UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
DEPARTAMENTO DE COMPUTAÇÃO

Henrique Cordeiro Frajacomo

Seleção de SNPs utilizando Random Forests

São Carlos
2020

HENRIQUE CORDEIRO FRAJACOMO

SELEÇÃO DE SNPS UTILIZANDO RANDOM FORESTS

Trabalho de Conclusão de Curso submetido à Universidade Federal de São Carlos, como requisito necessário para obtenção do grau de Bacharel em Ciência da Computação

São Carlos, Julho de 2020

UNIVERSIDADE FEDERAL DE SÃO CARLOS

HENRIQUE CORDEIRO FRAJACOMO

Esta Monografia foi julgada adequada para a obtenção do título de Bacharel em Ciência da Computação, sendo aprovada em sua forma final pela banca examinadora:

Orientador(a): Prof. Dr. Ricardo Cerri
Universidade Federal de São Carlos - UFSCar

Prof. Dr. Murilo Coelho Naldi
Universidade Federal de São Carlos - UFSCar

Dr. Marcelo Gonçalves Narciso
Empresa Brasileira de Pesquisa Agropecuária - Embrapa

São Carlos, Julho de 2020

AGRADECIMENTOS

Agradeço imensamente ao meu orientador, Prof. Dr. Ricardo Cerri, por me acompanhar na minha jornada acadêmica, apontando ideias e conceitos que tornaram este projeto possível. Agradeço também a todos os docentes do Departamento de Computação da UFSCar, em especial os professores Estevam Hruschka Jr., Murilo Naldi e Diego Silva por me introduzirem à área de pesquisa que tanto amo. Agradeço também aos professores Murilo Homem e Rafael Aroca por encorajarem minha criatividade e me proporcionarem experiências únicas. Agradeço ao meu co-orientador Me. Renato Santos pela orientação, paciência e companheirismo por todos estes anos. Por fim, agradeço a todos os meus amigos que estiveram comigo, me apoiando e sendo uma fonte de inspiração crucial para esta jornada.

Resumo

Os Polimorfismos de Nucleotídeo Único (SNPs) são variações de base única na sequência de nucleotídeos de indivíduos diferentes ou entre sequências homólogas dentro de um ser vivo. Uma grande parte de variações genéticas ocorrem como SNPs. Muitas destas variações genéticas ocorrem em plantas, influenciando características diretamente ligadas com a produtividade de culturas, como por exemplo o arroz. O Brasil, além de ser o maior produtor dentre os países ocidentais, é também o maior consumidor per capita de arroz. O arroz é um dos principais alimentos para a nutrição humana, sendo a base alimentar para mais da metade da população mundial e majoritariamente produzido por países asiáticos, mas também largamente produzido no Brasil. O arroz faz parte do Programa de Melhoramento Genético da Empresa Brasileira de Pesquisa Agropecuária (Embrapa), que tem como objetivo a melhoria das safras de arroz mirando atingir o padrão de preferência de consumo do Brasil. A Seleção de SNPs que estão fortemente relacionadas com o teor de amilose do arroz é um dos problemas a serem resolvidos no programa de Melhoramento Genético da Embrapa. A Seleção de SNPs pode ser modelada computacionalmente utilizando ferramentas de Aprendizado de Máquina, subárea da Inteligência Artificial, tornando a análise mais rápida e menos custosa. Assim, o objetivo desta pesquisa é desenvolver um método capaz de realizar a tarefa de Seleção de SNPs. Isto é, dado uma característica de um organismo, o método deve encontrar os SNPs relacionados com a dada característica. Como caso de teste, o método será aplicado nos SNPs do conteúdo genômico de diferentes safras de arroz, com o objetivo de encontrar quais SNPs tiveram maior impacto em seu teor de amilose. O método desenvolvido se mostrou eficiente em resolver o problema da Seleção de SNPs. As análises do método destacaram um SNP que foi validado experimentalmente pela Embrapa como importante para o teor de amilose.

Palavras-Chave: Aprendizado de Máquina, Aprendizado Multi-Classe, Seleção de SNPs, *Random Forests*, Bioinformática.

Abstract

Single Nucleotide Polymorphisms (SNPs) are single-base variations in the nucleotide sequence of different individuals or between homologous sequences within a living being. A large part of genetic variations occur as SNPs. Many of these genetic variations occur in plants, influencing characteristics directly linked to crop productivity, such as rice. In addition to being the largest producer among Western countries, Brazil is also the largest per capita consumer of rice. Rice is one of the main foods for human nutrition, being the food base for more than half of the world population and mostly produced by Asian countries, but also widely produced in Brazil. Rice is part of the Genetic Improvement Program of the Brazilian Agricultural Research Corporation (Embrapa), which aims to improve rice crops with the goal of reaching the consumption preference pattern in Brazil. The Selection of SNPs that are strongly related to the amylose content of rice is one of the problems to be solved in Embrapa's Genetic Improvement program. The Selection of SNPs can be modeled computationally using Machine Learning tools, a subarea of Artificial Intelligence, making analysis faster and less costly. Thus, the objective of this research is to develop a method capable of performing the SNP Selection task. That is, given a characteristic of an organism, the method must find the SNPs related to the given characteristic. As a test case, the method will be applied to the SNPs of the genomic content of different rice crops, in order to find out which SNPs had the greatest impact on their amylose content. The developed method proved to be efficient in solving the SNP Selection problem. The analysis of the method highlighted an SNP that was validated experimentally by Embrapa as important for the amylose content.

Lista de abreviaturas e siglas

SNP - do inglês, *Single Nucleotide Polymorphism*

Embrapa - Empresa Brasileira de Pesquisa Agropecuária

AM - Aprendizado de Máquina

RF - do inglês, *Random Forest*

MDGI - do inglês, *Mean Decrease in Gini Index*

Lista de ilustrações

Figura 1 – Exemplo de Árvore de Decisão 23
Figura 2 – Exemplo de Random Forest 23
Figura 3 – Exemplo de Bootstraping 24
Figura 4 – Exemplos de divisões puras e impuras 25
Figura 5 – Exemplo de indução de uma Árvore de Decisão 26
Figura 6 – Treino e pontuação com o método EnGENE 34
Figura 7 – Ordem de processamento interna do método 34
Figura 8 – SNPs importantes para distinguir o teor alto de amilose 40
Figura 9 – SNPs importantes para distinguir o teor intermediário de amilose . . . 41
Figura 10 – SNPs importantes para distinguir o teor baixo de amilose 41
Figura 11 – SNPs importantes para distinguir o teor muito baixo de amilose 42
Figura 12 – SNPs importantes para o teor de amilose 42
Figura 13 – Interface Rosalind vazia 44
Figura 14 – Interface Rosalind com dados treinados 45

Lista de tabelas

Tabela 1 – Exemplo de dados de treino para a geração de uma Árvore de Decisão 26
Tabela 2 – Conjuntos de dados utilizados 32
Tabela 3 – Exemplo de dados para ser aplicado as transformações 32
Tabela 4 – Resultado da transformação Um-Contra-Todos, tendo a classe 'Alto' como principal. 33
Tabela 5 – Resultado da transformação Um-Atributo-por-Valor. 33
Tabela 6 – Conjuntos de dados transformados 37
Tabela 7 – Medidas de avaliação para os conjuntos de teor alto 39
Tabela 8 – Medidas de avaliação para os conjuntos de teor intermediário 39
Tabela 9 – Medidas de avaliação para os conjuntos de teor baixo 39
Tabela 10 – Medidas de avaliação para os conjuntos de teor muito baixo 40

Sumário

1 INTRODUÇÃO 19

2 METODOLOGIA TEÓRICA E TÉCNICA 21
2.1 Polimorfismos de Nucleotídeo Único (SNP) 21
2.2 Aprendizado de Máquina 21
2.3 Árvores de Decisão 22
2.4 Random Forest 22
2.5 Mean Decrease in Gini Index 27

3 TRABALHOS CORRELATOS 29

4 DETALHAMENTO DO DESENVOLVIMENTO 31
4.1 Seleção de SNPs 31
4.2 Descrição dos Dados 31
4.3 O Método Proposto 32
4.4 Experimentos 36

5 RESULTADOS 39

6 CONCLUSÃO 47

REFERÊNCIAS 49

1 Introdução

Os Polimorfismos de Nucleotídeo Único (SNPs) são variações de base única na sequência de nucleotídeos de indivíduos diferentes ou entre sequências homólogas dentro de um indivíduo. O SNP pode ser usado para a distinção entre indivíduos e espécies, análise genética de doenças e características complexas, avaliação do desequilíbrio de ligação, geração de mapa de haplótipos, farmacogenômica, etc [Matukumalli et al. 2006].

O genoma humano contém aproximadamente três bilhões de pares de bases de DNA, chamados nucleotídeos. Quase 99% deles são idênticos entre todos os seres humanos, e apenas 1% varia entre os indivíduos. Uma grande parte dessas variações genéticas ocorrem como Polimorfismos de Nucleotídeo Único (SNPs) [Batnyam, Gantulga e Oh 2013]. Muitas destas variações genéticas ocorrem em plantas, influenciando características diretamente ligadas com a produtividade de culturas, como por exemplo o arroz.

O Brasil, além de ser o maior produtor dentre os países ocidentais, é também o maior consumidor per capita de arroz [Batista 2019]. O arroz é um dos principais alimentos para a nutrição humana, sendo a base alimentar para mais da metade da população mundial e majoritariamente produzido por países asiáticos, mas também largamente produzido no Brasil [Korres et al. 2017].

O amido é um polissacarídeo composto de duas estruturas moleculares complementares: amilose e amilopectina. Aumentos ou decréscimos no teor de uma delas reflete-se de forma inversa no teor da outra e resultam em tendências de comportamento igualmente inverso em relação às propriedades de cocção ou de processamento do arroz [VIEIRA 1998].

O teor de amilose do arroz exerce, reconhecidamente, uma influência marcante no desempenho de cozimento. A escala para classificação do teor de amilose do arroz, utilizada no Programa de Seleção de Linhagens da Empresa Brasileira de Pesquisa Agropecuária (Embrapa) Arroz e Feijão [Martínez 1989], considera os seguintes valores:

- Teor alto: atribuído ao arroz com conteúdo de amilose entre 28% e 32%;
- Teor intermediário: atribuído ao arroz com conteúdo amilótico entre 23% e 27%;
- Teor baixo: atribuído ao arroz com conteúdo amilótico entre 8% e 22%.
- Teor muito baixo: atribuiído ao arroz com conteúdo amilótico menor do que 8%

Para atender as preferências de consumo no Brasil, buscam-se cultivares com conteúdo de amilose intermediário a alto, cujos grãos, quando cozidos, apresentam-se secos e soltos [VIEIRA 1998]. A identificação de quais SNPs são importantes para o teor de

amilose alto de grãos de arroz é crucial para estudos que visam o melhoramento genético do grão, como o Programa de Melhoramento Genético da Embrapa, com o objetivo de aumentar sua produtividade atingindo o padrão de preferência de consumo do Brasil.

