Já com relação aos métodos de procedimentos de pesquisa, Gil (2002) descreve as seguintes alternativas: pesquisa bibliográfica, pesquisa documental, pesquisa experimental, pesquisa ex-post facto, estudo de corte, levantamento, estudo de campo, estudo de caso, pesquisa-ação e pesquisa participante.

Para Lacerda *et al.* (2013), diferentes procedimentos de pesquisa devem ser utilizados em conjunto e aplicados ao que denomina de Pesquisa em Ciência de Projeto (em inglês Design Science Research – DSR). Essa embasa a criação e validação de sistemas, elementos e métodos que ainda não existem e visam prescrever soluções para uma determinada classe de problemas, sendo particularmente aplicável ao campo da Engenharia de Produção. De maneira geral as etapas da DSR são: a conscientização, que, baseada em pesquisa na literatura, fornece os subsídios necessários à definição de classes de problemas; a proposição de artefatos para a solução de determinada cada classe de problemas; a validação destes artefatos no contexto para o qual foram definidos.

O Quadro 2 resume e situa o trabalho nas classificações discutidas, incluindo os instrumentos de pesquisa utilizados para a sua realização. Entretanto, demonstra-se na próxima seção, que a DSR orientou o uso combinado dos procedimentos de pesquisa utilizados.

Quadro 2 - Indicação das escolhas metodológicas do Trabalho

Método de Pesquisa	**Indutivo** Dedutivo Dialético Hipotético-Dedutivo	**Métodos de Procedimentos de Pesquisa**	**Pesquisa Bibliográfica** Pesquisa Documental **Pesquisa experimental** Pesquisa ex-post facto Estudo de Corte Levantamento Estudo de Campo Estudo de Caso Pesquisa - Ação Pesquisa Participante
Natureza de Pesquisa	Básica **Aplicada**		
Abordagem do Problema	**Quantitativa** Qualitativa		
Objetivo Gerais da Pesquisa	Exploratória **Descritiva** **Explicativa**	**Instrumentos de pesquisa**	**Softwares** **Modelos Simulação** Entrevistas

Fonte: Elaborado pelo Próprio Autor

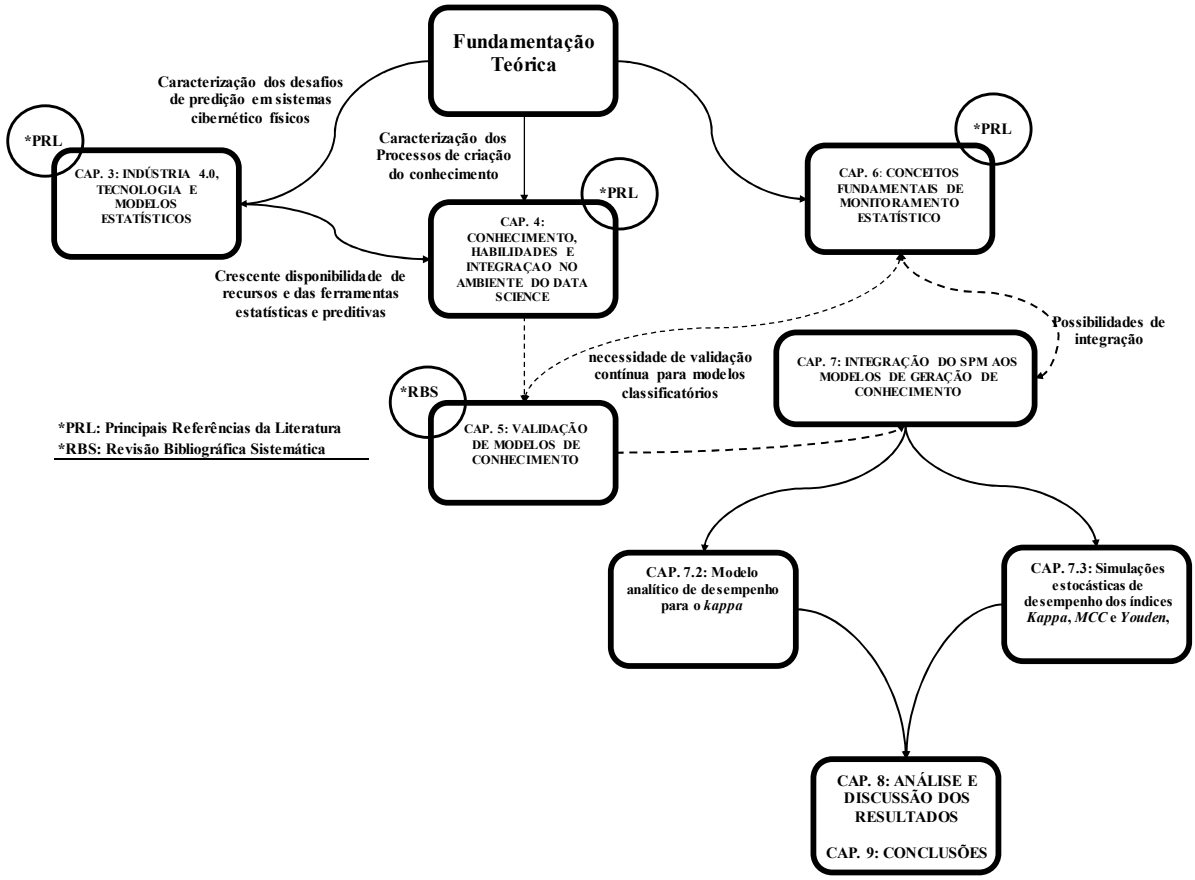
2.2 PROCEDIMENTO DE PESQUISA

A trilha percorrida durante a elaboração do trabalho está demonstrada nas Figuras 1 e 2. Estas Figuras são mapas mentais que descrevem as etapas gerais aplicadas e as questões avaliadas nesse processo (o que fazer e porque fazer). É possível identificar nestes mapas três grandes blocos: Fundamentação teórica, Elaboração da proposta de integração do SPM a DS e Validação da proposta. Com relação à DSR, estes blocos alinham-se respectivamente as etapas de conscientização, elaboração e validação do artefato. Definindo-se o artefato como o método ou abordagem, a ser proposta, para integração entre SPM e DS.

No bloco de fundamentação teórica: *i)* demonstra-se o desenvolvimento e a complexidade dos sistemas cibernéticos – físicos e o crescimento das possibilidades de uso das ferramentas de análise estatística; *ii)* compreende-se os vários conteúdos e escopos de aplicação da DS e sua relação com a criação de modelos preditivos; *iii)* avalia-se as relações da DS com os processos de geração de modelos, identificando-se lacunas pertinentes à manutenção dos modelos e monitoramento da qualidade de seus resultados preditivos.

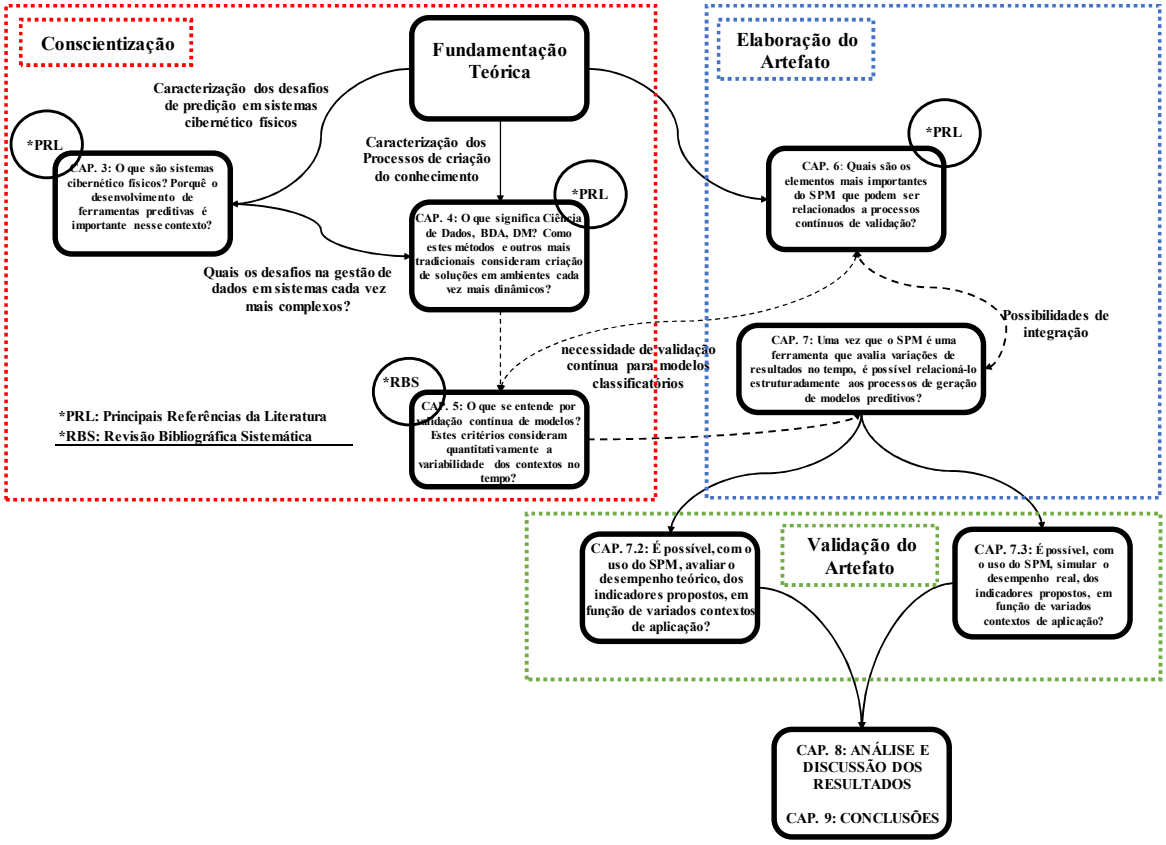
No bloco descrito como proposta de integração do SPM a DS, relacionam-se os fundamentos da teoria do SPM à lacuna identificada, propondo-se aplicar as Fases I e II do SPM no monitoramento de modelos. Para tanto, utiliza-se o *Kappa* (que avalia quantitativamente os acertos e erros de previsão de um modelo) nos gráficos de controle, desenvolvendo-se cálculos analíticos de desempenho baseados no comprimento médio de sequência (em inglês Average Run Length – ARL), bem como simulações computacionais estocásticas do ARL, a partir da geração de dados "reais". Simulações equivalentes são criadas para verificar o desempenho de gráficos de controle baseados nos índices *MCC e Youden.*

Figura 1 - Mapa mental: o que fazer



Fonte: Próprio autor

Figura 2 - Mapa mental: Por que fazer



Fonte: Próprio autor

Mais especificamente, demonstra-se nos capítulos 3 e 4 o contexto em que a pesquisa se insere. Utilizando-se as principais referências da literatura, a opinião dos autores auxilia a definir e delimitar elementos que serão aplicados no presente trabalho, e que por sua vez carecem de consenso sobre a sua caracterização. Dentre eles podem-se citar os sistemas Cibernéticos – Físicos, a DS, o BDA, a DM.

Ainda nos capítulos 3 e 4 demonstra-se que a complexidade da geração e troca de dados nos sistemas cibernéticos físicos, dificulta a evolução das capacidades preditivas destes sistemas, devido os desafios técnicos existentes. Considerando-se a complexidade, a possibilidade de erros e o dinamismo do contexto delimitado, demonstra-se no Capítulo 5 a importância de métodos que garantam a qualidade classificatória de modelos preditivos binários ao longo de todo o seu ciclo de vida.

Assim, com o objetivo de compreender quais as principais práticas de validação desses modelos, e como estas se adequam, ou não, ao uso destes modelos por longos períodos, uma revisão bibliográfica foi aplicada as bases Scopus e Web of Science (WoS) nos dias 30/11/20 e 01/12/20. Aplicou-se a string: TS=((("continuous validation") OR ("temporal validation") OR

(("Data mining")) AND (("Statistical Process Control") OR ("Statistical Process Monitoring")))) sem limitação de intervalo de tempo. Além desta consulta, alertas foram cadastrados nestas bases, monitorando qualquer publicação relevante inserida após a pesquisa.

Resume-se no capítulo 6 os principais fundamentos do SPM, e breve discussão de possíveis aplicações no contexto da I 4.0. Utiliza-se o SPM pois é um método de monitoramento de parâmetros ao longo da linha do tempo, e essa característica preenche lacunas identificadas para validação de modelos ao longo do seu ciclo de vida - ou validação contínua. Deste modo, a revisão dos fundamentos das Fases I e II do SPM embasa a definição de pontos aderentes entre os métodos de geração de modelos e o SPM, possibilitando sua integração com a DS em um sentido mais amplo.

Comparando-se os constructos desenvolvidos ao longo dos capítulos 3 a 5 com os elementos do SPM presentes no capítulo 6, elabora-se no capítulo 7 uma abordagem integrando-se o SPM aos processos de construção de modelos. Demonstra-se a aplicabilidade do SPM, integrado a DS, como instrumento de validação contínua e como ferramenta de apoio ao processo decisório de elaboração e utilização de modelos.

2.3 PROCEDIMENTO DE ANÁLISE

Nas seções 5.2 e 5.3, os fundamentos necessários para a compreensão de diversos, e relevantes, índices de validação de modelos são apresentados, entre estes, o *Kappa.* Resumidamente, no caso de modelos binários classificatórios, para calcular *Kappa,* é necessária uma amostra de n resultados de previsão e seus respectivos resultados reais da classificação, que, ordenados em uma matriz de contingência fornecem os dados para o cálculo desse índice e de seu desvio padrão.

Uma vez proposta a integração do SPM ao contexto da DS, com a utilização do *Kappa*, é fundamental desenvolver critérios para sua operacionalização. Para tanto, nas seções 7.2 e 7.3, complementa-se a abordagem proposta com a análise das quantidades e tamanho de amostras adequados ao cálculo desse índice, e dos limites de controle de seus gráficos. Estes critérios são desenvolvidos com base na análise de desempenho dos gráficos em questão, avaliados pelo comprimento médio de sequência (em inglês Average Run Length – ARL).

O comprimento de sequência (em inglês Run Length - RL) refere-se à quantidade de lançamentos, em um gráfico de controle, até que um destes lançamentos ocorra fora dos limites

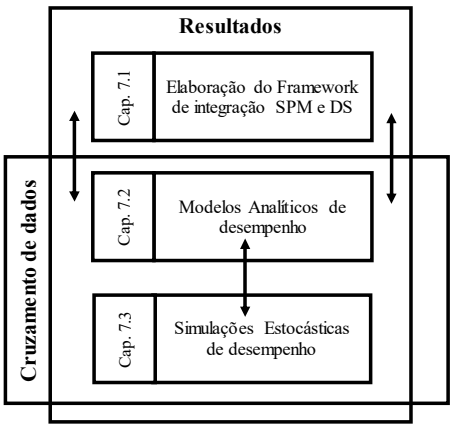
de controle do gráfico (neste caso, o sucesso é a ocorrência do ponto fora dos limites de controle). O ARL é a média de vários destes RL´s.

Na seção 7.2 elabora-se a formulação matemática para o cálculo dos limites de controle de *Kappa,* e consequentemente do ARL teórico, sob duas condições: quando os parâmetros necessários ao cálculo do índice são conhecidos e quando são desconhecidos.

Para a situação de parâmetros conhecidos, uma planilha em Excel® é desenvolvida. Nesta planilha é possível calcular o ARL teórico em função de: n, da probabilidade esperada de concordância - *Pe* (probabilidade de acertos decorrente do acaso, e que é detalhada na seção 5.3.1) e de δ. O valor δ representa o desvio do *Kappa* de seu resultado nominal, assim para $\delta = 0$, temos um processo sob controle, e conforme δ aumenta, *Kappa* desvia-se do seu valor, indicando um processo fora de controle. Espera-se que ARL seja o maior possível na situação de $\delta = 0$, e o menor possível na situação de $\delta > 0$.

Para o caso de parâmetros desconhecidos, um programa em Maple® é codificado. Este considera n, δ, *Pe* e a quantidade m de *Kappas* utilizados no cálculo dos limites de controle. No caso de parâmetros desconhecidos, a cada amostra retirada pode ser elaborado um gráfico de controle diferente. Assim, o programa calcula o ARL com base em todos os resultados possíveis para *Kappa* (e seu desvio padrão), dado um determinado m. Demonstra-se que para m = 1000, os valores de ARL, com parâmetros desconhecidos, aproximam-se dos valores com parâmetros conhecidos. Essas ferramentas possibilitam o cálculo do ARL para diferentes cenários de aplicação (n, m, *Pe* e δ), que são detalhados e desenvolvidos na seção 7.2.

Figura 3 - Procedimento de Análise

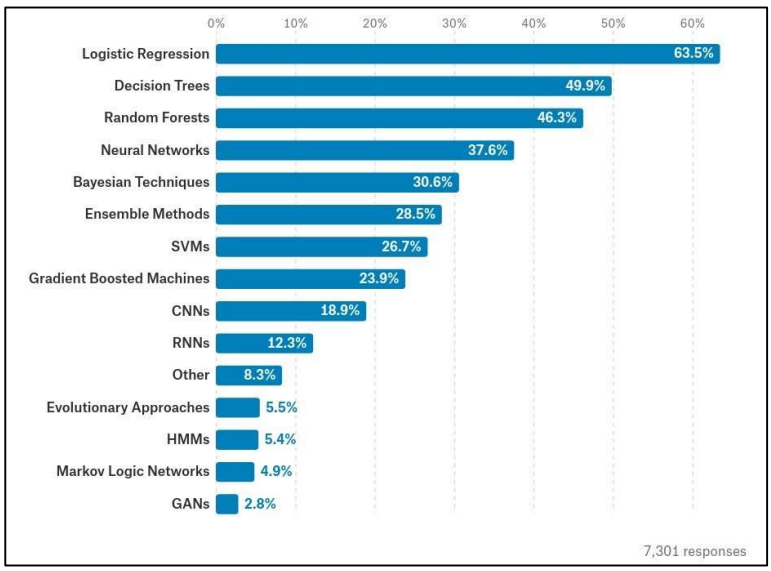


Fonte: Elaborado pelo Próprio Autor

Para complementar o procedimento de análise, conforme ilustrado na Figura 3, são desenvolvidas simulações estocásticas.

As técnicas de simulação utilizadas, detalhadamente explicadas na seção 7.3, visam avaliar, na prática, o desempenho teórico calculado na seção 7.2. Para a simulação, criou-se um programa que gera dezenas de milhares de previsões, com suas respectivas classificações "reais". Esse programa expande um banco de dados rotulado (ou seja, em que se conhece o resultado real da classificação), bastante simples que possui duas variáveis, uma explicativa e uma dependente. A variável explicativa é a quantidade de meses que um funcionário possui de experiência e a variável dependente o sucesso ou falha na execução de uma tarefa. Com base no banco de dados inicial, elabora-se um modelo de previsão, utilizando-se regressão logística. Esta foi escolhida devido à frequência de utilização da ferramenta em trabalhos acadêmicos, fato demonstrado na seção 5.1, e, também da sua utilização no meio não acadêmico, demonstrada pelo levantamento feito no site kaggle.com, ilustrado na Figura 4.

Figura 4 - Aplicações da Regressão Logística no site Kaggle.com



Fonte: Oliveira (2018)

O resultado da previsão (banco de dados expandido) é comparado com o resultado conhecido, o que gera informações sobre os acertos deste modelo, possibilitando a geração da matriz de contingência e o cálculo do *Kappa*. Na simulação, uma parte dos dados é utilizado para a criação de cartas de controle do referidos índice, com base em m amostras de tamanho n. A outra parte dos dados é utilizada para calcular valores que são comparados com os limites das cartas de controle previamente geradas. Como este procedimento é executado muitas vezes, calcula-se o ARL para várias condições de amostragem.

Nas simulações, um elemento denominado $\boldsymbol{\eta}$ foi inserido para criar dados expandidos onde a relação entre meses de experiência, e o sucesso na execução da tarefa, é diferente

daquelas utilizadas nos cálculos dos limites de controle. Como $\boldsymbol{\eta}$ pode variar, é possível simular perturbações de contexto maiores ou menores, e com isso avaliar o ARL também em função dessas mudanças de contexto. Neste caso $\boldsymbol{\eta}$ e δ são equivalentes, pois refletem situações fora de controle, mas não exatamente iguais, aspecto que é explicado na seção 7.3, juntamente com os detalhes dos cenários analisados.

Os resultados da simulação, analisados e comparados aos resultados do método analítico proposto, possibilitam definir os tamanhos de m e n a serem utilizados na abordagem elaborada, resultando em um método prescritivo (artefato) para o monitoramento de modelos classificatórios binários (classe de problema).

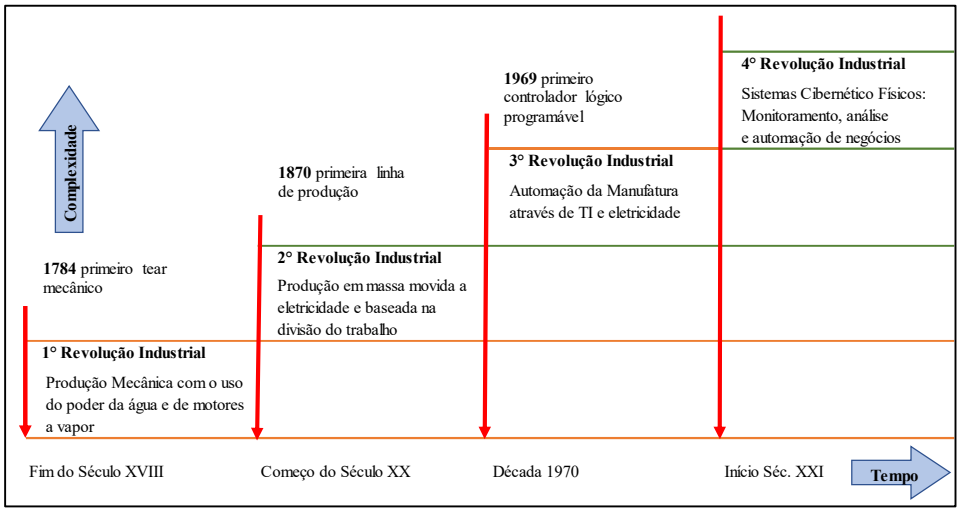
Para complementar a validação do artefato, o programa criado para a simulação estocástica dos valores de *Kappa* foi adaptado para os índices *MCC* e *Youden*. Possibilitando: *i)* avaliar a quantidade de registros (n) e quantidade de amostras (m) mais adequados na análise de cada um dos índices; *ii)* avaliar as diferenças de sensibilidade entre os diferentes índices; *iii)* identificar se existem diferenças significativas no uso de um ou outro índice.

3 INDÚSTRIA 4.0, TECNOLOGIA E MODELOS ESTATÍSTICOS

O desenvolvimento da indústria, como demonstrado na Figura 5, sempre esteve ligado a avanços nas tecnologias de geração de energia e geração, transmissão e processamento de dados. Amparados por estes avanços, outros campos da ciência têm se desenvolvido e auxiliado a melhoria de processos e produtos, um destes campos é o da probabilidade estatística. Durante o início da segunda revolução industrial Shewhart (1931) desenvolve as bases do controle estatístico de processo (em inglês Statistical Process Control – SPC), entretanto ainda eram necessárias ferramentas de cálculo adequadas para maior difusão do uso das técnicas estatísticas. Essa questão foi solucionada com o advento dos computadores após a Segunda Guerra Mundial, o que possibilitou a intensificação da aplicação da estatística na indústria (SALSBURG, 2009). Nas décadas de 1960/70, 30 anos após a apresentação dos gráficos de controle por Shewhart, engenheiros dos Laboratórios Bell discutiam como procurar problemas aleatórios nas grandes quantidades de dados gerados pelo monitoramento das linhas telefônicas, o que demonstra a importância de ferramentas adequadas para o tratamento eficiente das informações.

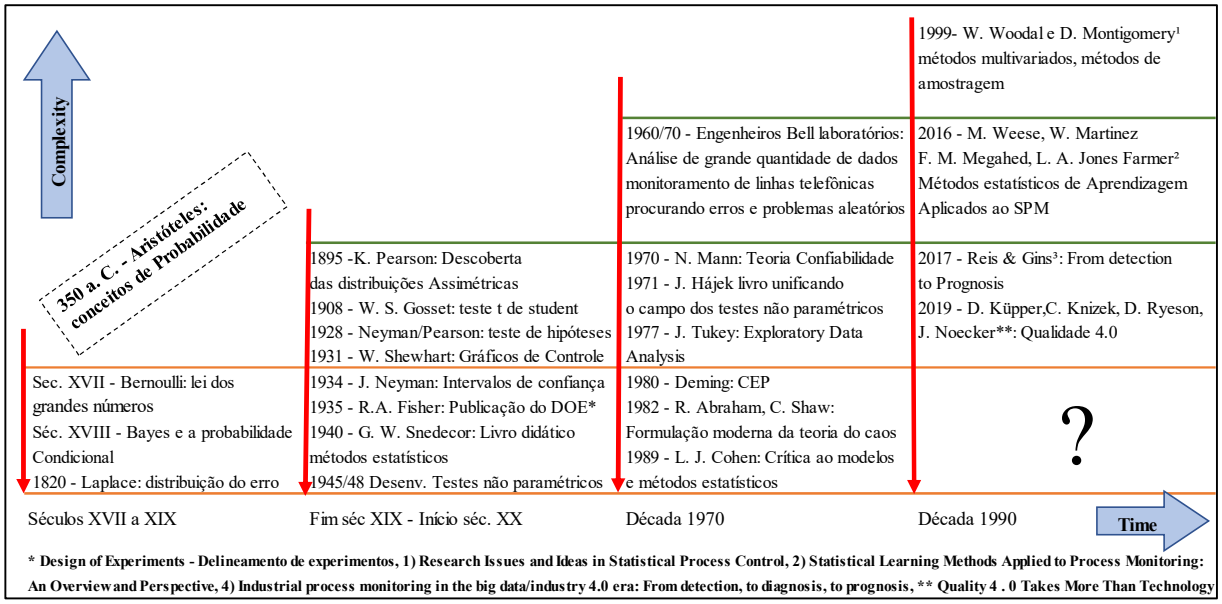
A partir da década de 1980 Deming reforça a importância do uso de métodos estatísticos para melhoria de qualidade e produtividade e, nas décadas seguintes, a intensificação da automação de processos gera dados que permitem a aplicação de técnicas de monitoramento multivariado, monitoramento de perfis e a aplicação de métodos estatísticos de aprendizagem (WOODALL; MONTGOMERY, 1999; WEESE *et al*., 2016; REIS; GINS, 2017). A Figura 6 ilustra alguns destes eventos.

Figura 5 - Etapas de desenvolvimento da Indústria



Fonte: Preuveneers; Ilie-Zudor, 2017

Figura 6 - Desenvolvimento da Probabilidade Estatística



Fonte: adaptado de Salsburg (2009)

Nas seções seguintes apresentam-se os principais elementos referentes a Indústria 4.0, e demonstra-se que, independentemente de haver consenso sobre esse movimento se tratar da quarta revolução industrial, o uso de dados, e das tecnologias de informação e comunicação (TIC´s), é elencado como fundamental para o aumento de competitividade nos mais variados tipos de operações, que não somente as industriais.

Demonstra-se também que o aumento da produtividade deve ser fruto do atingimento de metas que visam cada vez mais o uso eficiente dos recursos. As operações deverão se apoiar no uso de informações, dados e técnicas estatísticas que alterem a lógica presente de reação a problemas, para prevenção de problemas. Assim, discute-se que o acesso a uma quantidade de dados cada vez maior, denominado Big Data e a busca pelo desenvolvimento de sistemas cibernéticos físicos, serão efetivos desde que possibilitem a antecipação de problemas, ou seja, os ambientes evoluam de reativos para preditivos. O grande desafio é criar estas condições em operações reais, sujeitas a dificuldades técnicas e restrições financeiras não consideradas por modelos acadêmicos.

3.1 INDÚSTRIA 4.0: UM NOVO MODELO INDUSTRIAL?

O texto a seguir trata do termo Indústria 4.0, o que impulsiona seu desenvolvimento, suas tecnologias componentes, demais elementos que a caracterizam e aspectos relacionados a metas de desempenho dentro deste contexto. Estes elementos compõem parte do embasamento

teórico necessário para demonstrar o crescimento da importância do uso de modelos preditivos, da DS e, consequentemente, dos processos de monitoramento contínuo destes modelos.

O termo “Indústria 4.0” foi apresentado pela primeira vez em Hannover, no ano de 2011, durante uma reunião de negócios entre profissionais da indústria e elementos do governo alemão, neste evento discutia-se a relevância do uso de tecnologia como um fator de competitividade (KAGERMANN; WAHLSTER; HELBIG, 2013; SANDERS; ELANGESWARAN; WULFSBERG, 2016).

Apesar de alguns autores alegarem que o termo Indústria 4.0 é uma nova roupagem para tecnologias já existentes e que se desenvolvem continuamente (APREDA *et al*., 2016), governos, ao redor do mundo, criaram programas para discutir o assunto, propondo linhas de estudo e ação, com o objetivo de aumentar a competitividade de suas indústrias, alguns destes programas são elencados no Quadro 3.

Quadro 3 - Iniciativas Governamentais de Suporte ao Desenvolvimento da Indústria 4.0

Ano	País	Iniciativa	Origem	Objetivos	Referência Bibliográfica
2011	Estados Unidos	Advanced Manufacturing Partnership (AMP)	Governo	Garantir que EUA estejam preparados pra liderar a nova geração da Manufatura	(REIF; SHIRLEY; LIVERIS, 2014)
2012	Alemanha	High - Tech Strategy 2020	Governo	Desenvolvimento de tecnologias de ponta	(KAGERMANN; WAHLSTER; HELBIG, 2013)
2013	França	La Nouvelle France Industrielle	Governo	Priorização de Políticas Industriais	(CONSEIL NATIONAL DE L´INDUSTRIE, 2013)
2013	Reino Unido	Future of Manufacturing	Governo	Plano de longo prazo (2050) com políticas para suportar o crescimento da manufatura nas próximas décadas	(FORESIGHT, 2013)
2014	União Europeia	Factories of the Future (FoF)	União Europeia	Competitividade Manufatura Europeia	(EUROPEAN COMMISSION, 2016)
2014	Coréia do Sul	Innovation in Manufacturing 3.0	Governo	Ações estruturadas em 4 estratégias e 13 atribuições objetivando saltos de produtividade na Indústria Coreana	(KANG *et al*., 2016)
2015	China	Made in China 2025	Governo	Acelerar a informatização e industrialização	(LI, 2015)
2015	Japão	5° Plano Básico de Ciência e Tecnologia	Governo	Inclusão da Manufatura no Plano "Sociedade Super Inteligente"	(CABINET OFFICE, 2015)
2016	Singapura	Plano RIE 2020 (Pesquisa, Inovação e Empreendimentos)	Governo	Fortalecimento de Indústrias Chave	(NATIONAL RESEARCH FOUNDATION, 2016)
2017	Brasil	Agenda Brasileira para a Indústria 4.0	Governo	Proposta de Agenda Nacional para o Tema	(MINISTÉRIO DA INDÚSTRIA E COMÉRCIO E SERVIÇOS, 2019)

Fonte: Elaborado pelo Autor

Como Indústria 4.0 trata-se de tema relativamente recente, não há padronização ou consenso a respeito de todos seus elementos. Termos correlatos, demonstrados no Quadro 4,

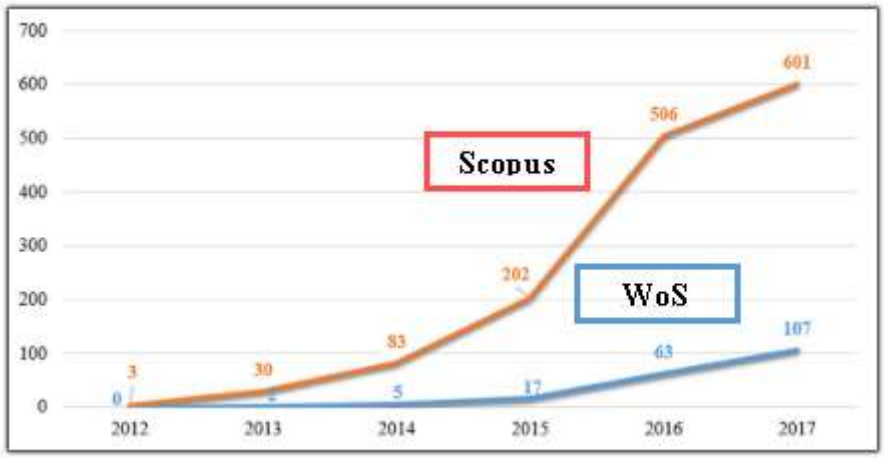
são frequentemente utilizados e as publicações sobre o tema tem crescido desde 2012. A Figura 7, elaborada a partir das plataformas Scopus e Web of Science (WoS), ilustra esse crescimento.

Quadro 4 - Termos correlatos utilizados para denominar a Indústria 4.0

Autor \ Termos sinônimos	Indústria 4.0	Manufatura inteligente	Produção Inteligente	IoT	IIoT	Sistemas Cibernético Físicos (CPS)	M-CPS	Sistemas Cibernético Físicos de Produção (CPPS)
Ahuett, G. e Kurfess, T. (2018)	√					√		
Babiceanu e Seeker (2016)		√	√				√	√
Dagnino (2019)					√			
Fujishima et al.(2016)				√				
Ji et al. (2018)	√	√						
Kamble et al. (2018)	√	√	√	√				
Lasi et al. (2014)	√	√				√		√
Liao et al. (2016)	√			√		√		
Muhuri et al. (2019)	√							
Pilloni, V. (2018)	√					√		√
Wang et al. (2016)	√	√				√		

Fonte: Elaborado pelo Autor

Figura 7 - Evolução das Publicações sobre Indústria 4.0



Fonte: Muhuri, Shukla e Abraham (2019)

Tabela 1 - Categorias de pesquisa sobre Indústria 4.0

Categorias de Pesquisa	Número de Publicações
Conceitos e Perspectivas da Indústria 4.0	18
Indústria 4.0 baseada em sistemas Cibernético Físicos	12
Interoperabilidade da Indústria 4.0	11
Tecnologias Chave da Indústria 4.0	20
Aplicações da Indústria 4.0	27
Total	88

Fonte: Lu (2017)

A respeito da Indústria 4.0, LU (2017) identificou 5 categorias de pesquisa, demonstradas na Tabela 1. Com relação a essa taxonomia, compreende-se que esta tese se enquadra na categoria intitulada: Indústria 4.0 baseada em sistemas Cibernético Físicos. Maiores detalhes podem ser encontrados nos trabalhos bibliométricos realizados por Liao *et al*. (2017) e Muhuri, Shukla e Abraham (2019), aplicados as bases WoS, Scopus e Science Direct.

Dada a importância do tema, as seções a seguir abordam os fatores que estimulam a aplicação das tecnologias pertencentes ao que se denomina Indústria 4.0, quais são essas tecnologias e como podem apoiar a melhoria dos processos operacionais. Pretende-se demonstrar que a melhoria nos resultados de desempenho destes estruturas industriais, demanda a aplicação intensiva das TIC´s em conjunto ao uso de modelos preditivos.

3.2 DEMANDAS COMPETITIVAS E OS OBJETIVOS DA INDÚSTRIA 4.0

Para Li *et al*. (2017), Pilloni (2018) e Wang *et al*.(2016) o desenvolvimento da indústria 4.0 é o resultado da crescente demanda por customização em massa, quando produtos atendem personalizações dos clientes, independentemente de suas escalas de produção (FOGLIATTO; DA SILVEIRA; BORENSTEIN, 2012). Lasi *et al*. (2019) acrescenta como fatores impulsionadores a necessidade de: *i)* prazos cada vez menores para o desenvolvimento de produtos, *ii)* maior flexibilidade produtiva, *iii)* descentralização do processo decisório e *iv)* uso cada vez mais eficiente dos recursos.

Entre autores que afirmam a Indústria 4.0 tratar-se da 4° Revolução Industrial, Pilloni (2018) considera existir 3 elementos de inovação habilitadores à tal transformação: acesso total a Internet, comunicação máquina a máquina e análise avançada de dados. Wang *et al*. (2016) afirma que o fator determinante é a aplicação de tecnologias de Internet das Coisas e Serviços (IoTS). Para Preuveneers e Ilie - Zudor (2017) a 4° Revolução Industrial será alavancada por Sistemas Cibernético Físicos (em inglês Cyber Physical Systems - CPS) que irão monitorar, analisar e automatizar processos de negócios. Considerando um contexto mais amplo, Muhuri, Shukla e Abraham (2019) afirmam que a 4° Revolução não é somente sobre a Indústria, e sim, sobre uma transformação geral, onde, através da integração digital e engenharia inteligente, as máquinas se comunicarão entre si, colaborando e definindo meios de melhor executar uma tarefa. Para Schwab (2016) o momento é de disrupção em relação à indústria 3.0, pois os seguintes fatores impulsionam as transformações: *i)* a *velocidade*, disrupções evoluem em taxas exponenciais, resultado de um mundo interconectado, onde technologia gera mais technologia, *ii) amplitude e profundidade* das mudanças, uma vez que a mudança de paradigmas atinge a

economia, os negócios, a sociedade e os indivíduos e *iii) impacto aos sistemas* que serão transformados por inteiro, atravessando barreiras entre países, companhias, indústrias e a sociedade como um todo.

Baseado nestas afirmações, é possível afirmar que a resposta para a elevação dos níveis de exigência dos mercados consumidores está na integração constante entre os recursos físicos da empresa e os recursos de comunicação e dados. E que o atendimento das demandas fomenta maiores níveis exigência, que por sua vez fomentam o uso de novas tecnologias. No Quadro 5 elencam-se objetivos de desempenho retratados como sendo pertinentes ao contexto da Indústria 4.0, essa relação possibilita traçar-se um paralelo com aqueles tradicionalmente discutidos na literatura de estratégia de manufatura: qualidade, velocidade, confiabilidade, flexibilidade e custos (NEELY; GREGORY; PLATTS, 2005; MOEUF *et al*., 2017). Destacam-se a sustentabilidade, uso eficiente de recursos e melhorias do desempenho organizacional como os objetivos mais citados entre os autores retratados, mas estes podem ser classificados como objetivos de custos. Da mesma forma, os objetivos referentes a menores tamanhos de lote, integração de clientes e parceiros na cadeia de valor, capacidade de inovação e disponibilidade de produtos podem ser enquadrados nos objetivos velocidade/confiabilidade/flexibilidade. Já Saúde, segurança e qualidade no local de trabalho, estariam englobados no objetivo Qualidade.

Quadro 5 - Dimensão: Objetivos de Desempenho Indústria 4.0

Autor	Sust. e uso eficiente de Recursos	Menores tamanhos de lote	Resiliência	Melhoria desempenho Org.	Segurança e saúde do local de trabalho	Qualidade do ambiente de trabalho	Integração cliente e parceiros na cadeia de valor	Capacidade de Inovação	Disp. de produtos
Kamble et al. (2018)	√			√					
Pilloni, V. (2018)						√			√
Wang et al. (2016)	√	√		√					√
Muhuri et al. (2019)	√								
Lasi et al. (2014)	√			√	√	√			√
Ji et al. (2018)	√			√					
Ahuett, G. e Kurfess, T. (2018)	√				√				
Babiceanu e Seeker (2016)			√						
Waibel et al. (2017)	√						√		
Riminucci, M (2018									
Hecklau et al. (2016)								√	
Joppen et al. (2019)									
Samir et al. (2018)	√								
Shamsuzzoha et al (2017)									
Ante et al. (2018)	√								
Kang et al. (2016)				√					

Fonte: Elaborado pelo Autor

Dentre os autores citados, somente Babiceanu e Seker (2016) afirmam que a resiliência é um objetivo de desempenho, e a definem como a capacidade das operações de manufatura lidarem com eventos complexos, respondendo em tempo aceitável. Nos dicionários, resiliência significa retornar a forma inicial após um choque. Baseado nestas definições não se identifica paralelos adequados com nenhum outro objetivo de desempenho tradicional, indicando que este refere-se diretamente a Indústria 4.0.

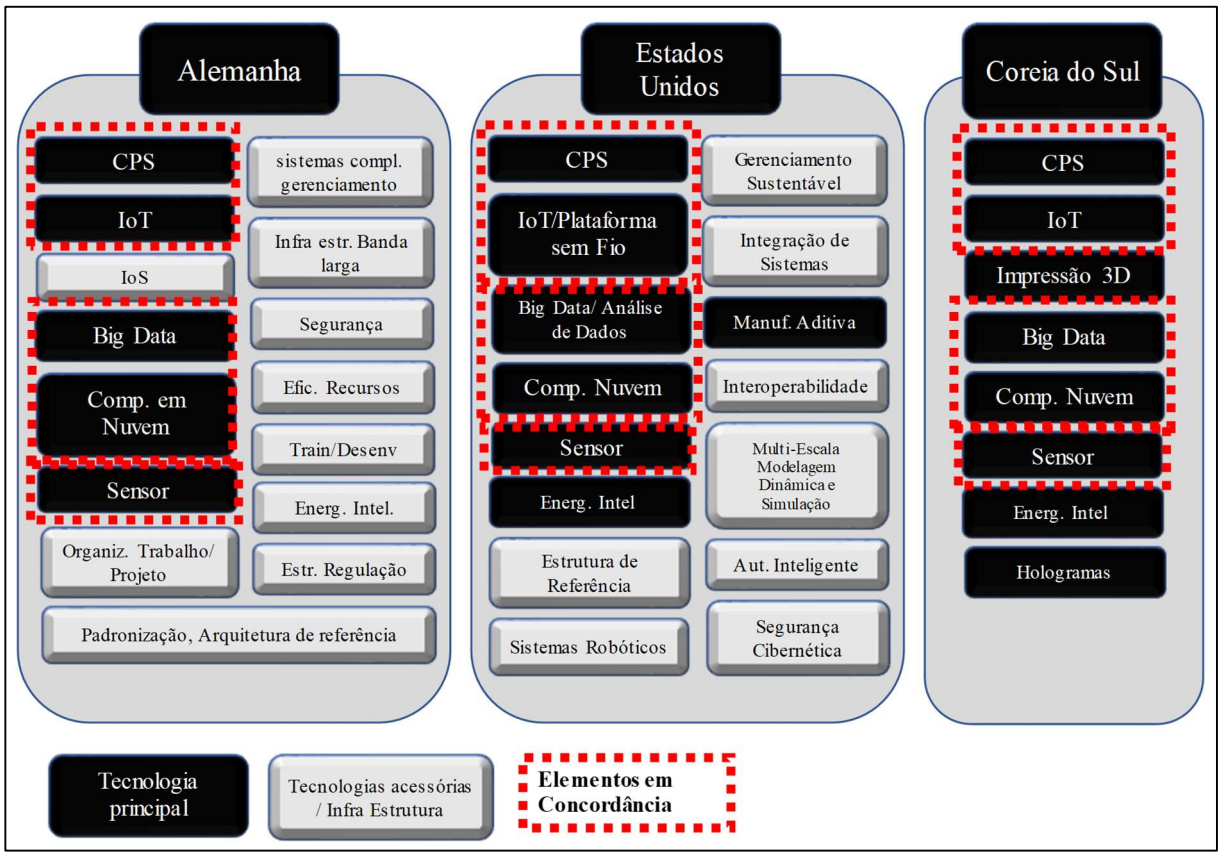
Independentemente das classificações, fica claro que resiliência, o uso eficiente de recursos e o aumento na velocidade de resposta às mudanças são consenso entre os autores como objetivos de desempenho relevantes. Usar os dados para conhecimento prévio de uma situação, possibilita a alocação mais racional de recursos, o bloqueio ou a mitigação de problemas antes destes virem a ocorrer, ou seja, melhoria de desempenho nestes importantes objetivos (ZHONG et al., 2017; SANTOS et al., 2018; JOPPEN et al., 2019).

Neste cenário de desenvolvimento tecnológico acelerado, pode-se elaborar a seguinte proposição: as organizações encontrarão grandes oportunidades de melhorias em seus resultados de desempenho através do uso de dados, especialmente aplicados com o objetivo do aumento da capacidade preditiva de suas operações, uma vez que o conhecimento propicia melhor alocação dos recursos e redução de riscos em geral. As seções seguintes auxiliarão na fundamentação dessa proposição, ao mesmo tempo em que delineiam os desafios técnicos que devem ser considerados nos processos de transformação de ambientes reativos para preditivos.

3.3 ELEMENTOS DA INDÚSTRIA 4.0

Como visto, o termo Indústria 4.0 é abrangente, elencar seus elementos constituintes é um desafio, todavia, pode-se relacionar aqueles que são consenso em trabalhos relevantes da área. Kang *et al*. (2016) analisaram programas de desenvolvimento da Indústria 4.0 dos governos Norte Americano, Alemão e Coreano. A partir dessa análise elencaram as tecnologias principais e as tecnologias acessórias, ou infraestrutura necessária, sob o escopo de cada programa. A Figura 8, adaptada deste trabalho, demonstra que os elementos CPS, Iot, Big Data, Computação em nuvem, e sensores são consideradas tecnologias principais em todos os programas analisados por estes autores.

Figura 8 - Classificação e Seleção de Tecnologias para a Indústria 4.0



Fonte: adaptado de Kang *et al.* (2016)

Além dos referidos programas Norte Americano e Alemão, Tao e Zhang (2017) analisaram o programa Chinês, afirmando que, todos possuem o objetivo comum de atuar na integração dos recursos de manufatura, através da aplicação das tecnologias de comunicação e informação, para desenvolver o que denominam de Manufatura Inteligente. Esta deve satisfazer demandas de socialização, personalização, servitização, inteligência e sustentabilidade ambiental.

Wang *et al.* (2016a) compararam as características técnicas de linhas de produção baseadas no paradigma da indústria 3.0 e as funcionalidades e estruturas físicas que devem estar presentes no que denominam fábricas inteligentes, capazes de se adequar ao presente ambiente competitivo. Demonstra-se no Quadro 6 que existe aderência das características definidas por estes autores com os elementos apresentados na Figura 8.

Quadro 6 - Características técnicas de uma Fábrica Inteligente comparadas com a Fábrica Tradicional

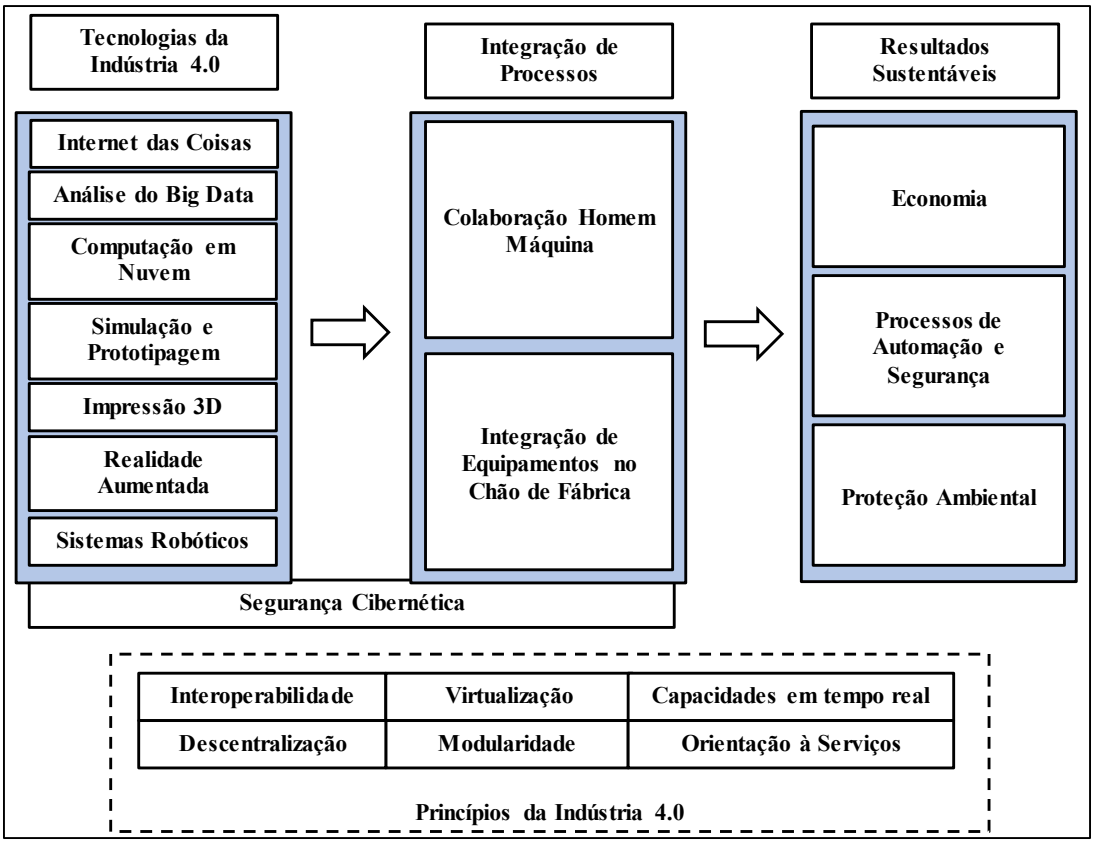
Sistema de Produção da Fábrica Inteligente	Linha de Produção Tradicional
Recursos Diversos. Para produzir vários tipos de produtos em pequenos lotes, mais recursos de diferentes tipos devem ser capazes de coexistir no sistema	**Recursos limitados e predeterminados.** Para construir uma linha fixa para produção em massa de um tipo de produto especial, os recursos necessários são cuidadosamente calculados, adaptados e configurados para minimizar a redundância de recursos.
Roteamento Dinâmico. Ao alternar entre diferentes tipos de produtos, os recursos necessários e a rota para vincular esses recursos devem ser reconfigurados automaticamente e on-line. CPS	**Roteamento fixo**. A linha de produção é fixa, a menos que reconfigurada manualmente por pessoas com desligamento do sistema.
Conexões Abrangentes. As máquinas, produtos, sistemas de informação e pessoas estão conectadas e interagem entre si através da infraestrutura de rede de alta velocidade. IoT	**Rede de Controle de Chão de Fábrica**. A comunicação das máquinas com um controle central existe, mas a comunicação entre as máquinas não é necessária.
Convergência Profunda. A fábrica inteligente opera em um ambiente em rede onde a Rede sem fio industrial e a nuvem integram todos os artefatos físicos e sistemas de informação para formar a IoT e os serviços. Comp. em nuvem	**Camadas separadas**. Os dispositivos de campo são separados dos sistemas de informação superiores.
Auto-organização. A função de controle é distribuída para várias entidades. Essas entidades inteligentes negociam entre si para se organizarem e lidar com a dinâmica do sistema. CPS	**Controle Independente.** Cada máquina é pré-programada para executar as funções atribuídas. Qualquer defeito em um único dispositivo impactará na linha completa.
Big Data. Os artefatos inteligentes podem produzir enormes quantidades de dados, a rede de banda larga pode transferi-los, e a nuvem pode processar o big data. Big Data	**Informações isoladas.** A máquina pode registrar suas próprias informações de processo. Mas essa informação raramente é usada por outros.

Fonte: adaptado de Wang *et al*. (2016a)

Já Kamble, Gunasekaran e Gawankar (2018) elaboraram o Framework apresentado na Figura 9. Para eles o uso das tecnologias da Indústria 4.0 resultam em maior integração dos processos e consequentemente em resultados sustentáveis. A implantação adequada dessas tecnologias garantirá um ambiente operacional onde os princípios de interoperabilidade, descentralização, virtualização, modularidade, capacidade em tempo real e orientação a serviços, estarão presentes. A interoperabilidade refere-se a capacidade de se executar a mesma função, mesmo que equipamentos sejam trocados por outros de mesma origem de fabricação ou não, e a modularidade refere-se a sistemas que podem ser adaptar pelo aumento ou redução de módulos (QIN; LIU; GROSVENOR, 2016; LU, 2017). A descentralização é definida como

a capacidade das empresas, equipe de operações e até mesmo das máquinas, tomarem decisões independentes. A virtualização significa criar uma cópia virtual do mundo físico, onde simulações e controles podem ser executados, sendo também referido na literatura como CPS. De acordo com Kamble, Gunasekaran e Gawankar (2018) capacidade em tempo real diz respeito ao uso de todas as informações disponíveis para adequação rápida de recursos, enquanto orientação a serviços é resultado da integração de todos os processos e disponibilidade de informações resultando em flexibilidade e melhores níveis de serviço. Estes últimos autores concluem que o desenvolvimento do BDA e IoT auxiliarão o desempenho organizacional no futuro, que será baseado na antecipação das demandas dos clientes.

Figura 9 - Framework Ambiente Sustentável Indústria 4.0



Fonte: Kamble, Gunasekaran e Gawankar (2018)

Conclui-se que a Indústria 4.0, ou de maneira mais abrangente, as "Operações 4.0", desenvolvem-se pela intensificação da geração, transmissão, manipulação, armazenamento e processamento de dados. Essa intensificação é resultado do maior sensoriamento das operações e da aplicação de tecnologias referentes ao Big Data, IoT, CPS e computação em nuvem. Os programas governamentais, ao redor do mundo, e os demais trabalhos apresentados nessa seção, corroboram a proposição apresentada ao fim da seção 3.2. Dessa forma, as seções a seguir visam

caracterizar em maior detalhe essas tecnologias, suas relações com a DS e com a necessidade de aumento da capacidade preditiva das operações.

3.4 IoT e BIG DATA: DESAFIOS PARA UMA NOVA INDÚSTRIA

A evolução das tecnologias de comunicação no ambiente fabril seguiu caminhos diferentes de outras áreas, pois nestes ambientes as redes precisam de: latências reduzidas, que é o intervalo de tempo entre um sinal ser coletado, transmitido e chegar ao destino (por exemplo, as imagens de uma transmissão ao vivo da Europa demoram cerca de 30 segundos para chegarem aos televisores no Brasil); tolerância a falhas sem hardware adicional; suporte para maior segurança e interoperabilidade entre as soluções de diferentes fabricantes. Tudo isso para garantir a qualidade dos dados desde a geração até sua utilização. Szymańska (2018) classifica essa qualidade através dos critérios de acuracidade, sincronia dos dados, consistência e integridade, aspectos resumidos e exemplificados no Quadro 7.

Quadro 7 - Características da Qualidade dos Dados

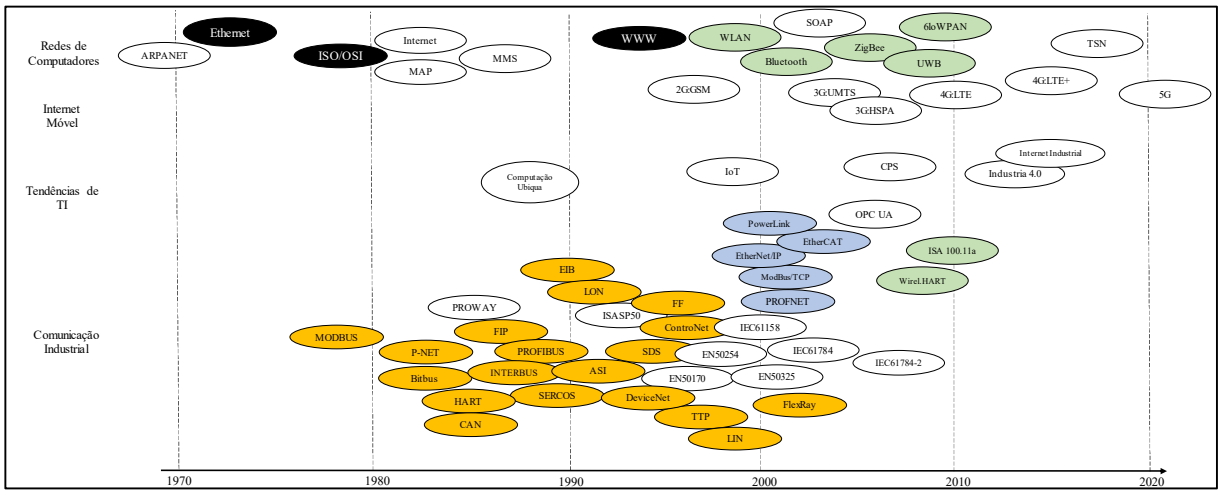
Dimensão	**Características**	**Artefatos de dados correspondentes**
Acuracidade	Quão bem os dados correspondem aos valores "reais"?	Outliers, erros em valores únicos
Sincronia	Os dados estão atualizados	Desalinhamento, erros na integração dos dados
Consistência	Quão bem os dados estão estruturados	Ruído, dados redundantes / colineares, dados ausentes, erros na integração de dados
Integridade	Quão completos são os dados?	Dados faltantes, erros na integração dos dados

Fonte: adaptado de Szymańska (2018)

O desenvolvimento da Indústria 4.0 despertou o interesse dos fabricantes de tecnologias de informação, que, a partir de 2012, buscam alternativas mais viáveis para atender as necessidades específicas do setor. Para demonstrar a complexidade do uso de dados em ambientes industriais, Wollschlaeger, Sauter e Jasperneite (2017) analisaram a evolução das principais tecnologias de comunicação aplicadas na indústria desde a década de 1970 a 2020. Na Figura 10 é possível identificar a evolução das tecnologias, bem como a diversidade dessas tecnologias e suas sobreposições no tempo. Fica claro que diferentes elementos físicos e protocolos de comunicação estão presentes em um mesmo parque industrial ou estrutura operacional (de qualquer outra área). Isto deve-se tanto a presença de equipamentos de diferentes fabricantes (e que podem utilizar tecnologias de informação e comunicação mais

defasadas, ou avançadas), quanto pela presença de equipamentos com diferentes datas de fabricação, o que também pode implicar em diferentes protocolos e meios físicos para comunicação de dados. No Brasil, por exemplo, o tempo de uso médio de máquinas ultrapassa 17 anos (ABIMAQ, 2014). Essa complexidade causada pela diversidade de padrões é o grande desafio para criação de ambientes operacionais com interoperabilidade abrangente a todos os sistemas, e consequentemente, para a garantia da qualidade dos dados, uma vez que aumentam a probabilidade de falhas de geração e transmissão de dados.

Figura 10 - Classificação e Seleção de Tecnologias para a Manufatura Inteligente



Fonte: (WOLLSCHLAEGER; SAUTER; JASPERNEITE, 2017)

Esse desafio cresce à medida que mais componentes e sistemas são integrados por meio dessas tecnologias. A Indústria 3.0 foi um avanço, em relação a Indústria 2.0, pelo surgimento de equipamentos com maior grau de automação e flexibilidade, se comparadas as iniciativas de automação anteriores que demandavam altos investimentos e se restringiam a operações específicas, ou seja, com baixa possibilidade de alteração do previamente projetado. A automação presente na Indústria 4.0 supera a automação da Indústria 3.0, pois baseia - se na troca de informações em tempo real, entre equipamentos, entre equipamentos e sistemas de controle, entre equipamentos e produtos, entre equipamentos e o ser humano, com o objetivo do desenvolvimento de sistemas industriais preditivos e sensíveis ao contexto operacional de curto prazo (PREUVENEERS; ILIE-ZUDOR, 2017).

Bi, Xu e Wang (2014) utilizam o termo Internet das Coisas (IoT) à essa troca de informações, e a aplicação da Tecnologia de Informação (TI), para: medir, identificar, posicionar, rastrear e monitorar objetos, conectando – os em uma rede e criando condições para sua interação. Para Babiceanu e Seker (2016) e Kang *et al.*(2016) IoT é o sistema onde objetos

do mundo físico são conectados à Internet por redes sem fio ou cabeadas. Para Zhong *et al*. (2017), além das características anteriores, a IoT integra os recursos de produção com o objetivo destes se adaptarem a diferentes situações. Já Fujishima *et al*. (2016) e Kamble, Gunasekaran e Gawankar (2018) usam IoT como termo correlato de Indústria 4.0.

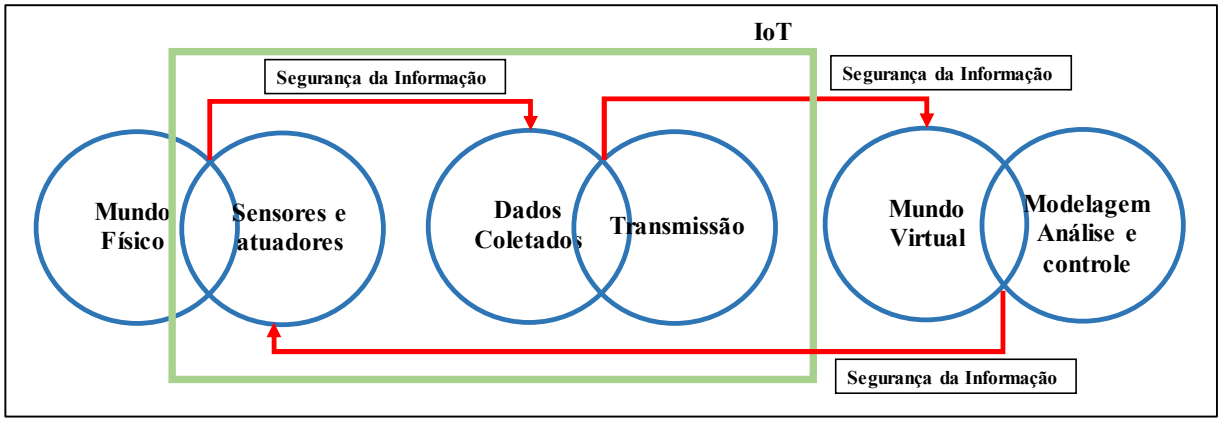
Wan *et al*. (2016), Flath e Stein (2018) e Dagnino (2018) utilizam o termo Internet Industrial das Coisas (IIoT). Para Wan *et al*. (2016) os dados coletados devem ser enviados para centros de dados remotos e seguros (computação em nuvem) e processados em circuito fechado, possibilitando a atualização de parâmetros e tornando os sistemas capazes de detectar falhas e acionar processos de manutenção. Nesse sentido IIoT é aplicado como termo correlato a sistemas Cibernético Físicos, conceito a ser tratado na próxima subseção deste texto.

Além dos desafios de integração das várias TIC´s apresentadas no início desta seção, Bi, Xu e Wang (2014) definem 5 características principais a serem endereçadas na integração completa dos processos físicos e virtuais: uso de RFID´s e redes de sensores sem fio (tornando mais flexível a criação das conexões com o mundo virtual); uso de arquitetura de rede dinâmica de modo a possibilitar reconfigurações para diversos usos dos mesmos recursos; uso do modelo de computação em nuvem (maior possibilidade de escalabilidade e recursos de processamento avançados); Integração do ser humano, representando seu comportamento em um ambiente virtual; e a integração do fluxo de dados aos sistemas de gestão da empresa.

Wan *et al*. (2016) discutem as dificuldades referentes às redes de dados tradicionais, onde a implantação de novos mecanismos de cooperação entre seus elementos demanda a atualização individual dos protocolos de comunicação para cada dispositivo; e as dificuldades no gerenciamento de sistemas e sensores heterogêneos, para que trabalhem adequadamente em conjunto.

Embora existam diferenças sobre a conceituação e a abrangência do termo IoT, neste trabalho ela refere-se ao conjunto de sensores e equipamentos interligados em rede estruturada para a troca de informações em tempo real, possibilitando a conexão entre o mundo físico (das coisas) com o mundo virtual da informação. Nesse cenário, quanto maior a intensidade de captura, transmissão, processamento e análise dos dados em tempo real, e maior a retroalimentação do sistema, maior o grau de integração entre o mundo físico e o mundo virtual. A Figura 11 demonstra esta integração, percebe-se que quanto maiores as intersecções dos conjuntos, maior a integração entre os dois mundos (físico e virtual).

Figura 11 - Integração entre Mundo Físico e Virtual



Fonte: Próprio Autor

He e Wang (2018a) e O'Donovan *et al.*(2015) afirmam que a evolução da IoT, resultará em crescimento acelerado dos dados disponíveis – fenômeno que é denominado de Big Data (BD) - e a manufatura, para melhorar seu desempenho, necessitará gerenciar este crescimento exponencial dos dados, bem como sua capacidade analítica para extrair significado de grandes bancos de dados.

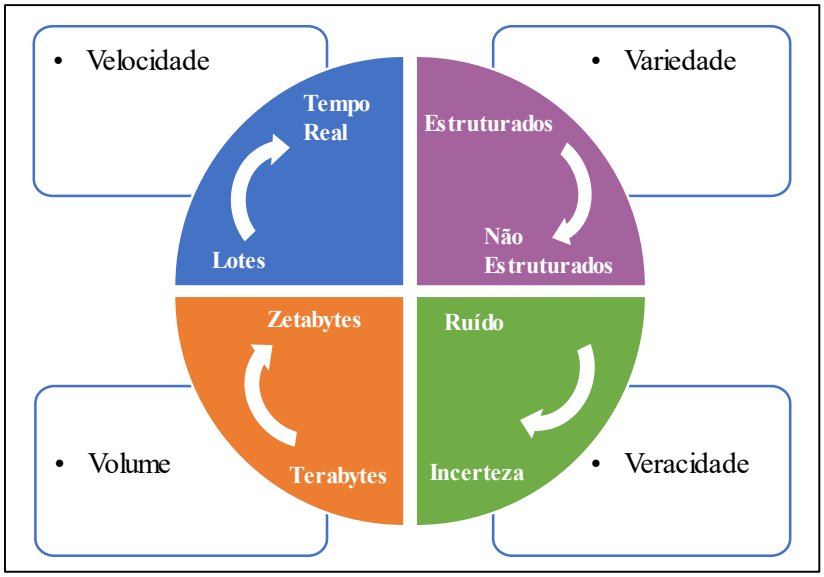
He e Wang (2018a), Brandenburger *et al* (2016) e Mauro *et al.*(2016) diferenciam o Big Data dos bancos de dados tradicionais pelas seguintes dimensões: Volume, Velocidade, Variedade e Veracidade – ou 4V´s. Weese *et al.*(2016) reduzem esta diferenciação às dimensões Volume, Velocidade e Variedade, alegando que veracidade e valor são dimensões importantes para qualquer conjunto de dados e não somente do Big Data.

Cabe assim ressaltar que o Big Data é diferente do conceito de bancos de dados tradicional, pois, do ponto de vista de volume, existe um crescimento exponencial na quantidade de dados coletada, resultado da intensificação na aplicação de sensores e sua integração com redes internas que habilitam o armazenamento destas informações. À dimensão Velocidade inclui-se o aspecto contínuo da captura dos dados, e, também a alta frequência desta captura. Com relação à Variedade, a captura de dados tem evoluído de dimensões escalares para dimensões de ordens mais altas, como imagens, espectro de imagens, informações cromatográficas, informações de equipamentos de ressonância, distribuição de tamanho de partículas e análises de perfis (REIS; GINS, 2017).

Na literatura encontram-se discussões mais detalhadas sobre os desafios resultantes de dados cada vez mais heterogêneos (CHIANG; LU; CASTILLO, 2017; JIRKOVSKÝ; OBITKO; MARÍK, 2016; GANDOMI; HAIDER, 2015). Com relação a dimensão Veracidade, as características pertinentes aos dados englobam confiabilidade estatística dos dados,

rastreabilidade e autenticidade da sua origem, bem como medidas de proteção a acessos não autorizados, o que garante consistência e confiança nesta dimensão, (BABICEANU; SEKER, 2016). A Figura 12 representa os pontos de inflexão entre os bancos de dados tradicionais e o Big Data.

Figura 12 - 4V's do Big Data



Fonte: adaptado de He e Wang (2018)

Babiceanu e Seker (2016) definem outras dimensões do Big Data como valor, visão, volatilidade, verificação, validação e variabilidade. Wan *et al.*(2017) propõe outra forma de classificação de Big Data, utilizando como dimensões dados de produto, dados de dispositivos e dados de comando. Já Mauro *et al*. (2017) classificam o Big Data em 4 temas: Informação, Tecnologia, Método e Impacto.

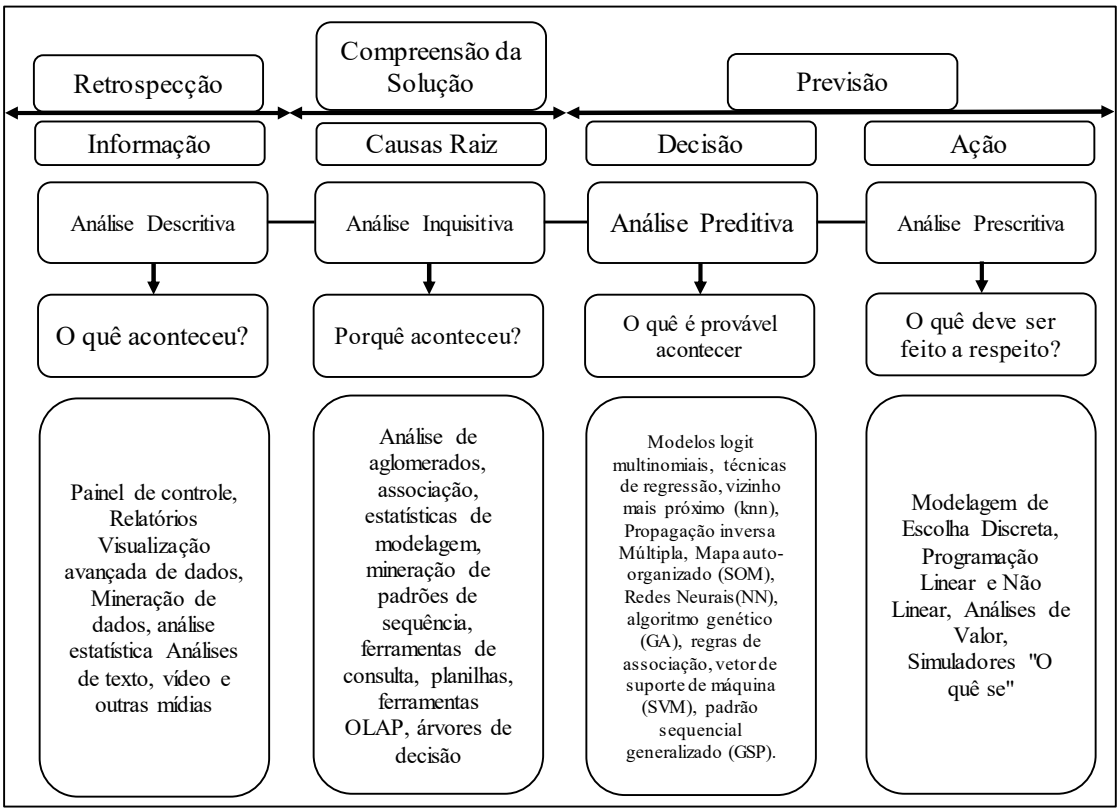
Nesse sentido, He e Wang (2018a), Gandomi e Haider (2015), Kamble, Gunasekaran e Gawankar (2018) e Samir *et al.*(2018) afirmam que análise do Big Data é a ciência que usa de dados para encontrar padrões e relações até então desconhecidas, e que sejam úteis no desenvolvimento de soluções de problemas, ou na tomada de decisões. Chiang, Lu e Castillo (2017) afirmam que o importante é transformar os dados em melhores decisões operacionais sendo o BDA uma jornada para transformar dados em informações relevantes para decisões operacionais , táticas e estratégicas.

Para orientar como transformar dados em melhores decisões, Belhadi *et al*. (2019) classificam os diferentes tipos de problemas de análise em: descritivos, inquisitivos, preditivos e prescritivos. Para eles, a analítica descritiva fornece uma visão retrospectiva sobre o estado

atual de uma situação da empresa. Utilizam-se programas de inteligência de negócios (em inglês business intelligence – BI) com a geração dos inúmeros relatórios disponíveis neste tipo de ferramenta (JOSEPH; JOHNSON, 2013; SIVARAJAH *et al*., 2017). Sendo assim, a análise descritiva revela "o que aconteceu".

A análise inquisitiva avalia “porque algo ocorreu”, sendo alimentada por resultados da análise descritiva ou dados adicionais, quando necessários, a fim de revelar as causas raiz de um problema (BANERJEE; BANDYOPADHYAY; ACHARYA, 2013). A análise preditiva utiliza os dados para a previsão e a modelagem estatística, buscando "o que é provável que aconteça", com base em modelos de aprendizagem supervisionados, não supervisionados, e semi supervisionados. Os modelos prescritivos complementam a análise, pois auxiliam a definição do conjunto de ações que devem ser tomadas para otimizar um determinado processo de negócio. A Figura 13 demonstra a organização destes conceitos.

