IVAN DE PAIVA BARBOSA

EFICIÊNCIA PREDITIVA DE TÉCNICAS DE INTELIGÊNCIA COMPUTACIONAL EM FUNÇÃO DA COMPLEXIDADE DE CARACTERES SOB CONTROLE DE GENES COM EFEITO EPISTÁTICO

Tese apresentado à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento para o título de *Doctor Scientiae.*

Orientador: Cosme Damião Cruz

Coorientador: Moysés Nascimento

VIÇOSA - MINAS GERAIS
2021

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
B238s
2021

Barbosa, Ivan de Paiva, 1990-
Eficiência preditiva de técnicas de inteligência computacional em função da complexidade de caracteres sob controle de genes com efeito epistático / Ivan de Paiva Barbosa. – Viçosa, MG, 2021.
95 f. : il. (algumas color.) ; 29 cm.

Orientador: Cosme Damião Cruz.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Plantas - Melhoramento genético. 2. Inteligência computacional. 3. Predição. I. Universidade Federal de Viçosa. Departamento de Agronomia. Programa de Pós-Graduação em Genética e Melhoramento. II. Título.

CDD 22. ed. 631.52

Bibliotecário(a) responsável: Renata de Fatima Alves CRB6/2578

IVAN DE PAIVA BARBOSA

EFICIÊNCIA PREDITIVA DE TÉCNICAS DE INTELIGÊNCIA COMPUTACIONAL EM FUNÇÃO DA COMPLEXIDADE DE CARACTERES SOB CONTROLE DE GENES COM EFEITO EPISTÁTICO

Tese apresentado à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento para o título de *Doctor Scientiae*.

APROVADA: 29 de julho de 2021

Assentimento:

Ivan de Paiva Barbosa
Autor

Cosme Damião Cruz
Orientador

Aos meus pais Antônio Siqueira Barbosa e Verônica de Paiva Barbosa, pelo amor, carinho e exemplo de vida na qual eu me orgulho muito.

Aos meus irmãos Igor Luís Barbosa e Israel de Paiva Barbosa pela amizade e companheirismos em todos os momentos da minha vida.

A minha sobrinha Ana Beatriz, que tanto amo, pelo carinho e alegria.

A minha cunhada Ana Lucia, pela amizade.

A minha avó Conceição pelo amor, carinho, preocupação e conselhos.

A minha companheira Adriana pelo apoio, amor e amizade.

DEDICO E OFEREÇO

AGRADECIMENTOS

A Deus, por me acompanhar sempre, dando força e segurança para superar os desafios da vida. Agradeço também pela família que tenho.

A meus pais Antônio Siqueira Barbosa e Verônica de Paiva Barbosa, que tanto amo, pelo carinho, amizade, amor, educação e por sempre me apoiarem e estarem dispostos a me ajudar em todas as etapas da minha vida.

A meus irmãos Igor Luís Barbosa e Israel de Paiva Barbosa pela amizade, apoio e suporte a quem posso confiar e buscar apoio.

Minha linda e amada sobrinha Ana Beatriz, pelo amor e carinho.

À minha companheira Adriana pelo apoio, amor e amizade.

À Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Melhoramento de Plantas pela oportunidade de cursar a graduação e mestrado.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela concessão da bolsa de estudos.

Ao meu orientador professor Cosme Damião Cruz que tanto admiro, pelos ensinamentos, apoio, suporte, amizade, momentos de alegria e descontração e pela oportunidade de trabalhar com uma equipe de laboratório com valor inestimado em relação a amizade, confiança, apoio e suporte.

Ao meu coorientador professor Moysés Nascimento, pelo suporte, apoio, grandes ensinamentos, descontração e alegria que proporciona.

Às professoras e pesquisadoras Camila Ferreira Azevedo, Isabela de Castro Sant'Anna e Michele Jorge da Silva, pela amizade e disponibilidade em participar da banca contribuindo para melhoria deste trabalho.

Aos grandes amigos, membros do laboratório de Bioinformática e do Laboratório de Inteligência Computacional e Aprendizado Estatístico (LICAE), pela união, companheirismo, ensinamentos e momentos de descontração e alegria.

A pesquisadora Maria Aparecida Sediyama, pelos ensinamentos, oportunidades, apoio, confiança e principalmente a amizade, contribuindo de forma significativa para que pudesse chegar até aqui.

Aos amigos inseparáveis Edipo, Edison (Titito) e Willian pelo companheirismo, fidelidade e a quem posso depositar confiança e buscar ajuda e suporte em qualquer momento de minha vida.
A meus familiares, por estarem sempre dispostos e de braços abertos a me ajudar.
A todos os amigos que de alguma forma fazem parte da minha história.
MUITO OBRIGADO!

$$B > \frac{1}{n}\sum_{i=1}^{n} X_i$$

Be better than average

RESUMO

BARBOSA, Ivan de Paiva, D.Sc., Universidade Federal de Viçosa, julho de 2021. **Eficiência preditiva de técnicas de inteligência computacional em função da complexidade de caracteres sob controle de genes com efeito epistático**. Orientador: Cosme Damião Cruz. Coorientador: Moysés Nascimento.

Diante dos desafios enfrentados pelos melhoristas de plantas para seleção de indivíduos superiores, a seleção genômica ampla (*Genome Wide Selection* - GWS) é uma técnica que têm ganhado destaque. A GWS consiste na utilização de um grande número de marcadores moleculares para a predição de valores genéticos e têm se mostrado altamente relevante para o melhoramento genético. Na elaboração do projeto desta pesquisa, duas questões básicas foram formuladas. A primeira é que métodos diferentes proporcionariam resultados diferentes e segundo é que a complexidade da característica analisada poderia influenciar no desempenho das técnicas. Para isso, as análises foram divididas em dois artigos. No primeiro artigo, o objetivo foi avaliar e comparar o desempenho preditivo de métodos estatísticos (RR-Blup e BayesB) e métodos de inteligência computacional, MLP (*Multilayer Perceptron*) e RBF (*Radial Basis Function*), e aprendizado de máquina, árvore de regressão com refinamentos BO (*Boosting*), BA (*Bagging*) e RF (*Random Forest*) por meio de GWS em populações simuladas, com características apresentando diferentes níveis de herdabilidade e números de QTL (*Quantitative Trait Loci*), na presença de dominância e efeitos epistáticos. Já segundo artigo, os níveis de alguns cenários foram reduzidos e novos cenários de complexidade foram assumidos, e o objetivo foi comparar o desempenho seletivo e preditivo do RR-BLUP (*Random Regression Best Linear Unbiased Predictor*) e métodos de inteligência computacional, MLP (*Multilayer Perceptron*) e RBF (*Radial Basis Function*), e aprendizado de máquina, árvore de regressão com refinamentos BO (*Boosting*), BA (*Bagging*) e RF (*Random Forest*), em populações simuladas para diferentes cenários em relação ao número de QTL, à distribuição dos QTL nos grupos de ligação (GL), ao grau médio de dominância, à herdabilidade e aos diferentes modelos de expressão aditivo e não aditivos. Para ambos os artigos, o genoma simulado compreendeu 2010 SNP (*Single Nucleotide Polymorphism*) e distribuídos equitativamente em 10 grupos de ligação. Os métodos RBF, RF e *Bagging*, de uma maneira, geral apresentaram bons e consistentes resultados de R^2 e REQM. Esses métodos apresentaram resultados iguais ou

superiores à média geral dos métodos para todos os cenários avaliados. O aumento no número de QTL, até 88 para a configuração do genoma simulado, afetou positivamente os resultados de R^2 e REQM. A presença de efeito de dominância e a baixa herdabilidade provocou impactos negativos para os resultados de R^2 e REQM. Todos os métodos apresentaram uma redução no R^2 e aumento na REQM quando efeitos não aditivos eram importantes para a característica. Assim, nenhum método foi excepcionalmente eficiente para a captura deste efeito. A distribuição dos QTL nos GL foi o único cenário que afetou os resultados de R^2 e REQM de modo variado. Neste cenário os melhores resultados de R^2, foram observados quando os QTL estavam distribuídos em apenas um GL, enquanto que para os resultados de REQM os melhores resultados foram obtidos quando os QTL foram distribuídos em oito GL. Esse resultado mostra que a forma de distribuição dos QTL no genoma pode ser o principal atributo a ser avaliado quanto ao interesse na seleção ou na predição dos valores genéticos pelo melhorista. Os métodos de de inteligência computacional e aprendizado de máquina demonstram ser ferramentas poderosas para prever valores genéticos com controle de gene epistático, em características com diferentes graus de herdabilidade e diferentes números de genes de controle.

Palavras-chave: Inteligência Computacional. Predição. GWS. Seleção Genômica.

ABSTRACT

BARBOSA, Ivan de Paiva, D.Sc., Universidade Federal de Viçosa, July 2021. **Predictive efficiency of computational intelligence techniques as a function of the complexity of characters under control of genes with epistatic effect**. Advisor: Cosme Damião Cruz. Co-advisor: Moysés Nascimento.

Given the challenges faced by plant breeders to select superior individuals, genomic wide selection (GWS) is a technique that has gained prominence. GWS consists of the use of a large number of molecular markers for the prediction of genetic values and is highly relevant for genetic improvement. In the elaboration of this research project, two basic questions were formulated. The first is that different methods would provide different results and the second is that the complexity of the analyzed characteristic could influence the performance of the techniques. For this, the analyzes were divided into two articles. In the first article, the objective was to evaluate and compare the predictive performance of statistical methods (RR-Blup and BayesB) and computational intelligence methods, MLP (Multilayer Perceptron) and RBF (Radial Basis Function), and machine learning, Regression Tree with refinements BO (Boosting), BA (Bagging), and RF (Random Forest) through GWS in simulated populations, with characteristics presenting different levels of heritability and QTL (Quantitative Trait Loci) numbers, in the presence of dominance and effects epistatic. According to the article, the levels of some scenarios were reduced and new complexity scenarios were assumed, and the objective was to compare the selective and predictive performance of RR-BLUP (Random Regression Best Linear Unbiased Predictor) and computational intelligence methods, MLP (Multilayer Perceptron) and RBF (Radial Basis Function), and machine learning, Regression Tree with BO (Boosting), BA (Bagging) and RF (Random Forest) refinements, in simulated populations for different scenarios in relation to the number of QTL, the distribution of the QTL in the linkage groups (GL), the average degree of dominance, the heritability and the different additive and non-additive expression models. For both articles, the simulated genome comprised 2010 SNP (Single Nucleotide Polymorphism) and was evenly distributed into 10 linkage groups. The RBF, RF, and Bagging methods, in general, presented good and consistent results for R^2 and RMSE. These methods showed results equal to or higher than the general mean of the methods for all evaluated scenarios. The increase in the number of QTL, up to 88 for the simulated

genome configuration, positively affected the results of R^2 and RMSE. The presence of dominance effect and low heritability caused negative impacts on the results of R^2 and REQM. All methods showed a reduction in R^2 and an increase in RMSE when non-additive effects were important for the trait. Thus, no method was exceptionally efficient for capturing this effect. The distribution of QTL in GL was the only scenario that affected the R^2 and RMSE results in different ways. In this scenario, the best R^2 results were observed when the QTL were distributed in only one GL, while for the RMSE results the best results were obtained when the QTL were distributed in eight GL. This result shows that the form of distribution of QTL in the genome can be the main attribute to be evaluated regarding the interest in the selection or prediction of genetic values by the breeder. Computer intelligence and machine learning methods prove to be powerful tools for predicting genetic values with epistatic gene control, in traits with different degrees of heritability and different numbers of control genes.

Keywords: Computational Intelligence. Prediction. GWS. Genomic Selection.

SUMÁRIO

1. INTRODUÇÃO GERAL

O melhoramento de plantas pode ser conceituado como a arte, ciência e negócio estabelecidos pela alteração genética para benefício do homem (BERNARDO, 2002). No melhoramento busca-se a incorporação de novas formas alélicas ou pelo aumento da frequência de alelos favoráveis nas populações cultivadas de modo a atender às necessidades direta, ou indireta, do produtor, do consumidor e da indústria.

O cultivo de plantas é realizado com diversos objetivos, tais como a obtenção de grãos, frutos, tubérculos, forragens, óleos, fibras, ornamentação, dentre outros, relacionados com uma maior quantidade e qualidade. Para a agricultura, considerando as diversas etapas no processo desde o plantio à colheita, o objetivo final do produtor é o máximo rendimento, se possível com emprego mínimo de recursos e energia, ou seja, máxima eficiência produtiva. Também se espera uma maior qualidade final do produto, que pode ser a qualidade nutricional, resistência de fibras, teor de óleo, qualidade sensorial ou diversas outras características demandadas pelo mercado. Dessa forma, o papel do melhorista é o de introduzir, gerar ou aprimorar cultivares superiores para a agricultura de forma a atender às necessidades dos produtores (BUENO *et al*., 2006).

De acordo com a *World Resources Institute*, a agricultura deve ser capaz de fornecer alimento suficiente para atender ao volume crescente da população que poderá atingir quase 10 bilhões de pessoas até 2050 (SEARCHINGER *et al*., 2019). Além do aumento populacional, à medida que a população global cresce e a renda aumenta nas regiões em desenvolvimento, tem-se maior demanda *per capta*. Assim, a demanda geral por alimentos de origem vegetal está a caminho de aumentar em mais de 50% até meados deste século (SEARCHINGER *et al*., 2019). Nesse sentido, o melhoramento de plantas é uma das principais ferramentas para solucionar esses problemas, pela criação de plantas com maior potencial produtivo e com maior valor nutricional e, ainda, com adaptação às condições ambientais variáveis, possibilitando a produção de cultivares em regiões anteriormente de baixo potencial produtivo.

O melhoramento de plantas já ocorre há cerca de dez mil anos, ou seja, desde quando o homem deixou de ser nômade e passou a ser agricultor, dando início ao processo de domesticação das plantas (BORÉM, 2017). Certamente, o melhoramento

das plantas se deu de forma inconsciente pela seleção visual das plantas superiores para cultivo. As pessoas mais cuidadosas selecionavam as plantas mais desejáveis de forma empírica, o que denominamos como “arte”, ou seja, o melhoramento realizado sem nenhuma metodologia ou base científica (BUENO *et al.,* 2006). Mas foi a partir da constatação da existência de diferentes sistemas reprodutivos e a função do pólen na fertilização, que aumentaram os interesses pelos cruzamentos artificiais entre variedades e até entre espécies, possibilitando o homem explorar e estabelecer combinações híbridas de indivíduos com características de interesse (BUENO *et al.*, 2006). Assim, com o passar do tempo, o melhoramento deixou de ser apenas “arte” para se tornar também ciência, agora sim com desenvolvimento de métodos e base cientifica (ACQUAAH, 2007; BUENO *et al*., 2006).

O melhoramento pode ser praticado pelo uso de metodologias e ferramentas já testadas e comprovadas que são denominadas, por alguns autores, como melhoramento “clássico” ou “tradicional” (ACQUAAH, 2007; FERRÃO *et al.,* 2016; XAVIER *et al.,* 2018). Essa abordagem utiliza o cruzamento de plantas, ou hibridação, como a principal técnica para criação de variabilidade. Após os cruzamentos, a partir dos vários métodos de melhoramento e sem o auxílio de informações genômicas de forma diretas, as melhores plantas são identificadas e novas recombinações são realizadas dando início a um novo ciclo (ACQUAAH, 2007; BORÉM, 2017). Ao final de alguns, ou até vários, ciclos de seleção e recombinação os melhores genótipos são multiplicados e avaliados quanto ao desempenho antes da liberação para os produtores. No entanto, com o aumento dos ciclos de seleção as diferenças entre os indivíduos superiores e inferiores se tornam cada vez menores, e agora não mais com a necessidade de separar os indivíduos “bons” dos indivíduos “ruins”, mas é preciso discriminar os indivíduos “bons” dos “ótimos”.

Quando se estuda com características quantitativas, ou seja, aquelas controladas por muitos genes, a reprodutibilidade da característica ao nível de seleção é dificultada, devido à baixa herdabilidade, e como consequência disso obtém-se reduzidos ganho de seleção a cada ciclo de melhoramento. Esses fatores provocam uma demanda maior de tempo para obtenção de materiais genéticos superiores.

Visto o grande avanço já obtido para as principais plantas cultivadas que atualmente possuem elevado potencial produtivo, e que as características de interesse são controladas por muito genes, o “melhoramento clássico” tem uma

demanda cada vez maior por auxílio de ferramentas modernas para aumentar e acelerar os ganhos de seleção. Tais fatos exigem dos melhoristas novas habilidades e conhecimentos técnicos especiais.

A seleção visual é uma arte, mas pode ser facilitada por auxiliares de seleção, como marcadores genéticos, simplesmente herdados e facilmente identificados, que estão vinculados a características desejáveis, geralmente difíceis de identificar (ACQUAAH, 2007). Nesse contexto, as tecnologias moleculares surgem como um conjunto de poderosas ferramentas para análise e manipulação genética. Com os marcadores moleculares disponíveis é possível auxiliar o processo de seleção e tornar as atividades mais eficientes e eficazes (ACQUAAH, 2007).

Uma técnica que vem ganhando destaque e com bons resultados para seleção de indivíduos, com melhor desempenho a partir de informação de marcadores moleculares codominantes, é a seleção genômica ampla (*Genome Wide Selection* - GWS). Essa metodologia foi proposta por Meuwissen *et al.* (2001) e permite incorporar informações moleculares diretamente na predição do mérito genético dos indivíduos. Essa nova metodologia de seleção conseguiu se desenvolver graças aos amplos investimentos em genotipagem em grande escala e com o desenvolvimento dos marcadores do tipo SNP (*Single Nucleotide Polymorphism*) (BERNARDO e YU, 2007).

A seleção genômica ampla (*Genome Wide Selection* - GWS) foi proposta para estimar o valor genético genômico (VGG) de indivíduos que ainda não foram fenotipados por meio de informações de marcadores distribuídos em todo o genoma (MEUWISSEN *et al.,* 2001). A maior motivação para a utilização dessa técnica consiste na possibilidade de utilizar a genotipagem em grande escala e incorporar informações genômicas no processo de predição, de modo a aumentar a eficiência seletiva, obter ganhos genéticos de forma mais ágil, além da possibilidade de reduzir os custos (BERNARDO, 2009).

No entanto, devido à existência de centenas de milhares de marcadores no genoma, alguns desafios estatísticos são enfrentados pelos métodos baseados em GWS, tais como a alta dimensionalidade, uma vez que o número de marcadores (m) é muito maior que o número de observações (n) (m > n) e também a alta correlação existente entre os marcadores provocada pelo desequilíbrio de ligação (CROSSA *et al.*, 2017; LONG *et al.*, 2011).