O objetivo desta pesquisa é desenvolver um método capaz de realizar a tarefa de Seleção de SNPs. Isto é, dado uma característica de um organismo, o método deve encontrar os SNPs relacionados com a dada característica. Como caso de teste, o método será aplicado nos SNPs do conteúdo genômico de diferentes safras de arroz, com o objetivo de encontrar quais SNPs tiveram maior impacto em seu teor de amilose. Os resultados da análise sobre o arroz foram validados experimentalmente, no intuito de adicionar um novo marcador SNP do teor de amilose ao programa de Melhoramento Genético da Embrapa.

O método desenvolvido se mostrou eficiente em resolver o problema da Seleção de SNPs. As análises do método destacaram um SNP que foi validado experimentalmente pela Embrapa como importante para o teor de amilose.

O restante deste documento está organizado da seguinte maneira. O Capítulo 2 apresenta o referencial teórico necessário para abordar o problema de Seleção de SNPs no contexto dessa pesquisa. O Capítulo 3 sintetiza uma análise da literatura sobre trabalhos de Seleção de SNP utilizando métodos de AM. O Capítulo 4 descreve os dados associados ao problema e os métodos desenvolvidos para tratá-los. O Capítulo 5 apresenta os resultados descobertos durante a análise discutida no capítulo anterior. Finalmente, no Capítulo 6 foram apresentadas as conclusões.

2 Metodologia Teórica e Técnica

2.1 Polimorfismos de Nucleotídeo Único (SNP)

Espécies de seres vivos tendem a compartilhar certas características específicas, como estrutura corporal, constituição, ligações moleculares, entre outras. Se tomarmos humanos como exemplo, podemos encontrar características que definem a maioria dos seres humanos saudáveis. Todas estas características estão descritas como conjuntos de bases nitrogenadas que constituem o DNA das espécies.

Por outro lado, algumas características são mutáveis, como as cores dos olhos, formato do rosto, impressões digitais, tendências nutricionais e outras. Tais características variam de indivíduo para indivíduo, sendo fatores que nos permitem distinguir uns ao outro. Em nível molecular, tais variáveis são descritas como regiões do DNA que são mutáveis. O fato de uma região possuir bases que podem assumir diferentes valores de indivíduo para indivíduo causa diferenças únicas em suas características. Estas bases mutáveis são conhecidas como Polimorfismos de Nucleotídeo Único.

2.2 Aprendizado de Máquina

Definição: A área que realiza estudos de algoritmos de computador que melhoram automaticamente através da experiência [Mitchell et al. 1997].

O Aprendizado de Máquina (AM) possui diversas utilidades. Alguns exemplos recentes, são: aplicações em carros autônomos, inteligência para jogos, reconhecimento de padrões, previsão de ações, classificações estatísticas, traduções automáticas, detecção de fraude, sistemas de recomendação e aplicações em bioinformática. A abordagem investigada neste projeto foi a bioinformática, que visa resolver problemas da biologia através de ferramentas e métodos computacionais.

Os algoritmos de AM tiram seu aprendizado à partir de conjuntos de dados com amostras já rotuladas. A análise dos dados pelo algoritmo faz com que ele se torne cada vez mais eficiente em resolver a tarefa de classificação modelada. Este processo de melhoria do algoritmo, através da análise dos dados já rotulados, é conhecido como 'treino de um algoritmo'. Um algoritmo de AM pode ser avaliado em um processo chamado de 'teste de um algoritmo'. Neste processo, o algoritmo irá predizer o rótulo de amostras desconhecidas por ele. Suas predições serão comparadas com os verdadeiros rótulos dessas amostras. Ao final deste processo, o número de classificações verdadeiramente positivas, verdadeiramente

negativas, falsamente positivas e falsamente negativas permitirá o algoritmo a ser avaliado utilizando diferentes métricas.

A Classificação de Dados é uma das aplicações mais clássicas do AM. Neste projeto, serão utilizadas as classificações binária e Multi-Classe. A Classificação binária consiste em rotular amostras como pertencentes à uma classe ou não. Já a classificação Multi-Classe consiste em rotular exemplos em uma classe dentre um conjunto de três ou mais classes. Exemplificando, temos que o teor de amilose pode ser classificado em mais de duas classes diferentes: alto, intermediário, baixo e muito baixo.

Para lidar com a tarefa de classificação nesta pesquisa foi utilizado o algoritmo *Random Forest*, que constrói uma floresta de algoritmos mais simples chamados de Árvores de Decisão. Ambos os algoritmos serão explicados a seguir.

2.3 Árvores de Decisão

Uma Árvore de Decisão é um modelo de aprendizado que particiona recursivamente o conjunto de dados com base em atributos selecionados até que um critério de parada pré-estabelecido seja atingido. Esse critério de parada pode ser, por exemplo, a profundidade da árvore, o número de exemplos em um nó, ou o particionamento perfeito dos dados. No caso deste último critério, apesar dele funcionar para os dados de treinamento, pode tornar a árvore incapaz de classificar corretamente problemas que diferem dos dados utilizados, perdendo sua capacidade de generalização.

Uma Árvore de Decisão é constituída por nós de decisão, divisões (ou arestas) e nós folha (ou terminais). Os nós de decisão são caracterizados pela comparação de um atributo julgado importante para o particionamento das classes do conjunto de treinamento. As divisões são caminhos tomados na árvore de acordo com valores ou intervalos de valores do atributo que diz respeito ao nó de divisão. Os nós folha representam a classificação resultante do algoritmo de Árvore de Decisão.

O treino de uma Árvore de Decisão se baseia na busca do melhor atributo para ser colocado em um nó de divisão. O melhor atributo pode ser escolhido a partir de dois critérios: ganho de informação ou minimização da impureza. A Figura 1 ilustra um exemplo de árvore de decisão.

2.4 Random Forest

O modelo *Random Forest* (RF) é um conjunto de Árvores de Decisão que são treinadas com amostras aleatórias do conjunto de dados. As divisões de cada árvore da floresta são feitas à partir de subconjuntos aleatórios e com amostras repetidas, com o fim de criar árvores levemente diferentes. As classificações da RF são feitas pelo voto da maioria de

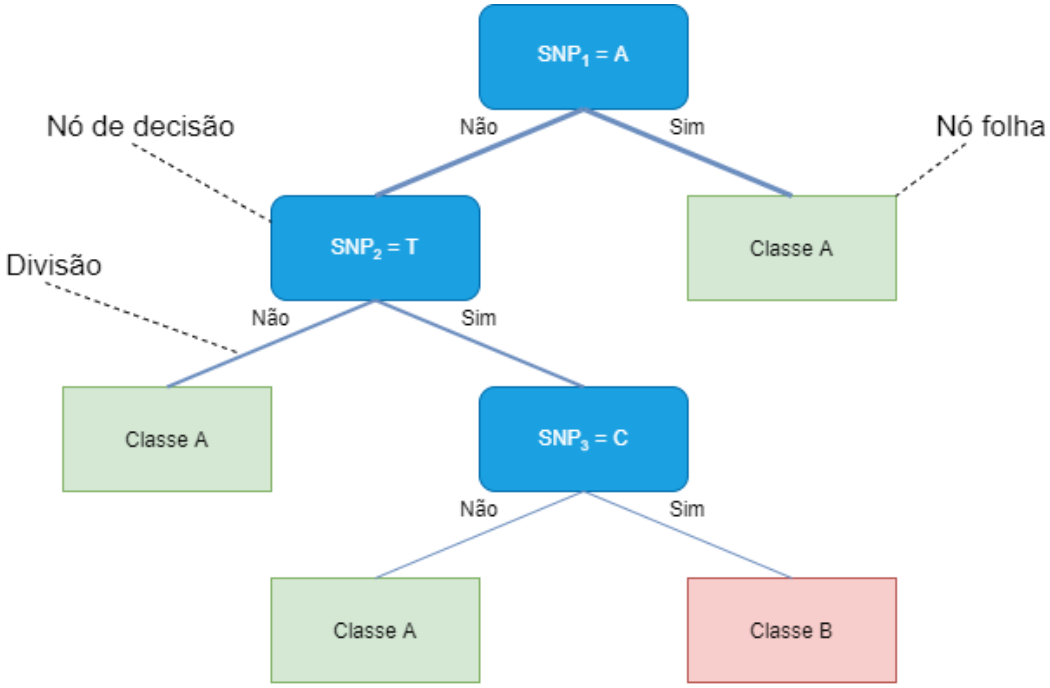


Figura 1 – Exemplo de Árvore de Decisão

suas árvores, diminuindo assim o sobreajuste que ocorreria caso uma única Árvore de Decisão fosse utilizada no problema. A Figura 2 ilustra uma *Random Forest* realizando uma predição. Note que, pelo fato das árvores da floresta serem levemente distintas, o caminho de decisão percorrido por elas (denotado por nós azulados) e os atributos levados em consideração em cada nó de divisão são diferentes, podendo atingir nós folha que dizem respeito a classes diferentes.