Figura 13 - Classificação dos problemas e técnicas referentes a análise do Big Data



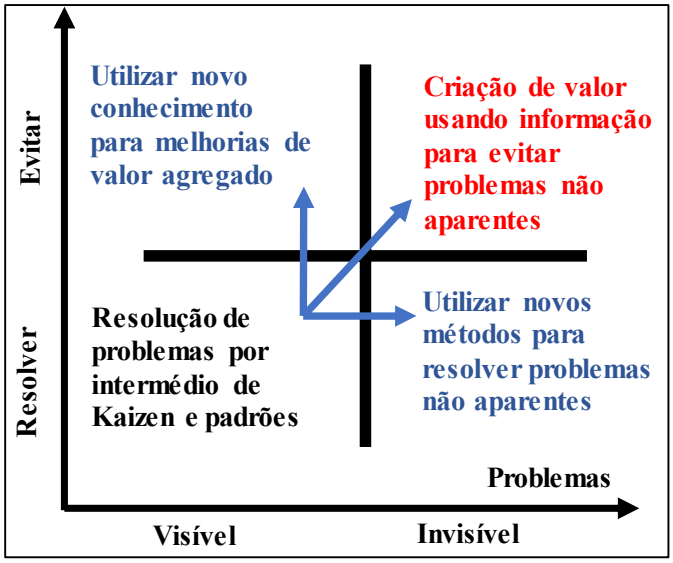
Fonte: Belhadi *et al*. (2019)

Outra abordagem é elaborada por Lee, Bagheri e Jin (2016), Flath e Stein (2018) e Lee *et al*. (2013). Estes autores afirmam que a análise adequada dos dados pode apoiar a identificação e solução de problemas conhecidos e problemas latentes. Com relação a estes 2

tipos de problemas, Lee *et al*. (2013) mapeiam as aplicações de análise de dados na manufatura em dois espaços: Visível e Invisível, a Figura 14 demonstra as oportunidades nestes espaços.

No espaço visível estão questões como defeitos em produtos, atrasos de fabricação e quedas na eficiência global dos equipamentos (em inglês Overall Equipment efficiency – OEE). A análise tradicional dos dados apoia a solução dos problemas visíveis (quadrante inferior esquerdo da Figura 14), a aplicação das usuais ferramentas de qualidade e o avanço na análise dos dados objetiva evitar a reincidência destes problemas. Ainda no espaço visível, no quadrante superior esquerdo, objetiva-se evitar a ocorrência de problemas (e não resolver problemas que já ocorreram) as iniciativas utilizam de técnicas preditivas, um exemplo é o uso de simulação na fundição de ligas metálicas, possibilitando a eliminação de causas de problemas de porosidade já no projeto do processo.

Figura 14 - Espaço de Oportunidades de Produtividade



Fonte: Adaptado de Lee *et al*.(2013)

No espaço invisível encontram-se questões como degradação de equipamentos, desgaste de componentes, aumento na probabilidade de falha de um airbag etc. Segundo Lee *et al*. (2013), os competidores mais eficientes iniciam ações no sentido de abranger estes quadrantes aplicando novas tecnologias ao Big Data. A atuação pode ocorrer dentro da empresa e, também, fora do ambiente de manufatura, identificando e corrigindo problemas não aparentes (quadrante inferior direito), ou evitando a ocorrência destes problemas e assim utilizando a informação para criação de valor para a empresa (quadrante superior direito). Como exemplo de aplicação no quadrante inferior direito pode-se citar os sensores que monitoram condições de lubrificação de um determinado equipamento, garantindo a lubrificação adequada e emitindo

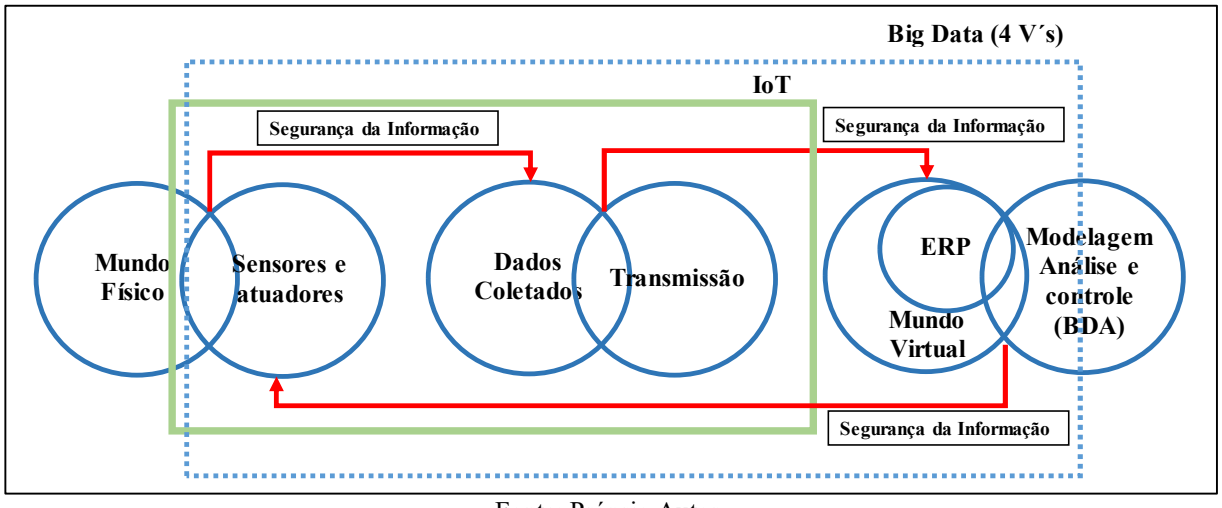
alertas ou impedindo sua operação em caso de problemas nesse sistema – resultando em aumento da vida útil da máquina. Como exemplo de avanço para o quadrante superior direito cita-se a análise de dados de telemetria de equipamentos agrários, onde o fornecedor orienta a melhor condição de operação para o referido equipamento, reduzindo problemas de assistência técnica ao mesmo tempo que melhora o nível de serviço ao cliente, criando valor para o seu produto.

Entretanto o uso do Big Data demanda adaptações em relação ao contexto tradicional. Wang *et al*. (2016) alertam sobre a necessidade de maior rigor no planejamento da coleta de dados. Para estes autores os objetivos de análise devem ser pré-definidos, orientando a busca dos dados e evitando retrabalhos.

Sharma *et al*. (2014) identificaram 4 diferenças entre a modelagem e análise de dados no contexto da Indústria 4.0 e os propósitos gerais da área de computação, que demandam maior trabalho de pesquisa: forte interação entre mundo físico e os controles de retroalimentação dos sistemas; necessidade de conhecimento rigoroso dos processos, falhas precisam ser corrigidas muito rapidamente nos programas; Indústria 4.0 opera de várias maneiras em função da situação, sistemas precisam identificar e atuar compativelmente; os modelos tradicionais de aprendizado de máquina funcionam adequadamente para análises padrão, mas, para aplicações no mundo da Indústria 4.0, precisam levar em consideração relações de custo-benefício e velocidade de análise adequada com realidade.

Com base nos conceitos apresentados, e na Figura 11, elabora-se a Figura 15, pela inserção do Big Data neste contexto.

Figura 15 - Integração entre Mundo Físico, Virtual e o Big Data



Fonte: Próprio Autor

A título de exemplo pode-se citar algumas ferramentas de análise do Big Data existentes no mercado. A General Eletric (GE) anunciou o Predix™, uma plataforma de serviços baseada na nuvem, que conta com ferramentas de análise de dados. Na mesma linha a National Instruments introduziu o Big Analog Data™ Solutions e o Watchdog Agent® para suportar o BDA.

Todas as classificações apresentadas demonstram a complexidade do assunto o que indica a necessidade de novos estudos que padronizem os conceitos e definam limites que caracterizem com maior clareza o ponto de transição de bancos de dados tradicionais para o Big Data. Entretanto, independentemente da padronização destes conceitos, conclui-se que o desenvolvimento da IoT implica na geração de maior quantidade e variedade de dados – Big Data – mas o que gerará retorno do investimento nestes elementos da Indústria 4.0 é o timing de processamento, a análise eficaz e o uso destas informações, de maneira que possibilite o aumento do desempenho dos agentes nestes ambientes.

Conclui-se que a IoT se refere; aos componentes físicos, como sensores e dispositivos que identificam cada elemento integrante de um ambiente operacional (dispositivos de transporte, produtos em processo, operadores, ferramentas e demais agentes): aos protocolos e TIC´s que possibilitam os fluxos de informação entre os elementos do mundo físico e entre o mundo físico e o mundo Virtual. Neste ambiente, os elementos fixos podem ser interligados a uma rede cabeada, mas os elementos móveis necessitam de conexões sem fio e dispositivos de identificação (como RFID´s). As redes e os locais de armazenamento de dados devem ser compatíveis com o volume de informação transportado e armazenado, e flexíveis o suficiente para variações nessa demanda. Essas redes devem disponibilizar os dados, em tempo real, para cada agente envolvido, possibilitando melhores decisões operacionais. Aspectos relativos à segurança de dados devem ser considerados, uma vez que o ambiente de manufatura se torna totalmente dependente da troca constante de informações, e incapacitado de operar caso ocorram falhas no fluxo das informações. Ou seja, é um ambiente tecnológico onde cada um dos elementos citados são os elos de uma longa corrente. Esse sistema é tão forte ou eficaz, quanto o mais fraco dos seus elos (coleta, transmissão, armazenagem, processamento, modelagem, análise dos dados e retroalimentação do sistema). O BDA que está inserido nesse contexto, será abordado em maiores detalhes na seção 4, mas em resumo refere-se ao processo de análise dessa grande massa de dados denominada Big Data, e assim é um elemento relevante na criação dos CPS´s, que são discutidos na seção 3.5.

Com base nas características abordadas sobre: o fluxo de informações, Big Data e o BDA elaboram-se as seguintes proposições: *i)* existe desenvolvimento de tecnologias que possibilitam cada vez maiores fluxos de informações entre os elementos do mundo físico e entre estes e o mundo virtual; *ii)* as tecnologias possibilitam aumento de quantidade, variedade, velocidade e frequência na coleta dos dados, aumentando a complexidade deste fluxo; *iii)* em ambientes dinâmicos, característico de operações em evolução, existem frequentes atualizações e trocas de equipamentos e sistemas, o que contribui para o aumento das dificuldades de interoperabilidade e da complexidade na captura e fluxo das informações, alterando continuamente as características de qualidade dos dados desse sistema (dados errados, dados faltantes, mistura de fontes de dados), assim os recursos para armazenagem e análise desses dados devem ser compatíveis, uma vez que o elo mais fraco de uma corrente determina sua força; i*v)* as informações apresentadas nessa seção corroboram a proposição apresentada ao fim da seção 3.2, uma vez que diversos autores afirmam a importância do uso de análises preditivas, identificando oportunidades para solução de problemas latentes, e para que os objetivos de desempenho sejam alcançados com maior velocidade e eficiência no uso dos recursos; *v)* a aplicação de ferramentas estatísticas preditivas é um dos pilares do BDA; *vi)* com base em *iv)* e *v)* existem elementos suficientes para afirmar que operações de manufatura, ou oriundas de outros ambientes de negócios, aplicarão cada vez mais modelos preditivos para o apoio a decisões, dessa forma, a assertividade desses modelos impactará de maneira mais significativa em suas operações; *vii)* o desenvolvimento e a aplicação de novas tecnologias implica em ambientes mais dinâmicos e complexos, onde avaliar e garantir continuamente a qualidade preditiva dos modelos utilizados será um novo desafio.

3.5 AMBIENTES CIBERNÉTICO – FÍSICOS, PESSOAS E MODELOS PREDITIVOS

O cenário tecnológico apresentado fornece as condições necessárias para o desenvolvimento de sistemas cibernético – físicos (em inglês, cyber physical systems – CPS), estes referem-se a sistemas com capacidades físicas e computacionais integradas entre si e ao ser humano de várias formas (JIRKOVSKÝ; OBITKO; MARÍK, 2016) e são prioridade de pesquisa em programas de vários países ao redor do mundo, conforme demonstrado na Figura 8. Neste contexto encontra-se o uso do termo gêmeo digital, referindo-se a simulações virtuais completas de todos os elementos de um sistema do mundo real (AHUETT-GARZA; KURFESS, 2018; ZHUANG; LIU; XIONG, 2018; CHEN, 2017). Para Ahuett e Kurfess (2018), Chen *et al*. (2015) e Babiceanu e Seker (2016) o CPS utiliza dados, modelagens

computacionais dos sistemas físicos e técnicas estatísticas, em tempo real, para criar múltiplos cenários de decisão, de modo a fornecer melhor visibilidade e controle aos processos, objetivando a melhoria de seu desempenho.

Para Tao e Zhang (2017) um dos desafios específicos da manufatura inteligente é a convergência entre o mundo da manufatura física e o mundo virtual. Esta convergência possibilitará a integração necessária à realização de operações inteligentes nos processos de manufatura, como interconexões inteligentes, interações inteligentes entre equipamentos e gerenciamento e controle inteligentes.

Seshia *et al*.(2016) afirmam que o CPS é um sistema híbrido, formado de elementos físicos e virtuais, onde os modelos e análises devem refletir aspectos discretos e dinâmicos; é um sistema heterogêneo que compreende diversas plataformas e modelos computacionais que necessitam de interoperabilidade garantida (troca de um elemento por outro sem falhas no processo); é um sistema distribuído (conectado via rede, mas com possíveis diferenças de localização); de larga escala (captura e armazenagem de dados cresce continuamente); é um sistema dinâmico (que deve evoluir) e adaptativo, pois se reorganiza na ocorrência de mudanças não planejadas; por fim, que interage com o ser humano devendo possuir interfaces adequadas à esse elemento.

Com base nesses argumentos e nas seções anteriores, conclui-se que o desenvolvimento do CPS demanda grande capacidade de processamento, armazenagem e visualização dos dados em tempo real. Segundo Chaari *et al*.(2016) os CPS´s possuem limitações em relação a essa estrutura, demandando a aplicação de recursos de computação em nuvem (em inglês: cloud computing) para executar papéis de processamento online dos dados dos sensores, suprir com ferramentas de análise do Big Data e apoiar a realização do objetivo da virtualização completa dos elementos físicos. Entre as vantagens estão a escalabilidade do sistema (pode-se aumentar a capacidade conforme a necessidade), confiabilidade e menores custos de processamento e acessibilidade (SHU *et al*. 2015; CHAARI *et al*. 2016). Em contrapartida Lu, Xu e Xu (2014) afirmam que poucas das soluções existentes conseguem se adaptar à mudanças nos ambientes de negócios e que aspectos de acessibilidade, segurança e interoperabilidade devem ser considerados.

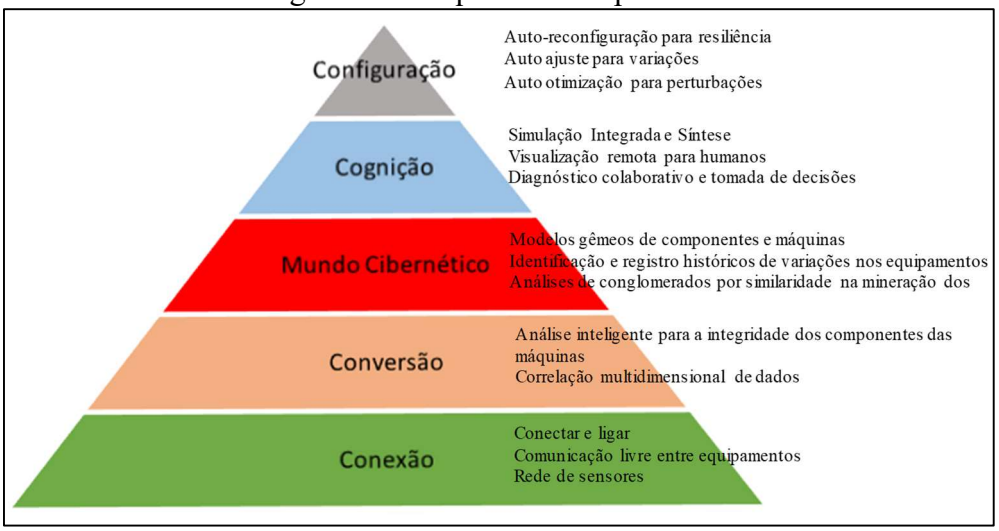
Vários termos são aplicados dentro do contexto do CPS, Babiceanu e Seker (2016) utilizam Manufatura em nuvem como o próprio CPS. Zhong *et al*.(2017) relacionam a manufatura em nuvem como modelos avançados de manufatura sob o suporte de computação

em nuvem. Shu *et al*.(2015) denominam de CCPS o CPS integrado à nuvem (em inglês: Cloud Cyber Physical System). Wu *et al*.(2015) utilizam o termo Manufatura e projeto baseado em Nuvem (em inglês: Cloud-based Design and Manufacturing – CBDM), definindo que neste ambiente o cliente poderia interagir desde a configuração do seu projeto individual até o recebimento de seu pedido. Estes autores afirmam que, embora esse modelo real ainda não exista, é possível pela integração de 4 modelos de serviço da computação em nuvem: Infraestrutura como serviço (IaaS), Plataforma como Serviço (PaaS), Hardware como serviço (HaaS) e Software-como serviço (SaaS).

Ahuett e Kurfess (2018) afirmam que uma fábrica inteira integrada como um CPS completo ainda é uma tarefa desafiadora. A aplicação de modelos preditivos na melhoria de desempenho de um CPS é exemplificado no trabalho de Chen *et al*. (2015), estes autores projetaram e aplicaram os conceitos de CPS a uma máquina CNC de usinagem. Adaptando sensores adicionais aos já existentes no equipamento, monitoraram todas as variáveis de entrada significativas para o referido processo. O estudo dos parâmetros monitorados, e sua relação com os resultados do processo, possibilitou a criação de modelos de otimização e geração de informações preditivas para melhor manutenção e projeto de ferramentas. Estes autores declaram que no futuro os equipamentos CNC serão equipados com dispositivos e sensores suficientes para se caracterizarem como um CPS o que corrobora a avaliação de Sharma *et al*. (2014) a respeito dos dados gerados, e analisados dentro de um CPS, como fonte de oportunidades de melhoria nas áreas de manutenção (proativa) e qualidade.

Segundo Lee, Jin e Bagheri (2017) um sistema de produção deve possuir características preditivas (antecipação a problemas) e serem dotados de mecanismos para encontrar ou auxiliar a busca das causas dos problemas, reconfigurando-se automaticamente diante de eventos de falha. Baseado nessas premissas, elaboraram uma estrutura com foco no gerenciamento de saúde e prognóstico de equipamentos (em inglês, Prognostics and Health Management - PHM), demonstrado na Figura 16. A estrutura baseia-se na integração dos 5C´s: Conexão, Conversão, Mundo Cibernético (virtual), Cognição e Configuração.

Figura 16 - Arquitetura 5C para CPS



Fonte: Lee, Jin e Bagheri (2017)

- Conexão: é a transmissão de dados dos sensores embarcados no equipamento para um servidor central, conforme demonstrado na seção 3.4, os desafios desta integração são os diferentes protocolos de informação existentes em diferentes fabricantes de equipamentos, de diversos anos de fabricação;
- Conversão: os dados brutos captados pelos sensores são convertidos em informação. Dados precisam ser processados para produzirem significado e servirem de base para a criação de modelos, neste caso visando a redução de tempo de parada do equipamento pela previsão de falhas conhecidas;
- Mundo Cibernético: Uma vez que as condições dos ativos são conhecidas, as projeções das condições de produção podem ser elaboradas. Os equivalentes virtuais (gêmeos virtuais) dos equipamentos são projetados a partir de modelagem baseada nos históricos de outros similares, e este conhecimento deve crescer com o tempo, possibilitando previsão de falhas não aparentes em contextos de processo dinâmicos;
- O nível de cognição é o que avalia as projeções de falha do equipamento (baseado no desempenho dos modelos virtuais, alimentados pelas informações de processo) e calcula/sugere o momento ótimo das intervenções visando aumento de produtividade e prolongamento da vida útil do ativo;
- Nível de configuração: sistema ajusta carga de trabalho do equipamento visando prorrogar falhas até intervenção adequada.

Apesar de elaborada para aplicações voltadas a questões pertinentes ao gerenciamento de ativos, a estrutura proposta pelos autores é facilmente extrapolada para outros CPS, e demonstra o papel dos modelos preditivos no desenvolvimento do CPS.

Além dos desafios de integração apresentados, o elemento humano, pouco abordado na maioria dos trabalhos, também deve ser considerado, pois interage, altera, se beneficia e é fator crítico de sucesso nesse novo ambiente (KÜPPER *et a*l., 2019). Seshia *et al*.(2016) e Zhong *et al*.(2017) denominam esta integração como: humanos no ciclo (em inglês human – in the – Loop), destacando que o projeto de ferramentas automáticas deve endereçar aspectos humanos, automatizando tarefas tediosas e permitindo que o ser humano libere sua criatividade ou evite erros.

Abell *et al*. (2017) apresenta um exemplo de interface entre o mundo virtual e o ser humano que utiliza realidade aumentada. Neste exemplo, um sistema de predição de problemas, na produção de células de bateria, monitora e analisa, no mundo virtual, as variáveis de entrada em diversas fases de produção dessas células. Como resultado indica-se para o operador quais delas tem grandes chances de falhas e, portanto, deveriam ser inspecionadas. Para orientar o operador o sistema gera feixes de luz, que incidem diretamente sobre a célula a ser inspecionada. Este é um exemplo de modelo classificatório binário, onde cada célula é categorizada como duvidosa (necessita de inspeção) ou não duvidosa (não necessita de inspeção), aumentando a produtividade desta etapa do processo de qualidade, ao restringir a necessidade de inspeção a uma menor quantidade de elementos. Aqui também é possível identificar a interferência do ser humano em processos apoiados por modelos preditivos, se o operador não atender a indicação luminosa, e "deixar passar" a inspeção, falhas futuras poderão ocorrer no produto, sem que o modelo tenha sido a causa raiz do problema e sim o elemento humano.

A respeito de outras formas de integração, Zhong *et al*.(2017) afirmam que a interface de voz é uma tecnologia adequada à integração do ser humano ao CPS. Longo, Nicoletti e Padovano (2017) validaram o uso de um assistente virtual que, via comando de voz, auxilia atividades de treinamento em tarefas complexas.

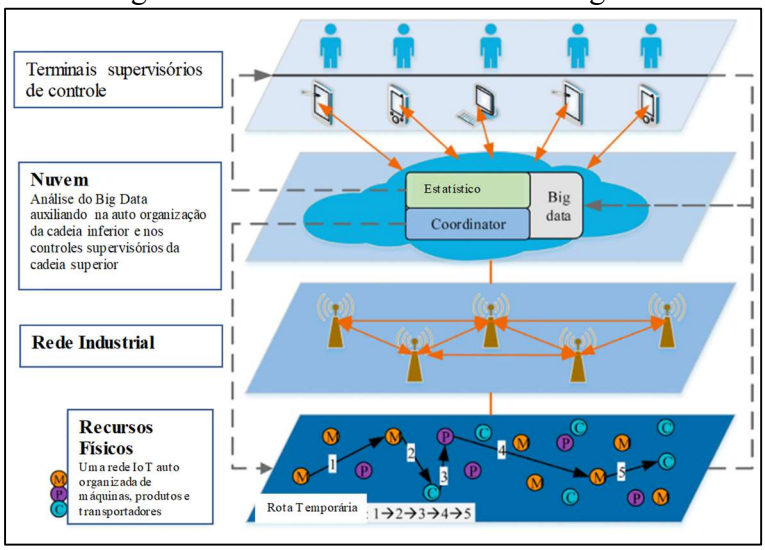
Pfeiffer (2016) realizou estudo em 5 plantas de montagem onde a questão principal era a possibilidade de substituir a mão de obra por agentes autômatos. Concluiu que os sistemas necessitam de pessoas para lidar com eventos e falhas não esperadas e agregar sua experiência aos processos. Para Gorecky *et al*.(2014) o usuário será um supervisor da execução das

estratégias de produção e a última instância no processo decisório, necessitando de ferramentas de integração adequadas.

Wang *et al*. (2016) definem a integração vertical como o fluxo de dados estruturado entre os diversos sistemas de gestão empresarial da empresa e o CPS. Entre esses sistemas estão aqueles dedicados à gestão de informações de clientes (em inglês Clients Requirements management - CRM), Sistemas de Gestão Empresarial (em inglês Enterprise Resources Planning - ERP), Sistemas de Execução de Manufatura (em inglês Manufacturing Execution Systems - MES), etc. Uma vez que estes sistemas também são alimentados por ações humanas e fornecem informações para tomada de decisões, pode-se dizer que a integração vertical é um elemento de integração entre o CPS e o ser humano. Zhong *et al*. (2017) discutem a importância da integração entre os níveis organizacionais, técnicos e gerenciais. O'Donovan *et al*. (2015a) demonstram que as fábricas inteligentes (CPS) são focadas em criar conhecimentos, a partir de dados em tempo real, que apoiam processos de decisão precisos e adequados no tempo, gerando impacto positivo em toda a organização.

Para Wang *et al*. (2016b) os CPS são supervisionados por pessoas capazes de atualizar os algoritmos (BDA) que coordenam a ação de seus agentes (transportadores, produtos, equipamentos), assim esses sistemas devem possuir transparência de dados, integração e facilidade de interface com os operadores (WANG *et al*., 2016a), isso fica claro na estrutura elaborada por estes autores e apresentada na Figura 17. Em estudo subsequente, Li *et al*. (2017) aplicam os conceitos propostos por Wang *et al*. (2016b) a uma linha protótipo avaliando a aplicação destes conceitos.

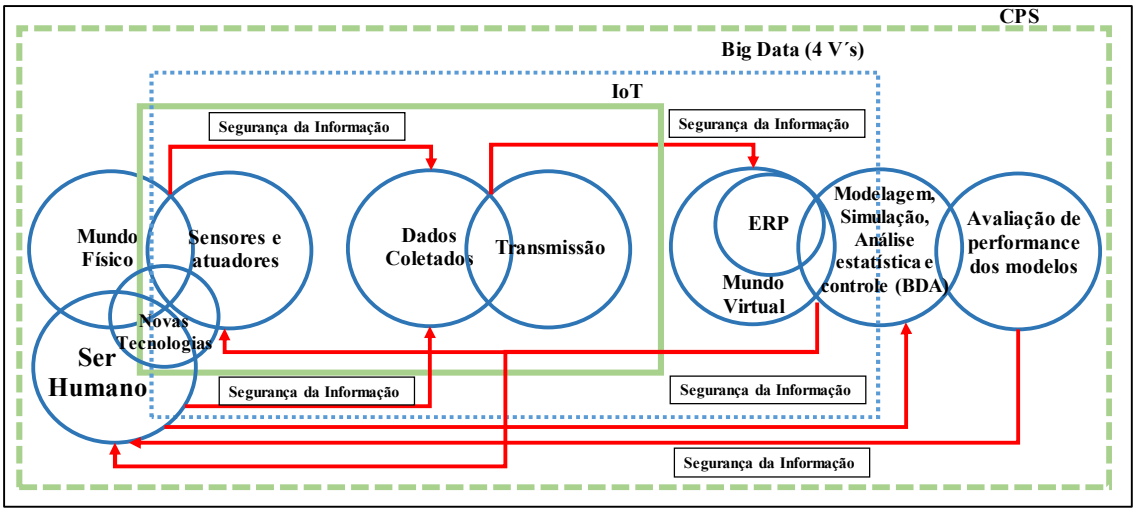
Figura 17 - Estrutura da Fábrica Inteligente



Fonte: Wang *et al*. (2016b)

Assim, além de ferramentas adequadas para a integração do Ser Humano aos ambientes virtuais, uma vez que ele será a última instância no processo decisório, são necessárias ferramentas para a avaliação de performance das decisões autônomas originadas nestes ambientes virtuais e, em vários casos, dependentes do ser humano. Demonstra-se na Figura 18 a complementação da Figura 15 pela inclusão da interação do ser humano às novas tecnologias, ao mundo físico e ao mundo virtual. Conclui-se que quanto maior a intersecção entre os conjuntos, maior o grau de integração de todo o sistema e que essa estrutura de CPS é aplicável tanto em ambientes de manufatura quanto em outras operações onde exista essa necessidade de integração.

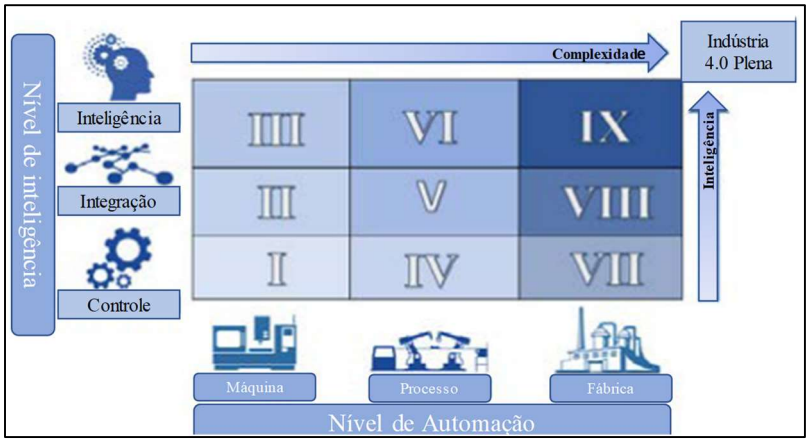
Figura 18 - Elementos integrantes de um CPS



Fonte: Próprio Autor

Demonstrando a relevância dos aspectos preditivos do CPS, Qin, Liu e Grosvenor (2016) propõem uma estrutura baseada em 2 dimensões para se avaliar o grau de maturidade da empresa com relação a utilização dos elementos da Indústria 4.0, a qual denominam roteiro para manufatura inteligente. No eixo horizontal estão os equipamentos, processos e a fábrica, e no eixo vertical estes elementos são classificados quanto ao seu grau de "inteligência" ordenados a partir do nível de controle até o nível de inteligência. No nível de controle a automação é utilizada para reduzir a necessidade de mão de obra e otimizar eficiências locais, no nível de integração as tecnologias de IoT são utilizadas integrando máquinas aos sistemas de gestão da empresa com o objetivo de aumentar o nível de informação e melhorar o desempenho. Já no nível de inteligência, os dados obtidos pela integração de IoT permitem criar planos e decisões, pelo uso de tecnologias avançadas de BDA, tornando os sistemas autoconscientes, auto otimizáveis e auto reconfiguráveis. A Figura 19 demonstra os vários níveis de integração possíveis para uma operação.

Figura 19 - Roteiro para Manufatura Inteligente



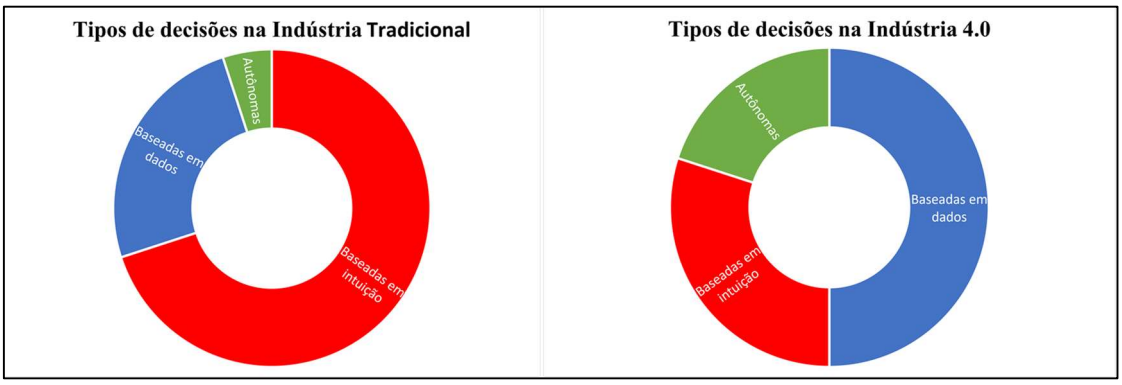
Fonte: Qin, Liu e Grosvenor (2016)

Conclui-se que o CPS possui várias definições como fábricas inteligentes, Smart Manufacturing, M – CPS, CPS, CPPS, Smart Production ou mesmo IoT/IIoT. Entretanto o CPS não é somente o resultado do investimento em TIC´s, mas sim a sua integração com o ser humano, novas tecnologias, processos de modelagem e análise de dados e, também, aos demais atores envolvidos nas relações com o sistema, como, por exemplo, fornecedores de SaaS.

No CPS a digitalização empregada visa criar um sistema mais visível e acessível, capaz de executar ações operacionais de maneira automática e aprender continuamente com as situações que se sucedem, aumentando sua resiliência. Desenvolve-se a capacidade de lidar com eventos complexos de maneira eficaz em curto espaço de tempo. Os controles digitais, presentes na Indústria 3.0, evoluem continuamente para um ambiente de inteligência integrada e acessível, resultando em auto reconfiguração e organização do sistema (BABICEANU E SEKER, 2016; CHEN *et al.*, 2017). Como consequência, a geração de informações preditivas torna-se fundamental e com maior impacto nos processos. O ser humano, apoiado por modelos preditivos, e integrado pelo acesso contínuo a informações (relevantes ao processo decisório), deverá contribuir na solução de problemas antes mesmo que ocorram. Outros elementos da Indústria 4.0 também serão integrados pela aplicação de modelos, resultantes da análise do Big Data por ferramentas matemáticas e estatísticas avançadas (aplicação da Ciência dos Dados na Manufatura). O uso de modelos como apoio a decisões em tempo real torna a análise de sua confiabilidade um aspecto crítico, diferentemente das aplicações tradicionais off-line que podem ser testadas a parte do processo, em que é possível a redução do impacto de possíveis erros. Tal disponibilidade de dados, e outras estruturas de suporte, resultam em melhorias no processo decisório com relação a indústria tradicional. Isso é ilustrado na Figura 20 onde avalia-

se o grau de integração, de uma determinada operação à Indústria 4.0, tomando-se como referência o aumento das decisões autônomas e das decisões baseadas em dados.

Figura 20 - Evolução da Estrutura de Decisão na Manufatura 4.0



Fonte: Próprio Autor

Entretanto existe um grande desafio à criação desta integração que são as características peculiares de cada sistema de manufatura, ou operações, que dificultam a elaboração de soluções padronizadas. Devido a isso, Waibel *et al.*(2017) destacam que os custos de investimento iniciais são altos, pois é necessário um sistema de TI com profissionais qualificados, redes de dados estáveis, rápidas e protegidas de ataques cibernéticos, uma vez que as consequências desses ataques são muito mais graves em ambientes digitais altamente integrados. Outro desafio refere-se aos recursos humanos, estes experimentarão o aumento da complexidade de suas tarefas diárias e se tornarão a ponte entre o mundo real e o virtual (LONGO; NICOLETTI; PADOVANO, 2017; RIMINUCCI, 2018; HECKLAU *et al.*, 2016; WAIBEL *et al.*, 2017; HELLENBRANDT *et al.*, 2019). Manyika *et al.* (2011), avaliam a necessidade de se criar estruturas e programas que facilitem a análise do Big Data por pessoal não especializado, pois a disponibilidade de profissionais capacitados em análise será insuficiente para a demanda a partir de 2018 (questão discutida na seção 4.1 em relação ao cenário de 2020). Baseado nessa afirmação, existirão cada vez mais profissionais inexperientes nesse campo, e os modelos gerados por eles necessitarão de processos mais robustos de validação.

Kaare e Otto (2015) afirmam que a "Fábrica escura" ainda está muito longe de ser viabilizada. Groover (2011, pg 11) relaciona outros aspectos dos impactos da automação sobre o ser humano, para ele, mesmo que tudo seja automatizado, o ser humano atuará em: manutenção de equipamentos, programação e operação dos computadores, trabalho de engenharia de projetos, gerenciamento da fábrica, indicando ser possível que habilidades

técnicas se tornem mais importantes que habilidades pessoais para os futuros gestores. O Quadro 8 resume os pontos de vista a respeito das habilidades dos trabalhadores nesse novo contexto, demonstrando a necessidade de trabalhadores com maior grau de conhecimento e preparo técnico.

Quadro 8 - Habilidades necessárias para adaptação do ser Humano à Indústria 4.0

Autor	Habilidades necessárias para adaptação à manufatura 4.0
Longo, Nicoletti e Padovano (2017)	Alta flexibilidade Capacidade de adaptação a ambientes de trabalho dinâmicos
Weyer et al. (2015)	Conhecimento de processos (para especificação e monitoramento)
Hecklau (2016)	Capacidade de coordenação Criatividade Solução de problemas e elaboração de estratégias
Waibel et al. (2017)	Planejamento Domínio de ferramentas de tecnologia da informação
Wang et al. (2016b)	Conhecimento estatístico
Gorecky et al. (2014)	Multifuncionalidade Conhecimento de processos (para especificação e monitoramento) Solução de problemas Flexibilidade
Hellebrandt et al. (2019)	Capacidade de monitorar a própria performance Capacidade de monitorar o sistema de produção Reagir a imediatamente a Turbulências Identificar e resolver problemas imediatamente

Fonte: Próprio Autor

As características e complexidades dos CPS´s, e sua integração com o ser humano, corroboram as proposições elaboradas ao final da seção 3.4, principalmente com relação a importância do uso de modelos preditivos. A necessidade de inclusão do elemento humano, sua interferência no desempenho desses sistemas devido as fontes de variabilidade identificadas, possibilitam uma nova proposição *viii*) uma vez que existe consenso sobre o aumento de complexidade das tarefas, e o ser humano é parte integrante destes complexos sistemas cibernético físicos, este, como parte integrante, também é fonte relevante de variação, tanto nos processos de coleta e transmissão dos dados, quanto nos processos de análise ou de execução das decisões baseadas nestas análises.

Na literatura muito há se desenvolvido no campo de análise de dados, dentro do contexto discutido, assim a seção seguinte busca contextualizar termos importantes como DS, DM, análise do Big Data (BDA) e KDD uma vez que estão relacionados a construção de modelos preditivos em diferentes áreas do conhecimento.

4 CONHECIMENTO, HABILIDADES E INTEGRAÇAO NO AMBIENTE DO DATA SCIENCE

É importante afirmar que processos de geração de conhecimento ou Descoberta de Conhecimento em Bancos de Dados (em inglês Knowledge Database Discovery – KDD), são estudados desde 1991, mesmo quando a disponibilidade de dados ainda não apresentava os padrões do Big Data (PIATETSKY-SHAPHIRO; FRAWLEY, 1991). Com o desenvolvimento do Big Data e a evolução das ferramentas de análise, surge o termo Ciência dos Dados, usado em diversos campos, entretanto relativamente recente do ponto de vista acadêmico.

As seções a seguir relacionam conexões da Ciência dos Dados com os elementos: Mineração de Dados (em inglês Data Mining – DM); Inteligência de Negócios (em inglês Business Intelligence - BI); Análise de negócios (em inglês Business Analytics - BA) e Análise do Big Data (em inglês Big Data Analytics - BDA). Demonstra-se que todos estes elementos estão relacionados a processos estruturados de análise de dados e geração de modelos já tratados pelos processos de KDD.

4.1 O AMPLO ESPECTRO DA CIÊNCIA DOS DADOS

Ainda não é possível definir uma caracterização padrão para as atividades da Ciência dos dados, entretanto, entende-se que seu objetivo é a geração de conhecimento relevante, insights significativos ou criação de valor para o negócio. Para esse fim necessita-se de conhecimento do contexto geral do problema, grandes quantidades de dados (estruturados ou não estruturados), ferramentas matemáticas, estatísticas, computacionais e processos sistemáticos de análise, que desenvolvam capacidades preditivas para apoiar decisões autônomas, ou mesmo que dependam do ser humano (AYANKOYA; CALITZ; GREYLING, 2014; FLATH; STEIN, 2018; STEINBERG, 2016; QIN, 2014). Wamba *et al*. (2015) classificaram em 5 pontos o significado de criar valor para o negócio a partir do Big Data, sendo: *i)* criar transparência; *ii)* habilitar a experimentação para descobrir necessidades, expor a variabilidade e melhorar o desempenho; *iii)* segmentar populações para personalizar ações; *iv)* substituir/apoiar a tomada de decisão humana com algoritmos automatizados; *v)* inovar em novos modelos de negócios, produtos e serviços. Szymańska (2018) compilou definições de termos aplicáveis ao contexto da Ciência dos dados, o que é apresentado no Quadro 9.

Quadro 9 - Termos e definições comuns no contexto da DS

Termo	Definição
Inteligência Artificial	Algoritmos para análise de dados para auxiliar a tomada de decisão que dá a aparência de inteligência
Inteligência de Negócios	Estratégias e métodos usados por empresas para coleta de dados e análise de informações de negócios
Quimiometria	Ciência de extração e análise de informações de dados químicos analíticos pela aplicação de algoritmos matemáticos e estatísticos adequados
Aprendizado Profundo	Algoritmos de aprendizado de máquina específicos, incluindo várias camadas de unidades de processamento não linear para identificar gradualmente as coisas em níveis mais altos de abstração
Integração de dados	O processo de combinar dados de diferentes fontes e apresentá-los em uma única exibição
Mineração de dados	A prática de extrair informações úteis de dados retirados de várias fontes
Computação de fronteira	Uma arquitetura de tecnologia da informação distribuída na qual os dados são processados na periferia da rede, o mais próximo possível da fonte de origem
Aprendizado misturado	Abordagens de aprendizado de máquina que combinam os resultados de muitos algoritmos diferentes, cujo voto combinado fornece uma saída mais confiável do que um método único
Computação evolucionária	Algoritmos inspirados na evolução biológica, parte da inteligência artificial e abordagens suaves de computação
Lógica Fuzzy	Algoritmos que lidam com informações aproximadas ou imprecisas, permitindo descrever dados com diferentes graus de associação em conjuntos predefinidos
Aprendizado de máquina	Algoritmos para extração automatizada de informações dos dados
Reconhecimento de padrões	Algoritmos que se concentram no reconhecimento de padrões e regularidades nos dados. Uma parte ou sinônimo de aprendizado de máquina. Eles podem ser supervisionados e não supervisionados
Raciocínio probabilístico	Algoritmos que usam a teoria da probabilidade para avaliar dados e inferir conclusões
Dados inteligentes	Dados prontos para uso que podem levar imediatamente a decisões
Computação Suave	Algoritmos que usam soluções não exatas para tarefas computacionais, incluindo lógica fuzzy, aprendizado de máquina, raciocínio probabilístico e computação evolutiva
Análise Supervisionada	Algoritmos que consistem em uma variável alvo / resultado que deve ser prevista a partir de um determinado conjunto de preditores
Análise não Supervisionada	Algoritmos sem qualquer alvo ou variável de resultado para prever. Seu objetivo é modelar a estrutura subjacente dos dados sem informações de análise prévia sobre eles

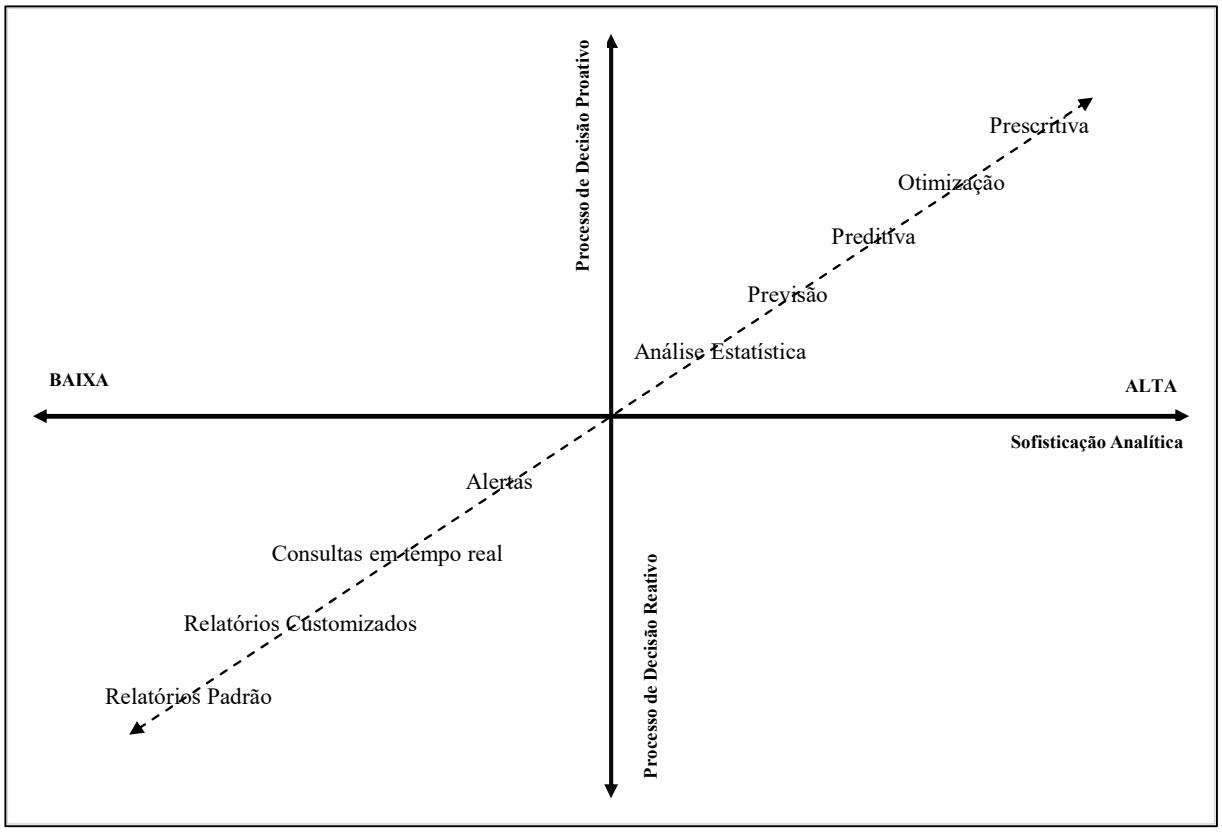
Fonte: adaptado de Szymańska (2018)

Para alguns autores os termos BDA, BA, e DS são correlatos (MAURO *et al*., 2016; FLATH; STEIN, 2018; WATSON, 2014; CHIANG; LU; CASTILLO, 2017). Entretanto, para Gandomi e Haider (2015), BDA refere-se somente a etapa de modelagem e interpretação dos dados. Enquanto que para Watson (2014) o BDA é a evolução das ferramentas e métodos de análise de dados, utilizados desde os anos 1970, e ilustrado na Figura 21. Alinhados com esta visão Banerjee, Bandyopadhyay e Acharya (2013) afirmam que o BDA é a evolução das ferramentas de análise, já disponíveis, para usos preditivos, apoiados pela maior facilidade de acesso, armazenagem e processamento dos dados, o que demonstra-se na Figura 22.

Figura 21 - Evolução dos Sistemas de Suporte à Decisão

Fonte: Watson (2014)

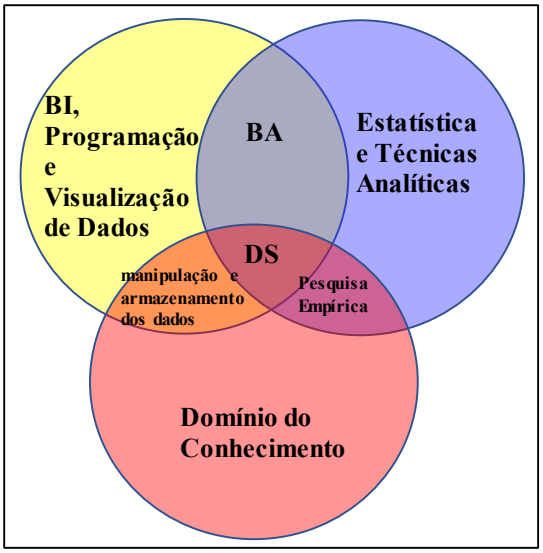
Figura 22 - Maturidade de uso das ferramentas de análise no processo decisório



Fonte: Banerjee, Bandyopadhyay, Acharya (2013)

Ayankoya, Calitz e Greyling (2014) classificam o BI como uma ferramenta aplicada a elaboração de informações descritivas, consultas específicas e relatórios, enquadrando o BDA como a união destas ferramentas com a estatística e demais técnicas de análise, sendo a DS a aplicação do BDA ao conhecimento empírico do contexto estudado, denominado domínio de conhecimento. A Figura 23 ilustra essas relações, o que também caracteriza o BDA como elemento da DS.

Figura 23 - Relacionamento entre DS, BI e BA



Fonte: Ayankoya, Calitz e Greyling (2014)

Song e Zhu (2015) afirmam que o BDA é um elemento da DS, juntamente com: a infraestrutura necessária de apoio ao Big Data; o ciclo de análise de big data; habilidades para gerenciamento de dados; disciplinas comportamentais. Em infraestrutura incluem-se tecnologias como Hadoop®, bancos de dados NoSQL, alta capacidade de processamento e soluções em nuvem. Compreende-se o ciclo de análise do BDA como: conhecimento do negócio, entendimento, preparação e modelagem dos dados, validação, desdobramento e monitoramento das soluções. Habilidades de gerenciamento de dados diz respeito a modelagem tradicional e conhecimento de bancos de dados relacionais. As habilidades comportamentais referem-se ao pensamento crítico, a capacidade de elaborar as perguntas adequadas aos problemas, a capacidade de comunicação com os especialistas da área alvo avaliada, e a capacidade de extrair resultados das análises.

Em conjunto com os dados e a tecnologia disponíveis, o elemento humano é o terceiro pilar da DS (SONG; ZHU, 2015). Desta forma, para se compreender a abrangência da DS, é necessário se caracterizar as competências e habilidades necessárias dos profissionais denominados de cientistas de dados (DAVENPORT E PATIL, 2012).

Para essa caracterização Mauro *et al*. (2017) utilizaram as informações do mercado de trabalho, contidas em descrições de cargo de anúncios de emprego postados no site Dice.com, e reconheceram 4 famílias de perfis profissionais distintas: Analistas de negócios, Cientistas de dados, Desenvolvedores e Gerentes de sistemas. De acordo com esses autores, as habilidades das famílias 1 e 2 apoiam a transformação da informação em oportunidades organizacionais.

Enquanto as habilidades das famílias 3 e 4 apoiam a transformação de dados crus em informação (captura, armazenagem e organização dos dados e softwares). Os autores conseguiram identificar as principais habilidades necessárias em cada família de conteúdo, o que está demonstrado no Quadro 10. Já Hecklau *et al*. (2016) estudaram as habilidades profissionais necessárias no contexto da Indústria 4.0 (não somente para o cientista de dados), classificando-as em competências: técnicas, metodológicas, sociais e pessoais, detalhadas no Quadro 11.

Quadro 10 - Conjunto de habilidades para as famílias de conteúdo de cargo

	Nuvem	Codificação	Gerenciamento de base de dados	Arquitetura	Gerencimento de Projeto	Gerenciamento de Sistemas	Processamento distribuído	Análise	Impacto no negócio
Analista de negócios			●		●●●	●		●	●●●
Cientista de Dados			●●		●●			●●●	●●
Desenvolvedor	●●	●●●	●	●		●●	●●	●	●
Engenheiro de dados	●●		●	●●●	●	●●	●●		●

Nota: o número de pontos em cada célula indica a relevância da habilidade para a família do cargo

Fonte: Mauro *et al*.(2017)

Quadro 11 - Detalhamento das habilidades e competências por categoria

Categoria	Competências Requeridas	Contexto
Competências Técnicas	Conhecimento de ponta	Devido ao aumento da responsabilidade do trabalho, o conhecimento está se tornando cada vez mais importante
	Habilidades técnicas	São necessárias habilidades técnicas abrangentes para passar de tarefas operacionais para tarefas mais estratégicas
	Compreensão do processo	A maior complexidade do processo exige uma compreensão mais ampla e profunda do processo
	Habilidades de mídia	O aumento do trabalho virtual requer que os funcionários sejam capazes de usar mídia inteligente, por exemplo, óculos inteligentes
	Habilidades de codificação	O crescimento de processos digitalizados cria uma necessidade maior de funcionários com habilidades de codificação
	Compreendendo a segurança de TI	O trabalho virtual em servidores ou plataformas obriga os funcionários a estarem cientes da segurança cibernética
Competências Metodológicas	Criatividade	A necessidade de produtos mais inovadores, bem como de melhorias internas, exige criatividade
	Pensamento empreendedor	Todo funcionário com tarefas mais responsáveis e estratégicas deve atuar como empreendedor
	Solução de problemas	Os funcionários devem ser capazes de identificar as fontes de erros e melhorar os processos
	Resolução de conflitos	Uma maior orientação para o serviço aumenta o relacionamento com o cliente; conflitos precisam ser resolvidos
	Tomada de decisões	Uma vez que os funcionários terão maior responsabilidade pelo processo, eles devem tomar suas próprias decisões
	Habilidades Analíticas	Estruturar e examinar grandes quantidades de dados e processos complexos torna-se obrigatório
	Habilidades de pesquisa	Precisa ser capaz de usar fontes confiáveis para aprendizagem contínua em ambientes em mudança
	Orientação para a eficiência	Problemas complexos precisam ser resolvidos de forma mais eficiente, por ex. analisando quantidades crescentes de dados
Competências Sociais	Habilidades interculturais	Compreender diferentes culturas, especialmente hábitos de trabalho divergentes, ao trabalhar
	Habilidades de linguagem	Ser capaz de entender e se comunicar com clientes e parceiros globais
	Habilidades de comunicação	A orientação para o serviço exige boas habilidades de escuta e apresentação, ao passo que aumentar o trabalho virtual requer habilidades de comunicação virtual suficientes
	Habilidades de rede	Trabalhar em uma cadeia de valor altamente globalizada e interligada requer as redes de conhecimento
	Capacidade de trabalhar em equipe	O crescimento do trabalho em equipe e o trabalho compartilhado em plataformas exigem a capacidade de seguir as regras da equipe
	Capacidade de comprometer e cooperar	As entidades ao longo de uma cadeia de valor se desenvolvem para parceiros iguais; todo projeto precisa criar situações em que todos ganham, especialmente em empresas com trabalho de projeto
	Capacidade de transferir conhecimento	As empresas precisam reter conhecimento dentro da empresa; especialmente com a atual mudança demográfica, o conhecimento explícito e tácito precisa ser trocado
	Habilidades de liderança	Tarefas mais responsáveis e hierarquias achatadas fazem com que cada funcionário se torne um líder
Competências Pessoais	Flexibilidade	O aumento do trabalho virtual torna os funcionários independentes no tempo e no local; a rotação de tarefas de trabalho exige ainda que os funcionários sejam flexíveis com suas responsabilidades
	Tolerância de ambigüidade	Aceitar mudanças, especialmente mudanças relacionadas ao trabalho devido à rotação de tarefas de trabalho ou reorientações
	Motivação para aprender	Mudanças mais frequentes relacionadas ao trabalho tornam obrigatório que os funcionários estejam dispostos a aprender
	Capacidade de trabalhar sob pressão	Os funcionários envolvidos em processos de inovação precisam lidar com o aumento da pressão, devido aos ciclos de vida mais curtos do produto e ao tempo reduzido de lançamento no mercado
	Mentalidade sustentável	Como representantes de suas empresas, os funcionários também precisam apoiar iniciativas de sustentabilidade
	Conformidade	Regras mais rígidas para segurança de TI, trabalho com máquina ou horário de trabalho

Fonte: adaptado Hecklau *et al*. (2016)

Waller *et al*. (2013) e Ayankoya, Calitz e Greyling (2014) afirmam que os cientistas de dados somente alcançarão resultados significativos se, além do conhecimento matemático e estatístico, existir conhecimento prévio do contexto do problema a ser tratado. Considerando-se o campo da gestão da cadeia de suprimentos (GCS), no Quadro 12 estão relacionadas tanto as habilidades mais importantes quanto as de menor importância para esta classe de profissionais.

Quadro 12 - Habilidades mais relevantes para o Cientista de Dados no GCS

	Conjunto de habilidades do cientista de dados	
Disciplina	**Mais Importante**	**Menos Importante**
Estatística	Amplo conhecimento de muitos métodos diferentes de estimativa e amostragem	Derivações de métodos e provas de estimativa de máxima verossimilhança
Previsão	Compreender a aplicação de métodos qualitativos e quantitativos de previsão	Compreensão dos processos estocásticos subjacentes
Otimização	Métodos numéricos de otimização	Encontrar soluções globais ideais
Simulação de eventos discretos	Projeto e implementação rápidos de modelos de simulação de eventos discretos	Teoria das filas
Probabilidade aplicada	Usando a teoria da probabilidade com dados reais para estimar o valor esperado de variáveis aleatórias de interesse	Teoria dos processos estocásticos
Modelagem matemática analítica	Usando métodos numéricos para estimar funções que relacionam variáveis independentes a variáveis dependentes	Prova de teoremas
Finanças	Orçamento de capital	Teoria de mercado eficiente
Economia	Determinando o custo de oportunidade	Teoria macroeconômica
Marketing	Ciência de marketing	Semiótica
Contabilidade	Contabilidade Gerencial	Lançamentos de débitos e créditos

Fonte: Waller *et al*. (2013)

Song e Zhu (2015) pesquisaram quais universidades Norte-Americanas já haviam implantado cursos para formação de cientistas de dados, bem como as disciplinas oferecidas. A Tabela 2 (dados de corte em 2014) apresenta a frequência das disciplinas oferecidas nestes cursos de bacharelado, e demonstra sua complexidade, que segundo eles é necessária. Entretanto essa complexidade é uma das causas da falta de profissionais nessa área, em que a demanda aumenta mais rápido que a oferta. Notícias como "Canadá abre 200 vagas para brasileiros", "Falta a Portugal alguns milhares de cientistas de dados", começam a ser recorrentes corroborando a previsão destes autores (TOMÉ, 2020; COUTO, 2020). Em pesquisa elaborada dia 27/10/20, através das palavras-chave "cientista de dados", o site "Linkedin" retornou 12.000 resultados no Brasil, sendo 376 referentes a vagas de emprego em aberto. A quantidade de vagas disponíveis é alta, visto que o país, naquele período, sofria os efeitos da pandemia de Coronavírus com recordes na média histórica de desemprego.

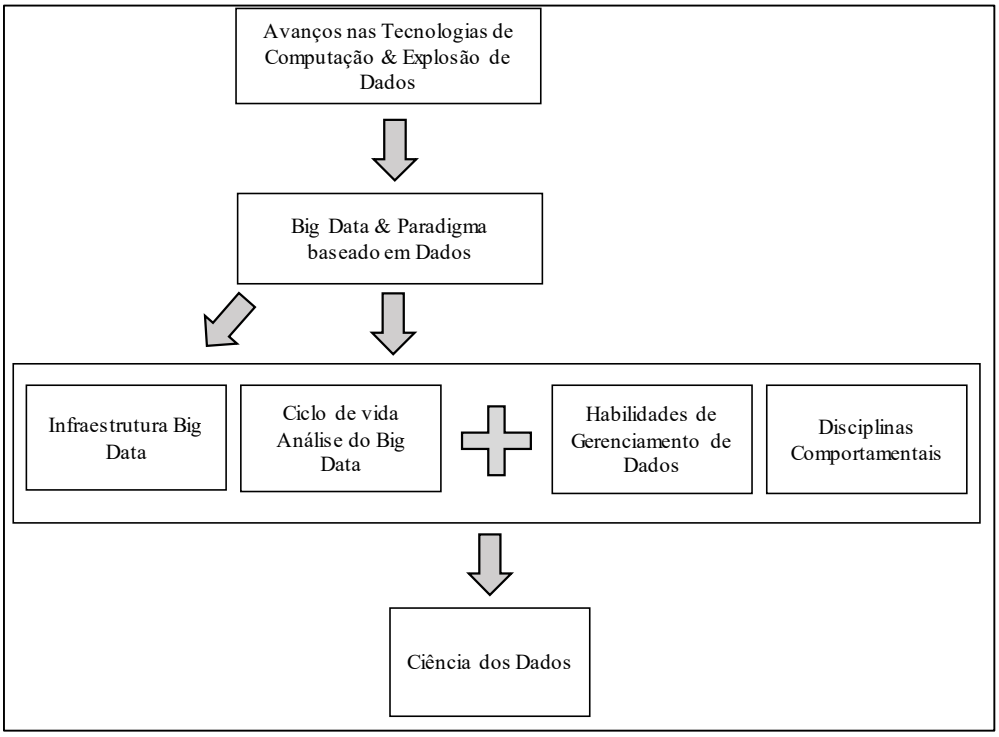
Tabela 2 - Principais disciplinas em programas de bacharelado

Disciplina	Quantidade de Universidades oferecendo a disciplina
Probabilidade e estatística	7
Mineração de dados	7
Programação	5
Matemática Discreta	4
Estrutura de dados e Algorítmos	4
Banco de Dados	4
Aprendizado de máquina	4
Modelagem Estatística	3
Visualização de Dados	3
Introdução à Ciência dos Dados	2
Inteligência Artificial	2
Segurança computacional	2

Fonte: Song e Zhu (2015)

Com base nos autores apresentados, conclui-se que o BDA é a evolução das ferramentas de BA e BI, aplicado a um ambiente de dados muito mais complexo – o Big Data – e apoiado em recursos de análise mais sofisticados, que objetivam a criação de processos decisórios proativos. O conceito de DS, um pouco mais amplo, será baseado em Song e Zhu (2015) e Ayankoya, Calitz e Greyling (2014), para estes autores o BDA refere-se as ferramentas aplicadas na etapa do processo referente a modelagem dos dados e, a DS diz respeito à todo o processo, desde; o conhecimento geral do contexto da aplicação; os recursos físicos e de informação para captura, armazenagem, tratamento dos dados; até a análise e extração efetiva de valor das informações geradas. A mineração de dados também é compreendida como uma fase dos processos de criação de modelos, e sua relação com a DS será abordada na seção 4.3. A Figura 24 resume os aspectos abordados de relações entre vários elementos que resultam na DS.

Figura 24 - O surgimento da Ciência dos Dados



Fonte: Adaptado de Song e Zhu (2015)

Concluindo, o cientista de dados necessita de uma formação técnica multidisciplinar, vide a Tabela 2, e grande conhecimento do ambiente de negócios. Retomando a proposição *viii,* apresentada ao final da seção 3.5: uma vez que existe consenso sobre o aumento de complexidade das tarefas, e o ser humano é parte integrante destes complexos sistemas cibernético físicos, este, como parte integrante, também é fonte relevante de variação; nos

processos de coleta e transmissão dos dados; nos processos de análise preditiva ou de execução das decisões baseadas nestas análises; e nos processos de modelagem dos dados. Assim, com relação aos processos de modelagem, devido à escassez de mão de obra especializada, existirão profissionais em níveis iniciais de competência, com isso a necessidade de ferramentas adequadas de validação contínua dos modelos criados por estes profissionais é uma alternativa para aferir o trabalho de modelagem e, avaliar problemas de super ajuste (em inglês – Overfitting) que ocorrem quando modelos se ajustam bem aos dados utilizados em sua elaboração, mas não aos futuros dados aos quais será aplicado (PARADY; ORY; WALKER, 2021).

A seção seguinte elenca aplicações no ambiente Industrial, demonstrando a complexidade dos problemas a serem solucionados, dos métodos a serem utilizados, do nível de conhecimento necessário aos profissionais envolvidos, e consequentemente validando a importância do uso de modelos preditivos nestes ambientes.

4.2 EXEMPLOS DE APLICAÇÕES DA DS NA INDÚSTRIA

A partir de 2012 a DS tornou-se um assunto relevante e muitos sites de bases de dados para a criação de modelos surgiram na segunda década do século XXI. Consultou-se os endereços kaggle.com, dados.gov.br, fivethirtyeight.com, portaldatransparencia.gov.br, reddit.com, buzzfeed.com, quandl.com, archive.ics.uci.edu/ml/index.php, cocodataset.org, fki.tic.heia-fr.ch/databases, moa.cms.waikato.ac.nz/ , procurando-se exemplos de aplicações na indústria nessa comunidade. Entretanto aplicações industriais nestes repositórios são escassas. Fato que pode ser explicado mais pelas dificuldades estruturais ainda existentes, como falta de profissionais experientes em análises avançadas de dados (KÜPPER *et al*., 2019), e questões de segurança de informação das empresas, que impedem a publicação de seus dados em bases públicas, do que devido a falta de oportunidades de aplicações.

Chiang, Lu e Castillo (2017) relacionam oportunidades de uso da DS nas indústrias de energia, semicondutores, farmacêutica e alimentação, que são resumidas no Quadro 13. Nesta relação fica bem claro que a DS utiliza com frequência modelos preditivos na gestão dos negócios e na análise de processos.

Quadro 13 - Oportunidades de aplicação da Ciência dos Dados nas Indústrias

Indústria	**Oportunidades de aplicação**
Energia	Estimativas de demandas energéticas, Otimização de gerenciamento de energia, Redução de impactos ambientais
Semicondutores	Melhoria de produtividade, modelos online para detecção de falhas e classificação (FCD) em processos de controle avançados (APC)
Farmacêutica	Modelos preditivos para suporte ao desenvolvimento de medicamentos, melhoria da qualidade de produto e redução de taxas de defeitos
Alimentação	Modelos preditivos para avaliação de matérias primas e produtos acabados (como exemplo previsão dos prazos de validade destes itens), modelos preditivos que evitam testes destrutivos, auxiliam na detecção de contaminação e adulterações, autenticação de qualidade de produtos (adição de açúcar no mel, óleos menos nobres no azeite extravirgem, procedência de vinhos)

Fonte: Adaptado de Chiang, Lu e Castillo (2017)

Na área industrial encontram-se aplicações que podem ser classificados dentro dos escopos de: melhoria da qualidade, manutenção preditiva, controle de desgaste de ferramental, auto recuperação de processos e previsão de desempenho.

No contexto da melhoria da qualidade, Kusiak e Kurasek (2001) estudaram o uso do que denominaram de mineração de dados (DM – Data Mining) para auxiliar na redução das possíveis causas de micro bolas de solda sob os componentes de placas de circuito impresso. O resultado foi a elaboração de um modelo, que a partir de informações de processo, gera três saídas: prediz se o problema irá ocorrer, não irá ocorrer, e a região de condições em que não se prevê adequadamente o resultado do processo. He e Wang (2018a) demonstram que o monitoramento de vários gráficos univariados conduzem a falhas na detecção, pois registros individuais de variáveis de processo, aparentemente dentro de controle, quando avaliados conjuntamente (análise multivariada), indicam processo fora de controle. Estes autores afirmam que novas tecnologias permitem o monitoramento de variáveis de processo transformadas, o que possibilita soluções para lidar com: multimodalidade, comportamento dinâmico de sistemas, relações não lineares entre variáveis, não gaussianicidade, outliers, falhas dos sensores etc.

Brandenburger *et al*. (2016) analisam os principais pontos com a ocorrência do defeito "pata de animal" em folhas de chapa de aço. No estudo divide-se cada chapa em grids, sobrepondo-se os dados de processo de centenas de chapas (milhões de dados), e desta forma encontrando os locais de concentração destes pontos (defeito pata de animal). Isso possibilita a correlação entre estes defeitos e os dados de tensão de processamento (variáveis de entrada). Os Autores afirmam que melhorias posteriores, embasadas nas análises e modelos elaborados, levaram a redução de 90% deste tipo de falha.

Stojanovic (2016) estudou desvios de forma da Hélice de resfriamento de um forno de micro-ondas. Dados de imagem 3D de peças produzidas são transformados e comparados com um padrão criado em autocad®, o que possibilita a segregação daquelas que se desviam significativamente da referência.

Guh e Shiue (2008) utilizam algoritmos de árvores de decisão para detecção online de variações na média de gráficos de controle multivariados. He *et al*. (2013) utilizam classificadores de árvore de decisão para identificação de falhas em Monitoramento multivariado de processos. Gajjar e Palazoglu (2016) usam técnicas de análise de componentes principais e gráficos de coordenadas paralelas para detecção de falhas. Uma vez identificada, um algoritmo de árvore de decisões é utilizado para o diagnóstico de causas.

Stojanovic e Milenovic (2018) criaram gêmeo digital de um equipamento de corte a laser, este modelo virtual, além de espelhar todas as operações do equipamento real, através de algoritmos de aprendizagem de máquina, baseados em métodos de aglomeração não supervisionados, analisa os seus padrões e os compara a uma base de dados de padrões anteriores, possibilitando a detecção de anomalias e atuação no sistema (alteração de parâmetros de entrada). Zhong *et al*. (2017) cita modelo desenvolvido pela Intel (fabricante de microprocessadores) que avalia online 5 TB/hora, de dados de equipamentos, para prever resultados de características de qualidade de produtos, reduzindo desta forma a necessidade de testes de qualidade. Lee *et al*. (2018) estudaram o uso de vários algoritmos de aprendizado supervisionado de máquina para prever problemas no processo de fundição de metal, tendo como dados de entrada os dados oriundos do sensoriamento do processo. Kim *et al*. (2012) estudaram a utilização de 7 algoritmos diferentes de aprendizado de máquina, que, aplicados as variáveis de entrada do processo de fabricação de semicondutores, estimam seus valores de saída e comparam com limites de controle previamente estabelecidos, indicando se o processo está ou não sob controle, sem a necessidade de leitura física das características de qualidade do produto (o que aumenta a quantidade de amostragem e reduz seu custo). Reis e Saraiva (2012)

estudam ferramentas de previsão de características de qualidade que são expressas por perfis, aplicando-as ao monitoramento das fibras do papel e aos resultados de espectrometria de vinhos. Reis e Bauer (2009) modelaram e monitoraram informações procedentes de imagens online na produção do papel, de forma a avaliar continuamente a distribuição das fibras de celulose e consequentemente a qualidade do produto.

No contexto das aplicações de DS em processos de manutenção preditiva, Çinar *et al*. (2020) afirmam que a utilização de modelos ampliou sua definição. Para estes autores o método da manutenção preditiva baseada em condição (em inglês condition based maintenance - CBM) refere-se à intervenção feita no equipamento somente quando seus sensores indicam que algum elemento monitorado apresenta falha funcional, atuando-se corretivamente neste momento, sem a oportunidade de realizar o planejamento da intervenção. O outro método preditivo, denominado de manutenção baseada em estatística (em inglês predictive maintenance - PdM), envolve a coleta de dados de sensores e a projeção do momento no tempo quando alguma falha funcional efetivamente ocorrerá nos elementos monitorados, possibilitando o planejamento da intervenção, o que garante o uso do ativo até o máximo de sua vida útil e evitam-se paradas não planejadas no sistema produtivo. Para aprofundamento no assunto, indica-se também a leitura de Carvalho *et al*. (2019), pois também relacionam métodos de aprendizado de máquina aplicados à manutenção preditiva.

He e He (2017) utilizaram métodos de aprendizado de máquina para diagnosticar defeitos em rolamentos, a partir de dados provenientes do monitoramento acústico do equipamento. Deutsch e He (2017) utilizaram dados de vibração do equipamento e métodos de aprendizado de máquina para a previsão de vida útil de componentes rotacionais em equipamentos. Luo *et al*. (2015) aplicaram técnicas de aprendizado de máquina para predição de contaminação dos equipamentos responsáveis por correção de defeitos em peças da indústria de semicondutores. Xu *et al*. (2017) criaram modelos para predição de falhas em bombas de poços de óleo e gás. O'Donovan *et al*. (2015a) discute as oportunidades e desafios para a aplicação da manutenção preditiva apoiada pela análise do Big Data (BDA). Bagheri, Zollanvari e Nezhivenko (2018) utilizaram algoritmos de aprendizado de máquina e técnicas de processamento de sinais para desenvolver um modelo de predição da condição de operação de transformadores, baseado em dados online de vibração, para detectar falhas de curtos-circuitos em estágios iniciais, antes destas se tornarem falhas mais graves.

Aplicações voltadas para predição de vida útil (ou de desgaste) de ferramental, também são relevantes para o desempenho do processo produtivo, tanto pelo uso completo do recurso

quanto por garantir os padrões de qualidade de processamento. Gao *et al*. (2017) comparam vários métodos para predição do desgaste de ferramentas de usinagem, em que utilizam algoritmos de redes neurais artificiais (em inglês Artificial Neural Network - ANN), Regressão de vetores de suporte (em inglês Support Vector Regression - SVR) e florestas aleatórias (em inglês Random Forests - RF), concluindo que a solução de RF é melhor. Yan *et al*. (2017) classificam o desgaste de ferramental como elemento da manutenção preditiva e demonstra o uso de ANN para previsão de desgaste de ferramenta de usinagem.

Na indústria química uma questão importante é a auto recuperação de processos, quando, após a detecção e o diagnóstico de uma falha (em inglês Fault detection and diagnosis - FDD), ocorre uma intervenção e o sistema retorna às suas condições padrão de operação (conceito tratado no âmbito da Indústria 4.0). Severson, Chaiwatanodom e Braatz (2016) abordam a utilização de modelos híbridos (combinação de modelos causais com modelos baseados em dados) tanto para a detecção e o diagnóstico da falha (válvulas com defeito por exemplo), quanto para ações de recuperação automática do processo (em inglês Fault-tolerant control - FTC). Os autores utilizam como exemplo o monitoramento, por imagens de estereomicroscopia de cristais dentro das tubulações. He e Wang (2018a) discutem que em muitos processos, a existência de sistemas e sensores redundantes permite, através de modelos, identificar se sinais de problemas no processo são mesmo problemas, falhas de leitura ou falhas de funcionamento dos sensores, evitando paradas e intervenções desnecessárias.

Favi *et al*. (2020) elaboraram um modelo geral para prever o tempo de desmontagem de elementos de união em componentes. O modelo prevê o tempo de desmontagem em função das dimensões e características técnicas (tipo da cabeça do parafuso por exemplo) destes elementos de união. Flath e Stein (2018) aplicaram algoritmos de aprendizado de máquina para responder ao desafio divulgado pela empresa Bosch. Através do site Kaggle.com a empresa disponibilizou informações sobre o resultado de qualidade de 2,4 milhões de tarefas fabris, propondo o desafio da criação de um modelo para a previsão dos defeitos em cada tarefa.