Visando contornar essas dificuldades, têm sido proposto o uso de metodologias capazes de lidar com problemas relacionados com a colinearidade e dimensionalidade dos dados, seja pela redução da dimensionalidade (por exemplo, quadrados mínimos parciais), ou pela propriedade de regularização/penalização (por exemplo, RR-BLUP), ou pela abordagem bayesiano (BayesA e BayesB), ou seleção de um subconjunto de marcadores por meio de procedimento específicos como sondagens ou regressão *Stepwise* ou baseadas em inteligência computacional, por meio dos métodos RBF (Rede de Base Radial) e MLP (*Perceptron* Multicamadas) de redes neurais (GONZÁLEZ-RECIO *et al.*, 2014; SANT'ANNA *et al.*, 2020; SILVA *et al.*, 2014; SOUSA *et al.*, 2021) com resultados muito animadores.

De acordo com Silva *et al*. (2014), as RNA mostraram superioridade em relação ao método tradicional para discriminar genótipos com base em seu valor genético real e que essa abordagem tem grande potencial para uso como um método alternativo para predizer valores genéticos e como ferramenta de seleção genotípica, uma vez que apresentaram acurácia preditiva, expressa por erro quadrático médio, superior à apresentada pelo RR-BLUP.

As metodologias de inteligência computacional baseadas em RNA como é o caso das redes *Perceptron* Multicamadas, Redes de funções de base radial, árvore de regressão, *Boosting*, *Random Forest* e *Bagging* apresentam características vantajosas para estudos de características complexas, com controle gênico envolvendo efeitos aditivos, dominantes e epistáticos, que são de interesse dos programas de melhoramento genético. No entanto, ainda não se sabe como essas metodologias se comportam em função da complexidade das características. Abordagens de redução de dimensionalidade já questionam se o excesso de marcadores moleculares em modelos estatísticos ou em abordagem de inteligência computacional possam prejudicar a eficácia de predição. Pode-se acrescentar a este questionamento, se estes possíveis efeitos seriam mais pronunciados em caracteres menos ou mais complexos.

Diante do exposto foi conduzido este trabalho com o objetivo de avaliar o potencial das técnicas *Perceptron* Multicamadas (MLP), Redes de funções de base radial (RBF), árvore de regressão, *Boosting*, *Random Forest* (RF) e *Bagging*, baseadas em inteligência computacional, para seleção e predição de valores genéticos a partir da análise de dados genômicos em relação à complexidade do

carácter. Para este fim, foram consideradas as análises de dados genômicos simulados para características com diferentes números de QTL e graus de herdabilidades na presença de efeito de dominância e epistasia.

2. REVISÃO DE LITERATURA

2.1. Predição

O propósito em um programa de melhoramento genético é selecionar indivíduos superiores em uma população para realizar recombinações desses e obter uma população melhorada, ou seja, uma população com maiores valores genéticos para as características de interesse econômico. Para isso, são necessárias a realização de experimentos e a utilização de procedimentos estatísticos de forma a realizar a predição dos valores genéticos que possibilitem o melhor ordenamento das plantas para fins de seleção, visando atingir o maior ganho de seleção na geração seguinte.

A variação biológica de uma característica é descrita pela variância fenotípica (Vf), que pode ser decomposta em dois componentes: a variância devida ao componente genético (Vg) e a devida ao componente ambiental (Ve). Dessa forma, a variância fenotípica pode ser descrita como: Vf = Vg + Ve. Embora o interesse dos melhoristas seja de modo geral os efeitos genéticos, os valores observados se referem ao efeito fenotípico, por esse motivo cada fonte de variação deve ser identificada e seus efeitos corretamente avaliados (BUENO *et al.*, 2006).

As características quantitativas apresentam variação fenotípica contínua que é resultado da expressão conjunta de vários genes que condicionam a manifestação de um genótipo, através do fenótipo (BUENO *et al.*, 2006). Apesar das dificuldades em se trabalhar com características controladas por muitos genes, estas seguem as mesmas leis básicas da genética. Dessa forma exibem os mesmos tipos de efeitos gênicos conhecidos: efeito aditivo, efeito de dominância e efeito epistático. Já se sabe que a maior, ou menor, influência de um ou outro efeito na herança de um caráter pode ter forte impacto sobre o a seleção realizada para esse caráter (BUENO *et al.*, 2006). A predominância do efeito aditivo é a situação que proporciona melhor resposta à seleção, pois garante que as progênies serão semelhantes aos progenitores

selecionados. Quando há predominância de efeito da dominância, infere-se que superioridade genotípica pode ser atribuída à heterose, havendo a possibilidade de que os progenitores selecionados segreguem na próxima geração, permitindo o aparecimento de genótipos inferiores, algumas vezes diminuindo a média da população segregante. Os efeitos epistáticos também podem ser perturbadores no processo de seleção, uma vez que podem fazer parte da fração não herdável da variância genotípica de uma população e assim, precisam ser adequadamente isolados no modelo de forma a permitir uma acurada estimação dos valores genéticos a partir de componentes aditivos e de interação aditiva por aditiva.

Portanto, a obtenção de estimadores dos parâmetros genéticos é fundamental, pois permite identificar a natureza da ação dos genes envolvidos no controle dos caracteres quantitativos. As variâncias aditivas e não aditivas, as correlações e as herdabilidades são os paramentos genéticos de maior importância para o programa de melhoramento de plantas (CRUZ *et al.*, 2014).

A herdabilidade (h^2) é o parâmetro genético que expressa a proporção da variação genética na variação fenotípica, ou seja, $h^2 = V_g / V_f$. Porém, quando se decompõem a variância genética, o componente aditivo (Va) pode ser empregado para obtenção de uma estimativa mais apropriada, para fins de predição de ganhos genéticos, da herdabilidade, denominada como herdabilidade no sentido restrito e pode ser expressa como $h^2_r = V_a / V_f$.

Com as diferentes estratégias de condução de experimentos e das análises estatísticas o melhorista será capaz de predizer o valor genético de um indivíduo por meio do modelo:

$$Y_g = \beta\left(Y_f - \bar{Y}\right) + \varepsilon$$

sendo

Y_g: valor genotípico

Y_f: valor fenotípica

$\bar{Y}$: média geral

β: coeficiente de regressão (h^2)

ε: resíduo desconhecido

Com a prática da seleção é possível proporcionar incremento no valor genético da população pela expressão:

$$\Delta\hat{Y}_g = \hat{\beta}\Delta_f$$

sendo

$\Delta\hat{Y}_g$: ganho genético (GS)

Δ_f: diferencial de seleção (DS)

$\hat{\beta}$: coeficiente de regressão, dado por $\hat{\beta} = Cov_g(U_T, U_M)/\hat{\sigma}_f^2$, com $U_{T,M}$ sendo a probabilidade de se ter alelos idênticos por ascendência simultânea considerando.

Pela genética quantitativa, sabe-se que:

$$Cov_g(U_T, U_M) = 2r_{T,M}\sigma_A^2 + U_{T,M}\sigma_D^2$$

em que

$r_{T,M}$: coeficiente de parentesco entre as unidades de teste e unidade melhorada.

$U_{T,M}$: probabilidade de se ter alelos idênticos por ascendência simultânea considerando dois genitores.

σ_A^2 *e* σ_D^2: variância aditiva e devida aos desvios da dominância, respectivamente.

Por meio das informações anteriores é possível obter as herdabilidades e a proporção da variância aditiva e assim, estimar o ganho genético, que pode ser expresso conforme a Equação 1.

$$GS = \frac{i\, h_r \sigma_A}{I_g} \qquad \text{Equação (1)}$$

em que:

$GS = \text{Ganho de Seleção}$

$i = \text{Intensidade de seleção} = \text{DS}/\hat{\sigma}_f$

$h_r = \text{Acurácia de seleção, no sentido restrito}$

$\sigma_A = \text{Desvio padrão aditivo}$

$I_g = \text{Intervalo de geração}$

No caso de avaliação de famílias, o desvio padrão aditiva (σ_A) pode ser estimado multiplicando o desvio-padrão genético (σ_g) pelo controle parental (*p*), sendo que, o valor de p será igual a 1/2 quando a unidade de seleção for igual à unidade de

recombinação e se utilizar todos os genótipos (selecionados e não selecionados) na obtenção da população melhorada; *p* será igual 1 se a unidade de seleção for também igual à unidade de recombinação, mas a população melhorada for originada pelos intercruzamentos entre indivíduos selecionados; e ainda *p* será igual a 2 se a unidade de seleção for diferente da unidade de recombinação (CRUZ *et al.*, 2012).

As diferentes estratégias em um programa de melhoramento vão afetar os parâmetros que compõe a Equação 1. A acurácia de seleção (h_r) é uma indicação de quão boa pode ser a estimativa do verdadeiro mérito genético. Assim, quanto mais acurada for as estimativas avaliadas pelo melhorista, maior será o ganho de seleção. A intensidade de seleção (i) irá imprimir o quanto dos genótipos superiores será levado para as futuras gerações, assim, quanto maior a intensidade de seleção, maior será o ganho de seleção devido ao avanço apenas de genótipos com elevado potencial genético. No entanto, com maior intensidade de seleção, menor será a diversidade genética mantida na população melhorada, esse fato pode reduzir os ganhos genéticos futuros dessa população e encurtar a vida útil do programa de melhoramento, visto a baixa diversidade genética na população, o que dificultará a seleção de genótipos. O desvio padrão genotípico (σ) é resultante da variação aditiva, dominante e epistática presente na população, assim, quanto maior os efeitos aditivos, maior será o desvio padrão genético da população. Já o Intervalo de geração (I_g), é o tempo necessário para a obtenção da população melhorada, que pode ser encurtado por avaliações precoces.

Dessa forma, as diferentes metodologias e abordagens estatísticas para a predição irão afetar principalmente a acurácia, o desvio e o intervalo de geração. Já a intensidade de seleção será uma estratégia do melhorista.

2.2. Predição por marcadores Fenotípicos/ Seleção indireta

Em um programa de melhoramento genético, a escolha do método de seleção que possibilite o desenvolvimento de genótipos superiores é uma das etapas mais importantes na busca de novos cultivares (ENTRINGER *et al.*, 2014). No entanto, com o desenvolvimento de cultivares superiores, é cada vez mais difícil a identificação de novos genótipos superiores (CRUZ *et al.*, 2013).

A seleção indireta é uma metodologia com grande êxito nos programas de

genética e melhoramento, que se consiste na seleção de caracteres que estão associados com outro de interesse a ser melhorado. Esta metodologia apresenta vantagem quando a herdabilidade do caráter auxiliar e a correlação deste com a característica de interesse são elevados (BERED *et al.*, 1997).

A partir da associação entre diferentes tamanhos com o padrão e pigmentação de sementes de Phaseolus vulgaris, Sax (1923) demonstrou a possibilidade do uso de marcadores fenotípicos como estratégia de seleção indireta em estudos de genética e melhoramento. Dessa forma, a identificação de caracteres com mecanismo de herança simples (marcadores) ligados a genes controladores de características oligogênicas e, ou, poligênicas puderam auxiliar os programas de melhoramento da grande maioria das espécies cultivadas, aumentando a sua eficiência e agilidade (CRUZ *et al.*, 2013).

Sendo assim, em um programa de melhoramento genético, a seleção em uma característica pode provocar alterações em outras quando existe correlação entre elas. Neste sentido, os estudos de correlações fornecem importantes informações, como a possibilidade de identificar associações entre características na qual pode tornar viável a seleção indireta de um caráter quantitativo de difícil ganho de seleção, por meio de outras características a ele correlacionadas e de maior herdabilidade ou de mais fácil mensuração (CRUZ *et al.*, 2012).

É importante ressaltar que, quando uma variável quantitativa é correlacionada simultaneamente a muitas outras, o coeficiente de correlação simples pode não ser uma solução adequada, isso devido a existência de efeitos indiretos, na qual a correlação entre uma variável qualquer está associada à variável quantitativa devido ao efeito de uma terceira variável (CRUZ *et al.*, 2012). Nesse caso, a seleção da variável secundária não necessariamente levaria a uma resposta direta sobre a variável principal e algumas análises complementares tais como, análise de trilha, correlações parciais e correlações canônicas, poderiam apresentar bons resultados (CRUZ *et al.*, 2012).

O conhecimento das relações entre caracteres, estimadas pelas correlações, possibilita desenvolver alternativas que aumentem o progresso genético com a seleção de vários caracteres, ou se a seleção em um deles apresenta complexidade, como é caso da produtividade, por ser uma característica de herança poligênica, e consequentemente é uma característica altamente influenciada pelo ambiente, e/ou

também pela dificuldade na mensuração e identificação. Portanto, a utilização das estimativas de correlação para aplicação da seleção indireta pode permitir maior eficiência e progresso na seleção, comparada com a seleção direta (RIOS *et al.*, 2012).

Para a predição dos valores genéticos de uma característica principal (P) a partir de variações genéticas em uma característica auxiliar (A), utiliza-se o modelo:

$$Y_{gP} = \beta Y_{gA}$$

Com a prática da seleção, são estabelecidas as seguintes variações:

$$\Delta Y_{gP} = \beta \Delta Y_{gA}$$

em que

$\Delta \boldsymbol{Y_{gP}}$: representa o ganho genético na característica principal

$\Delta \boldsymbol{Y_{gA}}$: representa o ganho genético na característica auxiliar

β: coeficiente de regressão genético dados por $\boldsymbol{\beta = Cov_g(Y_P, Y_A)/\sigma^2_{G(A)}}$

Assim, para predição do ganho indireto, podemos utilizar os estimadores conforme a Equação 2:

$$\boldsymbol{GS_{y(x)}} = \frac{\boldsymbol{i\, r_g h_{rx} \sigma_{Ay}}}{\boldsymbol{I_g}} \qquad \text{Equação (2)}$$

em que:

$\boldsymbol{GS_{y(x)}}$= Ganho de Seleção em y pela seleção no caráter auxiliar x

$\boldsymbol{i}$ = Intensidade de seleção

$\boldsymbol{r_g}$= Correlação genética entre os caracteres x (auxiliar) e y (principal)

$\boldsymbol{h_{rx}}$= Acurácia de seleção da característica auxiliar, no sentido restrito

$\boldsymbol{\sigma_{Ay}}$= Desvio padrão aditivo da característica principal y

$\boldsymbol{I_g}$= Intervalo de geração

Diversos trabalhos apresentaram boa eficiência de seleção de variáveis complexas a partir da seleção indireta de variáveis. No entanto, os marcadores

morfológicos podem causar efeitos indesejáveis (CRUZ *et al.*, 2013). O forte efeito dos genes determinantes de marcadores morfológicos pode afetar a análise genética de grande número de caracteres de importância agronômica; poucos caracteres podem ser estudados ao mesmo tempo devido aos efeitos das interações gênicas como a epistasia; e o ambiente pode modificar a expressão dos marcadores morfológicos, o que pode gerar equívocos durante e após a análise (PATERSON *et al.*, 1991; BERED *et al.,* 1997).

2.3. Predição auxiliada por marcadores genotípicos

Com o rápido aperfeiçoamento das técnicas moleculares, fundamentadas na amplificação de fragmentos de DNA, grandes avanços têm sido alcançados na área dos marcadores moleculares (CRUZ *et al.,* 2013).

A partir dos marcadores moleculares é possível mapear os grupos cromossômicos de ligação. Estes mapas de ligação podem contribuir de forma decisiva no mapeamento de características de interesse agronômico e auxiliar na seleção de marcadores para estudos de caracterização da variabilidade genética (BERED *et al.,* 1997). O mapeamento genético em uma população segregante é possível de ser feito após realizar estudos de segregação genética ou genotípica de pares de locos e testes estatísticos para a detecção de associações. Tanto os locos controladores de caracteres qualitativos quanto os quantitativos podem ser detectados, localizados e quantificados seus efeitos a partir de diferentes mecanismos. O mapeamento de locos controladores de características qualitativas pode ser realizado por meio dos testes de segregação conjunta de pares de locos, levando em consideração a frequência de recombinantes observada. Já o mapeamento de locos controladores de características quantitativas (QTL - *Quantitative Trait Loci*), está baseado em procedimentos estatísticos mais elaborados, geralmente, baseados em teoria de regressão por intervalo ou técnicas de máxima verossimilhança. A denominação QTL utilizada para nomear as regiões cromossômicas que contêm genes (ou loco) que controlam esses caracteres quantitativos (FALCONER e MACKAY, 1996).

Como descrito anteriormente, a eficiência da seleção indireta é dependente da correlação entre o caráter indireto selecionado com a característica de interesse e

também da herdabilidade do caráter indireto. Nesse sentido, a seleção por marcadores moleculares é uma forma de se realizar a seleção indireta no qual a característica selecionada irá se aproximar dos 100% de herdabilidade, uma vez que os marcadores moleculares não são influenciados pelo ambiente (BERED *et al.*, 1997).

Assim, o ganho de seleção através da seleção pode ser calculado de acordo com a Equação 3:

$$GS_y = \frac{ir_{g\hat{g}}\,\sigma_{Ay}}{I_g} \qquad \text{Equação (3)}$$

em que:

$GS_y = \text{Ganho de Seleção } em\ y$

$i = \text{Intensidade de seleção}$

$r_g = \text{Acurácia na estimativa dos valores genéticos genômicos (VGG)}$

$\sigma_{Ay} = \text{Desvio padrão aditivo da característica principal y}$

$I_g = \text{Intervalo de geração}$

Os marcadores moleculares apresentam amplo potencial de uso no melhoramento de plantas, sendo muito eficaz na identificação e discriminação de genótipos, quantificação da variabilidade genética e sua correlação com a expressão fenotípica, previsão de produtividade de híbridos a partir da avaliação das linhagens paternais, caracterização de germoplasma, construção de mapas genéticos, dentre outros (BUENO *et al.*, 2006).

2.4. Seleção assistida por marcadores (SAM)

Com o desenvolvimento dos marcadores moleculares e o avanço em técnicas de biologia molecular, as informações genotípicas (obtidas por meio dos marcadores moleculares), uma vez correlacionadas com características fenotípicas de interesse, passam a serem amplamente utilizadas na identificação e seleção de indivíduos com maiores valores genéticos (RESENDE Jr. *et al.*, 2013). Adicionalmente, a seleção com base em informações genotípicas pode ser realizada precocemente, o que no caso do

melhoramento de espécies vegetais perenes, tende a elevar os ganhos (RESENDE, 2008).

O primeiro método proposto para o uso de marcadores no melhoramento ficou conhecido como seleção assistida por marcadores (SAM) (RESENDE Jr. *et al*., 2013). A SAM consiste em integrar a genética molecular com a seleção fenotípica, através da procura de alelos desejáveis indiretamente por meio do uso de marcadores ligados. Quanto mais próximo o marcador encontra-se do gene, mais eficiente é o processo. Na SAM utiliza-se simultaneamente dados de marcadores moleculares e dados fenotípicos que estejam em ligação gênica próxima com alguns QTL. Assim, os dados de marcadores podem ser utilizados como covariáveis, de efeito fixo, na explicação dos valores fenotípicos dos genótipos candidatos à seleção, ou como efeitos aleatórios incorporados no modelo associado à fenotipagem (RESENDE, 2008).