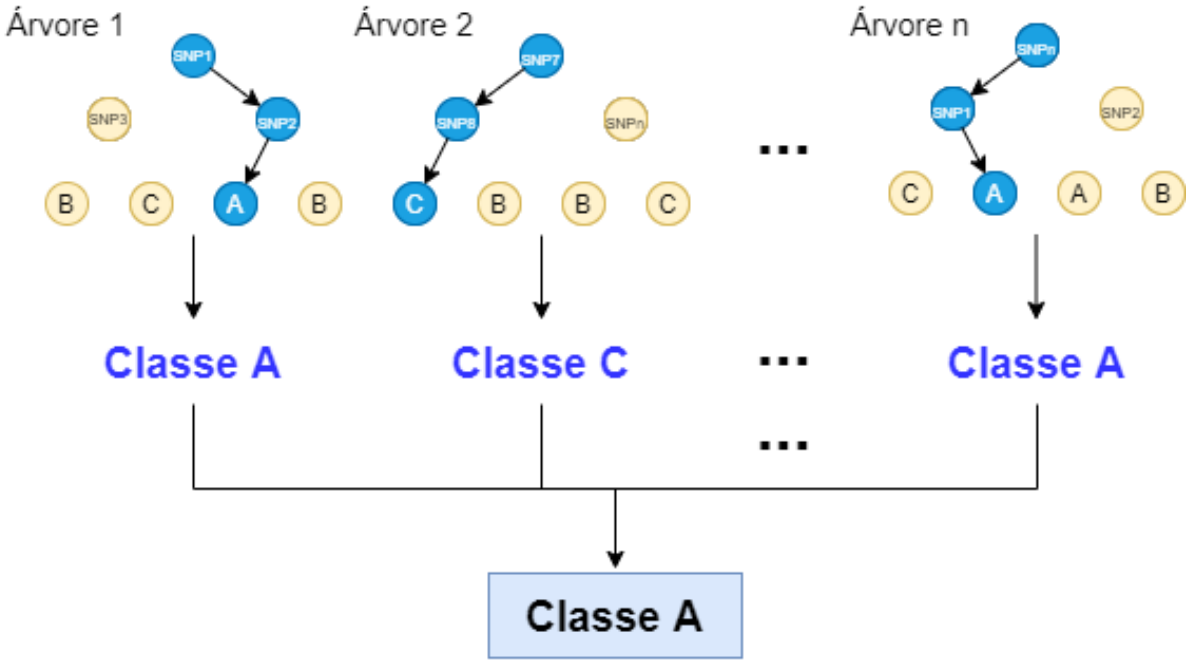


Figura 2 – Exemplo de Random Forest

Toda Árvore de Decisão da RF recebe um subconjunto de dados único para sua indução. Este subconjunto é um pedaço aleatório do conjunto de treinamento, com repetição. Exemplificando, imagine que o conjunto de treino possua elementos enumerados de 1 à 9, onde 2/3 dos dados sejam utilizados e o resto é deixado de lado. Um possível subconjunto passado para a indução de uma árvore pode ser: $\{1, 2, 3, 5, 7, 8\}$. Note como alguns elementos foram deixados de lado. O subconjunto precisa ter tamanho igual ao conjunto de treino original, portanto, deve-se repetir aleatoriamente elementos até o conjunto ter o mesmo tamanho do conjunto original. Completando a repetição, um subconjunto válido poderia ser: $\{1, 2, 2, 3, 5, 5, 7, 7, 8\}$. Este processo de aleatorização e repetição de amostras é chamado de *bootstraping*. A Figura 3 ilustra o processo de *bootstraping*. O subconjunto utilizado para a indução de uma árvore também pode ter apenas um subconjunto de atributos do conjunto original. As amostras deixadas de lado em uma árvore são chamadas de amostras *out-of-bag*. As mesmas podem ser usadas posteriormente para validar o erro de predição de suas respectivas árvores.

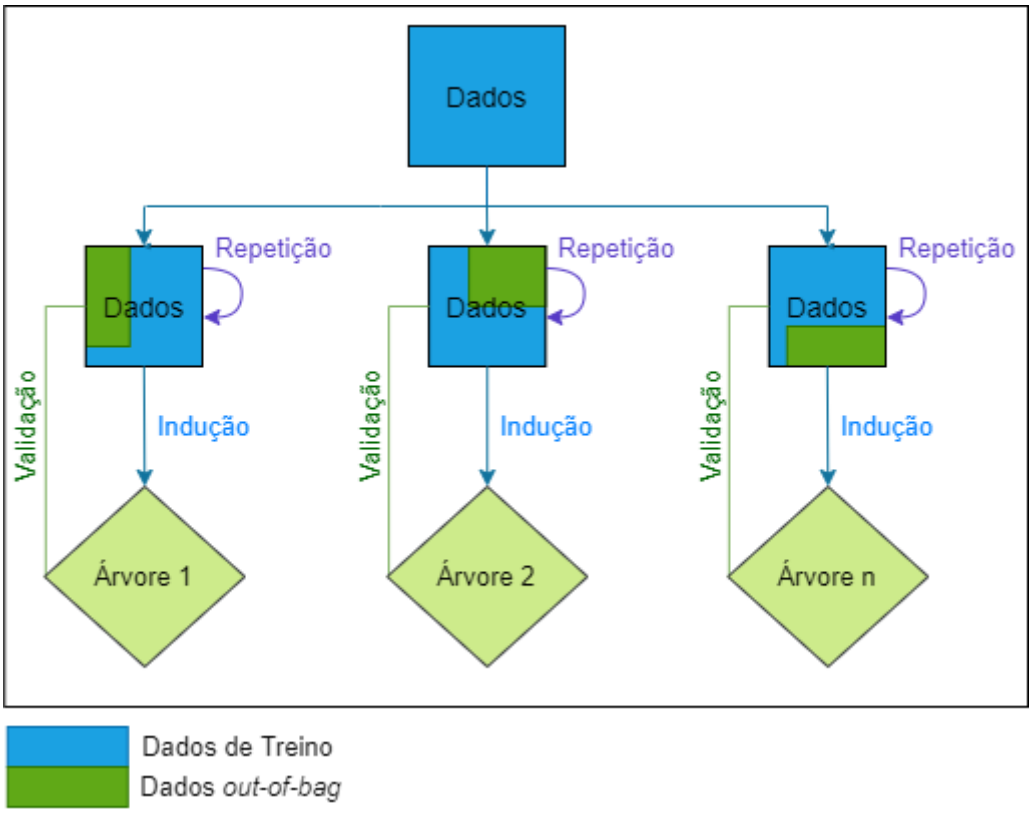
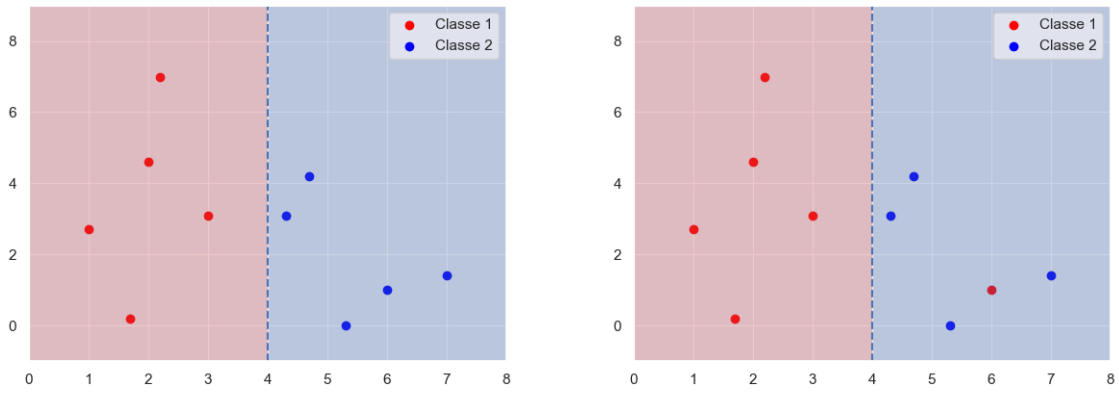


Figura 3 – Exemplo de Bootstraping

A indução das árvores do modelo apresentado neste projeto utiliza o critério de minimização de impureza, onde a medida de impureza escolhida foi o *Gini Index*. O *Gini Index* define a pureza de um nó na árvore, ou seja, o quanto este nó conseguiu separar uma classe das outras. A Figura 4 mostra a divisão no espaço causado por um nó de divisão no atributo $x > 4$. Um nó é totalmente puro quando divide um conjunto de exemplos

em subconjuntos com apenas exemplos de uma única classe. Um nó é impuro quando a divisão resulta em subconjuntos com exemplos de mais de uma classe



(a) Divisão pura (b) Divisão levemente impura

Figura 4 – Exemplos de divisões puras e impuras

A medida *Gini Index* é dada pela Equação 2.1, onde C representa o conjunto das classes do problema e $P(c|t)$ representa a probabilidade da classe c pertencer à partição t. Quanto menor o valor do *Gini Index*, mais pura é a partição.

$$GiniIndex(t) = 1 - \sum_{c \in C} [P(c|t)]^2 \tag{2.1}$$

Durante a indução das árvores, o melhor atributo é escolhido para criar uma divisão. A qualidade de uma divisão pode ser medida utilizando a média ponderada de Gini, e é dada pela Equação 2.2.

$$Gini_{ponderada} = \sum_{t \in T} \frac{N_t}{N} \cdot GiniIndex(t) \tag{2.2}$$

Onde N_t é o número de elementos na partição t, N o número de elementos no nó pai e T o conjunto das partições t geradas à partir da divisão atual. Quanto menor o valor da média ponderada de Gini, maior a qualidade da divisão.

Como exemplo, foi gerada uma Árvore de Decisão usando os dados fictícios da Tabela 1 utilizando o critério de parada 'particionamento perfeito dos dados'. O atributo e valor de comparação que possuem a menor medida de Gini Ponderada é o atributo e valor $SNP3 = C$, com o valor de Gini Ponderada $0,285$. Portanto, eles farão parte da primeira divisão da árvore. Como a partição onde $SNP3 = C$ possui medida de Gini Index 0, temos um nó terminal, onde a classe predita é a classe 'Alto'. Para a partição $SNP3 \neq C$, ainda temos um valor de Gini Index de $0,5$. Portanto, procuramos o próximo atributo e valor com menor medida de Gini Ponderado. Os atributos e valores $SNP1 = C$ e

Genótipo	SNP1	SNP2	SNP3	Classe
1	A	T	C	Alto
2	C	T	T	Baixo
3	C	T	A	Intermediário
4	A	G	C	Alto
5	T	G	C	Alto
6	C	G	T	Intermediário
7	A	T	T	Baixo