Os exemplos demonstram a gama de aplicações possíveis predominantemente baseadas em modelos de predição, mais especificamente os classificatórios. Estas aplicações se tornaram viáveis pelo aumento da capacidade e redução dos custos de: coleta, armazenagem e processamento de dados. Entretanto as oportunidades e desafios são concomitantes, pois os dados disponíveis muitas vezes utilizam protocolos de informação diferentes, são provenientes de fontes estruturadas e não estruturadas, são heterogêneos (cada processo tem uma particularidade), são gerados em frequências distintas, e dependentes da interferência humana.

Como consequência, processos para a análise e utilização desta informação necessitam considerar esse ambiente dinâmico, o que será considerado nas próximas seções.

4.3 PROCESSOS DE MODELAGEM

Independentemente das técnicas utilizadas, o uso modelos preditivos é importante no apoio a processos de tomada de decisão nas organizações e consequentemente na geração de valor para os negócios (SZYMAŃSKA, 2018; WAMBA *et al.*, 2015; KÜPPER *et al*., 2019). Para Fayyad, Piatetsky-Shapiro e Smyth (1996b) os limites entre predição e descrição não são nítidos, entretanto os principais objetivos preditivos englobam aspectos de:

i) Classificação – quando se necessita classificar uma instância dentro de categorias pré-determinadas, ou quando a meta é rotular corretamente uma instância (CHICCO; JURMAN, 2020);
ii) Regressão - quando é necessário predizer valores numéricos para uma determinada variável em função de outras variáveis;
iii) Agrupamento - quando se identifica uma série de categorias para descrever os dados;
iv) Sumarização – quando se encontra uma descrição completa para um subconjunto de dados;
v) Modelagem de dependência – quando se encontra um modelo que descreve dependências significativas entre variáveis;
vi) Detecção de mudança e desvio – para descobrir mudanças significativas comparando-se a dados históricos ou normativos.

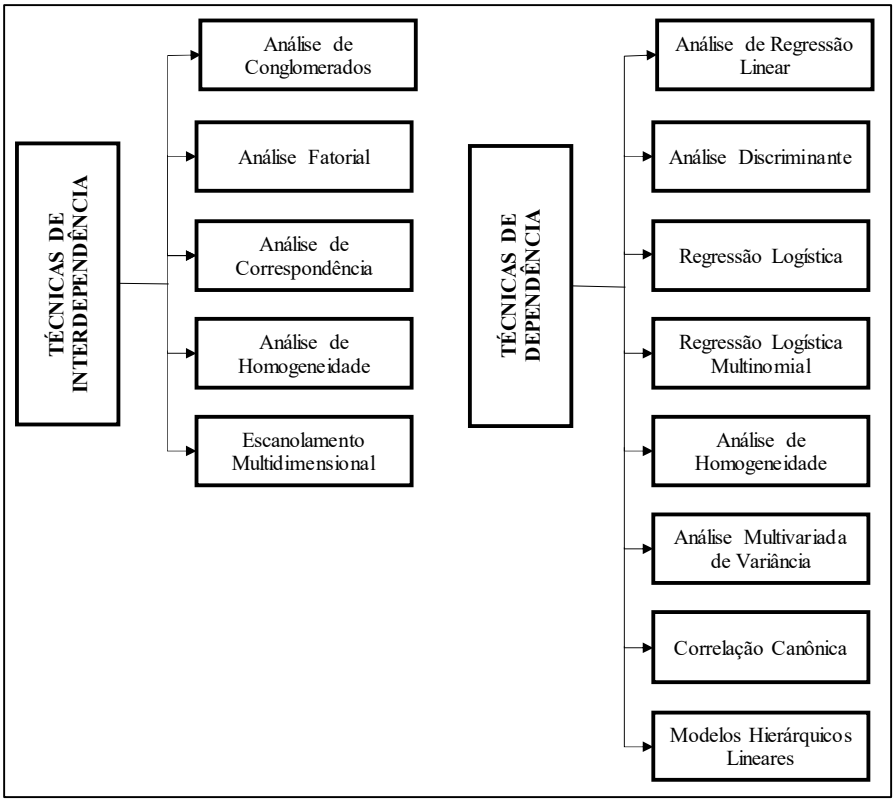
Os processos de KDD orientam de forma estruturada o alcance destes objetivos, e consequentemente são processos de apoio à criação de valor para os negócios, configurando-se em importante elemento de convergência com a DS.

As ferramentas de análise, utilizadas na elaboração dos modelos preditivos, dependem da abordagem aplicada sobre os dados disponíveis. As duas principais abordagens são denominadas: *i)* baseadas em conhecimento ou baseada em modelos (em inglês model based/knowledge based); *ii)* Modelos orientados por dados (em inglês data driven models).

A abordagem baseada em conhecimento busca construir o modelo preditivo através da aplicação de leis físicas, princípios fundamentais de engenharia, princípios fundamentais dos processos ou a aplicação de distribuições de probabilidades estatísticas a um determinado

conjunto de dados (GAO *et al*., 2017; CHIANG; LU; CASTILLO, 2017; VENKATASUBRAMANIAN; RENGASWAMY; YIN, 2003; REIS; GINS, 2017). Nessa abordagem as técnicas utilizadas buscam o estabelecimento de relações de interdependência ou dependência resumidas na Figura 25.

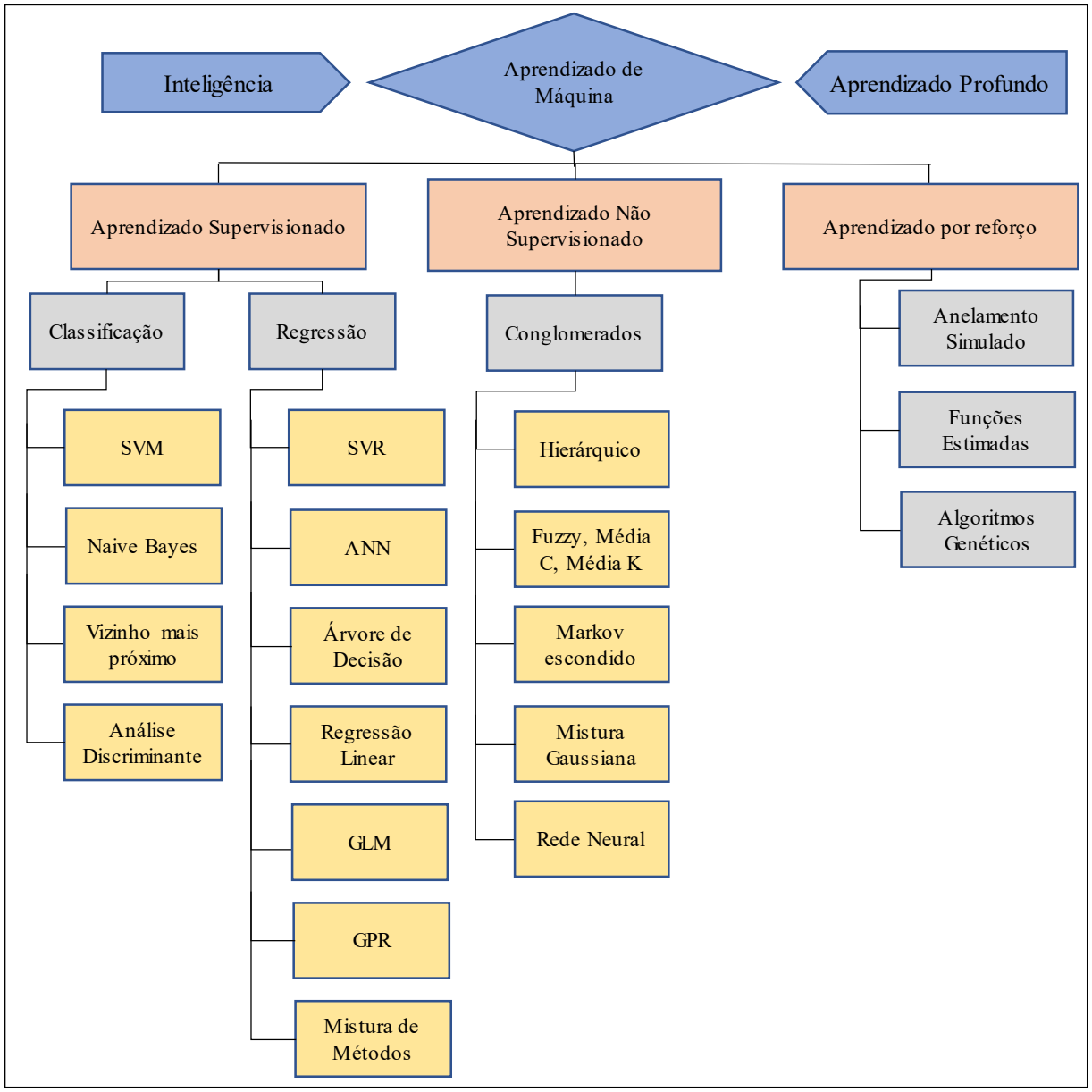
Figura 25 - Exemplos de técnicas para modelos baseados em conhecimento



Fonte: Adaptado de Hair *et al*. (2009)

Já a abordagem orientada por dados busca construir modelos preditivos utilizando algoritmos de aprendizado de máquina e grandes volumes de dados para sua criação/treinamento. Segundo Gao *et al*. (2017) a vantagem destas técnicas é que não existe a necessidade de um conhecimento profundo sobre o comportamento do sistema, em contrapartida, como não se assume nenhuma distribuição de probabilidade, estas técnicas podem não ser práticas para aplicações do mundo real. Algumas das ferramentas preditivas utilizadas para modelos baseados em dados são apresentadas na Figura 26. Algoritmos supervisionados são aqueles que utilizam dados rotulados para as etapas de elaboração e teste do modelo. Estes dados rotulados referem-se ao conjunto de dados onde há conhecimento prévio dos valores da variável dependente, isso torna possível sua comparação com os valores de predição.

Figura 26 - Exemplos de técnicas para modelos baseados em dados



Fonte: Çinar *et al*. (2020)

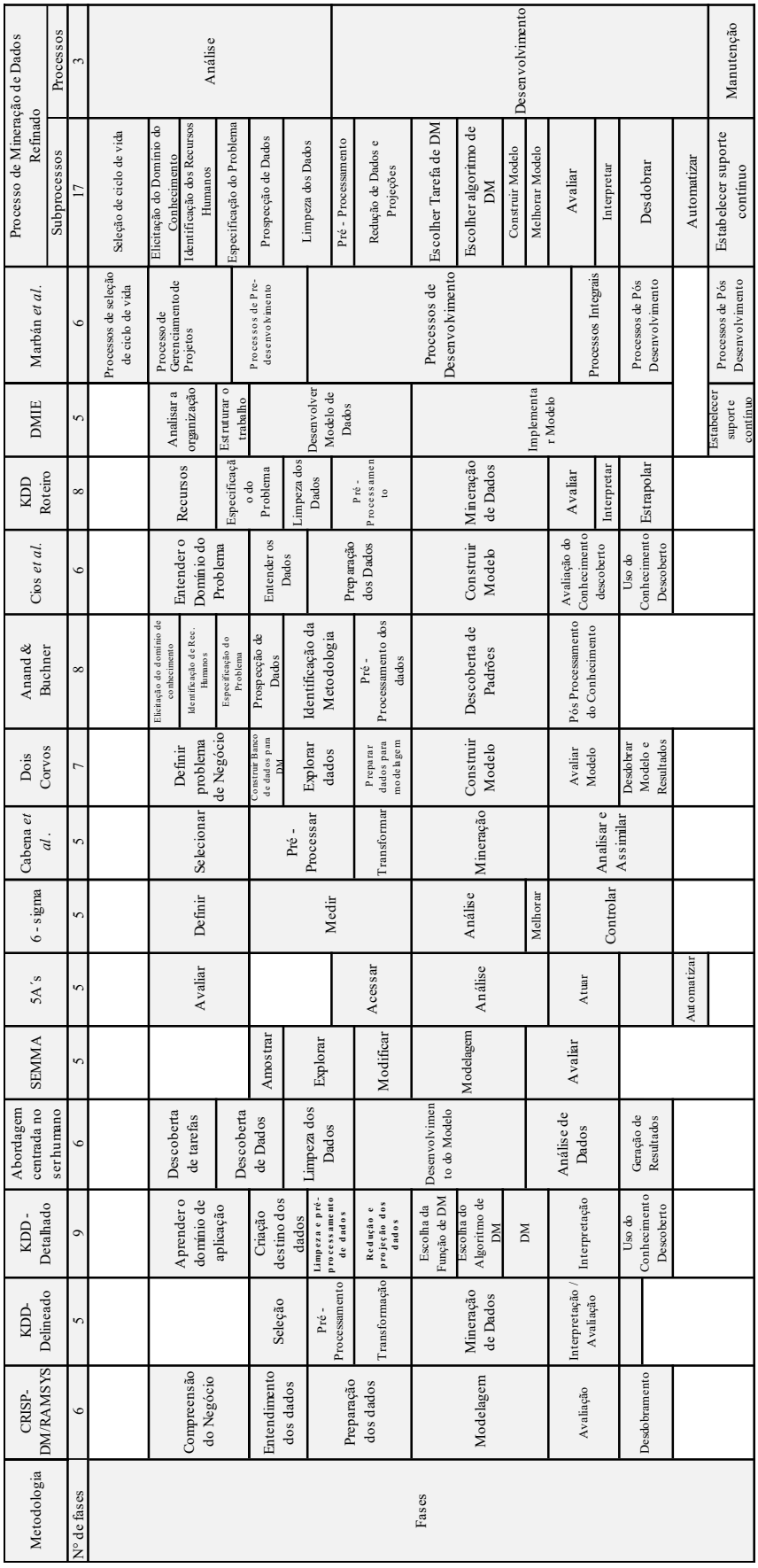
De acordo com Chiang, Lu e Castillo (2017) e Diez-Olivan *et al*. (2019), abordagens de caixa cinza (em inglês Gray-Box) deveriam ser utilizadas em aplicações onde as relações entre as variáveis estudadas são fortemente relacionadas à leis físicas. Neste contexto, o conhecimento adquirido pelos funcionários, através da experiência, deve ser conciliado com a modelagem dirigida por dados. Ambas as abordagens descritas têm como objetivo a realização de predições o mais próximas da realidade possível, obviamente quanto melhor a predição, melhor a qualidade do modelo.

Assim, a variedade de alternativas de análise e modelagem soma-se aos desafios apresentados e discutidos nas seções anteriores, com isso pode-se dizer que a qualidade dos

resultados de predição não depende somente do algoritmo aplicado, mas sim de todo o processo da elaboração do modelo preditivo, bem como de sua aplicação ao longo do seu tempo de uso. Reis (2018) propõe um modelo de avaliação de qualidade, denominado InfoQ, baseado na análise de 4 componentes: objetivo do modelo (descritivos, preditivos, relações causais, etc), banco de dados utilizado (diferentes origens e diferentes formatos), ferramenta preditiva aplicada (paramétrica, não paramétrica, probabilística, algoritmos, etc) e desempenho dos resultados preditivos. Em cada um destes componentes o autor propõe a avaliação dos dados com relação as dimensões de: resolução, estrutura, integração, seleção, generalização dos resultados, capacidade de operacionalização das soluções, tempo para completar todas as etapas de análise, comunicação clara do processo com todos os envolvidos em todos os níveis.

Desta forma compreende-se que a qualidade de um modelo é resultado: do uso de uma abordagem adequada para solução do problema de predição; da escolha da técnica a ser aplicada; do processo de análise que possibilite a avaliação dos elementos que interferem em todas as fases dessa modelagem. Nesse contexto as técnicas de KDD enfatizam que o conhecimento é o produto de um processo analítico utilizando diferentes pontos de vista, como métodos estatísticos, matemáticos, lógicos ou inteligência artificial (PIATETSKY-SHAPHIRO; FRAWLEY, 1991). Desde então vários modelos de processo de KDD foram desenvolvidos para auxiliar o processo de modelagem considerando toda a sua complexidade. Utilizam-se outros termos como sinônimos do KDD, dentre eles, os mais comuns, são a DM (em inglês Data Mining – DM), Descoberta de Conhecimento (em inglês knowledge Discovery - KD) e DM & KD. Segundo Agrawal e Shafer (1996) todos esses termos se referem ao resultado de pesquisa e uso de técnicas de extração de informação útil em grandes volumes de dados. Mariscal *et al*. (2010) avaliaram 14 diferentes processos de KDD, identificando duas abordagens principais: Descoberta de Conhecimento em Bancos de Dados (em inglês Knowledge Discovery in Databases - KDD) e o Processo padrão entre setores para mineração de dados (em inglês Cross-Industry Standard Process for Data Mining - CRISP-DM). Tomando como base de referência a abordagem CRISP - DM, compararam estes 14 processos, identificando lacunas e etapas faltantes. O resultado desta análise foi a criação do modelo definido como Processo Refinado de Mineração de Dados (em inglês Refined Data Mining Process - RDMP), que objetivou complementar o CRISP-DM com etapas relevantes encontradas nos demais processos de KDD avaliados. O resultado dessa pesquisa é apresentado no Quadro 14 e a descrição das etapas é resumida no Quadro 15.

Quadro 14 - CRISP-DM vs. demais abordagens e Processo Refinado de Mineração de Dados



Fonte: Mariscal, Marbán e Fernández (2010)

Quadro 15 - Descrição das Etapas do RDMP

Etapa	Descrição
1	Seleção do ciclo de vida: identifica, seleciona e ordena as etapas do projeto que serão desenvolvidas baseando-se no tipo de projeto a ser elaborado, bem como define critérios indicando quando cada fase pode ser iniciada.
2	Elucidação do domínio de conhecimento: melhor entendimento dos dados e do problema no contexto de sua aplicação, para restringir o espaço de busca dos algoritmos e o número de padrões a serem identificados
3	Identificação dos Recursos Humanos: uma vez conhecido o problema, escolha dos recursos humanos para execução do projeto
4	Especificação do Problema: melhor entendimento do problema e definição de objetivos do ponto de vista de DM. Tarefas, técnicas e ferramentas devem ser selecionadas nesta etapa, levando em consideração objetivos e metas
5	Prospecção de dados: 1° dos quatro processos referentes à preparação dos dados. Na realização desta tarefa os dados são coletados e campos adicionados para receber outros dados calculados a partir dos que foram coletados
6	Limpeza dos dados: preparar os dados para tarefas subsequentes como a procura e remoção dos erros, utilização de amostragem, correção de outliers, eliminação de dados faltantes ou não confiáveis e balanceamento de dados
7	Pré processamento: avaliação de padrões potenciais
8	Redução de dados e projeções: procura de características úteis e uso de métodos de redução de dimensionalidade, ou transformações nas variáveis consideradas
9	Escolha da função de DM: definir a finalidade do modelo e a técnica a aplicar (redução, classificação, regressão, agrupamento)
10	Escolha do algoritmo de DM: escolha das ferramentas específicas dentro de cada técnica aplicável
11	Construção do modelo, ou DM de acordo com várias abordagens: aplicação do modelo aos dados
12	Melhoria do modelo: revisão contínua do modelo
13	Avaliação: avaliar performance do modelo
14	Interpretação: análise dos resultados, que em caso de diferenças significativas do esperado pelos especialistas precisam ser depurados em busca de anomalias
15	Desenvolvimento: pode ser desde a organização dos dados em um relatório até a implementação de um procedimento repetitivo de DM
16	Automação: aplicação dos modelos a novos dados por não “experts”, uso em regime do modelo
17	Processo de Manutenção: diz respeito ao back up dos dados, manutenção de interfaces e atualização do modelo e softwares, pois os cenários podem mudar no tempo

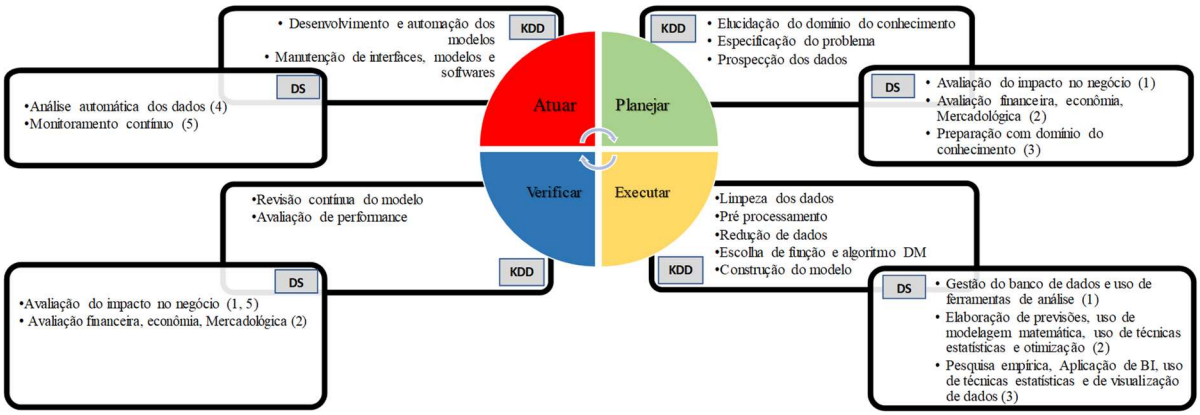
Fonte: Próprio autor

A análise do Quadro 14 auxilia a elaboração de algumas considerações: *i)* DM é um termo utilizado em três das abordagens como sinônimo da etapa de modelagem, mas isto não é consenso em toda literatura, por exemplo Kusiak e Kurasek (2001) e Fawcett (2013) empregam o termo no sentido mais amplo referindo-se ao estudo de grandes bases de dados; *ii)* todas abordagens possuem processos de validação baseados no uso de arquivos teste com resultados conhecidos (MARISCAL *et al*. 2010); *iii)* a etapa denominada suporte contínuo ou manutenção

do modelo recebe pouca importância individual, pois somente é detalhada como um subprocesso independente, em dois dos trabalhos compilados; *iv)* entre as questões a serem tratadas pelo subprocesso de suporte ao modelo está a verificação da necessidade de sua manutenção/revisão, entretanto os autores não indicam critérios para a identificação do momento no tempo em que a intervenção se faz necessária; *v)* ainda com relação à etapa de manutenção do modelo, o foco concentra-se na qualidade do algoritmo escolhido e aplicado o que torna a análise incompleta uma vez que vários elementos das etapas pregressas podem se tornar fontes de variação ao longo do tempo, interferindo na qualidade de predição dos algoritmos ou qualquer ferramenta estatística utilizada.

Conclui-se que tanto os processos de KDD, quanto o DS objetivam a extração de resultados a partir de grandes volumes de dados, a criação de ferramentas de apoio a decisão e consequentemente a melhoria contínua de processos, assim pode-se afirmar que são eles, em si, processos de melhoria contínua (FAWCETT; PROVOST, 2013). Baseado nisso compara-se alguns de seus elementos em relação ao ciclo de Shewhart ou PDCA (SOKOVIC; PAVLETIC; PIPAN, 2010; JAGUSIAK-KOCIK, 2017). A Figura 27 relaciona os conceitos de PDCA, DS e KDD, este último baseado no RDMP, dado que publicações mais recentes não apresentam diferenças significativas para estes processos (DOGAN; BIRANT, 2021; GÓMEZ-JIMÉNEZ *et al*., 2018).

Figura 27 - Demonstração de pontos de convergência entre DS e o RDMP



Fonte: Próprio Autor, sendo (1) (MAURO *et al*., 2017), (2) (WALLER *et al*., 2013), (3) (AYANKOYA; CALITZ; GREYLING, 2014), (4) (FAWCETT; PROVOST, 2013), (5) (SONG; ZHU, 2015)

É possível se compreender o DS como um processo particular de KDD, uma vez que este já era estudado anteriormente ao surgimento do Big Data e da DS. Entretanto autores

atribuem a DS, além dos processos de análise dos dados, a estrutura física e lógica necessária para o desenvolvimento das soluções (SONG; ZHU, 2015) .

Considerando-se o discutido a respeito de BDA, DM, DS, KDD e Big Data, conclui-se que as definições destes conceitos apresentam pequenas diferenças entre os vários autores. Assim, os termos DM e BDA, especificamente neste trabalho, serão utilizados como sinônimos da etapa de modelagem de dados. O termo KDD será referência para o processo completo de extração de conhecimento, independentemente do tamanho destas bases, uma vez que a literatura discute a existência de grandes bancos de dados já a partir de 1970. Aparentemente o conceito de tamanho de um banco de dados varia no tempo, em função das possibilidades tecnológicas presentes, enquanto o processo para sua gestão é muito mais perene. Já o termo DS será referência tanto para os processos de análise e geração de modelos, quanto para as estruturas físicas e lógicas para tal. Uma vez que o RDMP apresentado é resultado da união de várias técnicas, este será utilizado como referência de KDD ao longo do trabalho.

Dado que as ferramentas de modelagem e análise já tem seu uso consolidado em diversos campos, o que é recente é a aplicação destas ferramentas de forma automática, em tempo real e baseadas em número muito maior de variáveis e dados. Isso pode representar oportunidades, mas também riscos ao processo decisório, uma vez que quanto mais variáveis de entrada são utilizadas, maiores são as fontes de variação e sua capacidade de interferir no desempenho dos modelos desenvolvidos. A lacuna fica evidente, quando se demonstra que poucos trabalhos explicitam a relevância do processo de manutenção e monitoramento dos modelos (Quadro 14), e, quando o fazem, restringem-se a análise do algoritmo desenvolvido e não à todas as variáveis do processo decisório. Mesmo processos de análise mais recentes, como o Modelo DALM (em inglês Data Analytics Lifecycle Model), não abordam em maiores detalhes os demais fatores de variabilidade na qualidade de predição (SONG; ZHU, 2015). Assim, a seção seguinte analisa a importância dos processos de manutenção e validação de modelos, o que se entende por validação continua, impacto de erros de predição no processo decisório como um todo e métricas para essa avaliação.

5 VALIDAÇÃO DE MODELOS DE CONHECIMENTO

As atividades de validação não se restringem à checagem dos resultados do modelo, devendo estar presentes durante todo o seu processo de desenvolvimento, validação e manutenção. Smith e Petersen (2014), Eker *et al*. (2019), Collins e Lang (2018) e Rajamanickam *et al.* (2021) afirmam que os processos de validação possuem 3 dimensões: *i)* confiabilidade estatística, relacionada a variação dos resultados do modelo e a análise de sensibilidade de seus parâmetros, *ii)* confiabilidade metodológica, que leva em consideração a adequação dos algoritmos utilizados ao problema específico em estudo, o processo de coleta de dados (frequência, quantidade da amostra, processos de reamostragem), a qualidade estatística das informações (softwares utilizados e tipos de análises aplicadas) e demais métodos/procedimentos utilizados, *iii)* confiabilidade pessoal, que reflete até que ponto cientistas e modeladores possuem a confiança de seus pares. Rajamanickam *et al.* (2021) afirmam que uma maior atenção deve ser dedicada a estes aspectos, uma vez que não existem padrões definidos para processos de validação e o uso de modelos preditivos tem crescente relevância.

A lacuna identificada refere-se ao monitoramento contínuo da qualidade do resultado do modelo, o que a classificaria dentro do escopo da confiabilidade estatística. Conclui-se que este monitoramento é importante na identificação de oportunidades de melhoria nos modelos, e em todo o processo adotado para sua criação e aplicação, portanto, aumentando sua confiabilidade metodológica e sua confiabilidade estatística. A questão da confiabilidade pessoal não será abordada especificamente, pois depende de fatores socioculturais, além da técnica científica, apesar que esta, se adequadamente aplicada, deveria ser suficiente para os processos de validação.

Nas próximas seções serão discutidos os conceitos de validação interna, externa, temporal e contínua, quais os possíveis erros em modelos classificatórios binários e as métricas para sua avaliação denominadas *Kappa*, *AUC of ROC curve*, *MCC* e índice de *Youden.*

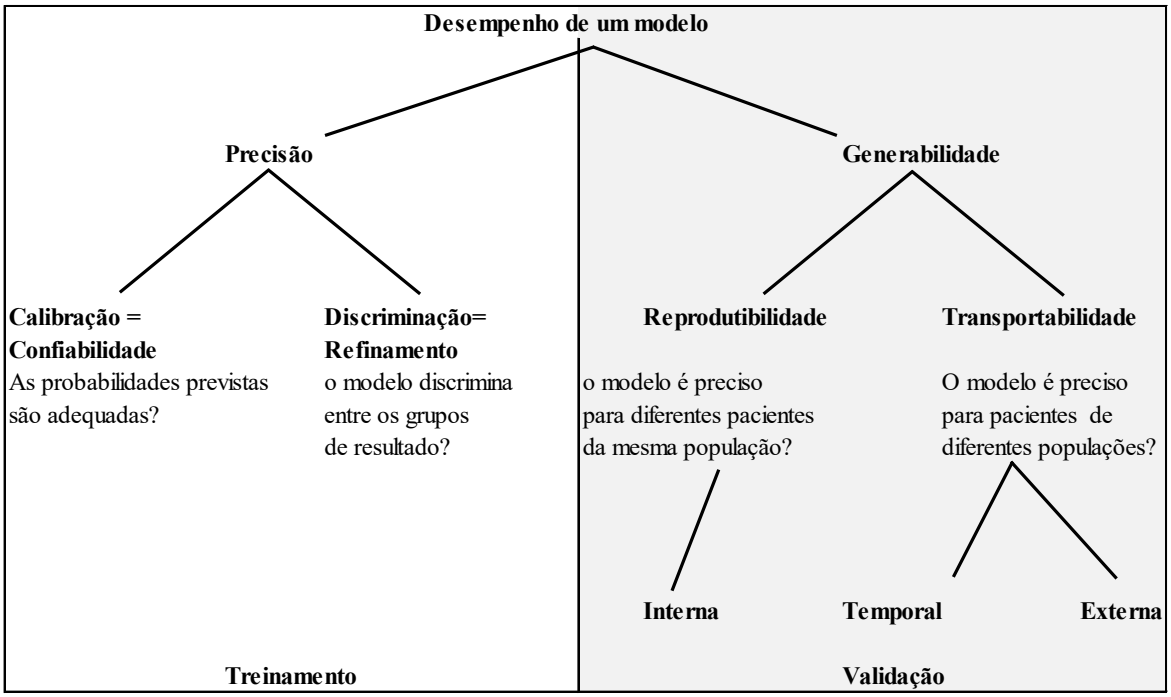
5.1 VALIDAÇÃO CONTÍNUA DE MODELOS PREDITIVOS CLASSIFICATÓRIOS

Dentro do escopo dos métodos para avaliação de resultados, Altman *et al*. (2009) os classificam em: validação interna, temporal e externa. Na validação interna, uma parte da base de dados disponível é segregada, em geral 30%, e posteriormente utilizada para avaliação do modelo (criado a partir dos 70% não segregados para validação). Segundo o mesmo autor,

quanto mais separados estes dados, melhor, pois reduz-se o efeito de super ajuste, situação na qual o modelo se ajusta muito bem aos dados utilizados na sua criação, mas não aos demais (HOFFMANN *et al*., 2019; FAWCETT; PROVOST, 2013; RAJAMANICKAM *et al*., 2021). Em bancos de dados reduzidos pode-se utilizar técnicas de validação cruzada ou expansão de dados (em inglês bootstrapping), com o objetivo da expansão das amostras (RAJAMANICKAM *et al*., 2021). Já na validação temporal, o critério para segregação entre o bloco de dados para elaboração do modelo e o bloco para avaliação de desempenho baseia-se no período de coleta das amostras (KÖNIG *et al*., 2007; ITAYA *et al*., 2022). Para estes autores, esse tipo de estratégia é intermediário entre a validação interna e a externa (quando dados similares são coletados de fontes diferentes daquelas dos dados utilizados na elaboração do modelo). Minne *et al*. (2012b) utilizam a mesma classificação, e definem a validação aparente onde o desempenho do modelo é avaliado com os dados utilizados em sua elaboração. Também relacionado à qualidade de resultados Van Vliet *et al*.(2016) utiliza o termo calibração em referência ao ajuste de parâmetros, para modelos já existentes, quando o contexto onde o modelo é aplicado é diferente do contexto de seu desenvolvimento, classificando esses métodos em: baseados no conhecimento de especialistas, calibração manual, processos automatizados de calibração e análises estatísticas. Woolley *et al*.(2012) aplicam os termos teste e validação como sinônimos afirmando que estes processos devem ser executados com dados independentes, ou seja, dados externos.

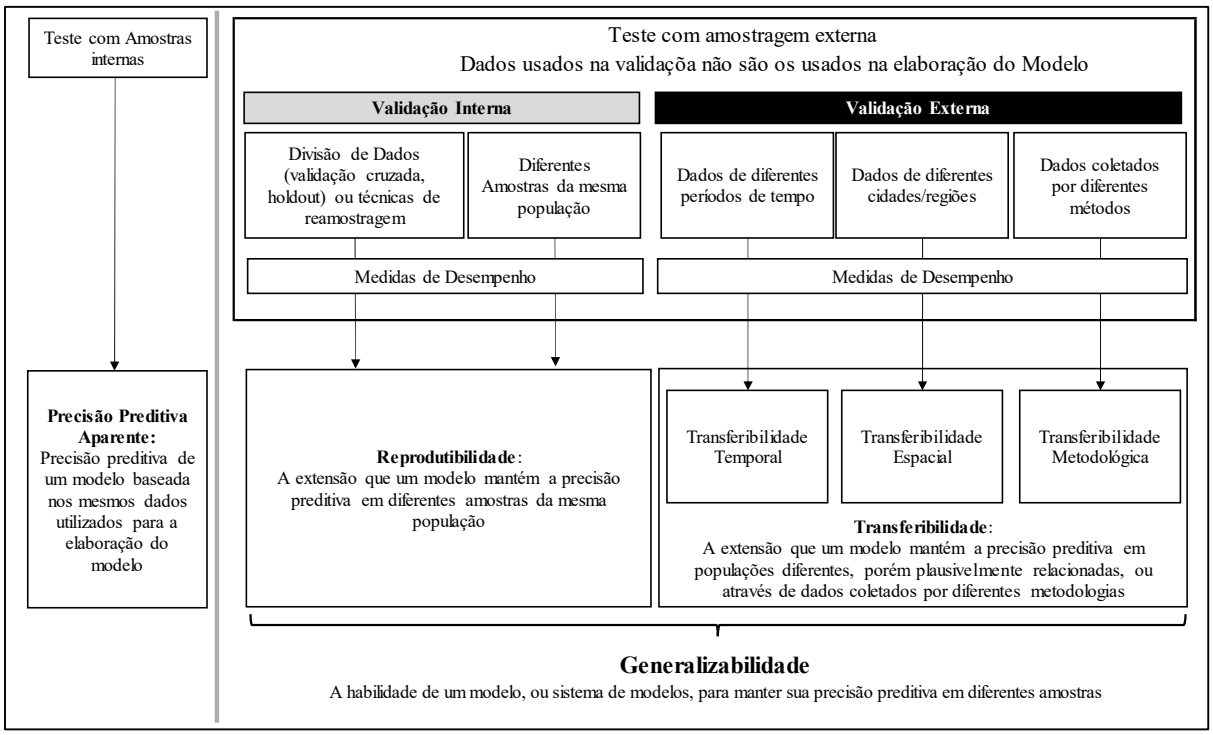
Na área de saúde König *et al*. (2007) avalia o desempenho de um modelo em função de sua precisão e generalização, demonstrados na Figura 28. Na área de transportes Parady, Ory e Walker (2021) apresentam vários métodos de expansão amostral e organizam os critérios de validação mais aplicados a modelos de transportes, concluindo que muito mais atenção deve ser dedicada aos processos de validação dos modelos da área. Esta organização é apresentada na Figura 29.

Figura 28 - Validação de Modelos Preditivos



Fonte: König *et al*. (2007)

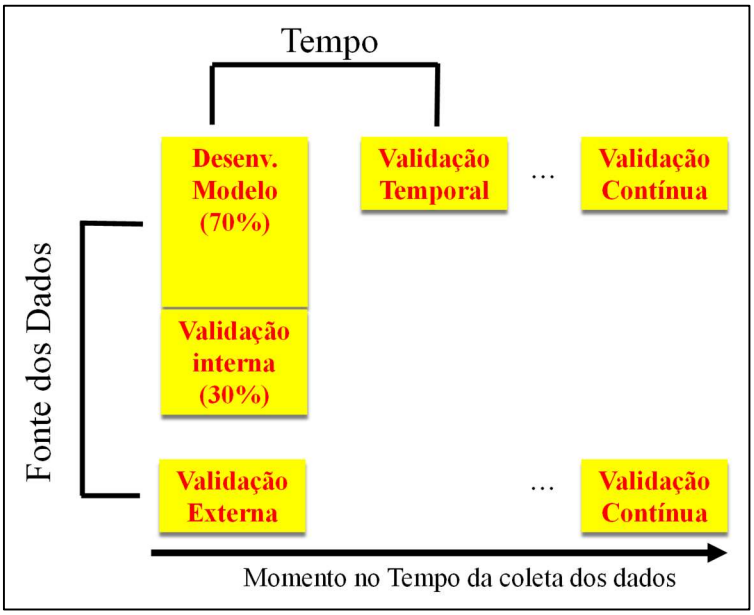
Figura 29 - Generalização como objetivo de validação de modelos estatísticos



Fonte: Parady, Ory e Walker (2021)

O conceito de validação contínua, aplica-se a checagem constante dos resultados dos modelos em relação a dados reais, considerando-se as Figuras 28 e 29, pode-se elaborar a Figura 30.

Figura 30 - Validação Contínua



Fonte: Próprio Autor

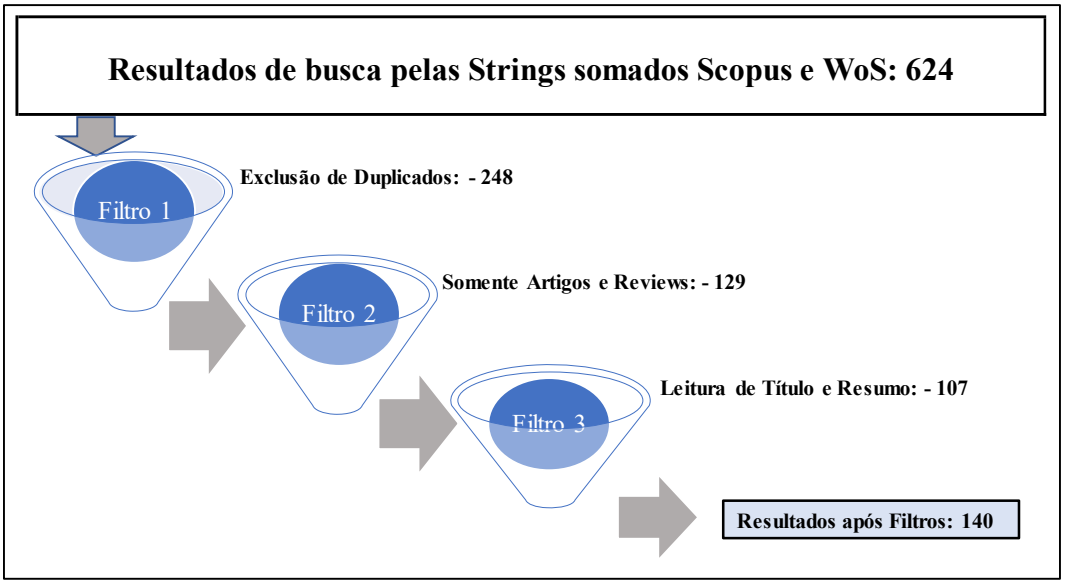
Entretanto, para melhor compreender os conceitos relacionados à validação contínua, pesquisou-se em profundidade o entendimento e as práticas sobre o tema. As buscas foram aplicadas as bases Scopus e Web of Science (WoS) nos dias 30/11/20 e 01/12/20 respectivamente. Em ambas as bases não foram aplicados limites de datas e os termos foram pesquisados nos títulos, resumos e tópicos, utilizando-se a seguinte string:

- (("continuous monitoring" AND KDD) OR "continuous validation" OR "temporal validation" OR ("concept drift" AND "continuous validation") OR ("statistical process monitoring" AND "model validation")) OR ("statistical process control" AND "model validation"))

O primeiro filtro aplicado foi a análise dos artigos que constavam das duas bases de dados (artigos duplicados), reduzindo de 624 para 376 trabalhos. O segundo filtro eliminou todos os textos que não pertenciam a classificação de artigo ou review, eliminando mais 129 documentos. Para os 247 documentos restantes foram analisados os resumos, visto que o tema só foi tratado em sua especificidade em alguns poucos trabalhos. Nos demais, os autores demonstram a importância de uma validação adequada, mas o foco específico são os modelos

e as variáveis estudadas, e não o processo de validação contínua. Nesta etapa excluiu-se os artigos fora do escopo do trabalho (107), e aqueles que se referiam a projeções de muito longo prazo (acima de décadas), validação de documentos, validação de softwares, validação de equipamentos ou sensores. A Figura 31 ilustra a aplicação dos filtros.

Figura 31 - Resultados da pesquisa



Fonte: Próprio Autor

A relação dos documentos selecionados com os nomes dos artigos, autores e ano está detalhada no apêndice A. Destes artigos, 30 foram publicados até 2011 e 110 de 2012 a 2020, essa evolução de publicações apoia a proposição já elaborada sobre o aumento da importância do uso de modelos preditivos. A Tabela 3 demonstra que a grande maioria dos trabalhos analisados pertence a área de saúde.

Tabela 3 - Trabalhos por área do conhecimento

Área	Quantidade de trabalhos
saúde	107
ecologia	20
hidrologia	3
agrícola	2
computação	1
energia	1
esportes	1
farmacêutica	1
geofísica	1
Marketing	1
química	1
rec. Hídricos	1
Total Geral	**140**

Fonte: Próprio autor

Outras informações relevantes: o termo validação temporal foi utilizado em 107 artigos, todos em concordância com o conceito apresentado na Figura 30; com relação às técnicas de modelagem a regressão logística é a ferramenta preponderante já que é citada ou aplicada em 39 trabalhos; o índice de desempenho mais aplicado é denominado de área sob a curva da curva característica de operação do receptor (em inglês Área Under Curve of Receiver Operating Characteristic Curve - *AUC* of ROC curve), presente em 59 trabalhos, que doravante será simplificado pela sigla *AUC*.

Quadro 16 - Artigos que abordam o tema Validação Contínua

Autor	Título	Ano
Dijkland S.A., Foks K.A., Polinder S., Dippel D.W.J., Maas A.I.R., Lingsma H.F., Steyerberg E.W.	Prognosis in moderate and severe traumatic brain injury: A systematic review of contemporary models and validation studies	2020
Huang C., Murugiah K., Mahajan S., Li S.-X., Dhruva S.S., Haimovich J.S., Wang Y., Schulz W.L., Testani J.M., Wilson F.P., Mena C.I., Masoudi F.A., Rumsfeld J.S., Spertus J.A., Mortazavi B.J., Krumholz H.M.	Enhancing the prediction of acute kidney injury risk after percutaneous coronary intervention using machine learning techniques: A retrospective cohort study	2018
Feidas H., Porcu F., Puca S., Rinollo A., Lagouvardos C., Kotroni V.	Validation of the H-SAF precipitation product H03 over Greece using rain gauge data	2018
Huang T., Mi H., Lin C.-Y., Zhao L., Zhong L.L.D., Liu F.-B., Zhang G., Lu A.-P., Bian Z.-X., Lin S.-H., Zhang M., Li Y.-H., Hu D.-D., Cheng C.-W., MZRW Group	Most-similar ligand based approach to target prediction	2017
Yin G., Li A., Jin H., Zhao W., Bian J., Qu Y., Zeng Y., Xu B.	Derivation of temporally continuous LAI reference maps through combining the LAINet observation system with CACAO	2017
Oduro A.R., Maya E.T., Akazili J., Baiden F., Koram K., Bojang K.	Monitoring malaria using health facility based surveys: Challenges and limitations	2016
Schmidt M., Schmidt S.A.J., Sandegaard J.L., Ehrenstein V., Pedersen L., Sørensen H.T.	The Danish National patient registry: A review of content, data quality, and research potential	2015
Munkholm S.B., Jakobsen C.-J., Mortensen P.E., Lundbye-Christensen S., Andreasen J.J.	Validation of post-operative atrial fibrillation in the western denmark heart registry	2015
Haltia U.-M., Bützow R., Leminen A., Loukovaara M.	FIGO 1988 versus 2009 staging for endometrial carcinoma: A comparative study on prediction of survival and stage distribution according to histologic subtype	2014
Oliveira A., Jesus G., Gomes J.L., Rogeiro J., Azevedo A., Rodrigues M., Fortunato A.B., Dias J.M., Tomas L.M., Vaz L., Oliveira E.R., Alves F.L., Den Boer S.	An interactive WebGIS observatory platform for enhanced support of integrated coastal management	2014
Katanoda K., Kamo K.-I., Saika K., Matsuda T., Shibata A., Matsuda A., Nishino Y., Hattori M., Soda M., Ioka A., Sobue T., Nishimoto H.	Short-Term projection of cancer incidence in Japan using an age-period interaction model with spline smoothing	2014
Richter A., Weber M., Burrows J.P., Lambert J.-C., Van Gijsel A.	Validation strategy for satellite observations of tropospheric reactive gases	2014
Foley R.N., Collins A.J.	The USRDS: What you need to know about what it can and can't tell us about ESRD	2013
Chu D., Pubu T., Norbu G., Sagar B., Mandira S., Guo J.	Validation of the satellite-derived rainfall estimates over the tibet	2011
Cheong, Y; de Gregorio, F; Kim, K	The Power of Reach and Frequency In the Age of Digital Advertising Offline and Online Media Demand Different Metrics	2010
Bolaños-Sanchez R., Sanchez-Arcilla A., Cateura J.	Evaluation of two atmospheric models for wind-wave modelling in the NW Mediterranean	2007
Manniën J., Van Der Zeeuw A.E., Wille J.C., Van Den Hof S.	Validation of surgical site infection surveillance in The Netherlands	2007
Spokoiny A., Shahar Y.	An active database architecture for knowledge-based incremental abstraction of complex concepts from continuously arriving time-oriented raw data	2007
Malan A.K., Martins T.B., Hill H.R., Litwin C.M.	Evaluations of Commercial West Nile Virus Immunoglobulin G (IgG) and IgM Enzyme Immunoassays Show the Value of Continuous Validation	2004
Kaufman Y.J., Tanré D., Gordon H.R., Nakajima T., Lenoble J., Frouin R., Grassl H., Herman B.M., King M.D., Teillet P.M.	Passive remote sensing of tropospheric aerosol and atmospheric correction for the aerosol effect	1997
Jerntorp P., Berglund G.	Stroke registry in malmö, sweden	1992
LANG, JR; BOLTON, S	A Comprehensive Method Validation Strategy For Bioanalytical Applications in the Pharmaceutical Industry - 2. Statistical Analyses	1991
Toole J.F., Teagle W.C., Howard V.J., Grizzle J., Chambless L.E.	Original contributions study design for randomized prospective trial of carotid endarterectomy for asymptomatic atherosclerosis	1989
Boehm B.W.	Seven basic principles of software engineering	1983

Fonte: Próprio autor

O Quadro 16 relaciona trabalhos que abordam a importância da validação contínua em diversas áreas do conhecimento. Vários autores afirmam que a verificação constante da qualidade preditiva é uma questão relevante, entretanto não definem critérios para sua implementação, como frequência da validação, tamanho das amostras ou critérios quantitativos que indiquem o momento em que é necessária a revisão do modelo (DIJKLAND *et al*., 2020;

HUANG *et al*., 2018; YIN *et al*., 2017; HALTIA *et al*., 2014; KATANODA *et al*., 2014; RICHTER *et al*., 2014; CHEONG; DE GREGORIO; KIM, 2010; BOLAÑOS-SANCHEZ; SANCHEZ-ARCILLA; CATEURA, 2007; MALAN *et al*., 2004; KAUFMAN *et al*., 1997; ODURO *et al*., 2016).

Os poucos trabalhos que definiram uma frequência para validação foram os de Chu *et al*. (2011) e Feidas *et al*. (2018) que avaliaram a qualidade de previsão de índices pluviométricos com base mensal. Na área da saúde Katanoda *et al*. (2014) discutem a checagem anual da qualidade de previsão de modelos de incidência de diversos tipos de câncer no Japão.

Huang *et al*. (2017) demonstraram a variabilidade dos resultados de previsão entre diferentes modelos ao longo de diferentes períodos. demonstrando que uma única medida de qualidade preditiva não é suficiente para avaliar diferentes modelos, pois são estimativas e varial em função das amostras utilizadas em seu cálculo .

Com relação aos dados de entrada utilizados na elaboração dos modelos, a validação contínua dessas bases de dados é identificada como uma questão relevante. Os dados de entrada devem refletir informações precisas e corretas, caso contrário o modelo não será elaborado adequadamente ou seus resultados serão prejudicados pela sua má qualidade (SCHMIDT *et al*., 2015; MUNKHOLM *et al*., 2015; FOLEY; COLLINS, 2013; *MANNIËN et al*., 2007; SPOKOINY; SHAHAR, 2007; JERNTORP; BERGLUND, 1992; TOOLE *et al*., 1989).

Em outras aplicações, não especificamente voltadas para modelos preditivos, como, análises estatísticas dos métodos de medição bioquímicos, a validação contínua é citada como uma etapa importante para a garantia da precisão dos resultados desses métodos (LANG; BOLTON, 1991a). E no caso dos processos de desenvolvimento de softwares, a validação contínua refere-se a verificação dos programas em cada etapa do seu desenvolvimento, para que no teste final de suas funcionalidades, a maioria das inconsistências, ou erros, já tenham sido corrigidas (BOEHM, 1983).

Os trabalhos de Minne *et al*. (2012a), Minne *et al*. (2012b) e Minne *et al*. (2012c) utilizam o termo validação temporal tanto referindo-se ao uso de dados de períodos diferentes, para a validação pontual do modelo, quanto para o processo de validação contínua. Estes autores utilizam o índice *AUC* para essa verificação. O valor destes índices é calculado a partir de diferentes bases de dados, geradas por simulação, para diferentes alternativas de modelos preditivos, e os valores lançados em gráficos de controle (um para cada modelo criado). O uso dos gráficos demonstra explicitamente a variabilidade de resultados, uma vez que cada um dos

valores dos índices é uma estimativa amostral. Estes mesmos autores concluíram que há necessidade de estudos relacionados a outros índices e ao tamanho das amostras a se utilizar nos cálculos das estimativas. Entretanto não discutem de modo detalhado a frequência de amostragem e que a variação nos resultados pode ter sua origem em outras fontes que não o modelo escolhido. Outro aspecto é o uso do índice *AUC*, como já visto amplamente utilizado na área de saúde e aprendizado de máquina, mas que na próxima seção demonstrar-se-á não ser a melhor das opções.

Verificou-se na literatura a importância e as várias ferramentas utilizadas para a validação pontual de um modelo, bem como a importância da validação contínua (ou constante monitoramento das previsões), entretanto sem o detalhamento de um método específico para orientar esse processo (COELHO *et al.*, 2016; KAKOSIMOS *et al*. 2010; VAN VLIET *et al*., 2016; SU *et al*., 2018; HOFFMANN *et al*., 2019; CINTOLO-GONZALEZ *et al*., 2017).

Como processo de validação contínua, Chikushi *et al*. (2020) propõe métodos automáticos de calibragem, ou troca contínua dos algoritmos, em função da variação de sua performance preditiva. Para algumas aplicações esta pode ser uma abordagem adequada, mas essa "caixa preta" também pode esconder problemas do processo decisório, como os já elencados anteriormente. As causas das variações devem ser diagnosticadas, analisadas e tratadas. Trocar, ou recalibrar o modelo, pode melhorar o desempenho preditivo momentaneamente, mas não auxilia no entendimento dos motivos das variações de qualidade das previsões.

Baseado no exposto conclui-se que o processo de validação tem se restringido à avaliação da qualidade matemática dos modelos criados. Demonstrou-se nas seções anteriores, que vários elementos são responsáveis pela variabilidade da qualidade das previsões. Estes elementos foram apresentados nas seções anteriores, à partir da abordagem do contexto da Indústria 4.0, mas cabe aqui elencá-los novamente: problemas na coleta das informações (falhas, falta de sensores, sensores inadequados, métodos inadequados para coleta de dados), problemas de processamento (falhas em comunicação, falhas na armazenagem de dados, problemas de capacidade, falhas de softwares, etc), intervenção humana inadequada (procedimentos e controles falhos), ou mesmo alteração na correlação entre as variáveis independentes utilizadas. Estas alterações podem ser resultado de mudanças nas condições operacionais sob os quais o modelo foi elaborado, como: alteração da qualidade da matéria prima fornecida, desgaste dos equipamentos pelo tempo de uso, entupimento de válvulas,

contaminações não monitoradas, troca de funcionários, problemas de treinamento, alterações em condições sociais, mudanças de hábitos das pessoas, etc.

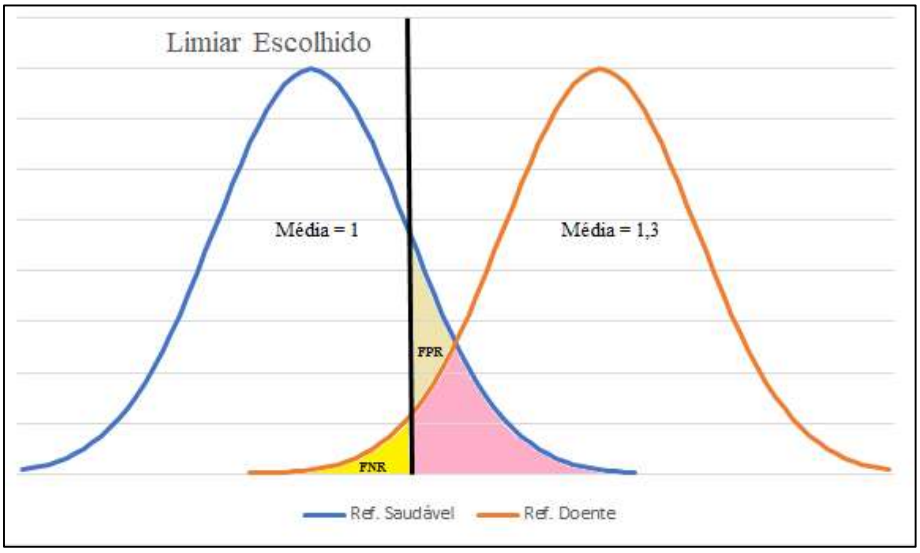
Dado que o escopo do trabalho são os modelos preditivos classificatórios binários, nas próximas duas seções serão apresentados os erros de classificação e índices que tem como objetivo quantificar estes erros.

5.2 ERROS DE CLASSIFICAÇÃO EM MODELOS BINÁRIOS

Modelos classificatórios envolvem atribuir a uma determinada instância, um rótulo referente a determinada categoria (ou classe), pertencente a um conjunto de k categorias distintas de elementos mutuamente exclusivos. Considerando problemas de classificação usando apenas duas classes (modelos binários), cada instância *I* é atribuída para um elemento do conjunto {p, n} de rótulos de classe positivos e negativos (ou sucesso e falha). Para esta atribuição alguns modelos de classificação produzem uma saída contínua, no caso da regressão logística, por exemplo, o resultado é um valor entre 0 e 1, indicando a probabilidade de sucesso P(E), para um determinado evento. A comparação do valor P(E), calculado para uma determinada instância, com um valor limite definido (L - Limiar) possibilita a classificação dessa instância como p, caso P(E) > L ou n se P(E)<= L. Outros modelos produzem um rótulo de classe discreto indicando apenas a classe prevista da instância.

Tomando um exemplo mais simples, se um exame de sangue resulta em valor acima do valor de referência para uma determinada característica ou elemento sanguíneo, a pessoa (instância, representada pela característica do sangue analisada) é classificada doente. Considerando-se os valores de referência, para pessoa saudável ou doente, comportando-se como variáveis aleatórias, existem 2 tipos de erros envolvidos: classificar uma pessoa sadia como doente (falso positivo - FP), ou uma pessoa doente como sadia (falso negativo - FN). A Figura 32 ilustra essas relações de comportamento das variáveis assumindo-se que se distribuem de conforme a Curva Normal. Percebe-se que maiores valores do limiar reduzem o FPR (em inglês False Positive Rate – FPR), mas aumentam o FNR (em inglês False Negative Rate – FNR) e vice-versa.

Figura 32 - Erros de classificação e o Limiar

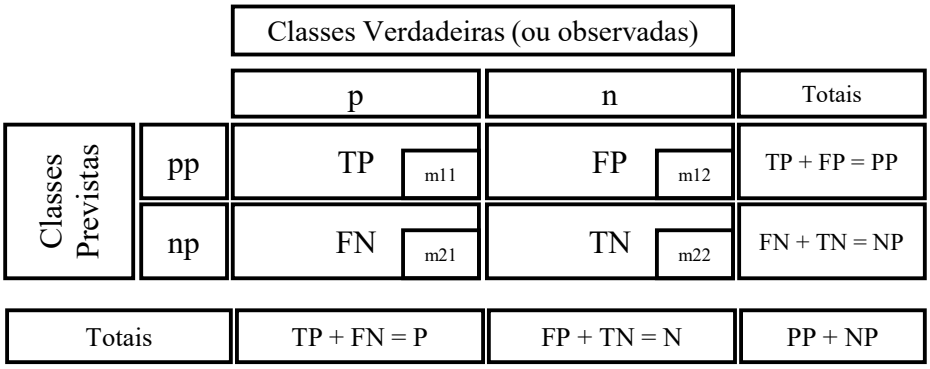


Fonte: Próprio autor

Uma Ferramenta bastante utilizada em análises de classificação é a matriz de confusão, também denominada de matriz de contingência ou concordância. Com base em Fawcett (2006) e Chicco e Jurman (2020) e Luque *et al* (2019), as medidas representadas pelas Equações de 1 a 6 podem ser extraídas da Matriz de confusão. Mas devem ser utilizadas com cautela, principalmente no caso de bancos de dados desbalanceados (mais de 90% de chance de um p ou n), em que o resultado da acuracidade pode ser razoável simplesmente devido a esse desbalanceamento (CHICCO; WARRENS; JURMAN, 2021).

No caso de modelos classificatórios binários a matriz é composta por quatro informações básicas, ilustradas na Figura 33: *i)* o valor TP – verdadeiros positivos (em inglês True Positive – TP), indicado na posição m11 da matriz de confusão, que se refere a quantidade de eventos que foram classificados como p (positivos, sucesso, classe p, etc) e realmente são p; *ii)* o valor FP - falsos positivos, indicado em m12, que se refere a quantidade de eventos classificados como p, mas que são n (negativos, falhas, classe n, etc); *iii)* o valor FN – falsos negativos, indicado em m21, que se refere a quantidade de eventos classificados como n, mas que são p; *iv)* o valor TN – verdadeiros negativos (do inglês True negative – TN), indicado em m22, que se refere a quantidade de eventos classificados como n, e que realmente são n; *v)* o valor P refere-se a quantidade total de eventos reais p; *vi)* o valor N refere-se a quantidade total de eventos reais n; *vii)* o valor PP é a quantidade de eventos classificados como p; e *viii)* o valor PN é a quantidade de eventos classificados como N. Desta forma P + N = PP + NP. Se FN e FP = 0, temos uma classificação perfeita, pois TP = P e TN = N. Sendo assim:

Figura 33 - Matriz de Confusão ou Concordância



Fonte: adaptado de Fawcett (2006)

$$FPR = FP/N \quad (1)$$

Taxa de Verdadeiros Positivos (do inglês True Positive Rate – TPR)

$$TPR = TP/P \quad (2)$$

$$\text{Precisão: } TP/ (TP + FP) \quad (3)$$

$$\text{Acuracidade: } (TP+TN) / (P + N) \quad (4)$$

$$\text{Medida F: } 2 / (1/\text{precisão} + 1/TPR) \quad (5)$$

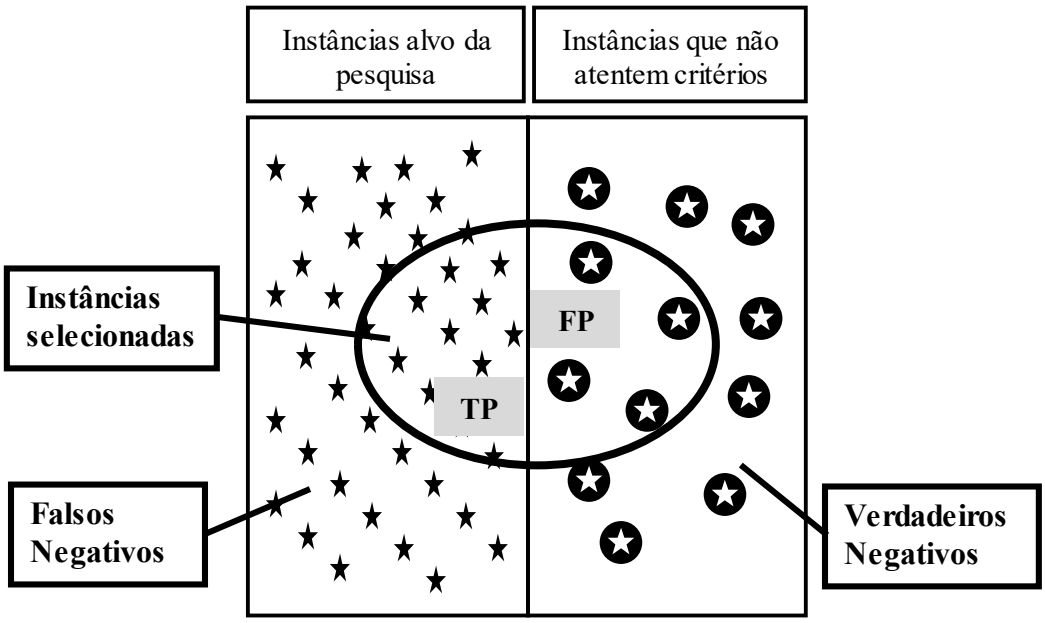
Em outras áreas de conhecimento o mesmo problema é analisado sob diferentes perspectivas. Na área de saúde índices bastante utilizados são denominados de sensibilidade e especificidade. A sensibilidade indica se indivíduos que testaram positivo para uma determinada doença (exame de covid por exemplo), realmente tem a doença, relacionando-se este conceito com a matriz de confusão temos a TPR – Equação 2. A especificidade diz respeito ao percentual de indivíduos que testaram negativo para uma doença e realmente não tem a doença (RAJAMANICKAM *et al*., 2021). Novamente relacionando-se a matriz de confusão a estes conceitos, temos a Equação (6). Em medicina denomina-se padrão ouro aos testes que apresentam resultados de sensibilidade e especificidade próximos de 100% (POWERS, 2020).

$$\text{Especificidade: } TNR = TN/N = 1 - FPR \quad (6)$$

Na área de Ciência da Informação é importante determinar a revocação e a precisão nos processos de buscas de informação dentro de um conjunto de instâncias alvo. A revocação refere-se à quantidade de instâncias alvo selecionadas dentro do conjunto de instâncias alvo possíveis, e a precisão a quantidade de instâncias alvo selecionadas entre todas as instâncias selecionadas. A Figura 34 ilustra a relação entre estes conjuntos de dados. Analisando-se sob o

ponto de vista da matriz de decisão, a revocação é a quantidade de verdadeiros positivos em relação a todos os positivos (ou instâncias alvo da busca), definido pela Equação 2,ou TPR. Já a precisão é definida pela Equação 3, referindo-se a relação entre os verdadeiros positivos e todas as instâncias selecionadas na busca. Exemplo mais específico são as buscas bibliográficas dentro dos bancos de dados acadêmicos, onde, para um determinado conjunto de critérios, deseja-se selecionar o máximo possível de trabalhos relevantes para o que se pesquisa, e o mínimo possível de trabalhos que não pertencem ao objetivo de pesquisa (OTTONI *et al*., 2013). Quanto mais abertos os critérios de busca, maior a revocação (ou a TPR), entretanto menor a precisão, pois elementos não relevantes serão selecionados, aumentando-se também o FPR refletido pela Equação 1. Os Falsos negativos são também denominados de erros de Omissão e os Falsos positivos de erros de Comissão (JIMÉNEZ-VALVERDE; LOBO, 2007).

Figura 34 - Revocação e Precisão



Fonte: próprio autor

Analisando-se a Figura 32 sob a ótica do SPM, compara-se a FPR ao erro Tipo I ou α, quando, em distribuições normais, o processo está sob controle, mas ocorrem eventos fora dos limites de controle. A FNR compara-se ao erro Tipo II ou β, quando o processo está fora de controle, mas ocorrem eventos dentro dos limites de controle. Dessa forma o SPM pode ser considerado um modelo classificatório, e os limiares representados pelos limites de controle. Mas existe uma premissa diferente no SPM, pois o processo ou está sob controle ou fora de controle teoricamente não coexistindo as duas situações ao mesmo tempo.

Pelo exposto fica evidente que a escolha do valor do Limiar interfere nos resultados classificatórios de um modelo (FPR e FNR). De acordo com Jiménez-Valverde e Lobo (2007), o critério usualmente aplicado é 0,5 para esse ponto de fronteira, ou, com menor frequência, um valor que maximize o valor de *Kappa* (índice que será detalhado na próxima seção). Esses autores também avaliam valores de limiar que minimizem a diferença entre sensibilidade e especificidade (em inglês minimized difference threshold - MDT) ou maximizem a soma de sensibilidade e especificidade (em inglês maximized sum threshold - MST), concluindo que o critério MDT é o de melhor desempenho na maioria das condições, principalmente quando erros de comissão tem maior impacto nos custos da decisão. Hernández-Orallo, Flach e Ferri (2012) afirmam que é necessário conhecer as condições de operação de um modelo (por exemplo, como os custos se comportam em função dos erros de classificação), para definir o limiar e a melhor métrica para sua avaliação.

5.3 MÉTRICAS PARA AVALIAÇÃO DE MODELOS CLASSIFICATÓRIOS BINÁRIOS

Modelos de classificação, como regressão logística, função discriminante, e algoritmos de aprendizado de máquinas têm sido amplamente utilizados para apoiar decisões na área de BI (AYANKOYA *et al.*, 2014). A literatura é clara ao mostrar a necessidade da calibração destes modelos, que, em essência, é avaliar a qualidade de seu ajuste. Esta etapa de calibração, elaboração ou ajuste dos parâmetros do modelo, é feita por meio de uma fração do banco de dados disponível, cuja recomendação é de 70% da base, reservando-se os outros 30% para sua validação (HAIR *et al*., 2009; SPENCER, 2014; EVERITT *et al*., 2011; AUSTIN, 2008). Um bom modelo necessariamente não é um modelo complexo, modelos simples também apresentam bons resultados, e modelos ideais são de fácil entendimento, clinicamente confiáveis e válidos estatisticamente por meio de dados capturados em intervalos de 3 ou 4 meses (KROIS *et al*., 2019; ALTMAN *et al*., 2009). Van Vliet *et al.* (2011) e Refsgaard e Henriksen (2004) são sugestões de trabalhos para se aprofundar nas mais variadas métricas de validação. No restante dessa seção serão apresentados os índices *Kappa* de Cohen, *AUC*, *MCC* e *Youden*, pois compreende-se serem índices capazes de considerar tanto as concordâncias de avaliação (TP e TN), quanto as discordâncias (FP e FN).

5.3.1 Índice *Kappa* de Cohen

Para os processos de validação de modelos o *Kappa* não é o mais aplicado na literatura principal referente a DS, BDA ou KDD, no entanto, é usado na área de ecologia (qualidade da

predição de: ausência ou existência de espécimes em uma região; uso da terra), na indústria automotiva (como importante indicador de concordância entre avaliadores) e especialmente no campo médico (VAN VLIET *et al.*, 2011; LOBO *et al.*, 2008; BENHAM *et al.*, 2002; PIRES; MEDEIROS, 2012).

Entre as alternativas de índices de concordância, escolheu-se o *Kappa* por excluir efeitos de acertos ao acaso, ser de fácil entendimento, estar presente em várias aplicações industriais, e ser disponível em softwares como Python e R (KUHN, 2008; PEDREGOSA *et al.*, 2015).

A Equação 7 demonstra o cálculo de *Kappa* (COHEN, 1960; ZHOU; RAZA; NELSON, 2020). Esse cálculo é baseado na matriz de confusão, lembrando que ela é formada pelas classificações reais das instâncias e pelas classificações previstas pelo modelo criado. Duas estatísticas são utilizadas: a primeira, denominada *Po*, é a proporção de concordância dos sucessos e das falhas entre o observado e o previsto pelo modelo. Baseado na Figura 33, *Po* = (TP + TN) / (PP + NP), note que para modelos binários essa é a mesma Equação 4 da acuracidade.

A segunda estatística denomina-se *Pe,* sendo a proporção esperada de concordância dos sucessos e falhas, caso a classificação fosse feita de maneira aleatória com base na chance, e não com base no modelo, os cálculos serão detalhados nos Quadros de 17 a 19.

$$Kappa = \frac{P_0 - P_e}{1 - P_e} \tag{7}$$

O Quadro 17 demonstra a matriz de confusão para o exemplo de classificação em duas categorias (binária), onde f_{xy} são as frequências em cada um dos quatro quadrantes (TP, FP, FN, TN).

Quadro 17 - Matriz de Confusão para Cálculo de *Po*

		Observado (ou real)		Total
		Sucesso	Falha	
Previsto	Sucesso	f_{11}= TP	f_{12} = FP	$T_{1.}$
	Falha	f_{21} = FN	$f_{22} = TN$	$T_{2.}$
Total		$T_{.1}$	$T_{.2}$	$T_{.1} + T_{.2}$ = n

Fonte: Próprio autor

P_0 é a frequência observada das concordâncias dividida pelo número de observações do conjunto de dados utilizado para validação do modelo, n; $P_0 = \sum P_{ii}$. Analisando o Quadro 18, $P_0 = P_{11} + P_{22}$.

Quadro 18 - Proporções baseadas na Matriz de Confusão para Cálculo de *Po*

		Observado (ou real)		Total
		Sucesso	Falha	
Previsto	Sucesso	$P_{11} = \frac{f_{11}}{n}$	$P_{12} = \frac{f_{12}}{n}$	$P_{1.} = \frac{T_{1.}}{n}$
	Falha	$P_{21} = \frac{f_{21}}{n}$	$P_{22} = \frac{f_{22}}{n}$	$P_{2.} = \frac{T_{2.}}{n}$
Total		$P_{.1} = \frac{T_{.1}}{n}$	$P_{.2} = \frac{T_{.2}}{n}$	$\frac{n}{n} = 1$

Fonte: Próprio autor

A proporção esperada $P_e = \sum P_{.i}P_{i.}$, corresponde a $P_e = P_{.1}P_{1.} + P_{.2}P_{2.}$ e é demonstrada no Quadro 19.

Quadro 19 - Proporções baseadas na Matriz de Confusão para Cálculo de *Pe*

		Observado		Total
		Sucesso	Falha	
Esperado	Sucesso	$P_{11} = \frac{T_{1.}T_{.1}}{n^2}$.	$P_{12} = \frac{T_{1.}T_{.2}}{n^2}$	$P_{1.} = \frac{T_{1.}}{n}$
	Falha	$P_{21} = \frac{T_{2.}T_{.1}}{n^2}$	$P_{22} = \frac{T_{2.}T_{.2}}{n^2}$	$P_{2.} = \frac{T_{2.}}{n}$
Total		$P_{.1} = \frac{T_{.1}}{n}$	$P_{.2} = \frac{T_{.2}}{n}$	$\frac{n}{n} = 1$

Fonte: Próprio autor

Brennan e Prediger (1981) analisaram o uso do *Kappa* desenvolvendo generalizações em relação ao seu cálculo, expressas pela Equação 8.

$$\beta = \frac{A_0 - A_c}{\max(A_0) - A_c} \quad (8)$$

Onde $A_0 = \sum P_{ii}$ (taxa de correção ou concordância), A_c é a chance aleatória de concordância e $\max(A_0)$ é o valor máximo que A_0 pode assumir. Note que $\beta = Kappa$, quando $\max(A_0) = 1$ e $A_c = \sum P_{.i}\ x\ P_{i.}$. Assim, ao analisar a Equação 7, o denominador indica

a máxima amplitude possível entre os acertos de classificação e os acertos aleatórios (devido a chance em função das probabilidades marginais), e o numerador indica a amplitude obtida. Quando *Kappa* = 1, a taxa máxima de concordância é obtida.

A equação 8 foi desenvolvida no contexto da concordância de classificação entre diversos avaliadores onde as instâncias pertencem a n classes, e os avaliadores conhecem, ou não, as proporções reais de ocorrência de cada uma dessas n classes no conjunto de instâncias avaliadas. Brennan e Prediger (1981) discutem o uso de diversos valores para A_c e $\max(A_0)$, por exemplo, no caso de 3 possíveis categorias para classificação das instâncias, uma das sugestões é tomar A_c = 1/n = 1/3 = 0,33. Como o objetivo deste trabalho são modelos classificatórios binários, o *Kappa* utilizado será o da Equação 7.