Algumas das vantagens dos marcadores no processo de seleção são: (i) não sofrem influência ambiental, (ii) são herdados mendelianamente e (iii) apresentam herdabilidade igual a um, o que pode aumentar muito a eficiência do processo de seleção dos genótipos superiores.

Para o emprego da SAM é necessário o mapeamento de caracteres de interesse agronômico de forma a maximizar a correlação genética. Esse é um procedimento demorado e que requer a construção de mapas de ligação genético saturados de marcadores (BERED *et al*., 1997). No melhoramento, existe a necessidade da determinação da posição, do número e dos efeitos dos QTL marcados. Entretanto, QTL de pequeno e raros efeitos normalmente não são detectados, o que resulta na capitação de apenas parte da variância genética explicada pelos marcadores/QTL, o que geralmente leva à subestimação desses efeitos (GODDARD e HAYES, 2007). Sendo assim, as seleções de genótipos baseadas nessas informações apresentam baixa precisão, pois sabe-se que, além de efeitos de QTL, os caracteres quantitativos ainda podem apresentar efeitos de interações interalélicas (epistasia) e efeitos de interação de genótipos com ambientes (PODLICH *et al*., 2004). Todos esses fatores afetam a precisão do mapeamento de QTL e, consequentemente, a utilização dessas informações na SAM. Nesse sentido, torna-se restrita a seleção genotípica para populações de distintos grupos gênicos, para diferentes ambientes ou por mais de um ciclo de seleção (BURGUEÑO *et al.,* 2012).

O modelo de predição de valores genéticos por seleção assistida a partir de um QLT supostamente localizado dentro de um intervalo delimitado por marcadores i e j, para uma população F2, é dado por:

$$Y = u + aX_{ij}^{*} + dZ_{ij}^{*} + E$$

Sendo a e d os efeitos associados a valores genotípicos de homozigotos e heterozigotos, respectivamente.

X_{ij}^{*}: Variável condicionadora, associada a a, que representa a probabilidade condicional da presença do QTL para um determinado padrão dos marcadores i e j;

Z_{ij}^{*}: Variável condicionadora, associada a d, que representa a probabilidade condicional da presença do QTL para um determinado padrão dos marcadores i e j.

De acordo com Resende (2008), a SAM pode apresentar superioridade em relação à seleção fenotípica apenas quando o tamanho da população genotipada é muito grande, da ordem de 500 genótipos ou mais. Além disso, suas vantagens sobre a seleção fenotípica são proporcionais à porcentagem da variância genética explicada pelo marcador. Desse modo, a questão chave é determinar quantos marcadores/QTL são responsáveis pela variação no caráter quantitativo e quantos são necessários para explicar a maior parte dessa variância genética (MORAES JÚNIOR, 2013). Na prática, a maioria dos marcadores explica muito pouco da variância genética e, portanto, o ganho em utilizá-los geralmente é muito pequeno. Desse modo, a SAM tem sido mais utilizada em programas de melhoramento apenas para casos específicos, como introdução de alelos de herança monogênica em germoplasma, seleção de plantas dentro de progênies e avanço de gerações em casa de vegetação, para caracteres de média ou alta herdabilidade (HOSPITAL *et al.*, 1997; HOLLAND, 2004).

Novas estratégias da aplicação dos marcadores moleculares no melhoramento de plantas foram desenvolvidas, como por exemplo, o *Genome-Wide-Selection* (GWS). Essa estratégia considera informações de todos os marcadores na seleção e não apenas daqueles associados ao caráter.

2.5. Seleção Genômica Ampla (GWS)

Com os avanços no desenvolvimento de tecnologias de genotipagem em larga escala, novos sistemas de marcadores moleculares como *Single Nucleotide Polymorphism* (SNP) têm permitido maior eficiência na avaliação genética em nível molecular, de modo que atualmente a maioria das espécies de interesse econômico dispõe de número elevado de marcadores passíveis de uso em programas de melhoramento (MORAES JÚNIOR, 2013; RESENDE Jr. *et al*., 2013). A redução do preço por *data point,* permitiu que grande número de marcadores fossem usados para várias culturas (JEHAN e LAKHANPAUL, 2006). Uma vez gerado grande número de marcadores (na ordem de centenas e milhares de marcadores) espalhados por todo o genoma de um indivíduo, alguns destes marcadores estarão em desequilíbrio de ligação (LD) com QTL (HASTBACKA *et al*., 1994; MORAES JÚNIOR, 2013).

Os maiores ganhos genéticos obtidos com a abordagem da GWS, em relação aos métodos tradicionais de seleção, são devido à redução do intervalo entre ciclos e maior acurácia de seleção (MORAES JÚNIOR, 2013). Esses ganhos têm justificado a sua utilização em programas de melhoramento vegetal (CROSSA *et al*., 2011; FRITSCHE NETO, 2011; HAYES *et al*., 2013; RESENDE JÚNIOR, 2010).

A teoria da genética de associação e da seleção genômica baseia-se no fato de que, com grande número de marcadores espalhados pelo genoma, aumenta-se a probabilidade de que QTL de interesse estejam em forte desequilíbrio de ligação (DL) com os marcadores (RESENDE Jr. *et al*., 2013). Na GWS os efeitos de todos os marcadores sobre as características de interesse são estimados simultaneamente e são elaborados modelos para predição do valor genético genômico dos indivíduos em gerações futuras (RESENDE Jr. *et al*., 2013). Dessa forma, quase a totalidade da variação genética do caráter quantitativo será capturada, uma vez que se utilizam todos os marcadores no modelo preditivo. Assim todos os QTL, sejam eles de grandes ou pequenos efeitos, estarão em DL com marcadores moleculares. Com isso a acurácia da estimativa do valor genético-genômico dos indivíduos apresenta resultados muito superiores daqueles observados para a SAM (RESENDE Jr. *et al*., 2013). Além disso a GWS pode ser aplicada em toda a população, não se restringindo a uma família específica.

Uma regressão linear simples para estimar o efeito individual de cada marcador pode ser descrita pela equação:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Onde:
y é o valor fenotípico e
x representa o valor genotípico, mm, Mm ou MM (-1, 0, 1).

Essa expressão pode ser generalizada para cálculo do efeito referente a todos os marcadores em uma equação multivariada com os β efeitos sendo estimados simultaneamente, assim tem-se a expressão:

$$y = \beta_0 + x_1\beta_1 + \cdots + x_m\beta_m + \varepsilon$$

Onde β_m representa o efeito para cada marcador.

A partir dos efeitos estimados é possível calcular o Valor Genético Genômico estimado (EGBV – *Estimated Genomic Breeding Values*), que de forma matricial é calculado conforme a expressão abaixo:

$$EGBV = X\hat{\beta}$$

em que:
X é a matriz de incidência, e
$\hat{\beta}$ é o vetor do efeito de marcadores.

Contudo, a implementação da GWS impõe desafios estatísticos e computacionais tais como a dimensionalidade do modelo, colinearidade entre marcas e a complexidade das características quantitativas (FERREIRA *et al*., 2018). Para contornar esses desafios, vários métodos têm sido propostos, na qual se diferem pelo tipo de suposição sobre o modelo genético associado ao caráter quantitativo (GONZÁLEZ-RECIO *et al*., 2014). Entre eles, os métodos de redução da dimensionalidade (por exemplo, quadrados mínimos parciais), ou de regularização/penalização (por exemplo, RR-BLUP), ou abordagem bayesiano (BayesA e BayesB), ou sondagens ou regressão *Stepwise* ou baseadas em inteligência computacional (RBF, MLP, *Bagging*, *Boosting*, *Random Forest*).

2.5.1. Codificação de SNPs

A codificação mais simples e mais comumente aplicada de genótipos SNP assume efeito "aditivo" de alelos que contribuem para o genótipo. Sob este modelo, o efeito do heterozigoto, Mm, é exatamente intermediário entre os homozigotos dominante e recessivo, MM e mm, respectivamente, de modo que os genótipos são

codificados de acordo com o número de alelos dominantes, denotados G_{ijA} (MM=0; Mm =1 e mm=2). Um modelo linear generalizado pode então ser expresso como:

$$g(E[y_i]) = \mu + \beta_A G_{ijA}$$

onde β_A é o efeito aditivo do alelo no SNP sobre a característica.

A estimativa de probabilidade máxima, $\hat{\beta}_A$, representa o efeito do alelo dominante em relação ao alelo recessivo no SNP sobre a característica. A interpretação precisa da estimativa do efeito dependerá do resultado sob investigação e do desenho do estudo. Para medidas quantitativas, para as quais a associação é avaliada por meio de um modelo de regressão linear, $\hat{\beta}_A$ pode ser interpretado como a mudança média no valor do traço para cada cópia do alelo dominante transportado por um indivíduo. Para desfechos binários de doença, para os quais a associação é avaliada por meio de um modelo de regressão logística, $\hat{\beta}_A$ pode ser interpretado como a razão de chances log do alelo dominante em relação ao alelo recessivo.

Uma codificação mais geral de genótipos permite desvios dos efeitos aditivos de alelos no SNP, incluindo um componente de "dominância", denominado G_{ijD} (MM=0; Mm =1 e mm=0). Sob esta parametrização de genótipos SNP, o modelo linear generalizado pode ser expresso como:

$$g(E[y_i]) = \mu + \beta_A G_{ijA} + \beta_D G_{ijD}$$

Nesta expressão, β_A é o efeito aditivo dos alelos no SNP sobre a característica, enquanto β_D é o efeito de dominância.

Sob esta parametrização do modelo genotípico geral, obtém-se estimativas de máxima verossimilhança dos efeitos aditivos e de dominância, denotados por $\hat{\beta}_A$ e $\hat{\beta}_D$, respectivamente. Isso possibilita estimar o efeito do heterozigoto, Mm, em relação ao homozigoto dominante, MM, por $\hat{\beta}_A + \hat{\beta}_D$, e o efeito do homozigoto recessivo, mm, em relação aos homozigotos dominante por $2\hat{\beta}_A$. Para resultados quantitativos, esses efeitos são interpretados em termos de diferenças nos valores médios de características entre os genótipos, enquanto para resultados de doenças binárias, eles podem ser interpretados como razões de chances logarítmicas (MORRIS e CARDON, 2019).

É importante destacar que a "dominância" é um termo geral para descrever um desvio da aditividade e não deve ser confundido com um modo de herança

"dominante", no qual os efeitos sobre a característica sob investigação são os mesmos para os genótipos heterozigotos e raros homozigotos. Se for de interesse avaliar a associação com um SNP sob um modo de herança dominante, pode-se considerar uma parametrização restrita do modelo linear generalizado para o qual $\beta_A = \beta_D$, com a estatística de teste resultante tendo uma distribuição aproximadamente qui-quadrado com um grau de liberdade. Da mesma forma, para um modo de herança recessivo, em que os efeitos sobre a característica sob investigação são os mesmos para os genótipos heterozigotos e homozigotos comuns, podemos considerar uma parametrização restrita do modelo linear generalizado para o qual $\beta_A = -\beta_D$, com a estatística de teste resultante tendo uma distribuição de aproximadamente qui-quadrado com um grau de liberdade (MORRIS e CARDON, 2019).

O modelo aditivo demonstra ser mais poderoso do que o modelo de genótipo geral, desde que o efeito do genótipo heterozigoto seja intermediário entre os dois homozigotos (MORRIS e CARDON, 2019). Além disso, a dominância será enfraquecida em SNPs que não estão em LD perfeito com uma variante causal não testada, mesmo se o verdadeiro modelo de associação subjacente não for aditivo na variante causal (SPENCER *et al*., 2009). Consequentemente, uma abordagem amplamente utilizada é usar um modelo aditivo para identificar SNPs associados e, em seguida, testar a evidência de dominância nesses SNPs selecionados. Não é recomendado considerar vários modelos (por exemplo, aditivo, dominante e recessivo) em todo o genoma, uma vez que isso incorrerá em uma penalidade para testes múltiplos que reduzirão o poder de detectar associação (MORRIS e CARDON, 2019).

2.5.2. Epistasia

Na genética mendeliana clássica, a epistasia (do grego *epi*, sobre; e *stasis*, parada, inibição) pode ser definida como um mascaramento, ou modificação, dos efeitos de genótipos em uma determinada características, e assim, os efeitos de QTL podem variar de acordo com o ambiente genético (MACKAY, 2001). Num contexto biológico, as interações do tipo epistáticas ocorrem quando dois ou mais genes determinam a produção de enzimas que catalisam diferentes etapas de uma mesma via biosintética. Vias biosintética são aquelas em que as enzimas produzidas por

determinados genes atuam, de maneira que uma substância inicial (substância precursora) é desdobrada em substratos até dar origem a um produto final, que pela ação do meio resultará num determinado fenótipo para aquele caráter. A epistasia envolve a supressão gênica inter-alélica, ou seja, os alelos de um loco gênico encobrem, ou suprimem, a expressão de outro alelo pertencente a outro loco gênico (não-alelo). O alelo (ou gene) que mascara a expressão do outro é denominado de epistático a esse alelo (ou gene). O alelo (ou gene) cuja ação suprimida é denominada de hipostático (MORRIS e CARDON, 2019). Quando se verifica epistasia entre dois locos gênicos bi-alélicos, o número de fenótipos entre os descendentes será menor que quatro. A proporção 9:3:3:1 se modifica dando origem a uma nova proporção que é uma combinação dessa.

Para a genética quantitativa o uso do termo epistasia é muito mais amplo e se refere a qualquer interação estatística entre genótipos em dois (ou mais) locos (MACKAY, 2001; CHEVERUD, 1995). A epistasia pode se referir a uma modificação dos efeitos homozigotos ou heterozigotos dos locos em interação, podendo ser sinérgica, na qual o fenótipo de um loco é realçado por genótipos em outro loco; antagonista, no qual a diferença entre os genótipos em um loco é suprimida na presença de genótipos no segundo loco; ou mesmo produzir novos fenótipos (MACKAY, 2001).

Epistasia extensa têm consequências práticas e teóricas, e com isso as estimativas dos principais efeitos de QTL poderão ser tendenciosas. A existência de epistasia que não se manifesta claramente em associação com características complexas não é surpreendente, uma vez que se espera que os mecanismos biológicos subjacentes sejam extremamente complexos, incorporando os efeitos conjuntos de múltiplos fatores de risco genéticos (MORRIS e CARDON, 2019). Qualquer via de desenvolvimento ou bioquímica que culmina na expressão de uma característica quantitativa é composta de redes de locos que interagem nos níveis genético e molecular. A variação genética em alguns desses locos está causalmente ligada à variação fenotípica da característica (MORRIS e CARDON, 2019).

De uma perspectiva estatística, a epistasia pode ser representada por uma interação entre genótipos em dois ou mais SNPs. Assim, o modelo de regressão linear generalizado para análise de associação de SNP único naturalmente se estende para levar em conta as interações (CORDELL, 2002). Por exemplo, para modelar a

interação entre dois SNPs, permitindo desvios de aditividade dentro e entre genótipos, pode ser representado da seguinte forma:

$$g(E[y_i]) = \mu + \beta_{1A}G_{i1A} + \beta_{1D}G_{i1D} + \beta_{2A}G_{i2A} + \beta_{2D}G_{i2D} + \beta_{12AA}G_{i1A}G_{i2A} + \beta_{12AD}G_{i1A}G_{i2D} + \beta_{12DA}G_{i1D}G_{i2A} + \beta_{12DD}G_{i1D}G_{i2D}$$

em que: G_{i1A} e G_{i2A} denotam a codificação aditiva de genótipos nos dois SNPs, respectivamente, e G_{i1D} e G_{i2D} denotam a codificação de dominância, respectivamente. Os parâmetros β_{jA} e β_{jA} correspondem aos efeitos principais aditivo e de dominância, respectivamente, do j-esimo SNP. Os quatro termos de interação, β_{12AA}, β_{12AD}, β_{12DA} e β_{12DD}, correspondem aos componentes aditivo- aditivo, aditivo-dominância, dominância-aditivo e dominância-dominância para epistasia entre os dois SNPs.

Na presença de mais de dois SNPs e expectativa de interações múltiplas, o modelo de regressão linear também pode ser estendido para incorporar termos de interação de ordem superior, embora esses efeitos sejam frequentemente difíceis de interpretar e de estimar sem grandes tamanhos de amostras (MORRIS e CARDON, 2019).

Na presença de epistasia entre SNPs, esperava-se que os efeitos de interação explícito nos modelos forneceriam um maior poder em relação àqueles que consideram apenas os efeitos principais. No entanto, a epistasia é muitas vezes desconsiderada em GWS porque: (i) o modelo menos parcimonioso não terá força, a menos que os efeitos da interação sejam grandes; (ii) limites de significância rigorosos são necessários para contabilizar o número de testes em varreduras pareadas do genoma; e (iii) a complexidade do modelo e dos números de pares SNP possíveis aumenta a carga computacional da análise (MARCHINI *et al.*, 2005). Porém, permitir efeitos principais aditivos e epistasia aditivo-aditiva para todos os pares de SNPs, em todo o genoma, demonstra aumentar o poder de detectar associação em análises de SNP individuais para uma gama de modelos de interação, apesar da carga adicional de testes múltiplos (MARCHINI *et al.*, 2005).

A teoria para estimar a contribuição das interações epistáticas para a variância fenotípica total de uma característica em cruzamentos entre linhagens endogâmica está bem desenvolvida. No entanto, a epistasia é difícil de detectar nesses projetos. Em parte, porque mesmo as interações epistáticas fortes contribuem pouco para a variância epistática, e em parte porque este termo tem uma variância de amostragem

muito alta, exigindo grandes tamanhos de amostra (MACKAY, 2001). Além disso, o conhecimento da variação significativa da interação epistática não é terrivelmente esclarecedor em relação à natureza dos efeitos individuais (MACKAY, 2001).

Testes de epistasia aditiva por aditiva requerem a construção de quatro genótipos homozigotos duplos em dois locos bialélicos, enquanto a estimativa de todas as classes de interações epistáticas envolve a síntese de nove genótipos de dois locos.

O poder de detectar epistasia entre QTL em populações de mapeamento é baixo, por várias razões: (a) mesmo grandes populações de mapeamento contêm poucos indivíduos nas classes mais raras de genótipos de dois locos; (b) a segregação para outro QTL pode interferir na detecção de epistasia entre o par de locos em consideração; e (c) depois de ajustar o limite de significância para os múltiplos testes estatísticos envolvidos na busca de interações epistáticas, apenas interações extremamente fortes permanecem significativas.