Tabela 1 – Exemplo de dados de treino para a geração de uma Árvore de Decisão

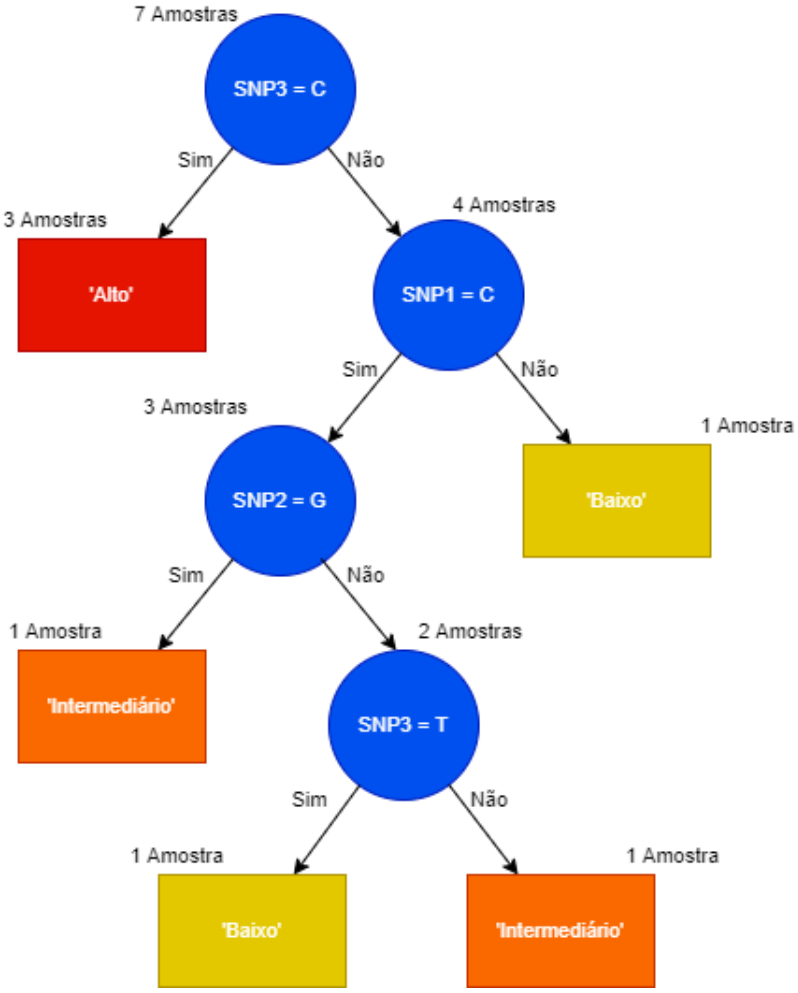


Figura 5 – Exemplo de indução de uma Árvore de Decisão

$SNP3 = T$ geram divisões com o mesmo valor de Gini Ponderada. Neste caso, escolhemos arbitratriamente a divisão $SNP1 = C$, com valor de Gini Ponderada $0,33$. Na partição $SNP1 = C$, temos um nó com Gini Index de $0,44$, e na partição $SNP1 \neq C$, temos uma partição pura com a classe 'Baixo'. A próxima divisão com menor Gini Ponderada é a divisão $SNP2 = G$, com Gini Ponderada de $0,33$. Com isso, separamos uma partição pura com a classe 'Intermediário' e uma partição impura com valor $0,5$ de Gini Index. Por fim, nossa última divisão acontece no atributo e valor $SNP3 = T$, com valor de Gini

Ponderada de 0, separando os dados em duas partições puras, induzindo completamente a Árvore de Decisão de exemplo (Figura 5). Note que a indução da árvore só parou quando todos as amostras do conjunto de dados foram particionadas corretamente. Este é o critério de parada conhecido como 'particionamento perfeito'.

2.5 Mean Decrease in Gini Index

A estratégia *Mean Decrease in Gini Index* (MDGI) é utilizada para medir a importância de cada atributo em uma RF. O método consiste em aleatorizar os valores dos atributos em nós de divisão, com o objetivo de verificar o quanto essa mudança afeta negativamente o desempenho da árvore. Atributos que causarem maior decrescimento no desempenho da árvore são considerados mais importantes, enquanto atributos que, ao serem aleatorizados, não causam alteração significativa, são considerados menos importantes. O MDGI calcula a média do decrescimento de cada atributo em cada uma das árvores da RF. Quanto maior o valor do MDGI para um dado atributo, maior sua importância.

Seja $MDGI(X_j)$ o cálculo da importância do atributo X_j para a predição de classes y. Temos que $p(t)\Delta i(t)$ é o decrescimento da impureza ponderada, onde $p(t)$ é a proporção N_t/N e N_t é o número de amostras que chegam na partição t. Sejam N e N_T o número de amostras do nó pai e o número de árvores na floresta respectivamente. Seja também $v(t)$ o atributo utilizado na partição t, A o conjunto de árvores da floresta e $\Delta i(t)$ o cálculo da variação do *Gini Index* para a partição t [Louppe et al. 2013]. Assim temos:

$$MDGI(X_j) = \frac{1}{N_T} \sum_{a \in A} \sum_{t \in a : v(t) = X_j} p(t)\Delta i(t) \tag{2.3}$$

Como exemplo de uso, foi aplicada a estratégia MDGI sob a Árvore de Decisão exemplo da Figura 5 construída à partir dos dados da Tabela 1. A aplicação vai ser feita considerando o caso trivial de uma RF com apenas uma árvore. Dada a árvore da Figura 5, a importãncia do SNP1 é calculado aleatorizando seu valor de divisão para, por exemplo, o valor A. Caso essa aleatorização ocorra, esta divisão continuará separando os atributos em uma partição com 3 amostras e outra com 1. Portanto, não houve decrescimento do Gini para a SNP1. Aleatorizando o valor da divisão da SNP2 para o valor T, a divisão continuará criando uma partição de 2 amostras e outra de 1. Portanto, não houve decrescimento do Gini para este SNP. Por fim, aleatorizando a primeira divisão do SNP3 na árvore para o valor A, resultamos em uma partição com 1 amostra e outra com 7. O Gini neste nó subiu para $0,571$, aumentando seu valor de impuridade em $0,286$. Considerando que a aleatorização do último nó da árvore iria gerar duas partições de um elemento cada, a medida de decrescimento para este nó continua sendo nula. Com isso, temos que o único SNP considerado importante, foi o atributo SNP3, com uma medida de MDGI de $0,286$.

3 Trabalhos Correlatos

Já se encontram trabalhos na literatura que exploraram o uso de AM para a tarefa de Seleção de SNPs. Em [Matukumalli et al. 2006] foi descrito o desenvolvimento do PolyBayes, um algoritmo de AM que é capaz de detectar SNPs de soja. Este algoritmo foi treinado usando conjuntos de dados de seis cultivares de soja homozigotas distintas. De forma similar a este projeto, [Matukumalli et al. 2006] também investiga as vantagens do uso de AM no problema de Seleção de SNPs. A pesquisa mostrou que métodos de AM se tornaram muito mais efetivos que os métodos considerados específicos para o problema de seleção. Foi também desenvolvida uma ferramenta, baseada no algoritmo de indução de árvores de decisão C4.5, que realiza a seleção de SNPs de forma genérica.

Em [Hess et al. 2017], foi estudado o uso de *Deep Boltzmann Machines* para aplicações de SNPs. O problema de usar SNPs como dados de entrada em preditores é que o número de SNPs pode ser muito grande, aumentando a dimensionalidade do problema. Este projeto buscou métodos de evoluir *Deep Boltmann Machines* para particionar o problema em diversos subespaços.

Uma das doenças mais estudadas da atualidade é a doença de Alzheimer. Alzheimer é uma doença cerebral identificada pela lenta e progressiva falha de memória, confusão e até morte. Em [Sherif, Zayed e Fakhr 2015], através do uso de Redes Bayesianas, foram descobertos SNPs altamente relacionados com a tendência ao Alzheimer em humanos. A descoberta foi uma contribuição para o estudo terapêutico da doença.

Em [Edelenyi et al. 2008] a abordagem de *Random Forests* foi utilizada para tratar da Seleção de SNPs da síndrome metabólica (associada à obesidade, aumento da triglicérides, pressão alta e outras doenças metabólicas). Apesar de ter obtido três SNPs promissores para a síndrome metabólica, a utilização de um único modelo *Random Forest* se tornou uma abordagem menos robusta da proposta neste projeto.

O estudo dos SNPs fortemente relacionados com o fator de crescimento de gado foi realizado em [Li et al. 2018]. Nesta pesquisa, os algoritmos *Random Forest* e *Gradient Boosting Machine* foram considerados eficientes para a Seleção de SNPs. Apesar do *Gradient Boosting Machine* ter resultados minimamente melhores, seu tempo de processamento é cerca de 12 vezes maior do que o da *Random Forest*.

Neste projeto, a proposta de um método robusto baseado em *Random Forests* leva a uma melhoria nos resultados, se aproximando ainda mais do desempenho obtido pelo *Gradient Boosting Machine* de [Li et al. 2018] em um período de tempo menor.

4 Detalhamento do Desenvolvimento

4.1 Seleção de SNPs

A Seleção de SNPs pode ser modelada como um problema de Aprendizado de Máquina Multi-Classe pelo fato de uma característica escolhida para a análise ter diversas possíveis classes atribuídas. Como exemplo, temos o teor de amilose do arroz, que pode ser classificado como alto, intermediário, baixo e muito baixo.

Independentemente do número de classes no problema, serão gerados conjuntos binários para cada conjunto Multi-Classe. Isto é feito para que as análises da Seleção de SNPs encontrem SNPs importantes para distinguir uma classe das outras. Exemplificando, poderemos encontrar SNPs que distinguem o teor de amilose alto dos outros, baixo dos outros, e assim por diante.

A tarefa de Seleção de SNPs deve levar em conta que existem características de organismos que dependem de apenas um SNP, e outras que dependem de uma quantidade maior. O modelo deve se adaptar à busca de não somente um SNP importante, e sim todos os possíveis SNPs associados à característica analisada.

4.2 Descrição dos Dados

Os conjuntos de dados utilizados nessa pesquisa foram fornecidos pela Embrapa. Eles contêm os SNPs de diferentes genótipos de arroz colhidos em diferentes regiões e anos, classificados em diferentes classes de teor de amilose.

A Tabela 2 apresenta os conjuntos de dados utilizados para a análise de SNPs do arroz. Todos os conjuntos possuem 536 amostras e 4709 SNPs levados em consideração. Por possuírem 4 classes de teor de amilose diferentes (Alto, Intermediário, Baixo e Muito Baixo), os conjuntos podem ser modelados como problemas de Aprendizado de Máquina Multi-Classe. O conjunto de dados conta com classes desbalanceadas, principalmente da classe 'Muito Baixo'.

Os SNPs assumem valores A, C, T ou G, sendo assim, atributos categóricos. Os mesmos foram enumerados de 1 até 4709 para facilitar a compreensão por este documento. Essa nomenclatura não condiz com seus nomes oficiais na literatura.

