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

Breno Osvaldo Funicheli

Seleção de SNPs em Culturas de Arroz Utilizando Aprendizado de Máquina

São Carlos
2024

Breno Osvaldo Funicheli

Seleção de SNPs em Culturas de Arroz Utilizando Aprendizado de Máquina

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

Área de concentração: Aprendizado de Máquina e Processamento de Línguas Naturais

Orientador: Ricardo Cerri

São Carlos
2024

UNIVERSIDADE FEDERAL DE SÃO CARLOS

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

Folha de Aprovação

Defesa de Dissertação de Mestrado do candidato Breno Osvaldo Funicheli, realizada em 02/02/2024.

Comissão Julgadora:

Prof. Dr. Ricardo Cerri (UFSCar)

Profa. Dra. Priscila Tiemi Maeda Saito (UFSCar)

Prof. Dr. Zanoni Dias (UNICAMP)

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

Agradecimentos

Agradeço e dedico a DEUS essa e todas a vitórias e conquistas da minha vida. Creio que todas as coisas estão nas mãos dELE e nada pode escapar de seu controle. Seja a vida ou seja a morte, tudo está nas mãos dELE. Obrigado meu Senhor e salvado Jesus Cristo pela saúde e por me ajudar a chegar até aqui.

Resumo

O arroz (*Oryza sativa*) é uma das maiores coleções de recursos genéticos entre as espécies vegetais de interesse econômico. Com o intuito de aumentar a produtividade desse cultivar, diversos estudos de variabilidade genética vêm sendo desenvolvidos. Nesse contexto, os polimorfismos de nucleotídeo único (SNPs), que são variações de base única nas sequências de DNA, têm sido amplamente estudados, pois atuam como marcadores moleculares vinculados à produtividade e resistência na cultura de arroz. No entanto, devido a não efetividade de métodos convencionais na tarefa de seleção de SNPs, métodos baseados em Aprendizado de Máquina (do inglês, "*Machine Learning*") (ML) vêm sendo utilizados. Para isso a seleção de SNPs é modelada como um problema de seleção de Atributos (do inglês, "*Feature Selection*") (FS). Embora a utilização de FS seja amplamente difundida na literatura, ainda há lacunas quanto à sua utilização no contexto de melhoramento genético de arroz. Em conjunto a isso, observa-se a necessidade de investigação dos SNPs selecionados por esses métodos em estudos de melhoramento genético, de modo a oferecer possíveis explicações biológicas vinculadas aos resultados gerados. Com o intuito de avançar nuances referentes a essa discussão, o presente trabalho propôs alguns métodos de *ensemble* para seleção de SNPs, a fim de combinar diversos algoritmos de FS para geração de um resultado robusto. Os métodos foram implementados de maneira a criar um *pipeline* para seleção de SNPs. O pipeline foi aplicado a um conjunto de dados com múltiplos fenótipos ligados à produtividade do arroz. Os métodos propostos foram contrastados a outros métodos presentes na literatura, demonstrando resultados superiores em alguns casos. Além disso, foi explorada a utilização do enriquecimento funcional como estratégia de explicação dos resultados. O conjunto de dados utilizado pertence à Coleção Nuclear de Arroz da Embrapa Arroz e Feijão, e foi cedido com o intuito de que os resultados gerados no presente trabalho fossem posteriormente investigados e utilizados no melhoramento genético de arroz.

Palavras-chave: SNP; Seleção de Atributos; Aprendizado de Máquina, Arroz.

Abstract

Rice (*Oryza sativa*) is one of the largest collections of genetic resources among plant species of economic interest. To increase the productivity of this cultivar, several genetic variability studies have been developed. In this context, single nucleotide polymorphisms (SNPs), which are single base variations in DNA sequences, have been widely studied, as they act as molecular markers linked to productivity and resistance in rice cultivation. However, due to the ineffectiveness of conventional methods in the task of selecting SNPs, methods based on Machine Learning (ML) have been used. For this purpose, the selection of SNPs is modeled as a Feature Selection (FS) problem. Although the use of FS is widespread in the literature, there are still gaps regarding its use in the context of rice genetic improvement. In conjunction with this, there is a need to investigate the SNPs selected by these methods in genetic improvement studies, to offer possible biological explanations linked to the results generated. To advance interesting points regarding this discussion, this work proposes some ensemble methods for selecting SNPs, to combine several FS algorithms to generate a robust result. These methods were implemented such as to create a pipeline for SNPs selection. The pipeline was applied to a dataset with multiple phenotypes linked to rice productivity. The proposed methods were compared to other methods present in the literature, demonstrating the best results in some cases. Furthermore, the use of functional enrichment as a strategy to explain the results was explored. The dataset used belongs to the Coleção Nuclear de Arroz of Embrapa Arroz e Feijão and was provided with the intention that the results generated in the present work would be subsequently investigated and used in the genetic improvement of rice.

Keywords: SNP; Feature Selection; Machine Learning; Rice..

Lista de ilustrações

Lista de tabelas

Lista de siglas

ML	Aprendizado de Máquina (do inglês, "*Machine Learning*")
API	Interface de Programação de Aplicação (do inglês, "*Application Program Interface*")
AUC	Área Sob a Curva (do inglês, "*Area Under Curve*")
BLUP	Best Linear Unbiased Prediction
CV	Validação Cruzada (do inglês, "*Cross Validation*")
CSV	Valores Separados por Vírgula (do inglês, "*Comma-separated Values*")
CMIM	Conditional Mutual Information Maximization
DAG	Grafo Acíclico Dirigido (do inglês, "*Directed Acyclic Graph*")
DNA	Ácido Desoxirribonucleico (do inglês, "*DeoxyriboNucleic Acid*")
FSDCI	Feature Selection Discretizada Consensual iterativa
FSCD	Feature Selection Consensual Direta
FSDCD	Feature Selection Discretizada Consensual Direta
FSCI	Feature Selection Consensual Iterativa
FS	Seleção de Atributos (do inglês, "*Feature Selection*")
FCS	Pontuação de Classe Funcional (do inglês, "*Functional Class Score*")
GWAS	Estudos de Associação de Genoma Completo (do inglês, "*Genome-Wide Association Studies*")
GBS	Genotipagem por Sequenciamento (do inglês, "*Genotyping by Sequencing*")
GSEA	Análise de Enriquecimento de Conjuntos de Genes (do inglês, "*Gene Set Enrichment Analysis*")
GO	*Gene Ontology*

HPO Otimização de Hiper-Parâmetros (do inglês, "*HyperParameter Otimization*")

IA Inteligência Artificial

KNN K-Vizinhos Mais Próximos (do inglês, "*K-Nearest Neighbors*")

KEGG banco de dados da Enciclopédia de Genes e Genomas de Kyoto (do inglês, "*Kyoto Encyclopedia of Genes and Genomes*")

LASSO Least Absolute Shrinkage and Selection Operator

miRNA Micro RNAs

MDI *Mean Decrease Impurity*

MDA *Mean Decrease in Accuracy*

MAE Erro Médio Absoluto (do inglês, "*Mean Absolute Error*")

MSE Erro Quadrático Médio (do inglês, "*Mean Square Error*")

RMSE Raiz do Erro Quadrático Médio (do inglês, "*Root Mean Square Error*")

NGS Sequenciamento de Nova Geração (do inglês, "*Next Generation Sequence*")

ORA Análise de Sobre-representação (do inglês, "*Over Representation Analysis*")

QTL Traços Quantitativos de Loci (do inglês, "*Quantitative Trait Loci*")

QTN Traços Quantitativos de Nucleotídio (do inglês, "*Quantitative Trait Nucleotides*")

RFE Eliminação Recursiva de Atributos (do inglês, "*Recursive Feature Elimination*")

RCV Validação Cruzada Repetida (do inglês, "*Repeated Cross-Validation*")

RBA Algoritmos Baseados em Relief (do inglês, "*Relief Based Algorithms*")

RF Floresta Aleatória (do inglês, "*Random Forest*")

ROC *Receiver Operating Characteristic*

REML Máxima Verossimilhança Residual

AUC-ROC Área Sob a Curva ROC (do inglês, "*Area Under ROC Curve*")

SGD Método do Gradiente Estocástico (do inglês, "*Stochastic Gradient Descent*")

SVM Máquina de Vetores de Suporte (do inglês, "*Support Vector Machine*")

SNP Polimorfismo de Nucleotídio Único (do inglês, "*Single Nucleotide Polymorphisms*")

SFS Seleção Sequencial de Atributos (do inglês, "*Sequential Feature Selection*")

XGB XGBoost ou *Extreme Gradient Boosting*

Sumário

Capítulo 1

Introdução

1.1 Contextualização e motivação

O arroz (*Oryza sativa*) possui mais de 120.000 variedades únicas armazenadas em bancos de germoplasma em todo o mundo, representando uma das maiores coleções de recursos genéticos entre as espécies vegetais de interesse econômico (KHUSH, 1997). Não obstante, grande parte da diversidade genética presente nos bancos de germoplasma tem se mostrado extremamente importante para o melhoramento genético (ZHAO et al., 2010). O uso exclusivo de genitores de elite na criação e cultivo resultou em um aumento anual de produção em torno de 1% por ano, índice inferior ao necessário para atender a demanda de consumo prevista para 2050, que seria de 2,4% por ano (RAY; MUELLER; PAUL, 2013). Contudo, aumentos nos ganhos de produtividade desse cultivar só foram possíveis através do estudo de variabilidade genética introduzidos nas populações de melhoramento. Em busca de alternativas melhores e mais eficientes para desenvolver novos cultivares que resultem em maior sustentabilidade da cadeia produtiva, linhas de arroz com as melhores combinações genéticas e alélicas devem ser identificadas, e essa variabilidade deve ser monitorada por marcadores moleculares ao longo das gerações durante o desenvolvimento de novas cultivares (RAY; MUELLER; PAUL, 2013).

As plataformas de sequenciamento e fenotipagem de DNA podem identificar marcadores moleculares relacionados a características quantitativas como produtividade, por meio da análise de Traços Quantitativos de Loci (do inglês, "*Quantitative Trait Loci*") (QTL) e Estudos de Associação de Genoma Completo (do inglês, "*Genome-Wide Association Studies*") (GWAS) (DUBOURG et al., 2018). Além disso, as estratégias de sequenciamento têm gerado uma enorme quantidade de informações (WANG et al., 2018) e com

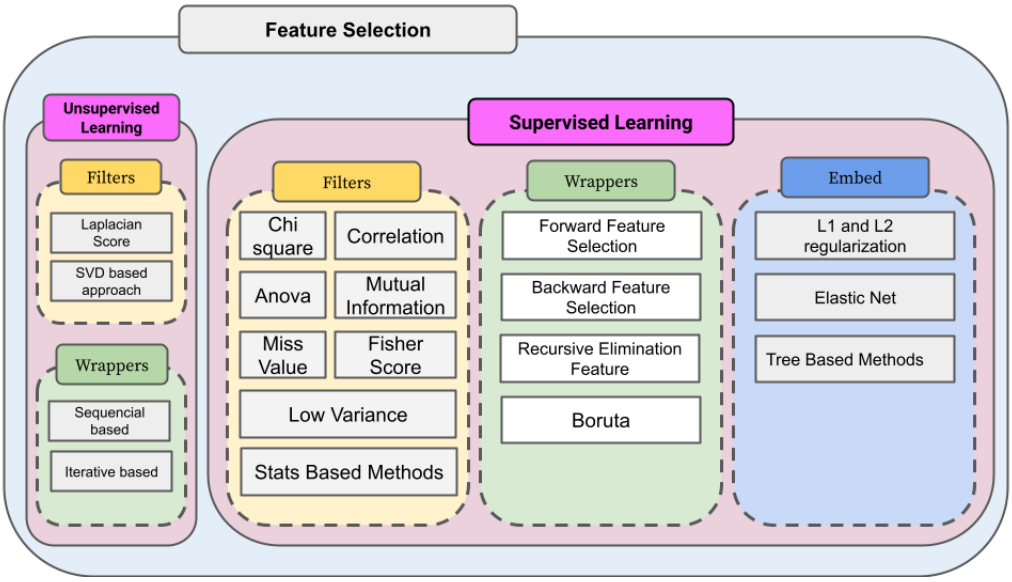
um nível de detalhamento capaz de identificar marcadores moleculares em genes, ou em regiões inter-gênicas que interferem na expressão gênica, podendo apontar elementos que afetam fenótipos de interesse. O SNP é um marcador molecular que têm sido amplamente utilizado em análises genéticas dessa natureza devido à sua grande distribuição genômica e baixo custo de implementação (VOSS-FELS; SNOWDON, 2016). Nos trabalhos de (HUANG et al., 2010) e (LU et al., 2015), através da análise GWAS realizada em arroz, foram identificados centenas de marcadores SNP relacionados a 14 e 12 fenótipos de interesse, respectivamente. Essa disponibilidade de marcadores em ampla cobertura demonstra que, após a etapa de validação, os SNPs têm importante papel como ferramenta de análise genética.

Todavia, a análise para identificação de marcadores associados a fenótipos de interesse era limitada pelo tempo e o alto custo envolvido antes do surgimento do Sequenciamento de Nova Geração (do inglês, "*Next Generation Sequence*") (NGS). Atualmente, o gargalo está na mineração de dados, ou extração de conhecimento útil a partir de enormes bancos de dados (XU; JACKSON, 2019). O uso de Traços Quantitativos de Loci (do inglês, "*Quantitative Trait Loci*") (QTL)s utilizando análises baseadas em mapas de ligação, ou Traços Quantitativos de Nucleotídio (do inglês, "*Quantitative Trait Nucleotides*") (QTN) por GWAS (NAVEED et al., 2018), incrementada pelo crescente sequenciamento de genomas, tem levado a descobertas de diversos alelos de interesse em cultivares estudados, ou seja, identificação de novas mutações que podem ser utilizadas no melhoramento através da seleção assistida de marcadores (GENTZBITTEL et al., 2019).

Contudo, fenótipos quantitativos influenciados por muitos *loci* de menores efeitos são, frequentemente, não preditos por QTLs identificados via mapas de ligação ou GWAS, visto que, tanto a seleção quanto a história de uma população têm influência importante na quantidade e padrões de variabilidade genética, e por consequência, populações com diferentes histórias genéticas podem ter diferenças nas frequências alélicas para muitos polimorfismos ao longo do genoma (FLOOD; HANCOCK, 2017). Se essas populações têm diferentes valores para o fenótipo, qualquer polimorfismo que difere na frequência entre duas populações será associado com o fenótipo, muito embora não sejam nem o causal nem aquele que está em forte equilíbrio de ligação com o polimorfismo causal (GENTZBITTEL et al., 2019).

Por esta razão, novas estratégias de Aprendizado de Máquina (do inglês, "*Machine Learning*") (ML) estão sendo aplicadas, com o objetivo de explicar a variação de características complexas, usando populações com alta variabilidade genética, isto é, populações com muitas mutações, a fim de possibilitar a avaliação de fenótipos quantitativos gerados por algumas dessas mutações (GENTZBITTEL et al., 2019). Ao incluir genótipos com alta variabilidade, as técnicas de ML podem buscar uma variação fenotípica melhor do que outros métodos estatísticos disponíveis para GWAS ou técnicas de mapeamento QTL (GENTZBITTEL et al., 2019).

Figura 1 – Categorias de métodos de seleção de atributos



ML envolve áreas de Ciência da computação, Inteligência Artificial (IA), Estatística computacional e Teoria da informação para construir algoritmos que possam aprender com conjuntos de dados existentes e fazer previsões sobre novos conjuntos de dados (WANG et al., 2018). A metodologia apoiada por algoritmos de ML permite a exploração de conceitos de Big Data em genômica vegetal, no entanto, isso implica grandes desafios relacionados à modelagem do problema de seleção de marcadores moleculares e extração de conhecimento a partir desses marcadores (SILVA et al., 2019).

Nesse contexto, estudos recentes têm usado técnicas de Seleção de Atributos (do inglês, "*Feature Selection*") (FS), uma etapa posterior ao processamento de dados em pipelines de ML, para selecionar estruturas genéticas que têm indicações de regulação de funções biológicas, como SNPs, genes, Micro RNAs (miRNA) e outras estruturas (SYLVESTER et al., 2018; TADIST et al., 2019). Entende-se que o uso de métodos de FS é adequado para modelagem de problemas de seleção de SNPs de interesse, pois esses métodos visam selecionar um subconjunto de variáveis preditoras que possuem um alto fator de relevância para o conjunto de dados (KUHN; JOHNSON, 2019). Além disso, existem vários métodos diferentes de FS disponíveis para seleção de atributos. Eles podem ser divididos em categorias, referenciados pela estratégia de execução que cada um deles adota. A Figura 1 ilustra a esquematização de alguns desses métodos.

Contudo, vale ressaltar que o uso indiscriminado de um algoritmo em detrimento de outro ou a configuração arbitrária de parâmetros de determinado método pode levar a problemas de redução de performance dos algoritmos utilizados (KUHN; JOHNSON, 2019). Além disso, mesmo com o uso correto de determinado método, os trabalhos de Saeys, Inza e Larranaga (2007), Zhou et al. (2019), Pudjihartono et al. (2022) mostraram que para tarefas de seleção de atributos aplicadas a contextos biológicos, uma investigação

cautelosa e robusta que combina diversos algoritmos para confirmação dos resultados deve ser aplicada, a fim de não ocasionar a escolha equivocada de uma determinada estrutura genética pouco relevante ao conjunto de dados.

Outrossim, em conjuntos de dados utilizados para melhoramento genético é comum haver mais de uma característica alvo, isto é, mais de um fenótipo associado aos atributos de entrada. Nesse cenário, é interessante que os algoritmos aplicados busquem atributos que sejam importantes para todos ou maior parte dos fenótipos estudados, aumentando as chances de encontrar SNPs causais, isto é, SNPs que de fato contribuem para regulação de um fenótipo complexo (AL-TASHI et al., 2020). Um exemplo para isso pode ser dado por um estudo que busca os SNPs que tem maior importância na produtividade do arroz. Nesse contexto, a produtividade é um fenótipo complexo, que pode depender de vários fatores, tais como crescimento da planta, resistência a pragas e outras características da planta. Sendo assim, uma estratégia para identificação inicial de SNPs de interesse seria realizar genotipagem, isto é mapeamento dos SNPs, em amostras com perfis genéticos diferentes, ou seja, em amostras com diferentes mutações e em seguida, realizar a análise de quais SNPs tem maior influência sobre os fenótipos estudados.

Além do que foi mencionado, vale ressaltar que fenótipos de interesse podem ser de natureza mista, isto é, compostos por características quantitativas e qualitativas (HAILU; ABDULKADIR, 2023; DROBNIČ et al., 2023). Um modo de lidar com isso é através de técnicas de discretização, que transformam valores contínuos de uma variável em valores categóricos, possibilitando uniformização na aplicação de algoritmos (GOSWAMI; CHAKRABARTI, 2014), todavia essa estratégia ainda não foi totalmente explorada na literatura.

Em conjunto aos procedimentos citados, há necessidade de explicação ou apontamentos intuitivos sobre o fator que associa determinados marcadores aos fenótipos estudados (ZHAO; RHEE, 2023). Em alguns trabalhos, são feitas investigações manuais a partir da busca de referências em sistemas biológicos, a partir dos marcadores selecionados (ALBAWI; MOHAMMED; AL-ZAWI, 2017; ZHANG et al., 2023). Junto a esses procedimentos, alguns estudos tem utilizado como suporte o enriquecimento funcional, adotado para validação estatística de vias tidas como importantes, pois esse método promove um teste estatístico que faz a verificação entre a associação de marcadores putativos e sistemas biológicos estatisticamente relevantes (DRAGO et al., 2015). Porém, por necessitar de algumas informações específicas, referentes ao genoma da espécie estudada, anotações sobre vias e sistemas biológicos catalogadas, essa técnica nem sempre é amplamente utilizada (ZHAO; RHEE, 2023).

Portanto, levando em consideração os diversos fatores envolvidos em um estudo de melhoramento genético utilizando algoritmos de ML, tais como padrões de execução de cada algoritmo, combinação das técnicas e o procedimento de investigação de resultados obtidos, observa-se a necessidade do desenvolvimento de um protocolo de padronização,

que pode ser proposto no formato de um *pipeline*, simplificando a aplicação e investigação dos melhores métodos de FS voltados ao melhoramento genético de arroz.

1.2 Hipótese e objetivos

Agrupando as necessidades levantadas na seção anterior referentes à utilização de ML em melhoramento genético de arroz, a hipótese desse projeto de mestrado pode ser apresentada da seguinte forma:

- ❑ É possível definir um *pipeline* que seja capaz de selecionar SNPs importantes para diferentes fenótipos de interesse de arroz, a partir da combinação de técnicas de FS.

Este trabalho tem como objetivo geral propor um *pipeline* contendo:

1. uma metodologia de filtragem e combinação de resultados obtidos em diferentes técnicas de FS na tarefa de seleção de SNPs;
2. uma padronização da avaliação de SNPs de interesse presentes em um conjunto de dados que possui múltiplos fenótipos;
3. uma estratégia que aponte estruturas funcionais enriquecidas, tais como genes, miRNAs, vias metabólicas e proteínas, a partir de um conjunto de SNPs associados aos fenótipos.

Para atingir o objetivo, foram implementados e avaliados alguns algoritmos de FS utilizados na literatura, sendo adaptados para execução em variáveis discretas, que foram geradas, tal como nas variáveis quantitativas já existentes, provenientes de conjuntos de dados com SNPs de arroz fornecidos pela Embrapa. Esses algoritmos foram então testados através de algumas metodologias de execução e filtragem, propostas pelo autor, de modo a evidenciar os SNPs que foram selecionados, consensualmente, pelos algoritmos aplicados, permitindo a comparação entre as pontuações e SNPs apontados como importantes por cada algoritmo. Foi desenvolvida uma metodologia de pontuação que combina os resultados obtidos pelos diversos algoritmos aplicados. Posteriormente foram criados scripts, disponibilizados na linguagem de programação Python, contendo os algoritmos e estratégias de filtragem e combinação adotadas. Por fim, é aplicada a abordagem de enriquecimento funcional sobre bancos de dados biológicos, permitindo a caracterização de estruturas funcionais a partir dos SNPs selecionados do arroz.

Adicionalmente, este trabalho tem como objetivo específico explorar de maneira sistemática o conjunto de dados gerado e disponibilizado pela Embrapa, a fim de:

1. Apontar possíveis SNPs tidos como marcadores fenotípicos de arroz ligados à produtividade e características válidas ao melhoramento do cultivar;

2. Contextualizar estruturas funcionais obtidas a partir dos SNPs selecionados, caracterizando funções biológicas putativas que são influenciadas;

3. Comparar os métodos de seleção de SNPs propostos com alguns métodos existentes na literatura.

1.3 Contribuições do trabalho

O presente trabalho apresenta contribuições importantes para ciência da computação e estudos de melhoramento genético. Aqui é realizado um estudo comparativo da utilização de diversos métodos de seleção de atributos para a tarefa seleção de SNPs de interesse. São comparados aqui tanto algoritmos utilizados na tarefa de classificação como de regressão. Para isso, foi realizada uma modelagem, que utilizou a discretização de características alvo, permitindo a comparação na performance entre técnicas semelhantes utilizadas nas tarefas de classificação e regressão.

Assim, este trabalho propôs procedimentos de investigação e seleção de SNPs utilizando múltiplas características alvos, isto é, múltiplos fenótipos. Sendo que essas etapas podem ser unidas em um *pipeline* sistematizado. Isso foi feito através da conciliação de diversos métodos de seleção de atributos. Além disso, os métodos propostos para a etapa de seleção dos SNPs foram comparados com métodos presentes na literatura. Alguns SNPs passíveis de investigação sistemática foram apontados, para os diferentes fenótipos. Além disso, o *pipeline* incorpora uma abordagem de enriquecimento funcional, que possibilitou a análise de vias e sistemas biológicos, possivelmente impactados pelos SNPs apontados como importantes.

Por fim, vale ressaltar que todos os dados utilizados nesse estudo, tanto brutos, como aqueles gerados pelas etapas propostas foram disponibilizados. Isso permite incorporações e evoluções futuras da proposta, tal como outras investigações a partir dos dados analisados.

1.4 Organização do documento

O restante deste documento está organizado da seguinte forma:

- Capítulo 2 - Seleção de Atributos: Apresenta conceitos básicos de FS utilizados para a modelagem do problema de seleção de SNPs. São demonstrados alguns dos principais paradigmas utilizados na literatura e seus respectivos algoritmos. Os algoritmos descritos são aqueles que foram implementados como parte do desenvolvimento desse estudo;

- Capítulo 3 - SNP: Apresenta conceitos importantes sobre o contexto biológico e aplicações relevantes que podem se beneficiar com os resultados do presente trabalho;
- Capítulo 4 - Enriquecimento Funcional: Apresenta conceitos e algoritmos convencionalmente usados após um estudo de análise genética, bem como alguns dos principais paradigmas utilizados na literatura e seus respectivos algoritmos;
- Capítulo 5 - Proposta: Retoma a motivação e os objetivos para o trabalho, além do detalhamento geral da proposta - dados, *pipeline* e implementação;
- Capítulo 6 - Experimentos e Resultados: Nesse capítulo são apresentados os resultados obtidos nos experimentos realizados nesse estudo;
- Capítulo 7 - Considerações Finais e Trabalhos Futuros: Nesse capítulo, são discutidos aspectos gerais sobre os resultados obtidos, tal como considerações sobre trabalhos futuros a serem desenvolvidos a partir do que foi estudado aqui;
- Apêndice A - Coleta e Preparação do Conjunto de Dados: Esse apêndice trata dos detalhes referentes ao conjunto de dados utilizado, tal como nuances de mensuração;
- Apêndice B - Referência aos Dados e Resultados: Esse apêndice apresenta tabela com as urls contendo os dados base utilizados nos experimentos;
- Apêndice C - Resultados Expandidos: Esse apêndice trata dos detalhes obtidos nos experimentos realizados.

Capítulo 2

Seleção de Atributos

Este capítulo traz uma fundamentação teórica a respeito do problema de FS, que será utilizado em parte da modelagem do problema de seleção de SNPs. Serão apresentados métodos utilizadas e explorados nesse trabalho, bem como os paradigmas aos quais cada método é pertencente. O capítulo está organizado em três seções que tratam, respectivamente, do problema de genérico de FS, as categorias e algoritmos de FS existentes, e metodologias de otimização de hiper-parâmetros. Nesse trabalho iremos explorar apenas métodos que utilizam algoritmos supervisionados, pois pretende-se explorar o conjunto de dados disponibilizados, com ênfase nos rótulos disponíveis.

2.1 Contexto

A seleção de atributos é um processo de preparação dos dados que consiste na redução de dimensionalidade, ou seja, na redução do número de variáveis de entrada em um conjunto de dados, sendo uma estratégia ótima para o aumento da eficiência dos algoritmos de mineração de dados e ML (LI et al., 2017). Os métodos de FS podem ser usados em diversos contextos, além disso é possível sintetizar dois principais objetivos de seu uso (CHANDRASHEKAR; SAHIN, 2014; JOVIĆ; BRKIĆ; BOGUNOVIĆ, 2015; KAMKAR et al., 2015):

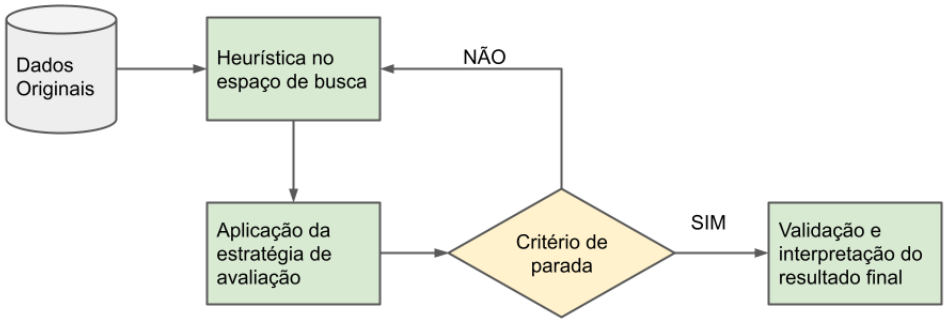
- Simplificação de modelos: torná-los mais fáceis de interpretar por pesquisadores e usuários, permitindo a busca e interpretação dos atributos, seja individualmente ou em grupo;
- Tempo de treino: com a eventual redução dos atributos, o tempo de treino de

algoritmos de ML e estatística se tornam mais curtos, o que em alguns casos pode ser um fator fundamental no que diz respeito à atualização periódica.

Avançando na discussão sobre FS, podemos apontar quatro etapas frequentemente presentes na seleção de atributos, que podem ser ilustradas pela Figura 2, são elas: (1) a geração de candidatos através de uma heurística de busca, (2) aplicação de uma avaliação de candidatos, (3) o estabelecimento de um critério de parada e (4) validação e interpretação dos resultados gerados (KUMAR; MINZ, 2014; KUHN; JOHNSON, 2019; VENKATESH; ANURADHA, 2019).

Na primeira etapa, os atributos são inicialmente selecionadas seguindo um critério de busca randomizado ou otimizado de acordo com heurísticas vinculadas ao espaço de estados do problema (VENKATESH; ANURADHA, 2019). Além disso, podemos caracterizar essa etapa como sendo ponto de partida para os algoritmos de seleção. Posterior a isso, seguimos com um critério de avaliação, que pode ser individual ou em grupo, sendo que a principal diferença é que no primeiro caso, cada atributo selecionado é avaliado de maneira individual em relação à sua importância, enquanto que no segundo caso, a cooperação dos atributos é avaliada como fator de permanência (KUMAR; MINZ, 2014). Não obstante, para seguir com a avaliação dos atributos é estabelecido um critério de parada que determina se o algoritmo deve seguir com o espaço de busca disponível, caso o critério seja atingido, então o conjunto de atributos de entrada selecionado é extraído e pode passar por uma exploração sistemática através de algoritmos de validação ou investigação manual (KUHN; JOHNSON, 2019).

Figura 2 – Procedimento de seleção de atributos



Fonte: Adaptado de (KUMAR; MINZ, 2014)

Compreendido o processo de seleção genérico, vale ressaltar que os métodos de seleção

de atributos podem ser agrupados em duas categorias de algoritmos, os supervisionados e não supervisionados (KUMAR; MINZ, 2014). No primeiro caso as variáveis alvo, isto é os rótulos e mensurações conferidos aos dados, são utilizadas no processo de seleção das melhores características, portanto, o processo de avaliação busca maximizar a representatividade das variáveis alvo através dos atributos de entrada (JOVIĆ; BRKIĆ; BOGUNOVIĆ, 2015). Em contraste a primeira categoria, os métodos que utilizam algoritmos não-supervisionados não utilizam os rótulos vinculados aos dados para seleção de características, mas concentram-se na manutenção da estrutura de dados do conjunto de variáveis (SOLORIO-FERNÁNDEZ; CARRASCO-OCHOA; MARTÍNEZ-TRINIDAD, 2020).

2.2 Categorias de algoritmos de seleção de atributos

A partir da Figura 1, observamos que existem algumas classificações passíveis aos algoritmos supervisionados de seleção de atributos. Essas classificações são relacionadas aos paradigmas de execução de cada um deles. As três principais classes de algoritmos são *wrapper*, *embeded* e *filter*. Nesse trabalho não discutimos métodos da categoria *filter*, pois um dos objetivos é gerar comparações entre métodos de FS que incorporam algoritmos de classificação e regressão e assim como é apresentado por Pudjihartono et al. (2022) os métodos *filter* fazem a seleção de atributos de modo independente do algoritmo de predição, isto é, sem levar em consideração métricas de performance.

2.2.1 Embeded

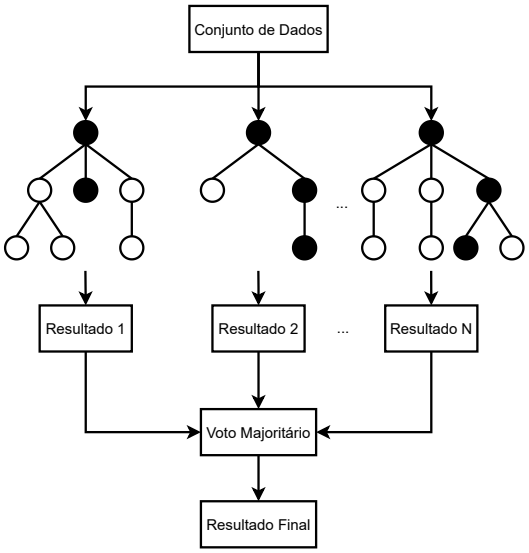
Os métodos Embeded, denominados métodos intrínsecos ou embutidos, são métodos de seleção de atributos que realizam a filtragem de atributos no momento em que o algoritmo de aprendizado é treinado e por conta disso possuem essa designação, ou seja, a estratégia de seleção é incorporada ao algoritmo (LI et al., 2017; SOLORIO-FERNÁNDEZ; CARRASCO-OCHOA; MARTÍNEZ-TRINIDAD, 2020). Nas subseções seguintes são apresentados os algoritmos representantes dessa categoria e que foram utilizados no presente trabalho.

2.2.1.1 Random Forest e Extra Tree

A Floresta Aleatória (do inglês, "*Random Forest*") (RF) é um algoritmo de ensemble (em português, "combinação"), que utiliza internamente o algoritmo de árvore de decisão (KOTSIANTIS, 2013). Esse algoritmo constrói vários modelos internos, que são treinados e ao final combinados a fim de produzir uma classificação final consensual entre os modelos (GOLDSTEIN; POLLEY; BRIGGS, 2011). Esse fato proporciona maior generalização do modelo e reduz o risco de *overfitting* (em português, "sobreajuste") (MOYANO et al.,

2018). A Figura 3 ilustra uma das estratégias de combinação dos resultados, chamada voto majoritário, onde o resultado da predição equivale ao resultado predito por N árvores.