Dados esses pensamentos contra a detecção de epistasia, muitos estudos relatam efeitos de QTL amplamente aditivos (ZENG *et al*., 2000) ou não testam a epistasia. Por outro lado, fortes interações foram observadas entre QTL afetando o número de cerdas de Drosophila (LONG *et al*., 1995), aptidão em cepas de dupla deleção de levedura (JASNOS *et al*., 2007), modular no metabolismo de leveduras (SEGRE *et al*., 2005), redes metabólicas de *Escherichia coli* e *Saccharomyces cerevisiae* (HE *et al*.,2010) e rendimento de grãos em arroz (LI *et al*., 1997).

Além do mais, estão ganhando destaques os métodos baseados em inteligência computacional, com propriedades que os tornam potencialmente atraentes para GWS (CRUZ e NASCIMENTO, 2018; SANT'ANNA *et al*., 2020; SILVA *et al*., 2014; SOUSA *et al*., 2021). Em modelos de inteligência computacional todos os marcadores têm uma chance de contribuir para o ajuste do modelo, incluindo aqueles com efeitos fracos e marcadores altamente correlacionados e interagentes. Além disso, esses métodos permitem que interações complexas entre marcadores sejam facilmente incluídas, como em modelos não aditivos com efeitos dominantes e epistáticos, e não fazem suposições distributivas sobre as variáveis preditoras.

2.6. Métodos estatístico utilizados em predições

Grandes são os desafios estatísticos devido ao nível de complexidade nos modelos de previsão de seleção genômica, uma vez que o número de marcadores (m) é maior que o tamanho da população (n) e além disso os preditores (marcadores) são altamente correlacionados. Essa condição torna impossível calcular estimativas de mínimos quadrados ordinários para efeitos de marcadores e predizer valores genéticos. Esta impossibilidade ocorre devido ao número insuficiente de graus de liberdade para ajustar todos esses efeitos de marcadores simultaneamente. Para solucionar esse problema algumas estratégias podem ser utilizadas, tais como o uso de regressão penalizada, seleção de variáveis e redução da dimensionalidade.

Sendo assim, na implementação da GWS existem desafios estatísticos computacionais, tais como a definição de métodos de predição genômica que possibilitem melhor tratamento dos dados, considerando a dimensionalidade, colinearidade entre marcadores e a complexidade dos caracteres quantitativos.

Além disso, a escolha dos métodos estatísticos para a predição dos efeitos de marcadores também pode afetar a acurácia do valor genético genômico. Assim, um desafio enfrentado está diretamente ligado às pressuposições acerca do modelo avaliado, tais como dimensionalidade das matrizes envolvidas, multicolinearidade entre os marcadores moleculares e a complexidade dos caracteres quantitativos em estudo, com a inclusão das interações intra e inter-alélicas (FERREIRA *et al*., 2018).

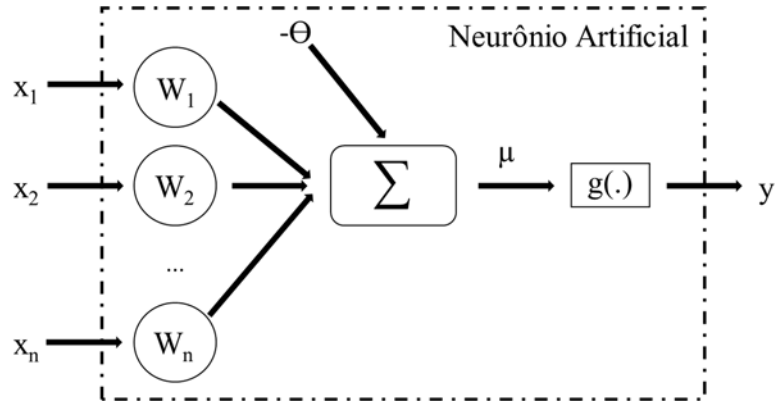
Os principais métodos estatísticos de seleção genômica pode ser divididos em três classes: regressão explícita, implícita e com redução dimensional (RESENDE *et al*., 2014). Os métodos da classe de regressão explícita podem ser divididos em dois grupos: (i) métodos de estimação penalizada, como RR-BLUP (MEUWISSEN *et al*., 2001; LASSO (TIBSHIRANI, 1996); (ii) métodos de estimação bayesiana, tais como BayesA, BayesB (MEUWISSEN *et al*., 2001). Os métodos de regressão com redução dimensional, por sua vez, compreendem os de componentes independentes, quadrados mínimos parciais e de componentes principais (SOLBERG *et al*., 2009). Na classe de regressão implícita, destacam-se o método semi-paramétrico RKHS (*Reproducing Kernel Hilbert Spaces*) (GIANOLA e DE LOS CAMPOS, 2008), e as redes neurais artificiais (GIANOLA *et al*., 2011) e a regressão Kernel não-paramétrica (GIANOLA *et al*., 2006).

2.7. Métodos de inteligência computacional utilizados em predições

A inteligência computacional é a área da ciência que estuda a teoria e a aplicação de algumas técnicas inspiradas na natureza: as redes neurais artificiais (RNA), a lógica nebulosa e a computação evolucionária (TOMAZ *et al*., 2018). A utilização da inteligência computacional nos programas de melhoramento genético é recente, sendo a grande vantagem a estrutura não linear e o processamento paralelo, que permite captar propriedades mais complexas dos dados (CRUZ e NASCIMENTO, 2018).

As RNA fundamentam-se em sistema que tem elementos que simulam o cérebro humano, inclusive em seu comportamento, ou seja, aprendendo, errando e realizando descobertas (TOMAZ *et al*., 2018). No modelo neural artificial, o desempenho está diretamente ligado às conexões entre os elementos que o compõe. Este modelo é composto por três elementos principais: um conjunto de sinapses, um somatório e uma função de ativação (HAYKIN, 2001) (Figura 1), e o processamento de uma RNA envolve três etapas primordiais: treinamento, aprendizado e validação, aliadas à escolha de uma arquitetura apropriada que possua funções de ativação eficientes, número de camadas ocultas e número de neurônios por camadas satisfatórios (SILVA *et al*., 2014).

Figura 1 **–** Modelo não linear de um neurônio artificial, em que x_1, x_2, ..., x_n são as entradas da rede; w_1, w_2, ..., w_n são os pesos, ou pesos sinápticos, associados a cada entrada; -Θ é o limiar de ativação (bias); μ é a combinação linear dos sinais de entrada; g (.) é a função de ativação; e y é a saída do neurônio



Fonte: O autor.

Dessa forma, quando se utiliza a abordagem de RNA para obter soluções de

problemas experimentais, existe a necessidade de mecanismos de aprendizado para proporcionar uma solução satisfatória para o problema apresentado, conhecido como processo de treinamento. O processo de aprendizado em uma RNA consiste em apresentar um conjunto grande de observações e de respostas desejáveis. Durante esse processo de treinamento, ocorre o ajuste de pesos entre as conexões dos neurônios artificiais. Estes são os parâmetros ajustáveis que variam à medida que o conjunto de treinamento é apresentado à rede (TOMAZ *et al*., 2018).

Quando redes neurais são aplicadas à GWS, a camada de entrada é a matriz com as informações de cada marcador, composta por um neurônio por marcador. Cada um dos neurônios (marcadores) na camada de entrada (sinal) é conectado a todos os neurônios na primeira camada, e a partir desse ponto cada sinal será multiplicado por um peso que indica a sua influência na saída da unidade, posteriormente é feito uma soma ponderada dos sinais, que produz um nível de atividade e se esse nível de atividade exceder certo limite, a unidade produz uma determinada resposta de saída (CRUZ e NASCIMENTO, 2018).

As RNAs caracterizam-se pela sua arquitetura e pelo ajustamento de seus pesos às conexões durante o processo de aprendizado (CRUZ e NASCIMENTO, 2018). A arquitetura de uma rede neural é definida pelo número de camadas (camada única ou múltiplas camadas), pelas conexões entre camadas, pelo número de neurônios em cada camada, pelo tipo de conexão entre eles (*feedforward* ou *feedback*) e pelo algoritmo de aprendizado (HAYKIN, 2001).

Nos modelos de rede propostos que apresentam apenas camada de entrada e saída, não é possível a formação de uma representação interna, e nesse caso a codificação proveniente da entrada não é suficiente para implementar mapeamento de saída, o que implica que padrões de entrada similares resultam em padrões de saída similares. Esse fato leva o sistema à incapacidade de aprender importantes mapeamentos, incluindo aqueles não linearmente separáveis (CRUZ e NASCIMENTO, 2018). Como resultado, padrões de entrada com estruturas similares, fornecidos do ambiente externo, que levem a saídas diferentes não são possíveis de serem mapeados por redes sem representações internas, isto é, sem camadas intermediárias (CRUZ e NASCIMENTO, 2018).

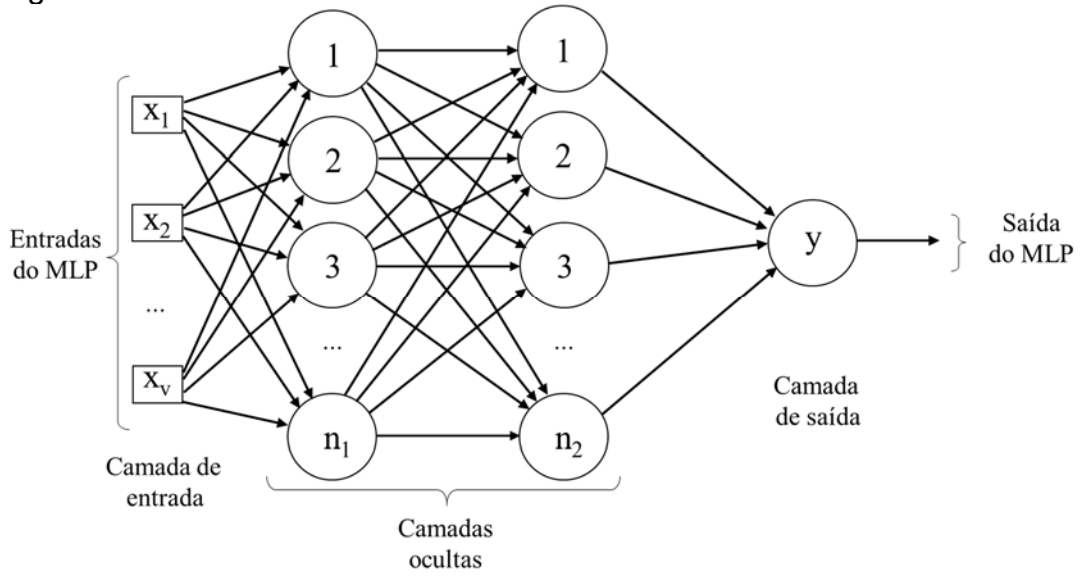
As Redes Neurais Artificiais do tipo Rede de Base Radial (RNA-RBF) e Redes *Perceptron* de Multiplas Camadas (RNA-MLP) são exemplos de redes com camadas

feedforward não-lineares (CRUZ e NASCIMENTO, 2018).

2.7.1. Redes *Perceptron* Multicamadas (MLP)

A Rede *Perceptron* Multicamadas ou *Multilayer Perceptron* (MLP) apresenta uma estrutura de redes neurais que se caracteriza pela existência de pelo menos uma camada oculta de neurônios, fundamentado no processo de aprendizagem denominado de *backpropagation* (CRUZ e NASCIMENTO, 2018). Além disso a rede MLP pode apresentar diversos neurônios na camada de saída, sendo que cada um destes representa uma das saídas do processo a ser mapeado. Assim, se um processo consistir de m saídas a rede MLP terá m neurônios em sua última camada neural (Figura 2).

Figura 2 **–** Ilustração de uma rede *Perceptron* Multicamadas com “v” variáveis na camada de entrada, “n1” neurônios na primeira camada oculta “n2” neurônios na segunda camada oculta e uma camadas de saída



Fonte: O autor.

A camada de entrada corresponde às informações disponíveis para ser apresentadas à rede a fim de seu treinamento. As camadas intermediárias funcionam como extratoras de características contidas no conjunto de dados apresentados. E a camada de saída recebe os estímulos das camadas intermediárias e constrói o padrão que será a resposta.

O número de camadas intermediárias (ocultas) e seu dimensionamento, ou seja, o número de neurônios (n) que as constituem, são objetos de investigação, com soluções diferentes para as diferentes áreas da pesquisa (CRUZ e NASCIMENTO, 2018). Geralmente, o número de neurônios para resolução de problemas na área das agrárias é definido de forma empírica, variando-se o número de neurônios até se encontrar uma solução ótima. Nesse caso, deve-se ter cuidado para não utilizar neurônios demais, o que pode levar a *overfitting,* ou seja, em vez de aprender, a rede memoriza os dados e decora o padrão específico de entrada e da saída (CRUZ e NASCIMENTO, 2018).

O algoritmo *backpropagation*, utilizado no treinamento das redes MLP consiste em duas fases: o passo onde a rede é alimentada para frente (*forward*) e o passo onde a rede é alimentada para trás (*backward*). Na etapa *forward*, os pesos sinápticos w(p) permanecem inalterados e os sinais funcionais da rede neural são calculados

para cada neurônio até que seja produzida a saída desejada na camada de saída (essa etapa também é conhecida como fase de propagação). A etapa *backward*, por sua vez, se inicia na camada de saída da rede, passando os sinais de erro para as camadas anteriores, de modo que os pesos sinápticos sejam recalculados de acordo com a regra Delta (Equação 4) até que se retorne à primeira camada oculta da rede (etapa também conhecida como fase de atualização de pesos ou retro-propagação).

$$\Delta w_{(t)} = \alpha \Delta w_{(t-1)} + \eta \delta_{(t)} y_{(t)} \qquad \text{Equação (4)}$$

onde

α é a constante de *momentum* com 0 < α < 1;

δ, é o gradiente local;

η, taxa de aprendizagem;

y, a saída da rede;

$\Delta w_{(t)}$ é o erro obtido pela rede neural na iteração t; e

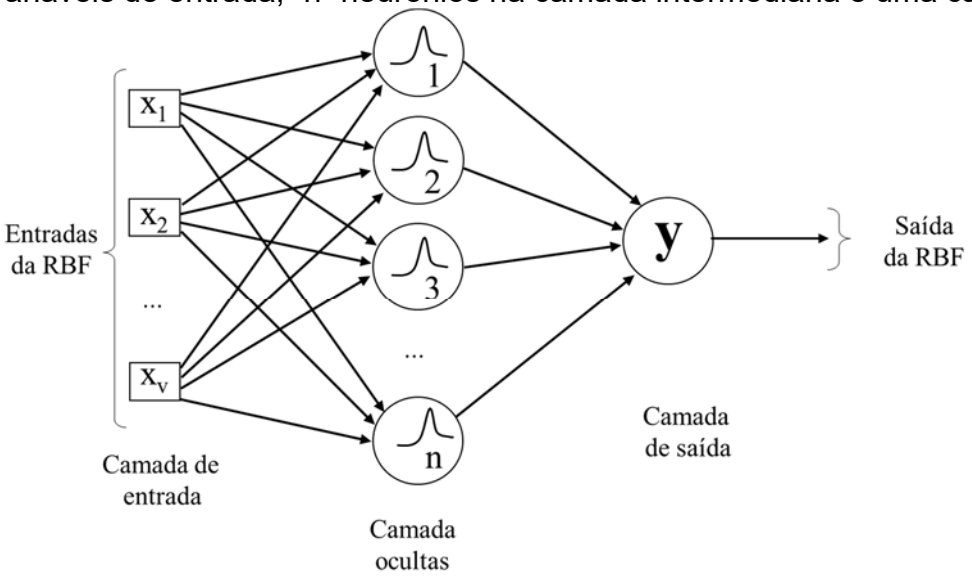
$\Delta w_{(t-1)}$ é o erro obtido pela rede neural na iteração anterior (t − 1).

2.7.2. Redes de funções de base radial (RBF)

Redes de RBF, diferentemente das redes de multicamadas, possui apenas uma camada oculta, sendo formada por neurônios com função de ativação de base radial local (HAYKIN, 2009). Com apenas uma camada intermediária na rede neural já é possível se calcular uma função arbitrária qualquer a partir de dados fornecidos. Para isso, a camada oculta deve ter por volta de (2i+1) neurônios, onde i é o número de variáveis de entrada (HECHT-NIELSEN, 1989). Esse tipo de estruturação é capaz de resolver problemas multivariáveis, no entanto, possui algumas restrições no que se refere a problemas complexos (HAYKIN, 2001).

Funções radiais representam uma classe especial de funções cujo valor diminui ou aumenta em relação à distância de um ponto central (BRAGA *et al.*, 2011). Em uma topologia básica (Figura 3), a camada de entrada conecta a RBF ao ambiente (agrupa os dados de entrada em *clusters*); sua camada oculta, aplica uma transformação não linear dos dados de entrada mapeando-os para um espaço de alta dimensionalidade (geralmente são utilizadas funções de ativação de base radial gaussianas); e a sua camada de saída, que é responsável por aplicar uma transformação linear nos dados, fornecendo uma saída da rede (HAYKIN, 2009).

Figura 3 - Arquitetura e topologia de uma Rede Funções de Base Radial com “v” variáveis de entrada, “n” neurônios na camada intermediária e uma camada de saída.



Fonte: O autor.

Assim como para a MLP, nas redes RBF a escolha da arquitetura a ser utilizada também fica a cargo do pesquisador. Sendo o algoritmo de aprendizado adotado, fator determinante para o treinamento da rede. Seu treinamento é realizado em duas etapas, e por isso as redes RBF podem ser classificadas como híbridas (PARK E SANDBERG, 1991). Na primeira etapa é adotado um método de aprendizagem auto-organizado ou não supervisionado onde o objetivo principal é formar grupos de indivíduos semelhantes para a posterior obtenção dos pesos das funções de base radial. Essa etapa é realizada com o auxílio de métodos de agrupamento de otimização provenientes da estatística multivariada, como o por exemplo o método de k-means. Já na segunda etapa, o treinamento é feito com base na regra delta generalizada, de modo similar ao utilizado quando se adota uma arquitetura de rede de múltiplas camadas (SILVA *et al*., 2010).

2.7.3. Árvore de regressão e seus refinamentos (*Boosting*, *Bagging* e *Random Forest*)

A árvore de regressão visa subdividir o conjunto de observações várias vezes de forma que os subgrupos subsequentemente formados sejam cada vez mais homogêneos (SOUSA *et al*., 2021). A estrutura da árvore de regressão foi feita

buscando a árvore que levaria à partição dos dados até a formação dos grupos homogêneos. Para isso, avalia-se o quão razoável é uma dada árvore T através de seu erro quadrático médio, conforme a equação abaixo:

$$P(T) = \sum_{R} \sum_{k \in R} (y_k - \hat{y}_R)^2$$

onde: $\hat{y}_R$ é o valor previsto para a resposta fenotípica do traço e y_k é o verdadeiro valor do traço de cada indivíduo dentro do grupo.