De acordo com Cohen (1960) e Brennan e Prediger (1981), para o uso do *Kappa*, devem ser consideradas as seguintes premissas: *i)* os n objetos classificados devem ser independentes; *ii)* os analistas e modelos, devem operar de forma independente. Observa-se, também, que nenhuma amostragem aleatória ou distributiva específica está necessariamente envolvida no uso do *Kappa*. Os usuários devem considerar aspectos adicionais; por exemplo, de acordo com o seu propósito, existe a possibilidade de obter altos índices *Kappa* quando há alta concordância em um dos quadrantes e baixa concordância no outro quadrante ($P_{11} >> P_{22}$ ou $P_{11} \ll P_{22}$) . Mesmo considerando essa "falha", Brennan e Prediger (1981) avaliam que o *Kappa* é um índice aceitável para avaliar concordâncias. No entanto, Delgado e Tibau (2019) indicam que certas composições de valores para TP, FP, FN e TN resultam em comportamento indesejado para *Kappa*, em que piores resultados de classificação resultam em melhores valores do índice. De acordo com Zhou, Raza e Nelson (2020) seu uso se justifica *i)* por sua simplicidade de entendimento e aplicação; *ii)* pela disponibilidade em softwares livres (pacote psych ou IRR do R) e em outros softwares como no proc. freq. do SAS; *iii)* pela correção da concordância casual. Entretanto afirmam que a estatística é limitada por sua aplicação a apenas dois avaliadores e a suscetibilidade a prevalências de doenças subjacentes extremas, ou seja, em condições de ocorrências muito desbalanceadas.

Com relação ao ponto de atenção demonstrado por Delgado e Tibau (2019), estes referem-se a situações muito particulares, em condições de classificação onde a frequência de ocorrência real de uma das categorias é muito superior da outra. Considerando-se os aspectos positivos e negativos pode-se afirmar que o *Kappa* é uma boa opção para a avaliação de qualidade de previsões, dado que outros índices, tanto os que serão apresentadas nas próximas

seções, quanto as demais possibilidades e variações, também possuem pontos positivos e negativos.

Quanto a valores de referência para *Kappa* Silva e Paes (2012) e Landis e Koch (1977) sugerem um critério prático para sua interpretação, indicado no Quadro 20, com a seguinte regra: para índices entre 0,6 e 0,79, a concordância é considerada substantiva; acima de 0,8 é considerada quase perfeita.

Quadro 20 - Interpretação do Índice Kappa

Valores de *Kappa*	**Interpretação**
< 0	Inexistência de Concordância
0,0 - 0,19	Concordância Pobre
0,2 – 0,39	Leve Concordância
0,4 – 0,59	Concordância Moderada
0,6 – 0,79	Forte Concordância
0,8 - 1	Concordância quase perfeita

Fonte: Silva e Paes (2012) e Landis e Koch (1977)

O limite máximo obtido para *Kappa* é 1, no entanto, o limite inferior pode, sob certas circunstâncias, situar-se: $-1 \leq Kappa < 0$, (COHEN, 1960). O *Kappa* também pode ser expresso em termos de frequência, dado pela Equação 9, sendo seu erro padrão expresso pelas Equações 10 e 11.

$$Kappa = \frac{f_0 - f_e}{1 - f_e} \tag{9}$$

$$\sigma_k = \sqrt{\frac{P_0(1-P_0)}{n(1-P_e)^2}} \tag{10}$$

$$\sigma_k = \sqrt{\frac{f_0(n-f_o)}{n(n-f_e)^2}} \tag{11}$$

Segundo Cohen (1960), a fórmula de desvio padrão de *Kappa* é uma aproximação, pois trata P_e como uma constante e P_0 como o valor da população. Quando o tamanho do bloco de registros é grande (n ≥100), o *Kappa* pode ser aproximado pela distribuição normal, com os

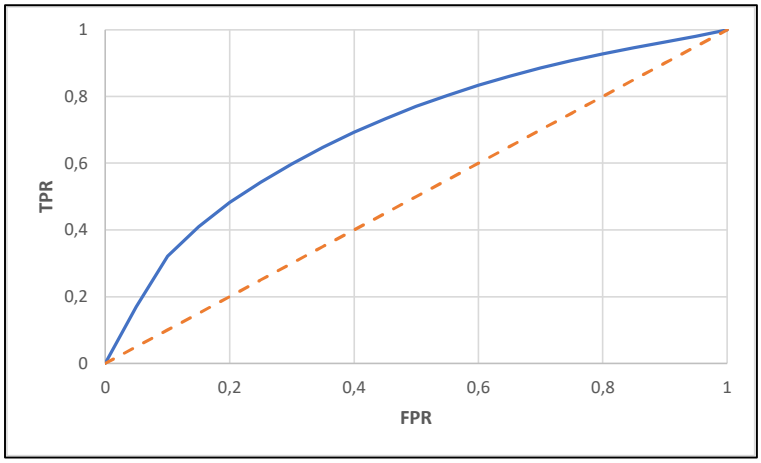
limites de controle superior e inferior (em inglês upper or Lower control Limits – UCL/LCL), determinados pela Equação 12, dado um determinado nível de confiança α.

$$CL = K_0 \pm Z_{\alpha/2}\sigma_k \quad (12)$$

5.3.2 Área sob a Curva da Curva Característica de Operação do Receptor

De acordo com Lobo, Jiménez-Valverde e Real (2008) a Curva Característica de Operação do Receptor (em Inglês Receiver Operating Characteristic Curve – ROC curve) foi desenvolvida durante a Segunda Guerra Mundial para avaliar o desempenho dos operadores de recepção de radar na detecção de sinais (para estimar o trade-off entre as taxas de acerto e as taxas de falso alarme). Seu uso, como critério de avaliação de desempenho para modelos classificatórios binários, tem se tornado comum desde então, fato demonstrado no tópico 5.1.

Figura 35 - Curva ROC

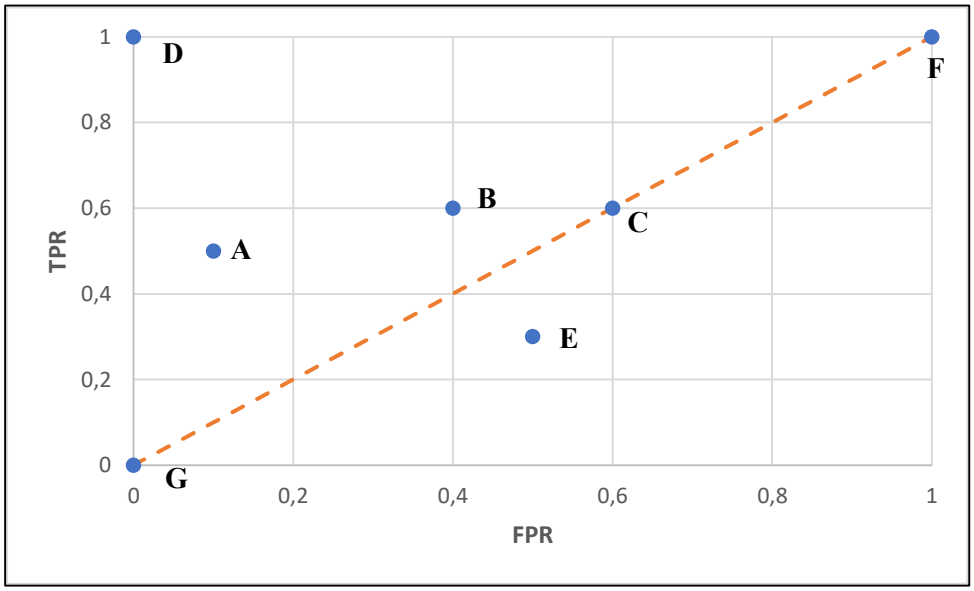


Fonte: adaptado de Fawcett (2006)

Fawcett (2006) define a curva ROC, ilustrada pela Figura 35, como gráficos bidimensionais em que a TPR é traçada no eixo Y e a FPR é traçada no eixo X, assim representando compensações relativas entre benefícios (verdadeiros positivos) e custos (falsos positivos). A *AUC* de um classificador (modelo preditivo) é equivalente à maior probabilidade de que o classificador ranqueie uma instância positiva, escolhida aleatoriamente, do que uma instância negativa, escolhida aleatoriamente. Segundo ele, entre as vantagens da curva ROC está o fato que ela possibilita visualizar e organizar o desempenho do modelo classificador, sem levar em conta as distribuições de classe ou custos de erro, também possibilita comparar classificadores que gerem pontuações em diferentes escalas (PEPE, 2012).

Para auxiliar na explicação utiliza-se os pontos particulares demonstrados na Figura 36. De modo geral, para traçar a curva ROC, varia-se o limiar aplicado a um modelo, e calcula-se a TPR e FPR para cada um dos limiares escolhidos. Esse intervalo de variação deve ser suficiente para que os pontos G(0,0) e F(1,1) sejam encontrados. O ponto G significa que a FPR é zero, no entanto a TPR também é zero. No ponto F a TPR é um (máxima revocação), entretanto a FPR também é um, assim, para aumentar os verdadeiros positivos selecionados, aumentam-se os falsos positivos. O Ponto D(0,1) representa a classificação perfeita, com TPR =1 e FPR = 0, desta forma, quanto mais próximos de D estiverem os pontos, melhor o desempenho de classificação. O ponto C está sobre a linha de não discriminação, ou seja, a situação hipotética em que o classificador escolhe aleatoriamente uma classe para a instância, baseado em determinada probabilidade. Por exemplo se o classificador escolher 50% das instâncias como positivas espera-se uma TPR de 0,5, mas também se espera uma FPR de 0,5. Com relação aos pontos A e B, acima da linha de não discriminação, a escolha do classificador dependerá se é mais importante ser conservador, classificando-se algo como positivo somente com evidências mais fortes (e assim reduzindo a FPR), ou mais liberal quando se aumenta a TPR (aumentando a revocação), mas também se aumenta a FPR. Finalmente, a respeito do ponto E, o resultado de classificação deve ser espelhado em relação a linha de não discriminação, o que leva o classificador para uma região de melhor desempenho.

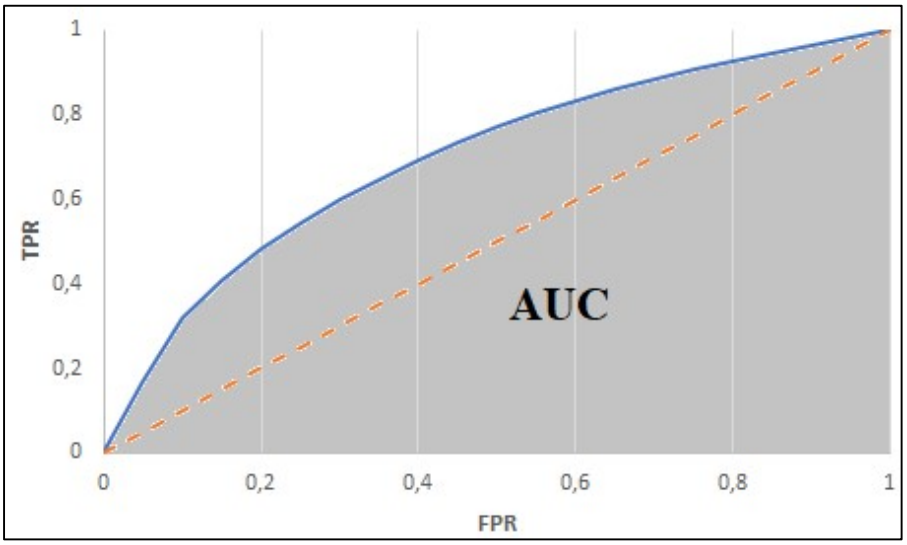
Figura 36 - Pontos Particulares da curva ROC



Fonte: adaptado de Fawcett (2006)

A Figura 37 demonstra a Área Sob a Curva (em inglês Área under Curve – *AUC*) da curva ROC, que também é denominada de *estatística c* (COOK, 2007; AUSTIN; STEYERBERG, 2017; WESSLER *et al*., 2019).

Figura 37 - *AUC* da Curva ROC



Fonte: próprio Autor

Como é uma medida única e facilmente replicável entre pesquisadores o uso da *AUC* é disseminado em várias áreas, em especial na área de saúde (BOORN *et al*., 2018; BOS *et al*., 2018),. Entretanto, Lobo, Jiménez-Valverde e Real (2008), Hand (2009) e Cook (2007) elencam deficiências deste indicador, desaconselhando o seu uso. Como muitos modelos preditivos geram um valor contínuo de probabilidade, para uma determinada instância, e sua classificação entre sucesso/falha, presença/ausência, positivo/negativo, é feita pela comparação deste valor com o limiar escolhido (acima do limiar é sucesso/presença/positivo, abaixo é falha/ausência/negativo), a *AUC* ignora o valor real de probabilidade, sendo insensível a transformação das probabilidades previstas que preservem a classificação das instâncias. Comparando-se com o *Kappa,* se as proporções reais de sucesso/falha se alterarem, o valor de *Pe* vai se alterar, e mesmo que mantidos os valores de TP e TN, o valor de *Kappa* se altera, indicando se o modelo alterou seu desempenho em função dessa variação de probabilidades. No caso da *AUC* um modelo pode melhorar ou piorar seu nível de acerto, enquanto a *AUC* pode não apresentar variações.

Outro ponto é que o teste de desempenho é realizado para todas as regiões do espaço ROC, todavia, raramente os pesquisadores estarão interessados em todas as situações possíveis (valores de limiares possíveis). Por exemplo, os lados direito e esquerdo extremos do espaço

ROC são geralmente inúteis, pois correspondem a altas taxas de falso-positivo e altas taxas de falso-negativo, respectivamente.

A *AUC* pondera erros de omissão e de comissão igualmente, enquanto em muitas aplicações de modelagem esses erros podem possuir pesos diferentes.

Utilizar a *AUC* é equivalente a medir o desempenho das regras de classificação usando métricas que dependem das regras que estão sendo medidas, ou seja, se a *AUC* depende do classificador, a métrica de avaliação entre classificadores, é diferente para cada um deles.

Segundo Austin (2008) a *AUC* não considera o tamanho do desacordo entre as respostas observadas e previstas.

Sokolova e Lapalme (2009) estudaram a variabilidade dos valores da *AUC* em função de 8 possíveis alterações dos valores de cada posição da matriz de confusão (TP, TN, FP, FN). Os resultados demonstraram a invariabilidade de resultado da *AUC* para 6 condições, e variação para 2 condições. Chama a atenção que o valor da *AUC*, não se alterou quando era esperado que se alterasse, e se alterou quando todas as posições da matriz de confusão foram multiplicadas por um mesmo escalar, ou seja, todas a proporções foram mantidas, significando que o modelo classificatório aplicado manteve seu desempenho. Este aumento amostral, sem alteração na matriz de confusão não alteraria o resultado do *Kappa*.

Nessa seção, importantes aspectos da *AUC* foram apresentados e apesar da ampla utilização da referida medida de desempenho, entende-se que as inconsistências elencadas são significativas o suficiente para se evitar seu uso, uma vez que várias outras medidas cumprem este papel.

5.3.3 Coeficiente de Correlação de Matthews (*MCC*)

De acordo com Jackson, Somers e Harvey (1989) e Chicco, Tötsch e Jurman (2021) não existe consenso sobre qual indicador estatístico é o melhor para se avaliar uma determinada matriz de confusão, existindo certa subjetividade na escolha deste índice, em função de cada conjunto de dados. Chicco, Tötsch e Jurman (2021) afirmam que o *MCC* é um indicador robusto, desde que os bancos de dados analisados sejam balanceados (ou seja, cada uma das classes possua probabilidades similares de ocorrência), e que valores de $MCC > 0{,}9$ indicam que os índices de precisão e *Youden* também terão bom desempenho. Powers (2020) demonstra a existência de viés em vários indicadores (revocação, precisão, medida F), afirmando que se deve compreender cada indicador antes de sua aplicação.

Delgado e Tibau (2019), Powers (2020) e Boughorbel, Jarray e El-Anbari (2017) demonstram que o *MCC* também aplica-se a bancos de dados não balanceados, bem como adaptam-se a problemas envolvendo várias classes.

Desenvolvido originalmente por Matthews (1975), Chicco (2017) e Boughorbel, Jarray e El-Anbari (2017) apresentam o *MCC* em função dos valores encontrados na matriz de confusão (TP, TN, FP, FN), como demonstrado na Equação 13.

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{13}$$

Avaliando-se Equação 13, verifica-se que o *MCC* não é definido se ao menos uma das quantidades TP+FN, TP+FP, TN+FP ou TN+FN forem zero, o que é um fator dificultador em alguns casos (CHICCO; JURMAN, 2020). O valor do *MCC* sempre estará no intervalo [-1,1], onde 1 significa total concordância, -1 total discordância e 0 indica que o desempenho do modelo preditivo não é melhor que o acaso.

Em bancos de dados desbalanceados, com 95% de positivos/sucessos, por exemplo, se todos os eventos forem classificados como positivos, o valor de precisão (Equação 3) indicará um bom resultado preditivo, mesmo que todos os eventos negativos sejam incorretamente classificados, o que não acontece com os valores do *MCC* para o mesmo caso. Pertinente a essa questão, Chicco (2017) orienta, em casos de bancos de dados muito desbalanceados, que a base de treino seja elaborada de forma que a quantidade de cada classe seja uma média entre 50% e a proporção real para cada classe. Por exemplo, se os positivos representam 90% dos casos, para a base teste, o arquivo deve possuir uma proporção de (90% + 50%)/2 = 70% de instâncias positivas, e, consequentemente 30% de instâncias negativas.

De acordo com Flath e Stein (2018) o *MCC* pode ser a referência para a otimização do modelo em função dos custos de falsos positivos e falsos negativos sendo o trade-off ótimo determinado pelo respectivo custo dos erros. Em ambientes com altos custos para defeitos não detectados (por exemplo, recalls de produtos), o número de falsos negativos será minimizado, em ambientes com altos custos para alertas errados (por exemplo, controle de qualidade complexo), o número de falsos positivos será minimizado. A capacidade de considerar a estrutura de custos do problema subjacente é um benefício adicional da métrica *MCC* principalmente no contexto de manufatura.

Chicco, Warrens e Jurman (2021) avaliam que o MCC é um índice mais confiável, mas também demonstram que para um classificador com resultados melhores que a classificação aleatória (MCC > 0) a diferença entre o *Kappa* e MCC é negligenciável.

Luque *et al*. (2019) demonstram que o MCC é a melhor escolha para bancos de dados desbalanceados para contextos em que os erros de classificação devam ser considerados.

5.3.4 Índice de Youden

Em consulta à WoS, no dia 06/01/2022, buscando em tópicos o termo "Youden Index", 2.294 artigos foram relacionados, cerca de 75% deles eram referentes à área de saúde, o que demonstra a relevância do indicador nessa área do conhecimento. Inicialmente o índice foi aplicado para a análise da qualidade de testes para detecção de doenças, exemplo são os testes de covid-19 já discutidos na introdução.

Youden (1950) o definiu conforme a Equação 14, que representa a média de dois conjuntos: a proporção dos indivíduos corretamente classificados como doentes, descontados os indivíduos doentes incorretamente classificados como sadios; e a proporção dos indivíduos corretamente classificados como sadios, descontados os indivíduos sadios incorretamente classificados como doentes. Com alguma simplificação matemática deduz-se a Equação 15. Os elementos desta equação são diretamente relacionados a matriz de confusão da Figura 33.

$$J = \frac{1}{2}\left[\frac{(TP-FN)}{(TP+F\ )} + \frac{(TN-FP)}{(TN+FP)}\right] \tag{14}$$

$$J = \frac{TP}{(TP+FN)} + \frac{TN}{(TN+F\ )} - 1 \tag{15}$$

Segundo Youden (1950) o indicador é uma estimativa, pois é baseado em uma amostra, sendo seu desvio padrão calculado pela Equação 16. Desta forma, a comparação entre os resultados classificatórios de 2 modelos diferentes deve se basear na comparação dos intervalos de confiança das estimativas. Estas podem ser aproximados pela normal, se a amostra for maior que 20 e o valor do índice não estiver nos extremos possíveis dos resultados (0 ou 1). O teste t de Student também é indicado nesse caso.

$$s_J = \sqrt{\frac{(TP*FN)}{(TP+FN)^3} + \frac{(TN*FP)}{(TN+F\ )^3}} \tag{16}$$

Hagan e Li (2018) propõem a Equação 17 para o cálculo do índice, de maneira a considerar a importância relativa (assimetrias) entre sensibilidade e especificidade, onde *w* é um peso definido pelo usuário. Note que se *w* = 0,5 a Equação 17 será igual a Equação 15.

$$J = 2[w\left(\frac{TP}{(TP+F\ )}\right) + (1-w)\left(\frac{TN}{(TN+FP)}\right)]\ -1 \quad (17)$$

Segundo Fluss, Faraggi e Reiser (2005) o índice é uma medida condensada da curva ROC e outra denominação para ele é informação do bookmaker. Uma comparação feita entre *MCC*, *Kappa* e *Youden* (J) demonstrou que o índice possui comportamento mais linear para bancos de dados desbalanceados, os autores deste estudo afirmam que diferente de J, *MCC* e *Kappa* não são indicados para bancos de dados desbalanceados, em contraposição ao apresentado nas seções 5.3.1 e 5.3.3 (ZHU, 2020).

Dado o exposto sobre métricas de validação, demonstrou-se a complexidade inerente a avaliação de desempenho de modelos classificatórios binários. Essa também é demonstrada por: Austin (2008) ao expandir um banco de dados (bootstrapping) em 1000 amostras e calcular a *AUC* para diversas opções de modelos; e Aggarwal (2019) que, para a validação de modelos classificadores de imagens, analisa o resultado médio e desvio padrão do *MCC*, especificidade, sensibilidade, F1 score, valor preditivo positivo (em inglês Predictive Positive Value – PPV) e Valor preditivo negativo (em inglês Negative predictive value), cabendo uma ressalva, o autor utilizou a fórmula de desvio padrão populacional e não amostral.

Conclui-se que nenhum dos indicadores apresentados é livre de restrições e que deveriam auxiliar na consideração da assimetria de importância entre falsos positivos ou falsos negativos. Sendo assim, avaliar um modelo de uso contínuo, baseado em uma única amostragem, e quiçá em um único indicador, pode ser insuficiente, pois os índices são variáveis aleatórias, baseadas em amostras da população e seu desempenho varia em função do contexto.

.

6 MONITORAMENTO ESTATÍSTICO DE PROCESSO, SPM

Para compor o background necessário, que será utilizado no capítulo 7, as subseções seguintes apresentam os fundamentos e as novas oportunidades de aplicação do SPM quando inserido no contexto do Big Data.

6.1 CONCEITOS BÁSICOS DE SPM

No início do século XX, o controle estatístico do processo (em inglês Statistical Process Control – SPC) foi a primeira metodologia a adotar técnicas estatísticas para monitorar a variabilidade dos processos na indústria, posteriormente referenciado como Monitoramento Estatístico do Processo (em inglês Statistical Process Monitoring – SPM), tornou-se uma ferramenta de qualidade cada vez mais relevante (MUKHERJEE, 2015; SRIKAEO; HOURIGAN, 2002; MOHAMMADIAN; AMIRI, 2013; WOODALL; MONTGOMERY, 2014; GOEDHART; WOODALL, 2021). O termo SPM caracteriza o controle do processo tanto pelas variáveis de saída (características do produto), quanto pelas variáveis de entrada ou parâmetros do processo (SALEH *et al*., 2021).

O potencial dos gráficos de controle, elemento fundamental das técnicas de SPM, reside em possibilitar a identificação e diferenciação entre causas comuns e causas especiais de variabilidade dos processos. As causas comuns referem-se a variáveis que não são passíveis, ou não são economicamente viáveis, de se identificar e controlar, resultando em aleatoriedade inevitável. Por sua vez, as causas especiais devem-se a fontes de variação que podem ser identificadas e removidas, como: defeitos mecânicos, manejo incorreto de máquinas e erros humanos (MUKHERJEE, 2015; HE *et al*., 2013).

A implantação do SPM envolve tradicionalmente 2 Fases (JONES-FARMER *et al*., 2017). A Fase I trata da definição da característica da qualidade e da coleta e análise dos dados para estimativa de parâmetros estatísticos, tais como a média e desvio padrão da média, que serão monitorados a partir de gráficos de controle. Nesta fase, a situação ideal é conhecer o parâmetro a ser controlado e seu desvio padrão. A partir destes dados calculam-se os limites de controle superior e inferior (em inglês Upper control limit – UCL e Lower Control limit – LCL) das cartas de controle que serão utilizadas na Fase II, ou etapa de monitoramento. As Equações 18 e 19 referem-se à situação em que tanto o parâmetro θ, quanto o desvio padrão do parâmetro (σ_θ), são conhecidos (na literatura condição KK, referentes a palavra em inglês Known)

$$UCL = \theta + h\sigma_\theta \quad (18)$$

$$LCL = \theta - h\sigma_\theta \quad (19)$$

Entretanto estas informações raramente estão disponíveis em aplicações reais (CHAKRABORTI; HUMAN; GRAHAM, 2009; OPRIME; MENDES, 2017; CHEN; SONG, 2012). Neste caso as Equações 20 e 21 são aplicáveis, onde $\hat{\theta}$ representa a estimativa de θ e $\widehat{\sigma_\theta}$ a estimativa de σ_θ (na literatura essa é a condição UU, referentes a palavra em inglês Unknown).

$$\widehat{UCL} = \hat{\theta} + h\widehat{\sigma_\theta} \quad (20)$$

$$\widehat{LCL} = \hat{\theta} - h\widehat{\sigma_\theta} \quad (21)$$

O procedimento usual para estimar os parâmetros da Fase I é tomar m = 25 amostras de tamanho n = 5 (OPRIME; MENDES, 2017). Onde:

$$\hat{\theta} = \frac{1}{m}\frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n}\theta_{ij} \quad (22)$$

Na Fase I, após calculados os limites de controle, os m dados coletados, calculados de acordo com a Equação 23, são lançados nas cartas.

$$\hat{\theta}_m = \frac{1}{n}\sum_{i=1}^{n}\overline{\theta}_{im} \quad (23)$$

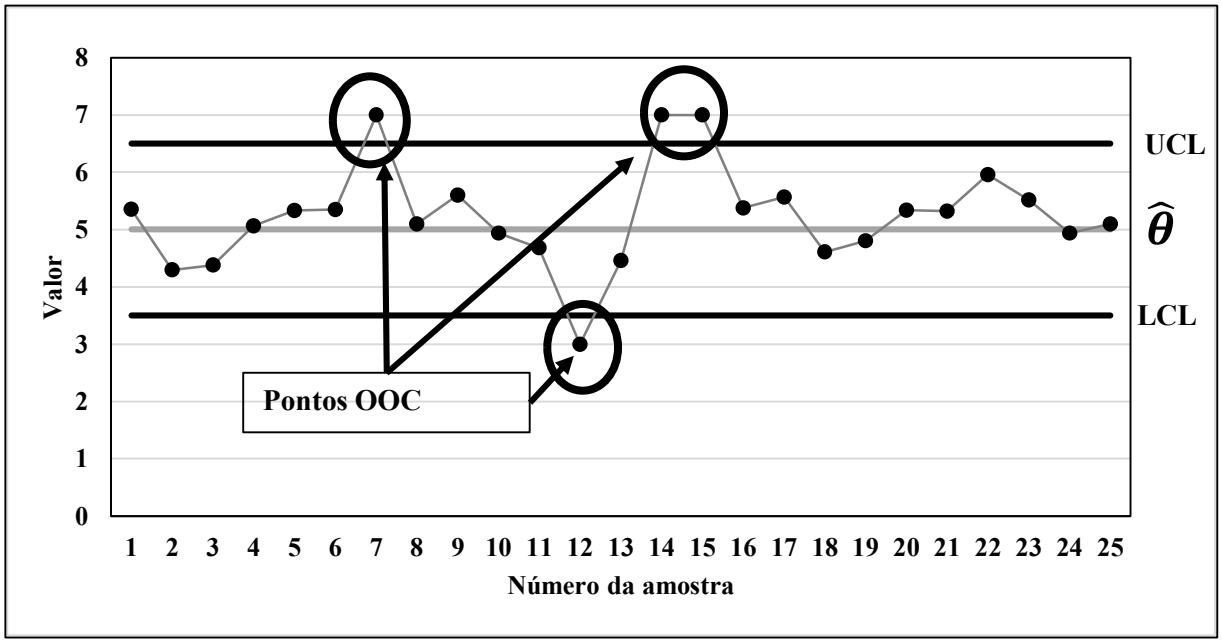
Se o processo estiver sob controle, ou seja, a causa de as variações serem devidas somente a causas comuns, espera-se que todos os dados estejam dentro dos limites de controle. Caso isso não ocorra pode significar processo fora de controle, devendo-se atuar na remoção das causas especiais, até que só existam causas comuns presentes e o processo mantenha-se estável, quando se executa nova coleta de dados para as cartas de controle finais. Pontos fora de controle (em inglês out of control points – OOC) podem ser identificados com o auxílio do Quadro 21. Uma vez o processo sob controle, e a carta definida, inicia-se a Fase II de monitoramento contínuo. Um exemplo de carta de controle é demonstrado na Figura 38.

Quadro 21 - Critérios de análise para Cartas de Controle $\bar{X}$ (X Barra)

Critério	Descrição
C1	1 ponto plotado acima do LSC ou um ponto abaixo do LIC, quando limites calculados com 3 desvios padrão
C2	2 pontos consecutivos entre $\bar{x}$ mais 2 desvios e $\bar{x}$ mais 3 desvios, ou $\bar{x}$ menos 2 desvios e $\bar{x}$ menos 3 desvios
C3	Idem C2 para 2 pontos em 3 pontos consecutivos
C4	3 pontos entre 4 consecutivos cair entre $\bar{x}$ mais 1,6 desvios e $\bar{x}$ mais 3 desvios, ou $\bar{x}$ menos 1,6 desvios e $\bar{x}$ menos 3 desvios
C5	8 pontos consecutivos acima ou abaixo de $\bar{x}$
C6	10 pontos consecutivos acima ou abaixo de $\bar{x}$

Fonte: adaptado de Costa, Epprecht, & Carpinetti (2004)

Figura 38 - Exemplo de Carta de Controle



Fonte: Próprio Autor

A escolha do valor do parâmetro h, das Equações 18 a 21, interfere no desempenho do gráfico, em aplicações práticas é comum o uso de h =3 (MOHAMMADIAN; AMIRI, 2013). Considerando-se os casos em que θ tende a uma distribuição normal, estatisticamente isso representa um intervalo de confiança bilateral teórico, onde α = 0,0027.

Isso porque o SPM é um teste de hipóteses contínuo para avaliar se a hipótese nula é verdadeira, dentro do nível de confiança α:

$$H_0: \widehat{\theta}_m - \widehat{\theta}_0 = 0$$

$$H_1: \widehat{\theta}_m - \widehat{\theta}_0 \neq 0$$

O desempenho esperado dos gráficos reflete a escolha entre os dois tipos de erros inerentes ao SPM (ou ao teste de hipóteses), os erros α e β. O erro α é a taxa de falsos positivos (quando o processo está sob controle, mas o valor encontrado do parâmetro monitorado está fora dos limites) e o erro β é a taxa de falsos negativos (quando o processo está fora de controle, mas o valor encontrado do parâmetro monitorado cai dentro dos limites). Assim, quanto menor o erro α aceito, maior o erro β. Para h = 3, parâmetros conhecidos e processo sob controle, somente 1 ponto em 370 deveria sair dos limites calculados, sendo ele denominado de falso alarme.

A estimativa de parâmetros, menores valores para m e n e o estimador escolhido para o desvio padrão, interferem no desempenho dos gráficos de controle, aumentando a taxa de falso alarme (SOBUE *et al*., 2020; JARDIM; CHAKRABORTI; EPPRECHT, 2019; JARDIM; CHAKRABORTI; EPPRECHT, 2018; JENSEN *et al*., 2006).

A medida mais importante para o cálculo de desempenho dos gráficos de controle é denominada comprimento médio de sequência (em inglês Average Run Length – ARL). Essa medida significa o número de amostras do processo que devem ser avaliadas até a ocorrência de um ponto fora dos limites de controle. Em função dessa característica o tamanho de sequência (em inglês Run Length – RL) é uma variável aleatória que segue uma distribuição geométrica com média 1/p, onde p é a probabilidade de sucesso do evento (ACOSTA-MEJIA, 1999). Em situações de processo sob controle, espera-se que este valor seja o maior possível, evitando paradas desnecessárias. Na Equação 24, p = α , onde α refere-se a probabilidade de um lançamento ocorrer fora dos limites de controle estabelecidos. O valor usual em gráficos de controle é α = 0,0027, assim ARL_0 = 370, para um processo com μ e σ conhecidos.

$$ARL_0 = 1/\alpha \tag{24}$$

Por outro lado, em situações de processo fora de controle, espera-se que esse valor seja o menor possível, identificando rapidamente variações anormais no processo. Nesse caso, o poder do teste é usado para calcular o ARL sendo demonstrado na Equação 25 (HARIDY; WU;

CASTAGLIOLA, 2011; JENSEN *et al*., 2006; JONES-FARMER *et al*., 2017; SOBUE *et al*., 2020; VICENTIN *et al.*, 2018; WOODALL, 1985)

$$ARL_{OOC} = 1/(\text{poder do teste}) = 1/ (1- \beta) \quad (25)$$

É importante ressaltar que na Fase I (fase retrospectiva), o principal interesse é entender melhor o processo e avaliar sua estabilidade. Em contrapartida, a Fase II (fase prospectiva) consiste em tentar controlar um processo através da análise histórica de dados e eliminação de quaisquer causas de variação atribuíveis (CHAKRABORTI; HUMAN; GRAHAM, 2009). A Figura 39 ilustra as etapas para a Fase I considerando-se que a distribuição de probabilidade pode ser conhecida ou não, e que o SPM também pode ser aplicado utilizando conceitos de estatística não paramétrica.

Figura 39 - Etapas Fase I SPM

Passo 1	Identificar ou escolher a distribuição de probabilidade (diga-se F(x)), ou considerá-la desconhecida mas contínua, neste caso isso conduzirá a se considerar um gráfico de controle não paramétrico. Identificar a estrutura de correlação das observações, se independentes e idênticamente distribuídas - i.i.d ou não i.i.d. Ou seja se o resultado de uma medida é dependente do resultado da medida anterior
Passo 2	Decidir sobre as características (parâmetros) que necessitam ser monitorados (medidas de locação como a média ou de variação como o desvio padrão). Decidir sobre as estatísticas calculadas a partir da amostra (e.g. média estimada, amplitude, desvio padrão estimado,etc). Denotar essa estatística por *Ti = h(xi1, ...xin),* onde h é alguma função.
Passo 3	Decidir sobre sobre o tipo de gráfico de controle (e.g. Shewhart, CUSUM* or EWMA**) ou qualquer um dos seus aprimoramentos. Esta decisão dependerá do grau de sensibilidade de detecção à causas especiais e de outras considerações práticas. Exemplo: Assumindo que a estatística escolhida se comporta de acordo com a curva normal, use um gráfico de CUSUM para Ti = X barra para detectar pequenas mudanças na média.
Passo 4	Obter a distribuição do ARL para o gráfico de controle e assim encontrar a constante *h* que atenda uma determinada performance desejada.

* Soma Acumulada (em inglês Cumulative Sum), ** Média móvel exponencialmente ponderada (em inglês Exponentially Weighted Moving Average).
Fonte: Adaptado de Chakraborti e Graham (2019)

Na indústria, é comum controlar a fração de defeitos em um processo produtivo. Neste caso, os parâmetros aplicados aos limites de controle, demonstrados pelas equações 20 e 21 referem-se a esta fração de defeituosos. A literatura indica que a distribuição binomial é aplicável nestes casos e os limites de controle são definidos pelas Equações 26 e 27 (HAGAN;

LI, 2018; ACOSTA-MEJIA, 1999). Para n$p \geq 5$ e n(1-p) ≥ 5, a distribuição binomial pode ser aproximada pela normal, simplificando os cálculos de probabilidade (GRANT, 1964).

$$\widehat{UCL} = \min\{\hat{p} + h\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, 1\} \quad (26)$$

$$\widehat{LCL} = máx\{0,\ \hat{p} - h\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\} \quad (27)$$

Chen e Song (2012) demonstram que o tamanho m e n, das amostras utilizadas durante a Fase I, para a estimativa do parâmetro p ($\hat{p}$), interferem significativamente no valor de ARL. Mesmo para m e n grandes, h =3 e processo sob controle, o ARL não é exatamente 370, oscilando cerca de 10% em torno deste valor para diferentes m e n. Em processos fora de controle, espera-se que $ARL_0 > ARL_{OOC}$, entretanto estes mesmos autores demonstram que $ARL_0 < ARL_{OOC}$ para determinadas situações de p, m e n. Acosta-Mejia (1999) já havia discutido esse viés de desempenho dos gráficos, bem como alternativas para sua redução pela alteração dos limites de controle.

Em geral, quanto maiores p, m e n, o comportamento de ARL real aproxima-se do ARL teórico, entretanto a agregação dos dados, proporcionada por n grande, pode ser um problema, por mascarar variações referentes a causas especiais, que seriam detectados em amostras menores (GOEDHART; WOODALL, 2021).

Demonstrou-se que o objetivo do SPM é a redução da variabilidade dos processos pela redução das causas especiais. O desenvolvimento de sensores, a viabilidade do monitoramento dos parâmetros de entrada, o desenvolvimento da tecnologia da informação e das ferramentas de análise de dados possibilitam melhores configurações de aplicação para a ferramenta e consequentemente redução nos erros α e β. Com o monitoramento das variáveis de entrada o SPM evolui de uma ferramenta de detecção para uma ferramenta de diagnóstico, e melhores sensores e atuadores resultam na transformação viável de causas comuns em causas especiais, passíveis de remoção, reduzindo a variabilidade do processo. Essas considerações serão apresentadas em maiores detalhes na seção seguinte.

6.2 NOVAS APLICAÇÕES DO SPM

No contexto tradicional a aplicação do SPM: não ocorre em tempo real; poucas variáveis são consideradas nos estudos; a velocidade de resposta na análise de novos processos é lenta;

problemas mais complexos demandam o uso de "especialistas"; a maioria das aplicações monitora características de saída, assim, o produto não conforme, que está entre o início do processo e o momento da detecção da falha, resulta em retrabalho ou perda; o método tradicional é reativo pois "emite o alarme" somente após a ocorrência e detecção da falha; cada carta de controle criada implica em aumento da carga de trabalho do operador e procedimentos demandantes de grande disciplina em sua execução (lançamentos manuais dos dados nas cartas), fatores que dificultam sua implantação (PARA *et al*., 2019)

Em contraponto a esse contexto, as aplicações que se utilizam de modelos preditivos, apresentadas na seção 4.2, e os elementos da Indústria 4.0, aumentam a disponibilidade de dados e criam condições em que ambientes reativos podem se tornar ambientes preditivos/preventivos que integrariam as 3 dimensões: Detecção, Diagnóstico e Prognóstico de Falhas, resultando em melhor desempenho do SPM (O'DONOVAN *et al*. 2015b; O'DONOVAN *et al*., 2015a; REIS E GINS 2017).

Desta maneira, o desenvolvimento das tecnologias de informação facilita a aplicação do SPC tanto para características finais de um processo, quanto para seus recursos de entrada. Com isso o uso do termo Monitoramento Estatístico do Processo (em inglês Statistical Process Monitoring - SPM) torna-se mais aplicado. Outras definições também são discutidas, de acordo com Jones-Farmer e Stevens (2017), a nomenclatura utilizada poderia ser Monitoramento Estatístico (em inglês - Statistical Monitoring - SM), uma vez que existem aplicações onde os dados monitorados não são resultado de um processo, *e.g*. quantidade de posts em mídias sociais.

Em publicações da indústria química e de processos utilizam-se os termos Monitoramento de Processos (em inglês Process Monitoring - PM) e Monitoramento de Processos Industriais (em inglês Industrial Process Monitoring - IPM) pois muitos parâmetros de entrada são controlados (temperatura, pressão, velocidade dos fluídos nas tubulações, tempos de reação), de forma a garantir que as causas especiais não ocorram, ou sejam detectadas em seus estágios iniciais pelas variações desses parâmetros (KAMBLE; GUNASEKARAN; GAWANKAR, 2018; REIS e GINS, 2017)

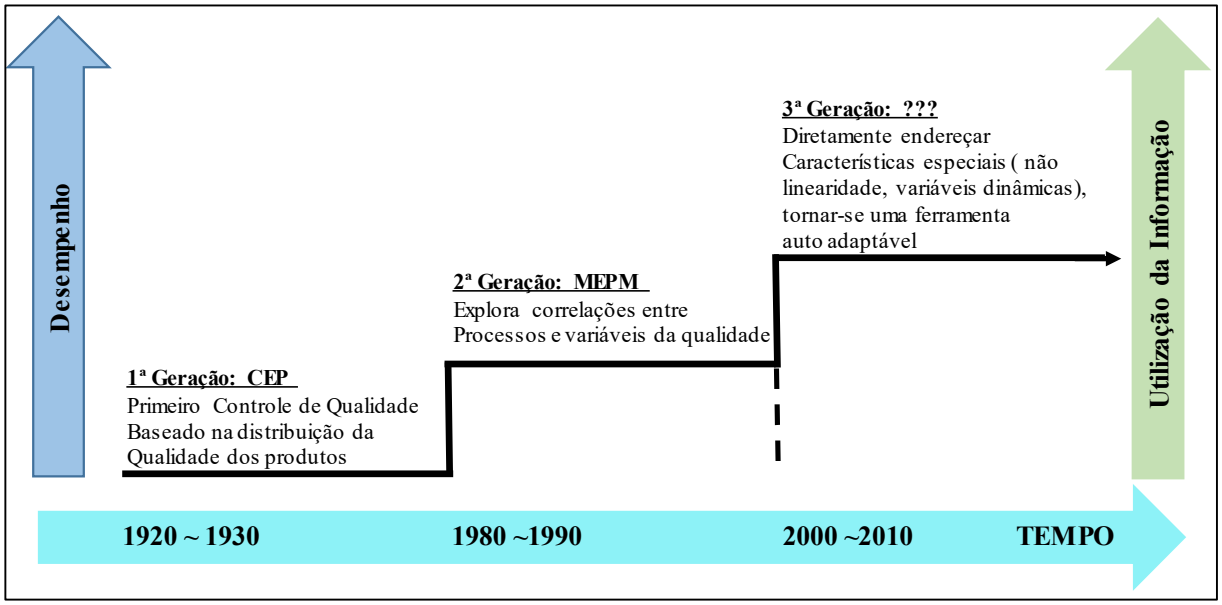
He e Wang (2018a) analisaram o desenvolvimento do SPM a partir de seus primórdios em 1920, classificando-o em 3 Gerações. A primeira Geração é caracterizada pelo uso dos gráficos de controle (SHEWHART, 1931). Denominado SPC, o método foi desenvolvido em uma época que recursos computacionais não existiam no chão de fábrica, em consequência deste contexto, formas de simplificar sua utilização pelos operadores eram consideradas. Os

gráficos baseavam-se nas premissas que os dados monitorados pertenciam à uma Gaussiana, eram independentes e identicamente distribuídos (*i.i.d*) e aplicavam-se ao produto de um processo.

Para Escobar *et al.* (2018) e Chang (2017), o monitoramento das variáveis, ou características de qualidade dos produtos em cada processo, carece de modelos que relacionem um determinado efeito à uma determinada causa. Gap preenchido pela 2ª geração de SPM que se caracteriza pelo desenvolvimento de ferramentas de monitoramento das variáveis de qualidade e das variáveis de processo, através do uso de análise multivariada. He *et al*. (2017), denominaram a 2ª Geração como Monitoramento Estatístico do Processo Multivariado (em inglês Multivariate Statistical Process Monitoring - MSPM). Assim houve evolução dos objetivos do SPM, de detecção de variações no produto, para identificação de causas raiz de mudanças, que impliquem em desvios de alvos do processo (HE; WANG, 2018b).

He e Wang (2018a) não definiram como denominar a 3ª Geração do SPM, mas afirmam que problemas como dados pertencentes a distribuições multimodais ou não Gaussianas, dados dinâmicos, variáveis correlacionadas não linearmente, características que variam com o tempo, outliers, erros grosseiros e falhas de leituras de sensores, deverão ser endereçados pelos novos modelos de SPM. Assim conclui-se que a garantia da qualidade dos dados de entrada tem sua importância alavancada no contexto do SPM. A Figura 40 demonstra a evolução do SPM.

Figura 40 - Evolução do Monitoramento Estatístico de Processo



Fonte: He e Wang, (2018a)

Ao buscar conexões entre SPM e Indústria 4.0, é possível encontrar interações entre

esses dois campos. Por exemplo, He e Wang (2018), afirmam que um importante foco da Indústria 4.0, além da geração dos dados, é sua análise em tempo real, permitindo decisões operacionais melhores e mais ágeis, e, Chang (2017), elenca oportunidades do SPM no contexto do Big Data, apresentadas no Quadro 22.

Quadro 22 - Implementação Clássica do SPM versus Oportunidades do Big Data

Funções do SPM	**Implementação Clássica do SPM**	**Oportunidades no Big Data**
Disponibilidade de dados	Escassos	Abundantes
Tipos de dados	Restrito à números	Expandido para voz e imagem em adição à números e Textos
Velocidade de coleta de dados	Lento	Rápido como no Torrent
Frequência de coleta de dados	Infrequente	Tempo real ou próximo a tempo real
Análise dos dados	Por amostragem	Usando todas as observações
Dimensão dos dados	Pequena (menor que 10)	Grande (centenas ou milhares)
Arquivo de dados	Passado Recente	Todo o histórico de observações
Visualização dos dados	Gráficos de controle convencionais	Espacial, temporal, multi-escalas
Assunto dos dados	produto	tanto paramêtros de processo quanto produtos

Fonte: Chang, (2017)

Alinhados às oportunidades apresentadas, alguns autores propõem novas alternativas para o SPM, entre elas o uso de algoritmos de aprendizado estatístico (popularmente denominados aprendizado de máquina) (WEESE *et al*.2016). Neste contexto propõem que, além dos métodos de monitoramento baseados em distribuições de probabilidade conhecidas e regras físicas (baseado em conhecimento), sejam aplicados métodos baseados em dados. No caso destes, como já descrito na seção 4.3, utilizam-se modelos e regras para detectar mudanças significativas a partir de algoritmos de aprendizado de máquinas baseados em dados, sem necessariamente existir conhecimento prévio das relações físicas entre variáveis de entrada e saída. Podendo serem classificados como Supervisionados, quando se relaciona um conjunto de dados a um rótulo (ou variável de saída), não Supervisionados, onde os dados são analisados na busca de padrões entre eles, ou por Reforço, referentes a algoritmos que “aprendem” por tentativa e erro. Dentro dessa abordagem, destaca-se o trabalho de Hryniewicz (2015), em que

se estuda a aplicação de diferentes algoritmos com o objetivo de predizer o resultado (se está conforme ou não conforme) de uma determinada variável indireta (vida útil de um equipamento por exemplo). Neste trabalho, a partir de diferentes distribuições de probabilidade, simulam-se valores para as variáveis de entrada e vários modelos criados têm sua qualidade medida pela probabilidade de se classificar erroneamente um determinado conjunto de instâncias de dados.

Segundo Severson, Chaiwatanodom e Braatz (2016), na indústria de processos, o uso dos modelos preditivos possibilitará em tempo real, a eliminação, mitigação e diagnóstico (localização) das falhas. Neste setor, os sistemas de controle dos processos, em geral, são capazes de corrigir o que os autores chamam de perturbações (alterações desconhecidas e não controláveis nas variáveis de entrada ou causas comuns), entretanto precisam ser complementados por sistemas que se antecipem à: *i)* falhas causadoras de desvios não permitidos em pelo menos uma propriedade ou parâmetro característico do sistema em suas condições de operação padrão (o que é definido como fault pelo autor); *ii)* interrupções permanentes da capacidade de um sistema de executar uma função exigida sob condições operacionais especificadas (o que é denominado como failure pelo autor). Qin (2012) afirma que estes sistemas também devem possibilitar que métodos de reconstrução de falhas estimem a magnitude destes eventos e as condições livres das falhas.

Comum em todas as aplicações citadas, o uso de modelos preditivos para o monitoramento e controle dos processos reflete a sua relevância para a indústria, uma vez demonstrado que os controles automáticos são uma realidade cada vez mais presente.

Em relação ao uso de modelos preditivos, Woodall e Montgomery (2014), Woodall (2016) e Jones-Farmer e Stevens (2016) afirmam que, apesar do desenvolvimento de métodos de aprendizagem estatística, os trabalhos nessa área não endereçam adequadamente o desenvolvimento da Fase I do SPM, e não orientam suficientemente sobre como os aplicar na prática ou como estabelecer as amostras, sob controle, de referência. Outro ponto de atenção é discutido por Steinberg (2016), o autor afirma que a atenção dedicada, pelos estatísticos, ao assunto qualidade dos dados não é condizente com a relevância do tema, com maiores carências no campo da indústria.

As aplicações demonstradas na seção 4.2 e as afirmações a respeito da importância da melhoria das ferramentas de diagnóstico permitem concluir que existem elementos no CPS que possibilitam o uso de modelos preditivos e que estes podem ser utilizados em conjunto, ou como apoio, ao SPM e vice-versa.

Como a DS está diretamente relacionada a criação de modelos preditivos, a integração dos conceitos de SPM à DS como ferramenta de validação é um assunto a ser investigado. Incluem-se neste escopo tanto os modelos criados pelos processos de geração do conhecimento, como todo o processo decisório onde o modelo estará inserido. Essa estrutura de integração apresenta-se na próxima seção.

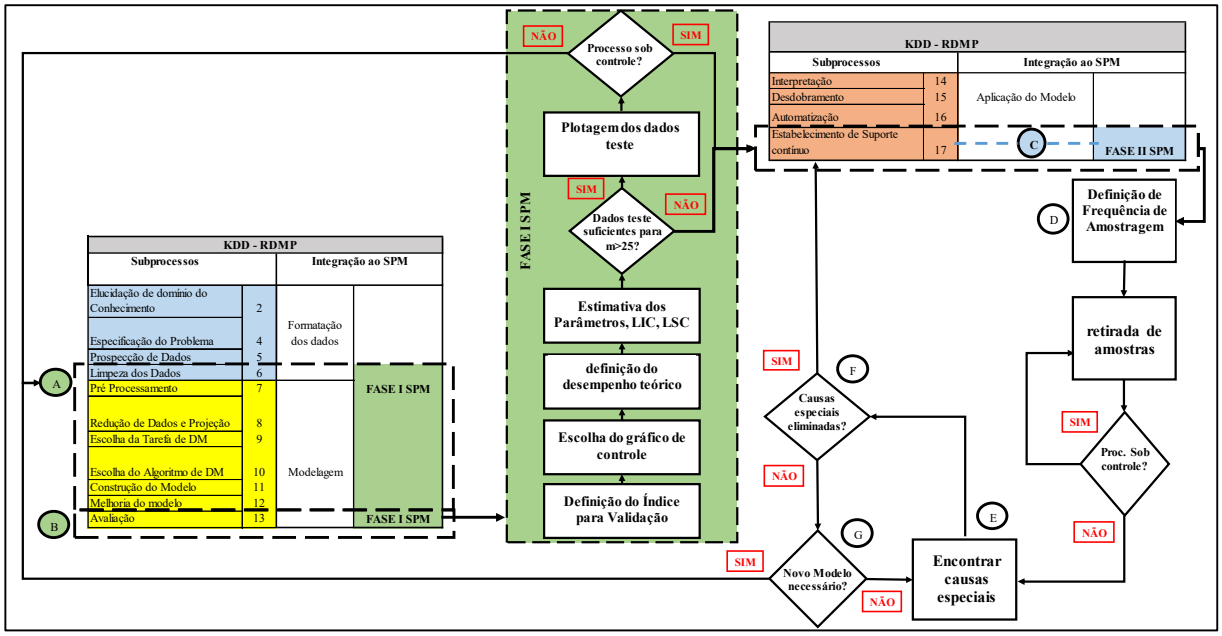
7 INTEGRAÇÃO DO SPM AOS MODELOS DE GERAÇÃO DE CONHECIMENTO

Nas seções a seguir utiliza-se o background apresentado para desenvolver uma estrutura de relacionamento entre os elementos dos processos de geração de conhecimento, parte fundamental da DS, e os elementos do SPM. O objetivo é utilizar seus fundamentos, como o monitoramento periódico da estabilidade de fenômenos, no preenchimento das lacunas anteriormente apresentadas. Para tanto desenvolve-se uma abordagem em que esses vários elementos são integrados. Propõem-se sua validação através do desenvolvimento de modelos analíticos de desempenho para gráficos de controle baseados no *Kappa* e de modelos de simulação computacional, para representar a realidade, que visam validar os resultados teóricos encontrados.

7.1 PROPOSTA DE ABORDAGEM DE INTEGRAÇÃO ENTRE O SPM E A CIÊNCIA DOS DADOS

Tomando-se como referência o processo de KDD denominado RDMP, resultado da comparação de vários processos de geração de conhecimento e detalhado no Quadro 15; considerando-se os dados apresentados na seção 5.1 em que demonstra-se a carência de procedimentos para verificação contínua dos resultados de um modelo; considerando-se as proposições e conclusões ao final das seções 3.4, 3.5, e 4.3 à respeito da crescente complexidade existente no gerenciamento dos dados, equipamentos e pessoas; considerando-se que os processos de KDD são elementos da DS; considerando-se que o SPM é uma ferramenta que avalia continuamente a estabilidade de um processo, demonstra-se na Figura 41 uma proposta de integração entre DS e SPM que fornece critérios para se validar continuamente a qualidade preditiva de processos decisórios baseados em modelos classificatórios binários, quaisquer que sejam os modelos aplicados. Aqui cabe reforçar que nesse tipo de processo decisório, além da capacidade preditiva do modelo (qualidade do modelo), outros elementos interferem na qualidade das previsões, e consequentemente na qualidade das decisões.

Figura 41 - Abordagem de Integração do SPM a DS



Fonte: Próprio autor

O ponto A, Figura 41, indica a primeira convergência entre os conceitos discutidos. Por exemplo, para a elaboração das cartas de controle de um processo industrial, o procedimento inicial é garantir que as máquinas e equipamentos estejam com a manutenção em dia, corretamente regulados, os operadores treinados, a matéria prima utilizada dentro das especificações e os instrumentos de medição calibrados, ou seja, as causas especiais de variação do processo precisam ser eliminadas, garantindo que este esteja sob controle no momento da coleta inicial dos dados. A eliminação de outliers, correção de dados faltantes, padronização de formatos de dados, seleção das variáveis relevantes para a elaboração do modelo e a avaliação de vários algoritmos de DM, presentes nas tarefas de 6 a 12, agrupadas no ponto A e presentes no processo de RDMP (Quadro 15), são ações que eliminam causas especiais no processo de criação de um modelo, e, sendo assim, são diretamente relacionadas ao SPM em sua Fase I.

O ponto B indica outra convergência em que os conceitos de SPM podem ser aplicados. Nesta etapa escolhe-se um índice para avaliar a qualidade de previsão, conforme discutido em 5.1 e 5.2. Independentemente do indicador escolhido, e como demonstrado na seção 5.3, o valor do índice calculado é uma variável aleatória, pois estima o desempenho a partir de uma amostra dos resultados de predição. Em casos de modelos de aplicação pontual, ou seja, aqueles que serão utilizados para uma única decisão, essa aleatoriedade tem interferência reduzida no processo decisório. Entretanto, para modelos preditivos de uso contínuo, sugere-se que os processos de validação considerem essa aleatoriedade, tanto em etapas iniciais, em que se

escolhe qual técnica é mais adequada para a elaboração do modelo de previsão, quanto para a avaliação da necessidade de manutenção do modelo escolhido. O SPM, por se tratar de uma técnica de revisão periódica da variabilidade de parâmetros, converge com essa necessidade.

Neste ponto sugere-se aplicar as etapas da Fase I do SPM, conforme demonstrado no Quadro 21. Neste caso, a característica da qualidade monitorada será o índice que demonstre a capacidade preditiva do modelo, compatível com o seu contexto de aplicação. Exemplos destes índices foram apresentados nas seções 5.2 e 5.3 e a avaliação da quantidade (m) e do tamanho das amostras (n) foi discutido na seção 6.1. Para uma quantidade de dados segregados suficiente para a elaboração de m amostras de tamanho n (nas próximas seções será demonstrado que esse valor depende diretamente da quantidade de sucessos da população), elabora-se a carta inicial de controle e verifica-se se os m valores calculados do índice estão dentro dos Limites. Caso não estejam, é necessário retornar ao ponto A e analisar novamente as etapas indicadas do KDD - RDMP para se eliminar a, ou as, causas especiais.

Se os valores calculados dos índices estiverem dentro dos limites entende-se que o processo está sob controle, e, após a execução das etapas de 14 a 16, inicia-se a etapa 17 de suporte contínuo ao modelo, ponto C de convergência com o SPM, nessa etapa sugere-se aplicar a Fase II de monitoramento, para avaliar continuamente se não surgem causas especiais no processo decisório.

Considerando-se que as etapas de 6 a 12 do KDD-RDMP foram adequadamente executadas, a eliminação de causas especiais está englobada nessas etapas. Aqui cabe a proposta de uma regra prática, se a quantidade de dados, segregada para essa validação inicial, não for grande o suficiente para a elaboração da quantidade de amostras (m) necessárias, sugere-se elaborar os gráficos de controle com os dados disponíveis, e iniciar a fase de monitoramento (C). Pois uma vez iniciado o monitoramento e obtidas a quantidade de amostras necessárias, o gráfico pode ser recalculado adequadamente, evitando uma taxa de alarme falso além da esperada para os limites de controle definidos.

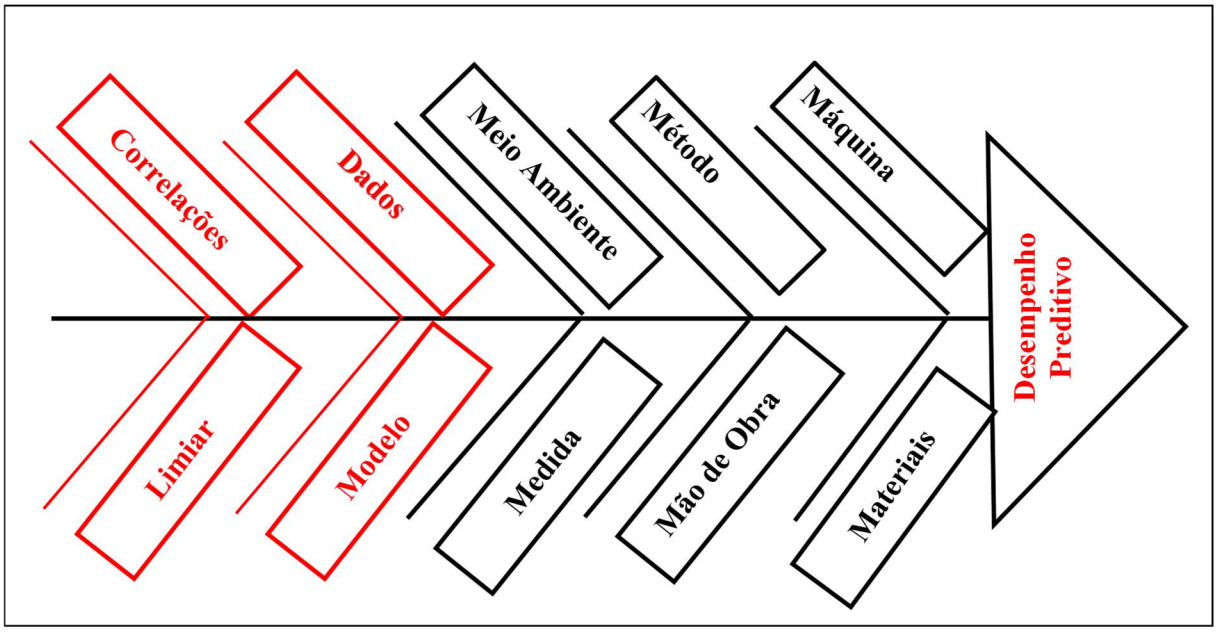
As proposições elaboradas ao final das seções 3.4, 3.5, 4.2 e 4.3, demonstram os vários elementos que podem interferir na qualidade dos processos preditivos e assim demonstram a importância do monitoramento contínuo desse desempenho, principalmente quando os modelos apoiam decisões, em regime automático e em tempo real. Isto posto, a Fase II do SPM, conforme indicado no ponto C na Figura 41, é uma necessidade e um elemento que complementa a etapa de suporte contínuo. O uso das cartas de controle indica precisamente o

momento no tempo quando uma intervenção pode se fazer necessária. A escolha do tipo de carta de controle tem relação ao grau de sensibilidade necessário em cada contexto de aplicação, e conforme abordado previamente, as cartas de CUSUM e EWMA são mais sensíveis, já as cartas tradicionais de Shewhart são mais simples e menos sensíveis.

No ponto D indica-se a necessidade de se definir a frequência da retirada das amostras para a plotagem na carta de controle. Essa frequência não será discutida por ser uma questão particular de cada aplicação, ou seja, depende de cada contexto. Obviamente, quanto mais alta a frequência de amostragem, mais rápido irá se detectar uma possível ocorrência de causa especial. Em aplicações onde a geração de dados é ininterrupta e o processo de geração de amostras e comparação aos limites de controle é automatizado (exemplo são os CPS's), as frequências viáveis podem estar na casa dos minutos. Em compensação, em aplicações onde a quantidade de dados é menor, ou é mais difícil a elaboração de bases de dados rotuladas (para o cálculo dos índices), estas frequências podem ser menores (diárias, semanais, mensais), em contrapartida a detecção também será mais lenta. Nas aplicações reais, o uso da Fase II do SPM possibilita um melhor entendimento do processo, e as frequências de amostragem podem ser alteradas em função de sua condição, reduzindo-se a frequência para processos que se demonstram mais estáveis e aumentando-se para processos menos confiáveis.

Os pontos E e F representam a análise de causas especiais, caso o processo fique fora de controle, ou seja, a estimativa do indicador monitorado apresente variabilidade superior à sua variabilidade padrão. Conforme apresentado em 5.1 a literatura usualmente aponta para o modelo utilizado como a principal causa de variação nessa estimativa. Essa análise nem sempre é correta, uma vez que em contextos de aplicação cada vez mais dinâmicos, restringir a causa de resultados preditivos insatisfatórios, somente à ferramenta de modelagem utilizada, desconsidera a ação de outros fatores envolvidos nesse processo, e que são relacionados no Diagrama de Ishikawa adaptado à análise de dados, apresentado na Figura 42. Neste diagrama, adicionam-se aspectos referentes a qualidade dos dados, escolha do Limiar, correlações entre as variáveis de entrada e o modelo em si, aos elementos tradicionais da análise de Ishikawa: problemas nas matérias primas (troca de fornecedores, erros de entrega, lotes defeituosos); falhas nas etapas manuais de coleta de dados (problemas com mão de Obra e procedimentos); equipamentos de medida descalibrados (erros de medidas); defeitos em equipamentos e alterações de condições ambientais.

Figura 42 - Diagrama de Ishikawa adaptado a análise de modelos



Fonte: Próprio autor

Com relação ao fator qualidade dos dados (Dados), este é descrito no Quadro 6 e no Quadro 23, e tem relação com a etapa de limpeza de dados executada no processo de KDD – RDMP (ponto A). Uma vez o modelo em uso, problemas na automação de troca de dados, atualização de sistemas, alterações de arquitetura de dados, interfaces para novos equipamentos, falhas de capacidade de armazenamento, incompatibilidade entre dispositivos, "bugs" de sistema, podem gerar dados incorretos que comprometam seu desempenho preditivo.

Quadro 23 - Dimensões de Qualidade dos Dados

Dimensão da Qualidade dos dados	**Descrição**
Acuracidade	Os dados estão livres de erro?
Atualidade	Os dados estão atualizados?
Consistência	Os dados estão no mesmo formato?
Completude	Estão faltando dados necessários?

Fonte: Hazen *et al*. (2014)

O fator limiar, definido na seção 5.2, tem relação com o ajuste inicial do modelo, contudo, melhorias no desempenho preditivo podem ser alcançadas pela redefinição de seus

valores, ajustando os resultados em função da relevância relativa da taxa de revocação e especificidade em cada situação específica, sem a necessidade da criação de outro modelo. Aqui cabe retomar que em algumas situações é mais viável economicamente encontrar os verdadeiros positivos, mesmo que a taxa de falsos positivos aumente, do que encontrar os verdadeiros negativos (ou vice – versa). Fazendo um paralelo com os testes de covid, é muito importante encontrar quem realmente tem covid (verdadeiros positivos), mesmo que alguns considerados positivos na realidade não tenham a doença (falsos positivos).

O fator correlação entre as variáveis de entrada, pode ser exemplificado pela situação em que o comportamento de um mesmo perfil de consumidor muda, ao longo do tempo, devido alterações nos padrões sociais, e.g. nas décadas de 1970 a 1990 fumar era uma demonstração de refinamento e status, nas décadas seguintes torna-se um vício grave. Neste caso, buscar novos modelos, ou calibrar o modelo utilizado com base nos dados iniciais, não irá resolver o problema, pois houve mudança nas relações entre as variáveis, não refletidas nos dados iniciais. Assim o modelo escolhido previamente pode ser reutilizado, mas precisa ser calibrado com dados mais atuais, que carreguem as alterações de relação entre as variáveis consideradas.

Isto posto, a revisão completa do modelo, indicado no ponto G, aplica-se após todos os fatores de variabilidade, indicados no Diagrama de Ishikawa adaptado, tiverem sido analisados, retornando-se ao ponto A. Com a coleta de dados recentes, recalibra-se ou reestuda-se o modelo, reavaliando-se a necessidade de se rever a Fase I do SPM. Cabe aqui as considerações anteriores sobre os tamanhos de amostra, e a continuidade da Fase II do SPM, sem necessariamente a revisão completa da Fase I.

Conclui-se que a abordagem proposta trata o modelo criado como parte integrante de um processo de tomada de decisão, relacionando-o aos vários fatores que causam variabilidade na qualidade de seu resultado preditivo, e, consequentemente na qualidade das decisões. Essa integração possibilita a elaboração de critérios precisos e replicáveis que apoiam a execução da do suporte contínuo. Os registros históricos de amostragem podem ser utilizados: na checagem de eliminação das causas especiais; no teste de alterações do modelo (para busca de melhorias), mesmo que causas especiais não estejam presentes; para comparação de modelos alternativos.

Na próxima seção são elaborados cálculos analíticos que visam auxiliar a elaboração de critérios quantitativos para a amostragem e uso do *Kappa* como elemento de validação contínua dos modelos apresentados.

7.2 MODELO ANALÍTICO DE DESEMPENHO PARA *KAPPA*

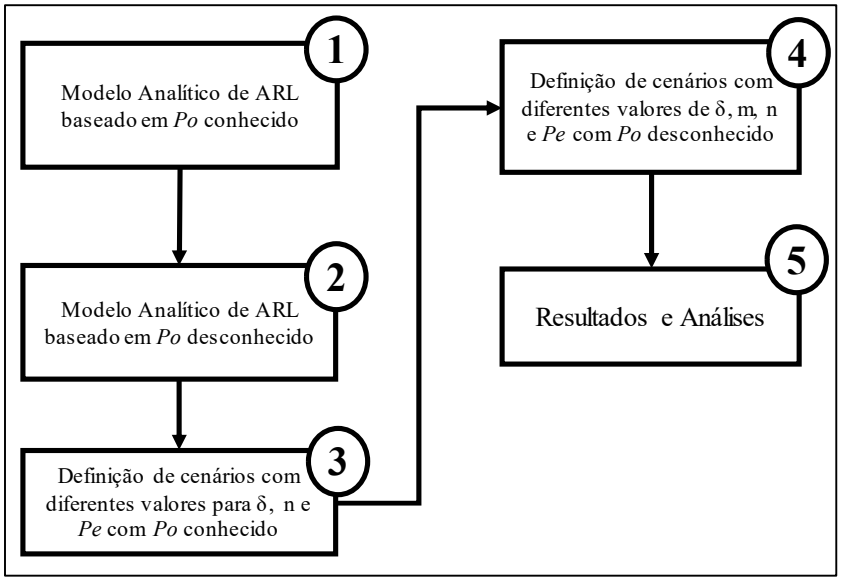
Baseando-se na teoria referente ao SPM, nesta seção desenvolvem-se modelos analíticos de desempenho para gráficos baseados em *Kappa* considerando-se duas situações para os valores de P_0, conhecido e desconhecido (lembrando que este parâmetro indica a proporção de concordâncias de avaliação, detalhadas na seção 5.3.1). Consideram-se como pontos OOC, somente aqueles cujo valor resulta fora dos limites de controle calculados. Isto deve-se à dificuldade matemática de se considerar todas as possibilidades de pontos OOC indicadas no Quadro 20.

7.2.1 Procedimento de Análise

Para aplicar os conceitos de SPM usando o *Kappa,* são necessárias algumas considerações. Apesar de sua não dimensionalidade, o valor *Kappa* está relacionado à fração P_0 (Equação 7). Ao plotar os valores de *Kappa* em um gráfico de controle, cujos limites são definidos na Fase I do SPM, isso representa a aplicação de testes de hipóteses ao longo do tempo. A hipótese nula $H_0: K = K_0$ significa que o processo está sob controle e não há variações significativas do valor *Kappa*. Caso contrário, a hipótese alternativa $H_1: K \neq K_0$ significa que o processo está fora de controle e existem causas especiais responsáveis por variações significativas nos valores de *Kappa*. Sugere-se esse mecanismo de teste de hipóteses estatísticas em intervalos de tempo como critério de aceitação contínua do modelo, criando elementos objetivos para identificar o momento em que causas especiais atuam no processo preditivo.

Aplicam-se os passos ilustrados na Figura 43 para validar a proposta de monitoramento do *Kappa* no contexto de modelos classificatórios preditivos binários. Inicialmente (passos um e dois), com base nos conceitos estatísticos do SPM, desenvolve-se um modelo analítico para se calcular os valores de ARL esperados das cartas de controle de *Kappa*, nas condições de P_0 conhecido e P_0 desconhecido (estimado).

Figura 43 - Etapas para Análise de Desempenho dos Gráficos de *Kappa*



Fonte: Próprio Autor

Para ambas as condições de P_0 (conhecido e desconhecido), o desempenho dos gráficos de controle de *Kappa* é calculado em função de diferentes níveis das variáveis definidas no Quadro 24, passos três e quatro.