Conjunto de dados	Amostras	Atributos	Alto	Intermediário	Baixo	Muito Baixo
Boa Vista 2004	536	4709	205	281	45	5
Goiania 2004	536	4709	309	195	24	8
Goiania 2005	536	4709	251	235	43	7
Pelotas 2005	536	4709	99	397	36	4
Teresina 2006	536	4709	398	105	28	5
Uruguaiana 2004	536	4709	58	429	45	4
Vilhena 2006	536	4709	268	246	19	3

Tabela 2 – Conjuntos de dados utilizados

4.3 O Método Proposto

O método proposto nessa pesquisa para realizar a tarefa de Seleção de SNPs chama-se *EnGENE*. Apesar de aplicado aqui especificamente para a tarefa de seleção de SNPs importantes para o teor de amilose do arroz, ele é genérico e pode ser aplicado para qualquer organismo.

O método *EnGENE* possui funções para transformação dos dados como a transformação Um-Contra-Todos (Tabela 4), que transforma o problema multi-classe para binário, e a transformação Um-Atributo-por-Valor (Tabela 5), que transforma atributos categóricos em vários binários, um para cada categoria. A transformação Um-Contra-Todos é utilizada para transformar um problema Multi-Classe em vários problemas binários, possibilitando que o método consiga discriminar os SNPs relevantes entre classes distintas. A transformação Um-Atributo-por-Valor é importante para passar os dados de entrada para o formato binário, o qual o método exige como entrada. Portanto o método sempre trata o problema como classificação binária (um problema binário para cada classe), e sempre trata os atributos como binários.

SNP1	SNP2	...	Classe
A	T	...	Alto
C	T	...	Baixo
C	T	...	Intermediário
A	G	...	Alto

Tabela 3 – Exemplo de dados para ser aplicado as transformações

O método *EnGENE* consiste em e execuções, onde uma RF distinta é induzida, utilizando diferentes separações do conjunto de dados, em cada execução. No fim de cada execução, o método *EnGENE* captura o SNP mais importante utilizando o método MDGI na floresta gerada.

A medida de pontuação *Score* de uma dada SNP i consiste no número de vezes que esta SNP foi considerada a mais importante de uma execução e. Na Figura 6 temos a

SNP1	**SNP2**	**...**	**Classe**
A	T	...	Alto
C	T	...	Outro
C	T	...	Outro
A	G	...	Alto

Tabela 4 – Resultado da transformação Um-Contra-Todos, tendo a classe 'Alto' como principal.

SNP1-A	**SNP1-C**	**SNP2-T**	**SNP2-G**	**...**	**Classe**
1	0	1	0	...	Alto
0	1	1	0	...	Baixo
0	1	1	0	...	Intermediário
1	0	0	1	...	Alto

Tabela 5 – Resultado da transformação Um-Atributo-por-Valor.

esquematização do processo de treino e pontuação de um método α_j. Ao fim de cada execução, o SNP considerado mais importante recebe +1 em sua medida de *Score*. Tomando a Figura 6 como exemplo, temos que o SNP3 possui valor de *Score* 16. Portanto, ele foi julgado, pelo método MDGI, como o SNP mais importante em 16 execuções diferentes.

Apesar da transformação Um-Atributo-por-Valor criar novos atributos de acordo com a quantidade de valores que cada SNP tenha, o método *EnGENE* trata o SNP em si como importante, desconsiderando seu valor. Exemplificando, se temos que o SNP1 foi considerado importante tendo valores diferentes (A e C) em duas execuções distintas, o valor de *Score* do SNP1 será 2, pois ele apareceu duas vezes como o mais importante, independentemente de seu valor.

A Figura 7 ilustra o processo dentro de uma etapa do método proposto. Em cada execução e, é gerada uma RF com um subconjunto aleatório dos dados. O método MDGI é aplicado para extrair o SNP mais importante em cada execução.

Os dados do arroz que foram analisados estão separados em regiões de cultivo e safras diferentes. Tais dados de diferentes regiões podem revelar padrões evolutivos distintos. Esses novos padrões devem seguir o conjunto de SNPs que descreve o teor de amilose. Portanto, os SNPs que aparecem mais frequentemente em métodos distintos são mais valiosos por serem considerados importantes até mesmo em padrões evolutivos distintos.

Levando isto em conta, foi introduzida a medida de *Score Cruzado*, uma nova métrica de pontuação de SNPs que leva em consideração sua importância em outros conjuntos de dados. O Algoritmo 1 mostra o cálculo do valor de Score Cruzado. O procedimento recebe um vetor de métodos *EnGENE* α treinados e aplica um multiplicador à soma de *Score* de um SNP i em todos os conjuntos de dados. O valor do multiplicador aumenta

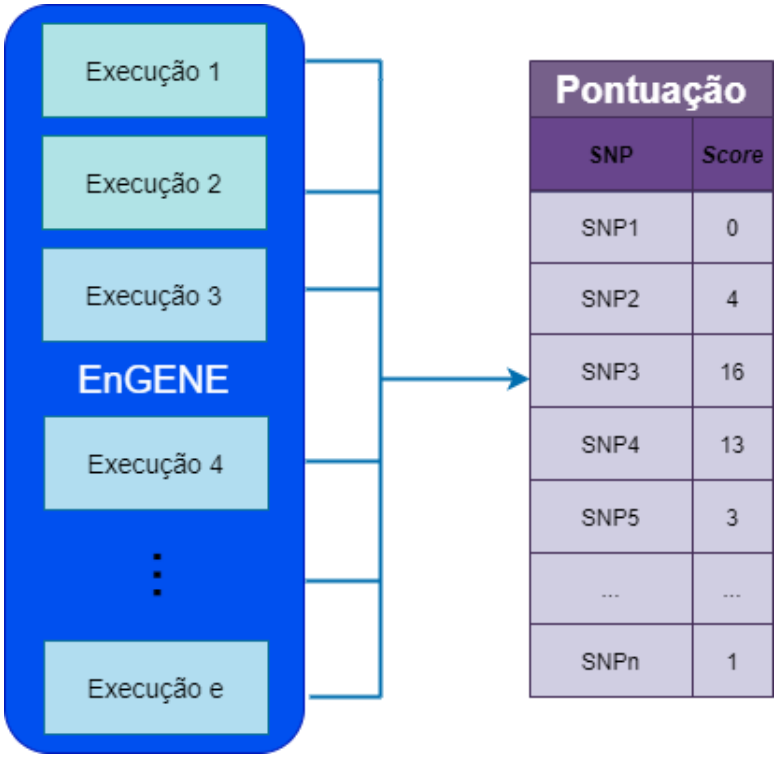


Figura 6 – Treino e pontuação com o método EnGENE

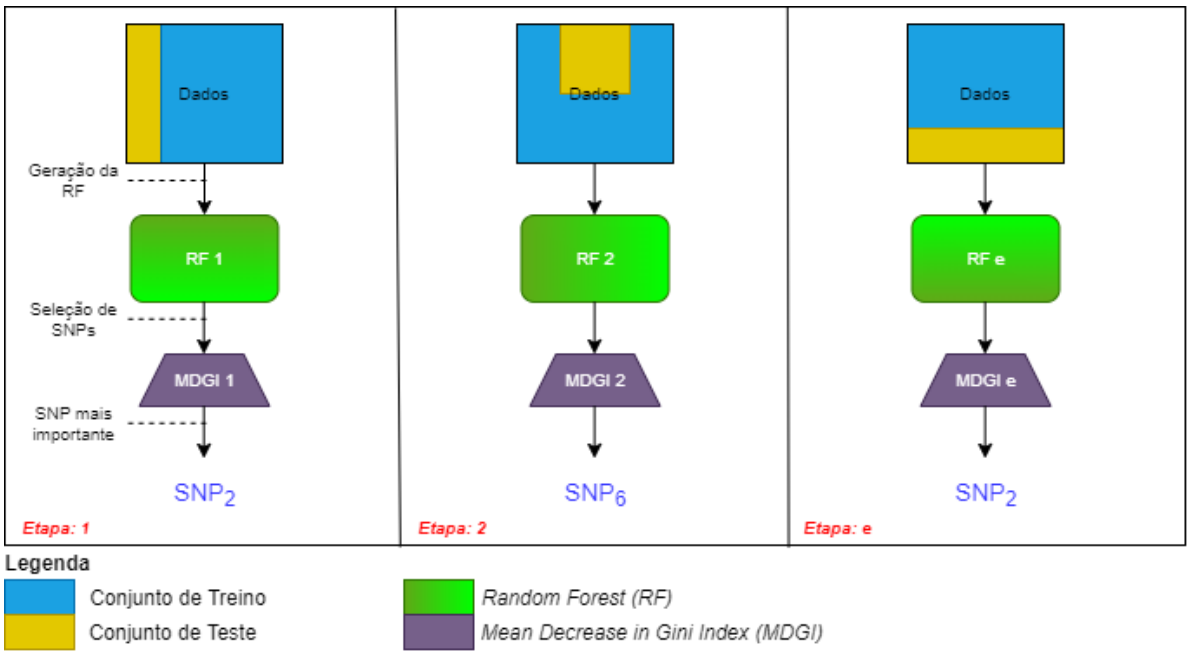


Figura 7 – Ordem de processamento interna do método

pela constante 0.3 para cada conjunto de dados que o SNP i apareceu como importante. A escolha da constante foi feita de forma que fosse possível destacar os SNPs mais relevantes em diferentes modelos sem desbalancear significantemente a distribuição de pontos em relação aos outros SNPs. Ainda no Algoritmo 1, a função Beta retorna a quantidade de modelos nos quais o SNP i apareceu como importante ao menos uma vez, e a função Score retorna o valor *Score* para um dado SNP i no método α_j.