Em uma segunda etapa, é realizada poda para tornar a árvore de regressão menor e menos complexa e diminuir a variância desse estimador. Cada nó é removido, um de cada vez, observando como o erro de predição varia no conjunto de validação. Posteriormente, com base nas observações, é decidida a menor árvore com o mínimo erro quadrático médio.

Geralmente, uma única árvore não tem uma boa precisão preditiva quando comparada a outras abordagens. Alguns refinamentos para melhorar o desempenho do método de árvore de regressão são apresentados na literatura e apresentam desempenhos superiores (SOUSA *et al.*, 2021). Desta forma, o desempenho preditivo dos métodos *Bagging* (BA), *Random Forest* (RF) e *Boosting* (BO) também são alternativas interessantes.

Um dos problemas apresentados pela árvore de regressão é a grande variabilidade entre os resultados obtidos. Para contornar este problema, BA é um modelo que aplica a técnica de *bootstrap*, ou seja, S amostras do conjunto de observações são obtidas, com reposição, adquirindo assim um número de S modelos f^1 (x), f^2 (x),..., f^S (x). A média aritmética desses modelos é o resultado final. O RF segue a mesma ideia que BA, porém, além do conjunto de observações, também altera o número de variáveis preditoras ($m = \sqrt{p}$) usadas em cada partição. Já o BO cria árvores sequencialmente usando informações de árvores anteriores, ao contrário do BA, que cria várias árvores independentes (JAMES *et al.*, 2013).

3. REFERÊNCIAS BIBLIOGRÁFICAS

ACQUAAH, G Principles of plant genetics and breeding. Blackwell, **Oxford**, p. 385, 2007.

BERED F; BARBOSA NETO JF; CARVALHO FIF. Marcadores moleculares e sua aplicação no melhoramento genético de plantas. **Ciência Rural**, Santa Maria, v. 27, n. 3, p. 513-520, 1997.

BERNARDO, Rex. Genomewide selection for rapid introgression of exotic germplasm in maize. **Crop Sci**. 49, 419–425, 2009.

BERNARDO, R.; YU, J. Prospects for Genomewide Selection for Quantitative Traits in Maize. **Crop Sci**. 47. 1082 – 1090, 2007

BERNARDO, R. **Breeding for quantitative traits in plants**. Woodbury: Stemma, 2002. 368p.

BORÉM A; MIRANDA GV; FRITSCHE-NETO R **Melhoramento de Plantas** - 7ª Edição. Editora UFV. p. 543, 2017.

BRAGA, A. de P.; CARVALHO, A.P. de L. e de.; LUDERMIR, T.B. **Redes neurais artificiais: teoria e aplicações.** 2.ed. Rio de Janeiro: LTC, 2011.

BUENO, L.C.S.; MENDES, A.N.G. CARVALHO, S.P. **Melhoramento Genético de Plantas**: Princípios e Procedimentos. Lavras: UFLA. 282p, 2001.

BURGUEÑO, J., G. DE LOS CAMPOS, K. WEIGEL, AND J. CROSSA. Genomic Prediction of Breeding Values when Modeling Genotype × Environment Interaction using Pedigree and Dense Molecular Markers. **Crop Sci.** 52:707-719, 2012.

CHEVERUD, J.M.; ROUTMAN, E.J. Epistasis and its contribution to genetic variance components. **Genetics**. 1995 Mar;139(3):1455-61. PMID: 7768453; PMCID: PMC1206471.

CORDELL, Heather J. Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans**. Human Molecular Genetics** 11, 2463–2468, 2002.

CROSSA, J.; PÉREZ-RODRÍGUEZ, P.; CUEVAS, J.; MONTESINOS-LÓPEZ, O.; JARQUÍN, D.; DE LOS CAMPOS, G.; BURGUEÑO, J.; CAMACHO-GONZÁLEZ, J.M.; PÉREZ-ELIZALDE, S.; BEYENE, Y.; DREISIGACKER, S. Genomic selection in plant breeding: Methods, models, and perspectives. **Trends in plant science**, 2017.

CROSSA, J.; PÉREZ, P.; CAMPOS, G. de LOS.; MAHUKU, G.; DREISIGACKER, S.; MAGOROKOSHO, C. Genomic Selection and Prediction in Plant Breeding. **Journal of Crop Improvement**, v. 25, n. 3, p. 239–261, 2011.

CRUZ, C.D.; NASCIMENTO, M. **Inteligência computacional aplicada ao melhoramento genético**. Viçosa – MG, Editora UFV, 414p. 2018.

CRUZ, C.D.; CARNEIRO, P.C.S.; REGAZZI, A.J. **Modelos biométricos aplicados ao melhoramento genético**. Editora UFV, V.2, p. 668, 2014.

CRUZ, C.D.; SALGADO, C.C.; BHERING, L.L. **Genômica aplicada**. Visconde do Rio Branco, MG: Suprema, 424f. 2013.

CRUZ, C.D.; REGAZZI, A.J.; CARNEIRO, P.C.S. **Modelos Biométricos Aplicados ao Melhoramento Genético**. 4. Ed. Viçosa: UFV, 2012. 1 v.

ENTRINGER, G.C.; SANTOS, P.H.A.D.; VETTORAZZI, J.C.F.; CUNHA, K.S.; PEREIRA, M.G. Correlação e análise de trilha para componentes de produção de milho superdoce. **Revista Ceres**. Viçosa, v. 61, n. 3, p. 356-361, mai./jun., 2014.

FALCONER, D.S.; MACKAY, T.F.C. **Introduction to quantitative genetics**. 4.ed. Edinburgh: Longman Group Limited, 464p., 1996.

FERREIRA, R.A.D.C.; SILVA, G.N., GLÓRIA, L.S.; SANT'ANNA, I. de C., RODRIGUES, H.S.; SILVA, F.F.E.; CRUZ, C.D. RNA – Aplicação em Estudos de Seleção Genômica Ampla. In Cruz, C.D.; Nascimento, M. (Eds.), **Inteligência computacional aplicada ao melhoramento genético** (1st ed., pp. 241–261). Editora UFV, 2018.

FERRÃO, R.G.; MOREIRA, S.O.; FERRÃO, M.A.G.; RIVA, E.M.; ARANTES, L.O.; COSTA, A.F.S.; CARVALHO, P.L.P.T.; GALVÊAS, P.A.O. Genetics and plant breeding: development and recommendation of cultivars with tolerance to drought in the State of Espírito Santo, Brazil. **Incaper em Revista**, Vitória, v. 6 e 7, n. 4, p. 51-71, jan 2015/dez 2016.

FRITSCHE NETO, R. Seleção genômica ampla e novos métodos de melhoramento do milho. 2011. 39 f. **Tese (Doutorado em Genéica e Melhoramento de Plantas) – Universidade Federal de Viçosa**, Viçosa, 2011.

GIANOLA, D.; OKUT, H.; WEIGEL, K.A.; ROSA, G.J. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. **BMC genetics**, 12(1), p.87, 2011.

GIANOLA, D.; DE LOS CAMPOS, G. Inferring genetic values for quantitative traitsnon-parametrically. **Genetics Research**, 90(6), pp.525-540, 2008.

GIANOLA, D.; FERNANDO, R.; STELLA, A. Genomic-assiste prediction of genetic value with semiparametric procedures. **Genetics** 173: 1761–1776, 2006.

GODDARD, M.E.; HAYES, B.J. **Genomic selection. Journal of Animal Breeding and Genetics**.V.124, n.6, p. 323–330, 2007.

GONZÁLEZ-RECIO, O.; ROSA, G.J.M.; GIANOLA, D. Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. **Livestock Science**, 166, 217– 231, 2014.

HASTBACKA, J.; CHAPELLE, A.; MAHTANI, M. M.; CLINES, G.; REEVE-DALY, M. P.; DALY, M.; HAMILTON, B. A.; KUSUMI, K.; TRIVEDI, B.; WEAVER, A. The

diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. **Cell**, v. 78, n. 1, p. 1073–1087, 1994.

HAYES, B.J.; COGAN, N.O.I.; PEMBLETON, L.W.; GODDARD, M.E.; WANG, J.; SPANGENBERG, G.C.; FORSTER, J.W. Prospects for genomic selection in forage plant species. **Plant Breeding**, v. 132, n. 2, p. 133–143, 2013.

HAYKIN, S. **Neural Networks and Learning Machines**. 3ed. Ontario:McMaster University Hamilton, 2009.

HAYKIN, Simon. **Redes neurais: princípios e prática**. 2ed. Porto Alegre: Bookman, 2001.

HE, X.; QIAN, W.; WANG, Z.; LI, Y.; ZHANG, J. Epistasia positiva prevalente em redes metabólicas de Escherichia coli e Saccharomyces cerevisiae. **Nat. Genet**. 42, 272-276, 2010.

HECHT-NIELSEN , R. Theory of the Backpropagation Neural Network; Neural Networks. IJCNN., **International Joint Conference**. pp 593 – 605. Washington, USA, 1989.

HOLLAND, J. B. Implementation of molecular markers for quantitative traits in breeding programs - challenges and opportunities. In: FISCHER, T. **New directions for a diverse planet: Proceedin GWS for the 4th International Crop Science Congress. Brisbane**. p. 1–13, 2004.

HOSPITAL, F.; MOREAU, L.; LACOUDRE, F.; CHARCOSSET, A.; GALLAIS, A. More on the efficiency of marker-assisted selection. **Theoretical and Applied Genetics**, v. 95, n. 1, p. 1181–1189, 1997.

JASNOS L. E KORONA R. Tamponamento epistático de perda de aptidão em cepas de dupla deleção de levedura. **Nat. Genet**. 39, 550-554, 2007.

JEHAN, T.; LAKHANPAUL, S. Single nucleotide polymorphism (SNP)– Methods and applications in plant genetics: A review. **Indian Journal of Biotechnology**, v. 5, n. 4, p. 435–459, 2006.

LI Z, PINSON SR, PARK WD, PATERSON AH, STANSEL JW. Epistasis for three grain yield components in rice (Oryza sativa L.). **Genetics** 145:453–65, 1997.

LONG, N.; GIANOLA, D.; ROSA, G.J.M.; WEIGEL, K.A. Dimension reduction and variable selection for genomic selection: application to predicting milk yield in Holsteins. **Journal of Animal Breeding and Genetics**, 128(4), 247-257, 2011.

MACKAY, Trudy F. C. The Genetic Architecture of Quantitative Traits. **Annual Review of Genetics**, 35(1), 303–339, 2001.

MARCHIORO, V.S.; CARVALHO, F.I.F.; OLIVEIRA, A.C.; CARGNIN, A.; LORENCETTI, C.; BENIN, G.; SILVA, J.A.G.; SIMIONI, D.; HARTWIG, I.; SCHIMIDT,

D. Peso de panícula como critério de seleção indireta, visando ao incremento do rendimento de grãos em aveia / panicle weight as a criterion of indirect selection for grain yield increase in oat. **Revista Ceres**, 51(298): 683-692, 2004.

MARCHINI, J.; DONNELLY, P.; AND CARDON, L.R. Genome-wide strategies for detecting multiple loci that influence complex diseases. **Nature Genetics** 37, 413–417, 2005

MEUWISSEN, T.H.E.; HAYES, B.J.; GODDARD, M.E. Prediction of total genetic value using genome wide dense marker maps. **Genetics,** v. 157, p. 1819-1829, 2001.

MORAES JÚNIOR, Odilon Peixoto de. Seleção genômoca ampla (GWS) aplicada ao melhoramento populacional**. Revisão bibliográfica apresentada à Coordenação do Programa de Pós-Graduação em Genética e Melhoramento de Plantas, da Universidade Federal de Goiás.** Goiânia, 2013.

MORRIS, A.P.; CARDON, L.R. Genome-Wide Association Studies. Handbook of Statistical Genomics, **Fourth Edition**, Volume 2. Edited by David J. Balding, Ida Moltke and John Marioni. 2019 John Wiley & Sons Ltd. Published 2019 by John Wiley & Sons Ltd, 2019.

PARK, J.; SANDBERG, I.W. Universal approximation using radial basis function networks. **Neural Comput**. 3ed, v2, p:246–259, 1991.

PATERSON, A.H.; TANKSLEY S.D.; SORRELLS M.E. DNA markers in plant improvement. **Advances in Agronomy**, San Diego, v. 46, p. 39-90, 1991.

PODLICH, D.W.; WINKLER, C.R.; COOPER, M. Mapping As You Go. **Crop Sci**. 44:1560-1571, 2004.

RIOS, S.A.; BORÉM, A.; GUIMARÃES, P.E.O.; PAES, M.C.D. Análise de trilha para carotenoides em milho. **Revista Ceres**. Viçosa, v. 59, n.3, p. 368-373, mai./jun. 2012.

RESENDE, M.D.V.; SILVA, F.F. e; Azevedo, C.F. **Estatística Matemática, Biométrica e Computacional:** Modelos Mistos, Multivariados, Categóricos e Generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição, Sobrevivência. Editora Suprema, Viçosa, 2014.

RESENDE JÚNIOR, M.F.R.; ALVES, A.A.; SÁNCHEZ, C.F.B.; RESENDE, M.D.V.; CRUZ, C.D. **Seleção Genômica Ampla**. In: Cruz, C.D.; Salgado, C.C.; Bhering, L.L. Genômica Aplicada. Visconde do Rio Branco MG, Suprema, 424f., 2013.

RESENDE JÚNIOR, M. F. R. Seleção genômica ampla no melhoramento vegetal. 2010. 78 f. **Dissertação (Mestrado em Genética e Melhoramento de Plantas) – Universidade Federal de Viçosa**, Viçosa, 2010.

RESENDE, Marcos Deon Vilela de. **Genômica quantitativa e seleção no melhoramento de plantas perenes e animais**. Colombo: Embrapa Florestas, p. 330, 2008.

RESENDE, Marcos Deon Vilela de. **Genética biométrica e estatística no melhoramento de plantas perenes**. Brasília, 2002.

SANT'ANNA, I. de C.; NASCIMENTO, M.; SILVA, G. N.; CRUZ, C. D.; AZEVEDO, C. F.; GLORIA, L. S.; SILVA, F. F. e. Genome-enabled prediction of genetic values for using radial basis function neural networks. **Functional Plant Breeding Journal** [Internet], 1(2595–9433), 1–8, 2020.

SAX, Karl. The association of size diferences with seed-coat pattern and pigmentation in Phaseolus vulgaris. **Genetics, Bethesda**, v. 8, p. 552-560, 1923.

SEARCHINGER, I.; WAITE, R.; HANSON, C.; RANGANATHAN, J. **Creating a Sustainable Food Future: A Menu of Solutions to Feed Nearly 10 Billion People by 2050**. Washington, DC, USA, July 2019. 556p.

SEGRE, D.; DELUNA, A.; CHURCH, G.; *et al*. Modular epistasis in yeast metabolism. **Nat Genet** 37, 77–83 2005.

SILVA, G.N.; TOMAZ, R.S.; SANT'ANNA, I. de C.; NASCIMENTO, M.; BHERING, L.L.; CRUZ, C. D. Neural networks for predicting breeding values and genetic gains. **Scientia Agricola**, 71(6), 494–498, 2014.

SILVA, I.N.; SPATTI, H.D.; FLAUZINO, R.A. **Redes Neurais Artificiais: para engenharia e ciências aplicadas**. São Paulo: Artliber, 399p. 2010.

SOLBERG, T.R.; SONESSON, A.K.; WOOLLIAMS, J.A.; MEUWISSEN, T.H.E. Reducing dimensionality for prediction of genome-wide breeding values. **Genetics Selection Evolution**, 41:299, 2009.

SOUSA, I.C.; NASCIMENTO, M.; SILVA, G.N.; NASCIMENTO, A.C.C.; CRUZ, C.D.; SILVA, F.F.E.; DE ALMEIDA, D.P.; PESTANA, K.N.; AZEVEDO, C.F.; ZAMBOLIM, L.; CAIXETA, E.T. Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. **Scientia Agricola**, 78(4), 2021.

TIBSHIRANI, Robert. Regression shrinkage and selection via the LASSO. **J. R. Stat. Soc. B** 58: 267–288, 1996.

TOMAZ, R.S.; ALVES, D.P.; NASCIMENTO, M.; CRUZ, C.D. Capítulo 2 - Inteligência Computacional. In: CRUZ, C.D.; NASCIMENTO, M. **Inteligência computacional aplicada ao melhoramento genético.** Viçosa – MG, Editora UFV, 414p. 2018.

XAVIER, T.D.A.; FILHO, L.N.; LOPES, S.S.S. **Análise prospectiva do algodão transgênico no brasil**. Cadernos de Prospecção, Salvador, v. 11, n. 3, p. 927-939, setembro, 2018.

ZENG, Z-B.; LIU, J.; STAM, L.F.; KAO, C-H.; MERCER, J.M.; *et al*. Genetic architecture of a morphological shape difference between two Drosophila species. **Genetics** 154:299–310, 2000..

LONG, A.D.; MULLANEY, S.L.; REID, L.A.; FRY, J.D.; LANGLEY, C.H.; *et al*. High resolution mapping of genetic factors affecting abdominal bristle number in **Drosophila melanogaster. Genetics** 139:1273–91, 1995.

4. ARTIGO 1

Genome-enabled prediction through machine learning methods considering different levels of traits complexity

Ivan de Paiva Barbosa[1]*, Michele Jorge da Silva[1], Weverton Gomes da Costa[1], Isabela de Castro Sant'Anna[2], Moysés Nascimento[3], Cosme Damião Cruz[1]

[1] Department of General Biology, Bioinformatics Laboratory, Federal University of Viçosa – UFV, Viçosa, MG, Brazil.

[2] Rubber Tree and Agroforestry Center, Agronomic Institute - IAC, Votuporana, SP, Brazil.

[3] Department of Statistics, Laboratory of Computational Intelligence and Statistical Learning, Federal University of Viçosa – UFV, Viçosa, MG, Brazil.