Figura 3 – Classificação utilizando Random Forest



Não obstante, a RF também é um algoritmo amplamente utilizado para seleção de atributos, tendo algumas aplicações voltadas a seleção de SNPs (SILVA et al., 2022; ZHOU et al., 2021). O processo interno para seleção de atributos, frequentemente implementado nesse algoritmo, se dá pela medida chamada *Mean Decrease Impurity* (MDI) (PEDREGOSA et al., 2011; HASAN et al., 2016). Há também uma medida denominada *Mean Decrease in Accuracy* (MDA), contudo, baseado em alguns trabalhos como de Botta et al. (2014) e Guo et al. (2023), a MDI foi escolhida. O funcionamento da MDI leva em conta o conceito de *impureza*. Esse conceito, em alguns contextos, recebe o nome de *índice de gini* ou *impureza de gini*, sendo calculado como a probabilidade de uma classe não ser corretamente classificada, caso selecionada aleatoriamente (LOUPPE et al., 2013). A Equação 1 apresenta como o *índice de gini* pode ser obtido para atributos categóricos. Nessa equação, $p(i)$ é a probabilidade de um exemplo ser classificado em uma classe específica, c é o número de classes e E é o atributo em avaliação. Se todos os exemplos estiverem vinculados a uma mesma classe, dizemos que o atributo é totalmente puro, enquanto que um desequilíbrio nessa condição é chamado de impureza. Para atributos de natureza quantitativa, a variância representada pela Equação 2 pode ser utilizada (LI et al., 2019).

$$GI(E) = 1 - \sum_{i=0}^{c} p(i)^2 \tag{1}$$

$$VAR(E) = 1\frac{1}{N(t)} \sum_{i=x_i}^{Rt} (y_i - y_n(t))^2 \tag{2}$$

No contexto de árvores de decisão, os nós produzidos durante o treinamento ramificam os diversos valores de um atributo. Sendo assim, a generalização do cálculo de *impureza* sobre um nó t de uma árvore T pode ser realizada segundo a Equação 3, onde a função *impurity(t)* que calcula a impureza pode ser substituída tanto pela *impureza de gini* como pela variância (LI et al., 2019). Compreendidos esses aspectos, podemos generalizar o cálculo de MDI para mensurar a importância de um atributo k para uma árvore de decisão T com a Equação 4, em que $\frac{N(t)}{n}$ é a proporção de exemplos vinculados a uma classe determinada do atributo em um nó t. Observamos que o MDI para uma única árvore T leva em consideração a recursão por cada um dos nós t dessa árvore.

$$\Delta r(t) = impurity(t) + \sum_{left=0}^{lefts} -\frac{N(left)}{N} impurity(left) \tag{3}$$

$$MDI(k,T) = \sum_{t \in I(T)} \frac{N(t)}{n} \Delta r(t) \tag{4}$$

Sendo assim, a partir do cálculo base estabelecido para um única árvore de decisão, a definição de MDI para o algoritmo RF, que leva em consideração várias árvores de decisão, é dado pela média dos valores MDI de cada árvore (LOUPPE et al., 2013; LI et al., 2019). Esse cálculo é apresentado na Equação 5, em que N_{tree} é o número total de árvores e T_n é a enésima árvore de decisão introduzida no modelo.

$$MDI(k) = \frac{1}{N_{\boldsymbol{tree}}} \sum_{n=1}^{N_{\boldsymbol{tree}}} MDI(k, T_n) \tag{5}$$

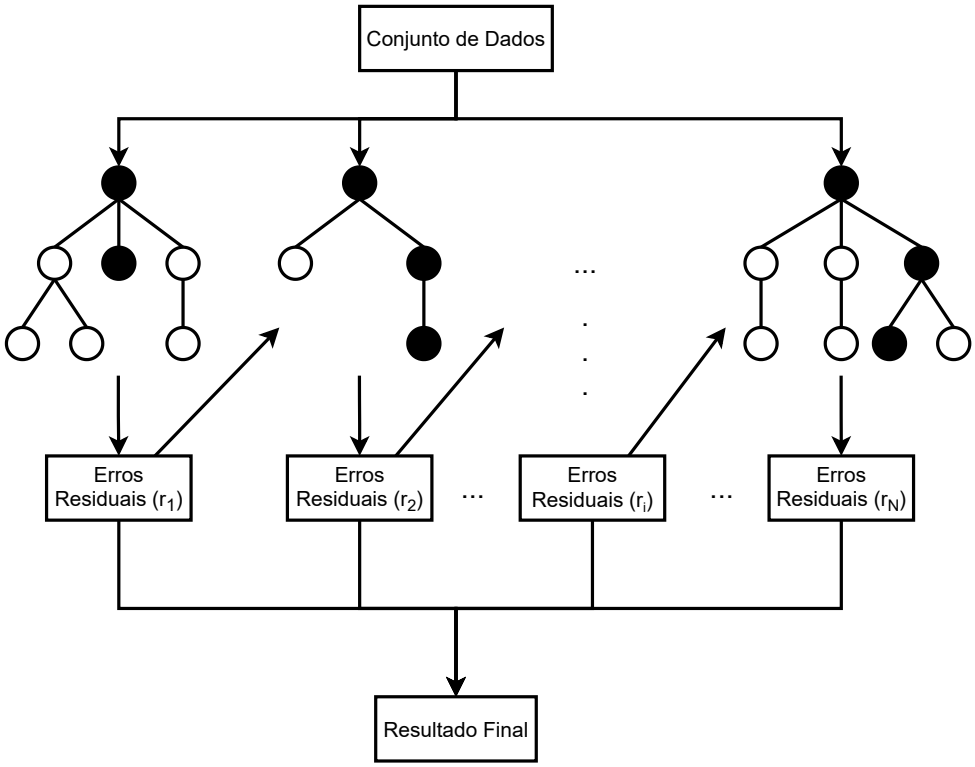
Uma variação do RF é o algoritmo *Extra Trees*, abreviação do termo *Extremely Randomized Trees* (em português, Árvores Extremamente Aleatorizadas), esse algoritmo assim como o RF também treina diversos modelos de árvore de decisão, utilizando o voto consensual como critério de escolha de determinada classe ou média de resultados em casos de regressão (GEURTS; ERNST; WEHENKEL, 2006). No entanto, não aplica a divisão dos dados com reposição de amostras, ou seja, os dados são divididos através de cortes aleatórios (GEURTS; ERNST; WEHENKEL, 2006). Semelhante a RF, este método utiliza amplamente o MDI como critério de relevância de cada atributo.

2.2.1.2 XGBoost

O XGBoost ou *Extreme Gradient Boosting* (XGB) é um modelo de *ensemble* utilizado para tarefas de classificação e regressão (CHEN; GUESTRIN, 2016). O XGB, assim como outras estratégias de *ensemble*, utiliza internamente árvores de decisão. O processo de treinamento consiste em adicionar, iterativamente, modelos que trabalham na predição e correção de erros residuais dos modelos anteriores. Esse procedimento é conhecido como *Gradient Boosting*, em português, gradiente aumentado (AZMI; BALIGA, 2020). O termo gradiente utilizado refere-se à estratégia de minimização de erro residual aplicada

nesse algoritmo (CHEN; GUESTRIN, 2016). A Figura 4 exemplifica o treinamento desse algoritmo. O XGB, assim como outros modelos baseados em árvores de decisão já apresentados nessa subseção, é utilizado em seleção de atributos calculando a importância dos atributos através de métricas de impureza, tal como a já citada MDI (MAGUIRE et al., 2022).

Figura 4 – Processo de treinamento XGBoost

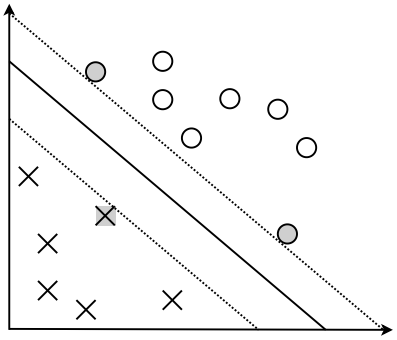


2.2.1.3 SVM

Quando utilizado como classificador, o algoritmo Máquina de Vetores de Suporte (do inglês, "*Support Vector Machine*") (SVM) constrói hiperplanos que buscam realizar uma separação adequada entre classes que ficam dimensionalmente posicionadas em um espaço vetorial definido pelo número de atributos (SANZ et al., 2018). Em problemas com N classes, onde N é um número inteiro maior que 2, uma estratégia alternativa é treinar N classificadores binários que resolverão problemas individuais (SANZ et al., 2018; CRISTIANINI; SHAWE-TAYLOR et al., 2000). A Figura 5 ilustra um hiperplano com margem máxima e seus vetores de suporte.

No entanto, é possível utilizar esse algoritmo no contexto de seleção de atributos. Para tanto, são definidas algumas estratégias de invólucro, assim como evidenciado na Seção 2.2.2, ou mesmo fazendo uso de métricas internas (KUMAR; MINZ, 2014; KUHN; JOHNSON, 2019). Uma estratégia utilizada pela biblioteca scikit-learn (PEDREGOSA et al., 2011) é o uso dos pesos associados aos atributos como critério de importância.

Figura 5 – Ilustração de hiperplano separado pelo algoritmo SVM



Nesse processo duas abordagens de penalização são, frequentemente, utilizadas. São elas a técnicas de regularização Lasso, chamada de regularização L1 (MUTHUKRISHNAN; ROHINI, 2016) e a técnica Ridge, chamada de regularização L2 (PAUL; DRINEAS, 2016). Os cálculos envolvendo essas técnicas são demonstrados, respectivamente, nas Equações 6 e 7.

$$RSS_{lasso} = \sum_{i=1}^{n}[y_p - (\omega_i \times x_i + \beta)]^2 + \underbrace{\alpha \sum_{j=1}^{p} \left|w_j^2\right|}_{lasso} \tag{6}$$

$$RSS_{ridge} = \sum_{i=1}^{n}[y_i - (w \times x_i + b)]^2 + \alpha \sum_{j=1}^{p} w_j^2 \tag{7}$$

As duas abordagens possuem semelhança no critério de otimização, no entanto, o fator chave para diferenciação está na soma final dos coeficientes de características altamente correlacionadas, pois a abordagem L1 assume que tais coeficientes serão zerados. Então, dizemos que esse modelo realiza seleção das atributos automaticamente, gerando vários coeficientes com peso zero (MUTHUKRISHNAN; ROHINI, 2016). No entanto, na abordagem L2 apenas os aproxima de zero (PAUL; DRINEAS, 2016), sendo necessário, então, definir um critério de corte (CRISTIANINI; SHAWE-TAYLOR et al., 2000). Para esse trabalho o SVM foi utilizado em conjunto com Eliminação Recursiva de Atributos (do inglês, "*Recursive Feature Elimination*") (RFE).

2.2.2 Wrapper

Os algoritmos classificados como *wrapper*, chamados de metodologias de invólucro, são métodos que utilizam como métricas de seleção de atributos valores obtidos através da execução de algoritmos de aprendizado de máquina (JOVIĆ; BRKIĆ; BOGUNOVIĆ, 2015; LI et al., 2017). Os métodos dessa categoria, em geral, têm maior tempo de execução e não têm resultados facilmente interpretáveis. No entanto, são métodos que apresentam melhor resultado do ponto de vista de representação do conjunto de dados, pois, utilizam métricas como acurácia para certificar que as características escolhidas têm boa taxa

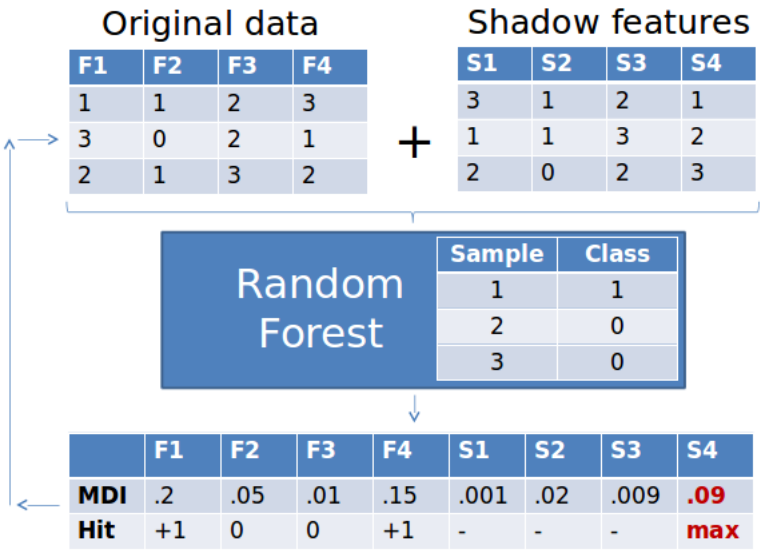
de acerto no modelo de classificação final (ABOUDI; BENHLIMA, 2016). As próximas subseções discutem sobre os principais métodos dessa categoria.

2.2.2.1 Boruta

O algoritmo denominado Boruta é um método baseado em um teste de relevância estatística realizado entre os dados originais e uma cópia contendo a permutação de suas variáveis (KURSA; JANKOWSKI; RUDNICKI, 2010). Para execução desse algoritmo inicialmente é criado um conjunto de dados contendo o mesmo número de atributos do conjunto original. No entanto, cada variável tem seus valores embaralhados. Esse procedimento cria um conjunto de atributos denominados atributos sombra, que são uma versão aleatória dos dados originais. Esse procedimento é realizado a fim de mitigar a correlação entre os atributos de entrada e variáveis alvo. Uma vez gerado, o conjunto de atributos sombra é unido aos atributos originais, formando assim um conjunto de dados mesclado.

Esse conjunto de atributos mesclado é submetido a um algoritmo de aprendizado *ensemble* baseado em árvores de decisão, tal como o RF, que irá ser executado durante n iterações. Em cada iteração é calculada a importância dos atributos e verificado se os atributos originais têm maior pontuação de relevância em relação ao atributo sombra que obteve maior pontuação de relevância. A cada vez que isso acontece é dito que determinado atributo obteve um *hit* (KURSA; JANKOWSKI; RUDNICKI, 2010). O cálculo de importância, amplamente utilizado nesses algoritmos *ensemble* é o MDI (NEMBRINI; KÖNIG; WRIGHT, 2018), apresentado na Subseção 2.2.1.1.

Figura 6 – Aplicação do algoritmo Boruta



Fonte: Extraído de (KURSA; JANKOWSKI; RUDNICKI, 2010)

A Figura 6 retrata uma iteração do algoritmo Boruta. Podemos observar um conjunto de dados com 4 variáveis, que tem prefixo F seguido de uma indicação numérica que vai de 1 até 4. Essas variáveis são, portanto, embaralhadas gerando o conjunto de atributos sombra, apresentados ao lado com um prefixo S seguido de uma indicação numérica que vai de 1 até 4. O algoritmo RF foi, então aplicado com o MDI. Logo, podemos ver na tabela de indicação ao final da imagem, pela coluna MDI, a pontuação obtida por cada um dos atributos do conjunto de dados originais e sombra. Notamos que o atributo sombra S4 obteve o maior valor de importância entre todos os atributos sombra. A partir dele é possível comparar os demais atributos. Essa etapa demonstra que apenas os atributos F1 e F4 obtiveram pontuações maiores, exemplificando *hits* atribuídos a esses atributos.

Ao final das iterações definidas, o resultado dos *hits* é utilizado como critério de aceitação ou recusa de determinada variável. Nessa etapa o número de iterações bem como o número de *hits* de uma variável testada são passados, respectivamente, como parâmetros k e N no cálculo do valor de p utilizando a distribuição binomial (KURSA; JANKOWSKI; RUDNICKI, 2010). É dito aceito uma variável que possui o valor de p inferior ao critério p de significância definido. Uma variável também pode ser recusada caso não tenha *hits* suficientes até determinada iteração. O principal objetivo desse método é utilizar a aleatoriedade como critério de seleção de variáveis, removendo variáveis ditas irrelevantes através de um teste estatístico que determina se uma variável é tão relevante que sua permanência no modelo é capaz de sobressair outras variáveis inseridas ao acaso.

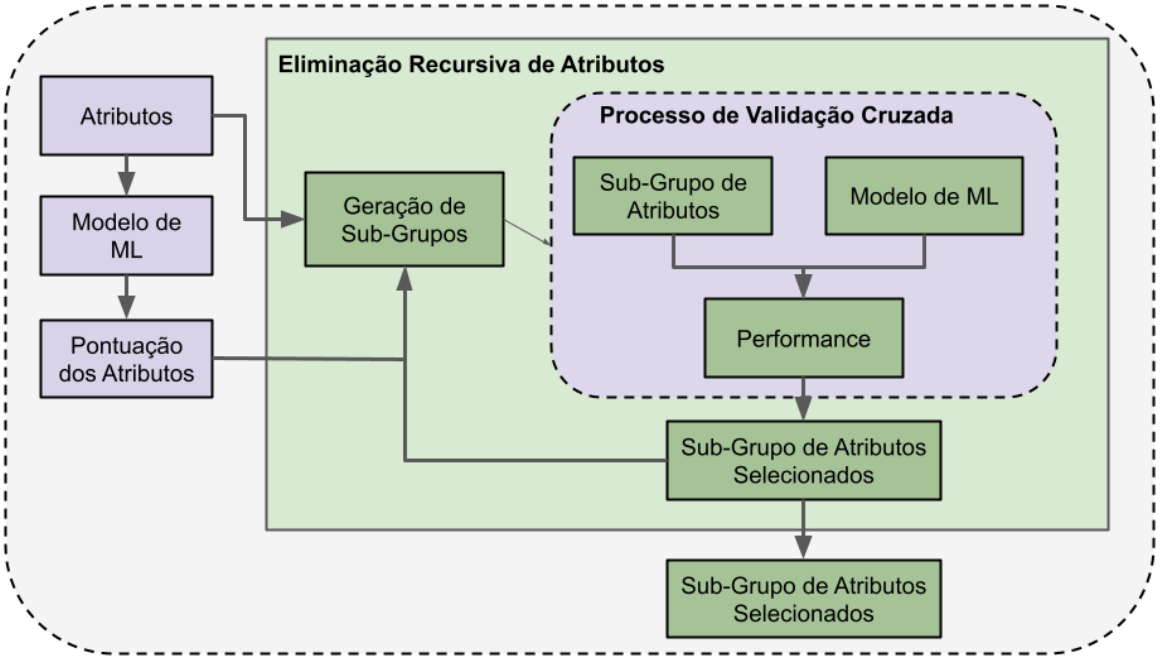
2.2.2.2 Eliminação Recursiva de Atributos

O método RFE consiste em um processo de eliminação que remove atributos menos importantes iterando sobre subconjuntos formados através do conjunto inicial de atributos explorado. A cada iteração o subconjunto resultante maximiza a precisão preditiva em relação ao subgrupo anterior. Nesse método são construídos modelos treinados com subconjuntos de atributos, permitindo a atribuição de coeficientes de importância a cada atributo pertencente a um subconjunto gerado. Isso permite a remoção dos atributos com menor pontuação final em cada subconjunto. Ao final de cada iteração, o mesmo procedimento é efetuado com o conjunto podado de atributos resultantes da etapa anterior (GUYON et al., 2002; KUHN; JOHNSON, 2019). Esse processo é efetuado n vezes, ou até que um número c de atributos especificado seja alcançado. Sendo assim, é fundamental ajustar o número de atributos a serem avaliados de modo a não sobrecarregar o modelo (DARST; MALECKI; ENGELMAN, 2018), pois esse método também se baseia em uma estratégia gulosa que pode ser desaconselhada em determinados contextos (KUHN; JOHNSON, 2019).

A definição dos parâmetros n e c é um importante passo na execução desse método (PUDJIHARTONO et al., 2022). Um exemplo arbitrário para escolha desses parâmetros em um estudo que utiliza esse método para seleção de SNPs pode ser o de iterar até que

se obtenha um número de SNPs menor que 20, onde esse valor pode estar atrelado a uma restrição de custos envolvendo uma validação experimental posterior, tal como o de Sondas Taqman (MCGUIGAN; RALSTON, 2002).

Figura 7 – Processo de eliminação recursiva de atributos - RFE



A Figura 7 exemplifica a execução desse método. Podemos ver que o conjunto total de atributos é treinado em um modelo de ML que recolhe a pontuação dos atributos a ser comparada. Posteriormente, o conjunto total de atributos é então subdividido em mais subgrupos que são treinados com a mesma versão do modelo de ML subdividindo o treino em k grupos, a fim de que uma validação cruzada seja realizada. Ao final, o subgrupo de atributos é então comparado. Se o subgrupo de atributos gerado atinge os parâmetros de parada estabelecidos, o subconjunto gerado é retornado. Caso os critérios de parada do método não tenham sido alcançados o processo continua. A RFE pode ser usada com diversos algoritmos de ML, permitindo assim critérios de comparação híbrida (KUHN; JOHNSON, 2019).

2.2.2.3 Seleção Sequencial de Atributos

Métodos de Seleção Sequencial de Atributos (do inglês, "*Sequential Feature Selection*") (SFS) removem ou adicionam determinados atributos ao subconjunto de atributos final de maneira sequencial (AHA; BANKERT, 1995). A função de avaliação aplicada em cada iteração desse método pode ser a acurácia, aplicada em problemas com algoritmos de classificação, ou erro quadrático médio, para algoritmos de regressão (RÜCKSTIESS; OSENDORFER; SMAGT, 2011). Esse método, assim como RFE, permite a utilização em conjunto com diversos algoritmos de aprendizado máquina. No entanto, algumas aplica-

ções comuns são feitas utilizando algoritmos como K-Vizinhos Mais Próximos (do inglês, "*K-Nearest Neighbors*") (KNN) e Máquina de Vetores de Suporte (do inglês, "*Support Vector Machine*") (SVM) com critério estabelecido por Método do Gradiente Estocástico (do inglês, "*Stochastic Gradient Descent*") (SGD) (PARK; KIM, 2015; BEULAH; PUNITHAVATHANI, 2018).

Apesar da designação genérica, os métodos de SFS podem ser subdivididos de acordo com a direção de início da busca. Existem, convencionalmente duas principais possibilidades, a saber (AHA; BANKERT, 1995; MARCANO-CEDEñO et al., 2010; RÜCKSTIESS; OSENDORFER; SMAGT, 2011):

- Seleção Direta: nessa estratégia os atributos testados são, iterativamente, adicionados ao conjunto final de atributos selecionados, de modo a atender o critério de satisfação;
- Seleção Inversa: nessa estratégia os atributos testados são, iterativamente, removidos do conjunto final de atributos selecionados. Começamos com um conjunto inicial de atributos equivalente ao número total de atributos a serem selecionados.

2.3 Métricas de avaliação

O processo de avaliação dos resultados de um método de FS para aplicação a determinado problema, de modo geral, é dado pela avaliação de performance de um algoritmo de aprendizagem treinado com os atributos selecionados pelo método de FS comparados ao treinamento do mesmo algoritmo com todos os atributos (HOSSIN; SULAIMAN, 2015), ou seja, é avaliado se os atributos escolhidos preservaram ou aumentaram alguma métrica de performance do algoritmo de aprendizado utilizado na tarefa de predição.

A Tabela 1 exibe algumas métricas amplamente utilizadas na avaliação de algoritmos utilizados em tarefas de classificação (HOSSIN; SULAIMAN, 2015). Nela são exibidos o nome da métrica, abreviação comum, fórmula que a representa e descrição do objetivo da métrica. O prefixo tp refere-se aos casos verdadeiros positivos, tn verdadeiros negativos, fp falsos positivos e fn falsos negativos. Não obstante, uma métrica que leva em conta a combinação das métrica anteriores é a Área Sob a Curva (do inglês, "*Area Under Curve*") (AUC) aplicada sobre a curva *Receiver Operating Characteristic* (ROC), que é uma curva de avaliação proveniente do cruzamento entre especificidade e revocação (HUANG; LING, 2005).

Para algoritmos treinados para tarefa de regressão, convencionalmente, são utilizados medidas de erro que mensuram como as predições realizadas por um modelo se distanciam do valor esperado (NASER; ALAVI, 2021). A Tabela 2 exibe algumas das principais métricas utilizadas nessa tarefa. São exibidas 4 informações sobre cada métrica, a saber, o nome, abreviação comum, fórmula que a representa e descrição do objetivo da métrica.

Tabela 1 – Métricas de avaliação de algoritmos de classificação

Nome	Abreviação	Fórmula	Descrição
Acurácia	acc	$\frac{tp+tn}{tp+tn+fp+fn}$	A métrica de acurácia mede o proporção de previsões corretas sobre o total número de exemplos avaliados.
Taxa de Erro	err	$\frac{fp+fn}{tp+tn+fp+fn}$	A métrica de taxa de erro para a classificação mede a proporção de previsões incorretas sobre o número total de exemplos avaliados.
Especificidade	sp	$\frac{tn}{tn+fn}$	Esta métrica é usada para medir a fração de exemplos negativos de um atributo alvo que são classificados corretamente.
Precisão	p	$\frac{tp}{tp+fp}$	A precisão é usada para medir os exemplos positivos que são previstos corretamente a partir do total de exemplos previstos em uma classe positiva, pertencente a um atributo alvo.
Revocação	r	$\frac{tp}{tp+tn}$	A revocação é uma métrica que indica, das amostras positivas existentes, quantas o modelo conseguiu classificar corretamente.
Métrica F1	$F1$	$\frac{2\times p\times r}{p+r}$	Esta métrica representa a média harmônica entre valores de revocação e precisão

Fonte: Extraído de (HOSSIN; SULAIMAN, 2015)

Embora as métricas apresentadas sejam úteis na mensuração de performance de algoritmos de aprendizado, uma prática comum aliada as métricas de avaliação é a Validação Cruzada (do inglês, "*Cross Validation*") (CV), que consiste em uma estratégia de divisão do conjunto de dados de treino e teste, a partir da permutação entre os exemplos disponíveis (YATES et al., 2023). Essa abordagem, de modo geral, denominada de K-fold CV, consiste em realizar a amostragem do conjunto de dados de k modos diferentes, treinando o algoritmo de aprendizagem com as diferentes segmentações dos dados, tirando uma média da performance do algoritmo nas diferentes divisões dos dados (BROWNE, 2000; FUSHIKI, 2011). Essa abordagem permite a observação de um panorama geral do comportamento do algoritmo em diferentes facetas dos dados (FUSHIKI, 2011).

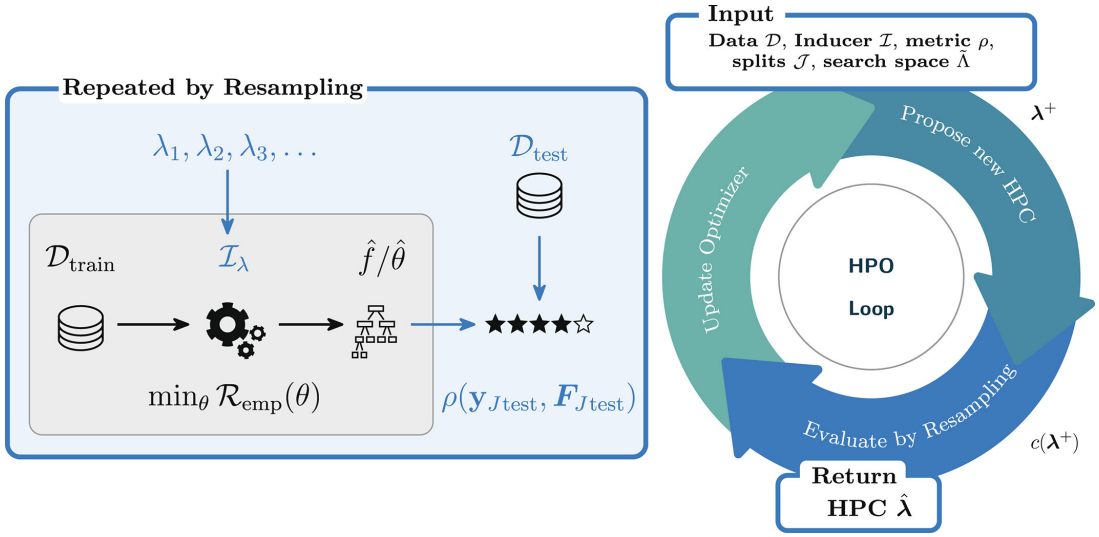
Tabela 2 – Métricas de avaliação de algoritmos de regressão

Nome	Abreviação	Fórmula	Descrição
Coeficiente de Determinação	R^2	$\frac{\sum_{i=1}^{n}(y_i-y)^2}{\sum_{i=1}^{n}(y_i-y)^2}$	Representa a proporção da variância (de y) que foi explicada pelas variáveis independentes do modelo. Ele fornece uma indicação da qualidade do ajuste e, portanto, uma medida de quão bem as amostras não vistas provavelmente serão previstas pelo modelo, por meio da proporção da variância explicada
Erro Médio Absoluto	MAE	$\frac{1}{n}\sum_{i=0}^{n}\lvert y_i - y\rvert$	Essa métrica mede a média da diferença absoluta entre o erro de cada predição e o valor esperado, onde n é o número de exemplos. Quando um modelo não apresenta erros, o MAE é igual a zero
Erro Quadrático Médio	MSE	$\frac{1}{n}\sum_{i=0}^{n-1}(y_i - y)^2$	Esse modelo mede a média de erro entre cada predição os valores esperados, porém penalizando os erros, através do quadrado da diferença entre eles. Quando um modelo não apresenta erros, o MSE é igual a zero
Raiz do Erro Quadrático Médio	$RMSE$	$\sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(y_i - y)^2}$	O RMSE leva em consideração a mesma ideia do MSE, porém aplica uma raíz quadrada ao resultado, permitindo ao resultado ficar na mesma escala do dado original

Fonte: Extraído de (HOSSIN; SULAIMAN, 2015; NASER; ALAVI, 2021)

2.4 Otimização de Hiper-Parâmetros

Figura 8 – Problema genérico de otimização de Hiper-Parâmetros



Fonte: Extraído de (BISCHL et al., 2023)

Assim como já foi apresentado, a modelagem de FS está ligada a utilização de algoritmos de ML, seja para a validação de atributos obtidos em uma etapa posterior ou mesmo como parte integrante da obtenção dos atributos. No entanto, um problema chave na utilização de algoritmos de ML diz respeito a Otimização de Hiper-Parâmetros (do inglês, "*HyperParameter Otimization*") (HPO), isto é, o ajuste de coeficientes, critérios de parada ou mesmo modos de execução do algoritmo de aprendizado, que permitem que os mesmos obtenham melhores resultados em uma tarefa de predição (ANDONIE, 2019; YANG; SHAMI, 2020). A cada um desses parâmetros, passíveis de customização, é dado o nome de hiper-parâmetro (YANG; SHAMI, 2020). O problema genérico de HPO pode ser descrito como a busca da combinação de hiper-parâmetros λ, entre o espaço de combinações de hiper-parâmetros I_λ, que maximiza a métrica de performance ρ de um algoritmo treinado com um conjunto de dados de treino D_{train}, avaliado em um conjunto de dados de teste D_{test} (BISCHL et al., 2023). A Figura 8 exemplifica o enunciado apresentado.

O processo de descoberta dos melhores hiper-parâmetros a serem utilizados pode ser realizado de diversos modos. No entanto, três algoritmos se destacam, pela ampla adesão em trabalhos (ANDONIE, 2019; YANG; SHAMI, 2020; BISCHL et al., 2023):

- ❏ Busca Aleatória: testa de modo randômico, combinações organizadas entre os hiper-parâmetros no espaço de possibilidade disponível. Apresenta um tempo de execução limitado ao número de combinações definidas;

- ❏ Busca em Grade: encontra sempre a escolha ótima, que maximiza o desempenho do algoritmo com os parâmetros disponíveis no espaço de busca;

- ❑ Otimização Bayesiana: utiliza o teorema de Bayes para construir um modelo probabilístico a partir de um conjunto de hiper-parâmetros. Então, ele usa um algoritmo de regressão, a fim de escolher iterativamente o melhor conjunto de hiper-parâmetros, a partir de um espaço de busca.

Entre os algoritmos apresentados, o único que sempre encontra a solução ótima, isto é, que maximiza a função de performance é a busca em grade (BISCHL et al., 2023). No entanto, a desvantagem desse algoritmo está no tempo de execução, que pode se tornar proibitivo, de acordo com o espaço de busca disponível, pois ele testa todas as combinações possíveis de parâmetros. A busca aleatória, é uma técnica simples, porém, pode não encontrar a combinação ótima de parâmetros para solução. Nesse contexto, estratégias semelhantes a otimização Bayesiana, onde heurísticas são utilizadas para encontrar os melhores parâmetros, têm sido amplamente utilizadas (YANG; SHAMI, 2020). Os dois primeiros algoritmos têm implementações disponíveis na biblioteca scikit-learn (PEDREGOSA et al., 2011), já a otimização Bayesiana pode ser utilizada a partir da biblioteca scikit-optamize (LOUPPE, 2017).

Em (SHAFIEE et al., 2021), a busca em grade foi utilizada para HPO em uma investigação de performance entre os algoritmos SVM e LASSO na tarefa de prever a produtividade dos grãos de trigo e encontrar os atributos mais relevantes para essa predição. Para comparação entre LASSO e SVM na seleção de atributos mais relevantes foi utilizada a técnica SFS aliada com SVM. Nesse trabalho os autores utilizaram a busca em grade para encontrar a melhor combinação de parâmetros para cada algoritmo. O teste de performance foi feito com a validação cruzada com 10 *folds* para os dois algoritmos utilizados. Os parâmetros investigados para o algoritmo SVM foram C, *gamma* e *kernel*, sendo que um intervalo de valores espaçados foi utilizado para os valores de C, isto é, no lugar de realizar o teste de todos valores possíveis de C variando de 1 até 1000, alguns valores crescentes nesse intervalo foram utilizados. Para o algoritmo LASSO, Um intervalo sequencial de valores C foram testados, indo de 0.001 até 1000. Essa estratégia de HPO demonstra que em determinados cenários, a busca em grade pode ser aplicada com intervalos sequenciais de valores para algoritmos que possuem poucos parâmetros, tal como LASSO, mas também em investigações que exploram combinações espaçadas de valores tal como foi realizado com SVM.

Yin e Li (2022) investigaram a eficácia do algoritmo de otimização Bayesiana para HPO dos algoritmos RF e XGB treinados para tarefa de predição de prospectividade mineral. Os algoritmos treinados também foram utilizados na tarefa de seleção dos atributos mais relevantes para o conjunto de dados utilizado. Os autores utilizaram como métrica de maximização o valor da AUC. A ênfase dada para o XGB está sobre os parâmetros relacionados ao booster, a saber, *min_child_weight*, *max_depth*, *sub_sample*, *colsample_bytree*, *gamma* e *lambda*. Para a RF os parâmetros explorados

foram max_depth, $max\_features$, $min\_samples\_split$ e $n\_estimators$. Para o XGB os resultados demonstraram ganhos de 7 pontos percentuais entre o algoritmo inicializado.

Capítulo 3

Seleção de Polimorfismos de Nucleotídio Único

Nesse capítulo será tratado o contexto biológico dos SNPs, bem como aplicações voltadas ao melhoramento de arroz a partir da seleção de SNPs associados a fenótipos. Aqui serão apresentadas estratégias utilizadas para seleção de SNPs presentes em alguns trabalhos que utilizaram técnicas de FS para essa tarefa. O capítulo está dividido em duas seções que tratam, respectivamente, do contexto de utilização de SNPs e trabalhos relacionados à seleção de SNPs em contextos de melhoramento de arroz e aplicações genéricas que exploraram marcadores SNPs.

3.1 Polimorfismo de Nucleotídio Único

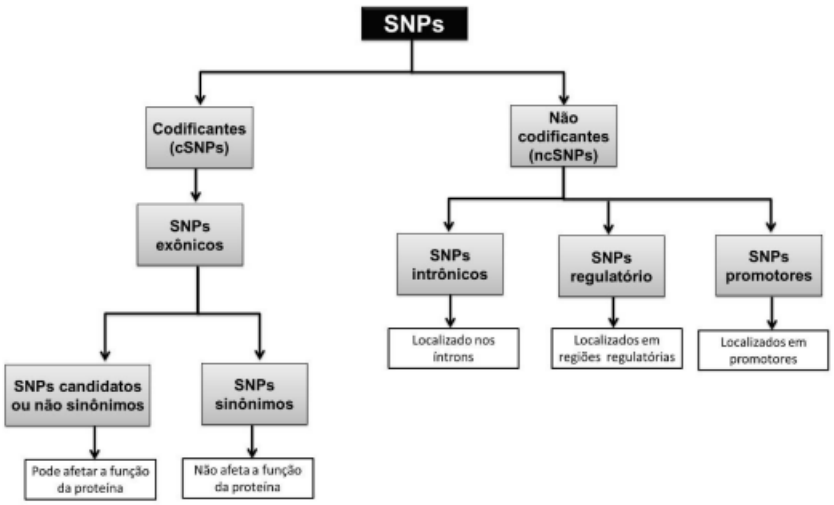
Em estudos genéticos são utilizados dois termos para representação do material genético e suas alterações. O primeiro é o denominado genótipo que pode ser definido como sendo a constituição genética de um indivíduo, isto é seu material genético (). O segundo termo é o fenótipo que refere-se ao conjunto de características morfológicas, fisiológicas ou representativas de um indivíduo (). Podemos tratar cada uma das características mencionadas como características fenotípicas.

Além disso, o genótipo de um indivíduo assemelha-se, em grande porção, ao genótipo de outros indivíduos da mesma espécie (ZHU; ZHOU, 2020). No entanto, é possível acontecerem alterações genéticas na sequência de DNA, que modificam o material genético do indivíduo, podendo ocasionar mudanças em funções biológicas. Quando uma variação ocorre de maneira pontual, isto é, em um único nucleotídio, chamamos tal variação

de SNP (KIM; MISRA, 2007; SHASTRY, 2007), essa variação pode ser, genericamente, denominada mutação ou polimorfismo. Uma alteração pontual na sequência de DNA, isto é, um SNP, ocorre entre bases purínicas (A/G) ou entre bases pirimidínicas (C/T). Além disso, as mutações acontecem, essencialmente, de duas maneiras: (1) como mutações somáticas, isto é, ocorrendo apenas em um tecido ou células específicas ou (2) como mutações germinativas, onde a mutação é ocasionada em um processo embrionário, tornando-se disponível no material genético para próximas gerações (KIM; MISRA, 2007).

Os SNPs são amplamente distribuídos no genoma e estão presentes em regiões codificadoras (éxons) e não codificadoras (íntrons e regiões intergênicas) (ZOLET et al., 2017). Além disso, a presença de determinado SNP pode provocar alteração do produto gênico, isto é, dos aminoácidos gerados no processo de codificação. Esse fato pode ocorrer caso uma nova base seja posicionada em uma região que será codificada, pois a combinação dessa base modificada com as bases adjacentes pode produzir um aminoácido diferente do que seria produzido naquela posição, em uma cadeia de DNA sem tal alteração. Esse tipo de polimorfismo é chamado de mutação não sinônima, visto que a presença do SNP acarretou uma alteração da proteína gerada (KIM; MISRA, 2007). Um efeito distinto do apresentado é a chamada mutação sinônima, onde a presença de um ou mais polimorfismos não altera o aminoácido gerado. Um exemplo para tal mutação é o de uma sequência de DNA com as bases TCG, que no processo de mutação sofreu uma substituição das bases, ficando com a sequência AGC. Ambas produzirão um códon Serina, de modo que as mutações presentes na segunda sequência não modificaram, diretamente, a proteína gerada.