Algorithm 1: Score Cruzado

```
procedure ScoreCruzado(EngeneVector α, int i) returns float;
    appearances ← Beta(α, i)
    multiplier ← 1 + (0.3 · appearances)
    newScore ← 0
    for j = 0; j < length(α); j++ do
        newScore ← newScore + Score(α_j, i)
    end
    newScore ← newScore · multiplier
    return newScore
```

Matematicamente, definimos o cálculo do *Score Cruzado* na Equação 4.1, onde $|\alpha|$ representa o número de modelos no vetor α.

$$ScoreCruzado(\alpha, i) = (1 + (0,3 \cdot Beta(\alpha, i))) \cdot \sum_{j=1}^{|\alpha|} Score(\alpha_j, i) \tag{4.1}$$

Assim que treinados, as métricas de validação de Precisão e Revocação são atualizadas como a média entre cada execução (Equações 4.4 e 4.5). Nas Equações 4.2 e 4.3, tp representa a quantidade de genótipos positivos classificados como positivos, fp a quantidade de genótipos negativos classificados como positivos, e fn a quantidade de genótipos positivos classificados como negativos. As métricas de validação são referentes às RFs de um método α_j em execuções e, onde E é o número total de execuções.

$$Precisao(\alpha_{je}) = \frac{tp}{tp + fp} \tag{4.2}$$

$$Revocacao(\alpha_{je}) = \frac{tp}{tp + fn} \tag{4.3}$$

$$PrecisaoMedia(\alpha_j) = \frac{\sum_{e=1}^{E} Precisao(\alpha_{je})}{E} \tag{4.4}$$

$$RevocacaoMedia(\alpha_j)) = \frac{\sum_{e=1}^{E} Revocacao(\alpha_{je})}{E} \tag{4.5}$$

4.4 Experimentos

Nesta etapa de experimentos, os conjuntos de dados do arroz foram passados pela transformação Um-Contra-Todos. Desta transformação, foram criados 4 métodos a partir de cada conjunto de dados, cada um resultado da transformação Um-Contra-Todos para cada classe do teor de amilose. Os conjuntos resultantes passaram pela transformação Um-Atributo-por-Valor, transformando seus atributos em binários, para que o método *EnGENE* consiga interpretá-los. Cada um destes conjuntos de dados foram utilizados para criar métodos *EnGENE* distintos. A Tabela 6 mostra todos os problemas binários gerados pela transformação Um-Contra-Todos nos conjuntos de dados da Tabela 2, apresentada anteriormente. A Tabela também descreve o número de amostras em cada classe, sendo a classe principal o teor de amilose isolado no problema e a classe 'Outros' definida como os teores de amilose restantes.

Após as transformações, cada método passou por 1000 execuções, em separações aleatórias e estratificadas, onde 90% dos dados eram para treinamento e 10% para teste. Cada execução gerava uma RF com 100 árvores induzidas. Os métodos treinados que faziam referência à mesma classe de teor de amilose foram sujeitos ao procedimento de cálculo do *Score Cruzado*. Com isso, foram obtidas listas de SNPs ranqueadas por seu Score Cruzado para cada teor de amilose. O cálculo do *Score Cruzado* também foi feito entre todos os métodos, independentemente do teor de amilose, para encontrar os SNPs mais relevantes como um todo para a distinção do teor de amilose.

Conjunto de dados	**Amostras**	**# da classe principal**	**# de Outros**
Boa Vista 2004 - Alto	536	205	331
Boa Vista 2004 - Intermediário	536	281	255
Boa Vista 2004 - Baixo	536	45	491
Boa Vista 2004 - Muito Baixo	536	5	531
Goiania 2004 - Alto	536	309	227
Goiania 2004 - Intermediário	536	195	341
Goiania 2004 - Baixo	536	24	512
Goiania 2004 - Muito Baixo	536	8	528
Goiania 2005 - Alto	536	251	285
Goiania 2005 - Intermediário	536	235	301
Goiania 2005 - Baixo	536	43	493
Goiania 2005 - Muito Baixo	536	7	529
Pelotas 2005 - Alto	536	99	437
Pelotas 2005 - Intermediário	536	397	139
Pelotas 2005 - Baixo	536	36	500
Pelotas 2005 - Muito Baixo	536	4	532
Teresina 2006 - Alto	536	398	138
Teresina 2006 - Intermediário	536	105	431
Teresina 2006 - Baixo	536	28	508
Teresina 2006 - Muito Baixo	536	5	531
Uruguaiana 2004 - Alto	536	58	478
Uruguaiana 2004 - Intermediário	536	429	107
Uruguaiana 2004 - Baixo	536	45	491
Uruguaiana 2004 - Muito Baixo	536	4	532
Vilhena 2006 - Alto	536	268	268
Vilhena 2006 - Intermediário	536	246	290
Vilhena 2006 - Baixo	536	19	517
Vilhena 2006 - Muito Baixo	536	3	533

Tabela 6 – Conjuntos de dados transformados

5 Resultados

As Tabelas 7, 8, 9 e 10 apresentam os valores médios das medidas de avaliação em cada classe de teor de amilose após 1000 execuções em cada conjunto de dados. A medida *Razão de Classes* representa o número de amostras do teor de amilose isolado no problema binário dividido pelo número de amostras na classe 'Outros'. Neste caso, um conjunto de dados é balanceado, ou seja, possui o mesmo número de amostras em ambas as classes, quando o valor da *Razão de Classes* for $1,0$.

Método	**Precisão Média**	**Revocação Média**	**Razão das Classes**
Boa Vista 2004 - Alto	76.893%	76.893%	0,619
Goiania 2004 - Alto	62.169%	62.169%	1,361
Goiania 2005 - Alto	68.583%	68.583%	0,88
Pelotas 2005 - Alto	82.181%	82.181%	0,226
Teresina 2006 - Alto	76.709%	76.709%	2,884
Uruguaiana 2004 - Alto	88.989%	88.989%	0,121
Vilhena 2006 - Alto	63.441%	63.441%	1,0

Tabela 7 – Medidas de avaliação para os conjuntos de teor alto

Método	**Precisão Média**	**Revocação Média**	**Razão das Classes**
Boa Vista 2004 - Inter.	74.341%	74.341%	1,101
Goiania 2004 - Inter.	64.404%	64.404%	0,571
Goiania 2005 - Inter.	66.393%	66.393%	0,78
Pelotas 2005 - Inter.	75.461%	75.461%	2,856
Teresina 2006 - Inter.	79.685%	79.685%	0,243
Uruguaiana 2004 - Inter.	81.215%	81.215%	4,009
Vilhena 2006 - Inter.	62.406%	62.406%	0,848

Tabela 8 – Medidas de avaliação para os conjuntos de teor intermediário

Método	**Precisão Média**	**Revocação Média**	**Razão das Classes**
Boa Vista 2004 - Baixo	91.813%	91.813%	0,091
Goiania 2004 - Baixo	96.211%	96.211%	0,046
Goiania 2005 - Baixo	92.809%	92.809%	0,087
Pelotas 2005 - Baixo	92.744%	92.744%	0,072
Teresina 2006 - Baixo	94.378%	94.378%	0,055
Uruguaiana 2004 - Baixo	91.839%	91.839%	0,091
Vilhena 2006 - Baixo	96.296%	96.296%	0,036

Tabela 9 – Medidas de avaliação para os conjuntos de teor baixo

Método	**Precisão Média**	**Revocação Média**	**Razão das Classes**
Boa Vista 2004 - M.B.	98.148%	98.148%	0,009
Goiania 2004 - M.B.	98.146%	98.146%	0,015
Goiania 2005 - M.B.	98.146%	98.146%	0,013
Pelotas 2005 - M.B.	99.998%	99.998%	0,007
Teresina 2006 - M.B.	98.148%	98.148%	0,009
Uruguaiana 2004 - M.B.	100.000%	100.000%	0,007
Vilhena 2006 - M.B.	100.000%	100.000%	0,005

Tabela 10 – Medidas de avaliação para os conjuntos de teor muito baixo

É possível analisar que as medidas de Precisão Média e Revocação Média são iguais em todos os modelos. Essa avaliação nos diz que o valor de falsos negativos é igual ao de falsos positivos. Vemos também que conjuntos de dados mais desbalanceados (com Razão das Classes muito menor ou muito maior que 1) possuem Precisão Média e Revocação Média maior. Isto vem do fato do modelo ter pouca informação sobre uma das classes para conseguir distinguir com total certeza suas amostras, acertando a maioria de suas predições como a classe predominante.

A análise de SNPs dos genótipos de arroz evidenciou uma lista de SNPs importantes para a distinção dos teores de amilose. As Figuras 8 até 11 representam os 10 SNPs mais bem pontuados para cada teor de amilose. A nomenclatura dos SNPs é apenas uma enumeração sequencial fornecida pela Embrapa. Esta identificação não reflete o nome oficial do SNP, encontrado na literatura.