*Corresponding author, e-mail: ivan.barbosa@ufv.br

Published in:
Crop Science. 2021;1–13.
DOI: 10.1002/csc2.20488

Received: 23 December 2020 | Accepted: 13 February 2021

DOI: 10.1002/csc2.20488

Crop Science

ORIGINAL RESEARCH ARTICLE

Crop Breeding & Genetics

Genome-enabled prediction through machine learning methods considering different levels of trait complexity

Ivan de Paiva Barbosa[1] | **Michele Jorge da Silva**[1] | **Weverton Gomes da Costa**[1] | **Isabela de Castro Sant'Anna**[2] | **Moysés Nascimento**[3] | **Cosme Damião Cruz**[1]

[1] Department of General Biology, Bioinformatics Laboratory, Federal University of Viçosa – UFV, Viçosa, MG, Brazil

[2] Rubber Tree and Agroforestry Center, Agronomic Institute - IAC, Votuporanga, SP, Brazil

[3] Department of Statistics, Laboratory of Computational Intelligence and Statistical Learning, Federal University of Viçosa – UFV, Viçosa, MG, Brazil

Correspondence
Ivan de Paiva Barbosa, Department of General Biology, Bioinformatics Laboratory, Federal University of Viçosa – UFV, Viçosa, MG, Brazil.
Email: ivan.barbosa@ufv.br

Assigned to Associate Editor Timothy Beissinger.

Abstract

Genomic-wide selection (GWS) consists of the use of a large number of molecular markers for the prediction of genetic values and has been shown to be highly relevant for genetic improvement. The objective of this work was to evaluate and compare the predictive performance of statistical (ridge regression-best linear unbiased predictor [RR-BLUP] and BayesB) and machine learning methods through GWS in simulated populations with traits presenting different levels of heritability and quantitative trait loci (QTL) numbers in the presence of dominant and epistatic effects. The simulated genome of population F_2 was formed by 1,000 individuals and genotyped with 2,010 single nucleotide polymorphism (SNP) markers. Twenty-six traits were simulated considering QTL numbers ranging from two to 88 and heritabilities of .3 and .6. The selective and predictive performances were evaluated using the multilayer perceptron (MLP), radial basis function (RBF), decision trees (DT), bagging (BA), random forest (RF), and boosting (BO) machine learning models and the classical RR-BLUP and BayesB methods. A high effect of heritability was observed for the results of selective accuracy when compared to the increased QTL number. In addition, the selective accuracy based on the number of QTL demonstrates that the application of alternative machine learning models, such as RBF, BA, BO, and RF, can be suitable for the analysis according to QTL number. Machine learning methods are powerful tools for predicting genetic values with epistatic gene control in traits with different degrees of heritability and different numbers of controlling genes.

1 | INTRODUCTION

The world population could reach almost 10 billion people by 2050 (Searchinger et al., 2019), and agriculture will experience challenges in meeting the growing demand for food. Plant breeding is one of the main tools to increase productive potential, nutritional value, and to ensure plants adapt to changing environmental conditions, enabling the production of cultivars in regions previously of low productive potential.

The selection of superior individuals is one of the main challenges for plant breeders. In this context, genomic-wide selection (GWS) allows for selection. GWS, as proposed by (Meuwissen et al., 2001), has become an important tool

Abbreviations: BA, bagging; BO, boosting; DT, decision trees; GWS, genomic-wide selection; MLP, multilayer perceptron; QTL, quantitative trait loci; RBF, radial basis function; RF, random forest; RR-BLUP, ridge regression-best linear unbiased predictor; SNP, single nucleotide polymorphism.

to help breeders due to its performance as a prediction model by associating marker information with phenotypic information (Alkimim et al., 2020; Crossa et al., 2017; Sousa et al., 2019). The combined genotypic and phenotypic data can be used to estimate the genetic merit or predict phenotypic values of the trait of interest. However, some statistical challenges are faced by traditional genomic prediction methods, such as high dimensionality, because the number of markers (m) is much greater than the number of observations (n; $m > n$; Akdemir et al., 2017; Azevedo et al., 2014; Crossa et al., 2017).

Different methods capable of addressing problems related to data dimensionality have been used, such as ridge regression-best linear unbiased predictor (RR-BLUP; Endelman, 2011) and Bayesian methods (Meuwissen et al., 2001). More recent approaches based on machine learning have also been adopted through the methods of radial basis function (RBF), multilayer perceptron (MLP; Ehret et al., 2015; Sant'Anna et al., 2020; Silva et al., 2014), and decision trees (DT) with their refinements boosting (BO), random forest (RF), and bagging (BA; de Sousa et al., 2021).

Machine learning-based methods also have other properties that make them potentially more attractive to GWS (Ferreira et al., 2018; Sant'Anna et al., 2020; Silva et al., 2014; de Sousa et al., 2021). This is because all markers have a chance to contribute to the adjustment of the model, including those with weak effects and highly correlated and interacting markers. Additionally, these methods allow complex interactions between markers to be easily included, such as in nonadditive models with dominant and epistatic effects, and do not make distributive assumptions about the predictor variables (I. C. de Ferreira et al., 2018; Sant'Anna et al., 2020; Silva et al., 2014; de Sousa et al., 2021).

However, genomic selection analysis information on the predictive efficiency of the different techniques and how they behave according to the number of controlling genes (oligogenic and polygenic traits) for different heritability scenarios and in the presence of nonadditive effects (epistasis and dominance) is scarce in the literature. In this context, two hypotheses were formulated. The first hypothesis is that the complexity of the analyzed traits can influence the performance of the techniques, and the second hypothesis is that different techniques provide different results due to this complexity. Thus, the present work aims to evaluate and compare the predictive response of statistical methods (RR-BLUP and BayesB) with machine learning methods through GWS in simulated populations, presenting traits with different levels of heritability and quantitative trait loci (QTL) numbers in the presence of dominant and epistatic effects.

Core Ideas

- Currently, there are many forecasting techniques whose comparative efficiency is still the subject of study.
- Adequate knowledge of the techniques on complex traits is useful for the researcher to concentrate efforts.
- The machine learning model can capture nonlinear relationships and does not require a priori distributions.

2 | MATERIALS AND METHODS

2.1 | Genome and simulated populations

Five F_2 populations of a diploid species ($2n = 2x = 20$) with an effective size of 1,000 individuals were simulated. This simulation involved the random combination of 5,000 gametes generated from contrasting homozygous parents (dominant P_1 and recessive P_2). The simulated genome was composed of 10 linkage groups with a size of 100 cM each and comprised 2,010 biallelic single nucleotide polymorphisms (SNPs) distributed equally and equidistantly. For the generation of gametes, the percentage of recombination equivalent to the distance between loci was 0.5 cM, providing linkage disequilibrium. To identify possible SNPs with significant segregation distortion in all five simulated populations, the chi-square test (χ^2) was performed at 5% probability with Bonferroni correction for multiple tests. However, no SNP showed significant distortions for the expected Mendelian segregation in all simulated populations, suggesting that the simulated data for the proposed tests were adequate, and therefore no markers were removed.

2.2 | Simulation of traits

Twenty-six traits were simulated, with heritabilities of .3 or .6 and with numbers of QTL ranging from 2 to 88, distributed between the first eight linkage groups (Table 1), representing different scenarios. The last two linkage groups were used to control for the quality of the evaluated models (Figure 1), contained 201 markers each, had no direct influence, and did not present linkage disequilibrium on the traits.

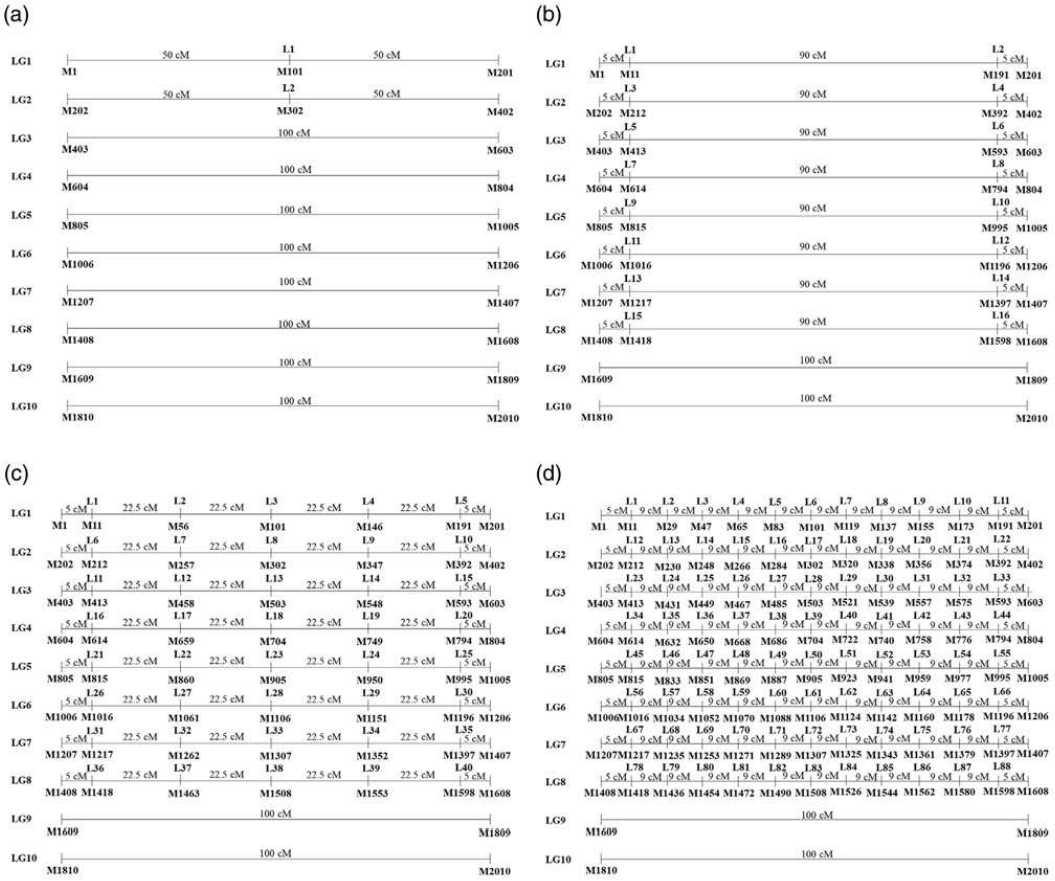


FIGURE 1 Loci distribution model within the linkage groups. Distribution model of the quantitative trait loci (QTL) in the T1 and T2 traits with the central markers in the first two linkage groups (a); in the T7 and T8 traits with 16 QTLs distributed equidistantly within the first eight linkage groups (b); in the T13 and T14 traits with 40 QTLs distributed equidistantly within the first eight linkage groups (c); and in the T25 and T26 traits with 88 QTLs distributed equidistantly within the first eight linkage groups (d). LG, linkage group; L, locus; M, marker

TABLE 1 Number of quantitative trait loci (QTLs) equally distributed in eight linkage groups and heritabilities (h^2) of the 26 simulated traits (T1 to T26)

	QTLs												
h^2	2	4	8	16	24	32	40	48	56	64	72	80	88
.3	T1	T3	T5	T7	T9	T11	T13	T15	T17	T19	T21	T23	T25
.6	T2	T4	T6	T8	T10	T12	T14	T16	T18	T20	T22	T24	T26

Two QTL controlled the T1 and T2 traits, defined by the central markers of the first two linkage groups (Figure 1a). Four QTL controlled the T3 and T4 traits which were represented by the central markers of the first four linkage groups. Eight QTL controlled traits T5 and T6 which were represented by the central markers of the first eight linkage groups. For the other traits (T7 to T26), shown in Table 1, the QTL were allocated to markers located 5 cM from the ends, and the other QTL were distributed equidistantly within the first eight linkage groups, exemplified for traits T7 and T8, T13 and T14, and T25 and T26 in Figures 1b, 1c, and 1d, respectively.

2.3 | Phenotypic values associated with the traits

All traits were simulated considering a mean equal to 100, a variation coefficient of 10%, an average degree of dominance ($\hat{d}/\hat{a}$) equal to 0.5, and an epistatic effect model, as described in the following equation:

$$Y_i = \mu + \sum_{j=1}^{n} \alpha_j + \sum_{j=1}^{n-1} \alpha_j \alpha_{j+1} + e_i$$

where Y_i is the phenotypic value for observation i; μ is a general mean; α_j assumes the values $u + a_j$, $u + d_j$ and $u - a_j$ for the genotypes associated with classes AA, Aa and aa, respectively, with u being the mean of the homozygotes, a_j is half the difference in genotypic value between both homozygotes, and d_j is the difference between the genotypic value of the heterozygote and the mean of the homozygotes. Coding 1, 0, or −1 identified these classes, respectively. In the above equation, the first summation of the expression refers to the contribution of the individual locus through its additive and dominant effects, and the second summation represents the multiplicative effects corresponding to the epistatic interactions between pairs of loci. In this summation $\alpha_j\ \alpha_{j+1}$ is the multiplicative effect of the favorable allele in locus j and j+1 to the manifestation of the trait under consideration (with n ranging from 2 to 88). The importance of the gene action was expressed by weights obtained of a uniform distribution so that each locus j had the same contribution to the trait under consideration. The e_i is the environmental effect, generated according to the variance structure of the residuals was given by $e \sim N(0, V_e)$, with $V_e = [V_g(1 - h^2)]/h^2$.

2.4 | Prediction of genetic values

From the simulated data, the effects of markers were estimated, and the genomic estimated breeding values (GEBVs) were predicted using the methods of MLP and RBF neural networks and DT with refinements BO, BA, and RF. The results found were compared to those obtained by the ridge regression-best linear unbiased predictor (RR-BLUP) and BayesB methods.

2.5 | RR-BLUP and BayesB

RR-BLUP was used as described by (Meuwissen et al., 2001) according to the following model:

$$\mathbf{y} = \mathbf{Wb} + \mathbf{Xm} + \mathbf{e}$$

where $\mathbf{y}$ is the vector of phenotypic observations; $\mathbf{b}$ is the vector of fixed effects (general average) with incidence matrix $\mathbf{W}$; $\mathbf{m}$ is the vector of the random effects of markers with incidence matrix $\mathbf{X}$ with $\mathbf{m} \sim N(0, I\sigma_g^2)$; and $\mathbf{e}$ is the vector of random errors with $\mathbf{e} \sim N(0, I\sigma_e^2)$, where σ_g^2 is the variance of the error. $\mathbf{X}$ is the incidence matrix composed of values 1, 0, and −1 according to the number of marker alleles of the MM, Mm, and mm genotypes, respectively.

The prediction equations were modeled assuming a priori that all loci explained equivalent amounts of the genetic variation, and therefore presented a common σ_g^2. Thus, the genetic variation explained by each locus is σ_g^2/n, where σ_g^2 is the total genetic variation and n is the number of markers used.

BayesB was also used as described by (Meuwissen et al., 2001), according to the model:

$$\mathbf{y} = \mu\mathbf{1_n} + \sum_{i=1}^{N} \mathbf{z_i} g_i + \mathbf{e}$$

where $\mathbf{y}$ is the vector of phenotypic values; μ is the general average; $\mathbf{1_n}$ is a vector of 1 of size n (number of observations); g_i is the random effect of marker i with a distribution of $N(0, \sigma_{gi}^2)$; N is the total number of SNP type markers; z_i is the design vector corresponding to g_i; and $\mathbf{e}$ is the vector of residual errors with a distribution of $N(0, \sigma_e^2)$. It was assumed that the variation of the effects of the marker (σ_{gi}^2) was equal to zero with a probability of $\pi = 0.8$ or followed an inverse χ^2 distribution with a probability of $(1 - \pi)$. Bayesian estimates were obtained via the Monte Carlo Markov chain (MCMC) method until 200,000 iterations. The first 20,000 interactions were discarded (burn-in). Sampling for the calculation of statistics a posteriori was carried out every five interactions (thin). Therefore, 36,000 samples of MCMC were used to construct the subsequent densities.

2.6 | MLP neural network

For MLP, the backpropagation training algorithm with Bayesian regularization was used. This algorithm is a modification of the Levenberg–Marquardt training algorithm to produce networks that generalize well and reduce the difficulty of determining the ideal network architecture (Demuth et al., 2000). The architecture of MLP involved a hidden layer and number of neurons (l) ranging from one to four ($l = 1\ldots, 4$). For the input layer, the matrix of molecular markers with 2,010 markers was provided so that the output layer returned to the predicted phenotypic value of each individual. Figure 2 shows the MLP network used. The activation function used was the hyperbolic tangent, defined by the following equation:

$$\tan(h\ x) = \frac{1 - \exp(-2x)}{1 + \exp(-2x)}$$

2.7 | RBF neural networks

RBF used a feedforward type architecture. This model consists of an input layer, a hidden layer, and an output layer. The input layer connects the network to the environment (groups the input data into clusters). The hidden layer applies a nonlinear transformation from the input space to a high-dimensional

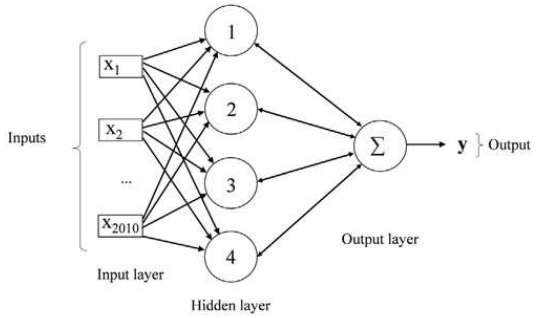


FIGURE 2 Multilayer Perceptron Neural Network with Backpropagation. The inputs X_1 to X_{2010} in the input layer refer to the markers, a hidden layer and the number of neurons ranging from one to four, and in the output layer, the network returns to the predicted value vector (y)

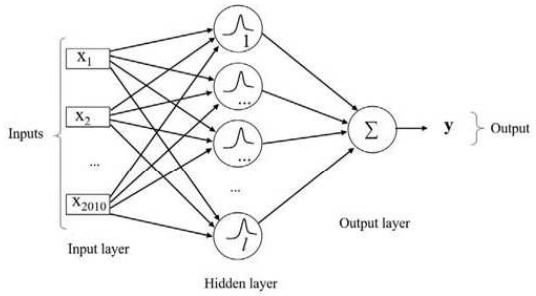


FIGURE 3 Radial Basis Function Neural Networks feedforward. Entries X_1 to X_{2010} in the input layer refer to the markers, a hidden layer with a radius ranging from 10 to 40 ($r = 10\ldots, 40$) and neurons (l) ranging from 10 to 60 ($l = 10\ldots, 60$). In the output, the network returns the vector of predicted values (y)

hidden space. Finally, the output layer applies a transformation to the hidden space, providing an output to the network.

The number of neurons in the hidden layer ranged from 10 to 60 neurons, and the size of the ray ranged in the same amplitude, from 10 to 60. The matrix with 2,010 molecular markers was used in the input layer so that the desired output of the network was the simulated phenotypic values (Figure 3). The activation function used in the hidden layer was Gaussian, according to the following equation:

$$g(u) = \exp\left[-(u-c)^2/2\sigma^2\right]$$

where c is the center of the Gaussian function, σ^2 is the variance of the Gaussian function, and u is the activation potential.