Figura 9 – Classificação dos SNPs



Fonte: Extraído de (ZOLET et al., 2017)

Baseado nas informações apresentadas, os SNPs podem ser classificados do seguinte

modo (ZOLET et al., 2017; MEKSEM; KAHL, 2006): (1) quanto à localização genômica, ou seja, se está presente em éxons, íntrons ou regiões intragênicas ou (2) quanto ao impacto, isto é, a atuação em regiões codificadoras, reguladores de determinado produto proteico ou mesmo na modificação de determinado fenótipo. A Figura 9 apresenta uma esquematização de classificação para SNPs quanto aos fatores apresentados.

É importante compreender que pesquisas que utilizam SNPs possuem duas etapas principais: a identificação e genotipagem, que são, respectivamente, responsáveis pela descoberta de polimorfismos em um conjunto de indivíduos, e a comparação desses polimorfismos com a população da determinada espécie (ZOLET et al., 2017). A Tabela 3 apresenta uma síntese dos principais métodos utilizados para essas tarefas.

Tabela 3 – Síntese dos métodos de identificação e genotipagem de SNP

Tipo de método	Nome do método	Referência
Identificação	PCR	(EDWARDS; JOHNSTONE; THOMPSON, 1991)
Identificação	eSNP	(PICOULT-NEWBERG et al., 1999)
Genotipagem	Hibridização de alelo	(HOWELL et al., 1999)
Genotipagem	Sonda TaqMan	(MCGUIGAN; RALSTON, 2002)
Genotipagem	SNP array	(PODDER et al., 2008)
Identificação e Genotipagem	Next Generation Sequence	(ORSOUW et al., 2007; BAIRD et al., 2008; TORKAMANEH; LAROCHE; BELZILE, 2016)

Embora a Tabela 3 apresente diversas métodos, a identificação de SNPs associados a determinados fenótipos de interesse era limitada pelo tempo e o alto custo vinculado a algumas das técnicas apresentadas. Contudo, com o surgimento das abordagens de NGS tornou-se possível a descoberta de centenas ou milhares de SNPs. Esse avanço também possibilitou a aplicação de metodologias de análise GWAS, onde grupos de indivíduos com características de interesse tem seu material genético sequenciado, permitindo a identificação de diversas variantes que são, posteriormente, comparadas a fim de estudar fenótipos destoantes em cada grupo (KORTE; FARLOW, 2013). A performance do GWAS permitiu que muitas características complexas presentes em indivíduos fossem estudadas e exploradas (ZOLET et al., 2017; TORKAMANEH; LAROCHE; BELZILE, 2016), uma vez que reduziu, significativamente os fatores limitantes de outros métodos. No entanto, este resultado culminou em outro problema, presente da área de mineração de dados, que é o desenvolvimento de metodologias de extração de conhecimento útil a partir de enormes conjuntos de dados (XU; JACKSON, 2019).

Apesar do GWAS ter se tornado amplamente utilizado em análises genéticas, os *pipelines* utilizados na análise dos dados resultantes ainda continuam em constante evolução (ADAM et al., 2021). Existem alguns *softwares* amplamente utilizados junto a esses *pipelines*. Contudo, são em geral utilizados em uma etapa inicial do GWAS, pois não são

eficazes em tarefas de filtragens de fenótipos quantitativos influenciados por muitos loci de menores efeitos, gerando subconjuntos finais com muitos SNPs para serem avaliados (NAVEED et al., 2018; GENTZBITTEL et al., 2019). A Tabela 4 apresenta uma revisão contendo alguns deles. Para cada *software* apresentado são exibidas 3 informações, a saber, nome da ferramenta, uma descrição da estratégia adotada pela ferramenta e a referência ao trabalho. É valido ressaltar que a Tabela 4 apresenta apenas soluções que trabalham com abordagens matemáticas e estatísticas, mas não aponta soluções que utilizam FS, pois esse tema é tratado nas seções seguintes.

Tabela 4 – Ferramentas utilizadas para seleção de SNPs

Nome	**Descrição**	**Referência**
EMMA	*Software* que utiliza combinação de Modelos Mistos (MLM) utilizando estratégias de divisão de amostras por população com controle de erros de tipo I e II para estudos de GWAS.	(YU et al., 2006)
PLINK	Pipeline que lida com gerenciamento de dados, estatísticas resumidas, estratificação populacional, análise de associação e estimativa de identidade por descendência.	(PURCELL et al., 2007)
TASSEL	Trabalha com Modelo Linear Geral (GLM) e Modelo Linear Misto para realizar mapeamento de associação.	(BRADBURY et al., 2007)
BLUP	Método para estimar efeitos aleatórios de um modelo misto. Este método foi originalmente desenvolvido no melhoramento animal para estimativa de valores genéticos, contudo atualmente é amplamente utilizado em diversas área.	(PIEPHO et al., 2008)
GBS PLAID	Pipeline que contém funções de análise de GWAS.	(SPINDEL et al., 2013)
GAPIT	*Software* que disponibiliza diversas implementações clássicas de modelos mistos e estratégias combinadas de desequilíbrio de ligação com abordagens bayesianas para aplicação em estudos de GWAS.	(WANG; ZHANG, 2021)

3.2 Estudos de seleção de SNPs

Avançando com exploração de trabalhos de melhoramento genético, esta seção apresenta alguns trabalhos que exploram a temática de seleção de SNPs em estudos de associação de fenótipos complexos. O conteúdo foi organizado em duas subseções. A primeira

subseção apresenta alguns estudos recentes de seleção de SNPs em arroz cujo enfoque não está em explorar as técnicas de FS, mas sim em explorar as metodologias utilizadas na literatura para seleção de SNPs em arroz. A segunda subseção, em complemento à primeira, apresenta alguns trabalhos que propuseram o uso de FS aplicado em outros contextos de seleção de marcadores gênicos. São apresentados pipelines completos para avaliação de SNPs associados a características complexas ou métodos que utilizam algoritmos de ML de maneira colaborativa para detecção de marcadores moleculares de interesse. É importante ressaltar que a segunda subseção não se restringe a estudos voltados ao arroz, mas explora os métodos utilizados em pipelines genéricos fornecendo uma visão ampla sobre as técnicas aplicadas.

3.2.1 Seleção de SNPs em arroz

A partir do que foi apresentado na seção anterior, entende-se que os SNPs são importantes marcadores moleculares (ZOLET et al., 2017), que devem ser explorados sobre o contexto de estudos que utilizam NGS, pois apontam características moleculares de alta granularidade, tais como variações genéticas que permitem a diferenciação de indivíduos de uma mesma espécie (BRAMMER, 2000). Por esse motivo, diversos estudos voltados a melhoramentos genéticos em arroz têm buscado encontrar SNPs que estão associados a determinados mecanismos que influenciam no rendimento e resistência a pragas específicas desse cultivar. A Tabela 5 apresenta uma revisão contendo alguns trabalhos que avaliaram genótipos associados a fenótipos do arroz, através de marcadores SNP. São exibidas quatro informações sobre cada trabalho, o nome principal ou abreviação atribuída ao trabalho, o número de amostras, ou seja, a quantidade de indivíduos utilizados como base ou validação, se o estudo utiliza abordagens de ML e a referência do estudo.

No trabalho de Yuan et al. (2020), 664 cultivares de arroz foram sequenciados e estudados a fim de obter SNPs que promovessem resistência ao sal. A identificação de SNPs foi feita pela combinação de abordagens baseadas em modelos mistos executados pelo *software* GAPIT. A execução dessa análise ocorreu levando em consideração as populações de arroz presentes no estudo, ou seja, *indica* e *japônica*. O GAPIT também foi executado com todas as amostras juntas. A etapa posterior contou com a determinação de genes a partir dos SNPs selecionados. Para essa etapa foi utilizada a ideia de desequilíbrio de ligação, implementada pelo *software* PLINK (PURCELL et al., 2007). A estratégia adotada mostrou dois pontos de atenção relevantes, a divisão populacional e a investigação de genes afetados.

Em (MORALES et al., 2020) foi realizado um estudo para seleção de SNPs em melhoramento de arroz. A validação foi realizada através do monitoramento da altura das plantas empregadas no estudo. O processo utilizado para avaliação de SNPs empregou os *softwares* TASSEL e GAPIT, apresentados na Seção 3.1. Apesar de se tratar de um

Tabela 5 – Estudos de seleção de SNPs para melhoramento de arroz

Nome	**Amostras**	**Utiliza ML**	**Referência**
Seleção de SNPs de resistência a sal	644	Não	(YUAN et al., 2020)
C7AIR – 7K SNP array	189	Não	(MORALES et al., 2020)
Análise de SNPs associados a características agronômicas	259	Não	(WANG et al., 2021)
Estudo de mecanismos genéticos da panícula de Arroz	406	Não	(ZHONG et al., 2021)
Estudo de Melhoramento de Arroz com 580.009 SNPs	417	Sim	(KIM et al., 2022)
Melhoramento Genético de Arroz através de 56.000 SNPs	192	Sim	(ZHANG et al., 2023)

estudo recente, observar-se que aqui outras técnicas de validação de SNPs selecionados não foram empregadas. Esse fato pode derivar da dificuldade na operacionalização das ferramentas de bioinformática existentes. Vale mencionar que no estudo de (WANG et al., 2021) alguns genes associados a 13 características agronômicas do arroz foram identificados utilizando a combinação dos *softwares* Best Linear Unbiased Prediction (BLUP) e EMMA.

Além dos estudos exploratórios voltados a investigação de associações entre fenótipos quantitativos, Zhong et al. (2021) demonstraram a utilização de SNPs para investigação de mecanismos vinculados ao rendimento de grãos, a partir de SNPs mapeados em genes que têm relação com panículas. Para identificação de SNPs, esse estudo combinou técnicas baseadas em modelos mistos, que apontaram alguns SNPs como tendo associação fenotípica. Assim como em outros trabalhos, não foi definido um critério padrão para a combinação dos modelos. As técnicas foram executadas seguindo padrões definidos pelos autores que posteriormente compilaram o resultado de modo manual. A partir disso alguns loci contendo SNPs de interesse foram mapeados para genes através da posição genômica ocupada pelo SNP, isto é, se estavam incidindo em genes ou promotores. O estudo aponta uma estratégia de mapeamento de genes incidentes já vista em (YUAN et al., 2020), evidenciando uma etapa comum, passível de generalização.

Um estudo com grande quantidade de SNPs foi realizado por Kim et al. (2022), onde 580.000 SNPs foram analisados utilizando outros estudos de base. O passo inicial da estratégia proposta consiste na utilização dos *softwares* TASSEL e GAPIT para uma análise de associação e filtragem inicial. Após isso, foi realizada a combinação do rrBLUP (ENDELMAN, 2011), derivado do BLUP (BAUER; REETZ; LÉON, 2006), que utiliza

a estratégia de penalização baseada em regularização, com redes neurais convolucionais (ALBAWI; MOHAMMED; AL-ZAWI, 2017). Nesse estudo, assim como os outros já mencionados, não foram estipulados critérios fixos para validação dos SNPs, mas uma análise manual apoiada por validação sequencial foi realizada.

No trabalho de Zhang et al. (2023) os SNPs foram avaliados utilizando combinação de nove modelos, entre eles variantes do BLUP (BAUER; REETZ; LÉON, 2006), e também regularização Least Absolute Shrinkage and Selection Operator (LASSO) e Ridge. Os modelos foram executados sobre dois fenótipos, vinculados às amostras. Para combinação das técnicas foi adotada uma avaliação manual a partir da acurácia dos modelos, onde a metodologia de validação cruzada foi utilizada para mensurar a acurácia. Embora algumas técnicas de FS, como as variantes de regularização, tenham sido utilizadas nesse estudo, não foi definido um critério robusto e generalizável para seleção de SNPs, sendo necessário uma tarefa de validação manual, anterior a etapa de validação dos SNPs.

Com bases nos trabalhos apresentados, é possível sintetizar três principais destaques em trabalhos recentes de seleção de SNPs de arroz. O primeiro ponto de atenção é que, apesar de se tratarem de estudos recentes, a ampla utilização de técnicas de FS ainda não está totalmente difundida, visto que grande parte os trabalhos se apoiam sobre a seleção e avaliação de SNPs contando unicamente com *softwares* clássicos baseados em modelos mistos de associação genotípica, indicando uma preferência por soluções de uso simples. Outro fator importante é que nos trabalhos de Kim et al. (2022) e Zhang et al. (2023), alguns algoritmos de FS foram aplicados, no entanto, critérios generalizáveis não foram sugeridos. Além disso, não é observado uma metodologia consensual para investigação de genes descobertos a partir dos SNPs. Portanto, observa-se que ainda há lacunas na tarefa de análise de dados resultantes de estudos de associação a partir de SNPs.

3.2.2 FS para seleção de SNPs em outros genomas

Nesta seção foi realizado um levantamento de trabalhos recentes que utilizaram algoritmos de ML na tarefa de FS aplicada a seleção de SNPs em outros contextos de melhoramento genético, isto é em outros genomas além do arroz. Apesar dos diversos trabalhos existentes, a definição do estado da arte é uma tarefa complexa, visto que os diferentes *pipelines* propõem alcançar objetivos distintos, tendo adaptações no conjunto de dados de entrada, métricas de eficácia e metodologias de validação diferentes. Apesar disso, pontos comuns e detalhes passíveis de melhoramento foram identificados.

Um grupo de algoritmos que vem sendo explorados são os Algoritmos Baseados em Relief (do inglês, "*Relief Based Algorithms*") (RBA)s, que surgiram como iniciativas de aprimoramento do algoritmo Relief (KIRA; RENDELL, 1992). O Relief não faz buscas exaustivas de atributos. Em vez disso, calcula a importância dos atributos levando em conta quão bem um atributo distingue instâncias que são de classes diferentes (por exem-

plo, caso base e controle) e que pertencem a mesma classe (URBANOWICZ et al., 2018a). Esse algoritmo foi inicialmente idealizado para lidar com problemas de classificação binária. Contudo, os RBAs rapidamente suplantaram a versão original desse algoritmo (URBANOWICZ et al., 2018a), propondo melhorias que permitiram lidar com problemas multi-classe e também de regressão. Entre os RBAs mais difundidos estão ReliefF (KONONENKO, 1994), SURF (GREENE et al., 2009), SURF* (GREENE et al., 2010), MultiSURF (URBANOWICZ et al., 2018b) e MuliSURF* (GRANIZO-MACKENZIE; MOORE, 2013). Em (URBANOWICZ et al., 2018b) os autores desenvolveram um *software* chamado ReBATE, que implementa os principais RBAs na linguagem Python. Os experimentos conduzidos mostraram que os RBAs tiveram performance melhor do que os métodos *Extra Trees* e RFE em conjunto ao *Extra Trees* (URBANOWICZ et al., 2018b). No entanto, assim como argumentado em (PUDJIHARTONO et al., 2022), apesar de suas vantagens, esses métodos possuem independência do classificador, isto é, selecionam conjuntos de atributos mais gerais, sem levar em consideração métricas de performance de classificadores.

Alzubi et al. (2017) propuseram uma método de seleção de atributos usando Conditional Mutual Information Maximization (CMIM) (SCHLITTGEN, 2011) e RFE-SVM para classificar pacientes saudáveis e doentes. Eles usaram conjuntos de dados SNPs para cinco classes (câncer de tireoide, autismo, câncer colorretal, deficiência cognitiva e câncer de mama). Os autores mostraram que para alguns casos, os SNPs selecionados pelo método CMIM em conjunto ao RFE-SVM produziram melhor desempenho de classificação do que usar separadamente CMIM (SCHLITTGEN, 2011) ou algoritmos como FCBF (YU; LIU, 2004), ReliefF (SPOLAÔR et al., 2013). Contudo, essa combinação não foi capaz de capturar interações entre SNPs, potencializando SNPs que tinham efeitos similares e estavam correlacionados.

O trabalho de Ramzan et al. (2020) propôs um pipeline para estudos GWAS que investigam SNPs associados a qualidade de ovos. A etapa inicial é uma filtragem dos SNPs utilizando o *software* PLINK. Após isso o algoritmo Boruta foi aplicado sobre o mesmo conjunto de dados para seleção de SNPs. Os SNPs apontados em ambos os pipelines foram os selecionados para etapa de investigação. A investigação de genes foi feita selecionando apenas os genes que tinham algum dos SNPs apontados pelo pipeline posicionados em sua extensão. Embora tenha sido apresentada uma estratégia que realiza a combinação de técnicas de ML com um *pipeline* clássico, abordagens de investigação de genes selecionados não foram propostas. Observa-se também a ausência de uma métrica robusta que integre mais algoritmos, a fim de promover outras bases de comparação e validação.

Outro modo interessante de identificar características complexas associadas aos SNPs é a integração de outros dados vinculados às amostras. No estudo realizado por Cueto-López et al. (2019), dados clínicos das amostras foram integrados, a fim de auxiliar na

predição de câncer colorretal, fornecendo explicações e indicativos, além dos polimorfismos. Neste estudo foram utilizados os algoritmos RF, RFE somada ao SVM com algoritmo interno de avaliação e regressão logística, para selecionar os atributos com maior relevância. Nesse estudo, os SNPs e dados clínicos são modelados como atributos. As estratégias de pontuação de importância dos atributos foram calculadas pelos algoritmos de maneira individual. Porém, os atributos foram selecionados pelo algoritmo que apresentou melhor performance, nesse caso o SVM e regressão logística. Observou-se que outros fatores além dos SNPs que foram apontados como causadores dos câncer, no entanto, não foi realizada uma investigação sobre a associação entre os atributos selecionados. Apesar disso, essa estrategia permitiu que outras informações fossem agregadas ao estudo, apontando possíveis explicações para os polimorfismos encontrados.

Em (SEO et al., 2021) foi aplicada uma estratégia com *softwares* de filtragem dos dados que foi sucedida de uma abordagem de refinamentos dos SNPs a partir dos algoritmos árvore de decisão, RF e adaboost. Essa abordagem explorou a combinação de técnicas de ML para filtragem dos SNPs. Por sua vez, o trabalho de (LI et al., 2022) realizou uma comparação entre o uso de técnicas de regularização LASSO e redes elásticas com variantes delas combinadas com modelos lineares. O conjunto de dados utilizado explorou 7957 SNPs, buscando melhorar 3 fenótipos ligados ao leite. Foi observado que a abordagem LASSO obteve resultados comparáveis a estratégia combinada para dois dos três fenótipos. Apenas em um dos fenótipos o resultado da combinação direta de modelos lineares apresentou um resultado melhor que os demais algoritmos. Isso indica que em determinados contextos a utilização das abordagens isoladas pode apresentar efetividade maior, do que em combinações diretas com modelos lineares mistos.

(RASCHIA et al., 2022) apresenta um estratégia que combina XGBoost, RF e LigthGBM para seleção de SNPs associados a qualidade do leite. Um total de 40.417 SNPs foram explorados. A estratégia de seleção de SNPs consistiu em usar as métricas internas de importância de atributos de cada algoritmo. Posterior a essa etapa, a filtragem dos SNPs foi realizada preservando apenas os SNPs, consensualmente, selecionados em todas as técnicas. Uma etapa de análise dos SNPs foi realizada utilizando a ferramenta PHANTER (THOMAS et al., 2003) para identificar genes codificadores de proteínas.

No trabalho de Lim et al. (2023) foi proposto um método que utiliza RFE empregando RF como algoritmo base, somado a 5 algoritmos de ML, que foram treinados e utilizados para filtragem do número de SNPs. O conjunto de dados utilizado contou com cerca de 70.000 SNPs investigados, a fim de prever polimorfismos capazes de identificar artrite e reumatoide. A seleção de SNPs foi modelada como um problema de FS que selecionou 13 SNPs importantes. Na etapa seguinte, 5 algoritmos foram combinados, a saber SVM, RF, Regressão Logística, Naive Bayes e XGBoost. Todos foram treinados utilizando os SNPs selecionados pela RF, na etapa anterior. A estratégia de refinamento dos SNPs a partir desses algoritmos consistiu em selecionar um número de SNPs que mantivesse a

curva ROC-AUC dos algoritmos implementados acima de 0.9. Essa estratégia permitiu a filtragem de 9 SNPs de interesse. Notavelmente, esse *pipeline* emprega uma técnica robusta de seleção de SNPs, no entanto, não foi empregado aqui uma estratégia de avaliação dos SNPs que leve em conta fatores biológicos e interação gênica.

Com base nos trabalhos apresentados, é possível observar que apesar dos algoritmos ML não estarem amplamente difundidos em estudos de melhoramento genético de arroz, bons resultados tem sido obtidos com sua utilização em estudos correlatos. Esse fato aponta para a necessidade de uma investigação e padronização do uso de algoritmos de FS em conjuntos de dados de arroz, possibilitando ampla aderência em estudos de associação fenotípica em novos estudos com essa temática.

Capítulo 4

Enriquecimento Funcional

É comum ao final de um experimento de investigação de algum sistema biológico envolvendo análises proteômicas, genéticas ou metabólicas, a exploração da relevância e atuação de SNPs, transcritos ou genes estudados em outros sistemas biológicos já conhecidos (HUNG et al., 2011). Para tal podemos verificar a presença de determinado elemento, contido no experimento, em bancos de dados de vias biológicas, os quais contêm informações referentes à interação entre elementos pertencentes a algum sistema biológico (DRAGO et al., 2015). Desse modo é possível buscar indicações e apontamentos sobre possíveis sistemas biológicos que estão sendo afetados, e por conseguinte, estão provocando determinada situação observada no estudo.

No entanto, esse tipo de abordagem tem uma eficácia baixa, uma vez que determinadas estruturas funcionais, tais como genes, podem aparecer em diversas vias e processos biológicos de maneiras distintas e podem até mesmo ser reguladas em conjunto a outros genes, caso os mesmos estejam mais ou menos expressos em determinado tecido ou contexto biológico (CARULLI et al., 1998; LIANG; PARDEE, 2003). Sendo assim, torna-se necessária a aplicação de uma abordagem robusta para atribuição de funções biológicas a determinadas estruturas funcionais, de modo a estabelecer não somente uma atribuição direta, mas também uma pontuação baseada em evidências estatísticas demonstrando que determinado conjunto de estruturas funcionais apresentam mais relação com um sistema biológico específico.

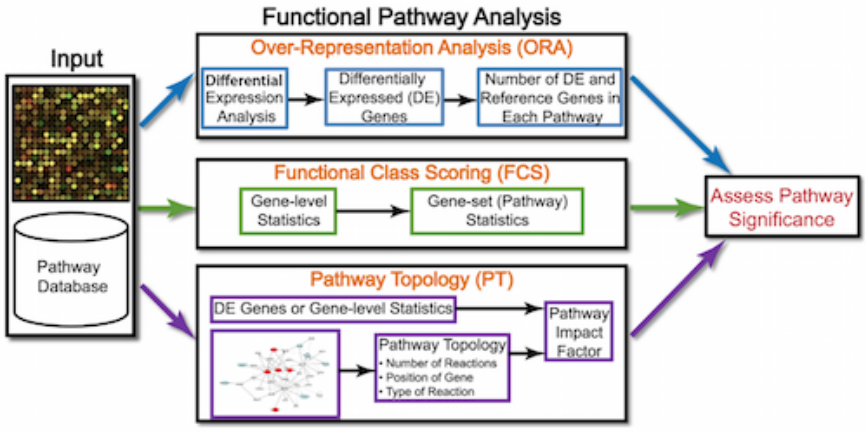
Nesse contexto, é aplicada uma metodologia de análise denominada enriquecimento funcional, que verifica se determinado grupo de estruturas biológicas, frequentemente genes, estão super-representados em determinados rótulos biologicamente relevantes (WANG, 2013; WIJESOORIYA et al., 2021). O termo enriquecimento é definido como o grau de relevância de determinada via ou família de genes para determinado grupo de ge-

nes estudados, levando em conta algum cálculo estatístico (WIJESOORIYA et al., 2021). Sendo assim, uma via dita mais enriquecida para um conjunto de genes é uma via que, estatisticamente, apresenta maior relevância para o conjunto de genes estudado. Existem diversas plataformas de anotações funcionais disponíveis (RAUDVERE et al., 2019), tal como ferramentas e metodologias distintas que implementam conceitos pertinentes a essa análise. O restante desse capítulo está organizado em três seções que tratam, respectivamente, das abordagens de enriquecimento, os bancos de dados frequentemente utilizados para os cálculos e algumas ferramentas que implementam essa metodologia.

4.1 Abordagens de Enriquecimento

Frequentemente, quando o termo análise funcional é utilizado, há uma associação indiscriminada dessa abordagem ao termo Análise de Sobre-representação (do inglês, "*Over Representation Analysis*") (ORA) executado sobre termos funcionais anotados. No entanto é válido mencionar que a Pontuação de Classe Funcional (do inglês, "*Functional Class Score*") (FCS) é distinta de uma análise ORA, pois a primeira leva em consideração um valor estatístico referente a conjuntos de genes isoladamente referenciados (FERNANDES; HUSI, 2021), enquanto que FCS utiliza estatísticas do nível de genes, para determinar interações sutis entre os genes, tal como é implementado uma Análise de Enriquecimento de Conjuntos de Genes (do inglês, "*Gene Set Enrichment Analysis*") (GSEA) (ACKERMANN; STRIMMER, 2009; KHATRI; SIROTA; BUTTE, 2012). Outra análise existente no contexto do enriquecimento em conjuntos de genes e vias biológicas é a análise de topologia, que utiliza informações relativas às interações entre os genes para determinar as relações mais relevantes (KHATRI; SIROTA; BUTTE, 2012). A Figura 10 apresenta uma síntese do que foi explicado.

Figura 10 – Tipos de abordagens de Enriquecimento Funcional



Fonte: Extraído de (KHATRI; SIROTA; BUTTE, 2012)

A subseção seguinte percorre os conceitos referentes às análises ORA. Esse trabalho não discute os conceitos vinculados à GSEA e metodologias baseadas em topologia, pois seu escopo excede as intenções e aplicações exploradas aqui.

4.1.1 Análise de Sobre-Representação

A ORA é uma abordagem utilizada no contexto de conjuntos de estruturas funcionais, que busca vias ou termos funcionalmente enriquecidos desprezando informações como expressão e força das interações (KHATRI; SIROTA; BUTTE, 2012; JIN et al., 2014). A ORA comumente utiliza os bancos de dados *Gene Ontology* (GO) (MI et al., 2018) e banco de dados da Enciclopédia de Genes e Genomas de Kyoto (do inglês, "*Kyoto Encyclopedia of Genes and Genomes*") (KEGG) (KANEHISA et al., 2015). Esses bancos de dados serão melhor explorados na Seção 4.2.

A análise ORA baseia-se no cálculo da probabilidade de que o número de genes observados em determinada via não foi estabelecido ao acaso. Para tal, o valor de p é computado utilizando a distribuição binomial negativa, assim como apresentado na Equação 8. Nesse contexto, N representa o número total de genes, denominado contexto de fundo ou universo de possibilidades. Aqui é possível utilizar o número total de genes existente para determinado organismo ou mesmo um fundo contendo os genes disponíveis em um sequenciamento de alto rendimento (MACARRON et al., 2011). O valor que substituirá M representa o número de genes que possuem alguma anotação funcional no banco de dados utilizado, seja direta ou indiretamente vinculada ao grupo gênico avaliado. Por fim, temos os valores de n e k, que significam, respectivamente, o número de genes de interesse e a quantidade dentre esses genes que estão atuando juntos em um grupo funcional ou via. Os valores de p obtidos devem ser ajustados para comparações múltiplas.

$$p = 1 - \frac{\sum_{i=0}^{k-1} \binom{M}{i} \binom{N-M}{n-i}}{\binom{N}{n}} \tag{8}$$

A interpretação estatística do cálculo presente na Equação 8 pode ser exemplificada. Considere a probabilidade de 9 genes (k) serem associados a uma via de reparo de DNA que possui 26 genes atuantes anotados (M), para uma lista resultante de 1470 genes (n), obtidos em um estudo aplicado ao genoma de uma espécie que contém ao todo 14.531 genes (N). Nesse caso o valor de p encontrado seria aproximadamente 0.0006, isto é, a via estaria enriquecida para os genes encontrados no estudo.

4.2 Bancos de dados de sistemas biológicos

As abordagens de enriquecimento funcional necessitam de anotações de grupos gênicos, isto é, grupos de genes que trabalham juntos para execução de alguma função biológica

ou são responsáveis por algum mecanismos do sistema biológico dos indivíduos de uma espécie. Sendo assim, é necessário utilizar bancos de dados com tais informações, capazes de fornecer os parâmetros necessários para a aplicação dessa abordagem. As subseções seguintes descrevem os dois principais bancos de dados utilizados em trabalhos recentes, a saber GO e KEGG.

4.2.1 Gene Ontology

O GO é um banco de dados de ontologias gênicas, isto é, representações formais de um corpo de conhecimento dentro de um determinado domínio (MI et al., 2018; PLESSIS; ŠKUNCA; DESSIMOZ, 2011), nesse contexto, o domínio biológico. O GO define o conhecimento do domínio biológico em três classes de aspectos, a saber (LOMAX, 2005; PLESSIS; ŠKUNCA; DESSIMOZ, 2011):

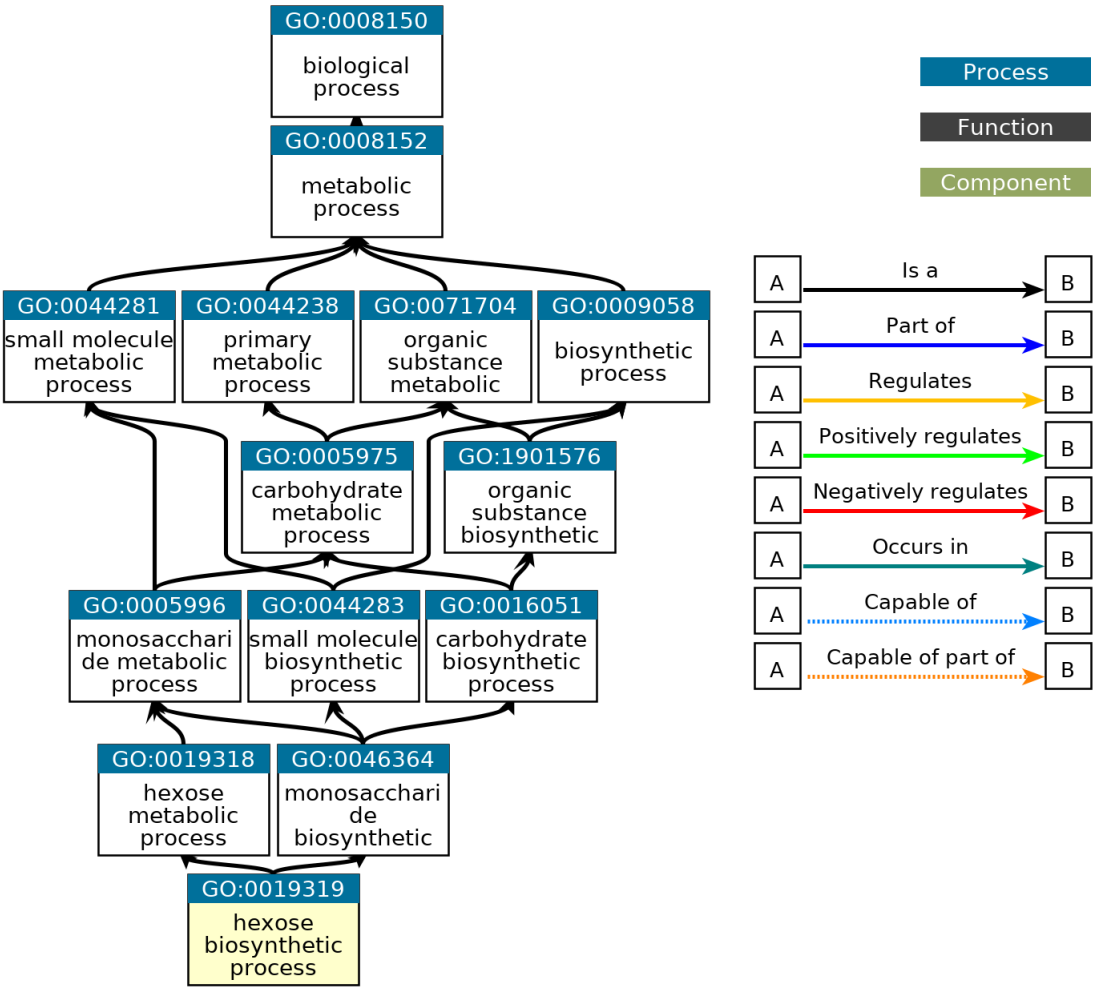
- Função Molecular: Atividades de nível molecular realizadas por produtos gênicos. Os termos de função molecular descrevem atividades que ocorrem no nível molecular, como “catálise” ou “transporte”. Os termos de função molecular GO representam atividades em vez das entidades (moléculas ou complexos) que realizam as ações e não especificam onde, quando ou em que contexto a ação ocorre;

- Componente Celular: As localizações relativas às estruturas celulares nas quais um produto gênico desempenha uma função, sejam compartimentos celulares (por exemplo, mitocôndria) ou complexos macromoleculares estáveis dos quais são partes (por exemplo, o ribossomo). Ao contrário dos outros aspectos da GO, as classes de componentes celulares não se referem a processos, mas sim a uma anatomia celular;

- Processo Biológico: São processos realizados por múltiplas atividades moleculares. Exemplos de termos pertencentes a essa classe são processos biológicos de reparo de DNA ou transdução de sinal. Exemplos de termos mais específicos são processos biossintéticos do núcleo. Observe que um processo biológico não é equivalente a uma via. Atualmente, a GO não tenta representar as dinâmicas ou dependências que seriam necessárias para descrever completamente uma via.

A estrutura do GO pode ser representada por um Grafo Acíclico Dirigido (do inglês, "*Directed Acyclic Graph*") (DAG), onde os nós são os termos ontológicos e as arestas representam as relações hierárquicas ou de cooperação entre os termos biológicos (MI et al., 2018; PLESSIS; ŠKUNCA; DESSIMOZ, 2011). Além disso, é válido informar que os termos ontológicos podem apresentar dois ou mais pais diretamente vinculados, bem como cooperação e participação. A Figura 11 representa as ideias expressas.

Figura 11 – Ontologia Gênica



4.2.2 KEGG

O KEGG é um banco de dados que contém informações de alto nível a respeito de sistemas biológicos (KANEHISA et al., 2015). Atualmente, o KEGG contém um total de 16 bancos de dados integrados amplamente categorizados com informações sistêmicas, químicas, genômicas e de saúde (KANEHISA et al., 2015). A Tabela 6 apresenta os bancos de dados integrados ao KEGG, bem como suas especificações.

4.3 Ferramentas para enriquecimento funcional

Com a ampla utilização do enriquecimento funcional em estudos biológicos, alguns *softwares* têm sido desenvolvidos com objetivo de simplificar a implementação de estratégias de ORA (CHICCO; JURMAN, 2022). Contudo, nem todas as ferramentas desenvolvidas possuem uma integração direta disponível com bancos de dados de conjuntos de genes e sistemas biológicos voltados ao arroz, gerando uma tarefa extra de aquisição e integração de dados, que pode dificultar a utilização desses *softwares* em estudos voltados

Tabela 6 – Bancos de dados integrados ao KEGG

Categoria	**Nome do Recurso**	**Conteúdo**
Sistêmica	KEGG PATHWAY	Mapas de vias
Sistêmica	KEGG BRITE	Tabela e hierarquias
Sistêmica	KEGG MODULE KEGG RModule	Modulações e reações disponíveis
Genômica	KEGG ORTHOLOGY KEGG Annotation	Anotações funcionais de estruturas que referenciam genes ortológos
Genômica	KEGG GENES KEGG SeqData	Genes e produtos gênicos, tal como proteínas
Genômica	KEGG GENOME KEGG Virus	Genomas de organismos celulares e vírus
Química	KEGG COMPOUND	Metabólitos e outras moléculas pequenas
Química	KEGG GLYCAN	Glicanos
Química	KEGG REACTION	Reações bioquímicas e classes de diferentes reações
Química	KEGG Enzyme	Nomenclatura enzimática com dados de sequência
Saúde	KEGG NETWORK	Variações de rede relacionadas a doenças e variantes de genoma
Saúde	KEGG DISEASE	Informações referentes a doenças
Saúde	KEGG DRUG	Mapeamento de medicamentos (ATC), Medicamentos (alvo) e anti-infecciosos

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

a arroz. Além disso, a forma de implementação dificulta a adoção de algumas dessas ferramentas pelos pesquisadores. A fim de investigar ferramentas passíveis de adoção e integração com os dados desse estudo, apresentamos um levantamento de alguns *softwares* utilizados na tarefa de enriquecimento funcional. A Tabela 7 apresenta quatro características avaliadas em cada um deles, a saber, i) se a ferramenta possui bancos de dados voltados ao arroz, ii) se possibilita integrações com bancos de dados externos, iii) se possui uma versão web, e iv) se possui integração via Interface de Programação de Aplicação (do inglês, "*Application Program Interface*") (API). Vale ressaltar que essa última característica foi investigada com o intuito de simplificar o processo de composição do *pipeline* proposto neste trabalho, pois a automação do processo de verificação de sistemas bioló-

gicos, a partir dos SNPs selecionados, é necessária a fim de evitar as tarefas manuais de copiar os identificadores SNPs apontados pelo pipeline, convertê-los em genes e aplicar o enriquecimento funcional.