Quadro 24 - Variáveis utilizadas no cálculo analítico *de Kappa*

Variável	Descrição
n	Tamanho da amostra utilizado para o cálculo de *Kappa*
m	Quantidade de amostras utilizadas no cálculo de *Kappa*
δ	distância dos valores *Kappa* da sua média
Pe	Probabilidade esperada, conforme Equação 7 e Quadro 18

Fonte: Próprio Autor

7.2.2 Desempenho Teórico dos gráficos de controle de *Kappa* com P_0 conhecido

O desempenho de um gráfico de controle na Fase II é avaliado pelo erro tipo I ou taxa de alarme falso (FAR) e pelo erro tipo II. O erro tipo I é indicado por α e o erro tipo II por β. É possível mostrar que a esperança matemática de *Kappa*, estimada de tempos em tempos, $\widehat{K}$, é o K_0 da Equação 12, sendo $\hat{P}_0$ uma estimativa de P_0 conforme indicado pela Equação 28.

$$E(\widehat{K}) = E\left(\frac{\hat{P}_0 - P_e}{1 - P_e}\right) = \frac{P_0 - P_e}{1 - P_e} \tag{28}$$

Similarmente, a variância de $\widehat{K}, V(\widehat{K})$, é facilmente obtida pela Equação 29, onde $V(\hat{P}_0) = \frac{P_0(1-P_0)}{n}$ e $V(P_e) = 0$, simplificando-se P_e como uma constante.

$$V(\widehat{K}) = V\left(\frac{\hat{P}_0 - P_e}{1-P_e}\right) = \left(\frac{1}{1-P_e}\right)^2 * V\left(\frac{\hat{P}_0 + Pe}{1}\right) = \frac{P_0(1-P_0)}{n(1-P_e)^2} \quad (29)$$

Para n suficientemente grande, obtém-se o erro tipo I, e o desempenho do gráfico Kappa, medido pelo comprimento de sequência (em inglês Run Length - RL), é calculado usando o FAR. Para simplificar, assume-se que o parâmetro P_0 é conhecido, desta forma, α e ARL são indicados pelas Equações 30 e 31.

$$\alpha = 1 - P\left(LIC \leq \widehat{K} \leq LSC\right) = 1 - P\left(K_0 - Z_{\alpha/2}\sigma_k \leq \widehat{K} \leq K_0 + Z_{\alpha/2}\sigma_k\right)$$

Fazendo o desenvolvimento correto, temos:

$$\alpha = 1 - \left[\theta\left(Z_{\alpha/2}\right) - \theta\left(-Z_{\alpha/2}\right)\right] \quad (30)$$

Seja θ a função cumulativa da distribuição normal padrão e α o erro nominal do tipo I, quando o modelo está discriminando adequadamente o fenômeno, em outros termos, o processo está sob controle (em inglês in control - IC), o desempenho do gráfico Kappa, medido por ARL_0 (comprimento médio de sequência, em função de FAR), é calculado por:

$$ARL_0 = \frac{1}{FAR} = \frac{1}{1-\theta(Z_{\alpha/2})+\theta(-Z_{\alpha/2})} \quad (31)$$

O erro tipo II ocorre quando o modelo não classifica adequadamente o fenômeno. É obtido de forma semelhante ao erro tipo I, acrescentando o fato que $K = K_0 + \delta$. Se $\delta \neq 0$, o modelo é inadequado. Cabe aqui uma observação, se $\delta > 0$, as causas especiais estão atuando para melhorar a capacidade preditiva do modelo, devendo ser estudadas para se manter o nível de desempenho preditivo alcançado. Assim, fazendo o desenvolvimento adequado e, considerando que o desvio padrão de *Kappa* é obtido por $\sqrt{\frac{P_0(1-P_0)}{n(1-P_e)^2}}$, chegamos ao cálculo do erro tipo II, β, indicado na Equação 32. Para:

$\left(LIC \leq \widehat{K} \leq LSC\right) = P\left(\frac{K_0 - Z_{\alpha/2}\sigma_k - K_0 - \delta}{\sigma_k} \leq \frac{\widehat{K} - K}{\sigma_k} \leq \frac{K_0 + Z_{\alpha/2}\sigma_k - K_0 - \delta}{\sigma_k}\right)$, segue que:

$$P\left(\frac{-Z_{\alpha/2}\sigma_k - \delta}{\sigma_k} \leq Z \leq \frac{+Z_{\alpha/2}\sigma_k - \delta}{\sigma_k}\right) = P\left(-Z_{\alpha/2} - \frac{\delta}{\sqrt{\frac{P_0(1-P_0)}{n(1-P_e)^2}}} \leq Z \leq +Z_{\alpha/2} - \frac{\delta}{\sqrt{\frac{P_0(1-P_0)}{n(1-P_e)^2}}}\right)$$

Depois verifica-se que:

$$\beta = P\left(-z_{\alpha/2} - \frac{\delta\sqrt{n}\,(1-Pe)}{\sqrt{Po(1-Po)}} \leq Z \leq z_{\alpha/2} - \frac{\delta\sqrt{n}\,(1-Pe)}{\sqrt{Po(1-Po)}}\right)$$

Sendo conservador, aplica-se o valor de 0,5 para P_0 (já que é o ponto de maior variância segundo a binomial), o que implica:

$$\left(\frac{1}{\sqrt{Po(1-Po)}}\right) = 2$$

As Equações 32 e 33 consideram esta condição específica. O número médio de amostras de *Kappa*, até a detecção de um sinal, obtido pela Equação 33, denomina-se ARL, uma medida de desempenho frequentemente utilizada.

$$P\big(LIC \leq \widehat{K} \leq LSC\big) = \theta\big(z_{\alpha/2} - 2\delta\sqrt{n}(1-Pe)\big) - \theta\big(-z_{\alpha/2} - 2\delta\sqrt{n}(1-Pe)\big) \quad (32)$$

$$ARL = \frac{1}{1-P(LIC\leq\widehat{K}\leq LSC)} = \frac{1}{1-\theta\left(Z_{\alpha/2}-2\delta\sqrt{n}(1-Pe)\right)+\theta\left(-Z_{\alpha/2}-2\delta\sqrt{n}(1-Pe)\right)} \quad (33)$$

7.2.3 Desempenho Teórico dos gráficos de controle de *Kappa* com P_0 desconhecido

Para os casos em que P_0 é estimado no cálculo de *Kappa* e dos limites de controle, *Kappa* é obtido pela expressão $\overline{K} = \frac{1}{m}\sum_{j=1}^{m} K_j$, onde m é a quantidade de *Kappas* individuais calculados. Para os casos em que m = 1, um único *Kappa*, baseado em amostra de tamanho n, é usado para estimar os limites de controle, sendo uma situação mais realista.

Considerando que os limites de controle são estimados. Então:

$$P\big(\widehat{LIC} \leq \widehat{K} \leq \widehat{LSC}\big) = P\big(\widehat{K}_0 - Z_{\alpha/2}\sigma_k \leq \widehat{K} \leq \widehat{K}_0 + Z_{\alpha/2}\sigma_k\big)$$

Assim, desenvolver esta expressão para $K = K_0 + \delta$, implica a seguinte equação:

$$P\big(\widehat{LIC} \leq \widehat{K} \leq \widehat{LSC}\big) = P\left(\frac{\widehat{K}_0 - K_0}{\sigma_k} - Z_{\alpha/2} - \delta\frac{\sqrt{n}\,(1-Pe)}{\sqrt{Po(1-Po)}} \leq Z \leq \frac{\widehat{K}_0 - K_0}{\sigma_k} + Z_{\alpha/2} - \delta\frac{\sqrt{n}\,(1-Pe)}{\sqrt{Po(1-Po)}}\right)$$

Por exemplo, da mesma forma que na Equação 33, consideramos a relação $\left(\frac{1}{\sqrt{Po(1-Po)}}\right) = 2$, e assumindo que $W\sigma_k = \big(\widehat{K}_0 - K_0\big)\sqrt{m}$, com a distribuição de W~N(0,1), é possível chegar à Equação 34 para $P\big(\widehat{LIC} \leq \widehat{K} \leq \widehat{LSC}\big)$:

$$\beta = P\big(\widehat{LIC} \leq \widehat{K} \leq \widehat{LSC}\big) = \theta\left(\frac{W}{\sqrt{m}} + z_{\alpha/2} - 2\delta\sqrt{n}(1-Pe)\right) - \theta\left(\frac{W}{\sqrt{m}} - z_{\alpha/2} - 2\delta\sqrt{n(1-pe)}\right) \quad (34)$$

O RL está condicionado à estimativa de *Kappa*, utilizada no cálculo dos limites de controle, com probabilidade $p\big(W, Z_{\alpha/2}, \delta, m, Pe\big)$. Para eliminar a condicionalidade na

probabilidade de detecção de um falso negativo, aplica-se a expressão do valor esperado de ARL, E(ARL) = AARL, dada pela Equação 35.

$$AARL = E(ARL(m, n, Z_{\alpha/2}, \delta, Pe)) = \int_{-\infty}^{\infty} \frac{1}{[1-\theta\left(\frac{W}{\sqrt{m}}+Z_{\alpha/2}-2\delta\sqrt{n}(1-Pe)\right)+\theta\left(\frac{W}{\sqrt{m}}+Z_{\alpha/2}-2\delta\sqrt{n}(1-Pe)\right)]} \varphi(w)dw \quad (35)$$

Onde φ denota a função densidade de probabilidade (em inglês probability density function - pdf) da distribuição normal (0,1).

O RL segue uma distribuição de probabilidade geométrica com desvio padrão igual à média, o que torna a média do RL (ARL) um indicador questionável devido à sua variabilidade. Assim, outras medidas estatísticas são recomendadas para avaliar o desempenho dos gráficos de controle, como a mediana e os quartis. Outro aspecto relevante, envolvendo a Equação 35, é que, como os limites de controle são estimados a partir de uma amostra (bloco de registro de dados), existe uma possibilidade infinita de limites de controle estimados (um limite de controle para cada amostra ou estimativa possível), e assim haverá uma distribuição de RL diferente para cada limite de controle. Em termos estatísticos, o RL e o ARL estão condicionados às estimativas dos limites de controle. Para resolver esse problema, é calculado o AARL, que é a média global do ARL, considerando todas as possibilidades amostrais.

Analisando a Equação 35, dado um determinado nível de confiança $Z_{\alpha/2}$ e aplicando-se diferentes valores de δ, m, n e *Pe*, verifica-se que o desempenho dos gráficos de controle, medido pelo AARL, será diferente para cada composição destes parâmetros. Na seção seguinte essa característica será avaliada em diversos cenários, com o objetivo de estabelecer de maneira objetiva as condições em que o *Kappa* pode ser utilizado como ferramenta de validação contínua dos modelos abordados.

7.2.4 Análise de desempenho dos Gráficos de Controle de *Kappa*

Nesta seção estão definidos cenários criados a partir de diversos valores de δ, m, n, e *Pe*, o que possibilita o cálculo de ARL conforme Equação 33 e AARL conforme Equação 35, definidas previamente. Os valores de *Pe* foram calculados em função de diferentes condições de balanceamento do banco de dados (ou dos eventos de interesse na população), mais especificamente com a probabilidade real de sucesso (em inglês Real Success Probability - RSP) variando de 0,01 a 0,5. Os valores de sensibilidade e especificidade (TPR e TNR) foram ajustados para que os valores de Kappa resultassem próximos a 0,55 em todos os cenários, o que não foi possível para RSP = 0,01.

Quadro 25 - Cenários analisados

	n = 100			n = 300			n = 500		
δ = 0,4	RSP	*Pe*	*Kappa*	RSP	*Pe*	*Kappa*	RSP	*Pe*	*Kappa*
	0.5	0.5	0.54	0.5	0.5	0.553	0.5	0.5	0.556
	0.4	0.512	0.528	0.4	0.512	0.542	0.4	0.512	0.545
	0.3	0.568	0.537	0.3	0.568	0.544	0.3	0.568	0.550
	0.2	0.668	0.518	0.2	0.668	0.548	0.2	0.668	0.548
	0.1	0.812	0.521	0.1	0.812	0.556	0.1	0.812	0.553
	0.05	0.905	0.473	0.05	0.902	0.557	0.05	0.901	0.574
	0.01	0.980	-0.01	0.01	0.980	0.494	0.01	0.980	0.696

	n = 100			n = 300			n = 500		
δ = 0,25	RSP	*Pe*	*Kappa*	RSP	*Pe*	*Kappa*	RSP	*Pe*	*Kappa*
	0.5	0.5	0.54	0.5	0.5	0.553	0.5	0.5	0.556
	0.4	0.512	0.528	0.4	0.512	0.542	0.4	0.512	0.545
	0.3	0.568	0.537	0.3	0.568	0.544	0.3	0.568	0.550
	0.2	0.668	0.518	0.2	0.668	0.548	0.2	0.668	0.548
	0.1	0.812	0.521	0.1	0.812	0.556	0.1	0.812	0.553
	0.05	0.905	0.473	0.05	0.902	0.557	0.05	0.901	0.574
	0.01	0.980	-0.01	0.01	0.980	0.494	0.01	0.980	0.696

	n = 100			n = 300			n = 500		
δ = 0,15	RSP	*Pe*	*Kappa*	RSP	*Pe*	*Kappa*	RSP	*Pe*	*Kappa*
	0.5	0.5	0.54	0.5	0.5	0.553	0.5	0.5	0.556
	0.4	0.512	0.528	0.4	0.512	0.542	0.4	0.512	0.545
	0.3	0.568	0.537	0.3	0.568	0.544	0.3	0.568	0.550
	0.2	0.668	0.518	0.2	0.668	0.548	0.2	0.668	0.548
	0.1	0.812	0.521	0.1	0.812	0.556	0.1	0.812	0.553
	0.05	0.905	0.473	0.05	0.902	0.557	0.05	0.901	0.574
	0.01	0.980	-0.01	0.01	0.980	0.494	0.01	0.980	0.696

	n = 100			n = 300			n = 500		
δ = 0,05	RSP	*Pe*	*Kappa*	RSP	*Pe*	*Kappa*	RSP	*Pe*	*Kappa*
	0.5	0.5	0.54	0.5	0.5	0.553	0.5	0.5	0.556
	0.4	0.512	0.528	0.4	0.512	0.542	0.4	0.512	0.545
	0.3	0.568	0.537	0.3	0.568	0.544	0.3	0.568	0.550
	0.2	0.668	0.518	0.2	0.668	0.548	0.2	0.668	0.548
	0.1	0.812	0.521	0.1	0.812	0.556	0.1	0.812	0.553
	0.05	0.905	0.473	0.05	0.902	0.557	0.05	0.901	0.574
	0.01	0.980	-0.01	0.01	0.980	0.494	0.01	0.980	0.696

	n = 100			n = 300			n = 500		
δ = 0,01	RSP	*Pe*	*Kappa*	RSP	*Pe*	*Kappa*	RSP	*Pe*	*Kappa*
	0.5	0.5	0.54	0.5	0.5	0.553	0.5	0.5	0.556
	0.4	0.512	0.528	0.4	0.512	0.542	0.4	0.512	0.545
	0.3	0.568	0.537	0.3	0.568	0.544	0.3	0.568	0.550
	0.2	0.668	0.518	0.2	0.668	0.548	0.2	0.668	0.548
	0.1	0.812	0.521	0.1	0.812	0.556	0.1	0.812	0.553
	0.05	0.905	0.473	0.05	0.902	0.557	0.05	0.901	0.574
	0.01	0.980	-0.01	0.01	0.980	0.494	0.01	0.980	0.696

	n = 100			n = 300			n = 500		
δ = 0,005	RSP	*Pe*	*Kappa*	RSP	*Pe*	*Kappa*	RSP	*Pe*	*Kappa*
	0.5	0.5	0.54	0.5	0.5	0.553	0.5	0.5	0.556
	0.4	0.512	0.528	0.4	0.512	0.542	0.4	0.512	0.545
	0.3	0.568	0.537	0.3	0.568	0.544	0.3	0.568	0.550
	0.2	0.668	0.518	0.2	0.668	0.548	0.2	0.668	0.548
	0.1	0.812	0.521	0.1	0.812	0.556	0.1	0.812	0.553
	0.05	0.905	0.473	0.05	0.902	0.557	0.05	0.901	0.574
	0.01	0.98	-0.01	0.01	0.980	0.494	0.01	0.980	0.696

Fonte: Próprio Autor

Utilizou-se o Excel® para o cálculo de ARL (Equação 33) e o software Maple® 16 (códigos detalhados no Apêndice B) para o cálculo do AARL (Equação 35). O Quadro 25 apresenta os valores de δ, n e *Pe*, definidos para cada cenário, sendo que para o AARL, em cada um dos cenários foram aplicados os valores de um, cinco, dez e mil para m.

Aqui, vale notar que as análises para $\delta > 0$ são demonstradas, mas os resultados são simétricos para variações de mesmo módulo de δ.

7.2.4.1 Resultados para n e m fixos, variando-se δ e *Pe*

Com base em n e m fixos, a Tabela 4 apresenta resultados para n= 100, 300 e 500, onde calculam-se os valores de ARL para P_0 conhecido e variância máxima de P_0 , $\left(\frac{1}{\sqrt{Po(1-Po)}}\right) = 2$, demonstrando esses valores no bloco denominado ARL max. Como na prática P_0 não é conhecido, é necessário estimá-lo. As implicações desta estimativa são demonstradas nos valores teóricos do AARL. Para o estudo, m=1, m=5 e m=10 foram utilizados por serem valores aplicáveis em situações reais, e o valor de m=1000 foi calculado para demonstrar a aproximação dos resultados estimados de AARL aos valores máximos de ARL quando m aumenta.

Pode se verificar na Tabela 4 que para maiores valores de *Pe* (aumento do desbalanceamento de dados) e menores variações de δ, o ARL max, e o AARL aumentam, principalmente em função de *Pe*. Para valores de *Pe* próximos de 1, os valores de AARL aproximam-se cada vez mais do valor de ARL máximo, conforme m aumenta.

Em termos práticos, valores de δ menores ou iguais a 0,15 necessitam de muitos ciclos para serem detectados. Para n = 100 e P_0 desconhecido, os resultados encontrados somente estão abaixo de 50 ciclos, usando m >= 5, δ >= 0,25 e *Pe* <=0,75. À medida que o tamanho da amostra aumenta de n=100 para n=500, encontramos valores de AARL abaixo de 50 ciclos com m >= 5 e *Pe* <=0,82, para δ >= 0,15.

Com *Pe* acima de 0.9 é necessário um grande aumento em m e n para manter o AARL abaixo de 50 ciclos. Por exemplo, com *Pe* = 0,9014, δ = 0,15 e m=1000, o AARL é 51,39 para n=1000 e 11,96 para n =3000 (usou-se o programa Maple® definido no Apêndice B).

Com *Pe* = 0,9014, δ = 0,15 e m=1 o AARL é 51,75 para n= 3000, ou 54,76 para m=50 e n=1000, ou seja, n precisa aumentar muito em bancos desbalanceados para pequenas variações de δ. Outra conclusão é, desde que n seja suficientemente grande, o AARL reduz para valores razoáveis em situações OOC. Para as mesmas condições de *Pe* = 0,9014 e δ = 0,15 quando se utiliza m=5 e n=3000, o valor de AARL é 19,35. Aqui cabe comentar que um valor de $\delta \geq 0,15$ é uma variação razoável para o *Kappa* que tem valor máximo de 1, assim espera-se que o AARL seja o menor possível, detectando rapidamente a situação fora de controle.

Tabela 4 - Resultados para n e m fixos, variando-se δ e *Pe*

n = 100	ARL max.					
Pe	**δ =0.4**	**δ =0.25**	**δ =0.15**	**δ =0.05**	**δ =0.01**	**δ =0.005**
0.5	1.188	3.241	14.96	155.2	352.9	365.8
0.512	1.223	3.475	16.05	160.1	353.7	366.0
0.568	1.479	4.988	22.62	184.7	357.2	367.0
0.668	2.736	11.09	44.31	235.4	362.5	368.3
0.812	14.85	50.65	131.4	314.6	367.8	369.7
0.905	79.17	165.5	261.5	354.5	369.7	370.2
0.980	329.2	353.2	364.0	369.6	370.3	370.3

n = 300	ARL max.					
Pe	**δ =0.4**	**δ =0.25**	**δ =0.15**	**δ =0.05**	**δ =0.01**	**δ =0.005**
0.5	1.000	1.101	2.908	60.68	322.0	357.1
0.512	1.000	1.123	3.113	63.89	324.1	357.7
0.568	1.001	1.297	4.443	81.57	333.2	360.4
0.668	1.057	2.220	9.883	127.7	347.6	364.4
0.812	2.886	11.75	46.37	238.9	362.8	368.4
0.902	19.88	63.34	151.5	323.7	368.3	369.8
0.980	267.4	322.9	351.9	368.2	370.3	370.3

n = 500	ARL max.					
Pe	**δ =0.4**	**δ =0.25**	**δ =0.15**	**δ =0.05**	**δ =0.01**	**δ =0.005**
0.5	1.000	1.004	1.566	33.40	295.7	348.7
0.512	1.000	1.007	1.645	35.50	298.6	349.7
0.568	1.000	1.034	2.176	47.59	311.9	354.0
0.668	1.001	1.312	4.549	82.81	333.7	360.5
0.812	1.558	5.418	24.36	190.1	357.8	367.1
0.901	9.242	34.60	101.9	297.4	366.8	369.5
0.980	223.5	296.9	340.5	366.8	370.2	370.3

n=100	AARL m=1					
Pe	**δ =0.4**	**δ =0.25**	**δ =0.15**	**δ =0.05**	**δ =0.01**	**δ =0.005**
0.5	1.852	14.83	59.43	127.1	139.6	140
0.512	2.02	16.31	61.84	127.7	139.6	140
0.568	3.382	25.5	73.56	130.3	139.7	140
0.668	11.55	49.33	95.49	134.2	139.9	140
0.812	59.17	99.54	123.8	138.2	140.0	140.1
0.905	112	128.3	135.7	139.6	140.1	140.1
0.980	138.7	139.6	139.9	140.1	140.1	140.1

n=300	AARL m=1					
Pe	**δ =0.4**	**δ =0.25**	**δ =0.15**	**δ =0.05**	**δ =0.01**	**δ =0.005**
0.5	1.003	1.464	12.68	104.8	138.5	139.7
0.512	1.004	1.561	14.01	106.2	138.5	139.7
0.568	1.022	2.385	22.06	112.7	138.9	139.8
0.668	1.285	8.107	45.52	123.2	139.4	139.9
0.812	12.54	51.26	96.88	134.4	139.9	140.0
0.902	69.15	106	126.7	138.5	140	140.1
0.980	136.1	138.5	139.5	140	140.1	140.1

n=500	AARL m=1					
Pe	**δ =0.4**	**δ =0.25**	**δ =0.15**	**δ =0.05**	**δ =0.01**	**δ =0.005**
0.5	1	1.047	3.892	86.54	137.4	139.4
0.512	1	1.061	4.37	88.52	137.5	139.5
0.568	1	1.189	7.817	97.67	138.1	139.6
0.668	1.024	2.464	22.66	113.1	138.9	139.8
0.812	3.843	27.28	76.07	130.8	139.7	140.1
0.901	43.34	87.7	118.2	137.5	140	140.1
0.980	133.4	137.5	139.1	140	140	140.1

n=100	AARL m=5					
Pe	**δ =0.4**	**δ =0.25**	**δ =0.15**	**δ =0.05**	**δ =0.01**	**δ =0.005**
0.5	1.249	4.105	25.08	173.5	234.6	236.9
0.512	1.294	4.461	27.14	176	234.7	236.9
0.568	1.62	6.876	39.38	187.6	235.4	237.1
0.668	3.355	17.78	75.17	206.4	236.3	237.3
0.812	24.86	84.17	159.7	227.1	237.2	237.5
0.905	117.8	178.8	214.2	234.9	237.5	237.6
0.980	230.1	234.7	236.6	237.5	237.6	237.6

n=300	AARL m=5					
Pe	**δ =0.4**	**δ =0.25**	**δ =0.15**	**δ =0.05**	**δ =0.01**	**δ =0.005**
0.5	1	1.138	3.607	97.16	228.6	235.4
0.512	1	1.167	3.912	101	229.1	235.5
0.568	1.003	1.386	5.986	120.2	230.9	235.9
0.668	1.083	2.617	15.53	157.3	233.6	236.6
0.812	3.575	19.02	78.16	207.5	236.4	237.3
0.902	34.32	100.4	171.5	229	237.3	237.6
0.980	215.8	228.8	234.4	237.3	237.6	237.6

n=500	AARL m=5					
Pe	**δ =0.4**	**δ =0.25**	**δ =0.15**	**δ =0.05**	**δ =0.01**	**δ =0.005**
0.5	1	1.009	1.732	58.16	222.9	233.8
0.512	1	1.013	1.836	61.59	223.5	234
0.568	1	1.052	2.556	79.9	226.5	234.8
0.668	1.004	1.406	6.158	121.4	231	236
0.812	1.722	7.591	42.54	189.9	235.5	237.5
0.901	14.35	60.14	138.5	223.3	237.1	237.5
0.980	202.5	223.1	232.3	237	237.6	237.6

n=100	AARL m=10					
Pe	**δ =0.4**	**δ =0.25**	**δ =0.15**	**δ =0.05**	**δ =0.01**	**δ =0.005**
0.5	1.217	3.615	19.2	176.2	272.6	276.5
0.512	1.257	3.9	20.72	179.8	272.8	276.6
0.568	1.544	5.783	30.03	196.8	273.9	276.9
0.668	3.008	13.85	60.64	225.8	275.5	277.3
0.812	19.03	69.24	157.1	259.7	277.1	277.7
0.905	104.8	183.8	238.2	273.1	277.7	277.8
0.980	264.8	272.7	276	277.7	277.9	277.9

n=300	AARL m=10					
Pe	**δ =0.4**	**δ =0.25**	**δ =0.15**	**δ =0.05**	**δ =0.01**	**δ =0.005**
0.5	1	1.119	3.213	82.34	262.3	273.9
0.512	1	1.144	3.46	86.4	263	274.1
0.568	1.002	1.339	5.097	107.6	266.2	274.9
0.668	1.07	2.397	12.2	153.8	270.9	276.1
0.812	3.187	14.75	63.45	227.5	275.6	277.3
0.902	26.13	85.71	173.4	262.9	277.2	277.7
0.980	240.9	262.6	272.3	277.2	277.8	277.9

n=500	AARL m=10					
Pe	**δ =0.4**	**δ =0.25**	**δ =0.15**	**δ =0.05**	**δ =0.01**	**δ =0.005**
0.5	1	1.007	1.642	45.4	252.6	271.3
0.512	1	1.01	1.732	48.37	253.7	271.6
0.568	1	1.043	2.346	65.11	258.7	272.9
0.668	1.003	1.356	5.23	109	266.3	274.9
0.812	1.633	6.328	32.52	200.2	274.1	277.7
0.901	11.34	47.11	129.6	253.2	276.8	277.6
0.980	219.6	253	268.6	276.8	277.8	277.9

n=100	AARL m= 1000					
Pe	**δ =0.4**	**δ =0.25**	**δ =0.15**	**δ =0.05**	**δ =0.01**	**δ =0.005**
0.5	1.189	3.244	15	155.7	351.6	364.2
0.512	1.224	3.479	16.1	160.6	352.3	364.4
0.568	1.48	4.996	22.69	185.2	355.7	365.3
0.668	2.739	11.12	44.45	235.8	360.9	366.6
0.812	14.89	50.83	131.9	314	366.1	368
0.905	79.45	166	261.6	353.2	368	368.4
0.980	328.3	351.9	362.4	367.9	368.6	368.6

n=300	AARL m= 1000					
Pe	**δ =0.4**	**δ =0.25**	**δ =0.15**	**δ =0.05**	**δ =0.01**	**δ =0.005**
0.5	1	1.101	2.911	60.9	321.3	355.7
0.512	1	1.124	3.116	64.12	323.3	356.3
0.568	1.001	1.298	4.449	81.87	332.3	358.9
0.668	1.058	2.222	9.903	128.1	346.3	362.8
0.812	2.889	11.78	46.52	239.2	361.2	366.7
0.902	19.93	63.57	152	323	366.6	368.1
0.980	267.5	322.2	350.6	366.5	368.5	368.6

n=500	AARL m= 1000					
Pe	**δ =0.4**	**δ =0.25**	**δ =0.15**	**δ =0.05**	**δ =0.01**	**δ =0.005**
0.5	1	1.005	1.567	33.5	295.4	347.5
0.512	1	1.007	1.646	35.61	298.3	348.5
0.568	1	1.035	2.178	47.75	311.4	352.7
0.668	1.002	1.313	4.556	83.11	332.8	359
0.812	1.559	5.426	24.43	190.6	356.4	368
0.901	9.26	34.71	102.4	297.1	365.2	367.7
0.980	223.9	296.6	339.4	365.1	368.5	368.6

Fonte: Próprio Autor

7.2.4.2 Resultados para n e δ fixos, variando-se m e *Pe*

Na análise dos cenários, com base em n e δ fixos, a Tabela 5 apresenta os resultados do ARL max e AARL para n= 100, 300 e 500, e para δ = 0,4, 0,25, 0,15, 0,05, 0,01 e 0,005. Nestas tabelas percebe-se que os valores de AARL se aproximam de ARL max, maior o m aplicado, e que para m $\geq$ 10 estes valores já se encontram razoavelmente próximos. Pode-se dizer que para um $\delta \leq$0,01, os resultados aproximam-se dos valores de AARLo (AARLo = AARL sob controle, quando δ = 0). O AARLo também pode ser calculado pelo programa Maple® 16, desta forma, utilizando a condição de δ = 0, n=100, m=1 e qualquer *Pe* , AARLo = 140,1 e para δ = 0, n=100, m=5 e qualquer *Pe,* AARLo = 237,6.

Para δ = 0,4, *Pe* = 0,812 e m=5, pode-se encontrar valores razoáveis de AARL (24,86 para n=100, 3,575 para n=300 e 1,722 para n=500). O problema é que um desvio de 0,4, em *Kappa,* é um grande desvio, uma vez que os valores aplicáveis desse indicador estão na faixa

de (0,5:1,0). Para a configuração de cenário, com δ = 0,05 (menor), *Pe* = 0,812, m=5 e n=100, AARL=227,1, um valor alto para uma aplicação prática. Amostras com m=5 e n=5000 seriam necessárias para um AARL = 36,57, que ainda seria bastante alto, ou seja, a detecção da variação demoraria muito em caso de baixa frequência de retiradas de amostras.

Simulando AARL, para δ = 0,05, *Pe* = 0,9014, m=5 e n= 50.000, o resultado é 22,06, o que, mesmo em contextos de Big Data, pode não ser um tamanho de amostra aplicável na prática. Observe que os valores de AARL aumentam rapidamente para valores menores de δ ou valores maiores de *Pe*, aumentando-se n, em geral, os valores de AARL crescem menos rapidamente.

Tabela 5 - Resultados para n e δ fixos, variando-se m e *Pe*

n = 100 **δ = 0,4**

RSP	*Kappa*	*Pe*	ARL max.	AARL m=1	AARL m=5	AARL m=10	AARL m=1000
0,50	0,540	0,5	1,189	1,852	1,249	1,217	1,189
0,40	0,529	0,512	1,224	2,020	1,294	1,257	1,224
0,30	0,537	0,568	1,480	3,382	1,620	1,544	1,480
0,20	0,518	0,668	2,737	11,550	3,355	3,008	2,739
0,10	0,521	0,812	14,852	59,170	24,860	19,030	14,890
0,05	0,474	0,905	79,174	112,000	117,800	104,800	79,450
0,01	-0,010	0,9802	329,210	138,700	230,100	264,800	328,300

n = 300 **δ = 0,4**

RSP	*Kappa*	*Pe*	ARL max.	AARL m=1	AARL m=5	AARL m=10	AARL m=1000
0,50	0,553	0,5	1,000	1,003	1,000	1,000	1,000
0,40	0,542	0,512	1,000	1,004	1,000	1,000	1,000
0,30	0,545	0,568	1,001	1,022	1,003	1,002	1,001
0,20	0,548	0,668	1,058	1,285	1,083	1,070	1,058
0,10	0,557	0,812	2,887	12,540	3,575	3,187	2,889
0,05	0,558	0,902	19,883	69,150	34,320	26,130	19,930
0,01	0,495	0,9802	267,499	136,100	215,800	240,900	267,500

n = 500 **δ = 0,4**

RSP	*Kappa*	*Pe*	ARL max.	AARL m=1	AARL m=5	AARL m=10	AARL m=1000
0,50	0,556	0,5	1,000	1,000	1,000	1,000	1,000
0,40	0,545	0,512	1,000	1,000	1,000	1,000	1,000
0,30	0,551	0,568	1,000	1,000	1,000	1,000	1,000
0,20	0,548	0,668	1,002	1,024	1,004	1,003	1,002
0,10	0,553	0,812	1,558	3,843	1,722	1,633	1,559
0,05	0,574	0,9014	9,243	43,340	14,350	11,340	9,260
0,01	0,697	0,9802	223,574	133,400	202,500	219,600	223,900

n = 100 **δ = 0,25**

RSP	*Kappa*	*Pe*	ARL max.	AARL m=1	AARL m=5	AARL m=10	AARL m=1000
0,50	0,540	0,5	3,241	14,830	4,105	3,615	3,244
0,40	0,529	0,512	3,475	16,310	4,461	3,900	3,479
0,30	0,537	0,568	4,989	25,500	6,876	5,783	4,996
0,20	0,518	0,668	11,096	49,330	17,780	13,850	11,120
0,10	0,521	0,812	50,659	99,540	84,170	69,240	50,830
0,05	0,474	0,905	165,554	128,300	178,800	183,800	166,000
0,01	-0,010	0,9802	353,264	139,600	234,700	272,700	351,900

n = 300 **δ = 0,25**

RSP	*Kappa*	*Pe*	ARL max.	AARL m=1	AARL m=5	AARL m=10	AARL m=1000
0,50	0,553	0,5	1,101	1,464	1,138	1,119	1,101
0,40	0,542	0,512	1,124	1,561	1,167	1,144	1,124
0,30	0,545	0,568	1,297	2,385	1,386	1,339	1,298
0,20	0,548	0,668	2,221	8,107	2,617	2,397	2,222
0,10	0,557	0,812	11,757	51,260	19,020	14,750	11,780
0,05	0,558	0,902	63,348	106,000	100,400	85,710	63,570
0,01	0,495	0,9802	322,948	138,500	228,800	262,600	322,200

n = 500 **δ = 0,25**

RSP	*Kappa*	*Pe*	ARL max.	AARL m=1	AARL m=5	AARL m=10	AARL m=1000
0,50	0,556	0,5	1,005	1,047	1,009	1,007	1,005
0,40	0,545	0,512	1,007	1,061	1,013	1,010	1,007
0,30	0,551	0,568	1,035	1,189	1,052	1,043	1,035
0,20	0,548	0,668	1,313	2,464	1,406	1,356	1,313
0,10	0,553	0,812	5,418	27,280	7,591	6,328	5,426
0,05	0,574	0,9014	34,610	87,700	60,140	47,110	34,710
0,01	0,697	0,9802	296,971	137,500	223,100	253,000	296,600

n = 100 **δ = 0,15**

RSP	*Kappa*	*Pe*	ARL max.	AARL m=1	AARL m=5	AARL m=10	AARL m=1000
0,50	0,540	0,5	14,968	59,430	25,080	19,200	15,000
0,40	0,529	0,512	16,058	61,840	27,140	20,720	16,100
0,30	0,537	0,568	22,625	73,560	39,380	30,030	22,690
0,20	0,518	0,668	44,312	95,490	75,170	60,640	44,450
0,10	0,521	0,812	131,441	123,800	159,700	157,100	131,900
0,05	0,474	0,905	261,527	135,700	214,200	238,200	261,600
0,01	-0,010	0,9802	364,061	139,900	236,600	276,000	362,400

n = 300 **δ = 0,15**

RSP	*Kappa*	*Pe*	ARL max.	AARL m=1	AARL m=5	AARL m=10	AARL m=1000
0,50	0,553	0,5	2,908	12,680	3,607	3,213	2,911
0,40	0,542	0,512	3,113	14,010	3,912	3,460	3,116
0,30	0,545	0,568	4,444	22,060	5,986	5,097	4,449
0,20	0,548	0,668	9,884	45,520	15,530	12,200	9,903
0,10	0,557	0,812	46,371	96,880	78,160	63,450	46,520
0,05	0,558	0,902	151,563	126,700	171,500	173,400	152,000
0,01	0,495	0,9802	351,955	139,500	234,400	272,300	350,600

n = 500 **δ = 0,15**

RSP	*Kappa*	*Pe*	ARL max.	AARL m=1	AARL m=5	AARL m=10	AARL m=1000
0,50	0,556	0,5	1,566	3,892	1,732	1,642	1,567
0,40	0,545	0,512	1,645	4,370	1,836	1,732	1,646
0,30	0,551	0,568	2,177	7,817	2,556	2,346	2,178
0,20	0,548	0,668	4,550	22,660	6,158	5,230	4,556
0,10	0,553	0,812	24,366	76,070	42,540	32,520	24,430
0,05	0,574	0,9014	101,997	118,200	138,500	129,600	102,400
0,01	0,697	0,9802	340,550	139,100	232,300	268,600	339,400

n = 100 **δ = 0,05**

RSP	*Kappa*	*Pe*	ARL max.	AARL m=1	AARL m=5	AARL m=10	AARL m=1000
0,50	0,540	0,5	155,224	127,100	173,500	176,200	155,700
0,40	0,529	0,512	160,108	127,700	176,000	179,800	160,600
0,30	0,537	0,568	184,709	130,300	187,600	196,800	185,200
0,20	0,518	0,668	235,497	134,200	206,400	225,800	235,800
0,10	0,521	0,812	314,661	138,200	227,100	259,700	314,000
0,05	0,474	0,905	354,568	139,600	234,900	273,100	353,200
0,01	-0,010	0,9802	369,684	140,100	237,500	277,700	367,900

n = 300 **δ = 0,05**

RSP	*Kappa*	*Pe*	ARL max.	AARL m=1	AARL m=5	AARL m=10	AARL m=1000
0,50	0,553	0,5	60,688	104,800	97,160	82,340	60,900
0,40	0,542	0,512	63,895	106,200	101,000	86,400	64,120
0,30	0,545	0,568	81,580	112,700	120,200	107,600	81,870
0,20	0,548	0,668	127,712	123,200	157,300	153,800	128,100
0,10	0,557	0,812	238,975	134,400	207,500	227,500	239,200
0,05	0,558	0,902	323,793	138,500	229,000	262,900	323,000
0,01	0,495	0,9802	368,264	140,000	237,300	277,200	366,500

n = 500 **δ = 0,05**

RSP	*Kappa*	*Pe*	ARL max.	AARL m=1	AARL m=5	AARL m=10	AARL m=1000
0,50	0,556	0,5	33,401	86,540	58,160	45,400	33,500
0,40	0,545	0,512	35,504	88,520	61,590	48,370	35,610
0,30	0,551	0,568	47,593	97,670	79,900	65,110	47,750
0,20	0,548	0,668	82,811	113,100	121,400	109,000	83,110
0,10	0,553	0,812	190,182	130,800	189,900	200,200	190,600
0,05	0,574	0,9014	297,458	137,500	223,300	253,200	297,100
0,01	0,697	0,9802	366,854	140,000	237,000	276,800	365,100

n = 100 **δ = 0,01**

RSP	*Kappa*	*Pe*	ARL max.	AARL m=1	AARL m=5	AARL m=10	AARL m=1000
0,50	0,540	0,5	352,931	139,600	234,600	272,600	351,600
0,40	0,529	0,512	353,726	139,600	234,700	272,800	352,300
0,30	0,537	0,568	357,217	139,700	235,400	273,900	355,700
0,20	0,518	0,668	362,510	139,900	236,300	275,500	360,900
0,10	0,521	0,812	367,836	140,000	237,200	277,100	366,100
0,05	0,474	0,905	369,741	140,100	237,500	277,700	368,000
0,01	-0,010	0,9802	370,370	140,100	237,600	277,900	368,600

n = 300 **δ = 0,01**

RSP	*Kappa*	*Pe*	ARL max.	AARL m=1	AARL m=5	AARL m=10	AARL m=1000
0,50	0,553	0,5	322,097	138,500	228,600	262,300	321,300
0,40	0,542	0,512	324,130	138,500	229,100	263,000	323,300
0,30	0,545	0,568	333,235	138,900	230,900	266,200	332,300
0,20	0,548	0,668	347,606	139,400	233,600	270,900	346,300
0,10	0,557	0,812	362,804	139,900	236,400	275,600	361,200
0,05	0,558	0,902	368,307	140,000	237,300	277,200	366,600
0,01	0,495	0,9802	370,313	140,100	237,600	277,800	368,500

n = 500 **δ = 0,01**

RSP	*Kappa*	*Pe*	ARL max.	AARL m=1	AARL m=5	AARL m=10	AARL m=1000
0,50	0,556	0,5	295,751	137,400	222,900	252,600	295,400
0,40	0,545	0,512	298,672	137,500	223,500	253,700	298,300
0,30	0,551	0,568	311,981	138,100	226,500	258,700	311,400
0,20	0,548	0,668	333,764	138,900	231,000	266,300	332,800
0,10	0,553	0,812	357,895	139,700	235,500	274,100	356,400
0,05	0,574	0,9014	366,882	140,000	237,100	276,800	365,200
0,01	0,697	0,9802	370,255	140,000	237,600	277,800	368,500

n = 100 **δ = 0,005**

RSP	*Kappa*	*Pe*	ARL max.	AARL m=1	AARL m=5	AARL m=10	AARL m=1000
0,50	0,540	0,5	365,888	140,000	236,900	276,500	364,200
0,40	0,529	0,512	366,100	140,000	236,900	276,600	364,400
0,30	0,537	0,568	367,022	140,000	237,100	276,900	365,300
0,20	0,518	0,668	368,398	140,000	237,300	277,300	366,600
0,10	0,521	0,812	369,755	140,100	237,500	277,700	368,000
0,05	0,474	0,905	370,234	140,100	237,600	277,800	368,400
0,01	-0,010	0,9802	370,391	140,100	237,600	277,900	368,600

n = 300 **δ = 0,005**

RSP	*Kappa*	*Pe*	ARL max.	AARL m=1	AARL m=5	AARL m=10	AARL m=1000
0,50	0,553	0,5	357,157	139,700	235,400	273,900	355,700
0,40	0,542	0,512	357,766	139,700	235,500	274,100	356,300
0,30	0,545	0,568	360,433	139,800	235,900	274,900	358,900
0,20	0,548	0,668	364,454	139,900	236,600	276,100	362,800
0,10	0,557	0,812	368,473	140,000	237,300	277,300	366,700
0,05	0,558	0,902	369,873	140,100	237,600	277,700	368,100
0,01	0,495	0,9802	370,377	140,100	237,600	277,900	368,600

n = 500 **δ = 0,005**

RSP	*Kappa*	*Pe*	ARL max.	AARL m=1	AARL m=5	AARL m=10	AARL m=1000
0,50	0,556	0,5	348,793	139,400	233,800	271,300	347,500
0,40	0,545	0,512	349,766	139,500	234,000	271,600	348,500
0,30	0,551	0,568	354,053	139,600	234,800	272,900	352,700
0,20	0,548	0,668	360,585	139,800	236,000	274,900	359,000
0,10	0,553	0,812	367,200	140,100	237,500	277,700	368,000
0,05	0,574	0,9014	369,514	140,100	237,500	277,600	367,700
0,01	0,697	0,9802	370,363	140,100	237,600	277,900	368,600

Fonte: Próprio Autor

7.2.4.3 Resultados para δ e m fixos, variando-se n e *Pe*

Na condição de análise para δ e m fixos, variando n e *Pe*, Tabela 6, a apresentação das entradas é ligeiramente simplificada. Na proporção RSP = 0,05, do balanço de dados, os valores de *Pe* calculados não são exatamente iguais para cada valor de n apresentado. Como referência, a Tabela 6 apresenta os valores de *Pe* = 0,902 (que é o valor para n=300) enquanto os valores de *Pe* utilizados nos cálculos, para n=100 e n=500, são 0,905 e 0,9014, respectivamente.

Tabela 6 - Resultados para δ e m fixos, variando-se n e *Pe*

ARL max.	**δ = 0,4**		
Pe	n=100	n=300	n=500
0.5	1.189	1.000	1.000
0.512	1.224	1.000	1.000
0.568	1.480	1.001	1.000
0.668	2.737	1.058	1.002
0.812	14.852	2.887	1.558
0.902	79.174	19.883	9.243
0.9802	329.210	267.499	223.574

AARL m=1	**δ = 0,4**		
Pe	n=100	n=300	n=500
0.5	1.852	1.003	1.000
0.512	2.020	1.004	1.000
0.568	3.382	1.022	1.000
0.668	11.550	1.285	1.024
0.812	59.170	12.540	3.843
0.902	112.000	69.150	43.340
0.9802	138.700	136.100	133.400

AARL m=5	**δ = 0,4**		
Pe	n=100	n=300	n=500
0.5	1.249	1.000	1.000
0.512	1.294	1.000	1.000
0.568	1.620	1.003	1.000
0.668	3.355	1.083	1.004
0.812	24.860	3.575	1.722
0.902	117.800	34.320	14.350
0.9802	230.100	215.800	202.500

AARL m=10	**δ = 0,4**		
Pe	n=100	n=300	n=500
0.5	1.217	1.000	1.000
0.512	1.257	1.000	1.000
0.568	1.544	1.002	1.000
0.668	3.008	1.070	1.003
0.812	19.030	3.187	1.633
0.902	104.800	26.130	11.340
0.9802	264.800	240.900	219.600

AARL m=1000	**δ = 0,4**		
Pe	n=100	n=300	n=500
0.5	1.189	1.000	1.000
0.512	1.224	1.000	1.000
0.568	1.480	1.001	1.000
0.668	2.739	1.058	1.002
0.812	14.890	2.889	1.559
0.902	79.450	19.930	9.260
0.9802	328.300	267.500	223.900

ARL max.	**δ = 0,25**		
Pe	n=100	n=300	n=500
0.5	3.241	1.101	1.005
0.512	3.475	1.124	1.007
0.568	4.989	1.297	1.035
0.668	11.096	2.221	1.313
0.812	50.659	11.757	5.418
0.902	165.554	63.348	34.610
0.9802	353.264	322.948	296.971

AARL m=1	**δ = 0,25**		
Pe	n=100	n=300	n=500
0.5	14.830	1.464	1.047
0.512	16.310	1.561	1.061
0.568	25.500	2.385	1.189
0.668	49.330	8.107	2.464
0.812	99.540	51.260	27.280
0.902	128.300	106.000	87.700
0.9802	139.600	138.500	137.500

AARL m=5	**δ = 0,25**		
Pe	n=100	n=300	n=500
0.5	4.105	1.138	1.009
0.512	4.461	1.167	1.013
0.568	6.876	1.386	1.052
0.668	17.780	2.617	1.406
0.812	84.170	19.020	7.591
0.902	178.800	100.400	60.140
0.9802	234.700	228.800	223.100

AARL m=10	**δ = 0,25**		
Pe	n=100	n=300	n=500
0.5	3.615	1.119	1.007
0.512	3.900	1.144	1.010
0.568	5.783	1.339	1.043
0.668	13.850	2.397	1.356
0.812	69.240	14.750	6.328
0.902	183.800	85.710	47.110
0.9802	272.700	262.600	253.000

AARL m=1000	**δ = 0,25**		
Pe	n=100	n=300	n=500
0.5	3.244	1.101	1.005
0.512	3.479	1.124	1.007
0.568	4.996	1.298	1.035
0.668	11.120	2.222	1.313
0.812	50.830	11.780	5.426
0.902	166.000	63.570	34.710
0.9802	351.900	322.200	296.600

ARL max.	**δ = 0,15**		
Pe	n=100	n=300	n=500
0.5	14.968	2.908	1.566
0.512	16.058	3.113	1.645
0.568	22.625	4.444	2.177
0.668	44.312	9.884	4.550
0.812	131.441	46.371	24.366
0.902	261.527	151.563	101.997
0.9802	364.061	351.955	340.550

AARL m=1	**δ = 0,15**		
Pe	n=100	n=300	n=500
0.5	59.430	12.680	3.892
0.512	61.840	14.010	4.370
0.568	73.560	22.060	7.817
0.668	95.490	45.520	22.660
0.812	123.800	96.880	76.070
0.902	135.700	126.700	118.200
0.9802	139.900	139.500	139.100

AARL m=5	**δ = 0,15**		
Pe	n=100	n=300	n=500
0.5	25.080	3.607	1.732
0.512	27.140	3.912	1.836
0.568	39.380	5.986	2.556
0.668	75.170	15.530	6.158
0.812	159.700	78.160	42.540
0.902	214.200	171.500	138.500
0.9802	236.600	234.400	232.300

AARL m=10	**δ = 0,15**		
Pe	n=100	n=300	n=500
0.5	19.200	3.213	1.642
0.512	20.720	3.460	1.732
0.568	30.030	5.097	2.346
0.668	60.640	12.200	5.230
0.812	157.100	63.450	32.520
0.902	238.200	173.400	129.600
0.9802	276.000	272.300	268.600

AARL m=1000	**δ = 0,15**		
Pe	n=100	n=300	n=500
0.5	15.000	2.911	1.567
0.512	16.100	3.116	1.646
0.568	22.690	4.449	2.178
0.668	44.450	9.903	4.556
0.812	131.900	46.520	24.430
0.902	261.600	152.000	102.400
0.9802	362.400	350.600	339.400

ARL max.	**δ = 0,05**		
Pe	n=100	n=300	n=500
0.5	155.224	60.688	33.401
0.512	160.108	63.895	35.504
0.568	184.709	81.580	47.593
0.668	235.497	127.712	82.811
0.812	314.661	238.975	190.182
0.902	354.568	323.793	297.458
0.9802	369.684	368.264	366.854

AARL m=1	**δ = 0,05**		
Pe	n=100	n=300	n=500
0.5	127.100	104.800	86.540
0.512	127.700	106.200	88.520
0.568	130.300	112.700	97.670
0.668	134.200	123.200	113.100
0.812	138.200	134.400	130.800
0.902	139.600	138.500	137.500
0.9802	140.100	140.000	140.000

AARL m=5	**δ = 0,05**		
Pe	n=100	n=300	n=500
0.5	173.500	97.160	58.160
0.512	176.000	101.000	61.590
0.568	187.600	120.200	79.900
0.668	206.400	157.300	121.400
0.812	227.100	207.500	189.900
0.902	234.900	229.000	223.300
0.9802	237.500	237.300	237.000

AARL m=10	**δ = 0,05**		
Pe	n=100	n=300	n=500
0.5	176.200	82.340	45.400
0.512	179.800	86.400	48.370
0.568	196.800	107.600	65.110
0.668	225.800	153.800	109.000
0.812	259.700	227.500	200.200
0.902	273.100	262.900	253.200
0.9802	277.700	277.200	276.800

AARL m=1000	**δ = 0,05**		
Pe	n=100	n=300	n=500
0.5	155.700	60.900	33.500
0.512	160.600	64.120	35.610
0.568	185.200	81.870	47.750
0.668	235.800	128.100	83.110
0.812	314.000	239.200	190.600
0.902	353.200	323.000	297.100
0.9802	367.900	366.500	365.100

ARL max.	**δ = 0,01**		
Pe	n=100	n=300	n=500
0.5	352.931	322.097	295.751
0.512	353.726	324.130	298.672
0.568	357.217	333.235	311.981
0.668	362.510	347.606	333.764
0.812	367.836	362.804	357.895
0.902	369.741	368.307	366.882
0.9802	370.370	370.313	370.255

AARL m=1	**δ = 0,01**		
Pe	n=100	n=300	n=500
0.5	139.600	138.500	137.400
0.512	139.600	138.500	137.500
0.568	139.700	138.900	138.100
0.668	139.900	139.400	138.900
0.812	140.000	139.900	139.700
0.902	140.100	140.000	140.000
0.9802	140.100	140.100	140.000

AARL m=5	**δ = 0,01**		
Pe	n=100	n=300	n=500
0.5	234.600	228.600	222.900
0.512	234.700	229.100	223.500
0.568	235.400	230.900	226.500
0.668	236.300	233.600	231.000
0.812	237.200	236.400	235.500
0.902	237.500	237.300	237.100
0.9802	237.600	237.600	237.600

AARL m=10	**δ = 0,01**		
Pe	n=100	n=300	n=500
0.5	272.600	262.300	252.600
0.512	272.800	263.000	253.700
0.568	273.900	266.200	258.700
0.668	275.500	270.900	266.300
0.812	277.100	275.600	274.100
0.902	277.700	277.200	276.800
0.9802	277.900	277.800	277.800

AARL m=1000	**δ = 0,01**		
Pe	n=100	n=300	n=500
0.5	351.600	321.300	295.400
0.512	352.300	323.300	298.300
0.568	355.700	332.300	311.400
0.668	360.900	346.300	332.800
0.812	366.100	361.200	356.400
0.902	368.000	366.600	365.200
0.9802	368.600	368.500	368.500

ARL max.	**δ = 0,005**		
Pe	n=100	n=300	n=500
0.5	365.888	357.157	348.793
0.512	366.100	357.766	349.766
0.568	367.022	360.433	354.053
0.668	368.398	364.454	360.585
0.812	369.755	368.473	367.200
0.902	370.234	369.873	369.514
0.9802	370.391	370.377	370.363

AARL m=1	**δ = 0,005**		
Pe	n=100	n=300	n=500
0.5	140.000	139.700	139.400
0.512	140.000	139.700	139.500
0.568	140.000	139.800	139.600
0.668	140.000	139.900	139.800
0.812	140.100	140.000	140.100
0.902	140.100	140.100	140.100
0.9802	140.100	140.100	140.100

AARL m=5	**δ = 0,005**		
Pe	n=100	n=300	n=500
0.5	236.900	235.400	233.800
0.512	236.900	235.500	234.000
0.568	237.100	235.900	234.800
0.668	237.300	236.600	236.000
0.812	237.500	237.300	237.500
0.902	237.600	237.600	237.500
0.9802	237.600	237.600	237.600

AARL m=10	**δ = 0,005**		
Pe	n=100	n=300	n=500
0.5	276.500	273.900	271.300
0.512	276.600	274.100	271.600
0.568	276.900	274.900	272.900
0.668	277.300	276.100	274.900
0.812	277.700	277.300	277.700
0.902	277.800	277.700	277.600
0.9802	277.900	277.900	277.900

AARL m=1000	**δ = 0,005**		
Pe	n=100	n=300	n=500
0.5	364.200	355.700	347.500
0.512	364.400	356.300	348.500
0.568	365.300	358.900	352.700
0.668	366.600	362.800	359.000
0.812	368.000	366.700	368.000
0.902	368.400	368.100	367.700
0.9802	368.600	368.600	368.600

Fonte: Próprio Autor

Analisando a Tabela 6, evidencia-se que, para δ <= 0,05 e Pe >= 0,668, o AARL com é muito alto, independentemente do tamanho n de amostra. Na condição de δ = 0,05, e qualquer valor de *Pe* e n, o AARL excede 33,5 ciclos, o que pode representar, dependendo da aplicação,

muitos ciclos para se detectar variações. Como já discutido, a solução é aumentar m e n, aumentando assim a sensibilidade de detecção (menor AARL). Nas Tabelas 7 e 8, o programa Maple® do apêndice B é utilizado para simular novas condições de n e m que resultam em valores de AARL mais compatíveis. A coluna denominada "AARL m=1.000" é usada como referência para o ARL max (calculado com base em *Po* conhecido). e a coluna "Reg." é o resultado da multiplicação m X n, indicando a quantidade de registros totais utilizados. Nessas tabelas, os valores de m e n foram escolhidos para resultar em valores de AARL mais baixos. Observa-se na Tabela 8 que um pequeno aumento em δ (de 0,05 para 0,1) reduz em cinco vezes os valores da coluna "Reg.", comparando-se a condição de δ = 0,05. No entanto, mesmo para δ = 0,1, o valor de n aumenta acentuadamente para valores mais altos de *Pe*. Usando o código em Maple® encontra-se AARL=2,509 para: δ = 0,1; *Pe* = 0,9; m=10; n = 20.000. Ou seja, n praticamente é dez vezes maior para *Pe* = 0,9 se comparado a *Pe* = 0,812, nas mesmas condições.

Tabela 7 - AARL para maiores valores de n com δ = 0,05

m	n	δ	*Pe*	AARL	AARL m=1.000	Reg.
10	2500	0.05	0.668	13.85	11.12	25.000
5	2500	0.05	0.668	17.78	11.12	12.500
1	2500	0.05	0.668	49.33	11.12	2.500
10	5000	0.05	0.668	4,409	3,895	50.000
5	5000	0.05	0.668	5,105	3,895	25.000
1	5000	0.05	0.668	18.85	3,895	5.000
10	7500	0.05	0.812	14.75	11.78	75.000
5	7500	0.05	0.812	19.02	11.78	37.500
1	7500	0.05	0.812	51.26	11.78	7.500
10	15000	0.05	0.812	4,693	4,124	150.000
5	15000	0.05	0.812	5,467	4,124	75.000
1	15000	0.05	0.812	20.2	4,124	15.000

Fonte: Próprio Autor

Tabela 8 - AARL para maiores valores de n com δ = 0,1

m	n	δ	*Pe*	AARL	AARL m=1.000	Reg.
10	500	0.1	0.668	19.83	15.45	5000
5	500	0.1	0.668	25.93	15.45	2500
1	500	0.1	0.668	60.44	15.45	500
10	1000	0.1	0.668	6.349	5.443	10000
5	1000	0.1	0.668	7.619	5.443	5000
1	1000	0.1	0.668	27.36	5.443	1000
10	1500	0.1	0.812	3.314	2.995	15000
5	1500	0.1	0.812	3.731	2.995	7500
1	1500	0.1	0.812	13.23	2.995	1500
10	2000	0.1	0.812	2.198	2.051	20000
5	2000	0.1	0.812	2.381	2.051	10000
1	2000	0.1	0.812	6.972	2.051	2000

Fonte: Próprio Autor

7.2.5 Considerações sobre os resultados

A partir da análise de AARL dos cenários calculados, conclui-se que amostras com m>5 (preferencialmente m=10) reduzem o erro α e melhoram o poder do teste $(1 - \beta)$. No entanto, o valor de n, que resulta em comprimentos de sequência adequados (ARL ou AARL) depende dos valores de *Pe e* δ (variação de *Kappa* em relação ao valor médio calculado na Fase I do SPM). Os resultados apresentados nas Tabelas 4 a 8 auxiliam a escolha adequada das amostras (m e n) a serem aplicadas nas Fases I e II do SPM para *Kappa,* fornecendo critérios quantitativos que poderão ser replicados em função dos parâmetros apresentados.

Para *Kappa* maior que 0,4 (o que representa uma concordância moderada) e lembrando que seu valor máximo é 1, um desvio $\delta \geq 0{,}05$ é significativo e deveria ser detectável. O problema é que, para este nível de detecção, considerando-se o critério conservador de máxima variância em *Po*, o tamanho da amostra n, que resulta em AARL abaixo de 15 ciclos, é alto e aumenta à medida que o *Pe* aumenta (Tabela 7). Isso pode inviabilizar o uso do indicador em alguns casos. Analisando a Tabela 8, para aplicações onde $\delta \geq 0{,}1$ é um nível de sensibilidade adequado, o valor de n cai consideravelmente em relação a $\delta \geq 0{,}05$, mas ainda é bastante alto para $Pe > 0{,}8$.

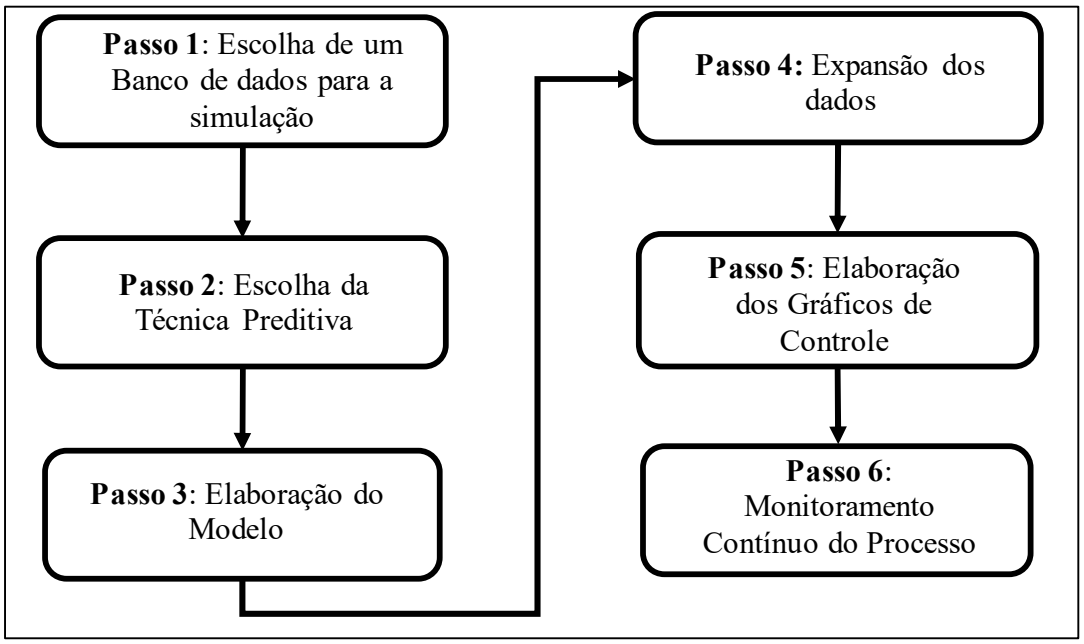
De maneira geral, conclui-se que o gráfico de controle de *Kappa* é adequado para situações em que $Pe < 0{,}7$ e $\delta \geq 0{,}1$, utilizando-se m=10 e n=500, sendo que m e n diminuem conforme *Pe* diminui e δ aumenta. Por outro lado, se $Pe \geq 0{,}7$ e $\delta < 0{,}1$, m e n crescem acentuadamente para um desempenho adequado dos gráficos de controle de *Kappa,* podendo inviabilizar sua utilização devido a necessidade de muitas amostras.

7.3 SIMULAÇÕES ESTOCÁSTICAS DE DESEMPENHO

As seções a seguir utilizam simulações para criar condições próximas a realidade com o objetivo de demonstrar a variabilidade das estimativas dos índices de qualidade aplicados aos modelos. A análise destas simulações, em conjunto com o método analítico desenvolvido na seção 7.2, fornecerá as informações necessárias para a validação do uso do SPM no seu monitoramento contínuo, e, consequentemente da abordagem apresentada no item 7.1. Justifica-se o uso de simulações, pois auxiliam no estudo prático de sistemas, buscando-se também a simplificação matemática e com isso um maior alcance fora do mundo acadêmico. No caso aplicado, questões complexas tratadas por Jardim, Chakraborti e Epprecht (2019) como problemas relativos à estimativa de parâmetros (μ e σ da população desconhecidos), podem ser

simplificadas com o uso dessa técnica. A estrutura geral da simulação está demonstrada na Figura 44.

Figura 44 - Estrutura geral das Simulações



Fonte: Próprio Autor

O primeiro passo foi a escolha de um banco de dados, já rotulado, para a elaboração de um modelo classificatório. Utilizou-se um exemplo didático extraído do software Statistica® (2007), composto por duas variáveis, uma independente, referente ao número de meses de experiência profissional, e a outra, tratada como dependente (ou rótulo), que registra o sucesso ou o fracasso da execução de tarefas de programação. Esses dados são demonstrados na Tabela C1 do Apêndice C e escolheu-se a referida base devido atender os seguintes critérios: *i*) simplicidade das relações de dependência; *ii*) acessibilidade de arquivos por outros profissionais; e *iii*) eliminação da necessidade de pré-processamento, como eliminação de dados incorretos, eliminação de outliers, necessidades de gerenciamento de dados ausentes etc.

Os passos 2 e 3 referem-se a escolha e a criação de um modelo de predição em que o objetivo é prever, baseado em sua experiência (meses), se um funcionário terá sucesso na execução de uma tarefa de programação. Utilizou-se a técnica de regressão logística, devido a grande quantidade de aplicações encontradas na literatura (demonstrado na seção 5.1 e na Figura 4), pela simplicidade de modelagem e facilidade de uso de softwares estatísticos, e, de acordo com Hair *et al*. (2009), por ser uma ferramenta mais robusta do que a análise discriminatória para a suposição de normalidade multivariada.

A regressão Logística prevê a probabilidade de sucesso – P(E) - de um evento (variável dependente) em função de um vetor X de variáveis explicativas (variáveis independentes). No caso estudado, para cada valor da variável explicativa experiência, calcula-se uma probabilidade – P(E) - de sucesso. A Regressão logística é expressa pela Equação 36, e nesse caso possui apenas dois parâmetros, β_0 e β_1 estimados pelo método de máxima verossimilhança.

$$P(E) = \frac{1}{1+e^{-(\beta_0+\sum\beta_i X_i)}} \quad (36)$$

O software Statistica® (2007) foi utilizado para calcular os parâmetros do modelo (β_0 e β_1), e o teste de Wald aplicado a cada um deles. As informações detalhadas do software encontram-se no Apêndice G, e a Tabela 9 resume os dados principais. Para o cálculo dos parâmetros do modelo, o banco de dados foi utilizado por completo, já que este era bem reduzido.

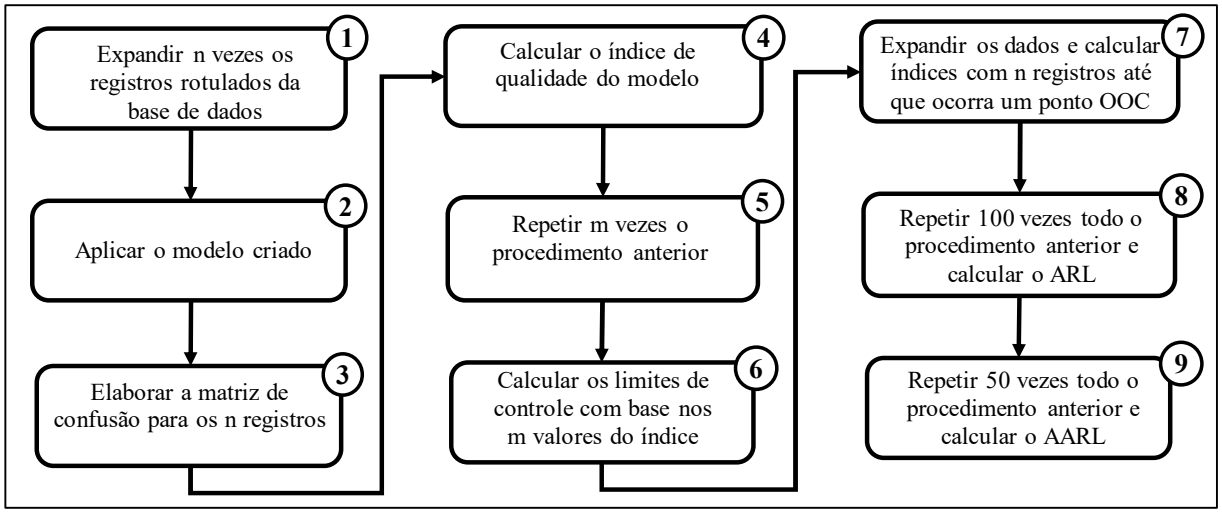
Tabela 9 - Informações dos parâmetros gerados

Elemento	média	Desvio padrão
Variável independente	18	8,3176
β_0	-3,0597	1.259592
β_1	0,1615	0,064995

Fonte: Próprio Autor

Na execução dos passos de 4 a 6 utilizou-se de ferramentas de programação para expandir o banco de dados e simular a aplicação das Fases I e II do SPM. Escolheu-se o software Python, devido sua maior simplicidade de programação, gratuidade e a crescente comunidade de especialistas usuários dessa ferramenta (o que facilita a troca de informações e resolução de problemas em algoritmos). Os códigos criados encontram-se nos Apêndices D a F, e a Figura 45 ilustra o Algoritmo geral utilizado.

Figura 45 - Algoritmo geral utilizado na programação em Phyton



Fonte: Próprio Autor

Para a expansão, exemplificada no Quadro 26, o campo mês de experiência recebe novos dados sorteando-os de acordo com uma normal de média 18 e desvio padrão 8,3176 (Tabela 9).

Para preencher o campo sucesso na realização da tarefa, aplica-se à Equação 36 o valor β_0 de -3,0597 (fixo); o valor mês de experiência criado (x); e o valor β_1 sorteado de acordo com a curva normal de média 0,1615 e desvio padrão 0,064995 * η, com η assumindo qualquer valor maior ou igual a zero, e que é definido em cada cenário de simulação. A ideia do uso do η é simular variabilidade nos resultados da regressão logística, se η=1, temos a variabilidade natural dos dados que a regressão está representando, se $\eta > 1$ temos uma variabilidade maior do que a esperada, e , obviamente, se $\eta < 1$ temos uma variabilidade menor do que a esperada.

Conforme já descrito, aplicados estes valores à Equação 36, a regressão logística resultará num valor entre 0 e 1, assim se P(E) $\geq$ 0,5, o campo será registrado como sucesso na execução da tarefa, e se P(E) < 0,5, será registrado como falha na execução da tarefa.

Este procedimento irá gerar as linhas de registro com dados rotulados, simulando dados reais. Na sequência, com base no valor de x (meses de experiência) sorteado, aplica-se novamente a Equação 36, mas nesse caso com β_0 = -3,0597 (fixo) e β_1 = 0,1615. Ou seja, aplica-se o modelo criado, onde valem as mesmas considerações a respeito de P(E), mas, neste caso, o campo que será preenchido é o campo de sucesso ou falha previsto.

Quadro 26 - Exemplo para a expansão de um registro

Expansão da Variável independente	**Expansão do campo rotulado que simula dados reais**	**Aplicação do modelo gerado**	**Análise para preenchimento da matriz de confusão**
Meses de Experiência (x)	Sucesso/Falha real na realização da tarefa (A)	Sucesso/Falha previsto na realização da tarefa (B)	
Valor sorteado de acordo com a normal de média 18 e desvio padrão 8.3176	Equação 36 com o valor de: β_0 de -3,0597 (fixo) mês de experiência criado (x) β_1 sorteado de acordo com a curva normal de média 0,1615 e desvio padrão 0,064995 * η. Exemplo $\beta_1 = 0{,}18$	Equação 36 como valor de: β_0 = -3,0597 (fixo) β_1 = 0,1615	Comparação entre A e B: A = S e B = S: TP A = S e B = F: FN A = F e B = F: TN A = F e B = S: FP
Exemplo x= 19	$P(E) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 * x)}}$ $= \frac{1}{1 + e^{-(-3{,}0597 + 0.18 * 19)}}$	$P(E)$ $= \frac{1}{1 + e^{-(-3{,}0597 + .1615 * 19)}}$	
	P(E) = 0,58911> 0,5, logo Sucesso	P(E) = 0,502199> 0,5, logo Sucesso	Análise resulta em TP

Fonte: Próprio Autor

Os cenários de simulação para *Kappa* são detalhados no Quadro 27, onde definem-se os valores das variáveis n, m, $\boldsymbol{\eta}$, VRL, VARL, VAARL e o Limiar.

Conforme a Figura 45, passos um e dois, os registros são expandidos seguindo o procedimento explicado no Quadro 26. A partir de n registros, executam-se os passos três e quatro, em que se cria a matriz de confusão e, a partir dela, calcula-se o índice sob análise (nesse caso, *Kappa, MCC ou Youden*). Os passos cinco e seis referem-se a Fase I do SPM, em que os passos um a quatro são repetidos m vezes criando-se m matrizes de confusão com n registros cada uma, possibilitando o cálculo de m valores do índice estudado. A média desses m valores é utilizada no cálculo da estimativa do índice e de seus limites de controle. O cálculo do desvio padrão será demonstrado na simulação de cada índice, pois é específico para cada estimativa de índice.

O passo sete refere-se à execução da Fase II do SPM. Uma vez definidos os limites de controle, é necessário iniciar o monitoramento, para isso o programa cria matrizes de confusão com n registros, e calcula os valores do índice estudado até que ocorra um ponto OOC. Por questão da programação, o limite de repetições (quantidade de índices calculados nesse loop) é limitado pela variável VRL, que no caso foi definida como 2500. Ou seja, se não ocorrer um ponto OOC até a repetição de número 2500, esse loop é interrompido e registra-se como RL o valor de 2500, se ocorrer antes, registra-se a repetição onde ocorreu o ponto OOC como o valor de RL. Programou-se a simulação para gerar 100 pontos nessa lista, definido através da variável VARL. A média desses 100 resultados é o valor ARL do limite de controle estimado. Todo este procedimento é repetido 50 vezes, definido pela variável VAARL. Resumindo, um limite de controle é calculado (a partir de m e n escolhidos), encontra-se o ARL deste limite de controle e registra-se, calculam-se outros 49 limites de controle e seus respectivos ARL´s. Ao fim tem-se uma lista com os ARL´s de 50 diferentes limites de controle, a média dos valores desta lista é a média dos ARL´s (em inglês Average Average Run length – AARL), ou o AARL.

Quadro 27 - Cenários de Simulação para *Kappa*

VRL =	**2500**		**VARL =**	**100**		**VAARL**	**50**		**Limiar =**	**0.5**	
n	**m**	**η**	**n**	**m**	**η**	**n**	**m**	**η**	**n**	**m**	**η**
100	1	1	300	1	1	500	1	1	1000	1	1
100	1	1.25	300	1	1.25	500	1	1.25	1000	1	1.25
100	1	1.5	300	1	1.5	500	1	1.5	1000	1	1.5
100	1	2	300	1	2	500	1	2	1000	1	2
100	1	3	300	1	3	500	1	3	1000	1	3
100	5	1	300	5	1	500	5	1	1000	5	1
100	5	1.25	300	5	1.25	500	5	1.25	1000	5	1.25
100	5	1.5	300	5	1.5	500	5	1.5	1000	5	1.5
100	5	2	300	5	2	500	5	2	1000	5	2
100	5	3	300	5	3	500	5	3	1000	5	3
100	10	1	300	10	1	500	10	1	1000	10	1
100	10	1.25	300	10	1.25	500	10	1.25	1000	10	1.25
100	10	1.5	300	10	1.5	500	10	1.5	1000	10	1.5
100	10	2	300	10	2	500	10	2	1000	10	2
100	10	3	300	10	3	500	10	3	1000	10	3
100	50	1	300	50	1	500	50	1	1000	50	1
100	50	1.25	300	50	1.25	500	50	1.25	1000	50	1.25
100	50	1.5	300	50	1.5	500	50	1.5	1000	50	1.5
100	50	2	300	50	2	500	50	2	1000	50	2
100	50	3	300	50	3	500	50	3	1000	50	3
100	1000	1	300	1000	1	500	1000	1	1000	1000	1
100	1000	1.25	300	1000	1.25	500	1000	1.25	1000	1000	1.25
100	1000	1.5	300	1000	1.5	500	1000	1.5	1000	1000	1.5
100	1000	2	300	1000	2	500	1000	2	1000	1000	2
100	1000	3	300	1000	3	500	1000	3	1000	1000	3

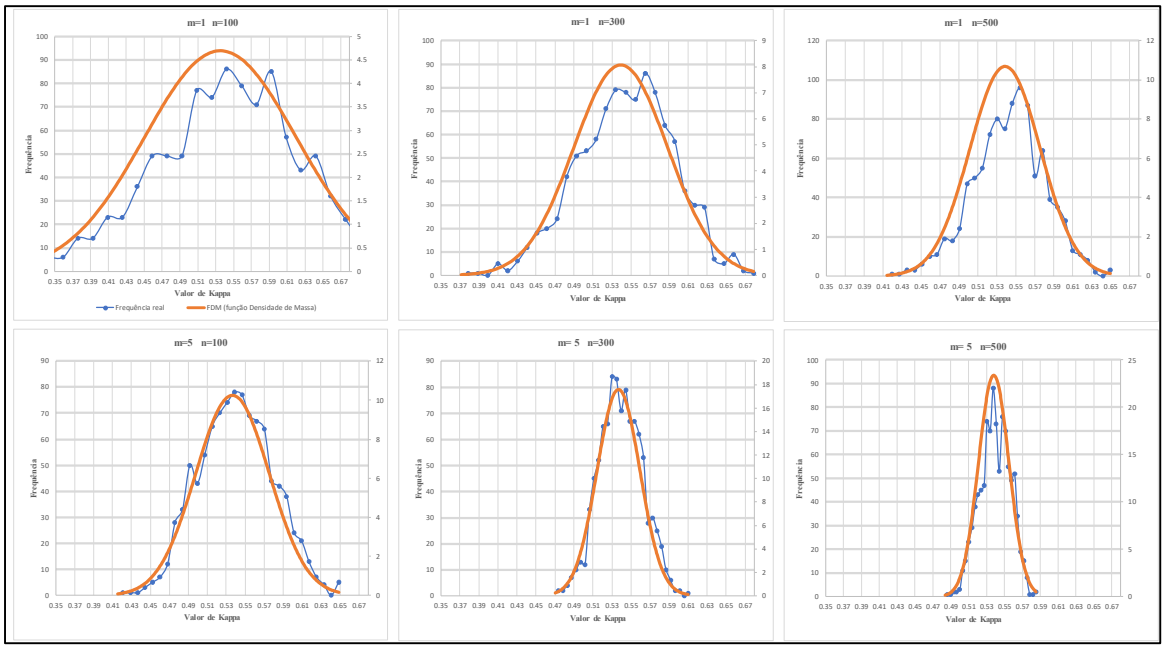
Fonte: Próprio Autor

7.3.1 Caso 1: Índice *Kappa* de Cohen

Elaborou-se a Figura 46 para demonstrar a normalidade dos valores de *Kappa* em função de m e n, com m assumindo valores de 1 ou 5 e n assumindo valores de 100, 300 ou 500. Para cada par m e n foram simulados 1000 valores de *Kappa.* O intervalo entre os valores extremos encontrados para *Kappa* foi dividido em 30 intervalos menores, e a frequência para cada um deles foi calculada. Calculou-se a média e o desvio padrão desses conjuntos de dados e elaborou-se a curva normal com estes parâmetros, sobrepondo-a aos dados de frequência. Assim demonstrou-se visualmente que os valores de *Kappa* seguem uma distribuição normal nos cenários simulados.