For the output layer, the linear activation function was used according to the following equation:

$$y_{ri} = g\left(x_0 w_0 + \sum_{j=1}^{q} f_{xj}\left(x_i\right) w_j\right)$$

where x_i is the ith input; w_j is the jth synaptic weight; and f_{xj} is hidden layer activation function associated with input x_i weighted by its respective weight.

2.8 | DT and its refinements (BO, BA, and RF)

Decision tree aims to subdivide the set of observations several times in such a way that the subsequently formed subgroups are increasingly homogeneous (de Sousa et al., 2021). The structure of the regression tree was created by searching for the tree that would lead to the partition of data until the formation of homogeneous groups was achieved. For this purpose, the equation was used to evaluate how reasonable a given T tree was through its mean square error:

$$P(\mathrm{T}) = \sum_{R}\sum_{k\in R}(y_k - \hat{y}_R)^2$$

where $\hat{y}_R$ is the predicted value for the phenotypic response of the trait, and y_k is the true value of the trait of each individual within the group.

In a second stage, pruning was performed to make the regression tree smaller and less complex and decrease the variance of this estimator. Each node was removed one at a time to observe how the prediction error varied in the validation set. Subsequently, based on the observations, the smallest tree with the minimum mean square error was decided.

Generally, a single tree does not have good predictive accuracy when compared to other approaches. Some refinements to improve the performance of the DT model are presented in the literature and present superior performances (de Sousa et al., 2021). In this way, the predictive performances of the BA, RF, and BO methods were also tested.

One of the problems presented by DT is the great variability between the results obtained. To address this problem, BA is a model that applies the bootstrap technique, that is, S samples from the set of observations are obtained with replacement, thus acquiring a number S models $f^1(x)$, $f^2(x)$,..., $f^S(x)$. The arithmetic mean of these models is the final model. Random forest follows the same idea as BA; however, in addition to the set of observations, it uses a smaller number of predictive variables in each split. Boosting creates trees sequentially using information from previous trees, unlike BA which creates multiple independent trees (James et al., 2013).

2.9 | Efficiency of techniques

To evaluate the efficiency of the techniques, the parameters used were the root mean square error (RMSE) and the square of correlation (r^2).

Selective accuracy is measured by the square of correlation (r^2) between the estimated values ($\hat{y}$) and true values (y); that is, it measures how much the estimate obtained is related to the real value of the parameter, which in quantitative genetics expresses the heritability of the traits (de Resende et al., 2010). Accuracy was given by the following equation:

$$r^2 = [\text{cor}\ (\hat{y},\ y)]^2$$

The RMSE was used to express the predictive accuracy of the models, as it has the advantage of presenting the error values on the same scale as the variable of interest, as described below:

$$\text{RMSE} = \sqrt{\frac{\sum (\hat{y} - y)^2}{n}}$$

2.10 | Training and validation

For the training and validation of the techniques used, cross-validation (k-fold) was performed with $k = 5$ partitions (Bengio & Grandvalet, 2004). In each of the five rounds, four of these subsets constituted the training population (80% of individuals), and the remaining subset constituted the validation population (20% of individuals). The techniques were compared based on the arithmetic mean of the five performance estimates of the validation sets.

2.11 | Statistical test

Each of the five simulated populations was considered a replicate. Analysis of variance of the data was performed, obtaining the mean squares of the factors by the F test. The RMSE and the r^2 of validation from each technique were subjected to the Tukey test at a 5% probability.

2.12 | Computational aspects

Population simulations were performed using GENES software (Cruz, 2013). The RR-BLUP, BayesB, DT, BO, BA, and RF techniques were performed with Genes software integrated with R software (Cruz, 2016; R Core Team, 2019) using the packages rrBLUP (RR-BLUP), BGLR (BayesB), tree (DT), randomForest (BA and RF), and gbm (BO) and the functions mixed.solve (RR-BLUP), BGLR (BayesB), tree (DT), randomForest (BA and RF), and gbm (BO). To perform the MLP and RBF methods, Genes software was integrated with MATLAB software (MATLAB, 2019; Cruz, 2016).

3 | RESULTS AND DISCUSSION

3.1 | Selective accuracy (r^2)

Higher selective accuracy (r^2) was observed in scenarios with greater heritability (Figure 4). However, there were responses to selective precision as a function of the number of QTL for both the .3 and .6 heritability scenarios (Figure 4).

The results obtained were able to demonstrate the strong effect of heritability on the results of r^2 when compared to the increase in the number of QTL. In addition, the results of r^2 as a function of QTL numbers demonstrated that the application of alternative models, such as RBF, BA, BO, and RF, presented superior responses (Figure 4).

The adverse effect of lower heritability on selective accuracy in several scenarios has been proven in previous studies that used the RR-BLUP (Coutinho et al., 2018; Moura et al., 2019), BayesB methods (Moura et al., 2019), BO (Ghafouri-Kesbi et al., 2017), RF (Ghafouri-Kesbi et al., 2017), RBF (Sant'Anna et al., 2020), and MLP methods (Coutinho et al., 2018). Guo et al. (2014) identified associations between the selective accuracy and genomic heritability of training and test sets. According to these studies, the increase in genomic heritability was due to greater genetic variations and consequently, there was lower environmental noise in these sets, contributing to accurate predictions of marker effects.

Selective accuracy was superior for the RBF model in scenarios of complexity with greater QTL numbers (32 to 88; Figure 4), although no statistically significant differences were observed (Table 2). In these scenarios, only the DT model was statistically inferior to the others for both heritabilities (Figure 4; Table 2). Boosting presented results similar to the other techniques in most scenarios but produced significantly lower results concerning RBF, MLP, RR-BLUP, and BayesB in some of the scenarios when the heritability was .3, demonstrating a greater sensitivity of BO for scenarios of low heritability. On the other hand, in scenarios with lower QTL numbers (2 to 8), DT, BA, and RF showed results that were higher in both heritability scenarios for selective accuracy in relation to other techniques (Figure 4; Table 2).

The low efficiency of DT was also observed by de Sousa et al. (2021) when evaluating real data of molecular markers to predict rust resistance in Arabica coffee. According to James et al. (2013), DT generally does not perform well when compared to other regression approaches, and its low precision is justified by the high variation in terms of forecasting. However, in the present study, these effects were only observed for the most complex scenarios, above 16 QTL. This can be

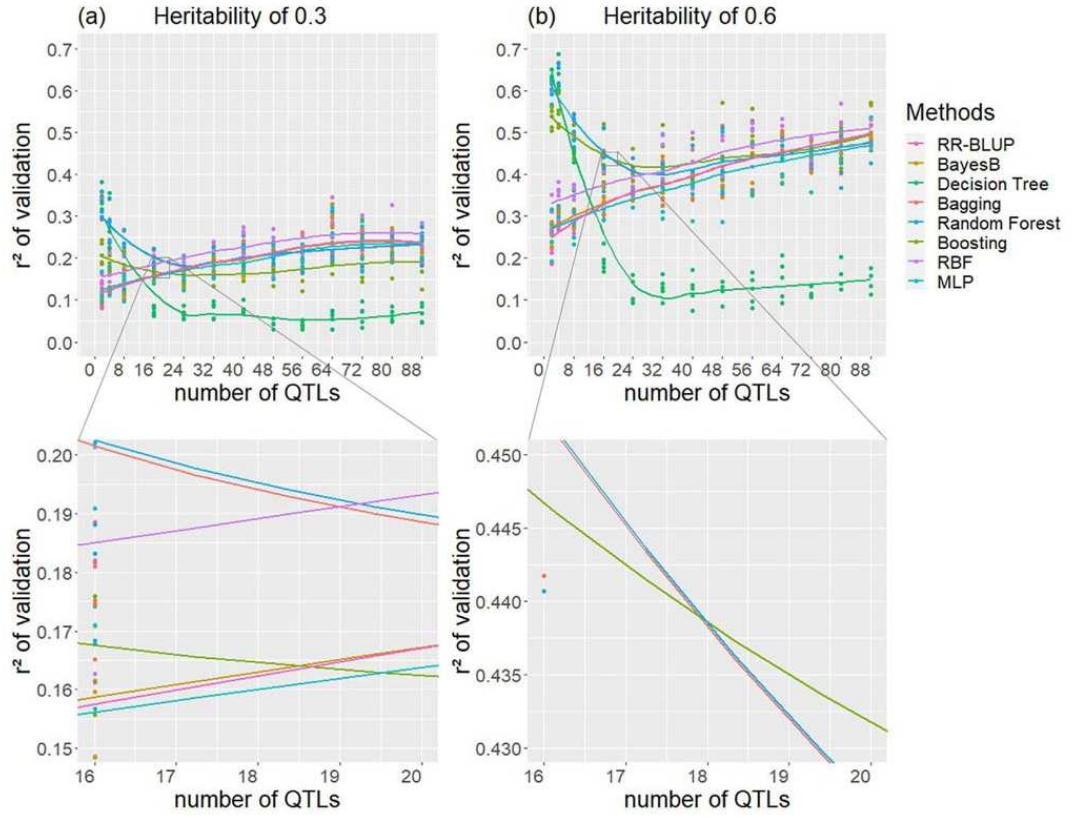


FIGURE 4 Selective accuracy of the techniques through square of correlation (r^2). Selective accuracy (r^2) of the ridge regression-best linear unbiased predictor (RR-BLUP), BayesB, decision tree, bagging, random forest, boosting, radial basis function (RBF), and multilayer perceptron (MLP) methods measured in scenarios with heritabilities of .3 (a) and .6 (b) as a function of quantitative trait loci (QTL) number with additive, dominant and epistatic effects

justified by the greater number of genotypic classes which reduces the representativeness of each genotypic combination in the training set and overparameterization of the model. For the traits controlled by a smaller number of QTL (2 to 8), the number of genotypic classes was drastically reduced, and thus may have increased the representativeness of each class within the data sets in training, resulting in less variation and high efficiency.

The good efficiency of the DT model for scenarios with low numbers of QTL suggests that DT can be used as a tool to signal whether a trait is controlled by a few or many genes when evaluated simultaneously with another technique; however, tests using real data still need to be investigated.

The DT, BA, RF, and BO methods were superior to the others in low QTL scenarios and equal to or lower in the scenario with the highest number of QTLs (Figure 4). In addition, almost identical results were observed between BA and RF for the different QTL values, and thus demonstrated a high overlap of the values for the graphical representation. We highlight the slight difference in the results of these two methods by magnification in Figure 4.

Computational intelligence-based methods do not require specification of the inheritance; therefore a first hypothesis would be that these methods might capture the nonadditive effects known to exist in the simulated genome that were not specified in the RR-BLUP and BayesB methods. In this sense, for traits with the lowest QTL number, the multiplicative effects of the controlling genes (epistasis) may be more important since the individual effect of each gene is greater than in the traits controlled by a greater QTL number. Thus, DT, BA, and RF were able to capture these epistatic effects in traits controlled by few QTL and presented better results (Figure 4).

An alternative way to compare the results of the various methods tested is to emphasize the disturbing effects that act on the analyzed traits, which are environmental noise,

TABLE 2 Selective accuracy (r^2) obtained by ridge regression-best linear unbiased predictor (RR-BLUP), BayesB, decision tree (DT), bagging, random forest (RF), boosting, radial basis function (RBF), and multilayer perceptron (MLP) with different heritability scenarios and quantitative trait loci (QTL) numbers

		r^2 of validation							
Heritability	**No. QTLs**	**RR-BLUP**	**BayesB**	**DT**	**Bagging**	**RF**	**Boosting**	**RBF**	**MLP**
$h^2 = .3$	2	.103c	.115c	.347a	.319a	.320a	.213b	.137c	.116c
	4	.140b	.144b	.273a	.279a	.278a	.191b	.181b	.142b
	8	.129c	.131c	.153bc	.229a	.229a	.185ab	.167bc	.132bc
	16	.174a	.174a	.074b	.185a	.186a	.163a	.191a	.172a
	24	.158a	.158a	.066b	.163a	.165a	.144a	.180a	.159a
	32	.187ab	.186ab	.076c	.195ab	.195ab	.166b	.217a	.174ab
	40	.218ab	.216ab	.080c	.219ab	.221ab	.177b	.247a	.210ab
	48	.200ab	.198ab	.043c	.202ab	.204ab	.158b	.225a	.192ab
	56	.212a	.209a	.042c	.194ab	.193ab	.152b	.227a	.200a
	64	.263a	.261a	.059b	.243a	.244a	.199a	.282a	.253a
	72	.240a	.237a	.065b	.217a	.219a	.198a	.258a	.235a
	80	.236ab	.234ab	.064c	.230ab	.229ab	.193b	.260a	.226ab
	88	.238ab	.236ab	.068c	.231ab	.232ab	.184b	.255a	.233ab
$h^2 = .6$	2	.225d	.255cd	.630a	.596a	.597a	.534b	.303c	.258cd
	4	.288d	.306d	.632a	.627a	.627a	.545b	.361c	.304d
	8	.278d	.285d	.438b	.503a	.504a	.469ab	.360c	.269d
	16	.347bc	.347bc	.194d	.431ab	.431ab	.436a	.382abc	.327c
	24	.356b	.356b	.119c	.397ab	.399ab	.414a	.380ab	.348b
	32	.378a	.377a	.129b	.413a	.412a	.436a	.421a	.363a
	40	.386a	.384a	.124b	.410a	.412a	.413a	.421a	.365a
	48	.417a	.416a	.122b	.421a	.421a	.433a	.438a	.395a
	56	.458a	.457a	.124b	.454a	.454a	.466a	.492a	.446a
	64	.455a	.454a	.149b	.458a	.455a	.454a	.491a	.429a
	72	.445a	.443a	.115b	.428a	.432a	.436a	.464a	.421a
	80	.485a	.483a	.159b	.463a	.462a	.473a	.509a	.459a
	88	.504a	.502a	.143b	.483a	.484a	.504a	.512a	.476a

Note: Means sharing a common letter within a row are not significantly different by the Tukey-test at the 5% level of significance.

epistatic effects, and the complexity of the traits. Decision tree-based methods benefit from situations in which the predictor variables can be partitioned into well-defined regions (de Sousa et al., 2021). In the present study, that situation occurred when few genes were acting on the control; therefore, the classes became evident. With a greater number of genes, the distribution becomes continuous, whereas the DT methods become less efficient.

The MLP and RBF methods are advocated to be efficient for capturing nonlinear effects, in this case, provided by interallelic interactions. However, these methods did not present significant differences concerning the RR-BLUP and BayesB methods. This less favorable performance may have been a consequence of the dimensionality of the problem in which there were 2,010 entries in the network. González-Camacho et al. (2012) reported that RBF proved to be efficient for predicting quantitative traits with complex underlying genetic action under various types of interactions with different environmental conditions. However, RBF showed degraded performance when a large number of nonsignal markers were added as can be seen in the results found by Sant'Anna et al. (2020) which showed that the use of stepwise regression before the use of RBF techniques led to an improvement in the accuracy of prediction of the genetic value and mainly, to a large reduction of the RMSE in addition to facilitating processing and analysis time due to a reduction in dimensionality.

In the present work, this problem may have been minimized in RBF due to its initial stage of clusters of genotypic values. On the other hand, MLP was severely hampered by the excess of markers. In this case, reducing the dimensionality to the most appropriate levels could offer a further increase in accuracy. However, additional investigations on the excess markers in the context of genomic prediction should be carried out.

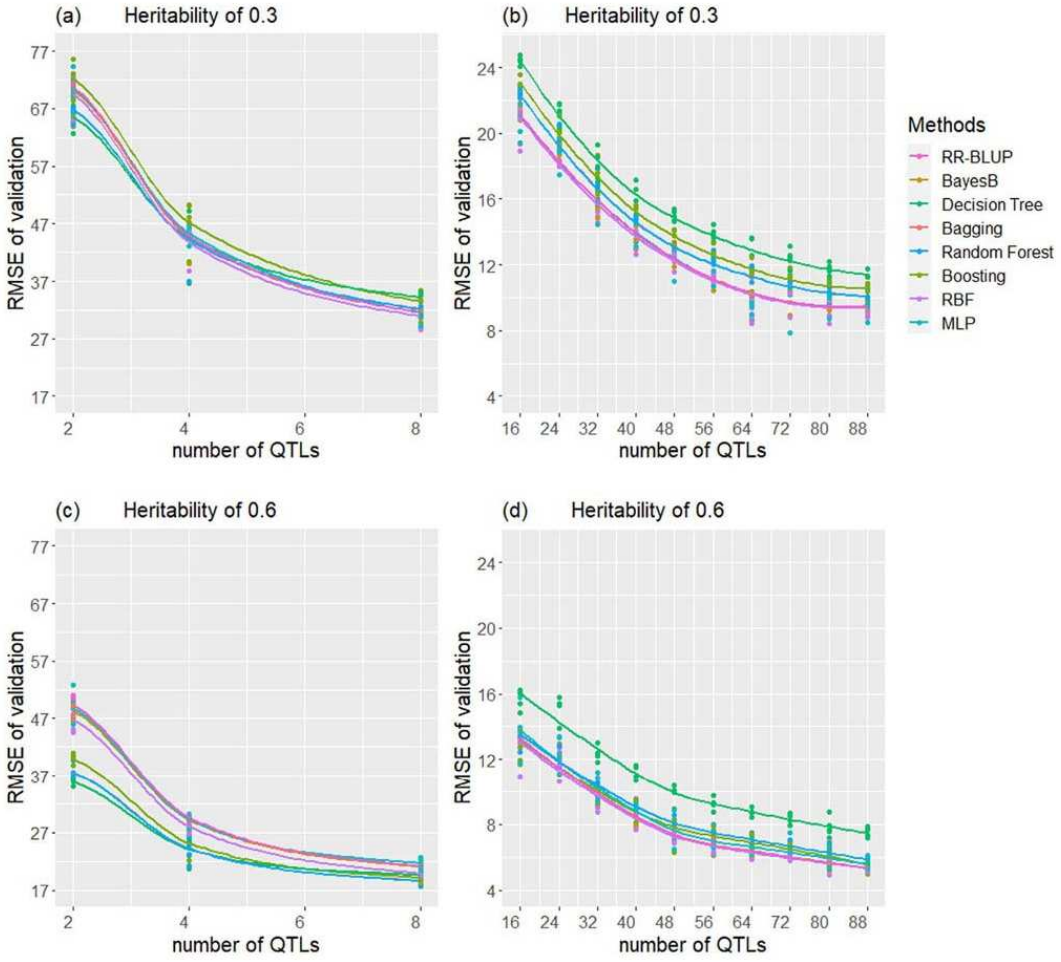


FIGURE 5 Predictive accuracy of the techniques through root mean square error (RMSE). Predictive accuracy (RMSE) obtained by the ridge regression-best linear unbiased predictor (RR-BLUP, BayesB, decision tree, bagging, random forest, boosting, radial basis function (RBF), and multilayer perceptron (MLP) methods, measured in scenarios with a heritability of .3 and QTL numbers ranging from 2 to 8 (a) and from 16 to 88 (b) and with a heritability of .6 and QTL numbers ranging from 2 to 8 (c) and from 16 to 88 (d) with additive, dominant, and epistatic effects

Radial basis function showed superiority, although not significant, in relation to RR-BLUP because it has a consistently higher selective accuracy value for increasingly complex traits. The Gaussian activation function of this technique has been shown to be more appropriate due to the behavior of the action of genes and the effect of the environment on the trait (Figure 4).