Tabela 7 – Ferramentas para enriquecimento funcional

Nome	Banco de Dados de Arroz Nativo	Possibilita Integrações	Possui Versão Web	Possui API	Referência
GSEApy	Não	Sim	Sim	Sim	(FANG; LIU; PELTZ, 2023)
PLant GSAD	Sim	Não	Sim	Não	(MA et al., 2022)
Orsum	Não	Sim	Não	Sim	(OZISIK; TÉRÉZOL; BAUDOT, 2022)
GProfiler	Sim	Sim	Sim	Sim	(RAUDVERE et al., 2019)
GOATOOLS	Não	Sim	Não	Sim	(KLOPFENSTEIN et al., 2018)
Enrichr	Não	Não	Sim	Sim	(KULESHOV et al., 2016)
gsea-MSigDB	Não	Sim	Sim	Não	(LIBERZON et al., 2015)
DAVID Tools	Não	Sim	Sim	Sim	(SHERMAN et al., 2022)

A partir das informações apresentadas na Tabela 7, conclui-se que o GProfiler possui todas as características levantadas como importantes. Devido a isso optamos por utilizá-lo como ferramenta para exploração de sistemas biológicos nesse estudo. Por conseguinte, apresentamos uma visão geral sobre suas características de uso.

4.3.1 GProfiler

A ferramenta GProfiler é uma das principais ferramentas utilizadas em análises de enriquecimento funcional (REIMAND et al., 2019). Ela realiza cálculos de ORA internamente valendo-se de diversos bancos de dados, os quais estão disponíveis para um contingente de 500 organismos, isto é, espécies biológicas (RAUDVERE et al., 2019). Isso a torna útil no contexto desse trabalho, que lida com o genoma da planta de arroz, que possui informações anotadas menos abrangentes em relação a outros organismos, tal como *Homo sapiens*. Essa ferramenta é muito versátil, uma vez que pode ser usada tanto em ambiente web (RAUDVERE et al., 2019), como através de uma API, permitindo a conexão através de diversas linguagens de programação. Em sua versão 2, essa ferramenta apresenta cinco funções que podem ser acessadas tanto pela plataforma web quanto pela API (KANEHISA et al., 2015), a saber:

- g:Gost - realiza análise ORA ou análise de enriquecimento de conjunto de genes na lista de genes de entrada. Além dos resultados tabulares detalhados, permite a criação de gráficos para exibição e interpretação dos resultados;
- g:Convert - permite a conversão entre vários genes, proteínas, sondas de microarray e vários outros tipos de espaços de nomes;
- g:Orth - traduz identificadores de genes entre organismos;

- g:SNPense - mapeia códigos que têm o prefixo *rs* (*Reference SNP*) para os SNPs presentes em genes, coordenadas cromossômicas e efeitos variantes;
- g:Random - busca uma lista (não realmente aleatória) de identificadores de genes, útil para fins de teste.

Não obstante, um uso interessante dessa ferramenta foi feito no trabalho de Chen et al. (2022), onde ela foi utilizada para investigar melhoramento de grãos de arroz e milho através da identificação de genes ortólogos entre as espécies. Para essa tarefa de conversão de genes foi utilizada a função *g:Orth*. Em seguida, o enriquecimento funcional foi realizado a partir da função *g:GOSt*, de modo a compreender alguns sistemas biológicos afetados por outros genes de interesse encontrados no estudo. Outrossim, nos trabalhos de Souza et al. (2023) e Thorburn et al. (2023), essa ferramenta foi integrada no *pipeline* de análise para a tarefa de exploração de sistemas biológicos.

Capítulo 5

Proposta

Como introduzido no Capítulo 1, este trabalho tem como objetivo propor um *pipeline* generalizável para seleção de SNPs de arroz associados a múltiplos fenótipos. O *pipeline* proposto nesse estudo foi idealizado levando em conta três passos:

- **Pré-Processamento**: etapa idealizada para uniformizar os dados de entrada, permitindo uso genérico para diversos conjuntos de dados de arroz;
- **Seleção de SNPs Associados a Fenótipos**: procedimento responsável pela seleção e filtragem dos SNPs mais importantes para determinado fenótipo;
- **Análise Exploratória de SNPs Selecionados**: procedimento de verificação e confirmação dos SNPs selecionados, através de confirmação em literatura, possíveis genes afetados e sistemas biológicos.

Para consolidar as melhores abordagens a serem utilizadas em cada etapa, alguns experimentos foram realizados. Esse capítulo irá delinear o procedimento de execução de cada um deles. Na Seção 5.1, é apresentado o conjunto de dados utilizado para realização dos testes, exploração dos resultados e recursos computacionais utilizados. Na Seção 5.2, é apresentada uma etapa de pré-processamento sugerida como passo de normalização para os dados. Em seguida, são descritas na Seção 5.3, quatro estratégias propostas para seleção de SNPs. Após isso, na Seção 5.4 é descrita a metodologia de avaliação adotada para verificação de performance dos métodos propostos. Por fim, na Seção 5.5, é proposta uma metodologia de exploração de conjuntos de genes e sistemas biológicos, a partir dos SNPs selecionados.

5.1 Conjunto de dados

Para esse trabalho foram utilizados dados genotípicos de 541 amostras de arroz e seis características fenotípicas relacionadas a essas amostras, que pertencem à Coleção Nuclear de Arroz da Embrapa Arroz e Feijão (ABADIE et al., 2005). Esses dados foram cedidos a fim de que fossem realizadas explorações de técnicas para a identificação de SNPs que propiciem o melhoramento do cultivar de arroz. Os dados são compostos por oito experimentos de caracterização da produtividade das amostras de arroz, denominados ensaios, localizados em seis unidades federativas: Goiás, Roraima, Rio Grande do Sul, Mato Grosso, Piauí e Rondônia. As cidades brasileiras onde os experimentos foram realizados são Boa Vista, Santo Antônio de Goiás, Goiânia, Vilhena, Teresina, Sinop, Uruguaiana e Pelotas. Os oito ensaios foram divididos em dois sistemas de cultivo, o irrigado e o sequeiro. Os experimentos foram realizados em três anos agrícolas (2004/2005, 2005/2006 e 2006/2007, denominados 2004, 2005 e 2006, respectivamente). A Tabela 8 mostra o local onde cada um desses experimentos foi realizado. São exibidas sete informações a respeito de cada um desses cultivos, o município onde foi realizado, o estado, ano agrícola de cada cultivo, sistema de cultivo, latitude, longitude e a altitude, que foi mensurada em metros.

Tabela 8 – Informações do conjunto de dados utilizado

Localidade	**Estado**	**Ano**	**Sistema**	**Latitude**	**Longitude**	**Altitude**
Goiânia	GO	2004/2005	Sequeiro	16° 28’ S	49° 17’ W	779
Sinop	MT	2005/2006	Sequeiro	11° 51’ S	53° 30’ W	345
Teresina	PI	2006/2007	Sequeiro	5° 05’ S	42° 48’ W	72
Vilhena	RO	2006/2007	Sequeiro	12° 47’ S	60° 05’ W	600
Santo Antônio do Goiás	GO	2004/2005	Irrigado	16°26’ S	49°23’ W	728
Boa Vista	RR	2004/2005	Irrigado	2° 48’ N	60° 39’ W	61
Uruguaiana	RS	2004/2005	Irrigado	29° 45’ S	57° 05’ W	74
Pelotas	RS	2005/2006	Irrigado	31° 52’ S	52° 21’ W	13

A informação genotípica, isto é, o DNA das 541 amostras, foi extraído usando o DNeasy 96 Plant Kit (QIAGEN) seguindo as instruções do fabricante (HANDBOOK, 2005). Para cada amostra foi utilizada uma planta individual. Os marcadores SNP foram obtidos pela metodologia Genotipagem por Sequenciamento (do inglês, "*Genotyping by Sequencing*") (GBS) (ELSHIRE et al., 2011) no Genomic Diversity Institute da Universidade de Cornell (EUA). A geração de dados foi realizada em uma Plataforma Genome Analyzer II (Illumina, Inc., San Diego, CA), e o sequenciamento foi do tipo single-end (VOELKERDING; DAMES; DURTSCHI, 2009). A filtragem dos dados brutos de sequenciamento, o alinhamento da sequência no genoma de referência do arroz Os-Nipponbare-Reference-IRGSP-1.0 (KAWAHARA et al., 2013), e o SNP calling a partir da genotipagem GBS

de baixa cobertura, foram realizadas usando o pipeline TASSEL GBS v5.0 (BRADBURY et al., 2007) fornecido pelo Buckler Lab for Maize Genetics and Diversity. As tarefas de filtragem dos dados brutos foram baseadas em estimativas anteriores de desequilíbrio de ligação e taxas de endogamia obtidas para o arroz. Após ancorar as sequências no genoma de referência, os SNPs foram identificados em cada amostra com uma frequência alélica menor (MAF) definida como 0,01. O valor do coeficiente de endogamia foi igual a 0,9, com mínima cobertura de locus igual a 0,1, que por sua vez corresponde à proporção de acessos com pelo menos uma etiqueta em um locus.

Os fenótipos utilizados nesse trabalho como atributos de classe do conjunto de dados de Arroz fornecido pela Embrapa são produtividade, floração, acamamento, altura, relação comprimento largura e percentual de grãos inteiros. A produtividade foi estimada pela pesagem dos grãos colhidos na área útil da parcela, convertida em kg por hectare. A predição dos valores genéticos de cada indivíduo foi realizada usando o procedimento BLUP. Os demais fenótipos avaliados foram dias até o florescimento (número de dias do plantio até 50% das panículas com as anteras expostas), altura (média em cm da superfície até a inserção da panícula do perfilho principal), percentagem de grãos inteiros, acamamento e relação comprimento x largura do grão. O processo de mensuração detalhado de cada fenótipo está descrito no Apêndice A. A Tabela 9 descreve cada característica fenotípica, apresentando abreviação e o nome do fenótipo, descrição e unidade em que o mesmo foi mensurado.

Figura 12 – Distribuição de SNPs por cromossomo

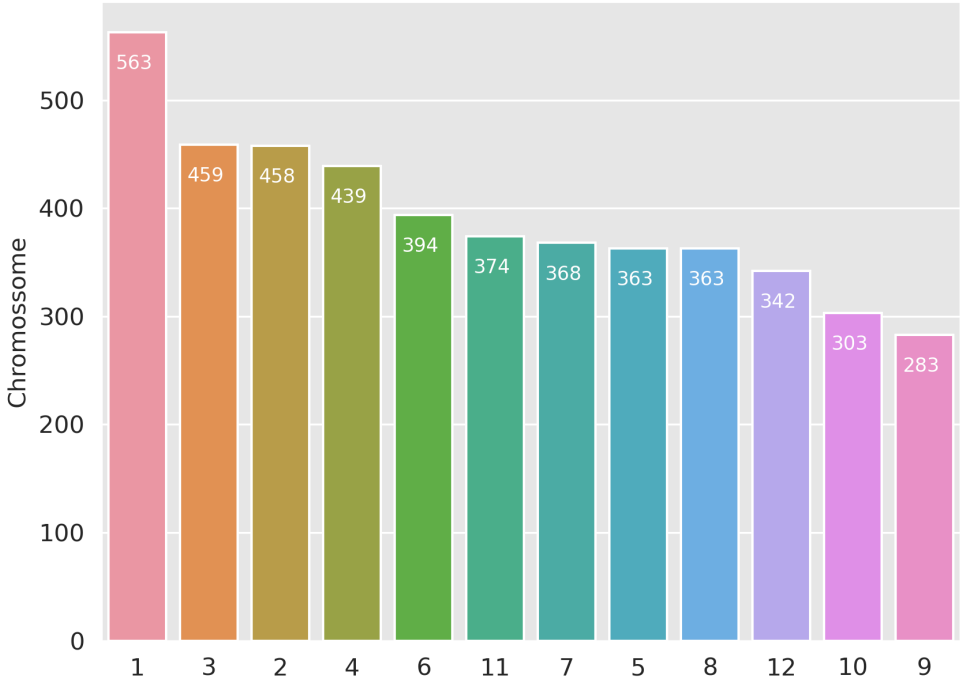


Tabela 9 – Características fenotípicas mapeadas para o arroz

Abreviação	**Nome**	**Descrição**	**Unidade**
Prod	Produtividade	A medida de produtividade foi estimada pelo peso total dos grãos, levando em consideração todos os grãos que foram destacados em cada planta de arroz.	Miligramas (mg)
Flo	Floração	Número de dias até a floração.	Dias
Aca	Acamamento	Este atributo é constituído pela queda ou arqueamento das plantas devido à flexão do caule ou à má ancoragem proporcionada pelas raízes. A nota 9 (nove) atribuiu-se a parcela onde todas as plantas permaneceram eretas até a ocasião de colheita e a nota 1 refere-se a uma parcela onde todas as plantas presentes sofreram acamamento. As notas intermediárias (2 a 8) distribuem-se entre as proporções relativas para os intervalos que representam o percentual de plantas da parcela que sofreram acamamento.	Inteiro de 0 a 9
Int	Porcentagem de Grãos Inteiros	Indica o percentual do número de grãos que no final do experimento ainda permaneceram consistentes. A escala indica melhores resultados para valores maiores.	Valor Percentual
C/L	Relação Comprimento/Largura do Grão	É um valor numérico obtido pela divisão do comprimento do grão pela sua largura.	Valor Contínuo Não Negativo
Alt	Altura	Tamanho do grão coletado, medido em centímetros.	Centímetros (cm)

O conjunto de dados contendo genótipo e fenótipos obtido para os experimentos foi previamente analisado por meio de modelos lineares mistos, e estimativas de componentes de variância foram obtidas pelo método Máxima Verossimilhança Residual (REML), com aplicação do procedimento BLUP para a estimativa dos valores genéticos dos efeitos aleatórios associados a cada uma das amostras (BUENO et al., 2012). Esses dados foram reunidos em arquivos de Valores Separados por Vírgula (do inglês, "*Comma-separated Values*") (CSV), junto com os dados de marcadores SNPs obtidos pela metodologia de GBS. O número de marcadores SNPs obtido a partir desse procedimento foi 445.589 SNPs (PANTALIAO et al., 2016), de onde foram selecionados 4.709 SNPs espaçados a cada 100 a 200 Kpb, número suficiente para uma espécie autógama, com elevado desequilíbrio de ligação (NORTON et al., 2018). A Figura 12 mostra o número de SNPs selecionados de cada cromossomo. Após essas etapas, os 4.709 SNPs obtidos juntamente com as seis características fenotipicas foram organizados em um arquivo CSV, onde cada linha representa uma amostra que foi explorada. As 4.709 primeiras colunas representam os SNPs observados para cada amostra e as seis colunas seguintes representam as características fenotipicas. O link de acesso a esse arquivo formatado está no Apêndice B.

Tabela 10 – Análise Inicial dos Dados

	Prod	**Flo**	**Aca**	**Int_%**	**C/L**	**Alt**
Média	3471.76	91.81	1.83	50.95	1.09	109.35
Desvio Padrão	205.12	0.99	0.28	2.68	0.69	1.61
Mínimo	2660.93	86.66	1.18	40.84	0.20	98.99
25%	3471.44	92.18	1.65	49.42	0.71	109.02
50%	3511.64	92.24	1.71	51.21	1.00	109.73
75%	3544.57	92.27	1.98	52.84	1.17	110.18
Máximo	4524.20	94.67	3.05	57.58	7.00	127.51

A Tabela 10 exibe uma análise descritiva das características fenotípicas apresentadas. São exebidas a média observada nos valores mensurados em cada características fenotípica, o desvio padrão dos valores observados, os valores máximos e mínimos de cada característica fenotípica e os percentis referentes a 25%, 50% e 75%. Essa informações são importantes para que após mensuradas métricas de erro seja possível realizar uma comparação. Além disso, essa informações oferecem a escala de valores presente de cada característica fenotípica.

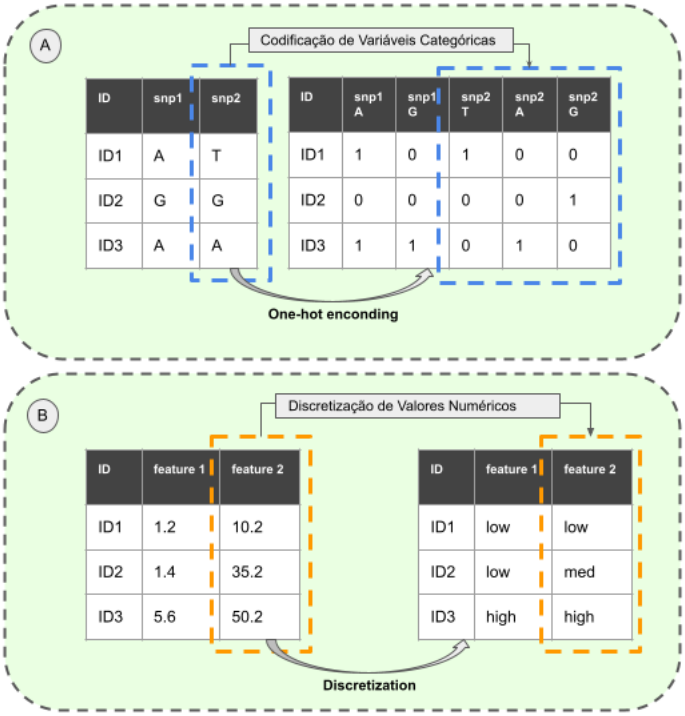
5.2 Pré-Processamento

Com o intuito de propor um *pipeline* que trate diferentes conjuntos de dados de maneira uniforme, foi proposta a aplicação da *codificação de atributos categóricos*, isto é, a

geração de uma representação numérica para atributos de entrada compostos por valores categóricos como letras ou classes (HANCOCK; KHOSHGOFTAAR, 2020; LI, 2019). Essa etapa é necessária, pois os conjuntos de dados contendo informações de SNPs, frequentemente estão organizados em arquivos tabulares, onde as colunas representam a identificação dos SNPs e linhas referenciam uma das amostras sequenciadas, de modo que o valor determinado por uma amostra i em relação a um SNP j é denotado por uma das quatro letras que representam as bases nitrogenadas, sendo que algumas bibliotecas como scikit-learn, utilizada nesse estudo, possuem a implementação de algoritmos de aprendizado dependentes de atributos de natureza numérica.

Para o *pipeline* proposto, a codificação foi feita com o módulo *preprocessing.OneHotEncoder* da biblioteca scikit-learn, que implementa o *One-Hot Enconding*, que consiste na transformação de um atributo categórico que possui cardinalidade n, isto é, possui n valores distintos, em n atributos, transformando cada valor distinto pertencente ao atributo em uma dimensão no conjunto de dados. Se um atributo do conjunto de dados que equivale ao nucleotídeo observado em uma determinada posição do genoma teve variações A, T e G, ao aplicar o *One-Hot Enconding*, teremos três atributos, onde o primeiro representa em quais exemplos do conjunto de dados observou-se o nucleotídeo A e os outros dois teriam efeito análogo para nucleotídios T e G. A parte A da Figura 13 apresenta um exemplo da aplicação do *One-Hot Enconding*. Outrossim, alguns trabalhos como os de (NASEER; SALEEM, 2018) e (SEGER, 2018) obtiveram bons resultados com a utilização dessa abordagem.

Figura 13 – Pré-processamento do conjunto de dados



A escolha dessa abordagem em detrimento de outra técnica de codificação de atributos se deu após a investigação da cardinalidade dos atributos presentes no conjunto de dados utilizado, pois observou-se que todos possuíam apenas duas variações de nucleótidos. Esse fato permitiu a utilização do módulo *preprocessing.OneHotEncoder* com uma adaptação do parâmetro *drop*, que possibilita a eliminação de um dos atributos gerados, de modo que a representação de um atributo binário se dá por apenas um dos atributos gerados, de modo que os valores 0 equivalem a ausência de uma das bases nitrogenada e presença da outra, enquanto que o valor 1 tem o sentido inverso. Porém, vale ressaltar que essa abordagem pode ser substituída por outra técnica de codificação mediante o conjunto de dados utilizado.

Baseado no trabalho de (SALMAN; KECMAN, 2012), onde problemas de regressão foram transformados em problemas de classificação, foi proposto nesse estudo uma etapa de discretização, a fim de gerar fenótipos divididos em classes, ou seja, categóricos, a partir dos atributos alvo presentes no conjunto de dados. Com isso uma segunda versão do problema original foi gerada, porém possibilitando a aplicação de algoritmos de FS que usem internamente estratégias de classificação. Esse procedimento foi realizado a fim de investigar possíveis nuances existentes na transformação de problemas de FS com algoritmos de regressão para uma versão que usa algoritmos de classificação. A ilustração da etapa de discretização pode ser vista na parte B da Figura 13.

A literatura aponta para duas abordagens frequentemente utilizadas para o processo de discretização, são elas (ARAUZO-AZOFRA; BENITEZ; CASTRO, 2008; PEDREGOSA et al., 2011):

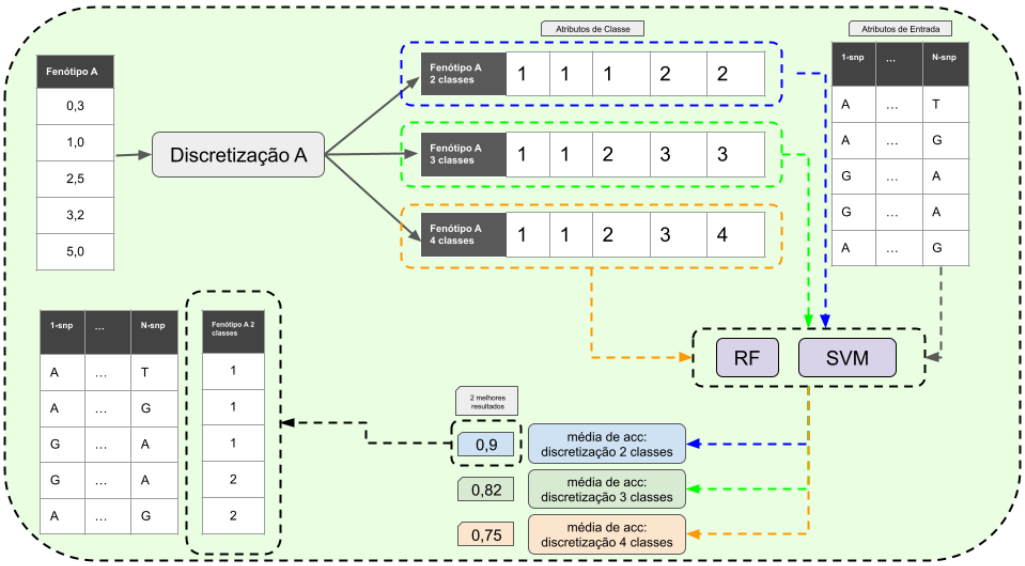
- Discretização Uniforme: consiste em traçar um intervalo de valores utilizando como limites o maior e menor valor observado em uma variável. Após isso, divide o intervalo em k espaços com mesmo tamanho. Por fim, atribui uma classe a cada exemplo, baseado em qual das divisões o valor do exemplo se encontra;

- Discretização por Agrupamento: os valores recebem uma classe discreta atribuída por um algoritmo de agrupamento executado sobre os valores dos atributos, levando em conta k centros. Essa abordagem em geral utiliza algoritmos não supervisionados para decisão do posicionamento dos centros.

Para a escolha da melhor abordagem de discretização a ser utilizada no conjunto de dados explorado nesse estudo, efetuamos testes utilizando o módulo *KBinsDiscretizer* da biblioteca scikit-learn (PEDREGOSA et al., 2011), que oferece uma implementação das duas abordagens de discretização. Cada uma das seis características fenotípicas foram submetidas as duas abordagens de discretização variando o valor k de classes geradas de 2 até 9. Para avaliar a quantidade de classes e o método de discretização que melhor dividiu os resultados fenotípicos, foi utilizada a média de acurácia obtida a partir do treinamento dos algoritmos RF e SVM para a tarefa de classificação utilizando os fenótipos

discretizados. Foram selecionados como atributos de classe discretizados para cada fenótipo as abordagens que apresentaram melhor resultado de acurácia. Essa escolha foi baseada no trabalho de (TSAI; CHEN, 2019), que investigou a melhor abordagem de discretização a ser aliada com algoritmos de FS. A Figura 14 ilustra o procedimento de escolha dos melhores atributos discretizados para uma única abordagem aplicada a uma das características fenotípicas.

Figura 14 – Exemplo da escolha do algoritmo de discretização



5.3 Métodos de seleção SNPs propostos

Para a segunda etapa do *pipeline*, seguindo diretrizes de alguns estudos que trabalharam com a seleção de SNPs apresentados na Seção 3.2.2, modelamos o problema de seleção de SNPs como uma tarefa de FS, em que os SNPs foram tratados como os dados de entrada, isto é, os atributos, e cada característica fenotípica como um atributo de classe a ser explorado. No entanto, a fim de compor uma etapa de seleção com critério robusto de escolha, como sugerido por Banerjee, Marathi e Singh (2020) e Chen et al. (2020), foram idealizados métodos de *ensemble*, isto é, métodos baseados na combinação de múltiplos algoritmos, como implementado nos trabalhos de Lin, Lin e Lane (2021). Os algoritmos utilizados foram selecionados através da investigação de trabalhos de seleção de SNPs, apresentados na Seção 3.2.2. Nesse estudo, oito algoritmos foram escolhidos para compor os métodos de FS combinados, são eles XGB, RF e *Extra Trees*, RFE junto aos algoritmos RF, SVM e *Extra Trees* e o algoritmo Boruta utilizando internamente os algoritmos *Extra Trees* e RF. Além disso, vale ressaltar que os algoritmos utilizados são amplamente utilizados, tal como ressaltado em (PUDJIHARTONO et al., 2022).

Com intuito de obter uma melhor performance de cada um dos algoritmos utilizados, aplicamos uma etapa de HPO, em que para cada um dos algoritmos base, isto é, algoritmos de ML internamente utilizados pelos algoritmos de FS, foram definidos hiper-parâmetros referentes a cada um dos fenótipos estudados, isto é, para cada combinação de algoritmo com fenótipo foi explorada uma configuração de hiper-parâmetros, de modo a melhorar a performance para o algoritmo em relação aquele fenótipo. Os algoritmos submetidos ao HPO foram XGB, RF, *Extra Trees* e SVM, pois o RFE e Boruta são abordagens *wrapper*, que utilizaram internamente os algoritmos citados como base para seleção de atributos. Os hiper-parâmetros explorados para cada um dos algoritmos são exibidos na Tabela 11. O procedimento de HPO foi aplicado tanto para os algoritmos na versão treinada para regressão, quanto para os mesmos algoritmos utilizando os fenótipos discretizados, treinados para a tarefa de classificação. Para o algoritmo *Extra Trees*, aplicado para classificação, o hiper-parâmetro *criterion* foi definido como gini, baseado nas métricas apresentadas em (JIANG et al., 2019). Os valores de HPO foram definidos com base em experimentos observados em alguns trabalhos relacionados como (PUDJIHARTONO et al., 2022), (SILVA et al., 2022) e (YIN; LI, 2022).

Tabela 11 – Hiper-Parâmetros explorados

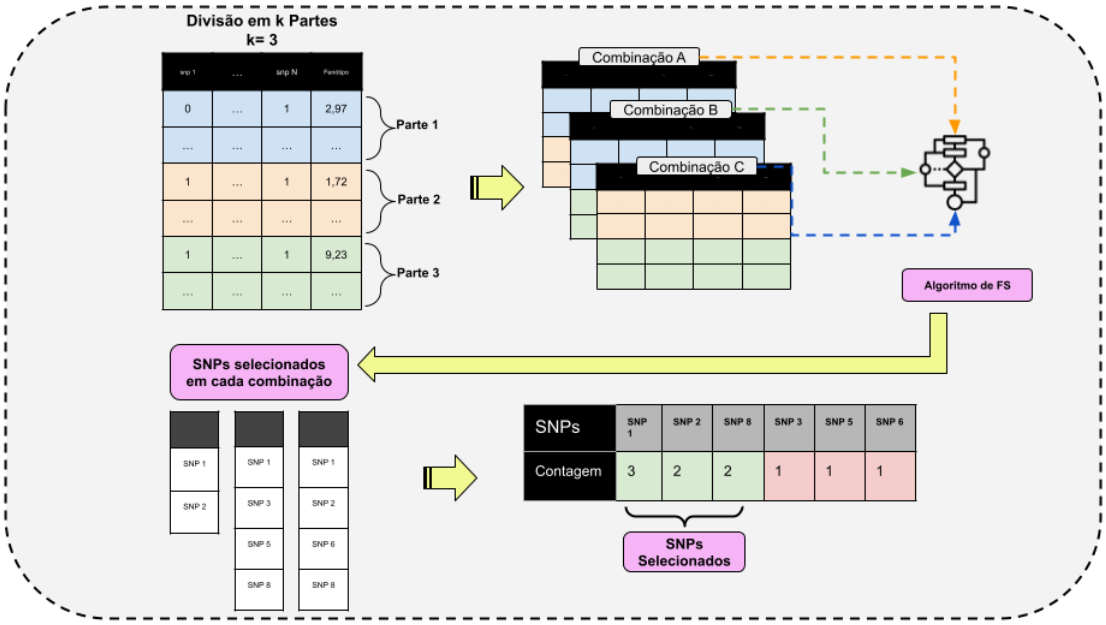
Algoritmo	**Parâmetros utilizados**	**Intervalo de Valores Explorados**
XGBoost	max_depth	[3, 5, 10, 15, 20]
	learning_rate	0.1 até 0.3 (variando 0.05)
	sub_sample	0.5 até 1 (variando 0.1)
	col_sample_bytree	0.4 até 1 (variando 0.1)
	colsample_bylevel	0.4 até 1 (variando 0.1)
	n_estimators	[100, 200, 500, 1000]
RF	bootstrap	[True, False]
	max_depth	[10, 20, 30, 40, 50, 60, 70, 80, 90, 100, None]
	max_features	[1, sqrt]
	min_samples_leaf	[1, 2, 4, 20, 50, 100]
	min_samples_split	[2, 5, 10,20,50]
	n_estimators	[200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000]
Extra Trees	bootstrap	[True, False]
	max_depth	[10, 20, 30, 40, 50, 60, 70, 80, 90, 100, None]
	max_features	['auto', 'sqrt']
	min_samples_leaf	[1, 2, 4, 20, 50, 100]
	min_samples_split	[2, 5, 10, 20, 50]
	n_estimators	[200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000]
	criterion	[squared_error, friedman_mse, absolute_error, gini ,poisson]
SVM	Kernel	[rbf]
	C	0.005 até 10 (Variando 0.1)
	Gamma	[1, 0.1 , 0.01, 0.001]

Em seguida, quatro métodos foram propostos, dois deles foram idealizados para trabalhar com as características fenotípicas em sua versão numérica, por conseguinte, utilizaram internamente algoritmos de ML para a tarefa de regressão. Os outros dois são semelhantes aos primeiros, contudo, utilizaram internamente os mesmos algoritmos de FS modificados para tarefa de classificação, executados com as características fenotípicas discretizados, gerados na etapa de processamento delineada na Seção 5.2.

O primeiro método, denominado Feature Selection Consensual Direta (FSCD), foi pensando da união de dois outros métodos. O primeiro é a FS baseados em voto majoritário (BOLÓN-CANEDO; ALONSO-BETANZOS, 2019) e seleção de atributos apoiada por amostragem sem substituição, que são algoritmos de FS aplicados em conjunto à Validação Cruzada (do inglês, "*Cross Validation*") (CV). A implementação da primeira estratégia foi explorada no trabalho de Bolón-Canedo et al. (2014), que combinou múltiplos algoritmos de FS para seleção dos melhores genes em um estudo de *microarray*. Não obstante, a FS em conjunto com CV tem sido utilizada em diversos trabalhos de seleção de SNP, como em (PUDJIHARTONO et al., 2022) e (PRAVEENA et al., 2023). Nesse método, para cada fenótipo, o conjunto de dados é dividido k partes, sendo que os algoritmos de FS foram aplicados sobre todas as combinações realizadas entre as partes, de modo a executar as respectivas estratégias de seleção. No conjunto de dados estudado no presente trabalho k foi definido como 3, seguindo as diretrizes dos trabalhos de Haeberle et al. (2017) e Aljouie, Schatz e Roshan (2019). Essa divisão aplicada sobre o conjunto de dados, que possui ao todo 541 exemplos, gera dois grupos com 180 amostras e um 181, número de amostras acima de 100, assim como sugerido por alguns trabalhos como (WASIKOWSKI; CHEN, 2010) e (HAN; WILLIAMSON; FONG, 2021). Após isso, foi aplicada uma etapa de filtragem onde, para cada algoritmo, apenas os SNPs que apareceram em mais um metade das vezes mais 1 conjuntos foram selecionados. Essa estratégia é dita voto majoritário. A estratégia por trás desse procedimento é a utilização de diferentes partes do conjunto de dados mantendo os atributos que tenham maior importância para os exemplos, assim como argumentado por Bolón-Canedo et al. (2014). A Figura 15 ilustra a primeira etapa desses métodos para um algoritmo executado sobre uma características fenotípicas, onde os dados são divididos e o algoritmo seleciona os SNPs mais importantes em cada uma das divisões.

Em seguida, foi realizada, para cada SNP, a contagem do número de algoritmos em que tal SNP apareceu. Os SNPs foram reunidos em grupos de consenso, isto é, os SNPs foram agrupados de acordo com o número simultâneo de algoritmos que os selecionaram. Esses grupos variaram de 1 até 8 (o número total de algoritmos de FS utilizados), sendo que um SNP selecionado por apenas dois algoritmos, simultaneamente, pode ser considerado como tendo uma evidência fraca em relação a outro SNP selecionado por quatro algoritmos simultaneamente. Ao final, para decidir o número de consenso mínimo a ser utilizado foi realizada uma contagem do número de SNPs em cada um desses grupos de consenso para

Figura 15 – Processo de divisão e aplicação dos algoritmos de FS



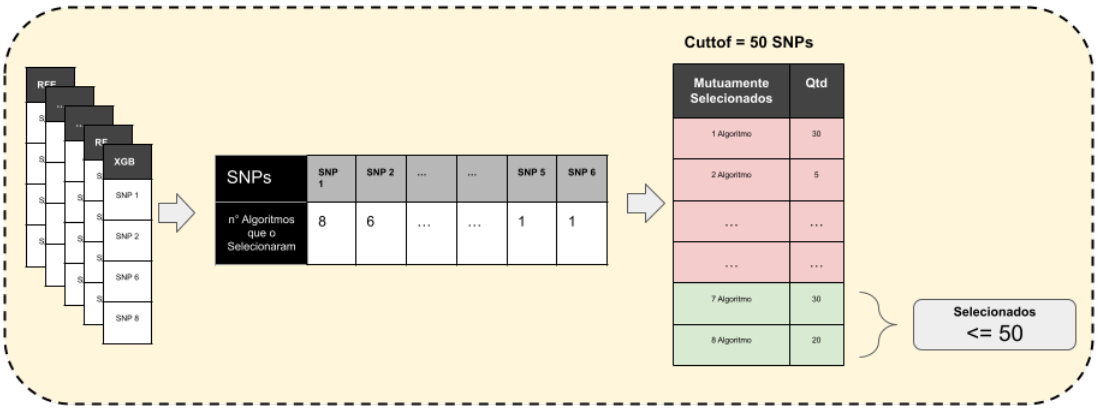
cada fenótipo. Então, foi definido um valor m, utilizado como valor filtro de quantidade de máxima de SNPs a serem selecionados. Para esse trabalho o valor de m foi limitado a 50, significando que o número máximo de SNPs selecionados seria 50. Esse valor máximo de SNPs foi definido por pesquisadores da Embrapa visando o custo de validação dos SNPs resultantes, visto que serão avaliados em estudos posteriores através de experimentos envolvendo sondas TaqMan (SCHLEINITZ; DISTEFANO; KOVACS, 2011) e técnicas de edição gênica como em (DOUDNA; CHARPENTIER, 2014). A Figura 16 ilustra esse procedimento de seleção consensual.