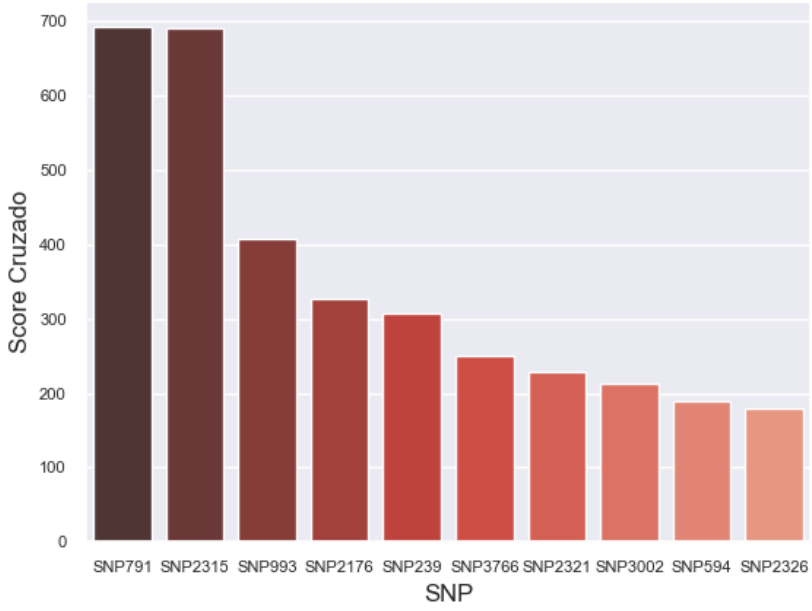


Figura 8 – SNPs importantes para distinguir o teor alto de amilose

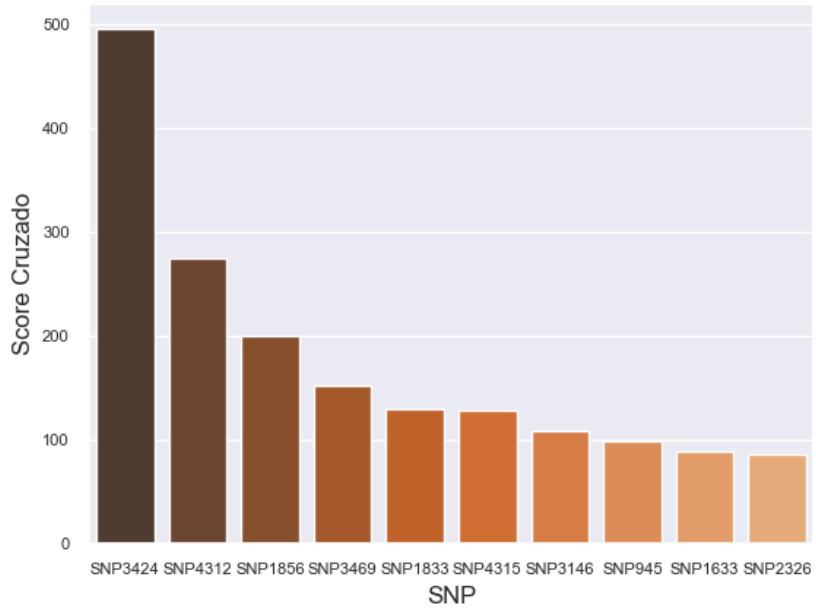


Figura 9 – SNPs importantes para distinguir o teor intermediário de amilose

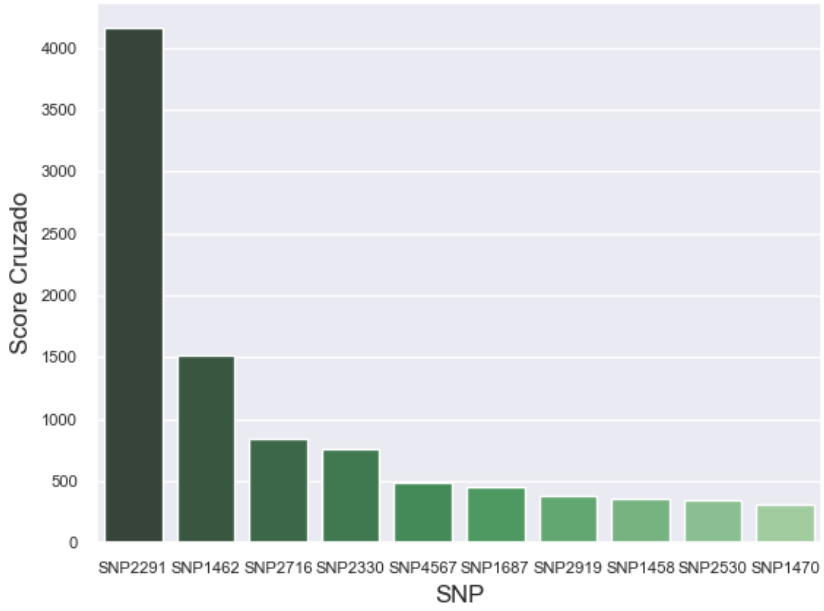


Figura 10 – SNPs importantes para distinguir o teor baixo de amilose

É possível analisar que os teores alto, intermediário e baixo possuem SNPs considerados bastante importantes para sua classificação, enquanto o teor muito baixo possui SNPs com *Score Cruzado* muito similar. Isso se dá ao fato de existirem poucas amostras de teor muito baixo nos conjuntos de dados, levando a uma predição desbalanceada e, consequentemente, maior dificuldade do modelo em encontrar os SNPs mais importantes na distinção do teor

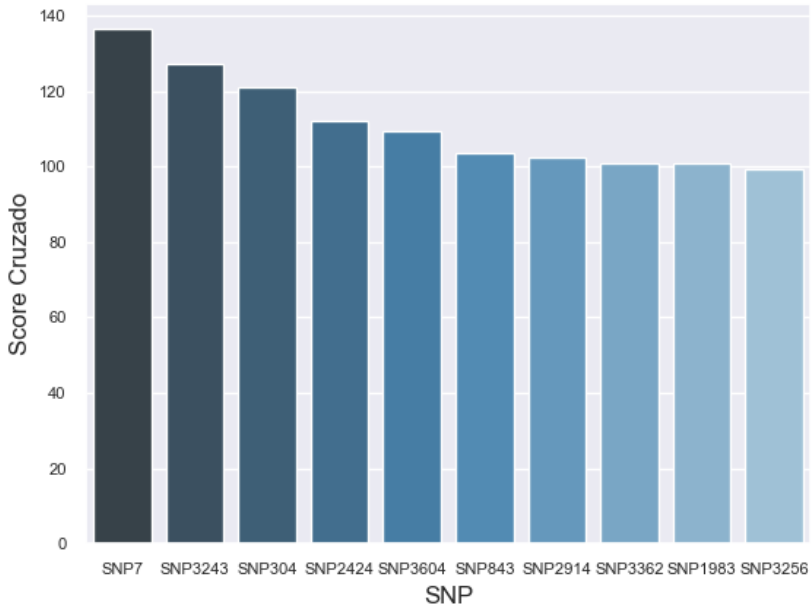


Figura 11 – SNPs importantes para distinguir o teor muito baixo de amilose

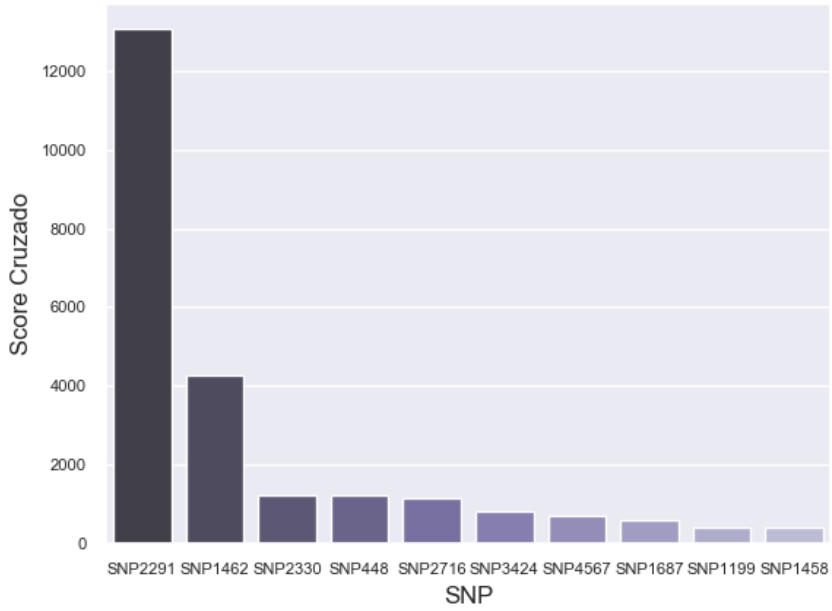


Figura 12 – SNPs importantes para o teor de amilose

de amilose muito baixo.

O cálculo do *Score Cruzado* entre todos os métodos, independentemente do teor de amilose, nos proporcionou uma lista dos SNPs mais importantes para a definição do teor de amilose como um todo (Figura 12). Os valores de *Score Cruzado* são bem maiores do que a análise individual de cada classe. Isto se dá pelo fato de existirem mais conjuntos

de dados levados em consideração na análise geral para calcular o *Score Cruzado*, dando uma pontuação maior aos SNPs que aparecem em diferentes regiões nas análises de teores de amilose diferentes. Estas 10 SNPs importantes para o teor de amilose do arroz foram reportadas para a *Embrapa* para serem validadas experimentalmente.

Um SNP em específico foi selecionado como um forte candidato para discriminar o teor de amilose baixo dos altos e intermediários. Este é o SNP1462. Na análise geral dos métodos *EnGENE*, o SNP atingiu o valor de aproximadamente 4000 *Score Cruzado* na análise geral e um *Score Cruzado* de aproximadamente 1500 na análise do teor baixo. Este SNP foi o segundo SNP mais bem votado no ranqueamento geral e no ranqueamento dos teores baixos. A validação experimental confirma a análise do método.

Estes resultados deixaram os pesquisadores da Embrapa empolgados. A criação de um novo método que, de forma rápida, encontra SNPs importantes para uma dada característica é muito mais efetivo do que os métodos atuais. Agora a Seleção de SNPs pode ser realizada de forma mais direta e menos custosa.

Com o propósito de facilitar as análises feitas com o método *EnGENE*, principalmente para profissionais da área da biotecnologia, a interface gráfica *Rosalind* foi desenvolvida. Seu nome é uma homenagem à cientista britânica *Rosalind Franklin*, conhecida pela descoberta da estrutura do DNA. Os procedimentos inclusos na interface *Rosalind* permitem o uso do método *EnGENE* para tratar o problema da Seleção de SNPs.

As Figuras 13 e 14 representam a interface do *Rosalind*. Dentro do ambiente gráfico, é possível criar métodos EnGENE para conjuntos de dados carregados, selecionar os SNPs que serão levados em consideração na análise, transformar o problema de Multi-Classe para binário com a transformação Um-Contra-Todos, transformar os atributos de categóricos para binários com a transformação Um-Atributo-por-Valor (*"dummification"*, como consta na Figura 13 é o nome popular para a transformação Um-Atributo-por-Valor). Ao carregar e pré-processar os conjuntos de dados, os métodos podem ser treinados, obtendo assim resultados como: A precisão e revocação média do método, a quantidade de vezes treinadas e a lista de SNPs importantes ordenadas pelo *Score*. Métodos já treinados podem ter suas pontuações de *Score* recalculadas na métrica de *Score Cruzado*, criando uma nova lista de SNPs ordenados, mas desta vez, pelo *Score Cruzado*. Os pesquisadores da Embrapa estão trabalhando no desenvolvimento de uma interface web para o método *EnGENE*.