Outra análise realizada foi a checagem da distribuição dos valores de RL dos gráficos de controle baseados no *Kappa*. Na Figura 47 demonstra-se a frequência de ocorrência de 1000 desses valores, e a distribuição geométrica equivalente. Para a simulação desses RL´s, parametrizou-se o programa (anexo D) com VRL = 2500, VARL =1, VAARL = 1000, n = 300 e m =5. Nessa configuração de simulação são calculados 1000 limites de controle e um RL para cada limite de controle diferente, buscando simular a condição mais extrema de aleatoriedade. A distribuição dos RL´s comportou-se próximo a distribuição geométrica com $\widehat{ARL}$ = 273 e $\hat{\sigma}_{RL}$ = 333. Para n = 2000 e m =50 a distribuição dos RL´s comportou-se ainda mais próximo a distribuição geométrica, com $\widehat{ARL}$ = 361.71 e $\hat{\sigma}_{RL}$ = 367.80 (valores próximos ao teórico onde ARL= σ_{RL} = 370 em função da probabilidade de sucesso de 0,27%).

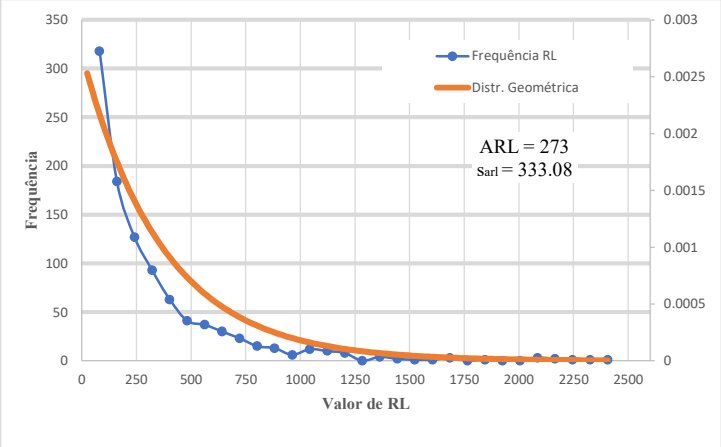
Figura 46 - Análise de normalidade de *Kappa* em função de m e n



Fonte: Próprio Autor

Em todas as simulações os limites de controle para *Kappa* foram calculados utilizando-se as Equações 7, 10, 12, 26 e 27 com $Z_{\alpha/2} = 3$.

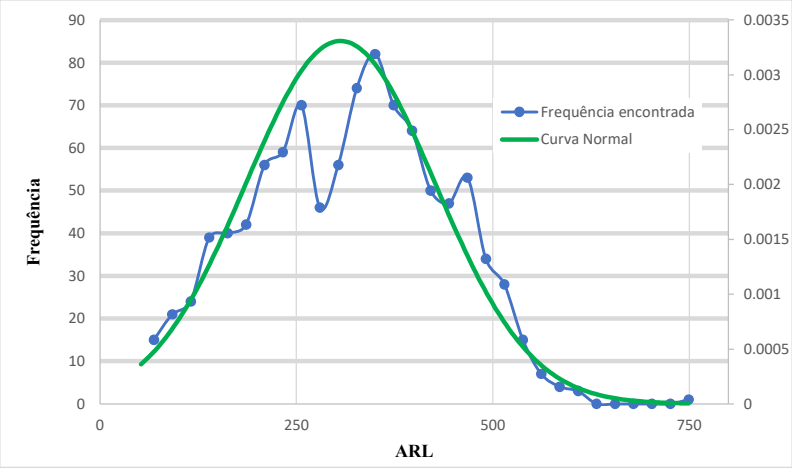
Figura 47 - Análise da distribuição dos RL´s - *Kappa*



Fonte: Próprio Autor

Também se analisou a distribuição dos valores de ARL. Para essa simulação parametrizou-se o programa com VRL = 2500, VARL =30, VAARL = 1000, n = 300 e m =10. Ou seja, para um limite de controle simulou-se trinta RL´s e calculou-se a sua média, encontrando-se o valor de seu ARL. Como VAARL=1000, esse procedimento foi executado mil vezes, resultando em 1000 ARL´s diferentes (um para cada limite de controle). A Figura 48 demonstra a distribuição de frequência destes 1000 ARL´s, e seu comportamento próximo a distribuição normal. Uma vez que existe um valor mínimo de 1 para o AARL, é esperado que a cauda à direita da curva normal seja maior do que à esquerda.

Figura 48 - Análise da distribuição do ARL´s - *Kappa*



Fonte: Próprio Autor

Verificadas essas condições, demonstram-se na Tabela 10 os resultados das simulações, executadas de acordo com os cenários definidos no Quadro 27. Nesta tabela, os valores indicados de *Kappa* médio referem-se a média dos *Kappas* médios calculados para os limites de controle (como VAARL=50, foram gerados cinquenta diferentes limites de controle). Os valores de RSP, *Po* e *Pe* seguem a mesma lógica e são as médias dos parâmetros utilizados nos cálculos dos valores de *Kappa* dos limites de controle. Para a estimativa do desvio padrão de *Kappa* aplicou-se a Equação 10, utilizando-se em *Po* e *Pe* a média dos m *Po's* e m *Pe's* utilizados nos cálculos de *Kappa* dos limites de controle.

Tabela 10 - Resultados de AARL nos cenários de simulação de *Kappa*

n	m	η	AARL	Kappa médio	*RSP*	*Po*	*Pe*
100	1	1	162.84	0.516	0.392	0.754	0.492
100	1	1.25	122.91	0.550	0.41	0.778	0.506
100	1	1.5	75.24	0.522	0.386	0.792	0.565
100	1	2	37.50	0.547	0.4026	0.774	0.501
100	1	3	4.65	0.541	0.4003	0.771	0.501
100	5	1	296.51	0.540	0.4028	0.773	0.507
100	5	1.25	111.21	0.542	0.3946	0.778	0.516
100	5	1.5	45.73	0.538	0.3956	0.778	0.520
100	5	2	8.77	0.526	0.3924	0.764	0.502
100	5	3	2.15	0.543	0.3912	0.779	0.517
100	10	1	343.44	0.546	0.3927	0.778	0.511
100	10	1.25	86.14	0.536	0.3990	0.770	0.504
100	10	1.5	31.78	0.531	0.3961	0.776	0.522
100	10	2	5.66	0.540	0.3952	0.781	0.523
100	10	3	2.26	0.533	0.3979	0.769	0.505
100	50	1	400.20	0.535	0.3971	0.773	0.512
100	50	1.25	73.96	0.538	0.3961	0.775	0.513
100	50	1.5	22.37	0.535	0.3962	0.770	0.506
100	50	2	5.64	0.536	0.3954	0.773	0.511
100	50	3	2.08	0.539	0.3973	0.776	0.514
100	1000	1	400.00	0.536	0.3967	0.774	0.512
100	1000	1.25	71.97	0.536	0.3964	0.773	0.512
100	1000	1.5	20.34	0.536	0.3969	0.774	0.512
100	1000	2	5.63	0.535	0.3964	0.774	0.513
100	1000	3	2.08	0.536	0.3963	0.773	0.512

n	m	η	AARL	Kappa médio	*RSP*	*Po*	*Pe*
300	1	1	144.47	0.533	0.3934	0.781	0.530
300	1	1.25	90.45	0.527	0.3959	0.766	0.505
300	1	1.5	16.22	0.537	0.4043	0.778	0.520
300	1	2	2.25	0.539	0.3949	0.773	0.508
300	1	3	1.15	0.536	0.4001	0.769	0.502
300	5	1	243.66	0.540	0.3974	0.774	0.508
300	5	1.25	42.19	0.539	0.3962	0.774	0.510
300	5	1.5	6.33	0.540	0.3969	0.772	0.505
300	5	2	1.77	0.533	0.3954	0.771	0.511
300	5	3	1.04	0.537	0.3964	0.772	0.508
300	10	1	312.60	0.541	0.3974	0.775	0.510
300	10	1.25	33.48	0.537	0.3959	0.772	0.507
300	10	1.5	5.36	0.539	0.3984	0.773	0.509
300	10	2	1.57	0.535	0.3992	0.773	0.511
300	10	3	1.03	0.541	0.3980	0.776	0.512
300	50	1	368.99	0.538	0.3962	0.775	0.512
300	50	1.25	27.18	0.537	0.3956	0.774	0.511
300	50	1.5	5.19	0.537	0.3964	0.774	0.511
300	50	2	1.50	0.539	0.3968	0.773	0.508
300	50	3	1.03	0.537	0.3970	0.773	0.510
300	1000	1	402.59	0.537	0.3968	0.775	0.514
300	1000	1.25	24.49	0.538	0.3965	0.774	0.511
300	1000	1.5	5.02	0.537	0.3965	0.774	0.511
300	1000	2	1.51	0.538	0.3966	0.774	0.511
300	1000	3	1.03	0.537	0.3963	0.774	0.511

n	m	η	AARL	Kappa médio	*RSP*	*Po*	*Pe*
500	1	1	171.56	0.539	0.39852	0.776	0.514
500	1	1.25	62.82	0.536	0.39328	0.768	0.501
500	1	1.5	14.78	0.530	0.39608	0.774	0.519
500	1	2	1.48	0.538	0.3956	0.772	0.507
500	1	3	1.01	0.546	0.39628	0.775	0.505
500	5	1	289.47	0.540	0.39759	0.798	0.561
500	5	1.25	21.33	0.539	0.39728	0.774	0.510
500	5	1.5	3.14	0.537	0.39722	0.774	0.511
500	5	2	1.18	0.532	0.39562	0.771	0.510
500	5	3	1.00	0.537	0.3979	0.772	0.509
500	10	1	294.71	0.537	0.3972	0.773	0.510
500	10	1.25	15.49	0.538	0.3963	0.775	0.512
500	10	1.5	3.02	0.535	0.3963	0.772	0.511
500	10	2	1.13	0.535	0.3960	0.773	0.512
500	10	3	1.00	0.538	0.3975	0.774	0.511
500	50	1	365.27	0.539	0.3966	0.774	0.511
500	50	1.25	13.99	0.536	0.3958	0.773	0.511
500	50	1.5	2.66	0.537	0.3972	0.774	0.512
500	50	2	1.10	0.538	0.3966	0.773	0.509
500	50	3	1.00	0.537	0.3962	0.773	0.510
500	1000	1	396.01	0.538	0.3965	0.774	0.510
500	1000	1.25	1.00	0.537	0.3962	0.773	0.510
500	1000	1.5	2.50	0.538	0.3966	0.774	0.510
500	1000	2	1.09	0.538	0.3966	0.774	0.510
500	1000	3	1.00	0.538	0.3965	0.774	0.511

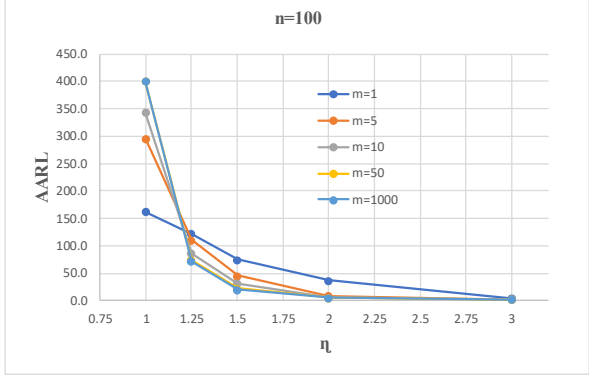
n	m	η	AARL	Kappa médio	*RSP*	*Po*	*Pe*
1000	1	1	150.84	0.532	0.3945	0.771	0.510
1000	1	1.25	24.30	0.538	0.39632	0.775	0.512
1000	1	1.5	2.91	0.536	0.39956	0.772	0.508
1000	1	2	1.00	0.542	0.39794	0.775	0.508
1000	1	3	1.00	0.541	0.39538	0.771	0.501
1000	5	1	254.77	0.539	0.39526	0.774	0.510
1000	5	1.25	8.48	0.534	0.39754	0.772	0.511
1000	5	1.5	1.34	0.538	0.39711	0.774	0.510
1000	5	2	1.00	0.538	0.39657	0.773	0.509
1000	5	3	1.00	0.536	0.3981	0.772	0.508
1000	10	1	319.19	0.538	0.3978	0.773	0.510
1000	10	1.25	4.66	0.540	0.3973	0.774	0.509
1000	10	1.5	1.28	0.539	0.3967	0.775	0.513
1000	10	2	1.00	0.538	0.3958	0.774	0.510
1000	10	3	1.00	0.539	0.3947	0.776	0.513
1000	50	1	381.57	0.537	0.3967	0.773	0.510
1000	50	1.25	4.97	0.538	0.3964	0.774	0.511
1000	50	1.5	1.28	0.538	0.3969	0.774	0.510
1000	50	2	1.00	0.538	0.3966	0.774	0.510
1000	50	3	1.00	0.538	0.3962	0.774	0.510
1000	1000	1	408.58	0.538	0.3968	0.774	0.510
1000	1000	1.25	4.68	0.538	0.3966	0.774	0.510
1000	1000	1.5	1.26	0.538	0.3965	0.774	0.510
1000	1000	2	1.00	0.538	0.3967	0.774	0.510
1000	1000	3	1.00	0.538	0.3966	0.774	0.510

Fonte: Próprio Autor

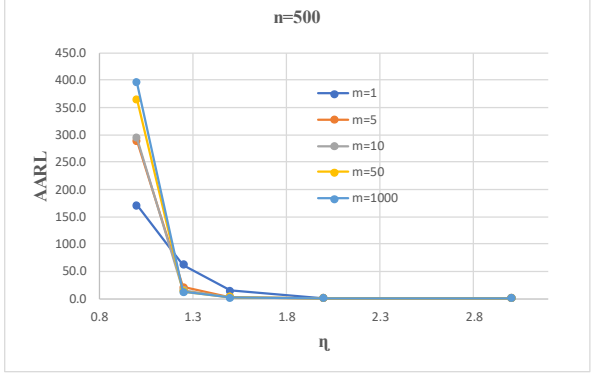
De forma a facilitar a visualização dos dados, foram elaboradas diversas Figuras. Na Figura 49 cada gráfico é fixo para um determinado valor de n, enquanto os valores de AARL são demonstrados em função de m e $\boldsymbol{\eta}$. Na Figura 50 cada gráfico é fixo para um determinado valor de $\boldsymbol{\eta}$, enquanto os valores de AARL são demonstrados em função de m e n. Na Figura 51 cada gráfico é fixo para um determinado valor de m, enquanto os valores de AARL são

demonstrados em função de n e η.

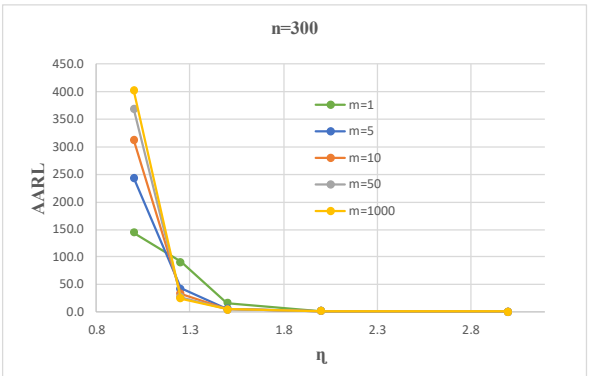
Figura 49 - AARL para n fixo e m e η variáveis



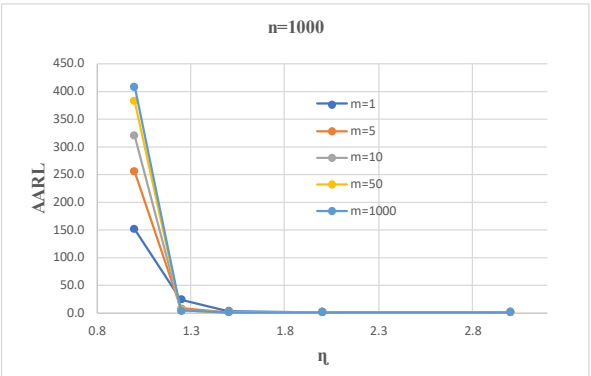
η	m=1	m=5	m=10	m=50	m=1000
1	162.84	296.51	343.44	400.20	400.00
1.25	122.91	111.21	86.14	73.96	71.97
1.5	75.24	45.73	31.78	22.37	20.34
2	37.50	8.77	5.66	5.64	5.63
3	4.65	2.15	2.26	2.08	2.08



η	m=1	m=5	m=10	m=50	m=1000
1	171.56	289.47	294.71	365.27	396.01
1.25	62.82	21.33	15.49	13.99	12.75
1.5	14.78	3.14	3.02	2.66	2.50
2	1.48	1.18	1.13	1.10	1.09
3	1.01	1.00	1.00	1.00	1.00



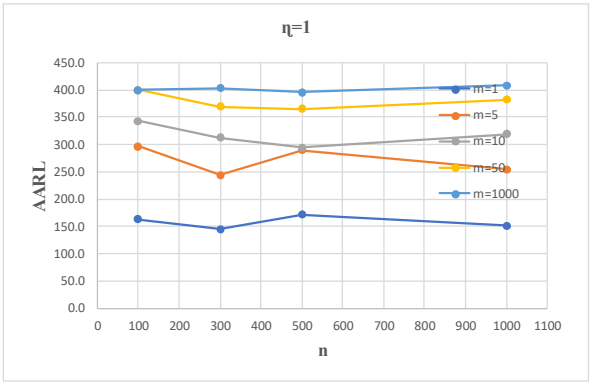
η	m=1	m=5	m=10	m=50	m=1000
1	144.47	243.66	312.60	368.99	402.59
1.25	90.45	42.19	33.48	27.18	24.49
1.5	16.22	6.33	5.36	5.19	5.02
2	2.25	1.77	1.57	1.50	1.51
3	1.15	1.04	1.03	1.03	1.03



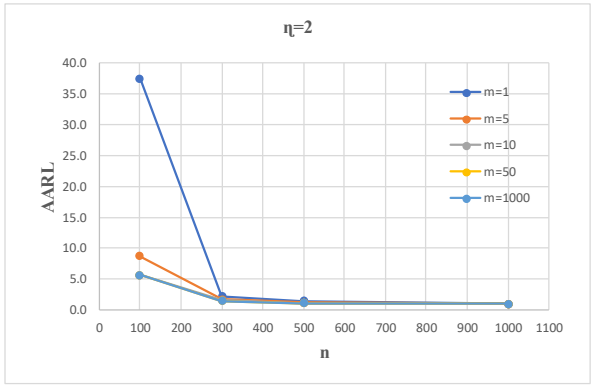
η	m=1	m=5	m=10	m=50	m=1000
1	150.84	254.77	319.19	381.57	408.58
1.25	24.30	8.48	4.66	4.97	4.68
1.5	2.91	1.34	1.28	1.28	1.26
2	1.00	1.00	1.00	1.00	1.00
3	1.00	1.00	1.00	1.00	1.00

Fonte: Próprio Autor

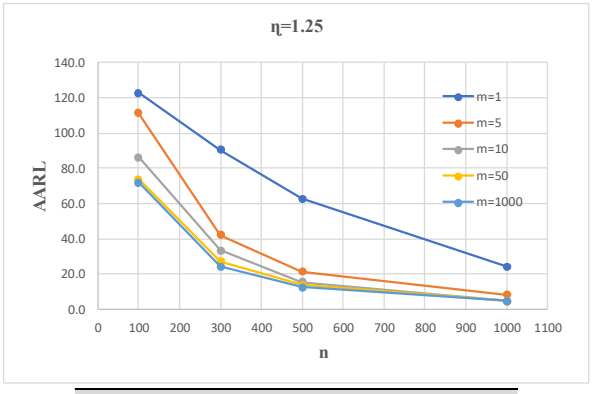
Figura 50 - AARL para η fixo e m e n variáveis



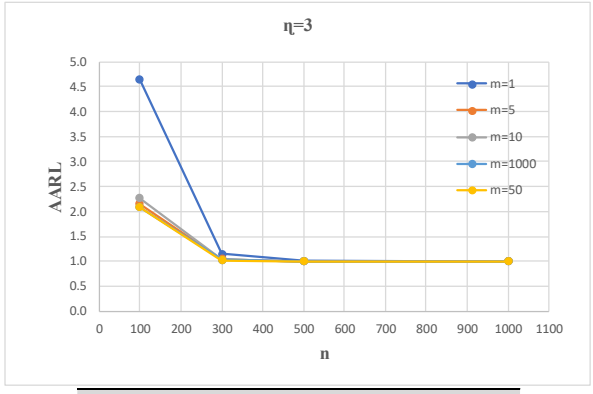
n	m=1	m=5	m=10	m=50	m=1000
100	162.84	296.51	343.44	400.20	400.00
300	144.47	243.66	312.60	368.99	402.59
500	171.56	289.47	294.71	365.27	396.01
1000	150.84	254.77	319.19	381.57	408.58



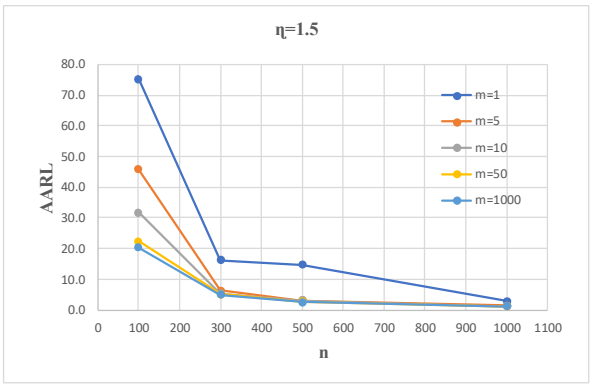
n	m=1	m=5	m=10	m=50	m=1000
100	37.50	8.77	5.66	5.64	5.63
300	2.25	1.77	1.57	1.50	1.51
500	1.48	1.18	1.13	1.10	1.09
1000	1.00	1.00	1.00	1.00	1.00



n	m=1	m=5	m=10	m=50	m=1000
100	122.91	111.21	86.14	73.96	71.97
300	90.45	42.19	33.48	27.18	24.49
500	62.82	21.33	15.49	13.99	12.75
1000	24.30	8.48	4.66	4.97	4.68



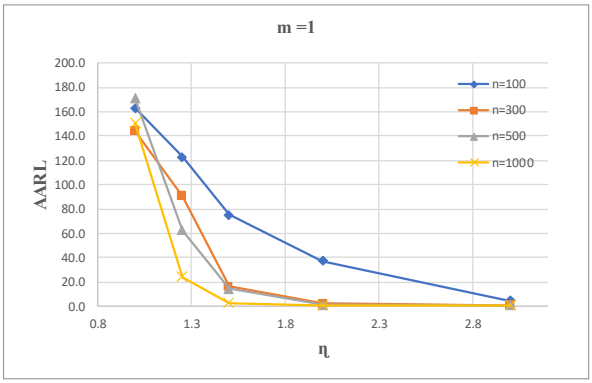
n	m=1	m=5	m=10	m=50	m=1000
100	4.65	2.15	2.26	2.08	2.08
300	1.15	1.04	1.03	1.03	1.03
500	1.01	1.00	1.00	1.00	1.00
1000	1.00	1.00	1.00	1.00	1.00



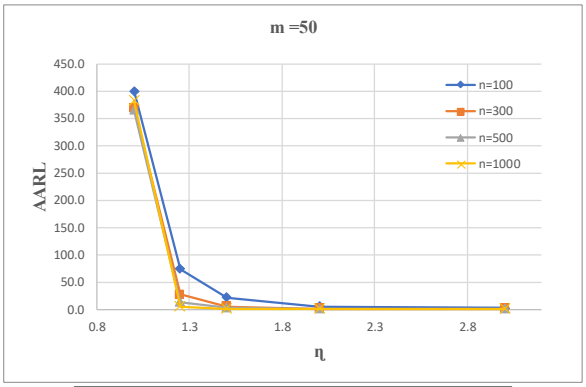
n	m=1	m=5	m=10	m=50	m=1000
100	75.24	45.73	31.78	22.37	20.34
300	16.22	6.33	5.36	5.19	5.02
500	14.78	3.14	3.02	2.66	2.50
1000	2.91	1.34	1.28	1.28	1.26

Fonte: Próprio Autor

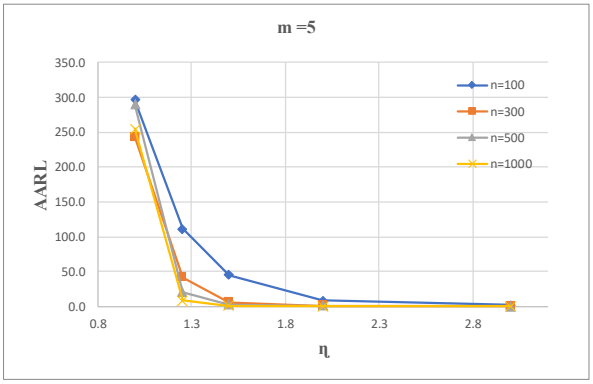
Figura 51 - AARL para m fixo e η e n variáveis



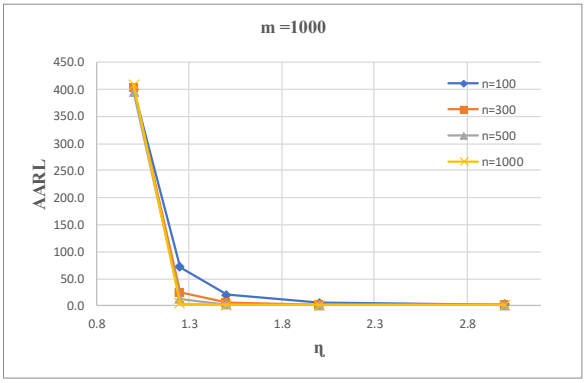
η	n=100	n=300	n=500	n=1000
1	162.84	144.47	171.56	150.84
1.25	122.91	90.45	62.82	24.30
1.5	75.24	16.22	14.78	2.91
2	37.50	2.25	1.48	1.00
3	4.65	1.15	1.01	1.00



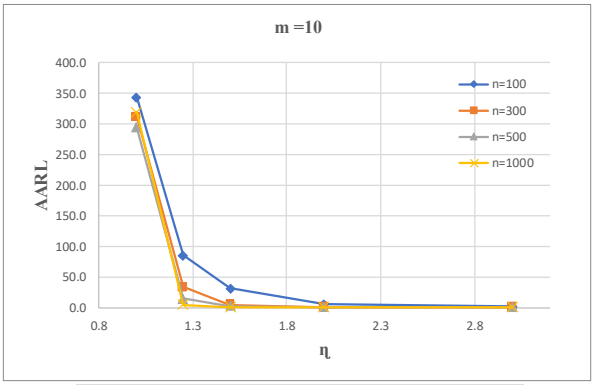
η	n=100	n=300	n=500	n=1000
1	400.20	368.99	365.27	381.57
1.25	73.96	27.18	13.99	4.97
1.5	22.37	5.19	2.66	1.28
2	5.64	1.50	1.10	1.00
3	2.08	1.03	1.00	1.00



η	n=100	n=300	n=500	n=1000
1	296.51	243.66	289.47	254.77
1.25	111.21	42.19	21.33	8.48
1.5	45.73	6.33	3.14	1.34
2	8.77	1.77	1.18	1.00
3	2.15	1.04	1.00	1.00



η	n=100	n=300	n=500	n=1000
1	400.00	402.59	396.01	408.58
1.25	71.97	24.49	12.75	4.68
1.5	20.34	5.02	2.50	1.26
2	5.63	1.51	1.09	1.00
3	2.08	1.03	1.00	1.00



η	n=100	n=300	n=500	n=1000
1	343.44	312.60	294.71	319.19
1.25	86.14	33.48	15.49	4.66
1.5	31.78	5.36	3.02	1.28
2	5.66	1.57	1.13	1.00
3	2.26	1.03	1.00	1.00

Fonte: Próprio Autor

Observa-se, que conforme m aumenta, os valores de AARL sob controle (AARLo ou AARL quando $\boldsymbol{\eta = 1}$) se aproximam de 400 e não 370, fenômeno demonstrado por Chen e Song (2012) para gráficos de p.

Aproximando-se a condição de processo sob controle para $\boldsymbol{\delta}$ = 0.005, verifica-se na Tabela 4 que AARLo = 140.1 para m=1 (e parâmetros desconhecidos). Com o apoio das Figuras percebe-se uma convergência para estes valores conforme n aumenta.

Analisando a Figura 50 para $\boldsymbol{\eta}$ = 1 (condição sob controle), verifica-se que para n = 1000, os valores de AARL aproximam-se dos valores encontrados na Tabela 4 em função de m (m=1, AARL=140; m=5, AARL= 237; m=10, AARL = 277; m=1000, AARL= 366).

Elaborou-se a Tabela 11 para relacionar cenários a valores de $\boldsymbol{\delta}$. Assim, foram simulados 1000 pontos fora de controle com as condições de n, m e $\boldsymbol{\eta}$ indicadas nessa tabela. As médias foram registradas no campo *Kappa* médio OOC possibilitando o cálculo de $\boldsymbol{\delta}$ em função da variação em $\boldsymbol{\eta}$. O conteúdo de cada campo é descrito no Quadro 28. Percebe-se pelos dados levantados que o AARL teórico ficou próximo ao AARL simulado.

Quadro 28 - Descrição dos Campos da Tabela 11

Kappa médio	*Kappa* médio encontrado para os limites de controle (Tabela 10)
Kappa médio OOC	*Kappa* médio dos mil Kappas calculados em condições fora de controle ($\boldsymbol{\eta > 1}$)
$\boldsymbol{\delta}$	*Kappa* médio - *Kappa* médio OOC
AARL Teórico	Interpolação do AARL teórico conforme demonstrado na Figura 52
AARL Simulado	Valores encontrados para os cenários, conforme Tabela 10
RSP simulado OOC	Valores encontrados para o RSP simulado nas condições fora de controle
Po simulado OOC	Média de *Po* para a situação fora de controle
Pe simulado OOC	Média de *Pe* para a situação fora de controle
RSP teórico	Tabela 4
Po Teórico	Tabela 4
Pe Teórico	Tabela 4

Fonte: Próprio Autor

Tabela 11- Análise de AARL em função de δ

análise	n	m	η	Kappa médio	Kappa médio OOC	δ	AARL teórico	AARL simulado	*RSP simulado OOC*	*Po Simulado OOC*	*Pe Simulado OOC*	*RSP Teórico*	*Po Teórico*	*Pe Teórico*
1	100	1	1.25	0.550	0.4743	0.076	109.99	122.91	0.395	0.745	0.514	0.400	0.770	0.512
2	300	1	1.25	0.527	0.4773	0.050	106.20	90.45	0.396	0.749	0.521	0.400	0.780	0.512
3	500	1	1.25	0.54	0.4766	0.060	80.51	62.82	0.398	0.745	0.512	0.400	0.780	0.512
4	100	1	1.5	0.522	0.4323	0.090	100.96	75.24	0.399	0.729	0.522	0.400	0.770	0.512
5	500	1	1.5	0.530	0.4349	0.095	31.08	14.78	0.397	0.724	0.512	0.400	0.780	0.512
6	500	1	2	0.538	0.3652	0.173	3.60	1.48	0.402	0.686	0.506	0.400	0.780	0.500
7	500	1	3	0.546	0.2842	0.262	1	1.01	0.414	0.648	0.508	0.400	0.780	0.512

Fonte: Próprio Autor

Figura 52 - Demonstração dos pontos interpolados

n=100 AARL m=1

Pe	**δ =0.4**	**δ =0.25**	**δ =0.15**	**δ =0.05**	**δ =0.01**	**δ =0.005**
0.5	1,852	14.83	59.43	127.1	139.6	140
0.512	2.02	16.31	61.84	127.7	139.6	140
0.568	3,382	25.5	73.56	130.3	139.7	140
0.668	11.55	49.33	95.49	134.2	139.9	140
0.812	59.17	99.54	123.8	138.2	140.0	140.1
0.905	112	128.3	135.7	139.6	140.1	140.1
0.980	138.7	139.6	139.9	140.1	140.1	140.1

1;4

n=300 AARL m=1

Pe	**δ =0.4**	**δ =0.25**	**δ =0.15**	**δ =0.05**	**δ =0.01**	**δ =0.005**
0.5	1,003	1,464	12.68	104.8	138.5	139.7
0.512	1,004	1,561	14.01	106.2	138.5	139.7
0.568	1,022	2,385	22.06	112.7	138.9	139.8
0.668	1,285	8,107	45.52	123.2	139.4	139.9
0.812	12.54	51.26	96.88	134.4	139.9	140.0
0.902	69.15	106	126.7	138.5	140	140.1
0.980	136.1	138.5	139.5	140	140.1	140.1

2

n=500 AARL m=1

Pe	**δ =0.4**	**δ =0.25**	**δ =0.15**	**δ =0.05**	**δ =0.01**	**δ =0.005**
0.5	1	1,047	3,892	86.54	137.4	139.4
0.512	1	1,061	4.37	88.52	137.5	139.5
0.568	1	1,189	7,817	97.67	138.1	139.6
0.668	1,024	2,464	22.66	113.1	138.9	139.8
0.812	3,843	27.28	76.07	130.8	139.7	140.1
0.901	43.34	87.7	118.2	137.5	140	140.1
0.980	133.4	137.5	139.1	140	140	140.1

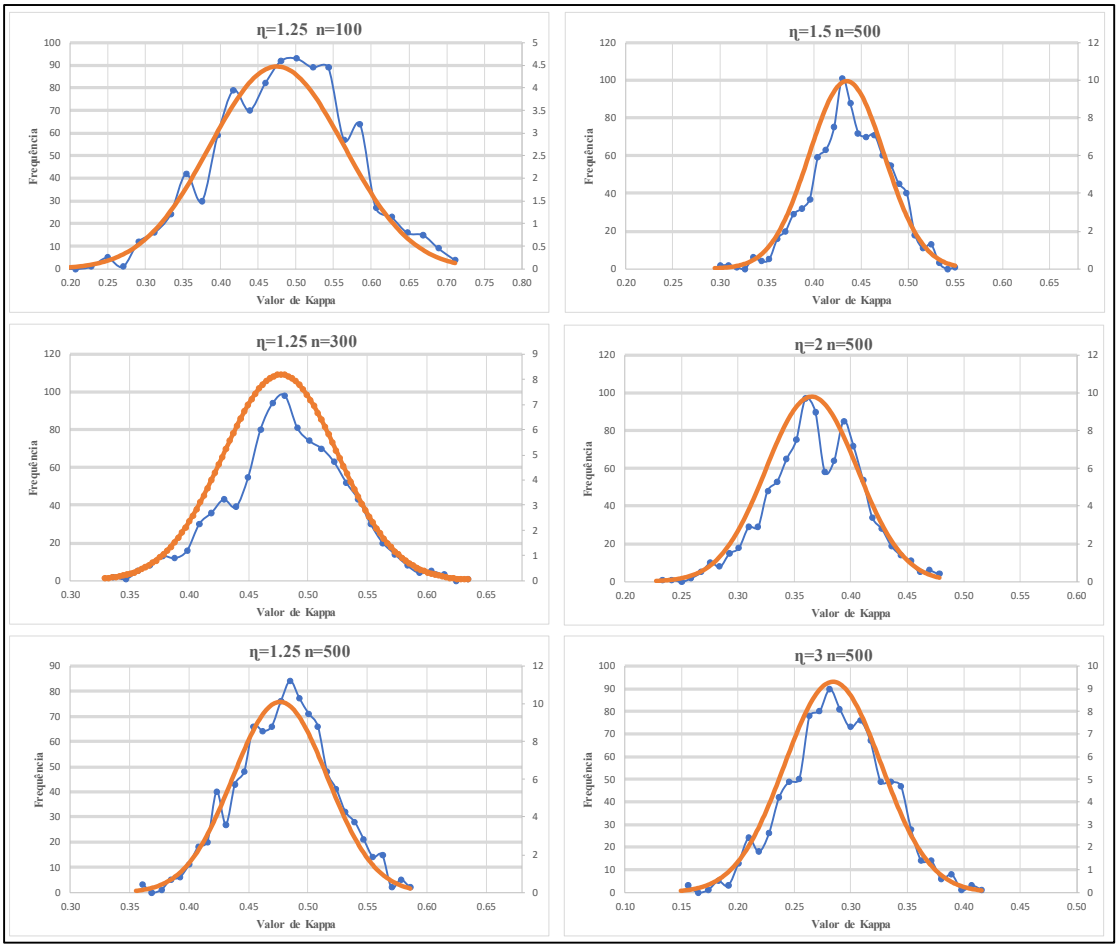
3;6;7

Fonte: Próprio Autor

A Figura 51 auxilia na análise dos processos fora de controle (AARLooc), pois cada gráfico está em função de $\boldsymbol{\eta}$ e n, para determinado m. Se m = 1 e $\boldsymbol{\eta} = 1{,}25$, os valores de AARLooc são muito altos, mesmo para n = 1000, AARL = 24,30. Este valor é 8.48 para m=5 e n=1000. Seria necessário m e n =1000 para que AARL fosse menor que 5. O que está em acordo com os valores teóricos calculados, que indicam a necessidade de m e n grandes quando o valor de $\boldsymbol{\delta}$ (simulado por valores de $\boldsymbol{\eta}$ maiores que 1) é menor que 0,05 (principalmente para *Pe* maior que 0.8). Entretanto para $\boldsymbol{\eta} = 1{,}5$, m = 5 e n=300, AARL = 6,33 enquanto para $\boldsymbol{\eta} = 1{,}5$, m =1, n=1000, AARL seria 2,91, alcançando uma sensibilidade melhor, com a necessidade de menos registros. Em linhas gerais para $\boldsymbol{\eta} \geq 1.5$ ($\boldsymbol{\delta} \approx 0.1$) , m = 5 e n=300 são tamanhos de amostra aplicáveis nas condições simuladas, para a elaboração da Fase I do SPM.

Utilizando-se os mesmos procedimentos aplicados para a Figura 46, elaborou-se a Figura 53 demonstrando a normalidade dos valores de *Kappa* simulados para condições fora de controle ($\boldsymbol{\eta} \geq 1$), e diferentes tamanhos de amostra n. Como, em teoria, referem-se aos pontos que serão plotados nos gráficos de controle, utilizou-se m=1.

Figura 53 - Análise de normalidade de *Kappa* OOC



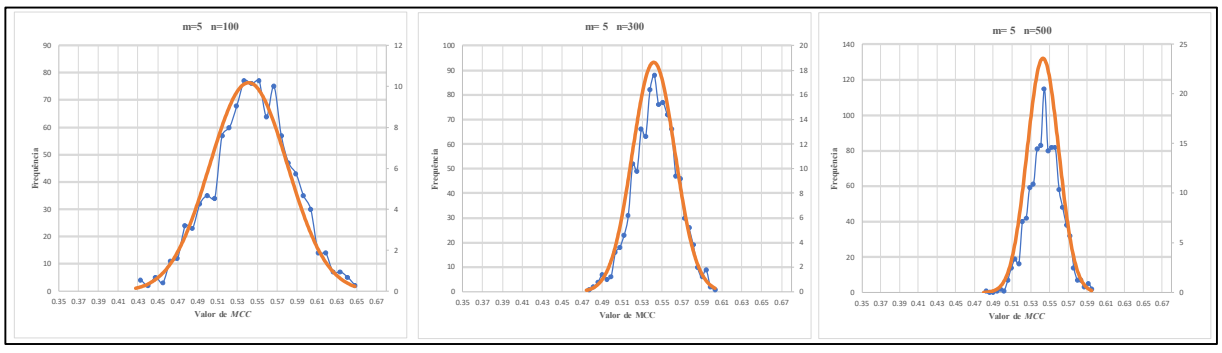
Fonte: Próprio Autor

7.3.2 Caso 2: Coeficiente de Correlação de Matthews

Semelhante às análises feitas para *Kappa*, elaborou-se a Figura 54 para demonstrar a normalidade dos valores de *MCC* em função de m e n. Com m = 5 e n assumindo valores de 100, 300 ou 500, simulou-se 1000 valores de *MCC* para cada par m e n. Em cada um desses conjuntos de 1000 dados, encontrou-se seus valores extremos e dividiu-se este intervalo em 30 intervalos menores, calculando-se suas frequências de ocorrência. Calculou-se a média, e o desvio padrão, para cada conjunto, e elaborou-se a curva normal com estes parâmetros, sobrepondo-a aos dados de frequência e demonstrando visualmente que os valores de *MCC* seguem essa distribuição nos cenários simulados.

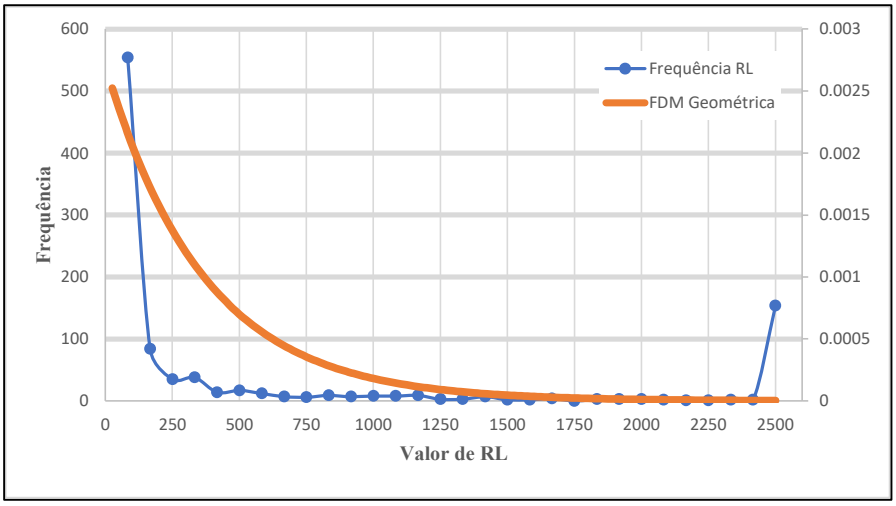
Outra análise realizada foi a checagem da distribuição dos valores de RL dos gráficos de controle baseados no *MCC*. Na Figura 55 demonstra-se a frequência de ocorrência de 1000 desses valores, e sua distribuição geométrica equivalente. Nessa simulação do cálculo dos RL´s, parametrizou-se o programa (anexo E) com VRL = 2500, VARL =1, VAARL = 1000, n = 1500 e m =5. Da mesma forma que na análise feita para o *Kappa*, cada RL é calculado para um limite de controle diferente, condição mais extrema de aleatoriedade. A distribuição dos RL´s comportou-se próximo a distribuição geométrica, entretanto identificou-se um comportamento anormal para a frequência dos valores mais altos de RL. O esperado era encontrar poucas ocorrências para o RL de 2500, entretanto encontrou-se 15.4% dos valores de RL nessa faixa. Cabe lembrar que o programa utilizado é truncado no valor máximo de 2500 (VRL=2500), o que significa que estes valores poderiam ser ainda maiores. Além disso, e a distribuição dos RL´s comportou-se com $\widehat{ARL}$ = 556,59 e $\hat{\sigma}_{RL}$ = 909,54 (valores afastados do teórico onde ARL= σ = 370).

Figura 54 - Análise de normalidade de *MCC* em função de m e n



Fonte: Próprio Autor

Figura 55 - Análise da distribuição dos RL´s - MCC



Fonte: Próprio Autor

Verificada a condição de normalidade, demonstram-se na Tabela 12 os resultados das simulações, onde os valores de AARL são calculados de acordo com os cenários definidos no Quadro 29. Na tabela 12, os valores indicados de *MCC* médio referem-se a média dos *MCC´s* médios utilizados para o cálculo dos limites de controle. Aqui cabe uma observação com relação ao cálculo do desvio padrão do *MCC*, este é calculado de acordo com a Equação 37, motivo pelo qual os cenários de MCC utilizam m ≥ 5.

$$s = \sqrt{\frac{\sum_1^m (MCC_i - \overline{MCC})^2}{(m-1)}} \qquad \text{EQ. (37)}$$

Onde,

$$\overline{MCC} = \frac{\sum_1^m MCC_i}{m}$$

Quadro 29 - Cenários de Simulação para *MCC*

VRL =	**2500**	**VARL = 100**	**VAARL = 50**	**Limiar =**	**0.5**

n	**m**	**η**
100	5	1
100	5	1.25
100	5	2
100	10	1
100	10	1.25
100	10	2
100	1000	1
100	1000	1.25
100	1000	2

n	**m**	**η**
300	5	1
300	5	1.25
300	5	2
300	10	1
300	10	1.25
300	10	2
300	1000	1
300	1000	1.25
300	1000	2

n	**m**	**η**
1000	5	1
1000	5	1.25
1000	5	2
1000	10	1
1000	10	1.25
1000	10	2
1000	1000	1
1000	1000	1.25
1000	1000	2

Fonte: Próprio Autor

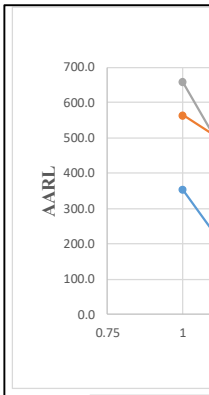
Tabela 12 - Resultados de AARL nos cenários de simulação de *MCC*

n	m	η	AARL	MCC médio	*RSP*
100	5	1	565.52	0.556	0.400
100	5	1.25	430.08	0.545	0.398
100	5	2	13.63	0.543	0.3941
100	10	1	660.78	0.544	0.4001
100	10	1.25	295.94	0.551	0.3971
100	10	2	11.17	0.538	0.3973
100	1000	1	355.15	0.542	0.3963
100	1000	1.25	61.68	0.541	0.3966
100	1000	2	5.27	0.541	0.3965
n	**m**	**η**	**AARL**	**MCC médio**	***RSP***
300	5	1	559.11	0.539	0.3974
300	5	1.25	227.98	0.545	0.3964
300	5	2	2.83	0.541	0.3939
300	10	1	581.02	0.547	0.3989
300	10	1.25	52.76	0.541	0.3948
300	10	2	1.57	0.542	0.3957
300	1000	1	388.61	0.542	0.3965
300	1000	1.25	22.55	0.542	0.3966
300	1000	2	1.47	0.542	0.3966
n	**m**	**η**	**AARL**	**MCC médio**	***RSP***
1000	5	1	551.66	0.544	0.3970
1000	5	1.25	71.55	0.543	0.3965
1000	5	2	1.02	0.541	0.3967
1000	10	1	448.14	0.540	0.3965
1000	10	1.25	29.92	0.540	0.3961
1000	10	2	1.00	0.540	0.3977
1000	1000	1	384.99	0.542	0.3965
1000	1000	1.25	4.43	0.542	0.3966
1000	1000	2	1.01	0.542	0.3967

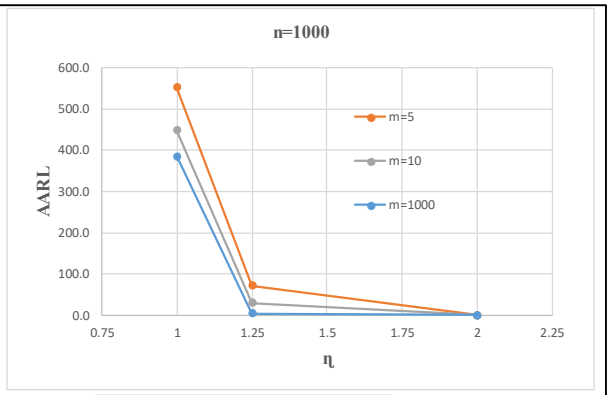
Fonte: Próprio Autor

De maneira semelhante a análise do *Kappa* elaborou-se a Figura 56 em que cada gráfico é fixo para um determinado valor de n, enquanto os valores de AARL são demonstrados em função de m e $\boldsymbol{\eta}$. Na Figura 57 cada gráfico é fixo para um determinado valor de $\boldsymbol{\eta}$, enquanto os valores de AARL são demonstrados em função de m e n. Na Figura 58 cada gráfico é fixo para um determinado valor de m, enquanto os valores de AARL são demonstrados em função de n e $\boldsymbol{\eta}$.

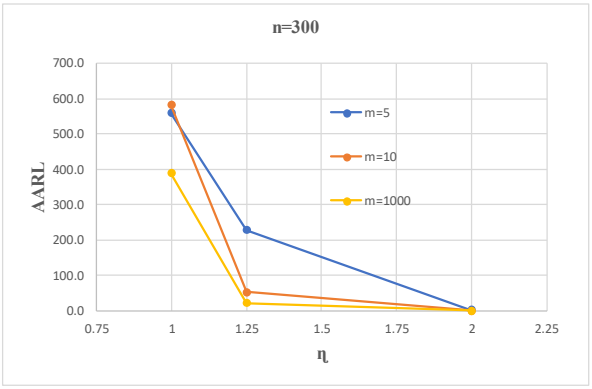
Figura 56 - AARL para n fixo e m e η variáveis



η	m=5	m=10	m=1000
1	565.52	660.78	355.15
1.25	430.08	295.94	61.68
2	13.63	11.17	5.27



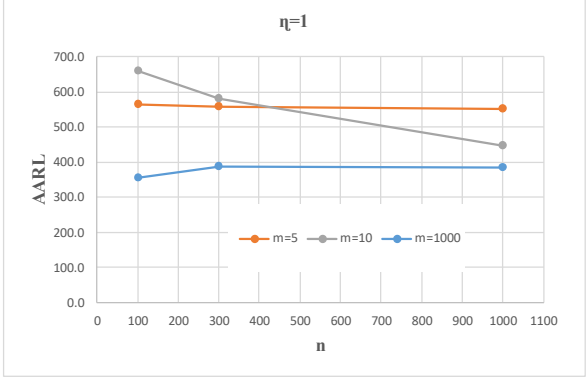
η	m=5	m=10	m=1000
1	551.66	448.14	384.99
1.25	71.55	29.92	4.43
2	1.02	1.00	1.01



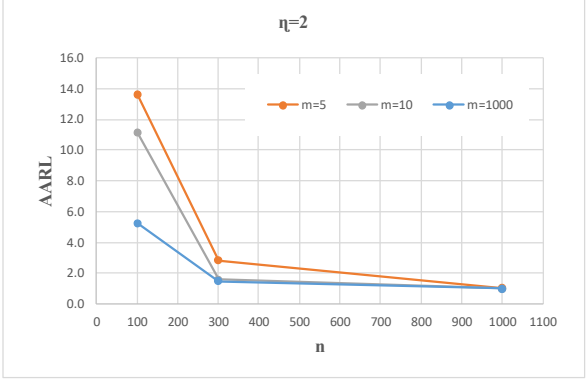
η	m=5	m=10	m=1000
1	559.11	581.02	388.61
1.25	227.98	52.76	22.55
2	2.83	1.57	1.47

Fonte: Próprio Autor

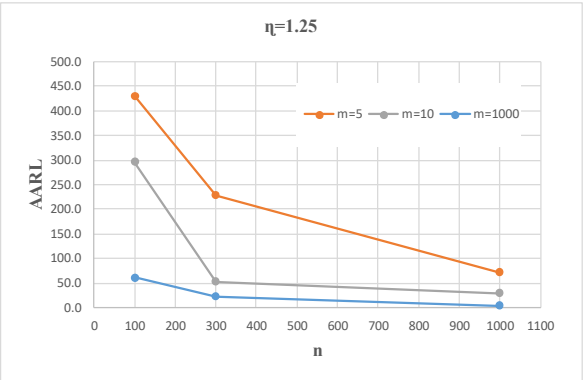
Figura 57 - AARL para η fixo e m e n variável



n	m=5	m=10	m=1000
100	565.52	660.78	355.15
300	559.11	581.02	388.61
1000	551.66	448.14	384.99



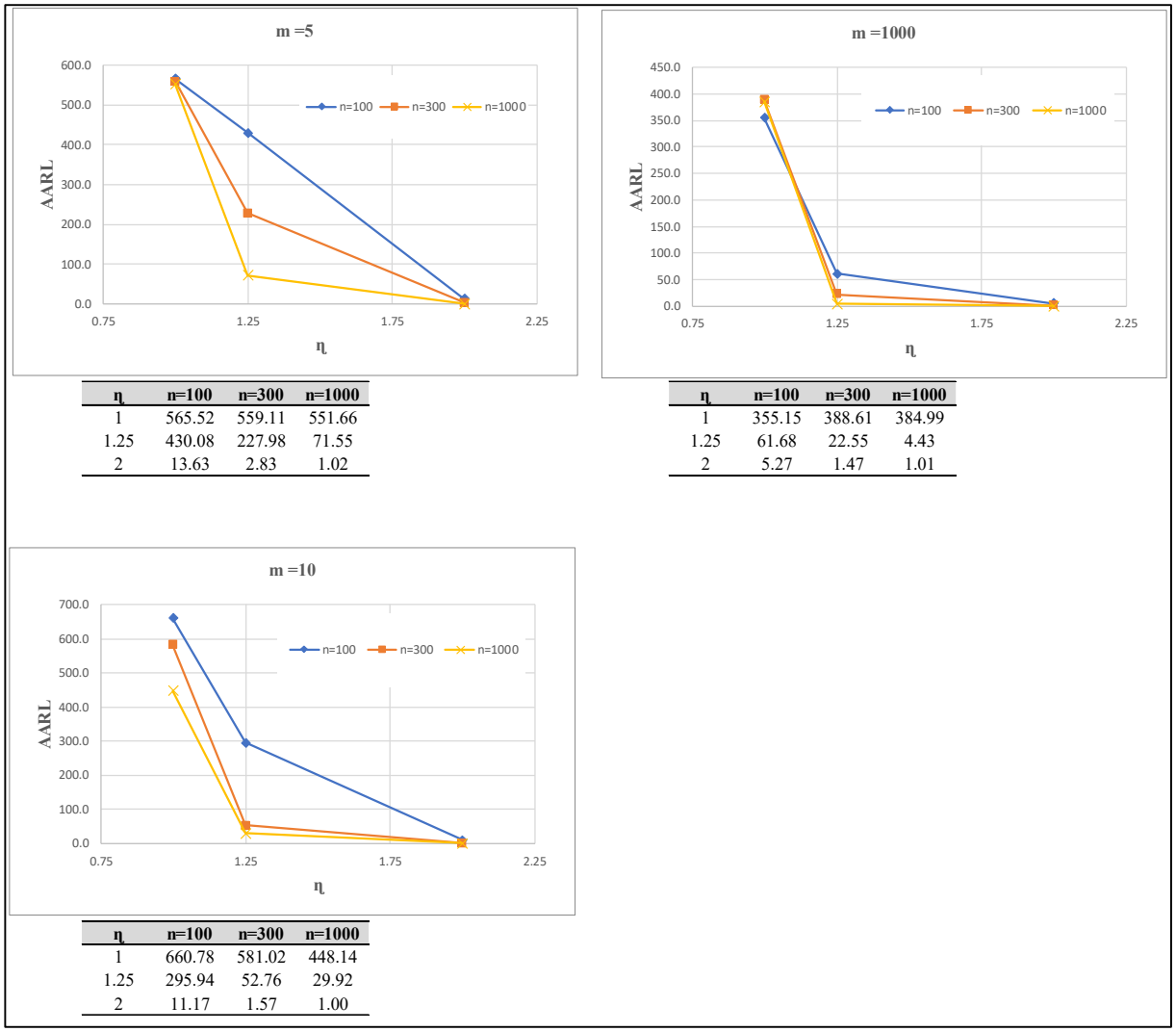
n	m=5	m=10	m=1000
100	13.63	11.17	5.27
300	2.83	1.57	1.47
1000	1.02	1.00	1.01



n	m=5	m=10	m=1000
100	430.08	295.94	61.68
300	227.98	52.76	22.55
1000	71.55	29.92	4.43

Fonte: Próprio Autor

Figura 58 - AARL para m fixo e η e n variável



η	n=100	n=300	n=1000
1	565.52	559.11	551.66
1.25	430.08	227.98	71.55
2	13.63	2.83	1.02

η	n=100	n=300	n=1000
1	355.15	388.61	384.99
1.25	61.68	22.55	4.43
2	5.27	1.47	1.01

η	n=100	n=300	n=1000
1	660.78	581.02	448.14
1.25	295.94	52.76	29.92
2	11.17	1.57	1.00

Fonte: Próprio Autor

Observa-se, nessas Figuras, que os valores de AARL sob controle (AARLo ou AARL quando $\eta = 1$) com m $\leq$ 10 são bem maiores que 370, aproximando-se desse valor somente quando m$\geq$1000. Na condição sob controle este fato reduziria a quantidade de falsos alarmes, entretanto na condição fora de controle, onde deseja-se o menor valor possível de AARL, verifica-se que a sensibilidade a variações é baixa, pois com η = 1.25 (situação fora de controle) n=100 e m=1000, o valor de AARL = 61.68, o que dificulta a identificação de causas especiais nas variações dos valores de *MCC*. Mesmo com n=1000 e m=10, ainda são necessárias mais de 29,92 rodadas para a identificação de uma variação de 0.25 em η.

Comparando os AARL's encontrados para *Kappa* e *MCC,* no cenário de m =10, n=300 e η = 1.25, os valores são respectivamente 33.48 e 52.76, nesse caso o *Kappa* tem um melhor desempenho. Nessa condição de η = 1.25 é necessário m=10 e n=1000 para o AARL do *MCC*

ser de 29.92 e aproximar-se do valor de AARL *Kappa* (33.48), ou seja, utilizam-se muito mais amostras.

Analisando a Figura 57 para $\boldsymbol{\eta}$ = 1 (condição sob controle) e m=5, verifica-se que o valor de AARL não apresenta mudanças significativas conforme aumenta-se o n. Já para $\boldsymbol{\eta}$ = 1.25, conforme aumenta-se n, reduz-se bastante o valor de AARL, o que é desejável. Entretanto para m=5 e n= 1000, o valor de AARL do *MCC* ainda é muito alto 71.55. Nas mesmas condições de m e n, o AARL baseado em *Kappa* é 8.48.

As simulações demonstram que apesar do *MCC* gerar uma quantidade menor de falsos alarmes, a quantidade e o número de amostras necessárias, para a detecção de condições fora de controle, são maiores que a quantidade e o número de amostras necessárias quando se utiliza o *Kappa*. Além disso o comportamento dos RL´s e dos ARL´s de *Kappa* aderem ao desenvolvimento teórico de forma mais regular que os RL´s e os ARL´s baseados no MCC, visto que para *MCC* a dispersão do RL é muito maior, conforme demonstrado na Figura 55.

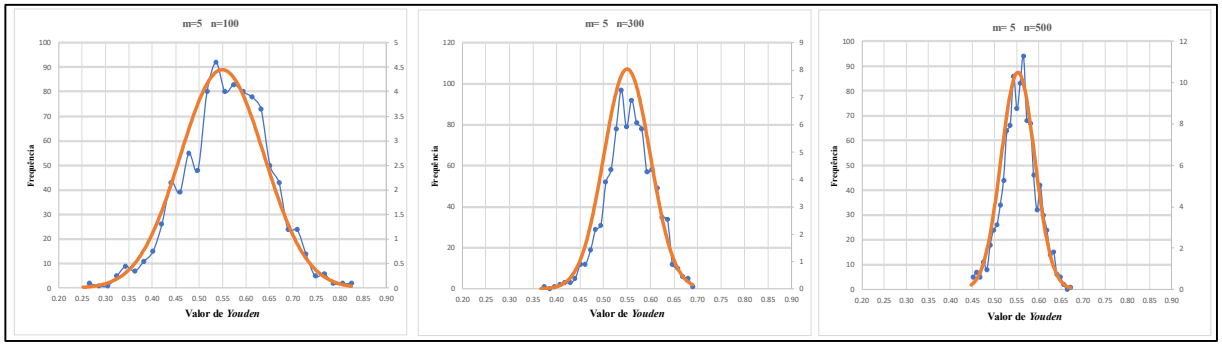
7.3.3 Caso 3: Índice de Youden

Semelhante às análises feitas para *Kappa*, elaborou-se a Figura 59 para demonstrar a normalidade dos valores de *Youden* em função de m e n. Com m = 5 e n assumindo valores de 100, 300 ou 500, simulou-se 1000 valores de *Youden* para cada um desses pares de m e n. Em cada um desses conjuntos de 1000 dados, encontrou-se os valores extremos e dividiu-se este intervalo em 30 intervalos menores, calculando-se suas frequências. Calculou-se a média e o desvio padrão, para cada conjunto, e elaborou-se a curva normal de mesmos parâmetros, sobrepondo-a aos dados de frequência e demonstrando visualmente que os valores de *Youden* seguem uma distribuição normal para os cenários simulados.

Outra análise realizada foi a checagem da distribuição dos valores de RL dos gráficos de controle baseados no índice de *Youden*. Na Figura 60 demonstra-se a frequência de ocorrência de 1000 desses valores, e sua distribuição geométrica equivalente. Nessa simulação, para cálculo dos RL´s, parametrizou-se o programa (anexo F) com VRL = 2500, VARL =1, VAARL = 1000, n = 1500 e m =5. Da mesma forma que na análise feita para o *Kappa*, cada RL é calculado para um limite de controle diferente, condição mais extrema de aleatoriedade. Neste caso a distribuição dos RL´s comportou-se próxima à distribuição geométrica. Entretanto o valor médio encontrado para os RL´s ($\widehat{ARL}$) simulados foi de 162.88 e $\hat{\sigma}_{RL}$ = 271,91, bem abaixo do valor de 370 da curva teórica, calculada em função da probabilidade de sucesso de

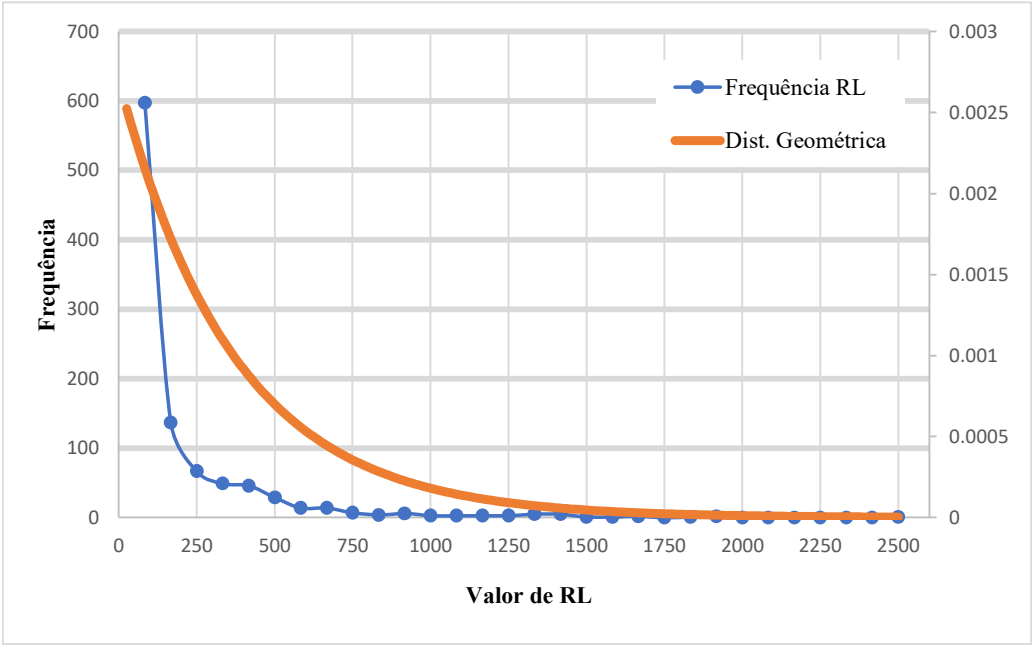
0,27% pertinente aos limites de controle com h=3 (Equações 20 e 21) .

Figura 59 - Análise de normalidade de *Youden* em função de m e n



Fonte: Próprio Autor

Figura 60 - Análise da distribuição dos RL´s - Youden



Fonte: Próprio Autor

Verificada a condição de normalidade, demonstram-se na Tabela 13 os resultados das simulações, onde os valores de AARL são calculados de acordo com os cenários definidos no Quadro 30. Na tabela 13, os valores indicados de *Youden* médio referem-se a média dos *Youden´s* médios utilizados para o cálculo dos limites de controle. Aqui cabe uma observação com relação ao cálculo do Índice de *Youden* e seu desvio padrão, estes foram, respectivamente, calculados de acordo com as Equação 15 e 16.

Em todas as simulações os limites de controle para *Youden* foram calculados considerando-se as equações 20, 21, 15 e 16 com $Z_{\alpha/2} = 3$.

Quadro 30 - Cenários de Simulação para *Youden*

VRL =	**2500**		**VARL =**	**100**		**VAARL = 50**		**Limiar =**	**0.5**
n	**m**	**η**	**n**	**m**	**η**	**n**	**m**	**η**	
100	1	1	300	1	1	1000	1	1	
100	1	1.25	300	1	1.25	1000	1	1.25	
100	1	2	300	1	2	1000	1	2	
100	5	1	300	5	1	1000	5	1	
100	5	1.25	300	5	1.25	1000	5	1.25	
100	5	2	300	5	2	1000	5	2	
100	10	1	300	10	1	1000	10	1	
100	10	1.25	300	10	1.25	1000	10	1.25	
100	10	2	300	10	2	1000	10	2	
100	1000	1	300	1000	1	1000	1000	1	
100	1000	1.25	300	1000	1.25	1000	1000	1.25	
100	1000	2	300	1000	2	1000	1000	2	

Fonte: Próprio Autor

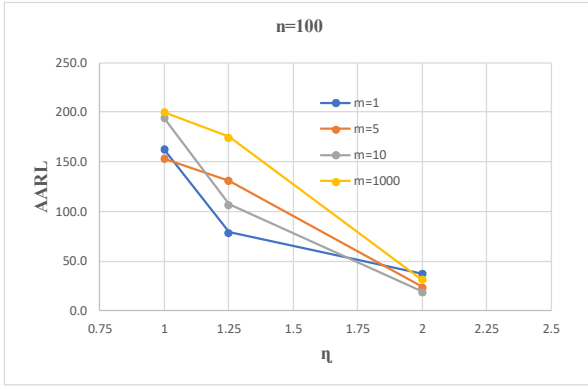
Tabela 13 - Resultados de AARL nos cenários de simulação de *Youden*

n	m	η	AARL	Youden médio	*RSP*
100	1	1	162.55	0.523	0.4046
100	1	1.25	78.85	0.572	0.4096
100	1	2	37.72	0.551	0.4002
100	5	1	153.49	0.558	0.397
100	5	1.25	131.14	0.558	0.39688
100	5	2	24.36	0.564	0.3994
100	10	1	194.07	0.575	0.3964
100	10	1.25	106.95	0.551	0.3969
100	10	2	19.59	0.538	0.3969
100	1000	1	199.82	0.549	0.3970
100	1000	1.25	175.14	0.559	0.3965
100	1000	2	31.49	0.543	0.3964
300	1	1	158.44	0.548	0.4022
300	1	1.25	73.09	0.546	0.3926
300	1	2	1.91	0.558	0.39367
300	5	1	112.25	0.557	0.39712
300	5	1.25	78.95	0.545	0.39545
300	5	2	1.57	0.565	0.39239
300	10	1	129.83	0.559	0.4307
300	10	1.25	100.04	0.539	0.3981
300	10	2	7.20	0.546	0.3955
300	1000	1	102.05	0.543	0.3968
300	1000	1.25	65.93	0.548	0.3966
300	1000	2	3.55	0.553	0.3968
1000	1	1	147.00	0.559	0.3949
1000	1	1.25	28.45	0.547	0.3973
1000	1	2	1.01	0.550	0.3991
1000	5	1	153.49	0.552	0.3935
1000	5	1.25	23.50	0.549	0.3952
1000	5	2	1.01	0.551	0.3958
1000	10	1	167.14	0.549	0.3955
1000	10	1.25	36.21	0.547	0.3974
1000	10	2	1.00	0.558	0.3961
1000	1000	1	161.99	0.555	0.3967
1000	1000	1.25	17.34	0.556	0.3966
1000	1000	2	1.01	0.551	0.3966

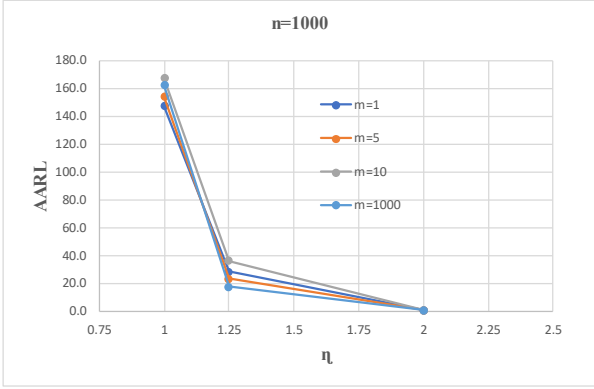
Fonte: Próprio Autor

Da mesma forma que para a análise do *Kappa*, foram elaboradas diversas Figuras. Na Figura 61 cada gráfico é fixo para um determinado valor de n, enquanto os valores de AARL são demonstrados em função de m e $\boldsymbol{\eta}$. Na Figura 62 cada gráfico é fixo para um determinado valor de $\boldsymbol{\eta}$, enquanto os valores de AARL são demonstrados em função de m e n. Na Figura 63 cada gráfico é fixo para um determinado valor de m, enquanto os valores de AARL são demonstrados em função de n e $\boldsymbol{\eta}$.