Outro método semelhante a esse, denominado Feature Selection Discretizada Consensual Direta (FSDCD) foi idealizado seguindo o mesmo procedimento de execução, contudo os fenótipos com valores numéricos foram discretizado e a seleção de atributos foi realizada com versões dos algoritmos treinadas para classificação. Esse método explorou as diferenças presentes na discretização das características fenótipicas discretas, a fim de investigar nuances relativas a esse procedimento discretização.

Para a implementação dos métodos FSCD e FSDCD, cada algoritmo de FS aplicado a um dos fenótipos presentes no conjunto de dados foi executado como um processo separado, de modo a armazenar os resultados individuais em um arquivo no servidor utilizado para execução. Isso permitiu maior velocidade de execução, uma vez que a máquina utilizada contava com múltiplos núcleos, permitindo a execução paralela dos algoritmos. Após a execução de todos os métodos, os resultados foram compilados seguindo a estratégia de junção consensual.

Um terceiro método para seleção de SNPs, denominado Feature Selection Consensual Iterativa (FSCI), foi proposto baseado na utilização de Validação Cruzada Repetida (do

Figura 16 – Procedimento de seleção de SNPs para 1 fenótipo



inglês, "*Repeated Cross-Validation*") (RCV), discutido nos trabalhos de Güney e Öztoprak (2018) e Abbas e El-Manzalawy (2020), onde o uso de RCV foi investigado em algoritmos de FS como procedimento para seleção robusta e balanceada. Esse método consiste em executar n vezes o procedimento de divisão CV, realizando a seleção de SNPs em cada um dos conjuntos de dados criados.

Contudo, ao final de cada seleção de SNPs efetuada com uma das divisões do conjunto de dados, são mensurados os valores de erro obtidos com o algoritmo interno do método de FS executado, utilizando apenas os SNPs selecionados. Esses valores ficam sendo associados a todos os SNPs selecionados pelo algoritmo em determinada iteração do CV. Esse procedimento é realizado para todos os algoritmos com cada um dos fenótipos.

Ao final, para cada algoritmo de FS aplicado a um fenótipo, são removidas as iterações onde a métrica de erro foi maior que a média de erro observada. Essa filtragem foi efetuada a fim de eliminar SNPs das iterações que contribuíram negativamente com as métricas de performance.

Em seguida, para cada algoritmo foram contados o número de vezes em que determinado SNP apareceu, então foram filtrados apenas SNPs que obtiveram um número de contagem maior que média. Essa filtragem foi implementada de modo a preservar apenas SNPs selecionados em um número de vezes significativamente grande.

Após isso o procedimento de seleção consensual de SNPs foi realizado de modo semelhante aos métodos FSCD e FSCI, em que os SNPs foram agrupados pela quantidade de algoritmos que os selecionaram simultaneamente e filtrados levando em conta um número n, definido para esse estudo como sendo 50. Assim como mencionado para os métodos FSCD e FSDCD o valor de 50 SNPs foi definido pela Embrapa a fim de possibilitar vali-

dações experimentais posteriores. O FSCI foi idealizado a fim de investigar se a aplicação repetida do método CV pode influenciar positivamente no processo de seleção de SNPs, de modo a trazer evidência para alguns SNPs mais relevantes. Vale ressaltar, apesar do número máximo ser 50, caso seja ultrapassado esse valor, escolhe-se um número menor que 50 para utilização do método. Em trabalhos futuros e de exploração outros procedimentos de escolha podem ser adotados, tal como realizar um ranqueamento baseado no número de vezes que determinado SNP foi escolhido de modo que sejam escolhidos exatamente o mesmo número de SNPs delimitados no problema, ao invés de u número inferior

5.4 Avaliação dos métodos

Para mensurar a performance dos métodos propostos, foram realizadas comparações entre a performance de algoritmos treinados para tarefas de predição, utilizando os SNPs selecionados pelos métodos propostos e utilizando todo o conjunto de SNPs, então os resultados de cada teste foram comparados. Nesse momento o conjunto de dados foi divido em Treino e Teste, na proporção de 80% das amostras sendo utilizadas para o treino dos algoritmos e 20% para teste. Não foi separado a partir desse conjunto de dados amostras utilizadas para validação isolada, devido ao número de amostras disponíveis. As repartições dos dados em cada divisão foram repetidas para os demais durante cada experimento individual.

Para os métodos FSCD e FSCI, cada conjunto de SNPs selecionados da combinação entre todos os algoritmos aplicados para determinada características fenotípica, foi treinado para tarefa de regressão por quatro algoritmos, são eles XGB, RF, SVM e *Extra Trees*, onde foram calculadas as métricas Erro Médio Absoluto (do inglês, "*Mean Absolute Error*") (MAE), Erro Quadrático Médio (do inglês, "*Mean Square Error*") (MSE) e Raiz do Erro Quadrático Médio (do inglês, "*Root Mean Square Error*") (RMSE), que já foram apresentadas na Tabela 2. Vale ressaltar que 4 algoritmos foram utilizados para as comparações, afim de realizar comparações mais gerais, ou seja, ao invés de simplesmente comparar os resultados de performance de cada método de seleção de SNPs através de um único algoritmo, uma comparação com mais algoritmos confere maior rigor para a análise. A escolha dos algoritmos XGB, RF, SVM e *Extra Trees* para essa tarefa se deu pelo fato de que esses são os algoritmos utilizados internamente nos métodos de seleção, isto é, no caso do método RFE foram utilizados internamente os algoritmos *Extra Trees*, RF, SVM. Por sua vez, o XGB também foi utilizado isoladamente para seleção de SNP.

Para os FSDCD e Feature Selection Discretizada Consensual iterativa (FSDCI), o mesmo procedimento foi realizado, utilizando os mesmos quatro algoritmos mencionados, contudo em sua versão adaptada para tarefa de classificação. As métricas comparadas para os métodos FSDCD e FSDCI foram acurácia e Área Sob a Curva ROC (do inglês, "*Area Under ROC Curve*") (AUC-ROC). No entanto, a título de comparação geral, os

SNPs selecionados por esses dois métodos com as características fenotípicas discretizados também foram utilizados em algoritmos de regressão, verificando se a discretização tem efetividade geral para selecionar SNPs em comparação aos métodos que utilizaram algoritmos de regressão. As métricas para essa segunda comparação foram as mesmas utilizadas para os métodos com fenótipos sem discretização, ou seja, MAE, MSE e RMSE.

Em seguida, utilizando as métricas de erro MAE, MSE e RMSE, os quatro métodos foram comparados em relação a cada fenótipo, de modo a observar qual método teve melhor resultado no contexto global dos fenótipos. Além disso, foi mensurado o tempo médio de execução de cada algoritmo em relação a cada um dos fenótipos, aferindo o desempenho e tempo viável de cada um dos métodos. As especificações do hardware utilizado são apresentadas na Tabela 12.

Tabela 12 – Especificações técnicas da máquina utilizada

Especificações	
Processador	AMD Rome EPYC 7282 CORE 3GHz
Armazenamento	2 SSD Micro 960GB NVMe PCIe 3.1
RAM	4 Unidades de Memória 32 GB DDR4 3200 HZ

Após a comparação entre os métodos FSCD, FSCI, FSDCD e FSDCI, o melhor método dentre eles foi comparado com outros métodos presentes na literatura para tarefa de regressão. Foram escolhidos para comparação três grupos de métodos presentes na literatura. O primeiro grupo testado foi o RFE utilizando internamente o RF, o SVM e Extra Trees. Além deles, foi comparado o método Boruta utilizando internamente o RF e também Extra Trees. Por fim, foram comparados cinco RBAs, a saber ReliefF (KONONENKO, 1994), SURF (GREENE et al., 2009), SURF* (GREENE et al., 2010), MultiSURF (URBANOWICZ et al., 2018b) e MultiSURF* (GRANIZO-MACKENZIE; MOORE, 2013). As implementações desses métodos estão disponíveis no pacote ReBATE proposto por Urbanowicz et al. (2018b).

Para essa comparação cada um dos métodos apresentados (ou seja, RBAs, baseados em RFE e baseados em Boruta), foi executado sobre o conjunto de dados, selecionando SNPs importantes para cada característica fenotípica. Após isso, os algoritmos XGB, RF, SVM e *Extra Trees* foram treinados para a tarefa de predição de cada um dos fenótipos, utilizando os SNPs selecionados como importantes para cada fenótipo. Esses algoritmos são os mesmos utilizados na avaliação dos 4 métodos propostos, que foi mencionada anteriormente. As métricas mensuradas em cada um desses 4 algoritmos foram MAE, MSE e RMSE. Por fim, os SNPs do método que apresentou melhor performance foram

selecionados para análise e exploração de sistemas biológicos.

Vale ressaltar que em grande parte dos casos, algoritmos de FS não são capazes de melhorar métricas de performance, de fato são usados para redução do número de atributos assim como mencionado na Seção anterior. Contudo, em alguns casos, para conjuntos de dados redudantes é possível observar melhoras nos resultados de performance, os trabalhos de Pullissery e Starkey (2023) e Goyal, Sota e Arora (2023).

5.5 Exploração de SNPs selecionados

Uma vez tendo o conjunto de SNPs selecionados pelo *pipeline* é necessário apontar critérios de impacto putativos para os SNPs selecionados. Baseado nos trabalhos de Kim et al. (2022), Zhong et al. (2021) e Zhang et al. (2023), definimos os seguintes critérios para exploração e avaliação dos SNPs:

- **Presença do SNP em bancos de dados de sequências**: procedimento de verificação da existência do SNP em outros estudos que já realizaram sequenciamento, evitando a necessidade de realização do sequenciamento do Ácido Desoxirribonucleico (do inglês, "*DeoxyriboNucleic Acid*") (DNA) adjacente a cada SNP;

- **Análise de efeito modificador**: análise feita a partir da posição do SNP, onde foi verificado se o mesmo encontra-se em uma região 3', 5' ou intron, apontando maior possibilidade de configurar-se como modificador de função (por alterar regiões promotoras (KIMURA-KATAOKA et al., 2012)). Foi verificado se está presente em um exon (revelando uma ação moderadora (WAITE; DARDICK, 2021)). Por fim, verificado se não está localizado em nenhuma posição gênica, apresentando baixo efeito putativo em detrimento dos demais, devido a não existência estrutura catalogada para exploração.

- **Enriquecimento Funcional**: análise ORA realizada a partir dos genes que possuíam SNPs os modificando, verificando possíveis sistemas biológicos enriquecidos sobre os genes identificados, favorecendo a efetividade de determinado SNP.

Para a investigação da presença de SNPs utilizamos como referência de sequência as variações encontradas na ferramenta RiceVarMap (ZHAO et al., 2015). Essa ferramenta foi escolhida em detrimento das demais, pois além de possuir como base o genoma referência utilizado nesse trabalho, isto é, Os-Nipponbare-Reference-IRGSP-1.0, também apresenta informações disponíveis e fácil integração em relação a estudos existentes, uma vez que possui versão web e dados disponibilizados para download e integração. Cada SNP do conjunto resultante retornado pelo *pipeline* foi procurado na ferramenta, de modo que os SNPs não encontrados foram removidos do conjunto de análise.

Em seguida, a partir do genoma de referência baixado do projeto RAPDB (SAKAI et al., 2013), foram recuperadas as informações sobre estruturas genômicas, tais como genes, exons, 3' UTR, 5'UTR e introns onde estão localizados os SNPs confirmados. Apenas os SNPs localizados nas regiões putativamente modificadoras e de exon que de fato contêm a presença de genes, foram investigados e permaneceram para análise. Por fim, os genes encontrados a partir dos SNPs restantes foram submetidos ao GProfiler (RAUDVERE et al., 2019) retornando as vias e sistemas biológicos enriquecidos, sendo que os SNPs incidentes em genes que não enriqueceram sistemas biológicos foram removidos, selecionando os SNPs passíveis de investigação intensa e putativamente mais relevantes.

Capítulo 6

Experimentos e Discussão

Esse capítulo apresenta e discute os experimentos realizados, e está organizado em cinco seções. Inicialmente, discute-se na Seção 6.1 sobre a composição dos dados após a etapa de pré-processamento e aplicação das metodologias de discretização que obtiveram melhor desempenho. Em seguida, na Seção 6.2 é apresentado o resultado obtido com HPO e também a avaliação de performance dos métodos propostos. São exibidas informações quantitativas com relação aos SNPs que foram selecionados consensualmente entre os métodos executados. A Seção 6.3 apresenta uma comparação com alguns métodos presentes na literatura. Na Seção 6.4 foram explorados sistemas biológicos estatisticamente relacionados aos SNPs selecionados como associados a determinados fenótipos. Por fim, são discutidos detalhes sobre os experimentos realizados e destaques observados.

6.1 Pré-Processamento

Como apresentado na Seção 5.2, a primeira etapa realizada foi a aplicação da abordagem *One-hot Enconding* que transformou as variáveis de entrada, os SNPs, em variáveis numéricas. Nesse caso, a dimensão original dos dados é preservada, uma vez que a cardinalidade dos dados é igual a 2, onde apenas uma das colunas é mantida, de modo que a variação de valores 0 e 1 é suficiente para representar a cardinalidade dos dados.

A Tabela 13 exibe o resultado observado com a discretização das características fenotípicas. A primeira coluna apresenta o nome do algoritmo utilizado para a discretização, na sequência é exibido o número de classes em que os dados foram divididos, as acurácias obtidas nos algoritmos treinados, média de acurácia entre os dois algoritmos e a abreviação do fenótipo discretizado. É apresentado o resultado de discretização por fenótipo, que

obteve melhor resultado de acurácia média para os algoritmos na tarefa de classificação do determinado fenótipo. O restante dos resultados, contendo os parâmetros utilizados e performances observadas são apresentados na Seção C.1 do Apêndice C, que contém os resultados expandidos de cada experimento realizado.

Tabela 13 – Métricas de avaliação de algoritmos de classificação

Estratégia Utilizada	Número de Classes	Acurácia SVM	Acurácia RF	Acurácia Média	Fenótipo Abreviado
Uniforme	2	0.870606	0.887272	**0.878939**	Aca
Uniforme	3	0.966605	0.966605	**0.966605**	C/L
Agrupamento	5	0.889083	0.902046	**0.895565**	Flor
Uniforme	2	0.807837	0.785615	**0.796726**	Int
Uniforme	4	0.950092	0.951944	**0.951018**	Alt
Uniforme	3	0.894700	0.887324	**0.891012**	Prod

É possível observar que a algoritmo de discretização uniforme apresentou melhor resultado para todos os fenótipos com exceção da Floração. Esse fato corrobora com a hipótese de que, em casos onde os dados possuem distribuição balanceada de observações, a divisão de classes em cortes uniformes pode ser uma boa estratégia. Outrossim, observa-se que o número de classes para os quais os dados foram divididos variou de 2 a 5.

6.2 Avaliação dos resultados

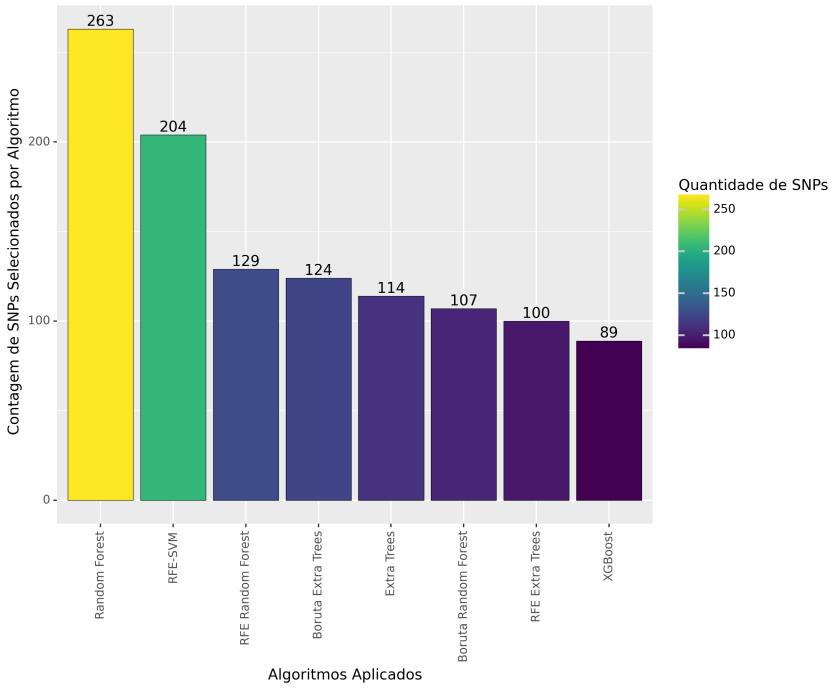
Essa seção irá discutir os resultados dos experimentos propostos no Capítulo 5. Dentre as etapas realizadas, a HPO aplicada em cada algoritmo não será discutida nessa seção por se tratar de configurações especificas, contudo os resultados detalhados são exibidos na Seção C.2 do Apêndice C. O restante dessa seção está divida em três subseções que tratam, respectivamente, dos resultados obtidos com os algoritmos utilizados com as características fenotípicas numéricas, os algoritmos avaliados com as características fenotípicas discretizados e uma comparação entre as duas estratégias.

6.2.1 FS Consensual Direta - FSCD

A Figura 17 exibe a soma de SNPs selecionados em todos os fenótipos para cada um dos algoritmos aplicados. É possível observar que para a *Random Forest*, que obteve a maior soma de SNPs selecionados em todos os fenótipos, a quantidade total de SNPs selecionados foi cerca de 5,5% do total de SNPs do conjunto de dados, representando uma filtragem significativa.

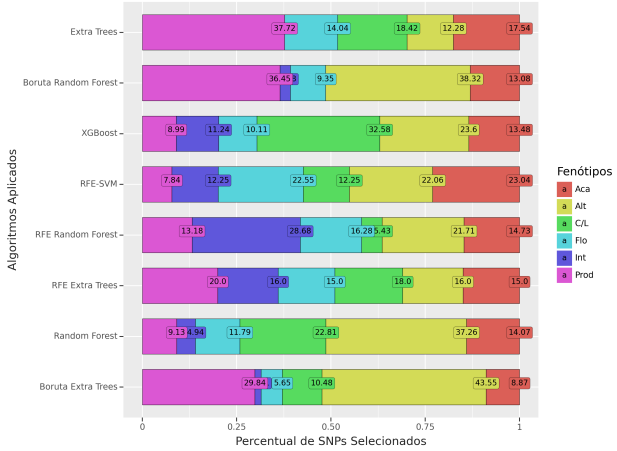
A Figura 18 exibe a soma percentual de SNPs selecionados pelos algoritmos, dividindo as somas por fenótipo. O eixo das abscissas representa a porcentagem de SNPs selecionados em relação ao todo, enquanto que no eixo das ordenadas é apresentado cada

Figura 17 – Contagem de SNPs selecionados - FSCD



algoritmo utilizado. Os diferentes fenótipos são distinguidos pelas cores apresentadas na legenda, sendo que próximo a cada barra há um número que equivale ao total de SNPs que determinado algoritmo selecionou como importante para cada fenótipo. É possível observar que para os algoritmos Extra Trees e RF (nas versões usando Boruta), e também na Random Forest usada isoladamente, a maior parte dos SNPs selecionados encontra-se direcionada ao fenótipo que denota a Altura.

Figura 18 – Percentual de SNPs selecionados por fenótipo - FSCD



Ao realizar uma avaliação consensual, podemos observar pela Tabela 14, que o número de SNPs consensualmente selecionados por mais de um algoritmo não superou o total de 50 SNPs em nenhum dos fenótipos. Os fenótipos que denotam floração e produtividade obtiveram SNPs selecionados por 5 entre os 8 algoritmos utilizados. Contudo, não houve

SNPs apontados como importantes por mais de dois algoritmos para o fenótipo de comprimento por largura dos grãos, isto é, os algoritmos executados, não selecionaram SNPs iguais entre si. Além disso, para FSCD foi possível utilizar todos os SNPs selecionados por mais de dois algoritmos, visto que em todos os casos a soma acumulada de SNPs não superou 50 em nenhum dos casos. Ainda nesse contexto, vale ressaltar que foram obtidos 139 SNPs, levando em conta a soma de SNPs selecionados para todos os fenótipos por pelo menos dois algoritmos. Ao agrupar todos os SNPs, gerados por todos os algoritmos, e remover os duplicados, obtivemos 109 SNPs distintos, ou seja diferentes entre si, sendo que 24 foram apontados como mais importantes para mais de um fenótipo com 18 deles sendo importantes para pelo menos 2 fenótipos e 6 sendo importantes para 3 fenótipos. O link de acesso para tabela com esses resultados é apresentada no Apêndice C.

Tabela 14 – SNPs selecionados consensualmente - FSCD

Fenótipo	**Aca**			**Alt**			**C/L**	**Flo**				**Int**		**Prod**			
Número de Algoritmos	2	3	4	2	3	4	2	2	3	4	5	2	3	2	3	4	5
Soma de SNPs Selecionados	11	2	9	21	7	11	13	7	2	6	2	9	3	10	11	11	4
Soma Acumulada	22	11		39	18		13	17	10	8		12		36	26	15	

Por fim, a Tabela 15 apresenta métricas de erro entre quatro algoritmos comparando suas execuções utilizando todos os SNPs e uma outra versão utilizando apenas os SNPs selecionados pelo FSCD. São destacados, em negrito, os valores que obtiveram menor valor de erro para cada algoritmo. É possível observar que nos experimentos de validação dos SNPs selecionados, levando em conta o MAE, todos os conjuntos de SNPs selecionados para os respectivos fenótipos obtiveram resultado de erro, para a média geral dos algoritmos, menor que o conjunto total de SNPs. Além disso, para MSE e RMSE os conjuntos de SNPs selecionados não obtiveram resultado médio geral melhor para todos os fenótipos, contudo as diferenças nas taxas de erros foram baixas. Portanto, isso aponta que a seleção de SNPs foi bem sucedida, uma vez que o número de SNPs foi significativamente reduzido, sem afetar as métricas de erro.

6.2.2 FS Consensual Iterativa - FSCI

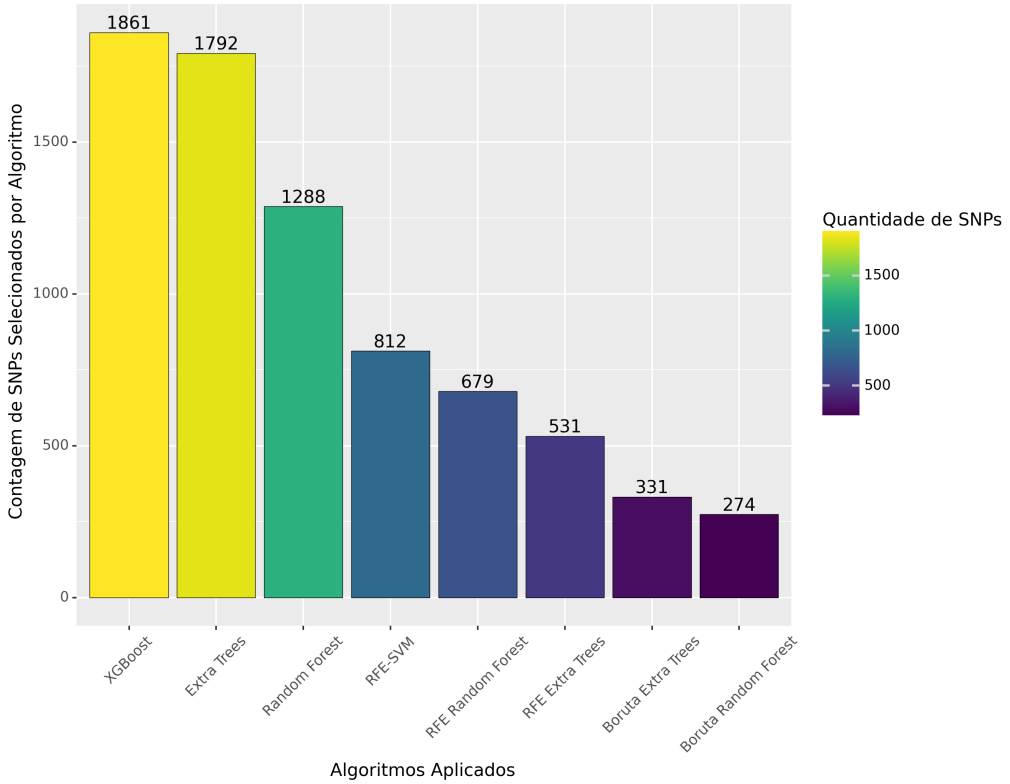
Após a execução dos algoritmos propostos e aplicação do processo de filtragem, foi observada uma redução significativa do número de SNPs selecionados como importantes para alguns fenótipo. A Figura 19 exibe as quantidades de SNPs selecionados em cada um dos algoritmos aplicados, levando em conta todos os fenótipos explorados. Observa-se que as abordagens baseadas em árvores de decisão, isto é, XGB, RF e *ExtraTree*, selecionaram mais SNPs que as demais, levando em consideração todos os fenótipos. Por

Tabela 15 – Métricas de erro comparadas - FSCD

	Erro Médio Quadrático - MSE									
	Extra Tree		Random Forest		SVM		XGBoost		Média Geral	
	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados
Aca	0.090679	**0.068577**	**0.054843**	0.054995	**0.052703**	0.053206	**0.059404**	0.073343	**0.064408**	0.062530
Alt	**2.702741**	4.137990	**1.868797**	1.959246	1.665701	**1.585413**	**2.096854**	2.352281	**2.083523**	2.508732
C/L	0.922409	**0.430562**	0.493227	**0.377639**	0.491065	**0.374816**	0.552142	**0.419000**	0.614710	**0.400504**
Flo	1.105070	**0.824166**	0.653096	**0.648069**	0.682541	**0.675460**	0.833779	**0.759325**	0.818621	**0.726755**
Int	10.364484	**6.960226**	**5.751960**	5.777289	5.939182	**5.694367**	7.201450	**6.825759**	7.314269	**6.314410**
Prod	45029.483113	**39405.819815**	**25613.221264**	31054.294092	43058.535631	**40461.436752**	28905.274906	35681.936871	35651.628729	**36650.871882**
	Raíz do Erro Médio Quadrático - RMSE									
	Extra Trees		Random Forest		SVM		XGBoost		Média Geral	
	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados
Aca	0.301130	**0.261873**	**0.234187**	0.234511	**0.229572**	0.230664	**0.243730**	0.270819	0.253787	**0.250061**
Alt	**1.644001**	2.034205	**1.367039**	1.399731	1.290621	**1.259132**	**1.448052**	1.533715	**1.443442**	1.583898
C/L	0.960421	**0.656172**	0.702301	**0.614524**	0.700760	**0.612223**	0.743062	**0.647302**	0.784035	**0.632854**
Flo	1.051223	**0.907836**	0.808144	**0.805027**	0.826160	**0.821864**	0.913115	**0.871392**	0.904777	**0.852499**
Int	3.219392	**2.638224**	**2.398324**	2.403599	2.437044	**2.386287**	2.683552	**2.612615**	2.704491	**2.512849**
Prod	212.201515	**198.508992**	**160.041311**	176.222286	207.505507	**201.150284**	**170.015514**	188.896630	**188.816389**	191.444174
	Erro Médio Absoluto - MAE									
	Extra Trees		Random Forest		SVM		XGBoost		Média Geral	
	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados
Aca	0.218968	**0.186213**	0.175662	**0.175431**	0.179680	**0.166678**	**0.179553**	0.187322	0.188466	**0.178911**
Alt	0.890829	**0.821612**	**0.681281**	0.696552	0.627249	**0.604372**	0.791045	**0.767495**	0.747601	**0.722508**
C/L	0.578046	**0.384334**	0.408184	**0.360909**	0.367873	**0.329889**	0.434424	**0.374666**	0.447132	**0.362449**
Flo	0.521900	**0.463263**	0.494301	**0.449719**	0.430634	**0.394256**	0.530098	**0.481480**	0.494233	**0.447179**
Int	2.556378	**1.952181**	1.824183	**1.813931**	1.872711	**1.791485**	2.066189	**1.941935**	2.079865	**1.874883**
Prod	139.733432	**109.094314**	**95.127386**	99.502040	120.414477	**117.488619**	108.448872	**105.841046**	115.931042	**107.981505**

sua vez, os algoritmos ExtraTree e Random Forest utilizados em conjunto a RFE tiveram uma filtragem mais restritiva. Além disso, o algoritmo Boruta, em conjunto com RF e *ExtraTree*, mostrou ser mais restritivo que os demais, selecionando um total de SNPs inferior a 10% do total de SNPs disponíveis.

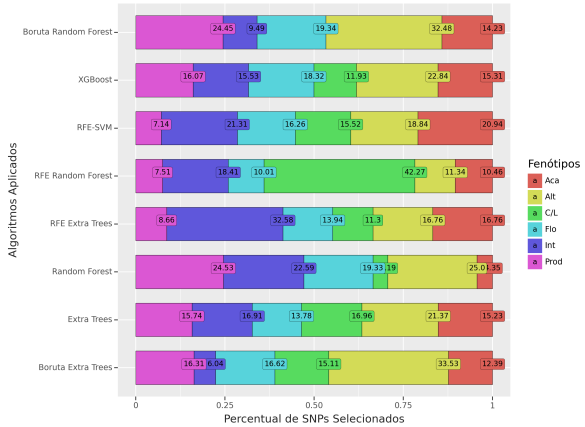
Figura 19 – Soma de SNPs selecionados - FSCI



Quando avaliamos as quantidades de SNPs selecionadas em cada algoritmo separando-

os por fenótipos, é possível observar uma proporção equilibrada de SNPs selecionados para cada fenótipo em grande parte dos algoritmos executados. As exceções foram o algoritmo *RFE-RF*, que apresentou um percentual de 42,27% dos SNPs selecionados para o fenótipo comprimento por largura, e o *RFE-ExtraTree*, com percentual de 32,58% dos SNPs selecionados para o fenótipo percentual de grãos inteiros. A Figura 20 exibe a contagem e proporção dos SNPs selecionados em cada algoritmo, realizando a divisão por fenótipo. Essa figura exibe quantidades e percentual de SNP, semelhantes a Figura 18. No gráfico da direita são apresentadas as quantidades de SNPs selecionados como importantes para cada fenótipo. O gráfico à esquerda trás os percentuais equivalentes às quantidades observadas no gráfico à direita.

Figura 20 – Percentual e quantidades de SNPs selecionados por fenótipo - FSCI



Uma informação pertinente com relação aos SNPs selecionados em cada abordagem é quantidade de SNPs consensualmente apontados como importantes para cada fenótipo, isto é, dentre todos os SNPs selecionados como importantes para um fenótipo como Produtividade, quais deles foram selecionados por mais de 1 algoritmo. A Tabela 16 apresenta essa informação, trazendo o nome do fenótipo, e o número de SNPs distintos selecionados por k algoritmos simultaneamente. Foram omitidos os SNPs apontados como importantes por apenas um algoritmo uma vez que essa informação equivale a soma de SNPs distintos apontados como importantes para todos os algoritmos em relação a determinado fenótipo. É possível observar que nenhum SNP foi consensualmente apontado como importante por todos os algoritmos. Contudo, para todos os fenótipos com exceção do fenótipo que denota comprimento por largura do grão, houve SNPs que foram selecionados por pelo menos 5 algoritmos. Podemos observar que a contagem acumulada de SNPs que foram selecionados por pelo menos 4 algoritmos não superou o valor de 50 SNPs para nenhum fenótipo. Além disso, para o fenótipo produtividade, houve dois SNPs selecionados por 6 dentre os 8 algoritmos. Por fim, vale ressaltar que foi obtido um total de 195 SNPs selecionados por pelo menos 3 algoritmos, dos quais 146 SNPs distintos com 37 SNPs selecionados consensualmente entre os fenótipos. O link de acesso para tabela com esses

resultados é apresentado no Apêndice C.

Tabela 16 – SNPs selecionados consensualmente - FSCI

Fenótipo	**Aca**				**Alt**				**C/L**			**Flo**				**Int**				**Prod**					**Total**
Número Algoritmos	2	3	4	5	2	**3**	4	5	2	3	4	2	3	4	5	2	3	4	**5**	2	3	4	5	6	
Número de SNPs Selecionados	85	13	33	**5**	210	30	38	3	124	53	3	142	7	43	**7**	173	35	14	**3**	87	7	30	14	2	1162
Soma Acumulada	136	51	**38**		282	72	**41**		180	56	**3**	199	57	**50**		225	52	**17**		140	53	**46**			**196**

A Tabela 17 apresenta as três métricas de erro avaliadas para os quatro algoritmos utilizados para validação dos SNPs selecionados. Para cada algoritmo utilizado na avaliação são exibidas a média de erro obtido pelo algoritmo treinado com todos os SNPs, designado pela coluna *Base* e do erro mensurado para o mesmo algoritmo treinado apenas com os SNPs selecionados consensualmente. Na última coluna é apresentada a média de erro dos quatro algoritmos. Além disso, estão destacados, em negrito, os menores valores de erro obtidos na comparação de cada algoritmo utilizando todos os SNPs e apenas os SNPs selecionados. Podemos observar que para as métricas de erro MSE e RMSE, com exceção do fenótipo altura, os conjuntos de SNPs selecionados pela filtragem consensual em cada fenótipo obtiveram valores de média geral de erro inferiores. Além disso, para MAE, os conjuntos de SNPs selecionados mostraram desempenho superior, pois em todos os fenótipos o resultado do erro médio geral foi inferior. Esse resultado denota um bom desempenho para a FSCI, pois, a quantidade de SNPs foi reduzida significativamente, melhorando o resultado dos algoritmos de aprendizado na tarefa de regressão.