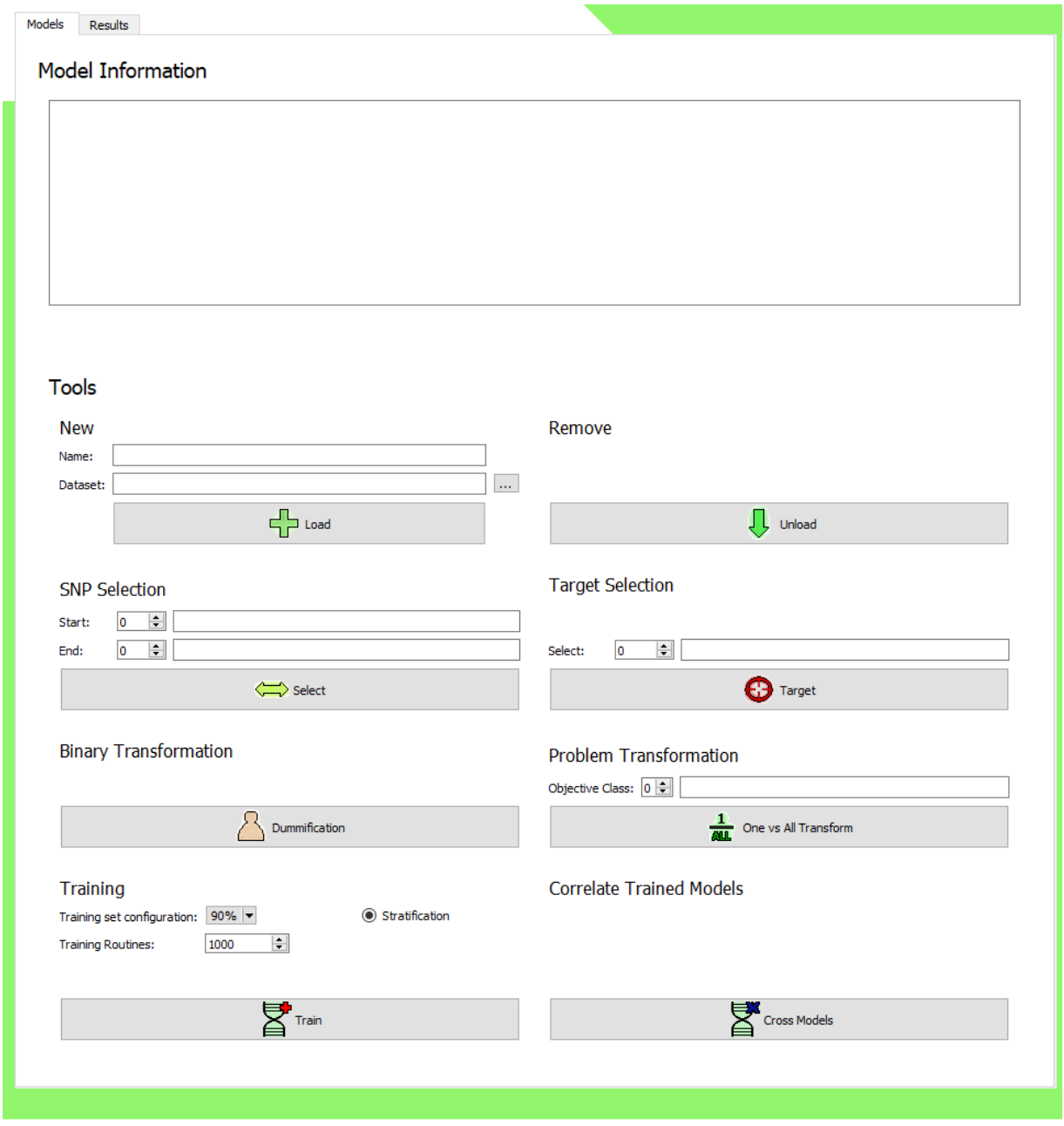


Figura 13 – Interface Rosalind vazia

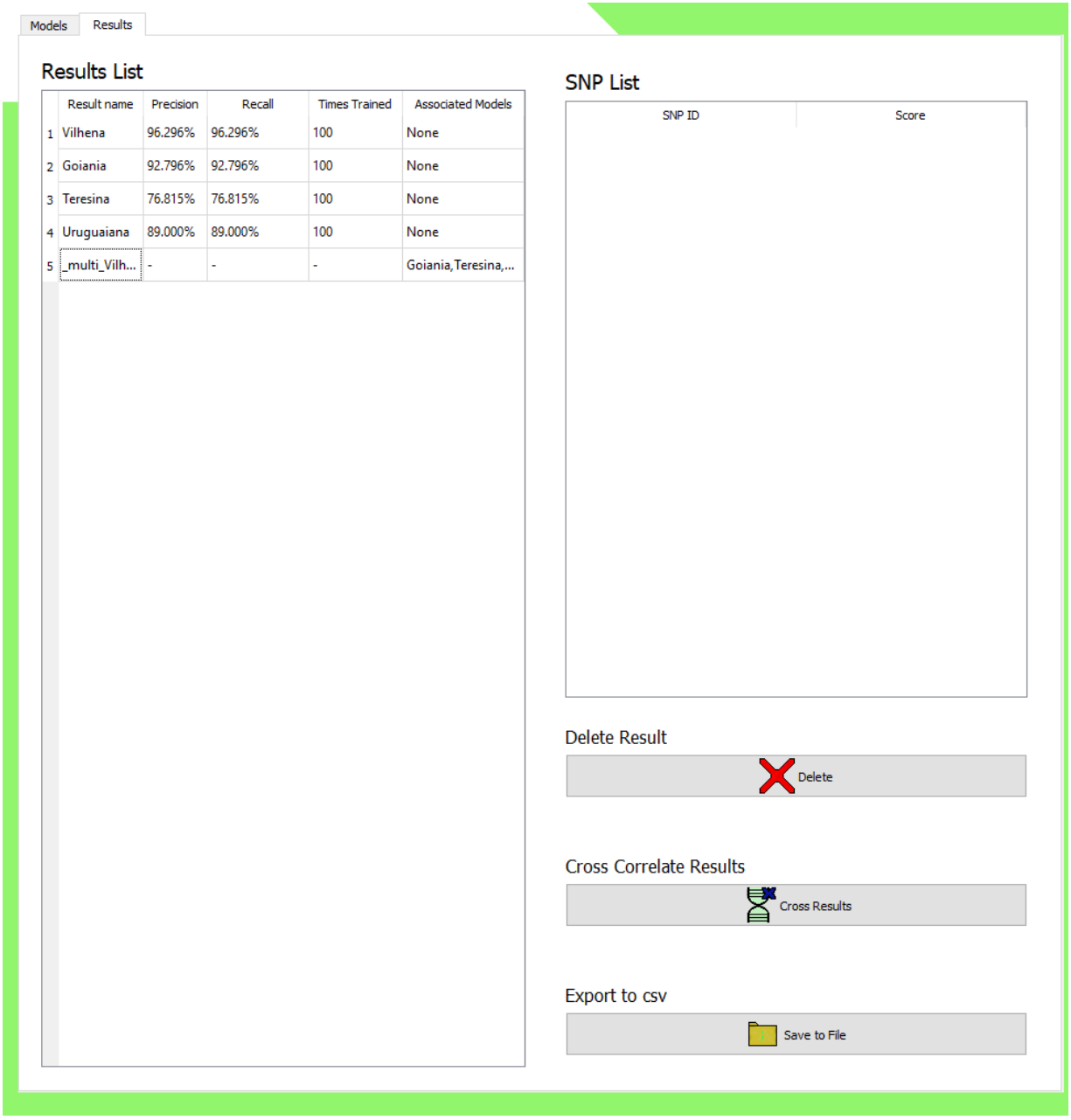


Figura 14 – Interface Rosalind com dados treinados

6 Conclusão

O método *EnGENE* se mostrou mais eficaz e rápido na tarefa de Seleção de SNPs comparado aos métodos aplicados até o momento pela Embrapa. Suas análises apontaram um candidato a novo marcador SNP fortemente relacionado à distinção do teor baixo de amilose dos teores alto e intermediário. Este novo marcador SNP, se confirmado em análises experimentais, entrará no programa de Melhoramento Genético da Embrapa. Os pesquisadores da Embrapa estão animados com os resultados mais efetivos do método, e procuram expandir as análises feitas com o mesmo para outros organismos, no intuito de acelerar o processo de Melhoramento Genético.

Com o intuito de facilitar a realização de futuras análises com o método, um ambiente gráfico foi desenvolvido para a manipulação de dados e de métodos *EnGENE*. Este *Software* simplificará a aplicação da metodologia proposta neste projeto para outras características de outros organismos. Como projetos futuros, considero o uso dos modelos *EnGENE* para analisar outras características do arroz, ou até mesmo outros organismos do Programa de Melhoramento Genético da Embrapa.

Referências

BATISTA, C. d. S. *Desenvolvimento de arroz integral de cozimento rápido: propriedades físico-químicas, tecnológicas e digestibilidade do amido*. Dissertação (Mestrado) — Universidade Federal de Pelotas, 2019. Citado na página 19.

BATNYAM, N.; GANTULGA, A.; OH, S. An efficient classification for single nucleotide polymorphism (snp) dataset. In: *Computer and Information Science*. [S.l.]: Springer, 2013. p. 171–185. Citado na página 19.

EDELENYI, F. S. de et al. Prediction of the metabolic syndrome status based on dietary and genetic parameters, using random forest. *Genes & nutrition*, v. 3, n. 3, p. 173, 2008. Citado na página 29.

HESS, M. et al. Partitioned learning of deep boltzmann machines for snp data. *Bioinformatics*, Oxford University Press, v. 33, n. 20, p. 3173–3180, 2017. Citado na página 29.

KORRES, N. et al. Temperature and drought impacts on rice production: An agronomic perspective regarding short-and long-term adaptation measures. *Water resources and rural development*, Elsevier, v. 9, p. 12–27, 2017. Citado na página 19.

LI, B. et al. Genomic prediction of breeding values using a subset of snps identified by three machine learning methods. *Frontiers in genetics*, Frontiers, v. 9, p. 237, 2018. Citado na página 29.

LOUPPE, G. et al. Understanding variable importances in forests of randomized trees. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2013. p. 431–439. Citado na página 27.

MARTÍNEZ, C. *Evaluación de la calidad culinaria y molinera del arroz*. [S.l.]: CIAT, 1989. Citado na página 19.

MATUKUMALLI, L. K. et al. Application of machine learning in snp discovery. *BMC bioinformatics*, Springer, v. 7, n. 1, p. 4, 2006. Citado 2 vezes nas páginas 19 e 29.

MITCHELL, T. M. et al. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, v. 45, n. 37, p. 870–877, 1997. Citado na página 21.

SHERIF, F. F.; ZAYED, N.; FAKHR, M. Discovering alzheimer genetic biomarkers using bayesian networks. *Advances in bioinformatics*, Hindawi, v. 2015, 2015. Citado na página 29.

VIEIRA, N. d. A. Qualidade de grãos e padrões de classificação de arroz. *Embrapa Arroz e Feijão-Artigo em periódico indexado (ALICE)*, Informe Agropecuário, Belo Horizonte, v. 25, n. 222, p. 94-100, 2004., 1998. Citado na página 19.