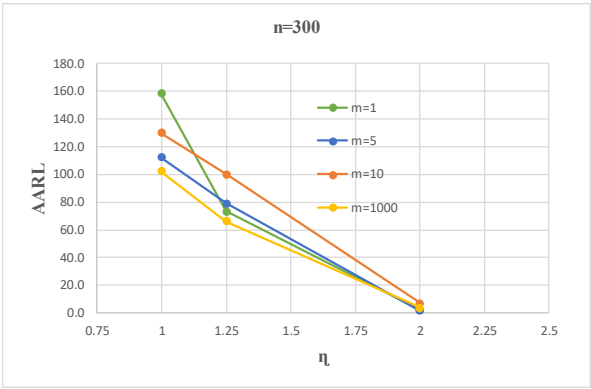
Figura 61 - AARL para n fixo e m e $\boldsymbol{\eta}$ variáveis - *Youden*



η	m=1	m=5	m=10	m=1000
1	162.55	153.49	194.07	199.82
1.25	78.85	131.14	106.95	175.14
2	37.72	24.36	19.59	31.49



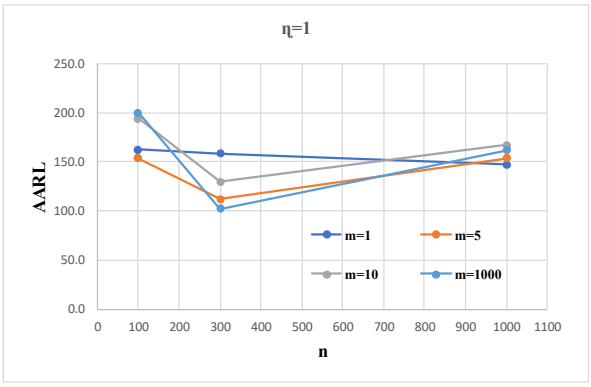
η	m=1	m=5	m=10	m=1000
1	147.00	153.49	167.14	161.99
1.25	28.45	23.50	36.21	17.34
2	1.01	1.01	1.00	1.01



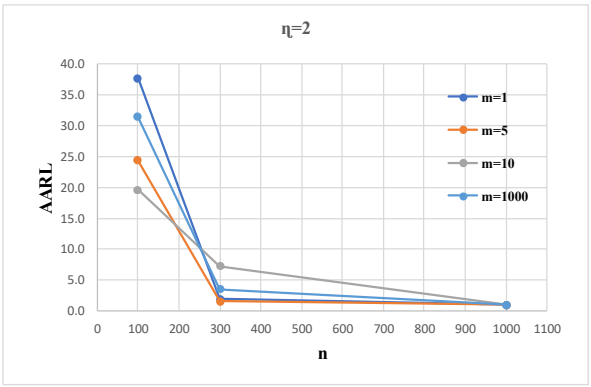
η	m=1	m=5	m=10	m=1000
1	158.44	112.25	129.83	102.05
1.25	73.09	78.95	100.04	65.93
2	1.91	1.57	7.20	3.55

Fonte: Próprio Autor

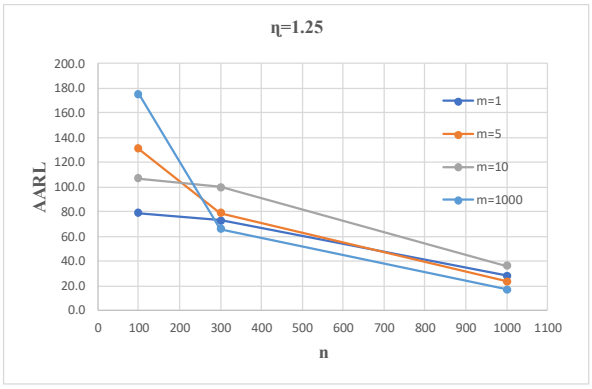
Figura 62 - AARL para η fixo e m e n variável - *Youden*



n	m=1	m=5	m=10	m=1000
100	162.55	153.49	194.07	199.82
300	158.44	112.25	129.83	102.05
1000	147.00	153.49	167.14	161.99



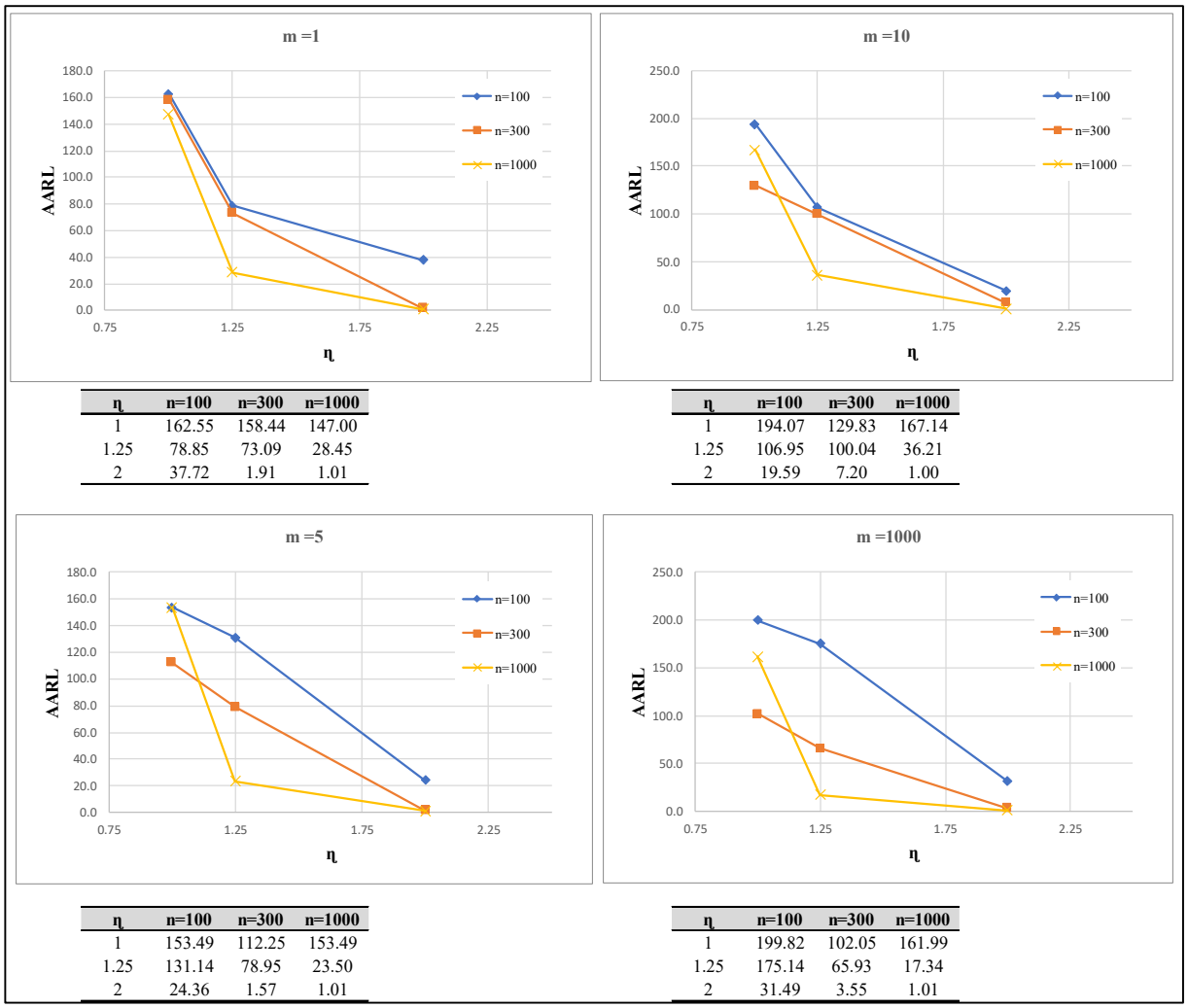
n	m=1	m=5	m=10	m=1000
100	37.72	24.36	19.59	31.49
300	1.91	1.57	7.20	3.55
1000	1.01	1.01	1.00	1.01



n	m=1	m=5	m=10	m=1000
100	78.85	131.14	106.95	175.14
300	73.09	78.95	100.04	65.93
1000	28.45	23.50	36.21	17.34

Fonte: Próprio Autor

Figura 63 - AARL para m fixo e η e n variável - *Youden*



η	n=100	n=300	n=1000
1	162.55	158.44	147.00
1.25	78.85	73.09	28.45
2	37.72	1.91	1.01

η	n=100	n=300	n=1000
1	194.07	129.83	167.14
1.25	106.95	100.04	36.21
2	19.59	7.20	1.00

η	n=100	n=300	n=1000
1	153.49	112.25	153.49
1.25	131.14	78.95	23.50
2	24.36	1.57	1.01

η	n=100	n=300	n=1000
1	199.82	102.05	161.99
1.25	175.14	65.93	17.34
2	31.49	3.55	1.01

Fonte: Próprio Autor

Nessas Figuras observa-se que, os valores de AARL sob controle (AARLo ou AARL quando $\boldsymbol{\eta = 1}$) são bem menores que 370 para todos os valores de m e n simulados, especificamente o maior valor encontrado foi de 199.82 para m=1000 e n =100. Como já observado, na condição sob controle, baixos valores de AARL aumentam a quantidade de falsos alarmes, o que não é desejável. Em condições sob controle, AARL abaixo de 200 também foi encontrado para o *Kappa,* no entretanto, quando a quantidade m de amostras aumenta de um para cinco, o AARL baseado em *Kappa* aumenta para valores acima de 243 (simulações), sendo que este valor é de 237 conforme o modelo analítico para m=5 (Tabela 4 com δ = 0,005). E o AARL de *Kappa* fica acima de 300, aumentando-se m=5 para m=10, aproximando-se dos valores teóricos de 370 conforme aumenta-se m e n, o que não acontece com o índice de *Youden*.

Na condição fora de controle, onde deseja-se o menor valor possível de AARL, verifica-se que a sensibilidade a variações no valor de *Youden,* em relação ao seu valor médio, é baixa,

pois com η = 1.25 (situação fora de controle) e n=100 o menor valor de AARL foi de 78,85, e este valor não diminuiu com o aumento de m. Para esse η (1.25), o menor valor AARL encontrado foi de 17,34, entretanto foram necessários n e m iguais a 1000. Tratando-se de *Kappa,* com η = 1.25, o valor de AARL = 15,59 é alcançado com m=10 e n=500, ou seja, uma amostra muito menor. Isso demonstra que o gráfico baseado em *Youden* é menos sensível que o baseado em *Kappa*, para pequenas variações em torno dos valores nominais.

Comparando os AARL's encontrados para *Kappa* e *Youden,* no cenário de m =5, n=100 e η = 2, os valores são respectivamente 8,77 e 24.36. Ou seja, para maiores η o *Kappa* também tem um melhor desempenho.

As simulações demonstram que o uso de gráficos baseados em *Youden* geram uma quantidade maior de falsos alarmes, tanto em relação aos gráficos baseados no *Kappa*, quanto aos gráficos baseados no *MCC*. A quantidade e o número de amostras necessárias, para a detecção de condições fora de controle, são maiores que a quantidade e o número de amostras necessárias quando se utiliza o *Kappa*. Além disso o comportamento dos RL´s e dos ARL´s de *Kappa* aderem ao desenvolvimento teórico de forma mais regular que os RL´s e os ARL´s baseados no *Youden.*

8 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Conforme demonstrado na seção 5.1, somente dados externos refletem as diferentes condições de operação a que um modelo está sujeito ao longo do tempo. Condições que podem se alterar significativamente, reduzindo ou melhorando o desempenho do modelo criado.

O controle possibilitado pelo SPM, aplicado aos indicadores estudados, em especial o *Kappa*, elimina a subjetividade de definição da variação aceitável, uma vez que este critério é objetivamente definido pelos Limites de Controle (LC).

Demonstrou-se até aqui, que essas condições operacionais se referem não apenas ao grau de correlação entre as variáveis, mas também às questões relativas a interferências na captura, transformação, padronização, armazenamento e processamento contínuo dos dados e aos demais itens indicados na Figura 42.

O SPM, integrado aos processos de criação de modelos (KDD/RDMP), atua como uma ferramenta de monitoramento e de apoio à decisão, cumprindo um papel ainda prescrito no campo da teoria do DS. Sua operacionalização, através do *Kappa*, possibilita definir precisamente quando variações nas condições operacionais interferem significativamente no resultado do processo decisório. O cálculo de *Kappa* independe do modelo classificatório criado, assim é possível aplicá-lo a outras técnicas preditivas como: análises discriminantes, algoritmos de árvore de decisão, redes neurais ou quaisquer outros métodos que se apliquem ao contexto estudado.

A simulação, criada para analisar o desempenho do gráfico de controle baseado em *Kappa*, indica que uma boa opção é usar quantidades de amostra maiores ou iguais a cinco ($m \geq 5$) e tamanhos de amostras maiores que 500 ($n \geq 500$), pois isso reduz o erro tipo I e o ARL (ou seja, aumenta o poder do teste - Equação 25), detectando rapidamente variações nas condições operacionais. Na criação dos gráficos de controle, quando valores de m maiores que um não forem aplicáveis na prática, o tamanho n da amostra deve ser acima de 500, ressalvando-se que pequenos valores de $\boldsymbol{\delta}$ (simulados por $\boldsymbol{\eta}$), não serão detectados rapidamente. Esses valores estão em concordância com as considerações feitas em 7.2.5 a respeito dos modelos analíticos desenvolvidos para *Kappa*. Estes valores de quantidade e número de amostras são bem maiores que o valor de oitenta, indicado na literatura, para o caso avaliado (vinte amostras para cada variável e para cada categoria de classificação possível, 20 x 2 x 2 = 80).

As simulações de distúrbios nas condições operacionais, criadas a partir de variações no fator $\boldsymbol{\eta}$, demonstrou que os gráficos de *Kappa* identificaram rapidamente variações de ordem prática ($\boldsymbol{\eta} \geq 1.5$), indicando não serem necessários gráficos mais sensíveis e mais complexos.

Com relação aos índices *MCC e Youden*, as simulações demonstram que o *Kappa* possui vantagens, pois demanda quantidades (m) e tamanhos de amostras (n) menores para desempenhos semelhantes em relação ao ARL, ao contrário do que é proposto pelos trabalhos mais recentes.

Outra característica importante, demonstrada pelas simulações, refere-se a distribuição dos valores de RL dos gráficos baseados em *Kappa*, sua distribuição aproxima-se da distribuição geométrica teórica (Figura 47), pois as estimativas de ARL e do desvio padrão de ARL tem valores próximos (conforme 7.3.1), tendendo a 370 com o aumento de m e n. Diferentemente disso, apesar da distribuição dos valores de RL de *MCC e Youden* terem perfil próximo a distribuição geométrica, os valores médios e desvios padrão, não se aproximam dos valores teóricos dessa distribuição da mesma forma como para os gráficos baseados em *Kappa*.

Elaborou-se comparação entre os índices, gerando-se 1000 valores de ARL, simulados com a seguinte parametrização dos programas: VRL = 2500, VARL =30, VAARL = 1000, n = 300, m =10 e $\boldsymbol{\eta}$ = 1 (condição sob controle). Relembrando, nessa parametrização de simulação, calcula-se um limite de controle com base em m=10 e n=300. Para esse limite de controle geram-se valores dos índices até que um saia fora dos LC´s, ou se chegue no valor máximo de repetições de 2500 (relembrando que se trunca o programa em 2500). Como VARL=30, para cada limite simulam-se 30 RL´s diferentes, a média destes RL´s é o ARL do limite de controle. Dado que VAARL=1000, o processo é feito para 1000 diferentes limites de controle, os dados estão resumidos na Tabela 14.

Tabela 14 - Comparações da Distribuição de ARL entre os indicadores (m=10, n=300, $\boldsymbol{\eta}$ = 1)

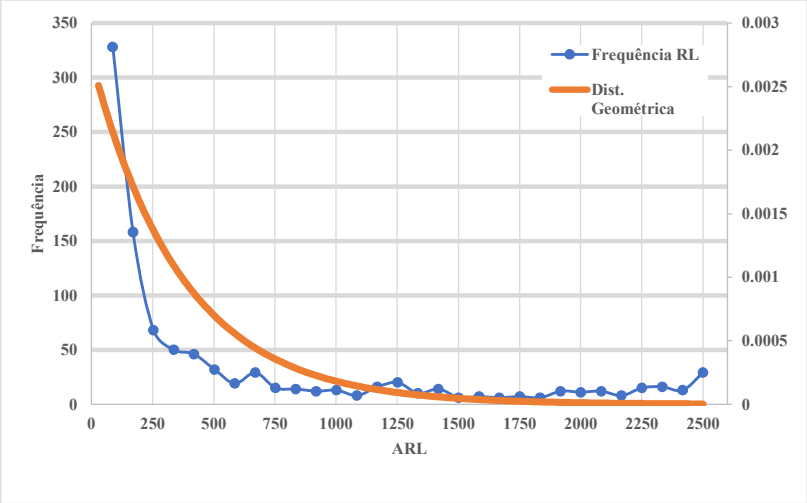
	Mínimo	1° quartil	2° quartil	3° quartil	4° quartil	média estimada	desvio padrão estimado
Kappa	45.13	213.6	312.77	395.7	749.43	306.26	120.58
MCC	3.3	57.87	180.07	781.93	2500	558.45	731.39
Youden	1.2	26.7	96	244.2	764.13	147.13	141.33

Fonte: Próprio Autor

Analisando-se a Tabela 14 percebe-se um comportamento muito mais regular do ARL baseado nos gráficos de *Kappa*, em relação aos demais índices. Os valores de ARL baseados em *MCC* e *Youden* são muito mais dispersos de uma maneira geral, e, quartil a quartil. Essa

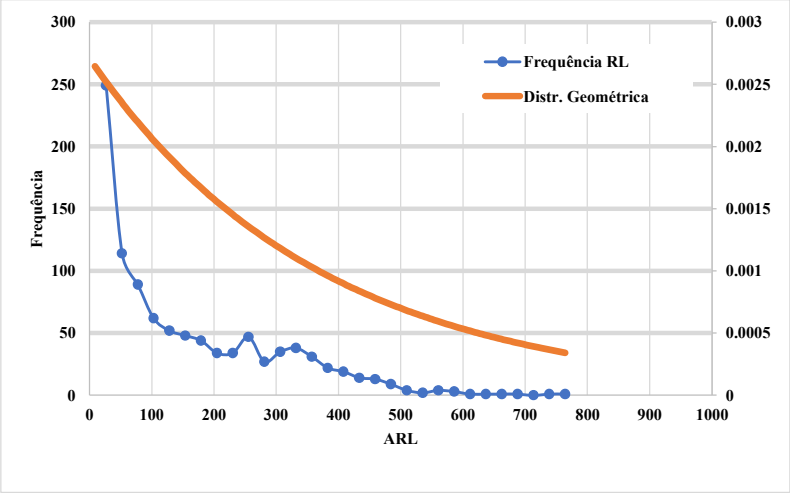
menor dispersão do ARL baseado em *Kappa* também pode ser verificada no comportamento normal da distribuição de seus ARL´s, demonstrado na Figura 48. O mesmo não acontece para a distribuição dos ARL´s baseados em *MCC* e *Youden*, que mantém o perfil de distribuição dos RL´s, mesmo aumentando-se a quantidade de RL´s gerados para cada limite de controle, e usados no cálculo do ARL. Isso é demonstrado nas Figuras 64 e 65, baseadas nos mesmos critérios utilizados na elaboração da Figura 47.

Figura 64 - Análise da distribuição do ARL´s - *MCC*



Fonte: Próprio Autor

Figura 65 - Análise da distribuição do ARL´s - *Youden*



Fonte: Próprio Autor

Finalmente, se forem comparadas as Figuras 46, 54 e 59, com m=5 e n=500, nota-se que: o *Kappa* tem uma dispersão de dados menor com relação ao *MCC;* e uma dispersão bem menor com relação ao *Youden*, sendo que as condições de simulação foram as mesmas.

9 CONCLUSÕES

O SPM, tradicionalmente utilizado para a melhoria de características físicas de produtos, foi estudado como elemento de garantia e melhoria da qualidade dos processos decisórios, apoiados por modelos preditivos classificatórios. Para tanto, propôs-se uma abordagem de integração entre a DS e o SPM, ilustrada pela Figura 41. Sua operacionalização, através da aplicação dos diferentes indicadores estudados, e técnicas analíticas e de simulação, apresentou como resultado uma abordagem prescritiva para o uso do *Kappa* na validação contínua de modelos. Assim, quanto a primeira questão de pesquisa, referente a utilização do SPM no contexto da DS, conclui-se que o SPM é aplicável à validação contínua do modelo, pois é elemento recomendado para o monitoramento das condições operacionais onde esse é utilizado. E demonstrou-se que modelos possuem inúmeras fontes de variabilidade.

As análises de desempenho dos gráficos de controle, baseadas nos índices *Kappa, Youden e MCC*, demonstraram que o *Kappa* necessita de menores quantidades e tamanhos de amostra que os demais índices. Entretanto essas simulações foram elaboradas para a condição de um banco de dados relativamente balanceado. Autores já indicados anteriormente, na seção 5.3, afirmam que o *Kappa* também é adequado para bancos de dados não balanceados, o que não deixa de ser verdade, mas os modelos analíticos propostos em 7.2 demonstram que o tamanho de m e n cresce bastante conforme o desbalanceamento do banco de dados aumenta, podendo inviabilizar o uso do índice em situações em que a quantidade de dados não é abundante.

Comparando-se ao *MCC*, o *Kappa* tem maior poder de detecção de causas especiais que tornam inadequado o modelo preditivo. Observa-se isso devido aos menores valores de AARL, para mesmos valores de m e n.

Com relação aos gráficos baseados em *Youden*, os valores de AARL, na situação fora de controle, foram maiores do que os valores de AARL baseados nos gráficos de *Kappa* (para mesmos tamanhos e quantidades de amostras). Inversamente, os valores de AARL na situação sob controle, dos gráficos baseados em *Youden*, são menores do que os valores de AARL baseados nos gráficos de *Kappa* (para mesmos tamanhos e quantidades de amostras). O que demonstra que os gráficos baseados em *Kappa* tem melhor desempenho com relação aos gráficos baseados em *Youden*.

Assim, com relação a segunda questão de pesquisa, concluiu-se que o *Kappa* pode ser utilizado na operacionalização do SPM como ferramenta de validação contínua desde que se

utilize os valores adequados de m e n. Apesar do desempenho inferior, os outros índices estudados também se adequaram ao cenário específico utilizado nas simulações, desde que também sejam aplicadas as considerações quanto à m e n específicas a cada um deles. Entretanto, para uma análise mais completa dos demais índices, é necessária a elaboração dos mesmos modelos analíticos desenvolvidos para *Kappa* em 7.2, o que se configura em oportunidade futura de pesquisa.

Não se comparou *Kappa* ao índice *AUC,* porque, apesar de muito utilizado, com base nos autores estudados, conclui-se que sua utilização, como medida de qualidade dos resultados de um modelo, é equivocada. Esta afirmação baseia-se principalmente em dois fatos: *i)* o *AUC* é calculado sobre todo o espectro do limiar, o que na prática não acontece, pois, um determinado modelo, após sua elaboração, é aplicado a um determinado limiar, e não a todo o seu espectro; *ii)* utilizar a *AUC* é equivalente a medir o desempenho das regras de classificação usando métricas que dependem das regras que estão sendo medidas, ou seja, se a *AUC* depende do classificador, a métrica de avaliação entre classificadores, é diferente para cada um deles, se é assim, não se está comparando as mesmas coisas.

Com relação a aderência do uso do SPM para a melhoria dos modelos preditivos, conclui-se que a técnica, fundamental na melhoria contínua de processos, também é aplicável nesse contexto. O uso do SPM estabelece uma base de referência, sobre as quais mensuram-se variações nos resultados dos indicadores. Demonstrou-se na seção 7.1 que essas variações podem ser devidas tanto ao modelo preditivo desenvolvido quanto a problemas na captação e fluxo de dados, e quanto maiores estes fluxos (como em ambientes de Big Data), maior a possibilidade de problemas. Assim, quando se verifica continuamente um modelo, na realidade, todos estes elementos do processo de decisão estão sendo checados. Especial atenção deve ser dedicada a tendências, ou variações, acima dos valores nominais estimados para os indicadores, uma vez que, quanto mais próximo de 1, maior a quantidade de acertos. Entender as causas que interferem na variação positiva, mesmo que dentro dos limites de controle, significa mais decisões assertivas.

O método de pesquisa utilizado orientou a definição de uma classe de problema específico, no caso, o uso do *Kappa* integrado ao SPM para a validação contínua de modelos. A elaboração e validação da abordagem de integração entre o SPM e a DS (ou seja, criação e validação do artefato) resultou em uma abordagem prescritiva que orienta a seleção da quantidade e o tamanho de amostras, em função da condição de balanceamento de dados, para se calcular adequadamente o *Kappa*. As poucas publicações, que discutiram o uso de

monitoramento estatístico não definiram critérios para a seleção da quantidade e tamanho das amostras necessárias, aspectos fundamentais do SPM, tendo o trabalho avançado nesse sentido.

Apesar de algumas críticas ao *Kappa*, delimitou-se objetivamente as fronteiras de aplicação onde o índice pode ser utilizado. Concluiu-se, mesmo que não se utilize a validação contínua, e sempre que viável, é melhor dividir o bloco de dados, utilizados na etapa de validação do modelo (Fase I do SPM), em várias amostras de tamanho $n \geq 500$, separadas no tempo, dada que a análise em vários pontos é mais adequada do que uma estimativa única. Além disso o *Kappa* é importante pois seu cálculo não é complexo, e está disponível em softwares como Python e R.

Como foi elaborado uma abordagem geral, a estrutura criada possibilita que a mesma análise possa ser realizada trocando-se *Kappa* por qualquer outro indicador de qualidade de resultados dos modelos: como a distância entre proporções; índices de precisão absoluta *(Po)*; variações do próprio índice *Kappa*; Score de Brier; entre outros. Cabe aqui outra observação, podem ser encontradas aplicações onde deseja-se comparar a qualidade e a estabilidade preditiva de classificação entre modelos automatizados e o julgamento empírico pessoal, ou mesmo uma condição híbrida entre estes elementos, situação que a abordagem se adequa sem necessidade de ajustes.

O estudo elaborado pode ser complementado nos seguintes aspectos: criando-se modelos analíticos para *MCC* e *Youden*; rodando simulações baseadas em bancos de dados com diferentes graus de desbalanceamento; aplicando a mesma estrutura com diferentes ferramentas de modelagem, como redes neurais artificiais, análise discriminante e árvores de decisão.

Dentro da ampla gama de índices, que avaliam a assertividade de modelos, comparou-se o desempenho dos índices *Kappa, MCC e Youden*, ou seja, ainda existem várias oportunidades de trabalhos futuros, considerando-se a aplicação da abordagem já definido, na análise de quaisquer outros índices pertinentes a este contexto, como os demonstrados na seção 5.2.

Outra oportunidade é utilizar a abordagem desenvolvida para avaliar a influência do uso de diferentes valores para o Limiar (baseados nos diferentes critérios descritos em 5.2), avaliando o desempenho (variabilidade) dos indicadores em função de diferentes ponderações entre erros de omissão e comissão.

Malan *et al*. (2004) usa intervalos de confiança para o cálculo de diversos índices de concordância para um mesmo modelo, entretanto a quantidade de amostras aplicadas é sempre

menor que 300. Apesar dos índices avaliados serem diferentes de *Kappa*, a análise desse artigo indica uma importante oportunidade de pesquisa, que é analisar em trabalhos similares o tamanho das amostras utilizadas e como os autores aplicam intervalos de confiança na comparação entre índices. Isso é relevante, pois se as amostras não são adequadas, os intervalos de confiança também não são, invalidando as comparações.

Os gráficos de controle utilizados demonstraram capacidade adequada na avaliação da variabilidade e, também com relação a sensibilidade na captura das mudanças. No entanto, este desempenho poderá ser comparado com outros tipos de gráfico mais adequados quando a quantidade de amostras é reduzida, como os gráficos EWMA e CuSum.

A base de referência foi a estatística paramétrica, assim, a exploração dos mesmos conceitos no campo da estatística não paramétrica é uma outra abordagem a ser avaliada.

Em resumo, as principais contribuições deste trabalho tanto para a teoria quanto para os praticantes podem ser sintetizadas nos seguintes aspectos: *i)* propõe a integração de DS e SPM; *ii)* desenvolve uma ferramenta para a validação contínua de modelos preditivos classificatórios; *iii)* compara diferentes índices de avaliação de qualidade de modelos, indicando suas vantagens e desvantagens; *iv)* define critérios de amostragem e procedimento para aplicação do SPM considerando-se as Fases I e II da técnica; *v)* a abordagem validada serve de base para as mais diferentes análises, o que possibilita a comparação objetiva entre todas as alternativas que forem desenhadas.

10 BIBLIOGRAFIA

ABELL, J. A. et al. Big Data-Driven Manufacturing - Process-Monitoring-for-Quality Philosophy. **Journal of Manufacturing Science and Engineering, Transactions of the ASME**, v. 139, n. 10, 2017.

ABIMAQ. **Medidas do Governo**. Disponível em: <http://www.abimaq.org.br/site.aspx/Abimaq-Informativo-Mensal-Infomaq?DetalheClipping=47&CodigoClipping=905>. Acesso em: 4 jan. 2021.

ACOSTA-MEJIA, C. A. Improved p charts to monitor process quality. **IIE Transactions (Institute of Industrial Engineers)**, v. 31, n. 6, p. 509–516, 1999.

AGGARWAL, L. P. Data augmentation in dermatology image recognition using machine learning. **Skin Research and Technology**, v. 25, n. 6, p. 815–820, 2019.

AGRAWAL, R.; SHAFER, J. C. Parallel Mining Of Association Rules. **Knowledge and Data Engineering**, v. 8, p. 962–969, 1996.

AHUETT-GARZA, H.; KURFESS, T. A brief discussion on the trends of habilitating technologies for Industry 4.0 and Smart manufacturing. **Manufacturing Letters**, v. 15, p. 60–63, 2018.

ALTMAN, D. G. et al. Prognosis and prognostic research: Validating a prognostic model. **BMJ (Online)**, v. 338, n. 605, p. 1432–1435, 2009.

APREDA, R. et al. Functional technology foresight . A novel methodology to identify emerging technologies. **European Journal of Futures Research**, v. 4, n. 13, 2016.

AUSTIN, P. C. A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. **Statistics in Medicine**, n. April, p. 4267–4278, 2008.

AUSTIN, P. C.; STEYERBERG, E. W. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. **Statistical Methods in Medical Research**, v. 26, n. 2, p. 796–808, 2017.

AYANKOYA, K.; CALITZ, A.; GREYLING, J. **Intrinsic relations between data science, big data, business analytics and datafication**. (SAICSIT, Ed.)SAICSIT 2014 - Empowered by Technology International Conference Proceeding Series. **Anais**...Centurion: 2014

BABICEANU, R. F.; SEKER, R. Big Data and virtualization for manufacturing cyber-physical systems : A survey of the current status and future outlook. **Computers in Industry**, n. 2015, p. 10, 2016.

BAGHERI, M.; ZOLLANVARI, A.; NEZHIVENKO, S. Transformer Fault Condition Prognosis Using Vibration Signals over Cloud Environment. **IEEE Access**, p. 1–13, 2018.

BANERJEE, A.; BANDYOPADHYAY, T.; ACHARYA, P. Data Analytics: Hyped Up Aspirations or True Potential? **Vikalpa**, v. 38, n. 4, p. 1–12, 2013.

BELHADI, A. et al. Understanding the capabilities of Big Data Analytics for manufacturing process : insights from literature review and multiple case study. **Computers & Industrial Engineering**, p. 106099, 2019.

BEN-DAVID, A. About the relationship between ROC curves and Cohen's kappa.

Engineering Applications of Artificial Intelligence, v. 21, n. 6, p. 874–882, 2008.

BENHAM, D. et al. **Measurement Systems Analysis - reference manual AIAG**. 3° ed ed. [s.l.] Daimler Chrysler Corporation, Ford Motor Company, General Motors Corporation, 2002.

BI, Z.; XU, L. DA; WANG, C. Internet of things for enterprise systems of modern manufacturing. **IEEE Transactions on Industrial Informatics**, v. 10, n. 2, p. 1537–1546, 2014.

BOEHM, B. W. Seven basic principles of software engineering. **The Journal of Systems and Software**, v. 3, n. 1, p. 3–24, 1983.

BOLAÑOS-SANCHEZ, R.; SANCHEZ-ARCILLA, A.; CATEURA, J. Evaluation of two atmospheric models for wind-wave modelling in the NW Mediterranean. **Journal of Marine Systems**, v. 65, p. 336–353, 2007.

BOORN, H. G. VAN DEN et al. Prediction models for patients with esophageal or gastric cancer : A systematic review and meta-analysis. **PLoS ONE**, p. 1–20, 2018.

BOS, J. M. et al. Prediction of clinically relevant adverse drug events in surgical patients. **PLoS ONE**, v. 13, n. 8, p. 1–12, 2018.

BOUGHORBEL, S.; JARRAY, F.; EL-ANBARI, M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. **PLoS ONE**, v. 12, n. 6, p. 1–17, 2017.

BRANDENBURGER, J. et al. Big Data Solution for Quality Monitoring and Improvement on Flat Steel Production. **IFAC-PapersOnLine**, v. 49, n. 20, p. 55–60, 2016.

BRENNAN, R. L.; PREDIGER, D. J. Coefficient kappa: Some uses, misuses, and alternatives. **Educational and Psychological Measurement**, v. 41, n. 3, p. 687–699, 1981.

BRYMAN, A. **Research Methods and Organization Studies**. London: Teylor & Francis Group, 2005.

CABINET OFFICE. **Report on The 5th Science and Technology Basic Plan**. Tokio: [s.n.].

CARVAJAL SOTO, J. A.; TAVAKOLIZADEH, F.; GYULAI, D. An online machine learning framework for early detection of product failures in an Industry 4.0 context. **International Journal of Computer Integrated Manufacturing**, v. 00, n. 00, p. 1–14, 2019.

CARVALHO, T. P. et al. A systematic literature review of machine learning methods applied to predictive maintenance. **Computers and Industrial Engineering**, v. 137, p. 10, 2019.

CHAARI, R. et al. Cyber-Physical Systems Clouds: A Survey. **computer Networks**, 2016.

CHAKRABORTI, S.; GRAHAM, M. A. **Non Parametric Statistical Process Control**. Hoboken, NJ - USA: John Wiley and Sons Ltd, 2019.

CHAKRABORTI, S.; HUMAN, S. W.; GRAHAM, M. A. Phase I statistical process control charts: An overview and some results. **Quality Engineering**, v. 21, n. 1, p. 52–62, 2009.

CHANG, S. I. **Approaches to implement statistical process control for manufacturing in Big data era**. Proceedings of the ASME 2017 12th International Manufacturing Science and Engineering Conference. **Anais**...2017

CHEN, B. et al. Smart Factory of Industry 4.0: Key Technologies, Application Case, and Challenges. **IEEE Access**, v. 6, p. 6505–6519, 2017.

CHEN, J. et al. CPS Modeling of CNC Machine Tool Work Processes Using an Instruction-

Domain Based Approach. **Engineering**, v. 1, n. 2, p. 247–260, 2015.

CHEN, Y. Integrated and Intelligent Manufacturing: Perspectives and Enablers. **Engineering**, v. 3, n. 5, p. 588–595, 2017.

CHEN, Y. C.; SONG, W. T. The effects of sample sizes in phases i and ii on control chart performance. **Communications in Statistics - Theory and Methods**, v. 41, n. 22, p. 4047–4068, 2012.

CHENG, F.-T. et al. Development of Advanced Manufacturing Cloud of Things (AMCoT)—A Smart Manufacturing Platform. **IEEE Robotics and Automation Letters**, v. 2, n. 3, p. 1809–1816, 2017.

CHEONG, Y.; DE GREGORIO, F.; KIM, K. The power of reach and frequency in the age of digital advertising: Offline and online media demand different metrics. **Journal of Advertising Research**, v. 50, n. 4, p. 403–415, 2010.

CHIANG, L.; LU, B.; CASTILLO, I. Big Data Analytics in Chemical Engineering. **Annual Review of Chemical and Biomolecular Engineering**, v. 8, n. 1, p. 63–85, 2017.

CHICCO, D. Ten quick tips for machine learning in computational biology. **BioData Mining**, v. 10, n. 1, p. 1–17, 2017.

CHICCO, D.; JURMAN, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. **BMC Genomics**, v. 21, n. 1, p. 1–13, 2020.

CHICCO, D.; TÖTSCH, N.; JURMAN, G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. **BioData Mining**, v. 14, n. 1, p. 1–22, 2021.

CHICCO, D.; WARRENS, M. J.; JURMAN, G. The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier score in binary classification assessment. **IEEE Access**, v. 9, n. Mcc, 2021.

CHIKUSHI, R. T. M. et al. Using spectral entropy and bernoulli map to handle concept drift. **Expert Systems with Applications**, 2020.

CHU, D. et al. Validation of the satellite-derived rainfall estimates over the tibet. **Acta Meteorologica Sinica**, v. 25, n. 6, p. 734–741, 2011.

ÇINAR, Z. M. et al. Machine learning in predictive maintenance towards sustainable smart manufacturing in industry 4.0. **Sustainability (Switzerland)**, v. 12, n. 19, 2020.

CINTOLO-GONZALEZ, J. A. et al. Breast cancer risk models: a comprehensive overview of existing models, validation, and clinical applications. **Breast Cancer Research and Treatment**, v. 164, n. 2, p. 263–284, 2017.

COELHO, V. N. et al. A self-adaptive evolutionary fuzzy model for load forecasting problems on smart grid environment. **Applied Energy**, v. 169, p. 567–584, 2016.

COHEN, J. A coefficient of agreement for nominal scales. **Educational and Psychological Measurement**, v. 20, n. 1, p. 37- 46 ST- A coefficient of agreement for nominal, 1960.

COLLINS, A. T.; LANG, S. H. A systematic review of the validity of patient derived xenograft (PDX) models: The implications for translational research and personalised medicine. **PeerJ**, v. 2018, n. 11, p. 1–22, 2018.

CONSEIL NATIONAL DE L´INDUSTRIE. **The New Face of Industry in France**. Paris:

[s.n.].

COOK, N. R. Use and misuse of the receiver operating characteristic curve in risk prediction. **Circulation**, v. 115, n. 7, p. 928–935, 2007.

CORPORATION, N. I. **Watchdog AgentTM**, [s.d.]. Disponível em: <http://www.ni.com/datasheet/pdf/en/ds-373>. Acesso em: 4 jul. 2019

COSTA, A. F. B.; EPPRECHT, E. K.; CARPINETTI, L. C. R. **Controle Estatístico de Qualidade**. São Paulo: Atlas S.A., 2004.

COUTO, M. **Canadá abre 200 vagas para brasileiros**. Disponível em: <https://br.financas.yahoo.com/noticias/canada-abre-200-vagas-para-brasileiros-salario-anual-medio-e-de-r-403-mil-190752741.html>. Acesso em: 8 out. 2020.

DAGNINO, A. Condition Monitoring of Rotating Machines in Power Generation Plants: A Real-World Example. **Analytics in the Industrial Internet of Things**, n. 2019, p. 138–150, 2018.

DAINES, L. et al. Systematic review of clinical prediction models to support the diagnosis of asthma in primary care. **npj Primary Care Respiratory Medicine**, v. 29, n. 1, p. 1–9, 2019.

DAVENPORT, T. H.; PATIL, D. J. **Data Scientist: The Sexiest Job of the 21st Century**. Disponível em: <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>. Acesso em: 27 out. 2020.

DEBUSE, J. C. W. et al. **Industrial Knowledge Management. Building the KDD Roadmap**. 1° ed ed. London: Springer-Verlag London Ltd., 2001.

DELGADO, R.; TIBAU, X. A. Why Cohen's Kappa should be avoided as performance measure in classification. **PLoS ONE**, v. 14, n. 9, p. 1–26, 2019.

DEUTSCH, J.; HE, D. Using Deep Learning-Based Approach to Predict Remaining Useful Life of Rotating Components. **IEEE Transactions on Systems, Man, and Cybernetics: Systems**, p. 1–10, 2017.

DIAZ-ROZO, J.; BIELZA, C.; LARRAÑAGA, P. Machine Learning-based CPS for Clustering High throughput Machining Cycle Conditions. **Procedia Manufacturing**, v. 10, p. 997–1008, 2017.

DIEZ-OLIVAN, A. et al. Data fusion and machine learning for industrial prognosis : Trends and perspectives towards Industry 4 . 0. **Information Fusion**, v. 50, n. October 2018, p. 92–111, 2019.

DIJKLAND, S. A. et al. Prognosis in moderate and severe traumatic brain injury: A systematic review of contemporary models and validation studies. **Journal of Neurotrauma**, v. 37, n. 1, p. 1–13, 2020.

DOGAN, A.; BIRANT, D. Machine learning and data mining in manufacturing. **Expert Systems with Applications**, v. 166, p. 114060, 2021.

DU, R. et al. Machine learning application for the prediction of SARS-CoV-2 infection using blood tests and chest radiograph. **Scientific Reports**, v. 11, n. 1, p. 1–13, 2021.

EKER, S. et al. Model validation: A bibliometric analysis of the literature. **Environmental Modelling and Software**, v. 117, n. March, p. 43–54, 2019.

ESCOBAR, C. A. et al. Process-Monitoring-for-Quality—Applications. **Manufacturing Letters**, v. 16, p. 14–17, 2018.

EUROPEAN COMMISSION. **Factories of the Future PPP: Towards Competitive EU Manufacturing.** Bruxelles: [s.n.].

EVERITT, B. S. et al. **Cluster Analysis**. 5th ed ed. London: Wiley & Sons Ltd, 2011.

EXAME, R. **Temos mais dados do que nunca. Como usá-los a nosso favor?** Disponível em: <https://exame.com/carreira/dados-uso-favor/>. Acesso em: 23 ago. 2021.

FAVI, C. et al. Big data analysis for the estimation of disassembly time and de-manufacturing activity. **Procedia CIRP**, v. 90, p. 617–622, 2020.

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861–874, 2006.

FAWCETT, T.; PROVOST, F. Data Science and Its Relationship to Big Data and Data-Driven Decision Making Data Science. **Big Data**, n. 1, p. 51–59, 2013.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The KDD Process Knowledge from Volumes of Data. **Communication of the ACM**, v. 39, n. 11, p. 27–34, 1996a.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. Knowledge Discovery and Data Mining: Towards a Unifying Framework. **Int Conf on Knowledge Discovery and Data Mining**, p. 82–88, 1996b.

FEIDAS, H. et al. Validation of the H-SAF precipitation product H03 over Greece using rain gauge data. **Theoretical and Applied Climatology**, v. 131, n. 1–2, p. 377–398, 2018.

FLATH, C. M.; STEIN, N. Towards a data science toolbox for industrial analytics applications. **Computers in Industry**, v. 94, p. 16–25, 2018.

FLUSS, R.; FARAGGI, D.; REISER, B. Estimation of the Youden Index and its associated cutoff point. **Biometrical Journal**, v. 47, n. 4, p. 458–472, 2005.

FOGLIATTO, F. S.; DA SILVEIRA, G. J. C.; BORENSTEIN, D. The mass customization decade: An updated review of the literature. **International Journal of Production Economics**, v. 138, n. 1, p. 14–25, 2012.

FOLEY, R. N.; COLLINS, A. J. The USRDS: What you need to know about what it can and can't tell us about ESRD. **Clinical Journal of the American Society of Nephrology**, v. 8, n. 5, p. 845–851, 2013.

FORESIGHT. **The Future of Manufacturing: A New era of Opportunity and Challenge for the UK**. London: [s.n.].

FUJISHIMA, M. et al. Study of sensing technologies for machine tools. **CIRP Journal of Manufacturing Science and Technology**, v. 14, p. 71–75, 2016.

GAJJAR, S.; PALAZOGLU, A. A data-driven multidimensional visualization technique for process fault detection and diagnosis. **Chemometrics and Intelligent Laboratory Systems**, v. 154, p. 122–136, 2016.

GANDOMI, A.; HAIDER, M. Beyond the hype : Big data concepts , methods , and analytics. **International Journal of Information Management**, v. 35, n. 2, p. 137–144, 2015.

GAO, R. X. et al. A Comparative Study on Machine Learning Algorithms for Smart Manufacturing: Tool Wear Prediction Using Random Forests. **Journal of Manufacturing Science and Engineering**, v. 139, n. 7, p. 071018, 2017.

GIL, A. C. **Como Elaborar Projetos de Pesquisa**. São Paulo: Atlas, 1991.

GIL, A. C. **Métodos e Técnicas de Pesquisa Social**. São Paulo: Atlas, 1999.

GIL, A. C. **Como Elaborar Projetos de Pesquisa**. 4° ed ed. São Paulo: Atlas, 2002.

GOEDHART, R.; WOODALL, W. H. Monitoring proportions with two components of common cause variation. **Journal of Quality Technology**, v. 0, n. 0, p. 1–31, 2021.

GÓMEZ-JIMÉNEZ, G. et al. **The OECD Principles for (Q)SAR Models in the Context of Knowledge Discovery in Databases (KDD)**. 1. ed. [s.l.] Elsevier Inc., 2018. v. 113

GORECKY, D. et al. Human-machine-interaction in the industry 4.0 era. **Proceedings - 2014 12th IEEE International Conference on Industrial Informatics, INDIN 2014**, p. 289–294, 2014.

GRANT, E. L. **Statistical Quality Control**. Nova York: McGraw Hill, 1964.

GROOVER, M. P. **Automação Industrial e Sistemas de Manufatura**. 3° ed ed. São Paulo: Hall, Pearson Prentice, 2011.

GUH, R. S.; SHIUE, Y. R. An effective application of decision tree learning for on-line detection of mean shifts in multivariate control charts. **Computers and Industrial Engineering**, v. 55, n. 2, p. 475–493, 2008.

HAGAN, J.; LI, B. Phase II Performance of P-Charts and P'-Charts. **Journal of Medical Statistics and Informatics**, v. 6, n. 1, p. 3, 2018.

HAIR, J. F. et al. **Análise Multivariada de Dados**. 6ª edição ed. Porto Alegre: Bookman, 2009.

HALTIA, U.-M. et al. FIGO 1988 versus 2009 staging for endometrial carcinoma: A comparative study on prediction of survival and stage distribution according to histologic subtype. **Journal of Gynecologic Oncology**, v. 25, n. 1, p. 30–35, 2014.

HAND, D. J. Measuring classifier performance: A coherent alternative to the area under the ROC curve. **Machine Learning**, v. 77, n. 1, p. 103–123, 2009.

HARIDY, S.; WU, Z.; CASTAGLIOLA, P. Univariate and multivariate approaches for evaluating the capability of dynamic-behavior processes (case study). **Statistical Methodology**, v. 8, n. 2, p. 185–203, 2011.

HAZEN, B. T. et al. Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. **International Journal of Production Economics**, v. 154, p. 72–80, 2014.

HE, M.; HE, D. A Deep Learning Based Approach for Bearing Fault Diagnosis. **IEEE Transactions on Industry applications**, n. 1, 2017.

HE, Q. P. et al. Statistical Process Monitoring for IoT-Enabled Cybermanufacturing: Opportunities and Challenges. **IFAC-PapersOnLine**, v. 50, n. 1, p. 14946–14951, 2017.

HE, Q. P.; WANG, J. Statistical process monitoring as a big data analytics tool for smart manufacturing. **Journal of Process Control**, v. 67, p. 35–43, 2018a.

HE, Q. P.; WANG, J. Statistics Pattern Analysis: A Statistical Process Monitoring Tool for Smart Manufacturing. In: EDEN, M. R.; IERAPETRITOU, M. G.; TOWLER, G. P. (Eds.). . **13th International Symposium on Process Systems Engineering (PSE 2018)**. Computer Aided Chemical Engineering. [s.l.] Elsevier, 2018b. v. 44p. 2071–2076.

HE, S. et al. Multivariate process monitoring and fault identification using multiple decision

tree classifiers. **International Journal of Production Research**, v. 51, n. 11, p. 3355–3371, 2013.

HECKLAU, F. et al. Holistic Approach for Human Resource Management in Industry 4.0. **Procedia CIRP**, v. 54, p. 1–6, 2016.

HELLEBRANDT, T. et al. Conceptual approach to integrated human-centered performance management on the shop floor. **Advances in Intelligent Systems and Computing**, v. 783, p. 309–321, 2019.

HERNÁNDEZ-ORALLO, J.; FLACH, P.; FERRI, C. A unified view of performance metrics: Translating threshold choice into expected classification loss. **Journal of Machine Learning Research**, v. 13, p. 2813–2869, 2012.

HOFFMANN, F. et al. Benchmarking in classification and regression. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 9, n. 5, p. 1–17, 2019.

HRYNIEWICZ, O. SPC of Processes with Predicted Data: Application of the Data Mining Methodology. **Frontiers in Statistical Quality Control 11**, v. 11, p. 219–235, 2015.

HUANG, C. et al. Enhancing the prediction of acute kidney injury risk after percutaneous coronary intervention using machine learning techniques: A retrospective cohort study. **PLoS Medicine**, v. 15, n. 11, 2018.

HUANG, T. et al. MOST: Most-similar ligand based approach to target prediction. **BMC Bioinformatics**, v. 18, n. 1, 2017.

ITAYA, T. et al. Temporal Validation of an Assessment Tool that Predicts a Possibility of Home Discharge for Patients with Acute Stroke. **Journal of Stroke and Cerebrovascular Diseases**, v. 31, n. 1, p. 106188, 2022.

JACKSON, D. A.; SOMERS, K. M.; HARVEY, H. H. Similarity Coefficients : Measures of Co-Occurrence and Association or Simply Measures of Occurrence ? **The American Naturalist**, v. 133, n. 3, p. 436–453, 1989.

JAGUSIAK-KOCIK, M. Pdca Cycle As a Part of Continuous Improvement in the Production Company - a Case Study. **Production Engineering Archives**, v. 14, p. 19–22, 2017.

JARDIM, F. S.; CHAKRABORTI, S.; EPPRECHT, E. K. effect of the amount of phase I data on the conditional performance of phase II $\bar{Xb}$chart. 2018.

JARDIM, F. S.; CHAKRABORTI, S.; EPPRECHT, E. K. $\bar{X}$ Chart with Estimated Parameters: The Conditional ARL Distribution and New Insights. **Production and Operations Management**, v. 0, n. 0, p. 1–13, 2019.

JARDIM, F. S.; CHAKRABORTI, S.; EPPRECHT, E. K. Two perspectives for designing a phase II control chart with estimated parameters: The case of the Shewhart Chart. **Journal of Quality Technology**, v. 52, n. 2, p. 198–217, 2020.

JENSEN, W. A. et al. Effects of Parameter Estimation on Control Chart Properties : A Literature Review. **Journal of Quality Technology**, v. 38, n. 4, p. 349–364, 2006.

JERNTORP, P.; BERGLUND, G. Stroke registry in Malmö, Sweden. **Stroke**, v. 23, n. 3, p. 357–361, 1992.

JIMÉNEZ-VALVERDE, A.; LOBO, J. M. Threshold criteria for conversion of probability of species presence to either-or presence-absence. **Acta Oecologica**, v. 31, n. 3, p. 361–369,

2007.

JIRKOVSKÝ, V.; OBITKO, M.; MARÍK, V. Understanding Data Heterogeneity in the Context of Cyber-Physical Systems Integration. **IEEE Transactions on Industrial Informatics**, p. 1–8, 2016.

JONES-FARMER, A. L.; STEVENS, N. T. Discussion of "Bridging the gap between theory and practice in basic statistical process monitoring". **Quality Engineering**, v. 29, p. 00–00, 2016.

JONES-FARMER, L. A. et al. An Overview of Phase I Analysis for Process Improvement and Monitoring. **Journal of Quality Technology**, v. 46, n. 3, p. 265–280, 2017.

JOPPEN, R. et al. Key performance indicators in the production of the future. **Procedia CIRP**, v. 81, p. 759–764, 2019.

JOSEPH, R. C.; JOHNSON, N. A. Big data and transformational government. **IT Professional**, v. 15, n. 6, p. 43–48, 2013.

KAARE, K. K.; OTTO, T. Smart health care monitoring technologies to improve employee performance in manufacturing. **Procedia Engineering**, v. 100, n. January, p. 826–833, 2015.

KAGERMANN, H.; WAHLSTER, W.; HELBIG, J. Recommendations for Implementing the Strategic Iniciative Industrie 4.0: Securing the Future of German Manufacturing Industry. **Final Reporte Industrie 4.0 Working Group of Acatech**, 2013.

KAKOSIMOS, K. E. et al. Operational Street Pollution Model (OSPM) - A review of performed application and validation studies, and future prospects. **Environmental Chemistry**, v. 7, n. 6, p. 485–503, 2010.

KAMBLE, S. S.; GUNASEKARAN, A.; GAWANKAR, S. A. Sustainable Industry 4.0 framework: A systematic literature review identifying the current trends and future perspectives. **Process Safety and Environmental Protection**, n. 117, p. 408–425, 2018.

KANG, H. S. et al. Smart Manufacturing : Past Research , Present Findings , and Future Directions. **International Journal of Precision Engineering and Manufacturing - Green Technology**, v. 3, n. 1, p. 111–128, 2016.

KATANODA, K. et al. Short-Term projection of cancer incidence in Japan using an age-period interaction model with spline smoothing. **Japanese Journal of Clinical Oncology**, v. 44, n. 1, p. 36–41, 2014.

KAUFMAN, Y. J. et al. Passive remote sensing of tropospheric aerosol and atmospheric correction for the aerosol effect. **Journal of Geophysical Research Atmospheres**, v. 102, n. 14, p. 16815–16830, 1997.

KIM, D. et al. Machine learning-based novelty detection for faulty wafer detection in semiconductor manufacturing. **Expert Systems With Applications**, v. 39, n. 4, p. 4075–4083, 2012.

KÖNIG, I. R. et al. Practical experiences on the necessity of external validation. **Statistics in Medicine**, v. 26, p. 5499–5511, 2007.

KROIS, J. et al. Evaluating Modeling and Validation Strategies for Tooth Loss. **Journal of Dental Research**, v. 98, n. 10, p. 1088–1095, 2019.

KUHN, M. Building predictive models in R using the caret package. **Journal of Statistical Software**, v. 28, n. 5, p. 1–26, 2008.

KUNCHEVA, L. L. Using control charts for detecting concept change in streaming data. **School of Computer Science, Bangor University, …**, p. 1–13, 2009.

KÜPPER, D. et al. **Quality 4 . 0 Takes More Than Technology**Boston Consulting Group, , 2019.

KUSIAK, A.; KURASEK, C. Data Mining of Printed-Circuit Board Defects. **Ieee Transactions on Robotics and Automation**, v. 17, n. 2, p. 191–196, 2001.

LACERDA, D. P. et al. Design Science Research: A research method to production engineering. **Gestão & Produção**, v. 20, n. 4, p. 741–761, 2013.

LAKATOS, E. M.; MARCONI, M. A. **Fundamentos da Metologia Científica**. São Paulo: Atlas, 1995.

LANDIS, J. R.; KOCH, G. G. The Measurement of Observer Agreement for Categorical Data. **Biometrics**, v. 33, n. 1, p. 159–174, 1977.

LANDSET, S. et al. A survey of open source tools for machine learning with big data in the Hadoop ecosystem. **Journal of Big Data**, p. 1–36, 2015.

LANG, J. R.; BOLTON, S. A comprehensive method validation strategy for bioanalytical applications in the pharmaceutical industry - 2. Statistical analyses. **Journal of Pharmaceutical and Biomedical Analysis**, v. 9, n. 5, p. 357–361, 1991a.

LANG, J. R.; BOLTON, S. A comprehensive method validation strategy for bioanalytical applications in the pharmaceutical industry -1. Statistical analyses. **Journal of Pharmaceutical and Biomedical Analysis**, v. 9, n. 6, p. 435–442, 1991b.

LASI, H. et al. Industry 4.0. **Smart Innovation, Systems and Technologies**, v. 132, p. 205–215, 2019.

LEE, J. et al. Recent advances and trends in predictive manufacturing systems in big data environment. **Manufacturing Letters**, v. 1, n. 1, p. 38–41, 2013.

LEE, J. et al. Implementation of cyber-physical production systems for quality prediction and operation control in metal casting. **Sensors (Switzerland)**, v. 18, n. 5, 2018.

LEE, J.; BAGHERI, B.; JIN, C. Introduction to Cyber Manufacturing. **Manufacturing Letters**, n. May, 2016.

LEE, J.; JIN, C.; BAGHERI, B. Cyber physical systems for predictive production systems. **Production Engineering**, p. 1–11, 2017.

LI, D. et al. A big data enabled load-balancing control for smart manufacturing of Industry 4 . 0. **cluster Comput**, n. 20, p. 1855–1864, 2017.

LI, J. P. et al. Machine learning and credit ratings prediction in the age of fourth industrial revolution. **Technological Forecasting and Social Change**, v. 161, n. September, p. 120309, 2020.

LI, K. **Made in China 2025**. Beijing: [s.n.].

LIAO, Y. et al. Past , present and future of Industry 4 . 0 - a systematic literature review and research agenda proposal. **International Journal of Production Research**, n. November, p. 3609–3629, 2017.

Linkedin. Disponível em: <linkedin.com>. Acesso em: 27 out. 2020.

LIU, C. L. et al. Predicting Short-term Survival after Liver Transplantation using Machine

Learning. **Scientific Reports**, v. 10, n. 1, p. 1–10, 2020.

LIZARELLI, F. L. et al. A bibliometric analysis of 50 years of worldwide research on statistical process control. **Gestão da Produção**, v. 23, n. 4, p. 853–870, 2016.

LOBO, J. M.; JIMÉNEZ-VALVERDE, A.; REAL, R. AUC: A misleading measure of the performance of predictive distribution models. **Global Ecology and Biogeography**, v. 17, n. 2, p. 145–151, 2008.

LONGO, F.; NICOLETTI, L.; PADOVANO, A. Smart operators in industry 4.0: A human-centered approach to enhance operators' capabilities and competencies within the new smart factory context. **Computers and Industrial Engineering**, v. 113, p. 144–159, 2017.

LU, Y. Industry 4.0: A survey on technologies, applications and open research issues. **Journal of Industrial Information Integration**, v. 6, p. 1–10, 2017.

LU, Y.; XU, X.; XU, J. Development of a Hybrid Manufacturing Cloud. **Journal of Manufacturing Systems**, v. 33, n. 4, p. 551–566, 2014.

LUO, M. et al. A data-driven two-stage maintenance framework for degradation prediction in semiconductor manufacturing industries. **Computers and Industrial Engineering**, v. 85, p. 414–422, 2015.

LUQUE, A. et al. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. **Pattern Recognition**, v. 91, p. 216–231, 2019.

MALAN, A. K. et al. Evaluations of Commercial West Nile Virus Immunoglobulin G (IgG) and IgM Enzyme Immunoassays Show the Value of Continuous Validation. **Journal of Clinical Microbiology**, v. 42, n. 2, p. 727–733, 2004.

MANNIËN, J. et al. Validation of Surgical Site Infection Surveillance in The Netherlands. **Infection Control & Hospital Epidemiology**, v. 28, n. 1, p. 36–41, 2007.

MANYIKA, J. et al. **Big data : The next frontier for innovation , competition , and productivity**. London: [s.n.]. Disponível em: <www.mckinsey.com/mgi>.

MARISCAL, G.; MARBÁN, Ó.; FERNÁNDEZ, C. A survey of data mining and knowledge discovery process models and methodologies. **Knowledge Engineering Review**, v. 25, n. 2, p. 137–166, 2010.

MATTHEWS, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. **BBA - Protein Structure**, v. 405, n. 2, p. 442–451, 1975.

MAURO, A. DE et al. A formal definition of Big Data based on its essential features. **emerald insight**, 2016.

MAURO, A. DE et al. Human resources for Big Data professions : A systematic classification of job roles and required skill sets. **Information Processing and Management**, v. 0, p. 1–11, 2017.

MAYNARD, A. D. "Navigating the Fourth Industrial Revolution". **Nature Nanotechnology**, p. 1005–1006, 2015.

MINISTÉRIO DA INDÚSTRIA E COMÉRCIO E SERVIÇOS. **Agenda Brasileira para a Indústria 4.0**. Disponível em: <http://www.industria40.gov.br/>. Acesso em: 20 jun. 2019.

MINNE, L. et al. Statistical process control for monitoring standardized mortality ratios of a classification tree model. **Methods of Information in Medicine**, v. 51, n. 4, p. 353–358, 2012a.

MINNE, L. et al. Statistical process control for validating a classification tree model for predicting mortality - A novel approach towards temporal validation. **Journal of Biomedical Informatics**, v. 45, n. 1, p. 37–44, 2012b.

MINNE, L. et al. Effect of changes over time in the performance of a customized SAPS-II model on the quality of care assessment. **Intensive Care Medicine**, v. 38, n. 1, p. 40–46, 2012c.

MOEUF, A. et al. The industrial management of SMEs in the era of Industry 4 . 0. **International Journal of Production Research**, n. September, p. 0, 2017.

MOHAMMADIAN, F.; AMIRI, A. Economic-statistical design of acceptance control chart. **Quality and Reliability Engineering International**, v. 29, n. 1, p. 53–61, 2013.

MUHURI, P. K.; SHUKLA, A. K.; ABRAHAM, A. Industry 4 . 0 : A bibliometric analysis and detailed overview. **Engineering Applications of Artificial Intelligence**, v. 78, n. September 2018, p. 218–235, 2019.

MUKHERJEE, P. S. On phase II monitoring of the probability distributions of univariate continuous processes. **Statistical Papers**, v. 57, n. 2, p. 539–562, 2015.

MUNKHOLM, S. B. et al. Validation of post-operative atrial fibrillation in the western denmark heart registry. **Danish Medical Journal**, v. 62, n. 12, p. 4, 2015.

MYLLYAHO, L. et al. Systematic literature review of validation methods for AI systems. **Journal of Systems and Software**, v. 181, p. 111050, 2021.

NAMDARI, M.; JAZAYERI-RAD, H. Incipient fault diagnosis using support vector machines based on monitoring continuous decision functions. **Engineering Applications of Artificial Intelligence**, v. 28, p. 22–35, 2014.

NATIONAL RESEARCH FOUNDATION. **Research, Innovation and Enterprise (RIE) 2015 Plan**. Singapore: [s.n.].

NEELY, A.; GREGORY, M.; PLATTS, K. Performance measurement system design: A literature review and research agenda. **International Journal of Operations and Production Management**, v. 25, n. 12, p. 1228–1263, 2005.

O'DONOVAN, P. et al. An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities. **Journal of Big Data**, v. 2, n. 1, p. 1–26, 2015a.

O'DONOVAN, P. et al. Big data in manufacturing: a systematic mapping study. **Journal of Big Data**, v. 2, n. 1, p. 1–22, 2015b.

ODURO, A. R. et al. Monitoring malaria using health facility based surveys: Challenges and limitations. **BMC Public Health**, v. 16, n. 1, 2016.

OLIVEIRA, E. M. **Uso da Regressão Logística: Desafios, datasets e comunidade no Kaggle**. Disponível em: <https://medium.com/industrial-insights/data-science-na-indústria-92063122dac7>. Acesso em: 1 dez. 2020.

OPRIME, P. C.; MENDES, G. H. DE S. The X-bar control chart with restriction of the capability indices. **International Journal of Quality and Reliability Management**, v. 34, n. 1, p. 38–52, 2017.

OTTONI, H. M. et al. Anais da Academia Brasileira de Ciências e o Ponto T de Goffman: estudo exploratório. v. 9, n. 1, p. 269–283, 2013.

ÖZKUNDAKCI, D. et al. Building a reliable evidence base: Legal challenges in environmental decision-making call for a more rigorous adoption of best practices in environmental modelling. **Environmental Science and Policy**, v. 88, n. June, p. 52–62, 2018.

PARA, J. et al. Analyze, Sense, Preprocess, Predict, Implement, and Deploy (ASPPID): An incremental methodology based on data analytics for cost-efficiently monitoring the industry 4.0. **Engineering Applications of Artificial Intelligence**, v. 82, n. March, p. 30–43, 2019.

PARADY, G.; ORY, D.; WALKER, J. The overreliance on statistical goodness-of-fit and under-reliance on model validation in discrete choice models: A review of validation practices in the transportation academic literature. **Journal of Choice Modelling**, v. 38, n. February 2020, p. 100257, 2021.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 19, n. 1, p. 29–33, 2015.

PEPE, M. S. Receiver Operating Characteristic Methodology. **Journal of the American Statistical Association**, v. 95, n. 449, p. 308–311, 2012.

PFEIFFER, S. Robots, Industry 4.0 and Humans, or Why Assembly Work Is More than Routine Work. **Societies**, v. 6, n. 2, p. 16, 2016.

PIATETSKY-SHAPHIRO, G.; FRAWLEY, W. **Knowledge Discovery in Database**. [s.l.] AAAI/MIT Press, 1991.

PILLONI, V. How data will transform industrial processes: Crowdsensing, crowdsourcing and big data as pillars of industry 4.0. **Future Internet**, v. 10, n. 4, 2018.

PIRES, S. R.; MEDEIROS, R. B. Mammographic system performance using an image reading qualification method. **Radiological Physics and Technology**, v. 5, n. 2, p. 213–221, 2012.

POWERS, D. M. W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. **International Journal of Machine Learning Technology**, p. 37–63, 2020.

PREDIX™. **General Eletric Predix**. Disponível em: <https://www.ge.com/digital/iiot-platform>. Acesso em: 30 jun. 2019.

PREUVENEERS, D.; ILIE-ZUDOR, E. The intelligent industry of the future : A survey on emerging trends , research challenges and opportunities in Industry 4 . 0. **Journal of Ambient Intelligence and Smart Environments**, n. 9, p. 287–298, 2017.

QIN, J.; LIU, Y.; GROSVENOR, R. A Categorical Framework of Manufacturing for Industry 4.0 and beyond. **Procedia CIRP**, v. 52, p. 173–178, 2016.

QIN, S. J. Survey on data-driven industrial process monitoring and diagnosis. **Annual Reviews in Control**, v. 36, n. 2, p. 220–234, 2012.

QIN, S. J. Process Data Analytics in the Era of Big Data. **AIChE Journal**, v. 60, n. 9, p. 3092–3100, 2014.

RAJAMANICKAM, V. et al. About Model Validation in Bioprocessing. **Processes**, p. 1–16, 2021.

REFSGAARD, J. C.; HENRIKSEN, H. J. Modelling guidelines - Terminology and guiding principles. **Advances in Water Resources**, v. 27, n. 1, p. 71–82, 2004.

REIF, R.; SHIRLEY, A. J.; LIVERIS, A. **Report To The President Accelerating U.S.**

Advanced Manufacturing. Washington, DC: [s.n.].

REIS, M. S. A Systematic Framework for Assessing the Quality of Information in Data-Driven Applications for the Industry 4.0. **IFAC-PapersOnLine**, v. 51, n. 18, p. 43–48, 2018.

REIS, M. S.; BAUER, A. Wavelet texture analysis of on-line acquired images for paper formation assessment and monitoring. **Chemometrics and Intelligent Laboratory Systems**, v. 95, n. 2, p. 129–137, 2009.

REIS, M. S.; GINS, G. Industrial process monitoring in the big data/industry 4.0 era: From detection, to diagnosis, to prognosis. **Processes**, v. 5, n. 35, p. 1–16, 2017.

REIS, M. S.; SARAIVA, P. M. Prediction of Profiles in the Process Industries. **industrial & Engineering Chemistry Research**, v. 51, p. 4254–4266, 2012.

RICHTER, A. et al. Validation strategy for satellite observations of tropospheric reactive gases. **Annals of Geophysics**, v. 56, n. FAST TRACK 1, p. 1–10, 2014.

RIMINUCCI, M. Industry 4.0 and Human Resources Development: A View from Japan. **E-Journal of International and Comparative Labour Studies**, v. 7, n. 1, 2018.

SALEH, N. A. et al. A Review and Critique of Auxiliary Information-Based Process Monitoring Methods. n. 1947, 2021.

SALSBURG, D. **Uma Senhora Toma Chá: como a estatística revolucionou a ciência no século XX**. 1° ed ed. Rio de Janeiro: Zahar, 2009.

SAMIR, K. et al. Key Performance Indicators in Cyber-Physical Production Systems. **Procedia CIRP**, v. 72, p. 498–502, 2018.

SANDERS, A.; ELANGESWARAN, C.; WULFSBERG, J. Industry 4.0 implies lean manufacturing: Research activities in industry 4.0 function as enablers for lean manufacturing. **Journal of Industrial Engineering and Management**, v. 9, n. 3, p. 811, 2016.

SANTOS, L. et al. The expected contribution of Industry 4 . 0 technologies for industrial performance. **Intern. Journal of Production Economics**, v. 204, n. August, p. 383–394, 2018.

SCHMIDT, M. et al. The Danish National patient registry: A review of content, data quality, and research potential. **Clinical Epidemiology**, v. 7, p. 449–490, 2015.

SCHWAB, K. **The Fourth Industrial Revolution**. 1 ed ed. Cologny/Geneva Switzerland: [s.n.].

SESHIA, S. A. et al. Design Automation of Cyber-Physical Systems : Challenges , Advances , and Opportunities. **IEEE transactions on computer - Aided Design of Integrated Circuits and Systems**, p. 1–15, 2016.

SEVERSON, K.; CHAIWATANODOM, P.; BRAATZ, R. D. Perspectives on process monitoring of industrial systems. **Annual Reviews in Control**, v. 42, p. 190–200, 2016.

SHARMA, A. B. et al. Modeling and analytics for cyber-physical systems in the age of big data. **ACM SIGMETRICS Performance Evaluation Review**, v. 41, n. 4, p. 74–77, 2014.

SHEWHART, W. A. **Economic Control of Quality of Manufactured Product**. 7th. ed. New York: D. Van Nostrand Company, Inc., 1931.

SHMUEI, G.; KOPPIUS, O. R. Predictive Analytics in Information Systems Research. **MIS**

Quartely, v. 35, n. 3, p. 553–572, 2011.

SHU, Z. et al. Cloud-Integrated Cyber-Physical Systems for Complex Industrial Applications. **Mobile network applications**, 2015.

SILVA, E. L.; MENEZES, E. M. **Metodologia de Pesquisa e Elaboração de Dissertação**. Florianópolis: [s.n.].

SILVA, R. DE S. E; PAES, A. T. Teste de Concordância Kappa. **Educaçao Continuada em Saúde einstein**, v. 10, n. 4, p. 165–166, 2012.

SIRYANI, J.; TANJU, B.; EVELEIGH, T. J. A Machine Learning Decision-Support System Improves the Internet of Things' Smart Meter Operations. **IEEE Internet of Things Journal**, v. 4, n. 4, p. 1056–1066, 2017.

SIVARAJAH, U. et al. Critical analysis of Big Data challenges and analytical methods. **Journal of Business Research**, v. 70, p. 263–286, 2017.

SMITH, L. A.; PETERSEN, A. C. Variations on Reliability: Connecting Climate Predictions to Climate Policy. In: M. BOUMANS, G. HON, A. C. P. (Ed.). . **Error and Uncertainty in Scientific Practice**. London: Pickering & Chatto, 2014. p. 137–156.

SOBUE, F. et al. Unconditional performance of the $\bar{X}$ chart: Comparison among five standard deviation estimators. **Quality and Reliability Engineering International**, v. 36, n. 5, p. 1808–1819, 2020.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. **Information Processing and Management**, v. 45, n. 4, p. 427–437, 2009.

SOKOVIC, M.; PAVLETIC, D.; PIPAN, K. K. Quality Improvement Methodologies - PDCA Cycle, RADAR Matrix, DMAIC and DFSS. **Journal of achievements un Materials and Manufacturing Engineering**, v. 43, n. 1, p. 476–483, 2010.

SOLUTIONS, B. A. D. **National Instruments**. Disponível em: <https://www.ni.com/pt-br/innovations/white-papers/13/big-analog-data--solutions.html>. Acesso em: 4 jul. 2019.

SONG, I.; ZHU, Y. Big data and data science: what should we teach? **Expert Systems**, v. 00, n. 00, 2015.

SORNETTE, D. et al. Algorithm for model validation: Theory and applications. **Proceedings of the National Academy of Sciences of the United States of America**, v. 104, n. 16, p. 6562–6567, 2007.

SPENCER, N. H. **Essentials of Multivariate Data Analysis**. 1º ed ed. Boca Raton, FL - EUA: CRC Press, 2014.

SPOKOINY, A.; SHAHAR, Y. An active database architecture for knowledge-based incremental abstraction of complex concepts from continuously arriving time-oriented raw data. **Journal of Intelligent Information Systems**, v. 28, n. 3, p. 199–231, 2007.

SRIKAEO, K.; HOURIGAN, J. A. The use of statistical process control (SPC) to enchance the validation of critical control points (CCPs) in shell egg washing. **Food Control**, v. 13, n. 4–5, p. 263–273, 2002.

STEINBERG, D. M. Industrial statistics : The challenges and the research. **Quality Engineering**, v. 28, n. 1, p. 45–59, 2016.

STETCO, A. et al. Machine learning methods for wind turbine condition monitoring : A

review. **Renewable Energy**, v. 133, p. 620–635, 2019.

STOJANOVIC, L. ET AL. Big-data- driven Anomaly Detection in Industry (4 . 0): an approach and a case study. **IEEE International Conference on Big Data**, p. 1647–1652, 2016.

STOJANOVIC, N.; MILENOVIC, D. Data-driven Digital Twin approach for process optimization: An industry use case. **Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018**, p. 4202–4211, 2018.

STROBL, A. N. et al. The next Generation of Clinical Decision-Making Tools: Development of a Real Time Prediction Tool for Outcome of Prostate Biopys in response to a continuously evolving Prostate Cancer Landscape. **Physiology & behavior**, v. 176, n. 1, p. 100–106, 2016.

SU, T.-L. et al. A review of statistical updating methods for clinical prediction models. **Statistical Methods in Medical Research**, v. 27, n. 1, p. 185–197, 2018.

SZYMAŃSKA, E. Modern data science for analytical chemical data – A comprehensive review. **Analytica Chimica Acta**, v. 1028, p. 1–10, 2018.

TAO, F. et al. Data-driven smart manufacturing. **Journal of Manufacturing Systems**, 2018.

TAO, F. E. I.; ZHANG, M. Digital Twin Shop-Floor : A New Shop-Floor Paradigm Towards Smart Manufacturing. **IEEE Access**, v. 5, 2017.

TOMÉ, J. **Falta a Portugal alguns milhares de cientistas de dados**. Disponível em: <https://www.dinheirovivo.pt/tecnologia/faltam-a-portugal-alguns-milhares-de-cientistas-de-dados/>. Acesso em: 21 out. 2020.

TOOLE, J. F. et al. Study design for randomized prospective trial of carotid endarterectomy for asymptomatic atherosclerosis. **Stroke**, v. 20, n. 7, p. 844–849, 1989.

VAN VLIET, J. et al. A review of current calibration and validation practices in land-change modeling. **Environmental Modelling and Software**, v. 82, p. 174–182, 2016.

VAN VLIET, J.; BREGT, A. K.; HAGEN-ZANKER, A. Revisiting Kappa to account for change in the accuracy assessment of land-use change models. **Ecological Modelling**, v. 222, n. 8, p. 1367–1375, 2011.