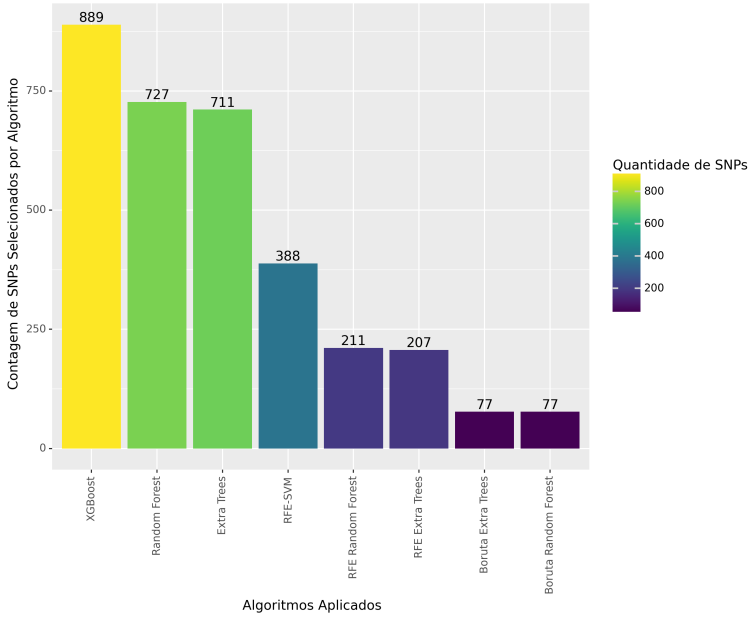
Tabela 17 – Métricas de erro comparadas - FSCI

	Erro Médio Quadrático - MSE									
	Extra Trees		**Random Forest**		**SVM**		**XGBoost**		**Média Geral**	
	Base	**SNPs Selecionados**	**Base**	**SNPs Selecionados**	**Base**	**SNPs Selecionados**	**Base**	**SNPs Selecionados**	**Base**	**SNPs Selecionados**
Aca	0.090537	**0.060336**	0.053208	**0.051009**	0.052703	**0.046925**	0.059404	**0.054881**	0.063963	**0.053288**
Alt	**2.717913**	3.537064	**1.828476**	1.939475	1.665701	**1.594325**	**2.096854**	2.166816	**2.077236**	2.309420
C/L	0.902999	**0.408858**	0.503444	**0.404657**	0.491065	**0.445248**	0.552142	**0.408868**	0.612412	**0.416908**
Flo	1.110039	**0.926611**	0.650200	**0.596417**	0.682541	**0.591639**	0.833779	**0.695487**	0.819140	**0.702539**
Int	10.384135	**7.339993**	**5.849938**	6.277830	5.939182	**5.250541**	**7.201450**	7.414490	7.343676	**6.570714**
Prod	44530.810868	**35451.491663**	25440.844427	**28807.205858**	43058.535631	**40557.677239**	**28905.274906**	31456.588933	35483.866458	**34068.240923**
	Raíz do Erro Médio Quadrático - RMSE									
	Extra Trees		**Random Forest**		**SVM**		**XGBoost**		**Média Geral**	
	Base	**SNPs Selecionados**	**Base**	**SNPs Selecionados**	**Base**	**SNPs Selecionados**	**Base**	**SNPs Selecionados**	**Base**	**SNPs Selecionados**
Aca	0.300894	0.245633	**0.230669**	**0.225852**	0.229572	0.216622	0.243730	**0.234267**	0.252909	**0.230841**
Alt	1.648609	1.880708	**1.352211**	1.392650	**1.290621**	1.262666	**1.448052**	1.472011	**1.441262**	1.519678
C/L	0.950263	**0.639420**	0.709538	**0.636127**	0.700760	**0.667269**	0.743062	**0.639428**	0.782568	**0.645684**
Flo	1.053584	0.962607	**0.806350**	**0.772281**	0.826160	**0.769181**	0.913115	**0.833959**	0.905063	**0.838176**
Int	3.222442	2.709242	**2.418665**	2.505560	**2.437044**	2.291406	**2.683552**	2.722956	2.709922	**2.563340**
Prod	211.023247	188.285665	**159.501863**	**169.726857**	207.505507	**201.389367**	**170.015514**	177.360055	188.371618	**184.575841**
	Erro Médio Absoluto - MAE									
	Extra Trees		**Random Forest**		**SVM**		**XGBoost**		**Média Geral**	
	Base	**SNPs Selecionados**	**Base**	**SNPs Selecionados**	**Base**	**SNPs Selecionados**	**Base**	**SNPs Selecionados**	**Base**	**SNPs Selecionados**
Aca	0.218626	**0.174648**	0.174559	**0.167469**	0.179680	**0.157261**	0.179553	**0.168667**	0.188105	**0.167011**
Alt	0.892869	**0.781895**	0.666748	**0.661200**	0.627249	**0.601889**	0.791045	**0.721892**	0.744478	**0.691719**
C/L	0.573958	**0.369434**	0.418259	**0.367646**	0.367873	**0.364695**	0.434424	**0.369433**	0.448629	**0.367802**
Flo	0.524915	**0.458078**	0.496546	**0.411810**	0.430634	**0.358282**	0.530098	**0.435962**	0.495548	**0.416033**
Int	2.559761	**2.026894**	**1.841544**	1.882314	1.872711	**1.735265**	2.066189	**2.015175**	2.085051	**1.914912**
Prod	138.723911	**102.830557**	94.593174	**94.543414**	120.414477	**117.752627**	108.448872	**97.861675**	115.545109	**103.247068**

6.2.3 FS Discretizada Consensual Direta - FSDCD

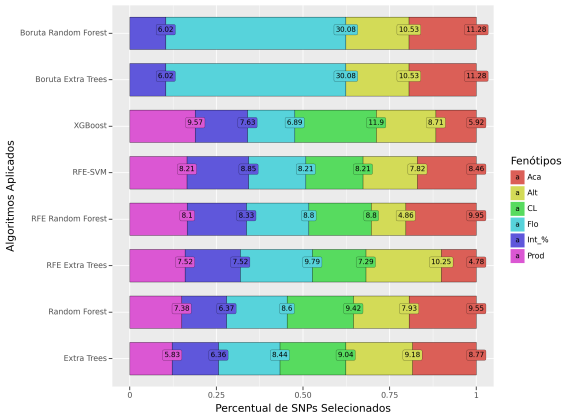
O método FSDCD foi aplicado a seis características fenotípicas discretizadas, provenientes das melhores abordagens de discretização executadas em cada um dos fenótipos do conjunto de dados original. A Figura 21 exibe a soma de SNPs selecionados em cada um dos algoritmos executados. É possível notar que as algoritmos baseados em árvores de decisão tiveram um total de SNPs mais acentuados que os demais algoritmos, seguidos dos algoritmos utilizados em conjunto a RFE. As variações do algoritmo Boruta obtiveram um resultado mais restritivo em relação à quantidade total de SNPs selecionados.

Figura 21 – Soma de SNPs selecionados - FSDCD



É possível observar na Figura 22 que os percentuais de SNPs selecionados para cada fenótipo foram uniformemente distribuídos uma vez que os percentuais variaram de aproximadamente 5% até 10%.

Figura 22 – Percentual de SNPs selecionados por fenótipo - FSDCD



A Tabela 18 exibe o número de SNPs selecionados como importantes para cada fenótipo. São exibidos o número de SNPs selecionados simultaneamente pelos algoritmos. A interpretação é feita da seguinte forma, a quarta coluna referente ao fenótipo acamamento tem o valor 12 na terceira linha, indicando que para o fenótipo acamamento, obtivemos 12 SNPs simultaneamente selecionados como importantes por quatro algoritmos, isto é, dentre todos os SNPs selecionados como importantes para o fenótipo acamamento, obtivemos 12 que foram selecionados por quatro algoritmos diferentes como importantes. A última linha apresenta a soma acumulada, que podemos interpretar da seguinte forma: para o fenótipo acamamento, temos 15 SNPs selecionados por pelo menos três algoritmos, sendo três SNPs selecionados por três algoritmos e 12 SNPs selecionados por quatro algoritmos.

Foram destacados na tabela, em negrito, o número de SNPs selecionados por pelo menos 3 algoritmos, em que a quantidade fosse menor que 50. Ao selecionar os SNPs destacados na tabela, total de SNPs selecionados por pelo menos três algoritmos consensualmente para todos os fenótipos foi de 92, entre os quais 85 foram SNPs distintos com 7 SNPs consensualmente selecionados em pelo menos 2 fenótipos.

Tabela 18 – SNPs selecionados consensualmente - FSDCD

Fenótipo	**Aca**				**Alt**				**CL**				**Flo**				**Int**				**Prod**		
Número de Algoritmos	1	2	3	4	2	3	4	5	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3
Número de SNPs Selecionados	417	31	3	12	39	7	9	1	449	67	10	1	404	27	11	27	378	33	3	6	408	41	2
Soma Acumulada	463	46	**15**		56	**17**	10		527	78	**11**		469	65	**38**		420	42	**9**		451	43	**2**

Por fim, é possível observar na Tabela 19 que os conjuntos de SNPs selecionados pela abordagem FSDCD obtiveram bons resultados de acurácia e AUC-ROC, para a média geral mensurada para todos os algoritmos. São exibidos os resultados obtidos para 4 algoritmos. Na coluna denominada Base é apresentado o resultado obtido por cada algoritmo, a partir do treinamento utilizando todos os SNPs. Ao lado de cada Base é apresentado o resultado do determinado algoritmo utilizando apenas os SNPs selecionados. Vale ressaltar que a AUC-ROC para os fenótipos comprimento por largura e altura apresentou melhora significativa, respectivamente, 5% e 32%. Estão destacados em negrito os melhores resultados, isto é, se para determinado algoritmo o conjunto de SNPs selecionado obteve melhor resultado na tarefa de classificação através do conjunto total de SNPs, ou com os SNPs selecionados.

A Tabela 20 apresenta os resultados de erro obtidos pelo treinamento de quatro algoritmos para a tarefa de regressão utilizando como valor esperado os valores observados nos fenótipos. A coluna denominada Base apresenta os resultados desses algoritmos utilizando todos os SNPs do conjunto de dados. A coluna denominada SNPs selecionados apresenta os resultados obtidos pelos algoritmos utilizando apenas os SNPs selecionados como importantes para os fenótipos. Foram destacados, para cada comparação, os menores valores de erro.

Tabela 19 – Métricas de performance - FSDCD

	AUC-ROC									
	Extra Trees		Random Forest		SVM		XGBoost		Média Geral	
	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados
Flo	0.835937	**0.850808**	**0.846295**	0.834362	0.817211	**0.826546**	**0.846033**	0.791340	**0.836369**	0.825764
Aca	**0.752783**	0.671615	**0.746632**	0.737588	**0.775634**	0.766681	**0.752780**	0.711194	**0.756957**	0.721769
Int	0.669234	**0.686561**	0.675378	**0.688025**	**0.692052**	0.675145	0.671640	**0.666161**	0.677076	**0.678973**
C/L	0.446886	**0.451584**	0.437112	**0.484567**	0.507728	**0.589473**	0.554392	**0.594925**	0.486529	**0.530137**
Prod	**0.787787**	0.716175	**0.796600**	0.716175	**0.733623**	0.642071	**0.793326**	0.728892	**0.777834**	0.700829
Alt	0.625565	**0.950771**	0.616586	**0.968720**	0.668681	**0.967832**	0.657416	**0.970172**	0.642062	**0.964374**
	Acurácia									
	Extra Trees		Random Forest		SVM		XGBoost		Média Geral	
	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados
Flo	**0.66832**	0.048272	**0.63118**	0.050124	**0.68674**	0.055680	**0.63128**	0.051986	**0.65438**	0.051516
Aca	**0.887364**	0.883651	**0.891058**	0.887364	0.879977	**0.894721**	**0.883661**	0.879977	0.885515	**0.886428**
Int	0.805893	**0.813331**	0.802220	**0.811490**	0.807755	**0.804072**	0.791130	**0.805944**	0.801750	**0.808709**
C/L	**0.966594**	0.961039	**0.966594**	0.962891	0.966594	**0.966594**	**0.966594**	0.962891	**0.966594**	0.963354
Prod	**0.896501**	0.868785	**0.896511**	0.868785	0.894659	0.868785	**0.889114**	0.868785	**0.894196**	0.868785
Alt	0.955483	**0.968498**	0.957345	**0.968508**	0.957345	0.968487	0.957345	**0.972191**	0.956880	**0.969421**

Em comparação aos resultados que utilizaram todos os SNPs, foram testados para cada algoritmo a performance utilizando apenas os SNPs selecionados como importantes. Observa-se que os algoritmos treinados com SNPs selecionados por esse método apresentam alguns resultados melhores, levando em consideração a métrica MAE, isto é, para a média geral de performance da métrica MAE, os SNPs selecionados tiveram menor erro para todos os fenótipos. Para MSE e RMSE, os SNPs selecionados mostraram performance superior para os fenótipos acamamento, comprimento por largura do grão, floração e percentual de grãos inteiros, enquanto que os demais mesmo tendo resultados inferiores, apresentaram diferenças próximas.

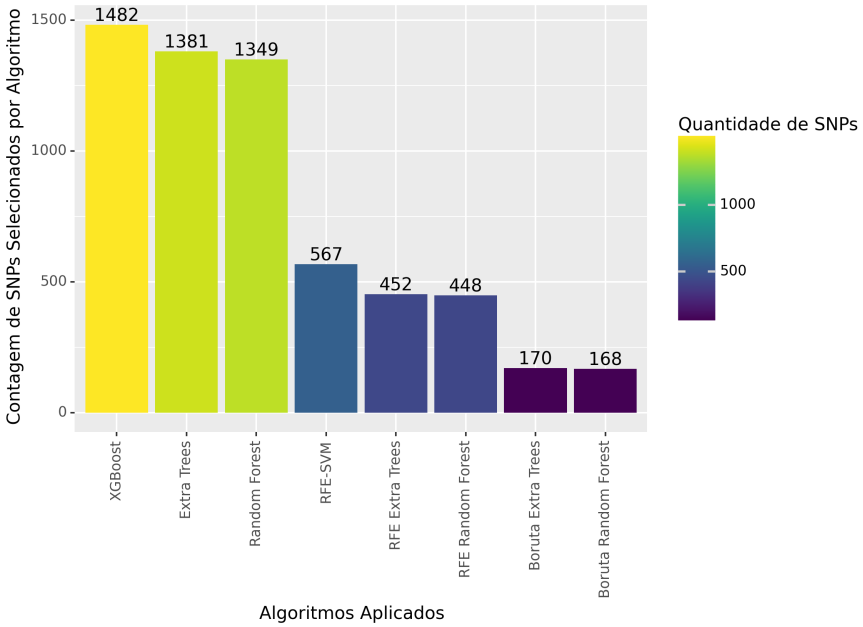
Tabela 20 – Métricas de erro para tarefa de regressão - FSDCD

	Erro Médio Quadrático - MSE									
	Extra Trees		Random Forest		SVM		XGBoost		Média Geral	
	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados
Aca	0.090126	**0.068455**	**0.053177**	0.054611	**0.052703**	0.053206	**0.059404**	0.073343	0.063853	**0.062404**
Alt	**2.713404**	4.163568	**1.828855**	1.958846	1.665701	**1.585413**	**2.096854**	2.352281	**2.076204**	2.515027
C/L	0.901985	**0.433996**	0.505537	**0.388066**	0.491065	**0.374816**	0.552142	**0.419000**	0.612682	**0.403970**
Flo	1.110346	**0.826391**	0.668212	**0.645139**	0.682541	**0.675460**	0.833779	**0.759325**	0.823719	**0.726579**
Int	10.334761	**6.970193**	5.868056	**5.811616**	5.939182	**5.694367**	7.201450	**6.825759**	7.335862	**6.325484**
Prod	44702.461190	**39488.114568**	**25056.933943**	30492.168199	43058.535631	**40461.436752**	**28905.274906**	35681.936871	**35430.801418**	36530.914098
	Raíz do Erro Médio Quadrático - RMSE									
	Extra Trees		Random Forest		SVM		XGBoost		Média Geral	
	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados
Aca	0.300211	**0.261640**	**0.230602**	0.233690	**0.229572**	0.230664	**0.243730**	0.270819	0.252691	**0.249808**
Alt	**1.647241**	2.040482	**1.352352**	1.399588	1.290621	**1.259132**	**1.448052**	1.533715	**1.440904**	1.585884
C/L	0.949729	**0.658784**	0.711011	**0.622949**	0.700760	**0.612223**	0.743062	**0.647302**	0.782740	**0.635586**
Flo	1.053730	**0.909061**	0.817442	**0.803206**	0.826160	**0.821864**	0.913115	**0.871392**	0.907590	**0.852396**
Int	3.214772	**2.640112**	2.422407	**2.410729**	2.437044	**2.386287**	2.683552	**2.612615**	2.708480	**2.515051**
Prod	211.429566	**198.716166**	**158.293822**	174.620068	207.505507	**201.150284**	**170.015514**	188.896630	**188.230713**	191.130621
	Erro Médio Absoluto - MAE									
	Extra Trees		Random Forest		SVM		XGBoost		Média Geral	
	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados
Aca	0.218912	**0.186118**	0.174872	**0.174285**	0.179680	**0.166678**	**0.179553**	0.187322	0.188254	**0.178601**
Alt	0.893233	**0.822604**	**0.666235**	0.695677	0.627249	**0.604372**	0.791045	**0.767495**	0.744440	**0.722537**
C/L	0.572385	**0.385334**	0.413288	**0.365482**	0.367873	**0.329889**	0.434424	**0.374666**	0.446993	**0.363843**
Flo	0.523133	**0.463506**	0.503896	**0.448829**	0.430634	**0.394256**	0.530098	**0.481480**	0.496940	**0.447018**
Int	2.548171	**1.953744**	1.849679	**1.814453**	1.872711	**1.791485**	2.066189	**1.941935**	2.084187	**1.875404**
Prod	139.432268	**109.394981**	**94.867566**	99.361591	120.414477	**117.488619**	108.448872	**105.841046**	115.790796	**108.021559**

6.2.4 FS Discretizada Consensual Iterativa - FSDCI

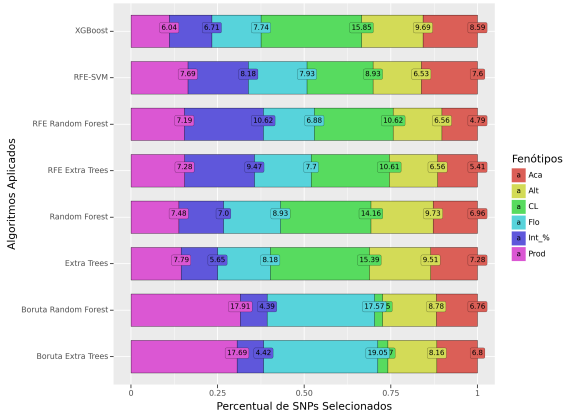
A execução do método FSDCI selecionou as quantidades de SNPs apresentadas na Figura 23. Os algoritmos que produziram o maior número de SNP foram os baseados em árvores de decisão, isto é, XGB, RF e *Extra Trees*. Os algoritmos que utilizaram RFE tiveram resultados mais restritivos, isto é, menor quantidade total de SNPs selecionados. Por fim, os algoritmos que utilizaram Boruta para seleção de SNPs foram os que apresentaram menor número de SNPs selecionados.

Figura 23 – Soma de SNPs selecionados - FSDCI



Além disso, é possível observar na Figura 24 que as distribuições de SNPs entre os fenótipos foi regularmente espaçada para todos os algoritmos, com exceção dos algoritmos baseados na abordagem Boruta, que apresentaram maior quantidade percentual de SNPs selecionados para os atributos produtividade e Floração.

Figura 24 – Percentual e quantidades de SNPs selecionados por fenótipo - FSDCI



A Tabela 21 exibe o número de SNPs selecionados como importantes para cada fenótipo. São exibidos o número SNPs selecionados simultaneamente pelos algoritmos, através do método FSDCI. A disposição dos dados é a mesma da apresentada para a Tabela 18. É possível observar que todos os fenótipos tiveram SNPs selecionados consensualmente por pelo menos cinco algoritmos, com exceção do acamamento que teve um conjunto de SNPs apontados como importantes por até quatro algoritmos. Além disso, as quantidades de SNPs selecionados por mais que três algoritmos apresentaram totais inferiores a 50 SNPs para todos os fenótipos.

Tabela 21 – SNPs selecionados consensualmente - FSDCI

Fenótipo	**Aca**				**Alt**				**CL**				**Flo**				**Int**				**Prod**			
Número de Algoritmos	1	2	3	4	2	3	4	5	2	3	4	5	2	3	4	5	2	3	4	5	2	3	4	5
Número de SNPs Selecionados	588	72	5	19	142	18	17	6	309	37	8	1	102	7	49	1	114	12	9	2	79	6	47	2
Soma Acumulada	684	96	**24**		183	41	**23**		357	48	**9**		159	57	**50**		137	23	**11**		134	55	**49**	

No que diz respeito às métricas de performance, é observado na Tabela 22 que os conjuntos de SNPs selecionados pela abordagem FSDCI obtiveram resultados de acurácia e AUC-ROC médios próximos aos dos algoritmos que utilizaram todos os SNPs. Estão destacados para cada algoritmo, os melhores valores de performance, observados, demonstrando se para determinado fenótipo o conjunto total de SNPs apresenta melhor resultado. Todavia, para os fenótipos altura e percentual de grãos inteiros, melhoras significativas de acurácia e AUC-ROC médias foram observadas.

Tabela 22 – Métricas de performance - FSDCI

	AUC-ROC									
	Extra Trees		**Random Forest**		**SVM**		**XGBoost**		**Média Geral**	
	Base	**SNPs Selecionados**	**Base**	**SNPs Selecionados**	**Base**	**SNPs Selecionados**	**Base**	**SNPs Selecionados**	**Base**	**SNPs Selecionados**
Flo	0.849545	**0.840408**	**0.851042**	0.834602	0.829537	**0.834701**	**0.846033**	0.818649	**0.844039**	0.832090
Aca	0.744748	**0.776234**	0.755244	**0.811144**	0.774284	**0.796191**	0.752780	**0.786979**	0.756764	**0.792637**
Int	0.682260	**0.682676**	0.669390	**0.729334**	0.692052	**0.781049**	0.671640	**0.697755**	0.678836	**0.722704**
C/L	0.448110	**0.484163**	**0.466799**	0.394794	0.457404	**0.552760**	**0.554392**	0.439575	**0.481676**	0.467823
Prod	**0.820752**	0.651518	**0.787886**	0.658839	**0.694223**	0.689799	**0.793326**	0.675563	**0.774047**	0.668930
Alt	0.644668	**0.931387**	0.647127	**0.948470**	0.525515	**0.956768**	0.657416	**0.953564**	0.618681	**0.947547**
	Acurácia									
	Extra Trees		**Random Forest**		**SVM**		**XGBoost**		**Média Geral**	
	Base	**SNPs Selecionados**	**Base**	**SNPs Selecionados**	**Base**	**SNPs Selecionados**	**Base**	**SNPs Selecionados**	**Base**	**SNPs Selecionados**
Flo	**0.59414**	0.048262	**0.64960**	0.050114	**0.68674**	0.051976	**0.63128**	0.044548	**0.64044**	0.048725
Aca	0.887364	**0.894731**	**0.889216**	**0.894741**	0.879977	**0.903980**	**0.883661**	0.889196	0.885055	**0.895662**
Int	**0.804062**	0.779998	**0.804072**	**0.787385**	0.807755	**0.809628**	0.791130	**0.781860**	**0.801755**	0.789718
C/L	0.966594	0.966594	0.966594	0.966594	0.966594	0.966594	0.966594	0.966594	**0.966594**	0.966594
Prod	**0.902046**	0.889094	**0.898353**	**0.896480**	0.894659	**0.907592**	**0.889114**	0.879834	**0.896043**	0.893250
Alt	0.955483	**0.964763**	0.957345	**0.962922**	0.957345	**0.961070**	0.957345	**0.966636**	0.956880	**0.963848**

Por fim, a Tabela 23 apresenta métricas de erro, comparando 4 algoritmos nas tarefas de regressão. Foram utilizados, para cada treino o conjunto total de SNPs, denotado pela coluna Base e os SNPs selecionados pelo método. Foram destacados os menores valores de erro. Ao mensurar os SNPs selecionados pelo método FSDCI em tarefas de regressão sobre os fenótipos presentes no conjunto de dados original, foi constatado que os SNPs selecionados apresentaram resultados melhores para quatro fenótipos utilizando

as métricas MSE e RMSE. Em relação a MAE, cinco dentre os seis fenótipos obtiveram resultados superiores utilizando os SNPs selecionados pelo método.

Tabela 23 – Métricas de erro para tarefa de regressão - FSDCI

	Erro Médio Quadrático - MSE									
	Extra Trees		Random Forest		SVM		XGBoost		Média Geral	
	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados
Flo	1.106100	**0.945809**	**0.669326**	0.685583	**0.682541**	**0.694103**	**0.833779**	0.850943	0.822936	**0.794110**
Aca	0.090589	**0.063892**	**0.053789**	0.056229	**0.052703**	**0.053081**	**0.059404**	0.059983	0.064121	**0.058296**
Int	10.412331	**8.330314**	**5.774933**	6.443013	5.939182	**5.376203**	**7.201450**	7.862017	7.331974	**7.002887**
C/L	0.901168	**0.751772**	0.495530	**0.336718**	0.491065	**0.392891**	0.552142	**0.512475**	0.609976	**0.498464**
Prod	44185.927831	**43741.473968**	**25641.685484**	**30726.370534**	43058.535631	40796.421504	**28905.274906**	**38497.168906**	**35447.855963**	38440.358728
Alt	**2.710104**	4.588862	**1.827004**	2.297863	**1.665701**	**1.768674**	2.096854	2.656982	**2.074916**	2.828095
	Raíz do Erro Médio Quadrático - RMSE									
	Extra Trees		Random Forest		SVM		XGBoost		Média Geral	
	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados
Flo	1.051713	**0.972527**	**0.818124**	0.827999	**0.826160**	0.833129	**0.913115**	0.922466	0.907158	**0.891128**
Aca	0.300981	**0.252769**	**0.231924**	0.237126	0.229572	**0.230393**	0.243730	**0.244914**	0.253222	**0.241446**
Int	3.226814	**2.886228**	2.403109	**2.538309**	2.437044	**2.318664**	2.683552	**2.803929**	2.707762	**2.646297**
C/L	0.949299	**0.867048**	0.703939	**0.580274**	0.700760	**0.626810**	0.743062	**0.715874**	0.781010	**0.706020**
Prod	210.204491	**209.144625**	160.130214	**175.289391**	207.505507	**201.981240**	**170.015514**	196.206954	**188.276010**	196.062130
Alt	**1.646239**	2.142163	**1.351667**	1.515870	**1.290621**	1.329915	**1.448052**	1.630025	**1.440457**	1.681694
	Erro Médio Absoluto - MAE									
	Extra Trees		Random Forest		SVM		XGBoost		Média Geral	
	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados	Base	SNPs Selecionados
Flo	0.521783	**0.457066**	0.500190	**0.432746**	0.430634	**0.375257**	0.530098	**0.480566**	0.495676	**0.436409**
Aca	0.218462	**0.186581**	**0.174454**	0.182578	0.179680	**0.173196**	**0.179553**	0.185730	0.188037	**0.182022**
Int	2.556755	**2.155016**	**1.834475**	1.923776	1.872711	**1.757266**	**2.066189**	2.090211	2.082533	**1.981567**
C/L	0.574200	**0.527086**	0.407218	**0.358894**	0.367873	**0.345084**	0.434424	**0.424920**	0.445929	**0.413996**
Prod	138.152207	**119.232545**	96.072803	**101.925130**	120.414477	**118.059372**	108.448872	**114.000450**	115.772090	**113.304374**
Alt	**0.894602**	0.987551	**0.669320**	0.805092	**0.627249**	0.652594	**0.791045**	0.870425	**0.745554**	0.828916

6.2.5 Comparação entre os métodos propostos

Nessa subseção são apresentados os resultados de comparação entre os métodos propostos para seleção de SNPs. Além disso, são exibidas as quantidades de SNPs que foram consensualmente selecionados entre os diferentes métodos, explorando a similaridade entre as escolhas. Por fim, é apresentado um resumo sobre o tempo de execução para cada método.

A Tabela 24 exibe a média geral das métricas de performance dos algoritmos aplicados sobre cada uma das características fennotípicas para a tarefa de regressão. É possível observar que para quatro dentre os seis fenótipos, o método FSCI apresentou melhor performance, isto é, métricas de erro menores que os demais métodos. Foram destacados os menores valores de erro, permitindo comparar qual método apresentou menor taxa de erro. Contudo, mesmo em relação aos fenótipos de comprimento por largura do grão e percentual de grãos inteiros, onde o método FSCD apresentou melhores resultados, a diferença entre as métricas médias de erro foram baixas, indicando similaridade na performance desses métodos.

Tabela 24 – Comparação das métricas de erro selecionados pelos métodos propostos

Fenótipos	MSE				RMSE				MAE			
	FSCD	FSCI	FSDCD	FSDCI	FSCD	FSCI	FSDCD	FSDCI	FSCD	FSCI	FSDCD	FSDCI
Aca	0.062530	**0.053288**	0.062404	0.058296	0.250061	**0.230841**	0.249808	0.241446	0.178911	**0.167011**	0.178601	0.182022
Alt	2.508732	**2.309420**	2.515027	2.828095	1.583898	**1.519678**	1.585884	1.681694	0.722508	**0.691719**	0.722537	0.828916
C/L	**0.400504**	0.416908	0.403970	0.498464	**0.632854**	0.645684	0.635586	0.706020	**0.362449**	0.367802	0.363843	0.413996
Flo	0.726755	**0.702539**	0.726579	0.794110	0.852499	**0.838176**	0.852396	0.891128	0.447179	**0.416033**	0.447018	0.436409
Int	**6.314410**	6.570714	6.325484	7.002887	**2.512849**	2.563340	2.515051	2.646297	**1.874883**	1.914912	1.875404	1.981567
Prod	36650.871882	**34068.240923**	36530.91409	38440.358728	191.444174	**184.575841**	191.130621	196.062130	107.981505	**103.247068**	108.021559	113.304374

Ainda nesse contexto, a Tabela 25 exibe as quantidade de SNPs selecionados para cada fenótipo após a filtragem consensual em cada um dos quatro métodos. São apresentadas as quantidades de SNPs que foram selecionados em cada fenótipo. São exibidas a quantidade selecionada por cada um dos métodos. A terceira coluna exibe a quantidade de SNPs comuns selecionados para os métodos que utilizaram os fenótipos numéricos, isto é, na primeira linha da terceira coluna, por exemplo, observamos que houveram 13 SNPs. Isso significa que dentre os SNPs selecionados pelos métodos FSCD e FSCI, 13 foram comuns aos dois métodos. A sexta coluna apresenta a mesma lógica de visualização, contudo para os métodos FSDCD e FSDCI. Por fim, na última coluna observamos que, com exceção dos fenótipos percentual de grãos inteiros e comprimento por largura, os demais apresentaram SNPs em comum para todos os métodos.

Tabela 25 – Comparação da quantidade de SNPs selecionados por característica fenotípica

Fenótipos	**Quantidade de SNPs por Método**						
	FSCD	**FSCI**	**FSCD e FSCI**	**FSDCD**	**FSDCI**	**FSDCD e FSDCI**	**Todos**
Aca	22	38	13	15	19	11	6
Alt	39	41	26	17	23	12	6
C/L	13	3	1	11	9	0	0
Flo	17	50	15	38	50	31	4
Int	12	17	7	9	11	0	0
Prod	36	46	32	2	49	2	1

Outro fator avaliado na execução conjunta dos algoritmos é o tempo utilizado por cada algoritmo para realizar a seleção dos SNPs. A Tabela 26 mostra o tempo médio de execução de uma rodada de seleção de SNPs para cada algoritmo em relação aos fenótipos em que o algoritmo foi executado, ou seja, o tempo utilizado para que determinado algoritmo fosse executado e selecionasse um conjunto de SNPs importantes. São exibidos os valores observados para a estratégia utilizada sobre os fenótipos originais, selecionados com algoritmos de regressão e também os fenótipos discretizados, selecionados com os algoritmos de classificação. Os valores informados estão em segundos. Sendo assim, o tempo total médio de execução dos métodos que FSDCI e FSCI é calculado pela soma dos tempos de todos os algoritmos executados sobre todos os fenótipos, multiplicada pelo número de vezes em que os algoritmos foram executados. Além disso, vale ressaltar que os métodos que utilizaram algoritmos de classificação apresentaram tempo médio de execução menor que as mesmas versões para a tarefa de regressão.

Em suma, o método FSCI apresentou melhores resultados de performance referentes à tarefa de regressão para os fenótipos acamamento, altura, floração e produtividade. Esse método teve métricas de erro significativamente próximas ao método FSCD, que teve os menores valores de erro para os outros dois fenótipos, assim como apresentado na Tabela 24. Além disso, em relação aos demais métodos, o FSCI apresentou maior número de SNPs

Tabela 26 – Tempo de treino para os algoritmos executados

FS com Algoritmos de Regressão									
Fenótipo	**Extra Trees**	**Random Forest**	**XGBoost**	**RFE-SVM**	**RFE Extra Trees**	**RFE-RF**	**Boruta Random Forest**	**Boruta Extra Trees**	**Todos**
Aca	2,60	13,93	53,99	1,03	6,87	17,72	86,29	56,70	239,13
Flo	2,89	6,01	9,78	1,27	5,32	11,48	63,84	59,66	160,25
Prod	4,73	3,12	20,37	1,37	11,75	6,82	49,54	63,46	161,16
Int	79,83	4,86	46,43	1,41	8,82	14,59	49,00	2069,67	2274,61
C/L	1,79	3,06	29,89	1,18	3,81	6,17	19,75	84,06	149,71
Alt	11,77	5,80	37,17	1,23	4,65	13,10	54,69	519,12	647,53
Todos	103,61	36,78	197,63	7,49	41,22	69,88	323,11	2859,67	3.632,39
FS com Algoritmos de Classificação									
Aca	7.70	4.20	20.21	1.71	12.48	20.16	61.54	75.88	203.88
Flo	6.54	1.78	23.77	1.46	11.07	18.13	57.27	33.04	153.06
Prod	1.39	5.72	21.20	1.91	1.76	24.41	39.80	40.65	136.84
Int	2.34	3.78	7.31	1.71	11.63	17.90	40.47	48.77	133.91
C/L	5.70	3.15	23.52	1.03	10.23	16.78	44.40	34.65	139.46
Alt	4.75	2.30	18.45	1.06	8.07	7.15	33.76	34.45	109.99
Todos	28.42	20.93	114.46	8.88	55.24	104.53	277.24	267.44	877.14

selecionados consensualmente por pelo menos quatro algoritmos. Outrossim, esse método apresentou alguns SNPs semelhantes ao método FSCD. No entanto, uma ressalva quanto ao FSCI é seu tempo de execução, que é o maior entre todos os métodos apresentados. Porém, levando em conta *pipelines* de bioinformática aplicados com outros propósitos, como pipelines de análise NGS (LEIPZIG, 2017), o tempo apresentado encontra-se em uma faixa esperada. Portanto, para etapa de seleção de SNPs, no *pipeline* proposto, selecionamos o método FSCI, uma vez que apresentou bons resultados referentes aos SNPs selecionados, não necessitando da etapa de discretização de atributos. Sendo assim, para as etapas de comparação com métodos da literatura e de exploração de vias biológicas, seguimos com os SNPs encontrados pelo método FSCI.

6.3 Comparação com métodos da literatura

Nesta seção são apresentados os resultados da comparação entre o método FSCI, proposto para seleção SNPs, em contraste com alguns métodos da literatura. Três grupos de métodos foram comparados utilizando os SNPs selecionados para treinar os algoritmos para tarefas de regressão. Os grupos de métodos usados na comparação são listados a seguir:

- ❏ Métodos Baseados em RFE: Grupo de métodos que utilizam RFE (CHEN; JEONG, 2007) para seleção de atributos. Composto pelos métodos RFE-SVM (SANZ et al., 2018), RFE-RF e RFE em conjunto com Extra Trees;

- ❏ Métodos Baseados em RBAs: Grupo de métodos baseados em Relief (KIRA; RENDELL, 1992). Nesse grupo cinco métodos foram comparados, a saber, ReliefF (KONONENKO, 1994), SURF (GREENE et al., 2009), SURF* (GREENE et al., 2010), MultiSURF (URBANOWICZ et al., 2018b) e MultiSURF* (GRANIZO-MACKENZIE; MOORE, 2013);

- ❏ Métodos Baseados em Boruta: Grupo de métodos que fazem uso do Boruta (KURSA; JANKOWSKI; RUDNICKI, 2010) para seleção de atributos. Aqui o método Boruta foi utilizado em conjunto com RF (implementação original) e Extra Trees (JAMEI et al., 2023).

Os métodos presentes em cada grupo foram escolhidos pela sua ampla utilização em trabalhos como (URBANOWICZ et al., 2018b) e (KUHN; JOHNSON, 2019), além de outros trabalhos da literatura. Além disso, esses métodos possuem implementações nativas em linguagem Python nos pacotes scikit-learn (PEDREGOSA et al., 2011) e ReBATE (URBANOWICZ et al., 2018b).

Antes da apresentação dos resultados das comparações realizados, vale ressaltar que as Tabelas 27, 28 e 29 apresentam a média de cada uma das métricas de erro (MSE, RMSE e MAE), observadas nos quatro algoritmos para determinado fenótipo. Isso pode ser ilustrado da seguinte forma, se determinado método (por exemplo, ReliefF) obteve na predição de um dos seis fenótipo (por exemplo, acamamento) os valores de MSE 0.035, 0.045, 0.05 e 0.04, o valor de MSE apresentado na tabela (coluna ReliefF e linha Aca) será de 0.0425. Os resultados são apresentados dessa forma a fim de gerar uma comparação visualmente mais simples. Contudo, os resultados completos de cada um dos métodos testados contendo as métricas MSE, RMSE e MAE observadas em cada método para cada um dos quatro algoritmos utilizados na comparação estão disponíveis através dos links do Apêndice B.

A Tabela 27 apresenta os resultados das médias das métricas de erro observadas em cada método de seleção em relação aos respectivos fenótipos. É exibido o grupo de algoritmos que utilizam Eliminação Recursiva de Atributos (do inglês, "*Recursive Feature Elimination*") (RFE). Para simplificar a visualização dos melhores resultados, estão destacados os menores valores de erro para cada fenótipo. É possível notar que o método FSCI obteve menores valores de MSE para quatro dentre os seis fenótipos. Além disso, em relação ao MAE, ele obteve melhores resultados, isto é, menor valor de erro, para cinco dentre os seis fenótipos. Esse fato indica efetividade desse método em relação a outros métodos da literatura, no que diz respeito a performance.

Seguindo com as comparações, a Tabela 28 apresenta os resultados das médias das métricas de erro observadas em cada método de seleção em relação aos respectivos fenótipos. É exibido o grupo de algoritmos que utilizam Algoritmos Baseados em Relief (do inglês, "*Relief Based Algorithms*") (RBA)s. Para simplificar a visualização dos melhores resultados, estão destacados os menores valores de erro para cada fenótipo. É possível notar que o método FSCI quando comparado aos outros cinco métodos, obteve menores valores de MSE para cinco dentre os seis fenótipos. Em relação ao MAE, o método FSCI obteve melhores resultados para cinco fenótipos.

Por fim, a Tabela 29 apresenta os resultados das médias das métricas de erro observadas

Tabela 27 – Comparação com métodos que utilizam RFE

	MSE			
	RFE-SVM	**RFE-RF**	**RFE-Extra Trees**	**FSCi**
Aca	0.073689	0.062815	0.057370	**0.053288**
Alt	1.573507	1.926395	**1.202786**	2.309420
CL	1.018711	0.756503	0.419807	**0.416908**
Flo	1.157541	**0.685608**	0.758261	0.702539
Int	10.695646	6.822437	7.807832	**6.570714**
Prod	48542.688339	41913.110786	46258.303793	**34068.240923**
	RMSE			
	RFE-SVM	**RFE-RF**	**RFE-Extra Trees**	**FSCi**
Aca	0.271456	0.250629	0.239521	**0.230841**
Alt	1.254395	1.387946	**1.096716**	1.519678
CL	1.009312	0.869772	0.647925	**0.645684**
Flo	1.075891	**0.828014**	0.870782	0.838176
Int	3.270420	2.611979	2.794250	**2.563340**
Prod	220.324053	204.726918	215.077437	**184.575841**
	MAE			
	RFE-SVM	**RFE-RF**	**RFE-Extra Trees**	**FSCi**
Aca	0.196063	0.191810	0.180883	**0.167011**
Alt	0.805620	0.960072	0.808714	**0.691719**
CL	0.486091	0.451320	**0.349165**	0.367802
Flo	0.595230	0.514947	0.544203	**0.416033**
Int	2.581251	2.066179	2.091230	**1.914912**
Prod	136.577007	121.884064	136.817917	**103.247068**

Tabela 28 – Comparação com métodos RBAs

	MSE					
	ReliefF	**SURF**	**SURF***	**Multi SURF**	**Multi SURF***	**FSCI**
Aca	0.081196	0.064883	0.066819	0.068243	0.067904	**0.053288**
Alt	4.412502	4.269439	4.080829	3.998007	2.601487	**2.309420**
C/L	0.516412	0.481730	**0.180518**	0.526594	0.619646	0.416908
Flo	0.808852	0.721350	1.012246	0.769773	0.924005	**0.702539**
Int	7.358718	7.083181	9.052655	8.111631	6.576016	**6.570714**
Prod	39859.543696	36853.198839	39400.093410	41953.010191	54629.204594	**34068.240923**
	RMSE					
	ReliefF	**SURF**	**SURF***	**Multi SURF**	**Multi SURF***	**FSCI**
Aca	0.284949	0.254721	0.258494	0.261233	0.260584	**0.230841**
Alt	2.100596	2.066262	2.020106	1.999502	1.612913	**1.519678**
C/L	0.718618	0.694068	**0.424874**	0.725668	0.787176	0.645684
Flo	0.899362	0.849323	1.006105	0.877367	0.961252	**0.838176**
Int	2.712696	2.661425	3.008763	2.848092	2.564374	**2.563340**
Prod	199.648550	191.971870	198.494568	204.824340	233.728913	**184.575841**
	MAE					
turn	**ReliefF**	**SURF**	**SURF***	**Multi SURF**	**Multi SURF***	**FSCI**
Aca	0.199867	0.184437	0.193891	0.194774	0.185925	**0.167011**
Alt	0.885765	0.978044	0.857216	0.828697	0.933779	**0.691719**
C/L	0.391087	0.430705	**0.310763**	0.429696	0.414459	0.367802
Flo	0.485327	**0.410480**	0.549952	0.446745	0.543680	0.416033
Int	2.037786	1.989498	2.214155	2.102814	1.930053	**1.914912**
Prod	126.039333	116.643293	129.675023	126.457824	146.309554	**103.247068**

em cada método de seleção em relação aos respectivos fenótipos. É exibido o grupo de algoritmos que utilizam Boruta. Para simplificar a visualização dos melhores resultados, estão destacados os menores valores de erro para cada fenótipo. É possível notar que o método FSCI quando comparado aos outros dois métodos, obteve menores valores de MSE para os fenótipos acamamento, altura, comprimento por largura e produtividade. Em relação ao MAE, o método FSCI obteve melhores resultados para cinco fenótipos, excetuando floração. Em suma, os resultados observados nos experimentos de comparação indicam que o método FSCI, em alguns casos, teve resultados de performance ligeiramente superiores a outros métodos presentes na literatura.

Tabela 29 – Comparação com métodos que utilizam Boruta

	MSE		
	Boruta - Extra Trees	**Boruta - RF**	**FSCI**
Aca	0.073583	0.060839	**0.053288**
Alt	3.212400	3.659749	**2.309420**
C/L	0.525599	0.579851	**0.416908**
Flo	**0.583129**	0.837396	0.702539
Int	7.004928	**5.837991**	6.570714
Prod	38204.157233	34432.365128	**34068.240923**
	RMSE		
	Boruta - Extra Trees	**Boruta - RF**	**FSCI**
Aca	0.271262	0.246656	**0.230841**
Alt	1.792317	1.913047	**1.519678**
C/L	0.724982	0.761480	**0.645684**
Flo	**0.763629**	0.915093	0.838176
Int	2.646683	**2.416194**	2.563340
Prod	195.458838	185.559600	**184.575841**
	MAE		
	Boruta - Extra Trees	**Boruta - RF**	**FSCI**
Aca	0.195373	0.178333	**0.167011**
Alt	0.756410	0.695652	**0.691719**
C/L	0.393500	0.394493	**0.367802**
Flo	**0.396171**	0.487406	0.416033
Int	2.049215	**1.883642**	1.914912
Prod	104.115905	**102.249402**	103.247068

6.4 Exploração de vias biológicas

A partir dos SNPs obtidos com o método FSCI, 130 SNPs foram confirmados pelo *software* RiceVarMap (ZHAO et al., 2015). Isso significa que esses SNPs também foram apontados como abrangentes nas variedades do arroz, isto é, são presentes em outras amostras de arroz sequenciadas em estudos de melhoramento genético. Essa etapa auxilia para que não seja necessário realizar um sequenciamento de DNA nas áreas subjacentes do SNP em outras amostras de arroz, a fim de verificar se está de fato presente em outras variedades. Entre os SNPs confirmados foram identificados 38 genes em todos os fenótipos estudados, sendo 28 deles distintos entre si. Entre os 28 genes distintos, 18 estavam

posicionados em regiões exonicas, cinco em regiões 3', dois em regiões 5' e três em regiões de introns. O total de genes encontrados representa que aproximadamente 21.53% dos SNPs selecionados estavam posicionados em regiões intragênicas de interesse. A Tabela 30 exibe as distribuições de genes encontradas para cada fenótipo em cada cromossomo. Não foram encontrados genes posicionado nos cromossomos 4, 5 e 11. Além disso na etapa de seleção incial o FSCI não identificou nenhum SNP posicionado no cromossomo 10, justificando a ausência de genes nesse cromossomo.

Tabela 30 – Quantidade de genes identificados por cromossomo em cada fenótipo

Fenótipos	**Cromossomo**								
	1	**2**	**3**	**6**	**7**	**8**	**9**	**12**	**Todos**
Aca	2	-	-	-	1	1	1	-	5
Alt	-	1	1	2	2	-	1	2	9
C/L	-	-	-	-	1	-	-	-	1
Flo	2	1	-	-	2	2	1	1	9
Int	2	-	-	1	1	-	-	1	5
Prod	2	1	1	-	3	1	1	-	9
Todos	8	3	2	3	10	4	4	4	38

Os genes identificados foram submetidos ao processo de enriquecimento funcional e 38 sistemas biológicos enriquecidos foram identificados entre todos os fenótipos. A Tabela 31 apresenta a distribuição de quantidades de sistemas biológicos enriquecidos em cada fenótipo. Para o arroz estão disponíveis 2 bancos de dados no GProfiler, a saber, GO e KEGG (RAUDVERE et al., 2019). É possível observar que o fenótipo comprimento por largura não apresentou nenhum sistema biológico enriquecido. Além disso, dentre todos os sistemas encontrados no GO, apenas foram identificadas ontologias referentes a funções moleculares.

Tabela 31 – Quantidade de sistemas biológicos enriquecidos

Fenótipos	**Gene Ontology**			**KEGG**	**Todos**
	Molecular Function (MF)	**Biological Process (BP)**	**Cellular Component (CC)**		
Aca	5	-	-	-	5
Alt	2	-	-	-	2
Flo	11	-	-	3	14
Int	4	-	-	3	7
Prod	9	-	-	1	10
Todos	31	0	0	7	38

Ainda nesse contexto, a Tabela 32 apresenta os sistemas biológicos afetados, onde são exibidos o identificador do sistema no respectivo banco de dados, o nome completo, o valor de significância p, o fenótipo que indica em quais genes o sistema está enriquecido, e o banco de dados proveniente. Vale ressaltar que os prefixos CC, BP e MF denotam as funções biológicas do respectivo sistema biológico apresentado, sendo respectivamente,

componentes celulares, processos biológicos e funções moleculares. Esses resultados obtidos não serão analisados nesse estudo, uma vez que fogem do escopo geral do trabalho que objetiva o desenho e demonstração dos resultados do *pipeline* proposto. Contudo, essas informações foram submetidas ao núcleo de investigação da Embrapa Arroz e Feijão que seguirá com as investigações a partir dos dados gerados.

Essa etapa de enriquecimento de vias acrescenta ao pipeline um filtro de validação que antecede a experimentação laboratorial, isto é, a Tabela 32 fornece uma lista de sistemas biológicos disponíveis em bancos de dados conhecidos que podem corroborar com investigações futuras ou desencadear novos experimentos de validação, possibilitando por exemplo, investigações da influência dos sistemas apresentados em genes de interesse do arroz, seja na expressão gênica ou maquinaria biológica que afeta o acamamento dos grãos. Isto oferece um alvo mais certo à pesquisa de validação.

Tabela 32 – Lista de sistemas biológicos enriquecidos

Fenótipo	Identificador	Nome	p-valor	Banco de Dados
Aca	GO:0016877	ligase activity, forming carbon-sulfur bonds	0.042162	GO:MF
	GO:0016405	CoA-ligase activity	0.039791	GO:MF
	GO:0016878	acid-thiol ligase activity	0.039791	GO:MF
	GO:0106290	trans-cinnamate-CoA ligase activity	0.039791	GO:MF
	GO:0016207	4-coumarate-CoA ligase activity	0.039791	GO:MF
Alt	GO:0052793	pectin acetylesterase activity	0.002474	GO:MF
	GO:0052689	carboxylic ester hydrolase activity	0.027388	GO:MF
Flo	KEGG:00230	Purine metabolism	0.042400	KEGG
	KEGG:01232	Nucleotide metabolism	0.042400	KEGG
	GO:0016877	ligase activity, forming carbon-sulfur bonds	0.039151	GO:MF
	GO:0016405	CoA-ligase activity	0.032844	GO:MF
	GO:0106290	trans-cinnamate-CoA ligase activity	0.032844	GO:MF
	GO:0016836	hydro-lyase activity	0.046781	GO:MF
	GO:0019205	nucleobase-containing compound kinase activity	0.032844	GO:MF
	GO:0016878	acid-thiol ligase activity	0.032844	GO:MF
	GO:0016776	phosphotransferase activity, phosphate group a...	0.032844	GO:MF
	GO:0016207	4-coumarate-CoA ligase activity	0.032844	GO:MF
	GO:0004089	carbonate dehydratase activity	0.032844	GO:MF
	GO:0004017	adenylate kinase activity	0.032844	GO:MF
	GO:0050145	nucleoside monophosphate kinase activity	0.032844	GO:MF
	KEGG:00730	Thiamine metabolism	0.032305	KEGG
Int	GO:0004017	adenylate kinase activity	0.043513	GO:MF
	GO:0050145	nucleoside monophosphate kinase activity	0.043513	GO:MF
	GO:0016717	oxidoreductase activity, acting on paired dono...	0.043513	GO:MF
	GO:0016705	oxidoreductase activity, acting on paired dono...	0.043513	GO:MF
	KEGG:01232	Nucleotide metabolism	0.042400	KEGG
	KEGG:00730	Thiamine metabolism	0.032305	KEGG
	KEGG:00230	Purine metabolism	0.042400	KEGG
Prod	KEGG:00310	Lysine degradation	0.019902	KEGG
	GO:0008276	protein methyltransferase activity	0.040935	GO:MF
	GO:0016878	acid-thiol ligase activity	0.037896	GO:MF
	GO:0140993	histone modifying activity	0.045150	GO:MF
	GO:0016405	CoA-ligase activity	0.037896	GO:MF
	GO:0106290	trans-cinnamate-CoA ligase activity	0.037896	GO:MF
	GO:0042054	histone methyltransferase activity	0.037896	GO:MF
	GO:0052793	pectin acetylesterase activity	0.037094	GO:MF
	GO:0016877	ligase activity, forming carbon-sulfur bonds	0.040935	GO:MF
	GO:0016207	4-coumarate-CoA ligase activity	0.037896	GO:MF

6.5 Discussões e destaques

Retomando os pontos avaliados nos experimentos, é possível apontar três grupos de contribuições diretas referentes à elaboração do *pipeline* proposto, a saber, a necessidade de uma etapa de pré-processamento dos dados, a junção de algoritmos de FS e análise exploratória de SNPs. Quanto à primeira etapa, a partir do conjunto de dados utilizado não foi possível obter grandes informações quanto à eficácia do processo de transformação e uniformização dos dados a partir do One-Hot Enconding, pois de fato não houve SNPs com mais de duas variações de valores. Contudo, no que diz respeito à discretização dos dados para utilização de algoritmos de classificação, os resultados observados pelos métodos que utilizaram fenótipos discretizados nesse estudo, apontaram que a seleção de SNPs não teve ganhos de performance relevantes no que diz respeito as métricas de avaliação empregadas.

Tabela 33 – Destaques dos experimentos realizados

Etapa	**Descrição**	**Destaques**	
		Positivos	**Negativos**
Pré-processamento	*One Hot Encoding*	- Não foi observado	- Não foi observado
	Discretização	- Não foi observado	- Etapa extra de discretização
Seleção de SNPs	*FSCD*	- Melhor resultado para 2 de 6 fenótipos - Melhor tempo de execução que FSCI	- Não obteve resultados de performance melhores que FSCI
	FSCI	- Melhor resultado para 4 de 6 fenótipos - Apresentou bons resultados em algumas comparações com outros métodos	- É que se torna lento de acordo com o número de iterações
	FSDCD	- Melhor tempo de execução que os demais	- Não obteve resultados de performance melhores que FSCI
	FSDCI	- Melhor tempo de execução que FSCI e FSCD	- Mais lento que FSDCD - Etapa extra de discretização
Exploração de SNPs Selecionados	*Cruzamento com outras plataformas (RiceVarMap)*	- Possibilitou identificar SNPs já catalogados	- Não há modo simples para verificação direta
	Enriquecimento de Vias	- Oferece possibilidade de verificar possíveis sistemas afetados	- Não foi observado

No que diz respeito à combinação de algoritmos de FS para seleção de SNPs em arroz, é possível notar que se trata de uma abordagem promissora para seleção de SNPs, pois o método FSCI obteve bons resultados quando comparados com outros métodos presentes na literatura. Contudo, os resultados observados nos experimentos não oferecem provas conclusivas sobre eficácia dos fenótipos numéricos em relação aos fenótipos discretizados, sendo que se tratam de resultados empíricos observados no conjunto de dados avaliado.

Quanto à análise exploratória de SNPs, nesse trabalho foi proposta a utilização da plataforma RiceVarMap (ZHAO et al., 2015) como ferramenta de validação de SNPs

catalogados. A plataforma confirmou 130 SNPs selecionados pelo método FSCI, evitando uma etapa posterior de sequenciamento. Além disso foi realizado o enriquecimento de vias utilizando o *software* GProfiler (RAUDVERE et al., 2019), brevemente analisado no Capítulo 4 e comparado a outras ferramentas que realizam esse procedimento. O GProfiler apontou vias e sistemas biológicos que possivelmente estão sendo afetados pelos SNPs estudados.

A fim de produzir uma síntese dos pontos levantados e observados nos experimentos realizados, a Tabela 33 apresenta um panorama resumido de cada etapa proposta e avaliada neste trabalho, trazendo destaques positivos e negativos em cada uma das etapas.

Capítulo 7

Considerações Finais e Trabalhos Futuros

Nesse trabalho exploramos a tarefa de identificação de SNPs de arroz associados a múltiplos fenótipos. Dois pontos principais foram estudados: 1) combinação (*ensemble*) de algoritmos de ML modelados para tarefa de FS; e 2) utilização de bancos de dados contendo informação de SNPs validados experimentalmente e aplicação de enriquecimento funcional para exploração de vias biológicas. Nesse contexto, foram propostos procedimentos de sistematização para uma seleção robusta de SNPs, onde quatro métodos de seleção de SNPs foram propostos e avaliados, sendo que dois deles investigaram a discretização de atributos para aplicação em fenótipos categóricos, onde foi exemplificado, que levando em conta o conjunto de dados utilizado, a combinação de algoritmos treinados para classificação, com fenótipos categóricos, tem desempenho ligeiramente pior ao das mesmas versões utilizando fenótipos numéricos. Vale ressaltar que os SNPs sugeridos foram submetidos à investigação e testes experimentais.

A partir desse estudo, conceitos referentes à padronização de métodos para seleção de SNPs foram levantados, tais como a efetividade da combinação de algoritmos de FS para a tarefa de seleção de SNPs de arroz e a utilização do enriquecimento funcional para identificação de vias putativamente afetas. Vale salientar que para esse trabalho a seleção de SNPs foi idealizada para os fenótipos de maneira isolada, sendo que os resultados foram compilados ao final da execução dos algoritmos. Contudo, a modelagem pode ser explorada pensando o problema como uma seleção de atributos multi-objetivo, tal como é citado nos trabalhos de (HASHEMI; DOWLATSHAHI; NEZAMABADI-POUR, 2021) e (SYED et al., 2021), onde algoritmos multi-objetivos foram implementados. Além disso

outros algoritmos podem ser explorados, sendo válido uma análise exaustiva de algoritmos disponíveis na literatura, afim de oferecer uma combinação robusta para seleção de SNPs de interesse.

Outra proposta de melhoria do trabalho realizado é a investigação de métodos de explicação de resultados dos algoritmos de ML utilizados apontando possíveis motivos de escolha de cada SNP. Isso pode favorecer a escolha de SNPs em um procedimento anterior à validação em bancos de dados. Uma sugestão para promover explicação do resultados fornecidos pode ser a utilização de *Shapley Values* (SCAVUZZO et al., 2022), que fornecem sugestões do comportamento de predições com uma estratégia baseada em teoria dos jogos.

Em trabalhos futuros é interessante a realização de uma coleta e avaliação outros conjuntos de dados que possuem fenótipos e genótipos de arroz. Isso permitiria a comparação dos métodos propostos em outros contextos. Além disso, como etapa de simplificação do uso dos métodos, o desenvolvimento de uma ferramenta web que implementa os procedimentos testados simplificaria não somente comparações, mas a navegação por bancos de dados com informações de arroz promovendo estudos consistentes e robustos de seleção de SNPs de maneira integrada, favorecendo a comparação e melhorando significativamente os resultados obtidos. Em conjunto a isso pode-se realizar análises complementares utilizandos características dos conjuntos de dados, de modo a observar se os grupos de experimentos de determinado conjunto de dados apresenta diferenças significativas, evidenciando tendências referentes a localidade do experimento.

Referências

ABADIE, T. et al. Constructing a rice core collection for brazil. **Pesquisa Agropecuaria Brasileira**, SciELO Brasil, v. 40, p. 129–136, 2005.

ABBAS, M.; EL-MANZALAWY, Y. Machine learning based refined differential gene expression analysis of pediatric sepsis. **BMC Medical Genomics**, BioMed Central, v. 13, n. 1, p. 1–10, 2020.

ABOUDI, N. E.; BENHLIMA, L. Review on wrapper feature selection approaches. In: **2016 International Conference on Engineering & MIS (ICEMIS)**. [S.l.: s.n.], 2016. p. 1–5.

ACKERMANN, M.; STRIMMER, K. A general modular framework for gene set enrichment analysis. **BMC Bioinformatics**, Springer, v. 10, n. 1, p. 1–20, 2009.

ADAM, Y. et al. Performing post-genome-wide association study analysis: overview, challenges and recommendations. **F1000Research**, Faculty of 1000 Ltd, v. 10, 2021.

AHA, D. W.; BANKERT, R. L. A comparative evaluation of sequential feature selection algorithms. In: FISHER, D.; LENZ, H.-J. (Ed.). **Pre-proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics**. [S.l.]: PMLR, 1995. (Proceedings of Machine Learning Research, R0), p. 1–7.

AL-TASHI, Q. et al. Approaches to multi-objective feature selection: A systematic literature review. **IEEE Access**, IEEE, v. 8, p. 125076–125096, 2020.

ALBAWI, S.; MOHAMMED, T. A.; AL-ZAWI, S. Understanding of a convolutional neural network. In: IEEE. **2017 International Conference on Engineering and Technology (ICET)**. [S.l.], 2017. p. 1–6.

ALJOUIE, A.; SCHATZ, M.; ROSHAN, U. Machine learning based prediction of gliomas with germline mutations obtained from whole exome sequences from tcga and 1000 genomes project. In: IEEE. **2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)**. [S.l.], 2019. p. 1–8.

ALZUBI, R. et al. A hybrid feature selection method for complex diseases SNPs. **IEEE Access**, IEEE, v. 6, p. 1292–1301, 2017.

ANDONIE, R. Hyperparameter optimization in learning systems. **Journal of Membrane Computing**, Springer, v. 1, n. 4, p. 279–291, 2019.

ARAUZO-AZOFRA, A.; BENITEZ, J. M.; CASTRO, J. L. Consistency measures for feature selection. **Journal of Intelligent Information Systems**, Springer, v. 30, p. 273–292, 2008.

AZMI, S. S.; BALIGA, S. An overview of boosting decision tree algorithms utilizing adaboost and xgboost boosting strategies. **Int. Res. J. Eng. Technol**, v. 7, n. 5, 2020.

BAIRD, N. A. et al. Rapid snp discovery and genetic mapping using sequenced rad markers. **PLOS one**, Public Library of Science San Francisco, USA, v. 3, n. 10, p. e3376, 2008.

BANERJEE, R.; MARATHI, B.; SINGH, M. Efficient genomic selection using ensemble learning and ensemble feature reduction. **Journal of Crop Science and Biotechnology**, Springer, v. 23, p. 311–323, 2020.

BAUER, A. M.; REETZ, T. C.; LÉON, J. Estimation of breeding values of inbred lines using best linear unbiased prediction (blup) and genetic similarities. **Crop Science**, Wiley Online Library, v. 46, n. 6, p. 2685–2691, 2006.

BEULAH, J. R.; PUNITHAVATHANI, D. S. A hybrid feature selection method for improved detection of wired/wireless network intrusions. **Wireless Personal Communications**, Springer, v. 98, n. 2, p. 1853–1869, 2018.

BISCHL, B. et al. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Wiley Online Library, v. 13, n. 2, p. e1484, 2023.

BOLÓN-CANEDO, V.; ALONSO-BETANZOS, A. Ensembles for feature selection: A review and future trends. **Information Fusion**, Elsevier, v. 52, p. 1–12, 2019.

BOLÓN-CANEDO, V. et al. A review of microarray datasets and applied feature selection methods. **Information Sciences**, Elsevier, v. 282, p. 111–135, 2014.

BOTTA, V. et al. Exploiting snp correlations within random forest for genome-wide association studies. **PLOS one**, Public Library of Science San Francisco, USA, v. 9, n. 4, p. e93379, 2014.

BRADBURY, P. J. et al. Tassel: software for association mapping of complex traits in diverse samples. **Bioinformatics**, Oxford University Press, v. 23, n. 19, p. 2633–2635, 2007.

BRAMMER, S. P. Marcadores moleculares: princípios básicos e uso em programas de melhoramento genético vegetal. Passo Fundo: Embrapa Trigo, 2000.

BROWNE, M. W. Cross-validation methods. **Journal of Mathematical Psychology**, Elsevier, v. 44, n. 1, p. 108–132, 2000.

BUENO, L. G. et al. Adaptabilidade e estabilidade de acessos de uma coleção nuclear de arroz. **Pesquisa Agropecuária Brasileira**, SciELO Brasil, v. 47, p. 216–226, 2012.

CARULLI, J. P. et al. High throughput analysis of differential gene expression. **Journal of Cellular Biochemistry**, Wiley Online Library, v. 72, n. S30–31, p. 286–296, 1998.

CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. **Computers & Electrical Engineering**, Elsevier, v. 40, n. 1, p. 16–28, 2014.

CHEN, C.-W. et al. Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. **Expert Systems**, Wiley Online Library, v. 37, n. 5, p. e12553, 2020.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining**. [S.l.: s.n.], 2016. p. 785–794.

CHEN, W. et al. Convergent selection of a wd40 protein that enhances grain yield in maize and rice. **Science**, American Association for the Advancement of Science, v. 375, n. 6587, p. eabg7985, 2022.

CHEN, X.-w.; JEONG, J. C. Enhanced recursive feature elimination. In: IEEE. **Sixth International Conference on Machine Learning and Applications (ICMLA 2007)**. [S.l.], 2007. p. 429–435.

CHICCO, D.; JURMAN, G. A brief survey of tools for genomic regions enrichment analysis. **Frontiers in Bioinformatics**, Frontiers, v. 2, p. 968327, 2022.

CRISTIANINI, N.; SHAWE-TAYLOR, J. et al. **An introduction to Support Vector Machines and other kernel-based learning methods**. [S.l.]: Cambridge University Press, 2000.

CUETO-LÓPEZ, N. et al. A comparative study on feature selection for a risk prediction model for colorectal cancer. **Computer Methods and Programs in Biomedicine**, Elsevier, v. 177, p. 219–229, 2019.

DARST, B. F.; MALECKI, K. C.; ENGELMAN, C. D. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. **BMC Genetics**, BioMed Central, v. 19, n. 1, p. 1–6, 2018.

DOUDNA, J. A.; CHARPENTIER, E. The new frontier of genome engineering with crispr-cas9. **Science**, American Association for the Advancement of Science, v. 346, n. 6213, p. 1258096, 2014.

DRAGO, A. et al. Enrichment pathway analysis. the inflammatory genetic background in bipolar disorder. **Journal of Affective Disorders**, Elsevier, v. 179, p. 88–94, 2015.

DROBNIČ, F. et al. Explained learning and hyperparameter optimization of ensemble estimator on the bio-psycho-social features of children and adolescents. **Electronics**, MDPI, v. 12, n. 19, p. 4097, 2023.

DUBOURG, A. et al. Giardia secretome highlights secreted tenascins as a key component of pathogenesis. **Gigascience**, Oxford University Press, v. 7, n. 3, p. giy003, 2018.

EDWARDS, K.; JOHNSTONE, C.; THOMPSON, C. A simple and rapid method for the preparation of plant genomic dna for pcr analysis. **Nucleic Acids Research**, Oxford University Press, v. 19, n. 6, p. 1349, 1991.

ELSHIRE, R. J. et al. A robust, simple genotyping-by-sequencing (gbs) approach for high diversity species. **PLOS one**, Public Library of Science San Francisco, USA, v. 6, n. 5, p. e19379, 2011.

ENDELMAN, J. B. Ridge regression and other kernels for genomic selection with r package rrblup. **The Plant Genome**, Wiley Online Library, v. 4, n. 3, 2011.

FANG, Z.; LIU, X.; PELTZ, G. Gseapy: a comprehensive package for performing gene set enrichment analysis in python. **Bioinformatics**, Oxford University Press, v. 39, n. 1, p. btac757, 2023.

FERNANDES, M.; HUSI, H. Ora, fcs, and pt strategies in functional enrichment analysis. In: **Proteomics Data Analysis**. [S.l.]: Springer, 2021. p. 163–178.

FLOOD, P. J.; HANCOCK, A. M. The genomic basis of adaptation in plants. **Current Opinion in Plant Biology**, Elsevier, v. 36, p. 88–94, 2017.

FUSHIKI, T. Estimation of prediction error by using k-fold cross-validation. **Statistics and Computing**, Springer, v. 21, p. 137–146, 2011.

GENTZBITTEL, L. et al. Whogem: an admixture-based prediction machine accurately predicts quantitative functional traits in plants. **Genome Biology**, Springer, v. 20, p. 1–20, 2019.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. **Machine Learning**, Springer, v. 63, n. 1, p. 3–42, 2006.

GOLDSTEIN, B. A.; POLLEY, E. C.; BRIGGS, F. B. Random forests for genetic association studies. **Statistical Applications in Genetics and Molecular Biology**, De Gruyter, v. 10, n. 1, 2011.

GOSWAMI, S.; CHAKRABARTI, A. Feature selection: A practitioner view. **International Journal of Information Technology and Computer Science (IJITCS)**, Citeseer, v. 6, n. 11, p. 66, 2014.

GOYAL, J.; SOTA, A.; ARORA, V. Feature selection for loan repayment prediction system using machine learning. **International Journal for Research in Applied Science & Engineering Technology (IJRASET)**, v. 11, 2023.

GRANIZO-MACKENZIE, D.; MOORE, J. H. Multiple threshold spatially uniform relieff for the genetic analysis of complex human diseases. In: SPRINGER. **Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics: 11th European Conference, EvoBIO 2013, Vienna, Austria, April 3-5, 2013. Proceedings 11**. [S.l.], 2013. p. 1–10.

GREENE, C. S. et al. The informative extremes: using both nearest and farthest individuals can improve relief algorithms in the domain of human genetics. In: SPRINGER. **European conference on evolutionary computation, machine learning and data mining in bioinformatics**. [S.l.], 2010. p. 182–193.

_____. Spatially uniform relieff (surf) for computationally-efficient filtering of gene-gene interactions. **BioData Mining**, Springer, v. 2, p. 1–9, 2009.

GÜNEY, H.; ÖZTOPRAK, H. Microarray-based cancer diagnosis: repeated cross-validation-based ensemble feature selection. **Electronics Letters**, Wiley Online Library, v. 54, n. 5, p. 272–274, 2018.

GUO, L. et al. Machine learning-based models for genomic predicting neoadjuvant chemotherapeutic sensitivity in cervical cancer. **Biomedicine & Pharmacotherapy**, Elsevier, v. 159, p. 114256, 2023.

GUYON, I. et al. Gene selection for cancer classification using support vector machines. **Machine Learning**, Springer, v. 46, n. 1, p. 389–422, 2002.

HAEBERLE, L. et al. Predicting triple-negative breast cancer subtype using multiple single nucleotide polymorphisms for breast cancer risk and several variable selection methods. **Geburtshilfe und Frauenheilkunde**, Georg Thieme Verlag KG, v. 77, n. 06, p. 667–678, 2017.

HAILU, T. G.; ABDULKADIR, T. Multidmet: Designing a hybrid multidimensional metrics framework to predictive modeling for performance evaluation and feature selection. **Research Square**, 2023.

HAN, S.; WILLIAMSON, B. D.; FONG, Y. Improving random forest predictions in small datasets from two-phase sampling designs. **BMC Medical Informatics and Decision Making**, BioMed Central, v. 21, n. 1, p. 1–9, 2021.

HANCOCK, J. T.; KHOSHGOFTAAR, T. M. Survey on categorical data for neural networks. **Journal of Big Data**, SpringerOpen, v. 7, n. 1, p. 1–41, 2020.

HANDBOOK, B. Qiagen. **Gmbh, Germany, June**, 2005.

HASAN, M. A. M. et al. Feature selection for intrusion detection using random forest. **Journal of Information Security**, Scientific Research Publishing, v. 7, n. 3, p. 129–140, 2016.

HASHEMI, A.; DOWLATSHAHI, M. B.; NEZAMABADI-POUR, H. Vmfs: A vikor-based multi-target feature selection. **Expert Systems with Applications**, Elsevier, v. 182, p. 115224, 2021.

HOSSIN, M.; SULAIMAN, M. N. A review on evaluation metrics for data classification evaluations. **International Journal of Data Mining & Knowledge Management Process**, Academy & Industry Research Collaboration Center (AIRCC), v. 5, n. 2, p. 1, 2015.

HOWELL, W. M. et al. Dynamic allele-specific hybridization. **Nature Biotechnology**, Nature Publishing Group, v. 17, n. 1, p. 87–88, 1999.

HUANG, J.; LING, C. X. Using auc and accuracy in evaluating learning algorithms. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 17, n. 3, p. 299–310, 2005.

HUANG, X. et al. Genome-wide association studies of 14 agronomic traits in rice landraces. **Nature Genetics**, Nature Publishing Group, v. 42, n. 11, p. 961–967, 2010.

HUNG, J.-H. et al. Gene set enrichment analysis: performance evaluation and usage guidelines. **Briefings in Bioinformatics**, v. 13, n. 3, p. 281–291, 09 2011.

JAMEI, M. et al. Surface water electrical conductivity and bicarbonate ion determination using a smart hybridization of optimal boruta package with elman recurrent neural network. **Process Safety and Environmental Protection**, v. 174, p. 115–134, 2023.

JIANG, L. et al. Prediction of snp sequences via gini impurity based gradient boosting method. **IEEE Access**, IEEE, v. 7, p. 12647–12657, 2019.

JIN, L. et al. Pathway-based analysis tools for complex diseases: a review. **Genomics, Proteomics & Bioinformatics**, Elsevier, v. 12, n. 5, p. 210–220, 2014.

JOVIĆ, A.; BRKIĆ, K.; BOGUNOVIĆ, N. A review of feature selection methods with applications. In: IEEE. **2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)**. [S.l.], 2015. p. 1200–1205.

KAMKAR, I. et al. Exploiting feature relationships towards stable feature selection. In: IEEE. **2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)**. [S.l.], 2015. p. 1–10.

KANEHISA, M. et al. KEGG as a reference resource for gene and protein annotation. **Nucleic Acids Research**, v. 44, n. D1, p. D457–D462, 10 2015.

Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice 6 (1): 4.

KHATRI, P.; SIROTA, M.; BUTTE, A. J. Ten years of pathway analysis: current approaches and outstanding challenges. **PLOS computational biology**, Public Library of Science San Francisco, USA, v. 8, n. 2, p. e1002375, 2012.

KHUSH, G. S. Origin, dispersal, cultivation and variation of rice. **Plant Molecular Biology**, Springer, v. 35, n. 1, p. 25–34, 1997.

KIM, K.-W. et al. Development of an inclusive 580k snp array and its application for genomic selection and genome-wide association studies in rice. **Frontiers in Plant Science**, Frontiers, v. 13, p. 1036177, 2022.

KIM, S.; MISRA, A. Snp genotyping: technologies and biomedical applications. **Annu. Rev. Biomed. Eng.**, Annual Reviews, v. 9, p. 289–320, 2007.

KIMURA-KATAOKA, K. et al. Genetic and expression analysis of snp s in the human deoxyribonuclease ii: Snp s in the promoter region reduce its in vivo activity through decreased promoter activity. **Electrophoresis**, Wiley Online Library, v. 33, n. 18, p. 2852–2858, 2012.

KIRA, K.; RENDELL, L. A. The feature selection problem: Traditional methods and a new algorithm. In: **Proceedings of the tenth national conference on Artificial intelligence**. [S.l.: s.n.], 1992. p. 129–134.

KLOPFENSTEIN, D. et al. Goatools: A python library for gene ontology analyses. **Scientific Reports**, Nature Publishing Group UK London, v. 8, n. 1, p. 10872, 2018.

KONONENKO, I. Estimating attributes: Analysis and extensions of relief. In: SPRINGER. **European conference on machine learning**. [S.l.], 1994. p. 171–182.

KORTE, A.; FARLOW, A. The advantages and limitations of trait analysis with gwas: a review. **Plant Methods**, BioMed Central, v. 9, n. 1, p. 1–9, 2013.

KOTSIANTIS, S. B. Decision trees: a recent overview. **Artificial Intelligence Review**, Springer, v. 39, n. 4, p. 261–283, 2013.

KUHN, M.; JOHNSON, K. **Feature engineering and selection: A practical approach for predictive models**. [S.l.]: CRC Press, 2019.

KULESHOV, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. **Nucleic Acids Research**, Oxford University Press, v. 44, n. W1, p. W90–W97, 2016.

KUMAR, V.; MINZ, S. Feature selection: a literature review. **SmartCR**, v. 4, n. 3, p. 211–229, 2014.

KURSA, M. B.; JANKOWSKI, A.; RUDNICKI, W. R. Boruta–a system for feature selection. **Fundamenta Informaticae**, IOS Press, v. 101, n. 4, p. 271–285, 2010.

LEIPZIG, J. A review of bioinformatic pipeline frameworks. **Briefings in Bioinformatics**, Oxford University Press, v. 18, n. 3, p. 530–536, 2017.

LI, C. Preprocessing methods and pipelines of data mining: An overview. **arXiv**, 2019.

LI, J. et al. Feature selection: A data perspective. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, v. 50, n. 6, dec 2017.

LI, S. et al. Genomic selection in chinese holsteins using regularized regression models for feature selection of whole genome sequencing data. **Animals**, MDPI, v. 12, n. 18, p. 2419, 2022.

LI, X. et al. A debiased mdi feature importance measure for random forests. **Advances in Neural Information Processing Systems**, v. 32, 2019.