VENKATASUBRAMANIAN, V.; RENGASWAMY, R.; YIN, K. A review of process fault detection and diagnosis Part I : Quantitati v e model-based methods. **Computers & Chemical Engineering**, v. 27, p. 293–311, 2003.

VICENTIN, D. S. et al. Monitoring process control chart with finite mixture probability distribution. **International Journal of Quality & Reliability Management**, v. 35, n. 2, p. 335–353, 2018.

WAIBEL, M. W. et al. Investigating the effects of Smart Production Systems on sustainability elements. **Procedia Manufacturing**, v. 8, n. October 2016, p. 731–737, 2017.

WALLER, M. A. et al. Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain. **Journal of Business Logistics**, v. 34, n. 2, p. 77–84, 2013.

WAMBA, S. F. et al. How “big data” can make big impact: Findings from a systematic review and a longitudinal case study. **International Journal of Production Economics**, v. 165, p. 234–246, 2015.

WAN, J. et al. Software-Defined Industrial Internet of Things in the Context of Industry 4 . 0. **IEEE Sensors Journal**, v. 16, n. 20, p. 7373–7380, 2016.

WAN, J. et al. A Manufacturing Big Data Solution for Active Preventive Maintenance. **IEEE Transactions on industrial Informatics**, v. 3203, n. c, 2017.

WANG, S. et al. Implementing Smart Factory of Industrie 4 . 0 : An Outlook. **International journal of distributed Sensor networks**, v. 2016, p. 10, 2016a.

WANG, S. et al. Towards smart factory for industry 4 . 0 : a self-organized multi-agent system with big data based feedback and coordination. **computer Networks**, v. 101, p. 158–168, 2016b.

WATSON, H. J. Tutorial: Big Data Analytics: Concepts, Technologies, and Applications. **Communications of the Association for Information Systems**, v. 34, n. 1, p. 1191–1208, 2014.

WEESE, M. et al. Statistical Learning Methods Applied to Process Monitoring: An Overview and Perspective. **Journal of Quality Technology**, v. 48, n. 1, p. 4–24, 2016.

WESSLER, B. S. et al. Clinical Prediction Models for Valvular Heart Disease. **Journal of the American Heart Association**, v. 8, n. 20, p. e011972, 2019.

WOLLSCHLAEGER, M.; SAUTER, T.; JASPERNEITE, J. The future of industrial communication: Automation networks in the era of the internet of things and industry 4.0. **IEEE Industrial Electronics Magazine**, v. 11, n. 1, p. 17–27, 2017.

WOODALL, W. H. The Statistical Design of Quality Control Charts. **The Statistician**, v. 34, n. 2, p. 155, 1985.

WOODALL, W. H. Bridging the gap between theory and practice in basic statistical process monitoring. **Quality Engineering**, v. 29, n. 1, p. 2–15, 2016.

WOODALL, W. H.; MONTGOMERY, D. C. **Research Issues and Ideas in Statistical Process Control**, 1999.

WOODALL, W. H.; MONTGOMERY, D. C. Some Current Directions in the Theory and Application of Statistical Process Monitoring. **Journal of Quality Technology**, v. 46, n. 1, p. 72, 2014.

WOOLLEY, T. et al. A review of logistic regression models used to predict post-fire tree mortality of western North American conifers. **International Journal Of Wildland Fire**, n. 2006, p. 1–35, 2012.

WU, D. et al. Cloud-based design and manufacturing: A new paradigm in digital manufacturing and design innovation. **CAD Computer Aided Design**, v. 59, p. 1–14, 2015.

XU, B. et al. Internet of things and big data analytics for smart oil field malfunction diagnosis. **2017 IEEE 2nd International Conference on Big Data Analysis, ICBDA 2017**, p. 178–181, 2017.

XU, L. DA; DUAN, L. Big data for cyber physical systems in industry 4 . 0 : a survey. **Enterprise Information Systems**, p. 0–22, 2018.

YAN, J. et al. Industrial Big Data in an Industry 4 . 0 Environment : Challenges , Schemes and Applications for Predictive Maintenance. **IEEE**, p. 1–9, 2017.

YIN, G. et al. Derivation of temporally continuous LAI reference maps through combining the LAINet observation system with CACAO. **Agricultural and Forest Meteorology**, v. 233, p. 209–221, 2017.

YOUDEN, W. J. Index for rating diagnostic tests. **Cancer**, v. 3, n. 1, p. 32–35, 1950.

ZHONG, R. Y. et al. Intelligent Manufacturing in the Context of Industry 4.0: A Review. **Engineering**, n. 3, p. 616–630, 2017.

ZHOU, T. J.; RAZA, S.; NELSON, K. P. Methods of assessing categorical agreement between correlated screening tests in clinical studies. **Journal of Applied Statistics**, v. 0, n. 0, p. 1–21, 2020.

ZHU, Q. On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset. **Pattern Recognition Letters**, v. 136, p. 71–80, 2020.

ZHUANG, C.; LIU, J.; XIONG, H. Digital twin-based smart production management and control framework for the complex product assembly shop-floor. **International Journal of Advanced Manufacturing Technology**, v. 96, n. 1–4, p. 1149–1163, 2018.

APÊNDICE A

Relação dos documentos selecionados na RBS

Autores	Título	Ano
Pivodic A., Hård A.-L., Löfqvist C., Smith L.E.H., Wu C., Bründer M.-C., Lagrèze W.A., Stahl A., Holmström G., Albertsson-Wikland K., Johansson H., Nilsson S., Hellström A.	Individual Risk Prediction for Sight-Threatening Retinopathy of Prematurity Using Birth Characteristics	2020
Dijkland S.A., Foks K.A., Polinder S., Dippel D.W.J., Maas A.I.R., Lingsma H.F., Steyerberg E.W.	Prognosis in moderate and severe traumatic brain injury: A systematic review of contemporary models and validation studies	2020
Kim, N; Chang, JS; Wee, CW; Kim, IA; Chang, JH; Lee, HS; Kim, SH; Kang, SG; Kim, EH; Yoon, HI; Kim, JW; Hong, CK; Cho, J; Kim, E; Kim, TM; Kim, YJ; Park, CK; Kim, JW; Kim, CY; Choi, SH; Kim, JH; Park, SH; Choe, G; Lee, ST; Kim, IH; Suh, CO	Validation and optimization of a web-based nomogramfor predicting survival of patients with newly diagnosed glioblastoma	2020
Martinez-Zayas G., Almeida F.A., Simoff M.J., Yarmus L., Molina S., Young B., Feller-Kopman D., Sagar A.-E.S., Gildea T., Debiane L.G., Grosu H.B., Casal R.F., Arain M.H., Eapen G.A., Jimenez C.A., Noor L.Z., Baghaie S., Song J., Li L., Ost D.E.	A prediction model to help with oncologic mediastinal evaluation for radiation: Homer	2020
Lombardo L., Tanyas H.	Chrono-validation of near-real-time landslide susceptibility models via plug-in statistical simulations	2020
Clift A.K., Coupland C.A.C., Keogh R.H., Diaz-Ordaz K., Williamson E., Harrison E.M., Hayward A., Hemingway H., Horby P., Mehta N., Benger J., Khunti K., Spiegelhalter D., Sheikh A., Valabhji J., Lyons R.A., Robson J., Semple M.G., Kee F., Johnson P., Jebb S., Williams T., Hippisley-Cox J.	Living risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus 19 in adults: national derivation and validation cohort study	2020
Tavakoli H., Chen W., Sin D.D., FitzGerald J.M., Sadatsafavi M.	Predicting Severe Chronic Obstructive Pulmonary Disease Exacerbations. Developing a Population Surveillance Approach with Administrative Data	2020
Paul S., Banerjee C., Nagesh Kumar D.	Evaluation Framework of Landsat 8-Based Actual Evapotranspiration Estimates in Data-Sparse Catchment	2020
Choe J., Lee S.M., Do K.-H., Kim S., Choi S., Lee J.-G., Seo J.B.	Outcome prediction in resectable lung adenocarcinoma patients: value of CT radiomics	2020
Bedoya A.D., Futoma J., Clement M.E., Corey K., Brajer N., Lin A., Simons M.G., Gao M., Nichols M., Balu S., Heller K., Sendak M., O'Brien C.	Machine learning for early detection of sepsis: An internal and temporal validation study	2020
Staub L.P., Aghayev E., Skrivankova V., Lord S.J., Haschtmann D., Mannion A.F.	Development and temporal validation of a prognostic model for 1-year clinical outcome after decompression surgery for lumbar disc herniation	2020
Mišić V.V., Gabel E., Hofer I., Rajaram K., Mahajan A.	Machine Learning Prediction of Postoperative Emergency Department Hospital Readmission	2020
Lemaire E., Schultz P., Vergez S., Debry C., Sarini J., Vairel B., de Bonnecaze G., Takeda-Raguin C., Cabarrou B., Dupret-Bories A.	Risk Factors for Pharyngocutaneous Fistula After Total Pharyngolaryngectomy	2020
Kim T.J., Lee J.S., Oh M.-S., Kim J.-W., Yoon J.S., Lim J.-S., Lee C.-H., Mo H., Jeong H.-Y., Kim Y., Lee S.-H., Jung K.-H., Kim L.Y., An M.R., Park Y.H., Lee T.S., Heo Y.J., Ko S.-B., Yu K.-H., Lee B.-C., Yoon B.-W.	Predicting Functional Outcome Based on Linked Data After Acute Ischemic Stroke: S-SMART Score	2020
Liu C.-L., Soong R.-S., Lee W.-C., Jiang G.-W., Lin Y.-C.	Predicting Short-term Survival after Liver Transplantation using Machine Learning	2020
Pan J., Adab P., Cheng K.K., Jiang C.Q., Zhang W.S., Zhu F., Jin Y.L., Thomas G.N., Steyerberg E.W., Lam T.H.	Development and validation of a prediction model for airflow obstruction in older Chinese: Guangzhou Biobank Cohort Study	2020
Ravidà A., Troiano G., Qazi M., Saleh M.H.A., Lo Russo L., Greenwell H., Giannobile W.V., Wang H.-L.	Development of a nomogram for the prediction of periodontal tooth loss using the staging and grading system: A long-term cohort study	2020
Alcorn S.R., Fiksel J., Wright J.L., Elledge C.R., Smith T.J., Perng P., Saleemi S., McNutt T.R., DeWeese T.L., Zeger S.	Developing an Improved Statistical Approach for Survival Estimation in Bone Metastases Management: The Bone Metastases Ensemble Trees for Survival (BMETS) Model	2020
Huang Y., Douiri A., Fahey M.	A Dynamic Model for Predicting Survival up to 1 Year After Ischemic Stroke	2020
Major V.J., Aphinyanaphongs Y.	Development, implementation, and prospective validation of a model to predict 60-day end-of-life in hospitalized adults upon admission at three sites	2020
Potash E., Ghani R., Walsh J., Jorgensen E., Lohff C., Prachand N., Mansour R.	Validation of a Machine Learning Model to Predict Childhood Lead Poisoning	2020
Hsu C.-N., Liu C.-L., Tain Y.-L., Kuo C.-Y., Lin Y.-C.	Machine learning model for risk prediction of community-acquired acute kidney injury hospitalization from electronic health records: Development and validation study	2020
Cai Y., Liu S., Lin H.	Monitoring the vegetation dynamics in the dongting lake wetland from 2000 to 2019 using the BEAST algorithm based on dense landsat time series	2020
Qin X., Wang H., Hu X., Gu X., Zhou W.	Predictive models for patients with lung carcinomas to identify EGFR mutation status via an artificial neural network based on multiple clinical information	2020
Choe A.R., Ryu D.-R., Kim H.Y., Lee H.A., Lim J., Kim J.S., Lee J.K., Kim T.H., Yoo K.	Noninvasive indices for predicting nonalcoholic fatty liver disease in patients with chronic kidney disease	2020
Choi H., Kim H., Hong W., Park J., Hwang E.J., Park C.M., Kim Y.T., Goo J.M.	Prediction of visceral pleural invasion in lung cancer on CT: deep learning model achieves a radiologist-level performance with adaptive sensitivity and specificity to clinical needs	2020
Gutierrez C.S., Passos S.C., Castro S.M.J., Okabayashi L.S.M., Berto M.L., Lorenzen M.B., Caumo W., Stefani L.C.	Few and feasible preoperative variables can identify high-risk surgical patients: derivation and validation of the Ex-Care risk model	2020
Aluvaala J., Collins G., Maina B., Mutinda C., Waiyego M., Berkley J.A., English M.	Prediction modelling of inpatient neonatal mortality in high-mortality settings	2020
Espada M., Leonardi M., Aas-Eng K., Lu C., Reyftmann L., Tetstall E., Slusarczyk B., Ludlow J., Hudelist G., Reid S., Condous G.	A Multicenter International Temporal and External Validation Study of the Ultrasound-based Endometriosis Staging System	2020
Liu W., Cheng Y., Liu Z., Liu C., Cattell R., Xie X., Wang Y., Yang X., Ye W., Liang C., Li J., Gao Y., Huang C., Liang C.	Preoperative Prediction of Ki-67 Status in Breast Cancer with Multiparametric MRI Using Transfer Learning	2020

Autores	Título	Ano
Backes Y., Schwartz M.P., Ter Borg F., Wolfhagen F.H.J., Groen J.N., De Vos Tot Nederveen Cappel W.H., Van Bergeijk J., Geesing J.M.J., Spanier B.W.M., Didden P., Vleggaar F.P., Lacle M.M., Elias S.G., Moons L.M.G.	Multicentre prospective evaluation of real-time optical diagnosis of T1 colorectal cancer in large non-pedunculated colorectal polyps using narrow band imaging (the OPTICAL study)	2019
Van Donkelaar C.E., Bakker N.A., Birks J., Veeger N.J.G.M., Metzemaekers J.D.M., Molyneux A.J., Groen R.J.M., Van Dijk J.M.C.	Prediction of Outcome After Aneurysmal Subarachnoid Hemorrhage: Development and Validation of the SAFIRE Grading Scale	2019
Pan Y.-X., Chen J.-C., Fang A.-P., Wang X.-H., Chen J.-B., Wang J.-C., He W., Fu Y.-Z., Xu L., Chen M.-S., Zhang Y.-J., Li Q.-J., Zhou Z.-G.	A nomogram predicting the recurrence of hepatocellular carcinoma in patients after laparoscopic hepatectomy	2019
Stamatelopoulos K., Georgiopoulos G., Athanasouli F., Nikolaou P.-E., Lykka M., Roussou M., Gavriatopoulou M., Laina A., Trakada G., Charakida M., Delialis D., Petropoulos I., Pamboukas C., Manios E., Karakitsou M., Papamichael C., Gatsiou A., Lambrinoudaki I., Terpos E., Stellos K.,	Andreadou I., Dimopoulos M.A., Kastritis E.	Reactive vasodilation predicts mortality in primary systemic light-chain amyloidosis	2019
Tsokos A., Narayanan S., Kosmidis I., Baio G., Cucuringu M., Whitaker G., Király F.	Modeling outcomes of soccer matches	2019
Song X., Waitman L.R., Hu Y., Yu A.S.L., Robins D., Liu M.	Robust clinical marker identification for diabetic kidney disease with ensemble feature selection	2019
Gattringer, T; Posekany, A; Niederkorn, K; Knoflach, M; Poltrum, B; Mutzenbach, S; Haring, HP; Ferrari, J; Lang, W; Willeit, J; Kiechl, S; Enzinger, C; Fazekas, F	Predicting Early Mortality of Acute Ischemic Stroke Score-Based Approach	2019
Mattonen S.A., Davidzon G.A., Benson J., Leung A.N.C., Vasanawala M., Horng G., Shrager J.B., Napel S., Nair V.S.	Bone marrow and tumor radiomics at 18F-FDG PET/CT: Impact on outcome prediction in non–small cell lung cancer	2019
Gattringer T., Posekany A., Niederkorn K., Knoflach M., Poltrum B., Mutzenbach S., Haring H.-P., Ferrari J., Lang W., Willeit J., Kiechl S., Enzinger C., Fazekas F.	Predicting early mortality of acute ischemic stroke: Score-based approach	2019
Krois J., Graetz C., Holtfreter B., Brinkmann P., Kocher T., Schwendicke F.	Evaluating Modeling and Validation Strategies for Tooth Loss	2019
Fernández-Labandera C., Calvo-Bonacho E., Valdivielso P., Quevedo-Aguado L., Martínez-Munoz P., Catalina-Romero C., Ruilope L.M., Sánchez-Chaparro M.A.	Prediction of fatal and non-fatal cardiovascular events in young and middle-aged healthy workers: The IberScore model	2019
Rotigliano E., Martinello C., Hernandéz M.A., Agnesi V., Conoscenti C.	Predicting the landslides triggered by the 2009 96E/Ida tropical storms in the Ilopango caldera area (El Salvador, CA): optimizing MARS-based model building and validation strategies	2019
Findlay J.M., Dickson E., Fiorani C., Bradley K.M., Mukherjee S., Gillies R.S., Maynard N.D., Middleton M.R.	Temporal validation of metabolic nodal response of esophageal cancer to neoadjuvant chemotherapy as an independent predictor of unresectable disease, survival, and recurrence	2019
Tacey M., Dinh D.T., Andrianopoulos N., Brennan A.L., Stub D., Liew D., Reid C.M., Duffy S.J., Lefkovits J.	Risk-Adjusting Key Outcome Measures in a Clinical Quality PCI Registry: Development of a Highly Predictive Model Without the Need to Exclude High-Risk Conditions	2019
Jorgensen S.C.J., Lagnf A.M., Bhatia S., Singh N.B., Shammout L.K., Davis S.L., Rybak M.J.	Diagnostic Stewardship: A Clinical Decision Rule for Blood Cultures in Community-Onset Methicillin-Resistant Staphylococcus aureus (MRSA) Skin and Soft Tissue Infections	2019
Gibertoni D., Rucci P., Mandreoli M., Corradini M., Martelli D., Russo G., Mancini E., Santoro A.	Temporal validation of the CT-PIRP prognostic model for mortality and renal replacement therapy initiation in chronic kidney disease patients	2019
Huang C., Murugiah K., Mahajan S., Li S.-X., Dhruva S.S., Haimovich J.S., Wang Y., Schulz W.L., Testani J.M., Wilson F.P., Mena C.I., Masoudi F.A., Rumsfeld J.S., Spertus J.A., Mortazavi B.J., Krumholz H.M.	Enhancing the prediction of acute kidney injury risk after percutaneous coronary intervention using machine learning techniques: A retrospective cohort study	2018
Huang, YQ; He, L; Dong, D; Yang, CY; Liang, CS; Chen, X; Ma, ZL; Huang, XM; Yao, S; Liang, CH; Tian, J; Liu, ZY	Individualized prediction of perineural invasion in colorectal cancer: development and validation of a radiomics prediction model	2018
Schwalbert R.A., Amado T.J.C., Nieto L., Varela S., Corassa G.M., Horbe T.A.N., Rice C.W., Peralta N.R., Ciampitti I.A.	Forecasting maize yield at field scale based on high-resolution satellite imagery	2018
de Waal, EEC; van Zaane, B; van der Schoot, MM; Huisman, A; Ramjankhan, F; van Klei, WA; Marczin, N	Vasoplegia after implantation of a continuous flow left ventricular assist device: incidence, outcomes and predictors	2018
De Waal E.E.C., Van Zaane B., Van Der Schoot M.M., Huisman A., Ramjankhan F., Van Klei W.A., Marczin N.	Vasoplegia after implantation of a continuous flow left ventricular assist device: Incidence, outcomes and predictors 11 Medical and Health Sciences 1103 Clinical Sciences	2018
Feidas H., Porcu F., Puca S., Rinollo A., Lagouvardos C., Kotroni V.	Validation of the H-SAF precipitation product H03 over Greece using rain gauge data	2018
Fukuma S., Shimizu S., Shintani A., Kamitani T., Akizawa T., Fukuhara S.	Development and validation of a prediction model for loss of physical function in elderly hemodialysis patients	2018
Moulton L.J., Eric Jelovsek J., Lachiewicz M., Chagin K., Goje O.	A model to predict risk of postpartum infection after Caesarean delivery	2018
Zarovska A., Evangelista A., Boccia T., Ciccone G., Coggiola D., Scarmozzino A., Corsi D.	Development and validation of a simplified BRASS index to screen hospital patients needing personalized discharge planning	2018
Rodríguez-Pérez R., Cortés R., Guamán A., Pardo A., Torralba Y., Gómez F., Roca J., Barberà J.A., Cascante M., Marco S.	Instrumental drift removal in GC-MS data for breath analysis: The short-Term and long-Term temporal validation of putative biomarkers for COPD	2018
Adderley N.J., Mallett S., Marshall T., Ghosh S., Rayman G., Bellary S., Coleman J., Akiboye F., Toulis K.A., Nirantharakumar K.	Temporal and external validation of a prediction model for adverse outcomes among inpatients with diabetes	2018
Dreijer, AR; Biedermann, JS; Diepstraten, J; Lindemans, AD; Kruip, MJHA; van den Bemt, PMLA; Vergouwe, Y	Development of a clinical prediction model for an international normalised ratio >= 4.5 in hospitalised patients using vitamin K antagonists	2018
Lenselink E.B., Ten Dijke N., Bongers B., Papadatos G., Van Vlijmen H.W.T., Kowalczyk W., Ijzerman A.P., Van Westen G.J.P.	Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set	2017
Huang T., Mi H., Lin C.-Y., Zhao L., Zhong L.L.D., Liu F.-B., Zhang G., Lu A.-P., Bian Z.-X., Lin S.-H., Zhang M., Li Y.-H., Hu D.-D., Cheng C.-W., MZRW Group	MOST: Most-similar ligand based approach to target prediction	2017
Yin G., Li A., Jin H., Zhao W., Bian J., Qu Y., Zeng Y., Xu B.	Derivation of temporally continuous LAI reference maps through combining the LAINet observation system with CACAO	2017

Autores	Título	Ano
Ivory S.E., Polkinghorne K.R., Khandakar Y., Kasza J., Zoungas S., Steenkamp R., Roderick P., Wolfe R.	Predicting 6-month mortality risk of patients commencing dialysis treatment for end-stage kidney disease	2017
Faxén J., Hall M., Gale C.P., Sundström J., Lindahl B., Jernberg T., Szummer K.	A user-friendly risk-score for predicting in-hospital cardiac arrest among patients admitted with suspected non ST-elevation acute coronary syndrome – The SAFER-score	2017
Faxen, J; Hall, M; Gale, CP; Sundstrom, J; Lindahl, B; Jernberg, T; Szummer, K	A user-friendly risk-score for predicting in-hospital cardiac arrest among patients admitted with suspected non ST-elevation acute coronary syndrome - The SAFER-score	2017
Sladkevicius P., Opolskiene G., Valentin L.	Prospective temporal validation of mathematical models to calculate risk of endometrial malignancy in patients with postmenopausal bleeding	2017
Stanhiser J., Chagin K., Jelovsek J.E.	A model to predict risk of blood transfusion after gynecologic surgery	2017
Fahey M., Rudd A., Béjot Y., Wolfe C., Douiri A.	Development and validation of clinical prediction models for mortality, functional outcome and cognitive impairment after stroke: A study protocol	2017
Tridente A., Bion J., Mills G.H., Gordon A.C., Clarke G.M., Walden A., Hutton P., Holloway P.A.H., Chiche J.-D., Stuber F., Garrard C., Hinds C., On behalf of the GenOSept and GAinS Investigators	Derivation and validation of a prognostic model for postoperative risk stratification of critically ill patients with faecal peritonitis	2017
Richardson A., Brearley S., Ahitan S., Chamberlain S., Davey T., Zujovic L., Hopkisson J., Campbell B., Raine-Fenning N.	Temporal validation of a simplified blastocyst grading system	2017
McAllister K.S.L., Ludman P.F., Hulme W., De Belder M.A., Stables R., Chowdhary S., Mamas M.A., Sperrin M., Buchan I.E., British Cardiovascular Intervention Society and the National Institute for Cardiovascular Outcomes Research	A contemporary risk model for predicting 30-day mortality following percutaneous coronary intervention in England and Wales	2016
Gorle J.M.R., Chatellier L., Pons F., Ba M.	Flow and performance analysis of H-Darrieus hydroturbine in a confined flow: A computational and experimental study	2016
Austin P.C., van Klaveren D., Vergouwe Y., Nieboer D., Lee D.S., Steyerberg E.W.	Geographic and temporal validity of prediction models: different approaches were useful to examine model performance	2016
Sarais V., Reschini M., Busnelli A., Biancardi R., Paffoni A., Somigliana E.	Predicting the success of IVF: External validation of the van Loendersloot's model	2016
Oduro A.R., Maya E.T., Akazili J., Baiden F., Koram K., Bojang K.	Monitoring malaria using health facility based surveys: Challenges and limitations	2016
Chen Y., Sun H., Li J.	Estimating daily maximum air temperature with MODIS data and a daytime temperature variation model in Beijing urban area	2016
Do Brasil P.E.A.A., Xavier S.S., Holanda M.T., Hasslocher-Moreno A.M., Braga J.U.	Does my patient have chronic chagas disease? Development and temporal validation of a diagnostic risk score	2016
Falasinnu T., Gilbert M., Gustafson P., Shoveller J.	A validation study of a clinical prediction rule for screening asymptomatic chlamydia and gonorrhoea infections among heterosexuals in British Columbia	2016
Schmidt M., Schmidt S.A.J., Sandegaard J.L., Ehrenstein V., Pedersen L., Sørensen H.T.	The Danish National patient registry: A review of content, data quality, and research potential	2015
Duane A., Piqué M., Castellnou M., Brotons L.	Predictive modelling of fire occurrences from different fire spread patterns in Mediterranean landscapes	2015
Dikaios N., Alkalbani J., Sidhu H.S., Fujiwara T., Abd-Alazeez M., Kirkham A., Allen C., Ahmed H., Emberton M., Freeman A., Halligan S., Taylor S., Atkinson D., Punwani S.	Logistic regression model for diagnosis of transition zone prostate cancer on multi-parametric MRI	2015
Dikaios N., Alkalbani J., Abd-Alazeez M., Sidhu H.S., Kirkham A., Ahmed H.U., Emberton M., Freeman A., Halligan S., Taylor S., Atkinson D., Punwani S.	Zone-specific logistic regression models improve classification of prostate cancer on multi-parametric MRI	2015
Terranova O.G., Gariano S.L., Iaquinta P., Iovine G.G.R.	GASAKe: Forecasting landslide activations by a genetic-algorithms-based hydrological model	2015
Weismüller T.J., Lerch C., Evangelidou E., Strassburg C.P., Lehner F., Schrem H., Klempnauer J., Manns M.P., Haller H., Schiffer M.	A pocket guide to identify patients at risk for chronic kidney disease after liver transplantation	2015
Murillo-García F.G., Alcántara-Ayala I.	Landslide susceptibility analysis and mapping using statistical multivariate techniques: Pahuatlán, Puebla, Mexico	2015
Kaijser, J	Towards an evidence-based approach for diagnosis and management of adnexal masses: findings of the International Ovarian Tumour Analysis (IOTA) studies	2015
Hoffmann V., Neubauer H., Heinzler J., Smarczyk A., Hellmich M., Bowe A., Kuetting F., Demir M., Pelc A., Schulte S., Toex U., Nierhoff D., Steffen H.-M.	A novel easy-to-use prediction scheme for upper gastrointestinal bleeding: Cologne-WATCH (C-WATCH) risk score	2015
Johnson W., Clancy T., Bastian P.	Child abuse/neglect risk assessment under field practice conditions: Tests of external and temporal validity and comparison with heart disease prediction	2015
Munkholm S.B., Jakobsen C.-J., Mortensen P.E., Lundbye-Christensen S., Andreasen J.J.	Validation of post-operative atrial fibrillation in the western denmark heart registry	2015
Van Calster B., Van Hoorde K., Valentin L., Testa A.C., Fischerova D., Van Holsbeke C., Savelli L., Franchi D., Epstein E., Kaijser J., Van Belle V., Czekierdowski A., Guerriero S., Fruscio R., Lanzani C., Scala F., Bourne T., Timmerman D., International Ovarian Tumour Analysis (IOTA) group	Evaluating the risk of ovarian cancer before surgery using the ADNEX model to differentiate between benign, borderline, early and advanced stage invasive, and secondary metastatic tumours: Prospective multicentre diagnostic study	2014
Bos L.D.J., Weda H., Wang Y., Knobel H.H., Nijsen T.M.E., Vink T.J., Zwinderman A.H., Sterk P.J., Schultz M.J.	Exhaled breath metabolomics as a noninvasive diagnostic tool for acute respiratory distress syndrome	2014
Allou N., Bronchard R., Guglielminotti J., Dilly M.P., Provenchere S., Lucet J.C., Laouénan C., Montravers P.	Risk factors for postoperative pneumonia after cardiac surgery and development of a preoperative risk score	2014

Autores	Título	Ano
Kruppa J., Liu Y., Diener H.-C., Holste T., Weimar C., König I.R., Ziegler A.	Probability estimation with machine learning methods for dichotomous and multicategory outcome: Applications	2014
Oliveira A., Jesus G., Gomes J.L., Rogeiro J., Azevedo A., Rodrigues M., Fortunato A.B., Dias J.M., Tomas L.M., Vaz L., Oliveira E.R., Alves F.L., Den Boer S.	An interactive WebGIS observatory platform for enhanced support of integrated coastal management	2014
Katanoda K., Kamo K.-I., Saika K., Matsuda T., Shibata A., Matsuda A., Nishino Y., Hattori M., Soda M., Ioka A., Sobue T., Nishimoto H.	Short-Term projection of cancer incidence in Japan using an age-period interaction model with spline smoothing	2014
Haltia U.-M., Bützow R., Leminen A., Loukovaara M.	FIGO 1988 versus 2009 staging for endometrial carcinoma: A comparative study on prediction of survival and stage distribution according to histologic subtype	2014
Falasinnu T., Gilbert M., Gustafson P., Shoveller J.	Deriving and validating a risk estimation tool for screening asymptomatic chlamydia and gonorrhea	2014
Rapacciuolo G., Roy D.B., Gillings S., Purvis A.	Temporal validation plots: Quantifying how well correlative species distribution models predict species' range changes over time	2014
Richter A., Weber M., Burrows J.P., Lambert J.-C., Van Gijsel A.	Validation strategy for satellite observations of tropospheric reactive gases	2014
Mihoub J.-B., Jiguet F., Lécuyer P., Eliotout B., Sarrazin F.	Modelling nesting site suitability in a popu?ation of reintroduced Eurasian black vultures Aegypius monachus in the Grands Causses, France	2014
Gol J.M., Burger H., Janssens K.A.M., Slaets J.P.J., Gans R.O.B., Rosmalen J.G.M.	PROFSS: A screening tool for early identification of functional somatic symptoms	2014
Foley R.N., Collins A.J.	The USRDS: What you need to know about what it can and can't tell us about ESRD	2013
Serbin S.P., Ahl D.E., Gower S.T.	Spatial and temporal validation of the MODIS LAI and FPAR products across a boreal forest wildfire chronosequence	2013
Van Calster B., Abdallah Y., Guha S., Kirk E., Van Hoorde K., Condous G., Preisler J., Hoo W., Stalder C., Bottomley C., Timmerman D., Bourne T.	Rationalizing the management of pregnancies of unknown location: Temporal and external validation of a risk prediction model on 1962 pregnancies	2013
Neumann L., Hoffmann V.S., Golgert S., Hasford J., Von Renteln-Kruse W.	In-hospital fall-risk screening in 4,735 geriatric patients from the LUCAS project	2013
Kong C.H., Guest G.D., Stupart D.A., Faragher I.G., Chan S.T.F., Watters D.A.	Recalibration and validation of a preoperative risk prediction model for mortality in major colorectal surgery	2013
Smits F.T., Brouwer H.J., Zwinderman A.H., van den Akker M., van Steenkiste B., Mohrs J., Schene A.H., van Weert H.C., Riet G.T.	Predictability of Persistent Frequent Attendance in Primary Care: A Temporal and Geographical Validation Study	2013
Van Den Boogaard M., Pickkers P., Slooter A.J.C., Kuiper M.A., Spronk P.E., Van Der Voort P.H.J., Van Der Hoeven J.G., Donders R., Van Achterberg T., Schoonhoven L.	Development and validation of PRE-DELIRIC (PREdiction of DELIRium in ICu patients) delirium prediction model for intensive care patients: Observational multicentre study	2012
Minne L., Eslami S., De Keizer N., De Jonge E., De Rooij S.E., Abu-Hanna A.	Effect of changes over time in the performance of a customized SAPS-II model on the quality of care assessment	2012
Minne L., Eslami S., de Keizer N., de Jonge E., de Rooij S.E., Abu-Hanna A.	Statistical process control for validating a classification tree model for predicting mortality - A novel approach towards temporal validation	2012
Minne L., Eslami S., de Keizer N., de Jonge E., de Rooij S.E., Abu-Hanna A.	Statistical process control for monitoring standardized mortality ratios of a classification tree model	2012
Van Middendorp J.J., Hosman A.J., Donders A.R.T., Pouw M.H., Ditunno Jr. J.F., Curt A., Geurts A.C., Van De Meent H.	A clinical prediction rule for ambulation outcomes after traumatic spinal cord injury: A longitudinal cohort study	2011
Bonderman D., Wexberg P., Martischnig A.M., Heinzl H., Lang M.-B., Sadushi R., Skoro-Sajer N., Lang I.M.	A noninvasive algorithm to exclude pre-capillary pulmonary hypertension	2011
Ramani S.E., Pitchaimani K., Gnanamanickam V.R.	GIS based landslide susceptibility mapping of Tevankarai Ar sub-watershed, Kodaikkanal, India using binary logistic regression analysis	2011
Johnson W.L.	The validity and utility of the California Family Risk Assessment under practice conditions in the field: A prospective study	2011
Carosella V.C., Grancelli H., RodríGUEZ W., Sellanes M., Cáceres M., Arazi H.C., Cárdenas C., Nojek C.	External and temporal validation 10 years after the development of the first latin-american risk stratification score for cardiac surgery (ArgenSCORE) [Primer puntaje de riesgo latinoamericano en cirugía cardíaca (ArgenSCORE): Validación externa y temporal a 10 años de su desarrollo]	2011
Chu D., Pubu T., Norbu G., Sagar B., Mandira S., Guo J.	Validation of the satellite-derived rainfall estimates over the tibet	2011
Timmerman D., Van Calster B., Testa A.C., Guerriero S., Fischerova D., Lissoni A.A., Van Holsbeke C., Fruscio R., Czekierdowski A., Jurkovic D., Savelli L., Vergote I., Bourne T., Van Huffel S., Valentin L.	Ovarian cancer prediction in adnexal masses using ultrasound-based logistic regression models: A temporal and external validation study by the IOTA group	2010
Den Eeckhaut M.V., Marre A., Poesen J.	Comparison of two landslide susceptibility assessments in the Champagne-Ardenne region (France)	2010
Cheong, Y; de Gregorio, F; Kim, K	The Power of Reach and Frequency In the Age of Digital Advertising Offline and Online Media Demand Different Metrics	2010
Medici C., Bernal S., Butturini A., Sabater F., Martin M., Wade A.J., Frances F.	Modelling the inorganic nitrogen behaviour in a small Mediterranean forested catchment, Fuirosos (Catalonia)	2010
Mitchell M.S., Gude J.A., Ausband D.E., Sime C.A., Bangs E.E., Jimenez M.D., MacK C.M., Meier T.J., Nadeau M.S., Smith D.W.	Temporal validation of an estimator for successful breeding pairs of wolves Canis lupus in the U.S. northern Rocky Mountains	2010

Autores	Título	Ano
Galve J.P., Gutiérrez F., Lucha P., Guerrero J., Bonachea J., Remondo J., Cendrero A.	Probabilistic sinkhole modelling for hazard assessment	2009
Chow E., Abdolell M., Panzarella T., Harris K., Bezjak A., Warde P., Tannock I.	Recursive Partitioning Analysis of Prognostic Factors for Survival in Patients With Advanced Cancer	2009
Chow E., Abdolell M., Panzarella T., Harris K., Bezjak A., Warde P., Tannock I.	Predictive model for survival in patients with advanced cancer	2008
Santhi C., Kannan N., Arnold J.G., Di Luzio M.	Spatial calibration and temporal validation of flow for regional scale hydrologic modeling	2008
Medici C., Butturini A., Bernal S., Vázquez E., Sabater F., Vélez J.I., Francés F.	Modelling the non-linear hydrological behaviour of a small Mediterranean forested catchment	2008
Binzenhöfer B., Biedermann R., Settele J., Schröder B.	Connectivity compensates for low habitat quality and small patch size in the butterfly Cupido minimus	2008
Haines T.P., Hill K., Walsh W., Osborne R.	Design-related bias in hospital fall risk screening tool predictive accuracy evaluations: Systematic review and meta-analysis	2007
Bolaños-Sanchez R., Sanchez-Arcilla A., Cateura J.	Evaluation of two atmospheric models for wind-wave modelling in the NW Mediterranean	2007
Manniën J., Van Der Zeeuw A.E., Wille J.C., Van Den Hof S.	Validation of surgical site infection surveillance in The Netherlands	2007
Spokoiny A., Shahar Y.	An active database architecture for knowledge-based incremental abstraction of complex concepts from continuously arriving time-oriented raw data	2007
Halligan S., Taylor S.A., Dehmeshki J., Amin H., Ye X., Tsang J., Roddie M.E.	Computer-assisted detection for CT colonography: external validation	2006
Malan A.K., Martins T.B., Hill H.R., Litwin C.M.	Evaluations of Commercial West Nile Virus Immunoglobulin G (IgG) and IgM Enzyme Immunoassays Show the Value of Continuous Validation	2004
Willinger R., Baumgartner D.	Human head tolerance limits to specific injury mechanisms	2003
Kaufman Y.J., Tanré D., Gordon H.R., Nakajima T., Lenoble J., Frouin R., Grassl H., Herman B.M., King M.D., Teillet P.M.	Passive remote sensing of tropospheric aerosol and atmospheric correction for the aerosol effect	1997
Miller M.E., Langefeld C.D., Tierney W.M., Hui S.L., Mcdonald C.J.	Validation of Probabilistic Predictions	1993
Jerntorp P., Berglund G.	Stroke registry in malmö, sweden	1992
LANG, JR; BOLTON, S	A COMPREHENSIVE METHOD VALIDATION STRATEGY FOR BIOANALYTICAL APPLICATIONS IN THE PHARMACEUTICAL-INDUSTRY .2. STATISTICAL-ANALYSES	1991
Toole J.F., Teagle W.C., Howard V.J., Grizzle J., Chambless L.E.	Original contributions study design for randomized prospective trial of carotid endarterectomy for asymptomatic atherosclerosis	1989
Boehm B.W.	Seven basic principles of software engineering	1983

APÊNDICE B

Maple®16 code

```
> RRC2 := proc(m, n, Z, δ, Pe) local i, fu, v, pn1, Beta, F, L, α, S, fv, ARL, ARL2, pn2, pn, u; pn1 := CDF(Normal(0, 1), u/sqrt(m) + Z − 2 δ
      ·sqrt(n) ·(1 − Pe), numeric); pn2 := CDF(Normal(0, 1), u/sqrt(m) − Z − 2 δ·sqrt(n) ·(1 − Pe), numeric); pn := pn1−pn2; fu :=
      PDF(Normal(0, 1), u, numeric); Beta := 1− evalf(int(pn·fu, u =− ∞..∞, numeric, digits = 5)); ARL2 := evalf(int(1/(1 − pn) ·fu, u =
      − ∞..∞, numeric, digits = 4)); α := evalf(1/ARL2, 4); print(ARL2) end proc
```

```
RRC2 := proc(m, n, Z, delta, Pe)                                                                                    (1)
    local i, fu, v, pn1, Beta, F, L, alpha, S, fv, ARL, ARL2, pn2, pn, u;
    pn1 := Statistics:-CDF(Normal(0, 1), u/sqrt(m) + Z − 2*delta*sqrt(n)*(1 − Pe), numeric);
    pn2 := Statistics:-CDF(Normal(0, 1), u/sqrt(m) − Z − 2*delta*sqrt(n)*(1 − Pe), numeric);
    pn := pn1 − pn2;
    fu := Statistics:-PDF(Normal(0, 1), u, numeric);
    Beta := 1 − evalf(int(pn*fu, u = −infinity..infinity, numeric, digits = 5));
    ARL2 := evalf(int(fu/(1 − pn), u = −infinity..infinity, numeric, digits = 4));
    alpha := evalf(1/ARL2, 4);
    print(ARL2)
end proc
```

APÊNDICE C

Tabela C1: Banco de Dados exemplo para Caso Hipotético

Reg.	Name	Experience	Failure/ Success	Reg.	Name	Experience	Failure/ Success
1	Frank	14	0	14	Tod	5	0
2	Henry	29	0	15	Take	20	1
3	Tom	6	0	16	Sam	13	0
4	Beth	25	1	17	Gail	9	0
5	Susan	18	1	18	Thomas	32	1
6	Harry	4	0	19	Theodore	24	0
7	Paul	18	0	20	Charles	13	1
8	Pete	12	0	21	Elizabeth	19	0
9	Diana	22	1	22	Lori	4	0
10	Louise	6	0	23	Ann	28	1
11	Fred	30	1	24	Valerie	22	1
12	Hank	11	0	25	Anke	8	1
13	Steven	30	1				

Fonte: Statistica® 8.0

APÊNDICE D
Código Fonte das simulações para o cálculo de AARL com base no *Kappa*

```
import random
import math
import numpy
import time

start = time.time()
# programa novo 2

#_____________________Variaveis_____________________

N = 1000 #numero de elementos
m = 10
media = 18
sig = 8.3176 #sigma do metodo
neta = 1
limiar = 0.5

RL_calculo = 2500 #kappa range (VRL)
ARL_calculo = 1 #RL (VARL)
AARL_calculo = 1000 #ARL (VAARL)

#_____________________Listas_____________________

marcador = 0

ARL_lista = []

KAPPA_medio = []

Po_m_list = []

Pe_m_list = []

real_S_lista_media = []
real_F_lista_media = []

for a in range(AARL_calculo):

    marcador = marcador + 1

    # _____________________Listas_____________________
    kappas = []

    Po_list = []
    Pe_list = []

    # ---------------------------------------------------------------------
    # -----------------KAPPA MEDIO E LIMITES DE CONTROLE-----------------

    kappa_medio_list = []

    Po_list_KM = []
    Pe_list_KM = []

    real_S_lista = []
    real_F_lista = []
```

```
    for b in range(m):  # Loop gerador de Po e Pe

        # _____________________Listas_____________________

        xp_list = []
        real_SF_list = []
        predict_SF_list = []

        # _____________________Data_Bank_____________________
        # cria o banco de dados de experiencia assim como o de sucesso e falha
real e põe nas suas respectivas listas

        for c in range(N):
            xp_list.append(math.sqrt((int(random.gauss(media, sig))) ** 2))

        for d in xp_list:
            RL = 1 / (1 + math.exp(-(-3.0597 + d * random.gauss(0.161486,
(0.064995)))))

            if RL > limiar:
                real_SF_list.append('S')
            else:
                real_SF_list.append('F')

        # _____________________Predição_____________________
        # preve se vai ser sucesso ou falha só com base na experiencia

        for e in xp_list:
            P = 1 / (1 + math.exp(-(-3.0597 + e * 0.161486)))

            if P > limiar:
                predict_SF_list.append('S')
            else:
                predict_SF_list.append('F')


        # _____________________Quantificação_____________
        # Determina a quantidade se sucesso ou falha em numero absoluto na lista
real e predita

        def quantificador(list, SF):  # usando a lista correta e a letra que quer
eliminar essa função retorna a quantidade de termos que sobrou
            new_words = [word for word in list.copy() if word not in SF]
            return len(new_words)


        real_S = quantificador(real_SF_list, 'F')
        real_F = quantificador(real_SF_list, 'S')

        real_S_m = real_S_lista.append(real_S)
        real_F_m = real_F_lista.append(real_F)

        predict_S = quantificador(predict_SF_list, 'F')
        predict_F = quantificador(predict_SF_list, 'S')


        # _____________________Acertos e erros_____________
        # Essa função retorna o "meio" da matriz com o cuzamento dos valores
```

```
        def match(x, y):
            number = 0
            real = real_SF_list.copy()
            previsto = predict_SF_list.copy()
            while number < len(real):
                if real[number] == x and previsto[number] == y:
                    number = number + 1
                else:
                    real.pop(number)
                    previsto.pop(number)

            return len(real)


        RS_PS = match('S', 'S')  # tp sucesso falha
        RF_PS = match('F', 'S')  # FP
        RS_PF = match('S', 'F')  # FN
        RF_PF = match('F', 'F')  # tn

        # _____________________Po_____________________

        Po_list_KM.append((RS_PS / len(real_SF_list) + (RF_PF /
len(real_SF_list))))


        # _____________________Pe_____________________

        def Pe_funcion(realSF, predictSF):

            SF_pe = ((realSF / len(real_SF_list)) * predictSF) / len(real_SF_list)

            return SF_pe


        Pe_list_KM.append(Pe_funcion(real_F, predict_F) + Pe_funcion(real_S,
predict_S))

        # ____________________________________________

    Po_m = numpy.mean(Po_list_KM)
    Pe_m = numpy.mean(Pe_list_KM)

    kp_list = []

    for f in range(m):
        Po = Po_list_KM[0]
        Pe = Pe_list_KM[0]

        kp_list.append((Po - Pe) / (1 - Pe))

        Po_list_KM.pop(0)
        Pe_list_KM.pop(0)

    # adiciona o kappa dividido por m a lista de kappas medio
    kappa_medio_m = sum(kp_list) / m
    kappa_medio_m_round = round(kappa_medio_m,5)
    KAPPA_medio.append(kappa_medio_m_round)
```

```
    # _____________________Po e Pe_____________________

    Po_m_list.append(round(Po_m,4))
    Pe_m_list.append(round(Pe_m,4))

    # _____________________SIGMA DO KAPPA_____________

    sigma_medio = (math.sqrt((Po_m * (1 - Po_m)) / (N * ((1 - Pe_m) ** 2))))



    # _____________________LIMITES DE CONTROLE________
    lsc0 = kappa_medio_m + (3 * sigma_medio)
    if lsc0 <= 1 :
        lsc = lsc0
    else:
        lsc = 1

    lic0 = kappa_medio_m - (3 * sigma_medio)

    if lic0 >= 0:
        lic = lic0
    else:
        lic = 0

    # ____________________________________________
    real_S_lista_media.append(numpy.mean(real_S_lista))
    real_F_lista_media.append(numpy.mean(real_F_lista))

    print('INICIO LOOP = ', marcador, '-----------------------------------------
--------')
    print('--------------------LIMITES DE CONTROLE-----------------------------')
    print('KAPA MEDIO = ', kappa_medio_m)
    print('SIGMA MEDIO = ', sigma_medio)
    print('LSC = ', lsc)
    print('LIC = ', lic)

    print('Real success=', numpy.mean(real_S_lista))
    print('Real failures=', numpy.mean(real_F_lista))
    #-------------------------------------------------------------------------
    #Ate aqui descobriu-se o LIC e LSC ----------------------------------
    #-------------------------------------------------------------------------

    RL_lista = []
    for g in  range(ARL_calculo):

        posicao = 1
        for o in range(RL_calculo):
            # ---------------------------------------------------------------
            # -----------------KAPPA ---------------------------------------

            # _____________________Listas_____________________

            xp_list = []
            real_SF_list = []
            predict_SF_list = []

            # _____________________Data_Bank__________________
```

```
            # cria o banco de dados de experiencia assim como o de sucesso e falha
real e põe nas suas respectivas listas

            for i in range(N):
                xp_list.append(math.sqrt((int(random.gauss(media, sig))) ** 2))

            for j in xp_list:
                RL = 1 / (1 + math.exp(-(-3.0597 + j * random.gauss(0.161486,
(neta * 0.064995))))) # acho que faltou o Neta multiplicando aqui

                if RL > limiar:
                    real_SF_list.append('S')
                else:
                    real_SF_list.append('F')

            # _____________________Predição_____________________
            # preve se vai ser sucesso ou falha só com base na experiencia

            for k in xp_list:
                P = 1 / (1 + math.exp(-(-3.0597 + k * 0.161486)))

                if P > limiar:
                    predict_SF_list.append('S')
                else:
                    predict_SF_list.append('F')

            # _____________________Quantificação_____________
            # Determina a quantidade se sucesso ou falha em numero absoluto na
lista real e predita

            def quantificador(list,SF):  # usando a lista coreta e a letra que
quer eliminar essa função retorna a quantidade de termos que sobrou
                new_words = [word for word in list.copy() if word not in SF]
                return len(new_words)

            real_S = quantificador(real_SF_list, 'F')
            real_F = quantificador(real_SF_list, 'S')

            predict_S = quantificador(predict_SF_list, 'F')
            predict_F = quantificador(predict_SF_list, 'S')

            # _____________________Acertos e erros___________
            # Essa função retorna o "meio" da matriz com o cuzamento dos valores

            def match(x, y):
                number = 0
                real = real_SF_list.copy()
                previsto = predict_SF_list.copy()
                while number < len(real):
                    if real[number] == x and previsto[number] == y:
                        number = number + 1
                    else:
                        real.pop(number)
                        previsto.pop(number)

                return len(real)

            RS_PS = match('S', 'S')  # tp sucesso falha
            RF_PF = match('F', 'F')  # tn
```

```
            # _____________________Po_____________________

            Po = ((RS_PS / len(real_SF_list) + (RF_PF / len(real_SF_list))))

            # _____________________Pe_____________________

            def Pe_funcion(realSF, predictSF):

                SF_pe = ((realSF / len(real_SF_list)) * predictSF) /
len(real_SF_list)

                return SF_pe


            Pe = (Pe_funcion(real_F, predict_F) + Pe_funcion(real_S, predict_S))

            # ____________________________________________________

            kappa = ((Po - Pe) / (1 - Pe))


            if kappa > lsc:
                RL_lista.append(posicao)
                break

            if kappa < lic:
                RL_lista.append(posicao)
                break

            if posicao == RL_calculo:
                RL_lista.append(posicao) # inseri esta linha para identificar
posiçao quando nenhum item sai fora de controle)

            else:
                posicao = posicao + 1
        #print(RL_lista)

    ARL = numpy.mean(RL_lista)
    print("ARL = ",ARL)
    ARL_lista.append(ARL)
    PE_medio = numpy.mean(Pe_m)

end = time.time()

print("#####################################################")
AARL = numpy.mean(ARL_lista)
print("AARL = ",AARL)
print('Tempo (s) = ',end-start)
print("List of ARL´s=", ARL_lista)
print("media de Kappas médios = ",  round((numpy.mean(KAPPA_medio)),5))
print("Lista de kappas medios = ", KAPPA_medio)
print('Lista PO = ', Po_m_list)
print('Lista PE = ', Pe_m_list)
print('PE Médio = ', PE_medio)
print('Media S = ', numpy.mean(real_S_lista_media),'\\Lista media S real =',
real_S_lista_media)
print('Media F = ', numpy.mean(real_F_lista_media),'\\Lista media F real =',
real_F_lista_media)
```

APÊNDICE E

Código python *MCC*

```
import random
import math
import numpy
import time
import statistics as st

start = time.time()

#_____________________Variaveis_____________________

N = 300 #numero de elementos
m = 10
media = 18
sig = 8.3176 #sigma do metodo
neta = 1
limiar = 0.5

RL_calculo = 2500 #mcc range
ARL_calculo = 30 #RL
AARL_calculo = 250 #ARL

#_____________________Listas_____________________

marcador = 0

ARL_lista = []

mcc_medio = []

real_S_lista_media = []
real_F_lista_media = []
predict_S_lista_media = []
predict_F_lista_media = []

#_____________________________________________

for a in range(AARL_calculo):
    marcador = marcador + 1

    # ------------------------------------------------------------------
    # -----------------INDICE DE MCC E LIMITES DE CONTROLE-----------------

    real_S_lista = []
    real_F_lista = []
    predict_S_lista = []
    predict_F_lista = []

    #--------------------
    A_list = []
    B_list = []
    C_list = []
    D_list = []

    for b in range(m):  # Loop gerador de A B C D

        # _____________________Listas_____________________
```

```
        xp_list = []
        real_SF_list = []
        predict_SF_list = []

        # _____________________Data_Bank_____________________
        # cria o banco de dados de experiencia assim como o de sucesso e falha
real e põe nas suas respectivas listas

        for c in range(N):
            xp_list.append(math.sqrt((int(random.gauss(media, sig))) ** 2))

        for d in xp_list:
            RL = 1 / (1 + math.exp(-(-3.0597 + d * random.gauss(0.161486,
(0.064995)))))

            if RL > limiar:
                real_SF_list.append('S')
            else:
                real_SF_list.append('F')

        # _____________________Predição_____________________
        # preve se vai ser sucesso ou falha só com base na experiencia

        for e in xp_list:
            P = 1 / (1 + math.exp(-(-3.0597 + e * 0.161486)))

            if P > limiar:
                predict_SF_list.append('S')
            else:
                predict_SF_list.append('F')


        # _____________________Quantificação_____________
        # Determina a quantidade se sucesso ou falha em numero absoluto na lista
real e predita

        def quantificador(list, SF):  # usando a lista correta e a letra que quer
eliminar essa função retorna a quantidade de termos que sobrou
            new_words = [word for word in list.copy() if word not in SF]
            return len(new_words)


        real_S = quantificador(real_SF_list, 'F')
        real_F = quantificador(real_SF_list, 'S')

        real_S_m = real_S_lista.append(real_S)
        real_F_m = real_F_lista.append(real_F)

        predict_S = quantificador(predict_SF_list, 'F')
        predict_F = quantificador(predict_SF_list, 'S')

        predict_S_m = predict_S_lista.append(predict_S)
        predict_F_m = predict_F_lista.append(predict_F)


        # _____________________Acertos e erros_____________
        # Essa função retorna o "meio" da matriz com o cuzamento dos valores
```

```
        def match(x, y):
            number = 0
            real = real_SF_list.copy()
            previsto = predict_SF_list.copy()
            while number < len(real):
                if real[number] == x and previsto[number] == y:
                    number = number + 1
                else:
                    real.pop(number)
                    previsto.pop(number)

            return len(real)

        RS_PS = match('S', 'S')  #A
        RF_PS = match('F', 'S')  #C
        RS_PF = match('S', 'F')  #B
        RF_PF = match('F', 'F')  #D

        A = RS_PS #TP
        B = RS_PF #FN
        C = RF_PS #FP
        D = RF_PF #TN

        A_list.append(A)
        B_list.append(B)
        C_list.append(C)
        D_list.append(D)

    # _____________________Media de MCC para servir de base_____________________
    #multiplica m*30
    #stddevi / raiz de m
    #

    mcc_list = []

    for f in range(m):
        mcc_base = (math.sqrt((A_list[0] + C_list[0]) * (A_list[0] + B_list[0]) *
(D_list[0] + C_list[0]) * (D_list[0] + B_list[0])))

        if mcc_base == 0:
            mcc_list.append(0)
        else:
            mcc_list.append( ((A_list[0] * D_list[0]) - (C_list[0] * B_list[0])) /
                             (math.sqrt((A_list[0] + C_list[0]) * (A_list[0] +
B_list[0]) * (D_list[0] + C_list[0]) * (D_list[0] + B_list[0]))))

        A_list.pop(0)
        B_list.pop(0)
        C_list.pop(0)
        D_list.pop(0)

    # adiciona o valor dividido por m a lista de mcc medio
    MCC_medio_m = sum(mcc_list) / m
    MCC_medio_m_round = round(MCC_medio_m, 5)
    mcc_medio.append(MCC_medio_m_round)

 # _____________________LIMITES DE CONTROLE________
    # media mcc +- 3 zigma st.stdev
    #print("mcc list",mcc_list)
```

```
    #print("stdev", st.stdev(mcc_list))

    lsc0 = MCC_medio_m_round + (3 * st.stdev(mcc_list))
    if lsc0 <= 1:
        lsc = lsc0
    else:
        lsc = 1

    lic0 = MCC_medio_m_round - (3 * st.stdev(mcc_list))

    if lic0 >= 0:
        lic = lic0
    else:
        lic = 0

    # _____________________________________________________
    real_S_lista_media.append(numpy.mean(real_S_lista))
    real_F_lista_media.append(numpy.mean(real_F_lista))
    predict_S_lista_media.append(numpy.mean(predict_S_lista))
    predict_F_lista_media.append(numpy.mean(predict_F_lista))

    print('INICIO LOOP = ', marcador, '----------------------------------------
--------')
    print('--------------------LIMITES DE CONTROLE-----------------------------')
    print("MCC = ", MCC_medio_m_round)
    print('LSC usado = ', lsc)
   # print('LSC real = ', lsc0)
    print('LIC usado = ', lic)
   # print('LIC real = ', lic0)
    print('Real success=', numpy.mean(real_S_lista))
    print('Real failures=', numpy.mean(real_F_lista))
    print("MCC LIST", mcc_list)

    # ------------------------------------------------------------------------
    # Ate aqui descobriu-se o LIC e LSC ---------------------------------------
    # ------------------------------------------------------------------------

    RL_lista = []
    for g in range(ARL_calculo):

        posicao = 1
        for o in range(RL_calculo):
            # ------------------------------------------------------------------

            # _____________________Listas_____________________

            xp_list = []
            real_SF_list = []
            predict_SF_list = []

            # _____________________Data_Bank_____________________
            # cria o banco de dados de experiencia assim como o de sucesso e falha
real e põe nas suas respectivas listas

            for i in range(N):
                xp_list.append(math.sqrt((int(random.gauss(media, sig))) ** 2))

            for j in xp_list:
                RL = 1 / (1 + math.exp(-(-3.0597 + j * random.gauss(0.161486, (
```

```
                        neta * 0.064995)))))

            if RL > limiar:
                real_SF_list.append('S')
            else:
                real_SF_list.append('F')

        # ____________________Predição____________________
        # preve se vai ser sucesso ou falha só com base na experiencia

        for k in xp_list:
            P = 1 / (1 + math.exp(-(-3.0597 + k * 0.161486)))

            if P > limiar:
                predict_SF_list.append('S')
            else:
                predict_SF_list.append('F')


        # ____________________Quantificação_____________
        # Determina a quantidade se sucesso ou falha em numero absoluto na
lista real e predita

        def quantificador(list,
                          SF):  # usando a lista coreta e a letra que quer
eliminar essa função retorna a quantidade de termos que sobrou
            new_words = [word for word in list.copy() if word not in SF]
            return len(new_words)


        real_S = quantificador(real_SF_list, 'F')
        real_F = quantificador(real_SF_list, 'S')

        predict_S = quantificador(predict_SF_list, 'F')
        predict_F = quantificador(predict_SF_list, 'S')


        # ____________________Acertos e erros____________
        # Essa função retorna o "meio" da matriz com o cuzamento dos valores

        def match(x, y):
            number = 0
            real = real_SF_list.copy()
            previsto = predict_SF_list.copy()
            while number < len(real):
                if real[number] == x and previsto[number] == y:
                    number = number + 1
                else:
                    real.pop(number)
                    previsto.pop(number)

            return len(real)


        RS_PS = match('S', 'S')  # A
        RF_PS = match('F', 'S')  # C
        RS_PF = match('S', 'F')  # B
        RF_PF = match('F', 'F')  # D
```

```
            A = RS_PS
            B = RS_PF
            C = RF_PS
            D = RF_PF

            # _____________MCC_____________________________________
            #databank N*m
            #a cada m eu tiro um mcc
            #pega no numemo m de mcc e tira a media

            conta = (math.sqrt((A + C) * (A + B) * (D + C) * (D + B)))

            if conta == 0:
                MCC == 0
            else:
                MCC = (((A * D) - (C * B)) /
                                (math.sqrt((A + C) * (A + B) * (D + C) * (D +
B))))

            if MCC > lsc:
                RL_lista.append(posicao)
                break

            if MCC < lic:
                RL_lista.append(posicao)
                break

            if posicao == RL_calculo:
                RL_lista.append(
                    posicao)  # inseri esta linha para identificar posiçao quando
nenhum item sai fora de controle)

            else:
                posicao = posicao + 1
        #print("RL list = ", RL_lista)

    ARL = numpy.mean(RL_lista)
    print("ARL = ", ARL)
    ARL_lista.append(ARL)

end = time.time()

print("###################################################")
AARL = numpy.mean(ARL_lista)
print("AARL = ", AARL)
print('Tempo (s) = ', end - start)
print("List of ARL´s=", ARL_lista)
#print("media de mcc médios = ", round((numpy.mean(mcc_medio)), 5))
print("Lista de mcc medios = ", mcc_medio, '\n')
print('Media S REAL = ', numpy.mean(real_S_lista_media), '\nLista media S real =',
real_S_lista_media)
print('Media F REAL = ', numpy.mean(real_F_lista_media), '\nLista media F real =',
real_F_lista_media, '\n')
print('Media S PREDITO = ', numpy.mean(predict_S_lista_media), '\nLista media S
predict =', predict_S_lista_media)
print('Media F PREDITO= ', numpy.mean(predict_F_lista_media), '\nLista media F
predict =', predict_F_lista_media)
```

APÊNDICE F
Código Python *Youden*

```
import random
import math
import numpy
import time

start = time.time()

#_____________________Variaveis_____________________

N = 1000 #numero de elementos
m = 100
media = 18
sig = 8.3176 #sigma do metodo
neta = 1
limiar = 0.5

RL_calculo = 2500 #youden range
ARL_calculo = 30 #RL
AARL_calculo = 100 #ARL

#_____________________Listas_____________________

marcador = 0

ARL_lista = []

YOUDEN_medio = []

real_S_lista_media = []
real_F_lista_media = []
predict_S_lista_media = []
predict_F_lista_media = []

for a in range(AARL_calculo):

    marcador = marcador + 1

    # _____________________Listas_____________________
    youdens = []

    # ------------------------------------------------------------------
    # -----------------INDICE DE YOUDEN MEDIO E LIMITES DE CONTROLE---------------
---

    youden_medio_list = []

    real_S_lista = []
    real_F_lista = []
    predict_S_lista = []
    predict_F_lista = []

    for b in range(m):  # Loop gerador de Po e Pe

        # _____________________Listas_____________________
```

```
        xp_list = []
        real_SF_list = []
        predict_SF_list = []

        # _____________________Data_Bank_____________________
        # cria o banco de dados de experiencia assim como o de sucesso e falha
real e põe nas suas respectivas listas

        for c in range(N):
            xp_list.append(math.sqrt((int(random.gauss(media, sig))) ** 2))

        for d in xp_list:
            RL = 1 / (1 + math.exp(-(-3.0597 + d * random.gauss(0.161486,
(0.064995)))))

            if RL > limiar:
                real_SF_list.append('S')
            else:
                real_SF_list.append('F')

        # _____________________Predição_____________________
        # preve se vai ser sucesso ou falha só com base na experiencia

        for e in xp_list:
            P = 1 / (1 + math.exp(-(-3.0597 + e * 0.161486)))

            if P > limiar:
                predict_SF_list.append('S')
            else:
                predict_SF_list.append('F')


        # _____________________Quantificação_____________
        # Determina a quantidade se sucesso ou falha em numero absoluto na lista
real e predita

        def quantificador(list, SF):  # usando a lista correta e a letra que quer
eliminar essa função retorna a quantidade de termos que sobrou
            new_words = [word for word in list.copy() if word not in SF]
            return len(new_words)


        real_S = quantificador(real_SF_list, 'F')
        real_F = quantificador(real_SF_list, 'S')

        real_S_m = real_S_lista.append(real_S)
        real_F_m = real_F_lista.append(real_F)

        predict_S = quantificador(predict_SF_list, 'F')
        predict_F = quantificador(predict_SF_list, 'S')

        predict_S_m = predict_S_lista.append(predict_S)
        predict_F_m = predict_F_lista.append(predict_F)


        # _____________________Acertos e erros_____________
        # Essa função retorna o "meio" da matriz com o cuzamento dos valores
```

```
        def match(x, y):
            number = 0
            real = real_SF_list.copy()
            previsto = predict_SF_list.copy()
            while number < len(real):
                if real[number] == x and previsto[number] == y:
                    number = number + 1
                else:
                    real.pop(number)
                    previsto.pop(number)

            return len(real)

        RS_PS = match('S', 'S')  #A
        RF_PS = match('F', 'S')  #C
        RS_PF = match('S', 'F')  #B
        RF_PF = match('F', 'F')  #D

        A = RS_PS
        B = RS_PF
        C = RF_PS
        D = RF_PF



    # _____________________Media de Youdens para servir de
base_____________________
    yd_list = []
    for f in range(m):
        yd_list.append(((A/(A+B))+(D/(C+D))-1))


    # adiciona o youden dividido por m a lista de youdens medio
    youden_medio_m = sum(yd_list) / m
    youden_medio_m_round = round(youden_medio_m,5)
    YOUDEN_medio.append(youden_medio_m_round)


    # _____________________SIGMA DO INDICE DE YOUDEN _____________

    sigma_medio = (math.sqrt(((A*B)/((A+B)**3))+((C*D)/((C+D)**3))))


    # _____________________LIMITES DE CONTROLE________
    lsc0 = youden_medio_m + (3 * sigma_medio)
    if lsc0 <= 1 :
        lsc = lsc0
    else:
        lsc = 1

    lic0 = youden_medio_m - (3 * sigma_medio)

    if lic0 >= 0:
        lic = lic0
    else:
        lic = 0

    # ____________________________________________________
    real_S_lista_media.append(numpy.mean(real_S_lista))
```

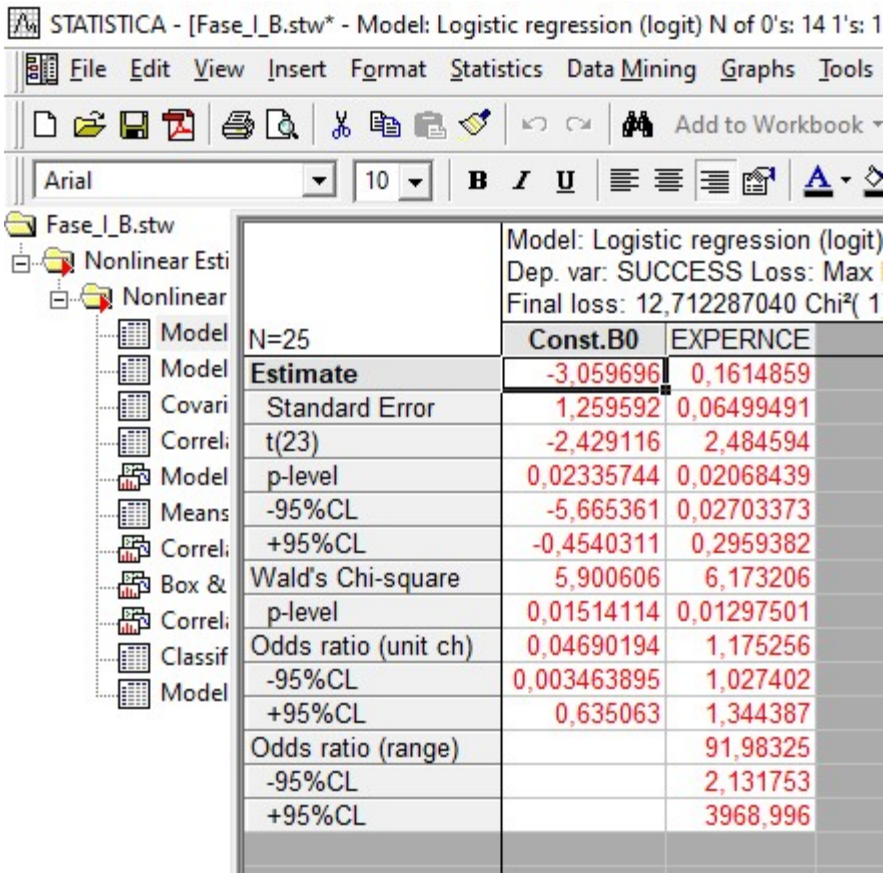
```
    real_F_lista_media.append(numpy.mean(real_F_lista))
    predict_S_lista_media.append(numpy.mean(predict_S_lista))
    predict_F_lista_media.append(numpy.mean(predict_F_lista))

    print('INICIO LOOP = ', marcador, '----------------------------------------
--------')
    print('--------------------LIMITES DE CONTROLE-----------------------------')
    print('YOUDEN MEDIO = ', youden_medio_m)
    print('SIGMA MEDIO = ', sigma_medio)
    print('LSC = ', lsc)
    print('LIC = ', lic)

    print('Real success=', numpy.mean(real_S_lista))
    print('Real failures=', numpy.mean(real_F_lista))
    #-------------------------------------------------------------------------
    #Ate aqui descobriu-se o LIC e LSC ------------------------------------
    #-------------------------------------------------------------------------

    RL_lista = []
    for g in  range(ARL_calculo):

        posicao = 1
        for o in range(RL_calculo):
            # ---------------------------------------------------------------
            # ----------------- INDICE DE YOUDEN ----------------------------
------------

            # _____________________Listas_____________________

            xp_list = []
            real_SF_list = []
            predict_SF_list = []

            # _____________________Data_Bank_____________________
            # cria o banco de dados de experiencia assim como o de sucesso e falha
real e põe nas suas respectivas listas

            for i in range(N):
                xp_list.append(math.sqrt((int(random.gauss(media, sig))) ** 2))

            for j in xp_list:
                RL = 1 / (1 + math.exp(-(-3.0597 + j * random.gauss(0.161486,
(neta * 0.064995))))) # acho que faltou o Neta multiplicando aqui

                if RL > limiar:
                    real_SF_list.append('S')
                else:
                    real_SF_list.append('F')

            # _____________________Predição_____________________
            # preve se vai ser sucesso ou falha só com base na experiencia

            for k in xp_list:
                P = 1 / (1 + math.exp(-(-3.0597 + k * 0.161486)))

                if P > limiar:
                    predict_SF_list.append('S')
                else:
                    predict_SF_list.append('F')
```

```
            # _____________________Quantificação_____________
            # Determina a quantidade se sucesso ou falha em numero absoluto na
lista real e predita

            def quantificador(list,SF):  # usando a lista coreta e a letra que
quer eliminar essa função retorna a quantidade de termos que sobrou
                new_words = [word for word in list.copy() if word not in SF]
                return len(new_words)

            real_S = quantificador(real_SF_list, 'F')
            real_F = quantificador(real_SF_list, 'S')

            predict_S = quantificador(predict_SF_list, 'F')
            predict_F = quantificador(predict_SF_list, 'S')


            # _____________________Acertos e erros_____________
            # Essa função retorna o "meio" da matriz com o cuzamento dos valores

            def match(x, y):
                number = 0
                real = real_SF_list.copy()
                previsto = predict_SF_list.copy()
                while number < len(real):
                    if real[number] == x and previsto[number] == y:
                        number = number + 1
                    else:
                        real.pop(number)
                        previsto.pop(number)

                return len(real)

            RS_PS = match('S', 'S')  #A
            RF_PS = match('F', 'S')  #C
            RS_PF = match('S', 'F')  #B
            RF_PF = match('F', 'F')  #D

            A = RS_PS
            B = RS_PF
            C = RF_PS
            D = RF_PF

            # ____________INDICE DE YOUDEN_____________________________________

            youden = ((A/(A+B))+(D/(C+D))-1)


            if youden > lsc:
                RL_lista.append(posicao)
                break

            if youden < lic:
                RL_lista.append(posicao)
                break

            if posicao == RL_calculo:
                RL_lista.append(posicao) # inseri esta linha para identificar
posiçao quando nenhum item sai fora de controle)
```

```
            else:
                posicao = posicao + 1
        #print(RL_lista)

    ARL = numpy.mean(RL_lista)
    print("ARL = ",ARL)
    ARL_lista.append(ARL)


end = time.time()

print("##################################################")
AARL = numpy.mean(ARL_lista)
print("AARL = ",AARL)
print('Tempo (s) = ',end-start)
print("List of ARL´s=", ARL_lista)
print("media de youdens médios = ",  round((numpy.mean(YOUDEN_medio)),5))
print("Lista de youdens medios = ", YOUDEN_medio,'\n')
print('Media S REAL = ', numpy.mean(real_S_lista_media),'\nLista media S real =',
real_S_lista_media)
print('Media F REAL = ', numpy.mean(real_F_lista_media),'\nLista media F real =',
real_F_lista_media,'\n')
print('Media S PREDITO = ', numpy.mean(predict_S_lista_media),'\nLista media S
predict =', predict_S_lista_media)
print('Media F PREDITO= ', numpy.mean(predict_F_lista_media),'\nLista media F
predict =', predict_F_lista_media)
```

APÊNDICE G

Parâmetros da Regressão Logística gerados pelo Software Statística®



N=25	Const.B0	EXPERNCE
Model: Logistic regression (logit)		
Dep. var: SUCCESS Loss: Max		
Final loss: 12,712287040 Chi²(1		
Estimate	-3,059696	0,1614859
Standard Error	1,259592	0,06499491
t(23)	-2,429116	2,484594
p-level	0,02335744	0,02068439
-95%CL	-5,665361	0,02703373
+95%CL	-0,4540311	0,2959382
Wald's Chi-square	5,900606	6,173206
p-level	0,01514114	0,01297501
Odds ratio (unit ch)	0,04690194	1,175256
-95%CL	0,003463895	1,027402
+95%CL	0,635063	1,344387
Odds ratio (range)		91,98325
-95%CL		2,131753
+95%CL		3968,996