LIANG, P.; PARDEE, A. B. Analysing differential gene expression in cancer. **Nature Reviews Cancer**, Nature Publishing Group, v. 3, n. 11, p. 869–876, 2003.

LIBERZON, A. et al. The molecular signatures database hallmark gene set collection. **Cell Systems**, Elsevier, v. 1, n. 6, p. 417–425, 2015.

LIM, A. J. et al. Robust snp-based prediction of rheumatoid arthritis through machine-learning-optimized polygenic risk score. **Journal of Translational Medicine**, BioMed Central, v. 21, n. 1, p. 1–17, 2023.

LIN, E.; LIN, C.-H.; LANE, H.-Y. Prediction of functional outcomes of schizophrenia with genetic biomarkers using a bagging ensemble machine learning method with feature selection. **Scientific Reports**, Nature Publishing Group UK London, v. 11, n. 1, p. 10179, 2021.

LOMAX, J. Get ready to go! a biologist's guide to the gene ontology. **Briefings in Bioinformatics**, Henry Stewart Publications, v. 6, n. 3, p. 298–304, 2005.

LOUPPE, G. Bayesian optimisation with scikit-optimize. In: **PyData Amsterdam**. [S.l.: s.n.], 2017.

LOUPPE, G. et al. Understanding variable importances in forests of randomized trees. **Advances in Neural Information Processing Systems**, v. 26, 2013.

LU, Q. et al. Genetic variation and association mapping for 12 agronomic traits in indica rice. **BMC Genomics**, BioMed Central, v. 16, n. 1, p. 1–17, 2015.

MA, X. et al. Plantgsad: a comprehensive gene set annotation database for plant species. **Nucleic Acids Research**, Oxford University Press, v. 50, n. D1, p. D1456–D1467, 2022.

MACARRON, R. et al. Impact of high-throughput screening in biomedical research. **Nature Reviews Drug Discovery**, Nature Publishing Group, v. 10, n. 3, p. 188–195, 2011.

MAGUIRE, T. et al. A review of feature selection and ranking methods. **Proceedings of the 19th RUG**, Rijksuniversiteit Groningen Groningen, The Netherlands, p. 15–20, 2022.

MARCANO-CEDEñO, A. et al. Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network. In: **IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society**. [S.l.: s.n.], 2010. p. 2845–2850.

MCGUIGAN, F. E.; RALSTON, S. H. Single nucleotide polymorphism detection: allelic discrimination using taqman. **Psychiatric Genetics**, LWW, v. 12, n. 3, p. 133–136, 2002.

MEKSEM, K.; KAHL, G. **The handbook of plant genome mapping: genetic and physical mapping**. [S.l.]: John Wiley & Sons, 2006.

MI, H. et al. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. **Nucleic Acids Research**, v. 47, n. D1, p. D419–D426, 11 2018.

MORALES, K. Y. et al. An improved 7k snp array, the c7air, provides a wealth of validated snp markers for rice breeding and genetics studies. **PLOS One**, Public Library of Science San Francisco, CA USA, v. 15, n. 5, p. e0232479, 2020.

MOYANO, J. M. et al. Review of ensembles of multi-label classifiers: models, experimental study and prospects. **Information Fusion**, Elsevier, v. 44, p. 33–45, 2018.

MUTHUKRISHNAN, R.; ROHINI, R. Lasso: A feature selection technique in predictive modeling for machine learning. In: IEEE. **2016 IEEE International Conference on Advances in Computer Applications (ICACA)**. [S.l.], 2016. p. 18–20.

NASEER, S.; SALEEM, Y. Enhanced network intrusion detection using deep convolutional neural networks. **KSII Transactions on Internet and Information Systems (TIIS)**, Korean Society for Internet Information, v. 12, n. 10, p. 5159–5178, 2018.

NASER, M.; ALAVI, A. H. Error metrics and performance fitness indicators for artificial intelligence and machine learning in engineering and sciences. **Architecture, Structures and Construction**, Springer, p. 1–19, 2021.

NAVEED, S. A. et al. Identification of qtn and candidate genes for salinity tolerance at the germination and seedling stages in rice by genome-wide association analyses. **Scientific Reports**, Nature Publishing Group UK London, v. 8, n. 1, p. 6505, 2018.

NEMBRINI, S.; KÖNIG, I. R.; WRIGHT, M. N. The revival of the gini importance? **Bioinformatics**, Oxford University Press, v. 34, n. 21, p. 3711–3718, 2018.

NORTON, G. J. et al. Genome wide association mapping of grain and straw biomass traits in the rice bengal and assam aus panel (baap) grown under alternate wetting and drying and permanently flooded irrigation. **Frontiers in Plant Science**, Frontiers Media SA, v. 9, p. 1223, 2018.

ORSOUW, N. J. V. et al. Complexity reduction of polymorphic sequences (crops™): a novel approach for large-scale polymorphism discovery in complex genomes. **PLOS one**, Public Library of Science San Francisco, USA, v. 2, n. 11, p. e1172, 2007.

OZISIK, O.; TÉRÉZOL, M.; BAUDOT, A. orsum: a python package for filtering and comparing enrichment analyses using a simple principle. **BMC Bioinformatics**, Springer, v. 23, n. 1, p. 293, 2022.

PANTALIAO, G. F. et al. Genome wide association study (gwas) for grain yield in rice cultivated under water deficit. **Genetica**, Springer, v. 144, n. 6, p. 651–664, 2016.

PARK, C. H.; KIM, S. B. Sequential random k-nearest neighbor feature selection for high-dimensional data. **Expert Systems with Applications**, Elsevier, v. 42, n. 5, p. 2336–2342, 2015.

PAUL, S.; DRINEAS, P. Feature selection for ridge regression with provable guarantees. **Neural Computation**, MIT Press, v. 28, n. 4, p. 716–742, 2016.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PICOULT-NEWBERG, L. et al. Mining snps from est databases. **Genome Research**, Cold Spring Harbor Lab, v. 9, n. 2, p. 167–174, 1999.

PIEPHO, H. et al. Blup for phenotypic selection in plant breeding and variety testing. **Euphytica**, Springer, v. 161, n. 1-2, p. 209–228, 2008.

PLESSIS, L. D.; ŠKUNCA, N.; DESSIMOZ, C. The what, where, how and why of gene ontology—a primer for bioinformaticians. **Briefings in Bioinformatics**, Oxford University Press, v. 12, n. 6, p. 723–735, 2011.

PODDER, M. et al. Robust snp genotyping by multiplex pcr and arrayed primer extension. **BMC Medical Genomics**, BioMed Central, v. 1, n. 1, p. 1–15, 2008.

PRAVEENA, V. et al. Bio-inspired ensemble feature selection and deep auto-encoder approach for rapid diagnosis of breast cancer. **Multimedia Systems**, Springer, p. 1–17, 2023.

PUDJIHARTONO, N. et al. A review of feature selection methods for machine learning-based disease risk prediction. **Frontiers in Bioinformatics**, Frontiers Media SA, v. 2, p. 927312, 2022.

PULLISSERY, Y. H.; STARKEY, A. Application of feature selection methods for improving classifcation accuracy and run-time: A comparison of performance on real-world datasets. In: IEEE. **2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)**. [S.l.], 2023. p. 687–694.

PURCELL, S. et al. Plink: a tool set for whole-genome association and population-based linkage analyses. **The American Journal of Human Genetics**, Elsevier, v. 81, n. 3, p. 559–575, 2007.

RAMZAN, F. et al. Combining random forests and a signal detection method leads to the robust detection of genotype-phenotype associations. **Genes**, MDPI, v. 11, n. 8, p. 892, 2020.

RASCHIA, M. A. et al. Methodology for the identification of relevant loci for milk traits in dairy cattle, using machine learning algorithms. **MethodsX**, Elsevier, v. 9, p. 101733, 2022.

RAUDVERE, U. et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). **Nucleic Acids Research**, v. 47, n. W1, p. W191–W198, 05 2019.

RAY, D. K.; MUELLER, N. D.; PAUL, C. West, and jonathan a. foley. **Yield Trends are Insufficient to Double Global Crop Production**, v. 2050, p. 1–8, 2013.

REIMAND, J. et al. Pathway enrichment analysis and visualization of omics data using g: Profiler, gsea, cytoscape and enrichmentmap. **Nature Protocols**, Nature Publishing Group, v. 14, n. 2, p. 482–517, 2019.

RÜCKSTIESS, T.; OSENDORFER, C.; SMAGT, P. v. d. Sequential feature selection for classification. In: SPRINGER. **Australasian Joint Conference on Artificial Intelligence**. [S.l.], 2011. p. 132–141.

SAEYS, Y.; INZA, I.; LARRANAGA, P. A review of feature selection techniques in bioinformatics. **Bioinformatics**, Oxford University Press, v. 23, n. 19, p. 2507–2517, 2007.

SAKAI, H. et al. Rice annotation project database (rap-db): an integrative and interactive database for rice genomics. **Plant and Cell Physiology**, Oxford University Press, v. 54, n. 2, p. e6–e6, 2013.

SALMAN, R.; KECMAN, V. Regression as classification. In: IEEE. **2012 Proceedings of IEEE Southeastcon**. [S.l.], 2012. p. 1–6.

SANZ, H. et al. Svm-rfe: selection and visualization of the most relevant features through non-linear kernels. **BMC Bioinformatics**, BioMed Central, v. 19, n. 1, p. 1–18, 2018.

SCAVUZZO, C. M. et al. Feature importance: Opening a soil-transmitted helminth machine learning model via shap. **Infectious Disease Modelling**, Elsevier, v. 7, n. 1, p. 262–276, 2022.

SCHLEINITZ, D.; DISTEFANO, J. K.; KOVACS, P. Targeted snp genotyping using the taqman® assay. **Disease Gene Identification: Methods and Protocols**, Springer, p. 77–87, 2011.

SCHLITTGEN, R. A weighted least-squares approach to clusterwise regression. **AStA Advances in Statistical Analysis**, Springer, v. 95, n. 2, p. 205–217, 2011.

SEGER, C. **An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing**. 2018.

SEO, D. et al. Identification of target chicken populations by machine learning models using the minimum number of snps. **Animals**, MDPI, v. 11, n. 1, p. 241, 2021.

SHAFIEE, S. et al. Sequential forward selection and support vector regression in comparison to lasso regression for spring wheat yield prediction based on uav imagery. **Computers and Electronics in Agriculture**, Elsevier, v. 183, p. 106036, 2021.

SHASTRY, B. S. Snps in disease gene mapping, medicinal drug development and evolution. **Journal of Human Genetics**, Nature Publishing Group, v. 52, n. 11, p. 871–880, 2007.

SHERMAN, B. T. et al. David: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). **Nucleic Acids Research**, Oxford University Press, v. 50, n. W1, p. W216–W221, 2022.

SILVA, J. C. F. et al. Machine learning approaches and their current application in plant molecular biology: A systematic review. **Plant Science**, Elsevier, v. 284, p. 37–47, 2019.

SILVA, P. P. et al. A machine learning-based snp-set analysis approach for identifying disease-associated susceptibility loci. **Scientific Reports**, Nature Publishing Group UK London, v. 12, n. 1, p. 15817, 2022.

SOLORIO-FERNÁNDEZ, S.; CARRASCO-OCHOA, J. A.; MARTÍNEZ-TRINIDAD, J. F. A review of unsupervised feature selection methods. **Artificial Intelligence Review**, Springer, v. 53, n. 2, p. 907–948, 2020.

SOUZA, I. P. de et al. Whole-genome resequencing of common bean elite breeding lines. **Scientific Reports**, Nature Publishing Group UK London, v. 13, n. 1, p. 12721, 2023.

SPINDEL, J. et al. Bridging the genotyping gap: using genotyping by sequencing (gbs) to add high-density snp markers and new value to traditional bi-parental mapping and breeding populations. **Theoretical and Applied Genetics**, Springer, v. 126, p. 2699–2716, 2013.

SPOLAÔR, N. et al. Relieff for multi-label feature selection. In: IEEE. **2013 Brazilian Conference on Intelligent Systems**. [S.l.], 2013. p. 6–11.

SYED, F. H. et al. Feature selection for semi-supervised multi-target regression using genetic algorithm. **Applied Intelligence**, Springer, v. 51, p. 8961–8984, 2021.

SYLVESTER, E. V. et al. Applications of random forest feature selection for fine-scale genetic population assignment. **Evolutionary Applications**, Wiley Online Library, v. 11, n. 2, p. 153–165, 2018.

TADIST, K. et al. Feature selection methods and genomic big data: a systematic review. **Journal of Big Data**, SpringerOpen, v. 6, n. 1, p. 1–24, 2019.

THOMAS, P. D. et al. Panther: a library of protein families and subfamilies indexed by function. **Genome Research**, Cold Spring Harbor Lab, v. 13, n. 9, p. 2129–2141, 2003.

THORBURN, D.-M. J. et al. Origin matters: Using a local reference genome improves measures in population genomics. **Molecular Ecology Resources**, Wiley Online Library, v. 23, n. 7, p. 1706–1723, 2023.

TORKAMANEH, D.; LAROCHE, J.; BELZILE, F. Genome-wide snp calling from genotyping by sequencing (gbs) data: a comparison of seven pipelines and two sequencing technologies. **PLOS one**, Public Library of Science San Francisco, CA USA, v. 11, n. 8, p. e0161333, 2016.

TSAI, C.-F.; CHEN, Y.-C. The optimal combination of feature selection and data discretization: An empirical study. **Information Sciences**, Elsevier, v. 505, p. 282–293, 2019.

URBANOWICZ, R. J. et al. Relief-based feature selection: Introduction and review. **Journal of Biomedical Informatics**, p. 189–203, 2018.

_____. Benchmarking relief-based feature selection methods for bioinformatics data mining. **Journal of Biomedical Informatics**, Elsevier, v. 85, p. 168–188, 2018.

VENKATESH, B.; ANURADHA, J. A review of feature selection and its methods. **Cybernetics and Information Technologies**, v. 19, n. 1, p. 3–26, 2019.

VOELKERDING, K. V.; DAMES, S. A.; DURTSCHI, J. D. Next-generation sequencing: from basic research to diagnostics. **Clinical Chemistry**, Oxford University Press, v. 55, n. 4, p. 641–658, 2009.

VOSS-FELS, K.; SNOWDON, R. J. Understanding and utilizing crop genome diversity via high-resolution genotyping. **Plant Biotechnology Journal**, Wiley Online Library, v. 14, n. 4, p. 1086–1094, 2016.

WAITE, J. M.; DARDICK, C. The roles of the igt gene family in plant architecture: past, present, and future. **Current Opinion in Plant Biology**, Elsevier, v. 59, p. 101983, 2021.

WANG, A. et al. Genome-wide association study-based identification genes influencing agronomic traits in rice (oryza sativa l.). **Genomics**, Elsevier, v. 113, n. 3, p. 1396–1406, 2021.

WANG, J. Functional enrichment analysis. In: _____. **Encyclopedia of Systems Biology**. New York, NY: Springer New York, 2013. p. 772–772.

WANG, J.; ZHANG, Z. Gapit version 3: boosting power and accuracy for genomic association and prediction. **Genomics, Proteomics & Bioinformatics**, Elsevier, v. 19, n. 4, p. 629–640, 2021.

WANG, W. et al. Genomic variation in 3,010 diverse accessions of asian cultivated rice. **Nature**, Nature Publishing Group, v. 557, n. 7703, p. 43–49, 2018.

WASIKOWSKI, M.; CHEN, X.-w. Combating the small sample class imbalance problem using feature selection. **IEEE Transactions on Knowledge and Data Engineering**, v. 22, n. 10, p. 1388–1400, 2010.

WIJESOORIYA, K. et al. Guidelines for reliable and reproducible functional enrichment analysis. **bioRxiv**, Cold Spring Harbor Laboratory, 2021.

XU, C.; JACKSON, S. A. **Machine learning and complex biological data**. [S.l.]: Springer, 2019. 1–4 p.

YANG, L.; SHAMI, A. On hyperparameter optimization of machine learning algorithms: Theory and practice. **Neurocomputing**, Elsevier, v. 415, p. 295–316, 2020.

YATES, L. A. et al. Cross validation for model selection: a review with examples from ecology. **Ecological Monographs**, Wiley Online Library, v. 93, n. 1, p. e1557, 2023.

YIN, J.; LI, N. Ensemble learning models with a bayesian optimization algorithm for mineral prospectivity mapping. **Ore Geology Reviews**, Elsevier, v. 145, p. 104916, 2022.

YU, J. et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. **Nature Genetics**, Nature Publishing Group US New York, v. 38, n. 2, p. 203–208, 2006.

YU, L.; LIU, H. Efficient feature selection via analysis of relevance and redundancy. **The Journal of Machine Learning Research**, JMLR. org, v. 5, p. 1205–1224, 2004.

YUAN, J. et al. Genetic basis and identification of candidate genes for salt tolerance in rice by gwas. **Scientific Reports**, Nature Publishing Group UK London, v. 10, n. 1, p. 9958, 2020.

ZHANG, C. et al. Rice3k56 is a high-quality snp array for genome-based genetic studies and breeding in rice (oryza sativa l.). **The Crop Journal**, Elsevier, v. 11, n. 3, p. 800–807, 2023.

ZHAO, H. et al. Ricevarmap: a comprehensive database of rice genomic variations. **Nucleic Acids Research**, Oxford University Press, v. 43, n. D1, p. D1018–D1022, 2015.

ZHAO, K.; RHEE, S. Y. Interpreting omics data with pathway enrichment analysis. **Trends in Genetics**, Elsevier, 2023.

ZHAO, K. et al. Genomic diversity and introgression in o. sativa reveal the impact of domestication and breeding on the rice genome. **PLOS one**, Public Library of Science San Francisco, USA, v. 5, n. 5, p. e10780, 2010.

ZHONG, H. et al. Uncovering the genetic mechanisms regulating panicle architecture in rice with gpwas and gwas. **BMC Genomics**, Springer, v. 22, p. 1–13, 2021.

ZHOU, J. et al. A correlation analysis between snps and rois of alzheimer's disease based on deep learning. **BioMed Research International**, Hindawi Limited, v. 2021, p. 1–13, 2021.

ZHOU, W. et al. Minor qtls mining through the combination of gwas and machine learning feature selection. **BioRxiv**, Cold Spring Harbor Laboratory, p. 712190, 2019.

ZHU, H.; ZHOU, X. Statistical methods for snp heritability estimation and partition: A review. **Computational and Structural Biotechnology Journal**, Elsevier, v. 18, p. 1557–1568, 2020.

ZOLET, A. C. T. et al. Marcadores moleculares na era genômica: metodologias e aplicações. Sociedade Brasileira de Genética, 2017.

Apêndices

APÊNDICE A

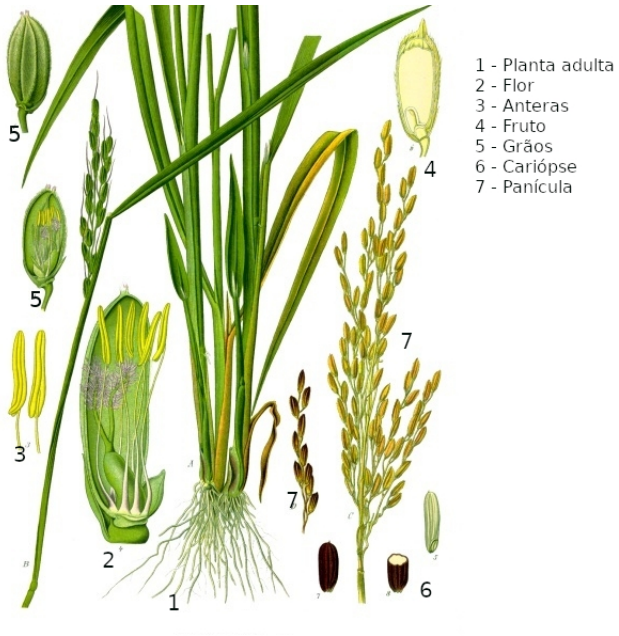
Coleta e Preparação do Conjunto de Dados

Neste capítulo iremos apresentar uma extensão técnica das abordagens utilizadas para coleta e mensuração de cada fenótipo atribuído a um genótipo estudado e comumente explorado em trabalhos dessa natureza. Os procedimentos aqui mencionados foram realizados pelos responsáveis da empresa Embrapa, não sendo assim resultados diretos do presente trabalho. Porém, é fundamental a compreensão dos mesmos para o entendimento dos atributos de classe gerados e utilizados nas abordagens de seleção de SNPs, bem como das metodologias de normalização aplicadas na etapa de pré-processamento.

A Figura 25 apresenta uma representação numerada das partes de uma planta de arroz. Nessa figura é possível observar algumas características fenotípicas mensuradas e avaliadas nesse trabalho, bem como fenótipos amplamente estudados em outros trabalhos correlatos. Nas seções a seguir, iremos apresentar os procedimentos de mensuração e anotações dos fenótipos avaliados nesse trabalho, desde a recepção da planta no momento da colheita até a anotação final das medidas no quadro de dados utilizado. O restante desse apêndice está organizado da seguinte forma:

- Floração A.1 - Apresenta o processo de mensuração da floração da planta de arroz, bem como imagens ilustrativas do cenário ao qual o procedimento foi realizado;
- Altura A.2 - Apresenta o processo de medição da altura da planta de arroz, bem como os equipamentos utilizados para essa tarefa;
- Número de Panículas A.3 - Apresenta o processo de medição do número de panículas da planta de arroz, bem como o procedimento realizado;

Figura 25 – Representação simbólica da planta de arroz



- ❏ Percentual de Grãos Inteiros A.4 - Apresenta o processo de avaliação do percentual de grãos inteiros desde sua colocação no equipamento até a medição final;
- ❏ Produtividade A.5 - Apresenta a medida utilizada e sua convenção;
- ❏ Relação Comprimento por Largura e Centro Branco A.6 - Apresenta o procedimento de avaliação das informações dos grãos de arroz através do equipamento de mensuração;

A.1 Floração

Figura 26 – Floração da planta de arroz

A medida de Floração é mensurada a partir da contagem dos dias desde o plantio até o dia em que 50% das flores estão abertas. Essa medida pode ser entendida como a quantidade de dias que uma planta leva do plantio até o florescimento. A Figura 26

mostra um exemplo de cultivo de plantas de arroz e os estados anteriores e posteriores ao primeiro florescimento. A imagem à esquerda demonstra uma plantação de arroz antes do florescimento. A imagem à direita apresenta uma planta de arroz com algumas flores.

A.2 Altura

O fenótipo denominado altura equivale, de fato, à altura das plantas no momento que elas estavam totalmente formadas. A altura foi medida com auxílio de uma trena. O valor apresentado equivale à distância obtida a partir da base do solo até a inserção da folha bandeira. A folha bandeira é aquela denominada mais próxima aos grãos. A haste de referência para essa medição foi perfilho mais central, que em termos práticos é o galho mais central. O valor aqui coletado foi registrado em centímetros. A Figura 27 mostra o processo de medição da planta de arroz descrito aqui.

Figura 27 – Medição da planta de arroz com a trena

A.3 Número de panículas

As Panículas de uma planta de arroz são os pequenos cachos que contêm aglomerados de grãos de arroz. Logo, a medição do número de panículas equivale à contagem desses cachos de grãos. Idealmente, cada perfilho deve conter uma panícula, sendo portanto o número de panículas equivalente ao número de perfilhos em um cenário acomodado.

A.4 Percentual de grãos inteiros

O percentual de grãos inteiros é obtido pela divisão do valor observado na pesagem dos grãos que ficam inteiros após passarem por um equipamento que descasca e poli os grãos. A Figura 28 mostra esse procedimento de descasco. Para essa medição, amostras de 100 gramas coletadas em cada planta foram avaliadas. Uma vez passados pelo equipamento,

Figura 28 – Equipamento de polimento e remoção da casca do grão de arroz

os grãos de cada amostra foram separados manualmente e depois pesados para, por fim, serem mensurados e anotados.

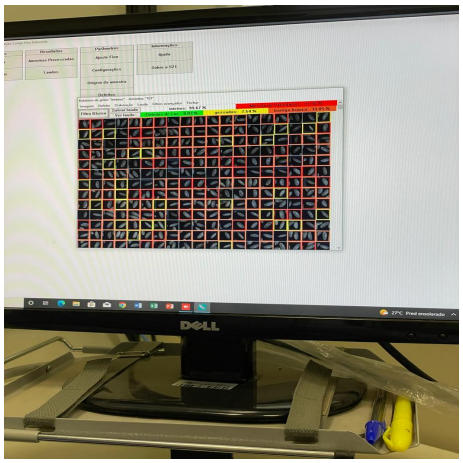
A.5 Produtividade

A medida de produtividade foi estimada pelo peso total dos grãos, levando em consideração todos os grãos que foram destacados em cada planta de arroz. Essa medida é importante no contexto geral, levando em consideração os aspectos produtivos do cultivo.

A.6 Comprimento por Largura e Centro Branco

Figura 29 – Avaliação de grãos de arroz pelo software SA-21

Figura 30 – Software SA21



As informações de comprimento, largura e centro branco de um grão são obtidas através de um equipamento chamado SA-21, que é um escâner inteligente para avaliação de informações relacionadas ao arroz, assim como mostrado na Figura 29. Esse equipamento

conta e avalia cada grão enviando as informações obtidas para um software que permite que as informações sejam salvas em uma planilha, assim como mostrado na Figura 30.

APÊNDICE B

Referência aos Dados e Resultados

Este apêndice tem como objetivo explicar e contextualizar alguns arquivos existentes no link Resultados, que contém os algoritmos e resultados provenientes de cada dos métodos propostos e comparações realizadas, o restante das seções está dividida da seguinte forma:

- Seção B.1 Conjunto de dados: contém os links contendo as informações dos SNPs e fenótipos tal como foram recebidos pela Embrapa.

- Seção B.2 Algoritmos Aplicados: contém os links que dão acesso aos algoritmos utilizados no presente trabalho.

B.1 Conjunto de dados

Os dados recebidos fornecem informações relativas as mutações pontuais que foram mapeados na fita direta do DNA. Essas informações estão dividas em dois arquivos tabulares assim como apresentados na tabela 34.

Tabela 34 – Links com dados brutos

Nome do Arquivo	Descrição	
base_features.csv	Arquivo que contém os SNPs detectados em cada amostra	
base_targets.csv	Arquivo com as características fenotípicas observadas em cada amostra	Caracte

B.2 Algoritmos

Os algoritmos utilizados nesse trabalho foram executados utilizando linguagem python. Todos os scripts levam em consideração os arquivos base apresentados na seção anterior. Os algoritmos estão disponíveis pelo link Algoritmos, que confere livre acesso a qualquer individuo possuidor do link. O diretório remoto leva para duas pastas, uma denominada Classificação e outra Regressão, sendo que ambas possuem os 8 algoritmos utilizados adaptados, respectivamente, para os fenótipos sem a discretização e com a discretização. A Tabela 35 explica qual algoritmo está contido em cada um dos arquivos.

Tabela 35 – Algoritmos aplicados

Nome do Arquivo	Descrição
boruta_random_forest.py	Algoritmo Boruta adaptado para utilizar internamente Random Forest *Random Forest* e *Extra-Trees*.
boruta_extra_trees.py	Algoritmo Boruta adaptado para utilizar internamente Extra Trees *Random Forest* e *ExtraTrees*.
extra_trees.py	Algoritmo *ExtraTree* para seleção de SNPs.
random_forest.py	Algoritmo *ExtraTree* para seleção de SNPs.
rfe_random_forest.py	Algoritmo RFE que utiliza internamente Random Forest para seleção de SNPs.
rfe_extra_trees.py	Algoritmo RFE que utiliza internamente *Extra Trees* para seleção de SNPs.
rfe_svm.py	Algoritmo RFE que utiliza internamente SVM para seleção de SNPs.
xgboost.py	Algoritmo XGBoost utilizado para seleção de SNPs.

APÊNDICE C

Resultados Expandidos

Esse apêndice está dividido em 2 seções que apresentam uma breve explicação sobre alguns detalhes extras dos resultados. A primeira seção trata da localização dos arquivos que contém os resultados detalhados dos experimentos, bem como uma breve explicação dos dados contidos em cada um. A segunda seção apresenta os resultados observados na etapa de otimização de hiper-parâmetros.

C.1 Arquivos com Resultados Expandidos

Esta seção apresenta 3 tabelas contendo os resultados expandidos observados nos experimentos realizados. A Tabela 36 apresenta uma explicação sobre o conteúdo dos arquivos encontrados no diretório acessível pelo link Resultados da Discretização, que detalham os resultados observados na etapa de discretização.

Tabela 36 – Arquivos com resultados da etapa de discretização

Nome do Arquivo	Descrição
results_Aca	Arquivo que contém os resultados de cada algoritmo de discretização executado sobre o fenótipo acamamento.
results_Alt	Arquivo que contém os resultados de cada algoritmo de discretização executado sobre o fenótipo altura.
results_CL	Arquivo que contém os resultados de cada algoritmo de discretização executado sobre o fenótipo comprimento do grão por largura.
results_Flo	Arquivo que contém os resultados de cada algoritmo de discretização executado sobre o fenótipo floração.
results_Int	Arquivo que contém os resultados de cada algoritmo de discretização executado sobre o fenótipo percentual de grãos inteiros.
results_Prod	Arquivo que contém os resultados de cada algoritmo de discretização executado sobre o fenótipo produtividade.

A Tabela 37 apresenta uma explicação sobre o conteúdo dos arquivos encontrados no diretório acessível pelo link SNPs Selecionados. Em cada arquivo estão detalhados os resultados contendo os SNPs que foram selecionados como importantes para cada fenótipo.

Por fim, a Tabela 38, apresenta o nome dos arquivos onde estão dispostas as métricas de erro expandidas de cada método comparado na Seção 6.3. Assim como foi mencionado,

Tabela 37 – Contagens de SNPs selecionados

Nome do Arquivo	Descrição
FSCd	Apresenta os SNPs selecionados como importantes para cada atributo pelo método FSCd. Estão dispostos o nome do fenótipo, a identificação do SNP, e a contagem de algoritmos que o selecionaram.
FSCi	Apresenta os SNPs selecionados como importantes para cada atributo pelo método FSCi. Estão dispostos o nome do fenótipo, a identificação do SNP, e a contagem de algoritmos que o selecionaram.
FSDCd	Apresenta os SNPs selecionados como importantes para cada atributo pelo método FSDCd. Estão dispostos o nome do fenótipo, a identificação do SNP, e a contagem de algoritmos que o selecionaram.
FSDCi	Apresenta os SNPs selecionados como importantes para cada atributo pelo método FSCd. Estão dispostos o nome do fenótipo, a identificação do SNP, e a contagem de algoritmos que o selecionaram.

são exibidos para cada método comparado nas Tabelas 27, 28 e 29, apenas o resultado referente à média dos 4 algoritmos utilizados na tarefa de predição, contudo nos arquivos apontados aqui, é possível verificar para cada método, quais foram os valores das métricas MSE, RMSE e MAE para cada um dos quatro algoritmos (XGB, RF, SVM e *Extra Trees*) obtidos em cada fenótipo. Esses arquivos podem ser encontrados através do link Métricas de Erro Expandidas.

Tabela 38 – Descrição dos arquivos contendo as métricas de erros

Nome do arquivo	Descrição
boruta_extra_trees	Contém as métricas de erro dos algoritmos utilizados para comparar Boruta em conjunto as Extra Trees, no capítulo de comparação.
boruta_random_forest	Contém as métricas de erro dos algoritmos utilizados para comparar Boruta em conjunto a Random Forest, no capítulo de comparação.
multi_surf	Contém as métricas de erro dos algoritmos utilizados para comparar Multi SURF, no capítulo de comparação.
multi_surf_start	Contém as métricas de erro dos algoritmos utilizados para comparar Multi SURF*, no capítulo de comparação.
reliefF	Contém as métricas de erro dos algoritmos utilizados para comparar ReliefF, no capítulo de comparação.
rfe_extra_trees	Contém as métricas de erro dos algoritmos utilizados para comparar Extra Trees, no capítulo de comparação.
rfe_random_forest	Contém as métricas de erro dos algoritmos utilizados para comparar Random Forest, no capítulo de comparação.
rfe_svm	Contém as métricas de erro dos algoritmos utilizados para comparar RFE em conjunto ao SVM, no capítulo de comparação.
surf	Contém as métricas de erro dos algoritmos utilizados para comparar SURF, no capítulo de comparação.
surf_start	Contém as métricas de erro dos algoritmos utilizados para comparar SURF*, no capítulo de comparação.

C.2 Resultado da Otimização de Hiper-Parâmetros

Esta seção apresenta os resultados referentes aos experimentos de otimização de hiper-parâmetros. Em cada tabela são apresentados os melhores hiper-parâmetros selecionados por fenótipo. Os resultados estão separados em regressão e classificação. Os resultados do grupo de regressão denotam os hiper-parâmetros selecionados para fenótipos com valores contínuos, isto é, dos fenótipos sem nenhum processamento, enquanto que os demais equivalem aos melhores hiper-parâmetros selecionados para os fenótipos discretizados.

Tabela 39 – Hiper-Parâmetros ExtraTrees

Tarefa de Aprendizado	**Fenótipo**	**min samples split**	**min samples leaf**	**bootstrap**	**max features**	**n estimators**	**max depth**	**criterion**
Regressão	Aca	2	2	True	sqrt	800	60	friedman_mse
	Int	20	1	True	auto	400	60	friedman_mse
	Alt	5	4	False	sqrt	400	50	absolute_error
	Flo	2	4	False	sqrt	800	90	poisson
	Prod	5	2	True	sqrt	1400	100	poisson
	C/L	20	20	False	sqrt	800	80	poisson
Classificação	uniforme_2_Aca	2	1	False	sqrt	100	None	gini
	uniforme_3_Prod	10	2	True	sqrt	500	200	gini
	uniforme_4_Alt	2	1	False	sqrt	200	80	gini
	uniforme_2_Int	10	4	True	sqrt	200	10	gini
	agrupamento_5_Flo	5	4	False	sqrt	500	10	gini
	uniforme_3_CL	2	1	True	sqrt	200	10	gini

Tabela 40 – Hiper-Parâmetros Random Forest

Tarefa de Aprendizado	**Fenótipo**	**min samples split**	**min samples leaf**	**bootstrap**	**max features**	**n estimators**	**max depth**
Regressão	Aca	2	1	True	sqrt	2000	None
	Int	2	4	True	sqrt	1800	90
	Alt	10	4	True	sqrt	1600	30
	Flo	5	4	False	sqrt	1800	30
	Prod	2	4	True	sqrt	800	100
	C/L	2	1	False	1	1800	80
Classificação	uniforme_3_Prod	10	4	True	sqrt	200	80
	uniforme_4_Alt	10	4	False	sqrt	1400	80
	uniforme_2_Int	10	4	True	sqrt	400	60
	agrupamento_5_Flo	5	4	False	sqrt	600	30
	uniforme_3_CL	5	2	True	sqrt	1600	10
	uniforme_2_Aca	10	4	False	sqrt	1600	50

Tabela 41 – Hiper-Parâmetros XGBoost

Tarefa de Aprendizado	**Fenótipo**	**colsample bylevel**	**colsample bytree**	**subsample**	**learning rate**	**n estimators**	**max depth**
Regressão	Aca	0.8	0.9	0.6	0.01	1000	15
	Int	0.7	0.6	0.6	0.01	1000	15
	Alt	0.5	0.4	0.5	0.01	1000	15
	Flo	0.9	0.8	0.8	0.10	200	10
	Prod	0.8	0.4	0.7	0.10	500	5
	C/L	0.6	0.4	0.5	0.01	100	20
Classificação	uniforme_3_Prod	0.5	0.4	0.9	0.01	200	3
	uniforme_4_Alt	0.5	0.5	0.9	0.01	500	3
	uniforme_2_Int	0.7	0.8	0.6	0.01	200	3
	agrupamento_5_Flo	0.6	0.7	0.5	0.01	500	3
	uniforme_3_CL	0.7	0.4	0.7	0.10	500	10
	uniforme_2_Aca	0.4	0.8	0.6	0.01	500	10

Tabela 42 – Hiper-Parâmetros SVM

Tarefa de Aprendizado	Fenótipo	C	Kernel	gamma
Regressão	Aca	0.005	poly	0.001
	Int	3.805	rbf	0.001
	Alt	0.105	poly	0.001
	Flo	4.805	rbf	0.001
	Prod	0.005	poly	0.010
	C/L	9.805	rbf	0.001
Classificação	uniforme_3_Prod	1.105	rbf	0.001
	uniforme_4_Alt	7.205	poly	0.100
	uniforme_2_Int	0.805	rbf	0.001
	agrupamento_5_Flo	0.105	poly	0.010
	uniforme_3_CL	7.105	sigmoid	0.001
	uniforme_2_Aca	2.605	rbf	0.001