An increase in the efficiency of all techniques was observed in scenarios ranging from 24 to 88 QTL (Figure 4). However, with an increase in QTL numbers, it is expected that the total genetic variation will be divided among the QTL; thus, the efficiency of the methods to estimate these small QTL effects will decrease, leading to a loss of precision (Ghafouri-Kesbi et al., 2017; Resende et al., 2012). On the other hand, the increase in efficiency can be explained by the excess marks with null effects in scenarios with low QTL which can hinder the accuracy of the methods (Resende et al., 2012). Therefore, a balance must be sought for the number of markers in the analytical process, according to the expectations regarding the QTL number and to adopt the best analytical strategy.

Although BayesB performs marker selection, eliminating a large number of markers, the performance of this method did not show significant differences in relation to RR-BLUP which considers all markers. BayesB presented a low difference in the average of results compared to RR-BLUP only for the scenarios controlled by two QTL numbers (Table 2). Due

TABLE 3 Predictive accuracy (root mean square error, RMSE) obtained by the ridge regression-best linear unbiased predictor (RR-BLUP), BayesB, decision tree (DT), bagging, random forest (RF), boosting, radial basis function (RBF), and multilayer perceptron (MLP) techniques with different heritability scenarios and quantitative trait loci (QTL) numbers

		RMSE of validation							
Heritability	**No. QTLs**	**RR-BLUP**	**BayesB**	**DT**	**Bagging**	**RF**	**Boosting**	**RBF**	**MLP**
$h^2 = .3$	2	70.62ab	70.16ab	65.52b	66.76ab	66.74b	72.09a	69.44ab	70.40ab
	4	44.84a	44.73a	44.42a	44.17a	44.19a	47.24a	43.76a	45.36a
	8	31.60a	31.57a	34.20a	32.16a	32.13a	33.52a	30.92a	31.54a
	16	20.99c	20.99c	24.42a	22.28bc	22.28bc	23.01ab	20.81c	20.82c
	24	18.78c	18.78c	21.30a	19.49bc	19.48bc	20.20ab	18.55c	18.69c
	32	15.52b	15.53b	18.03a	16.52ab	16.52ab	17.17a	15.31b	15.60b
	40	13.82bc	13.83bc	16.20a	14.46bc	14.44bc	15.12ab	13.61c	13.89bc
	48	12.57c	12.58c	14.99a	13.04bc	13.03bc	13.68b	12.42c	12.49c
	56	11.33d	11.35cd	13.85a	12.38bc	12.38bc	12.94ab	11.25d	11.37cd
	64	9.68b	9.69b	12.66a	10.90ab	10.90ab	11.39ab	9.63b	9.67b
	72	10.20b	10.21b	12.40a	10.91b	10.90b	11.25ab	10.10b	10.06b
	80	9.56c	9.57c	11.74a	10.31bc	10.31bc	10.73ab	9.43c	9.54c
	88	9.35c	9.35c	11.36a	10.01bc	10.01bc	10.51b	9.28c	9.27c
$h^2 = .6$	2	49.23a	48.28a	35.82c	37.50bc	37.43bc	39.84b	46.68a	48.76a
	4	29.63a	29.26a	24.10c	24.22c	24.22c	25.18bc	28.09ab	29.40a
	8	21.12ab	21.01ab	19.57c	18.57c	18.55c	19.10c	19.87bc	21.62a
	16	13.12b	13.12b	15.59a	13.26b	13.26b	12.94b	12.87b	13.66b
	24	11.88b	11.89b	15.23a	12.51b	12.50b	12.12b	11.76b	12.20b
	32	9.83b	9.84b	12.35a	10.09b	10.09b	9.65b	9.58b	10.17b
	40	8.43b	8.44b	11.15a	9.18b	9.17b	8.95b	8.31b	8.77b
	48	7.41b	7.42b	9.85a	8.02b	8.02b	7.78b	7.39b	7.75b
	56	6.62b	6.62b	9.26a	7.42b	7.42b	7.15b	6.52b	6.74b
	64	6.41cd	6.42cd	8.81a	7.20b	7.21b	7.06bc	6.29d	6.67bcd
	72	6.21b	6.22b	8.41a	6.68b	6.67b	6.52b	6.16b	6.58b
	80	5.66b	5.67b	7.83a	6.34b	6.34b	6.15b	5.61b	5.91b
	88	5.28b	5.29b	7.47a	5.82b	5.81b	5.50b	5.31b	5.53b

[a]Note: Means sharing a common letter within a row are not significantly different by the Tukey-test at the 5% level of significance.

to the existence of a high rate of linkage disequilibrium, since it is an F2 population, SNPs closer to QTL are not sampled often enough, so the QTL signal is captured by more distant SNPs. Consequently, the signal of a QTL to BayesB is blurred concerning RR-BLUP in which all markers are examined.

3.2 | Predictive accuracy (RMSE)

The RMSEs of the eight techniques according to the different combinations of heritability scenarios (.3 and .6) and QTL numbers (2, 4, 8, 16, 24, 32, 40, 48, 56, 64, 72, 80, and 88) are shown in Figure 5 and Table 3. The coverage of the genome with 2,010 markers was sufficient to capture the influence of the greater number of QTL (Figure 5) according to the reduction in RMSE relative to the increase in the number of QTL. The RMSEs of the eight techniques according to the different combinations of heritability scenarios (.3 and .6) and QTL numbers (2, 4, 8, 16, 24, 32, 40, 48, 56, 64, 72, 80, and 88) are shown in Figure 5. The coverage of the genome with 2010 markers was sufficient to capture the influence of the greater number of QTL (Figure 5) according to the reduction in RMSE to the increase in the number of QTL. The results showed that there was a reduction in RMSE, that is, fewer deviations between the predicted values in relation to the actual values in scenarios with a higher heritability value and a higher number of QTL (up to 88), and consequently a larger area of gametic phase imbalance since it is an F_2 population. Similar results were observed by Ghafouri-Kesbi et al. (2017), who in different scenarios and with 100 QTL, obtained better predictive accuracy results for the scenarios of greater heritability.

The reduction in RMSE due to an increase in the QTL number may have occurred due to the lesser influence of the

multiplicative effect between the additive and dominant effects which characterize epistatic effects in more complex traits (Coster et al., 2010; Ghafouri-Kesbi et al., 2017). However, Ghafouri-Kesbi et al. (2017) observed a reduction in predictive accuracy when changing from 100 to 1,000 QTL. According to Azevedo et al. (2016), there is an ideal number of markers for each data set, so that smaller predictive accuracy can be observed when using an excess of markers, especially characteristics controlled by a smaller number of QTL.

For the scenarios with the highest number of QTL (16 to 88), DT always presented higher RMSE values (Figure 5). The other methods did not show a significant difference in relation to RMSE (Table 3), with insufficient results to recommend a model at the expense of another. Similar results were also observed in several other studies (Ghafouri-Kesbi et al., 2017; González-Recio & Forni, 2011; Neves et al., 2012; Moser et al., 2009). In the scenarios with the lowest number of QTL (2 to 16), greater differences were observed (Table 3), similar to those found for the results of predictive accuracy.

The DT and RF methods showed significantly lower RMSE values only compared to BO for the scenario with a heritability of .3 and two QTL (Table 3). However, when the number of QTL increased to four and eight with the same heritability of .3 (Figure 5a), these differences were not significant (Table 3).

In scenarios with a larger number of QTL, between 16 and 88 and heritability .3, it was possible to observe an inversion in the quality of the methods (Figure 5b). Radial basis function was the model with the best average performance for the RMSE parameter, although significant differences were predominant only for DT and BO (Table 3).

For the scenarios with a heritability of .6 and a specific QTL number (2 to 8; Figure 5c), the differences between the tree-based methods and the others were better highlighted. Decision tree obtained the lowest RMSE values, with significantly better results in relation to RBF, MLP, RR-BLUP, and BayesB (Table 3). Nevertheless, similar to what occurred in scenarios with less heritability, there was a reduction in the differences between the techniques with the increase in the number of QTL, and still there were some inversion trends in the RMSE results (Figure 5d).

Boosting was the model that showed the greatest sensitivity to heritability and showed a substantial improvement in results in higher heritability scenarios (Figure 5). Boosting is an approach repeatedly trained on the same sample so that at each iteration, a measure of forecast error is calculated for each individual, and in the next iteration, individuals with greater errors receive greater weight in training the model (Drucker, 1997). Thus, in analyses with low heritability, it is likely that explaining the data based on the traits of individuals with the greatest influence of residual variance is not a good strategy.

4 | CONCLUSIONS

The performance of statistical and machine learning methods is influenced by the complexity of the analyzed traits, and different results can be found due to this complexity. Machine learning methods are powerful alternative tools for predicting genetic values with epistatic gene control in traits with different degrees of heritability and different numbers of controlling genes. The DT, BA, RF, and BO methods showed superior results in the prediction of genetic values in traits with dominant and epistatic effects for scenarios of two to eight QTL. The RBF, MLP, BA, RF, and BO methods were equally efficient in predicting genetic values in traits with dominant and epistatic effects compared to traditional RR-BLUP and BayesB methods for scenarios of 16 to 88 QTL. Boosting was the model that showed the highest sensitivity in terms of selective accuracy, in relation to heritability, and presented lower results with low heritabilities, whereas DT showed greater sensitivity with increased in QTL numbers.

ACKNOWLEDGMENTS

The authors are grateful for the financial support of the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES and the Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq, and to the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) for researcher fellowship to ICS (2018/26408-0).

AUTHOR CONTRIBUTIONS

Ivan de Paiva de Paiva Barbosa: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Validation; Visualization; Writing-original draft; Writing-review & editing. **Michele Jorge da Silva**: Conceptualization; Writing-review & editing. **Weverton Gomes da Costa**: Conceptualization; Formal analysis; Visualization; Writing-review & editing. **Isabela de Castro Sant'Anna**: Conceptualization; Writing-review & editing. **Moysés Nascimento**: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Supervision; Validation; Writing-review & editing. **Cosme Damião Cruz**: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Software; Supervision; Validation; Visualization; Writing-original draft; Writing-review & editing.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest.

ORCID

Ivan de Paiva Barbosa https://orcid.org/0000-0001-8266-4414

Michele Jorge da Silva https://orcid.org/0000-0001-8648-8825
Weverton Gomes da Costa https://orcid.org/0000-0003-0742-5936
Moysés Nascimento https://orcid.org/0000-0001-5886-9540

REFERENCES

Akdemir, D., Jannink, J. L., & Isidro-Sánchez, J. (2017). Locally epistatic models for genome-wide prediction and association by importance sampling. *Genetics Selection Evolution*, *49*(1), 1–14. https://doi.org/10.1186/s12711-017-0348-8

Alkimim, E. R., Caixeta, E. T., Sousa, T. V., de Resende, M. D. V., da Silva, F. L., Sakiyama, N. S., & Zambolim, L. (2020). Selective efficiency of genome-wide selection in Coffea canephora breeding. *Tree Genetics and Genomes*, *16*(3). https://doi.org/10.1007/s11295-020-01433-3

Azevedo, C. F., de Resende, M. D. V., Silva, F. F., Viana, J. M. S., Valente, M. S. F., Resende, M. F. R., & Oliveira, E. J. (2016). New accuracy estimators for genomic selection with application in a cassava (Manihot esculenta) breeding program. *Genetics and Molecular Research*, *15*(4), 1–14. https://doi.org/10.4238/gmr.15048838

Azevedo, C. F., Silva, F. F., de Resende, M. D. V., Lopes, M. S., Duijvesteijn, N., Guimarães, S. E. F., Lopes, P. S., Kelly, M. J., Viana, J. M. S., & Knol, E. F. (2014). Supervised independent component analysis as an alternative method for genomic selection in pigs. *Journal of Animal Breeding and Genetics*, *131*(6), 452–461. https://doi.org/10.1111/jbg.12104

Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of K-fold cross-validation. *Journal of Machine Learning Research*, *5*(1533–7928), 1089–1105.

Coster, A., Bastiaansen, J. W. M., Calus, M. P. L., Van Arendonk, J. A. M., & Bovenhuis, H. (2010). Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. *Genetics Selection Evolution*, *42*(1), 1–11. https://doi.org/10.1186/1297-9686-42-9

Coutinho, A. E., Neder, D. G., Da Silva, M. C., Arcelino, E. C., De Brito, S. G., & Filho, J. L. S. D. C. (2018). Prediction of phenotypic and genotypic values by BLUP/GWS and neural networks. *Revista Caatinga*, *31*(3), 532–540. https://doi.org/10.1590/1983-21252018v31n301rc

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J., & Varshney, R. K. (2017). Genomic selection in plant breeding: Methods, models, and perspectives. *Trends in Plant Science*, *22*(11), 961–975. https://doi.org/10.1016/j.tplants.2017.08.011

Cruz, C. D. (2013). GENES - Software para análise de dados em estatística experimental e em genética quantitativa. *Acta Scientiarum - Agronomy*, *35*(3), 271–276. https://doi.org/10.4025/actasciagron.v35i3.21251

Cruz, C. D. (2016). Programa genes – Ampliado e integrado aos aplicativos R, Matlab e Selegen. *Acta Scientiarum - Agronomy*, *38*(4), 547–552. https://doi.org/10.4025/actasciagron.v38i4.32629

Ferreira, R. A. D. C., Silva, G. N., Glória, L. S., Sant'Anna, I. de C., Rodrigues, H. S., Silva, F. F. E, & Cruz, C. D. (2018). RNA – Aplicação em Estudos de Seleção Genômica Ampla. In C. D. Cruz & M. Nascimento (Eds.), *Inteligência computacional aplicada ao melhoramento genético* (1st ed., pp. 241–261). Editora UFV.

Demuth, H., Beale, M., & Hagan, M. (2000). *Neural network toolbox User's guide: For Use with MATLAB*. The Mathworks.

Drucker, H. (1997). Improving regressors using boosting techniques. *14th International Conference on Machine Learning*, *August 1997*, 107–115.

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome*, *4*, 250–255. https://doi.org/10.3835/plantgenome2011.08.0024

Ehret, A., Hochstuhl, D., Gianola, D., & Thaller, G. (2015). Application of neural networks with back-propagation to genome-enabled prediction of complex traits in Holstein-Friesian and German Fleckvieh cattle. *Genetics Selection Evolution*, *47*(1), 1–9. https://doi.org/10.1186/s12711-015-0097-5

Ghafouri-Kesbi, F., Rahimi-Mianji, G., Honarvar, M., & Nejati-Javaremi, A. (2017). Predictive ability of random forests, boosting, support vector machines and genomic best linear unbiased prediction in different scenarios of genomic evaluation. *Animal Production Science*, *57*(2), 229–236. https://doi.org/10.1071/AN15538

González-Camacho, J. M., de los Campos, G., Pérez, P., Gianola, D., Cairns, J. E., Mahuku, G., Babu, R., & Crossa, J. (2012). Genome-enabled prediction of genetic values using radial basis function neural networks. *Theoretical and Applied Genetics*, *125*(4), 759–771. https://doi.org/10.1007/s00122-012-1868-9

González-Recio, O., & Forni, S. (2011). Genome-wide prediction of discrete traits using bayesian regressions and machine learning. *Genetics Selection Evolution*, *43*(1), 1–12. https://doi.org/10.1186/1297-9686-43-7

Guo, Z., Tucker, D. M., Basten, C. J., Gandhi, H., Ersoz, E., Guo, B., Xu, Z., Wang, D., & Gay, G. (2014). The impact of population structure on genomic prediction in stratified populations. *Theoretical and Applied Genetics*, *127*(3), 749–762. https://doi.org/10.1007/s00122-013-2255-x

James, G., Witten, D., Hastie, T., & Tibishirani, R. (2013). An introduction to statistical learning with applications in R (older version). *Springer Texts in Statistics*, *426*. https://doi.org/10.1007/978-1-4614-7138-7

Matlab. (2019). *Natick*. Massachusetts: The MathWorks.

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829.

Moser, G., Tier, B., Crump, R., Khatkar, M., & Raadsma, H. (2009). A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genetics Selection Evolution*, 41(1). https://doi.org/10.1186/1297-9686-41-56

Moura, E. G., Pamplona, A. K. A., & Balestre, M. (2019). Functional models in genome-wide selection. *PLOS ONE*, *14*(10), 1–27. https://doi.org/10.1371/journal.pone.0222699

Neves, H. H. R., Carvalheiro, R., & Queiroz, S. A. (2012). A comparison of statistical methods for genomic selection in a mice population. *BMC Genetics*, *13*. https://doi.org/10.1186/1471-2156-13-100

R Core Team. (2019). *R: A language and environment for statistical computing. 3*.

de Resende, M. D. V., Júnior, M. F. R. R., Aguiar, A. M., Abad, J. I. M., Missiaggia, A. A., Sansaloni, C., Petroli, C., & Grattapaglia, D. (2010). Computação da Seleção Genômica Ampla (GWS). *Série Documentos Da EMBRAPA Florestas*, *209*, 78.

Resende, J. F. R., Muñoz, P., de Resende, M. D. V., Garrick, D. J., Fernando, R. L., Davis, J. M., Jokela, E. J., Martin, T. A., Peter, G. F., & Kirst, M. (2012). Accuracy of genomic selection methods in a standard data set of loblolly pine (Pinus taeda L.). *Genetics*, *190*(4), 1503–1510. https://doi.org/10.1534/genetics.111.137026

Sant'Anna, I. de C., Nascimento, M., Silva, G. N., Cruz, C. D., Azevedo, C. F., Gloria, L. S., & Silva, F. F. e. (2020). Genome-enabled prediction of genetic values for using radial basis function neural networks. *Functional Plant Breeding Journal [Internet]*, *1*(2595–9433), 1–8. https://doi.org/10.35418/2526-4117/v1n2a1

Searchinger, T., Waite, R., Hanson, C., & Ranganathan, J. (2019). Creating a Sustainable Food Future: A Menu of Solutions to Feed Nearly 10 Billion People by 2050. *World Resources Report*, *1*(July).

Silva, G. N., Tomaz, R. S., Sant'Anna, I. de C., Nascimento, M., Bhering, L. L., & Cruz, C. D. (2014). Neural networks for predicting breeding values and genetic gains. *Scientia Agricola*, *71*(6), 494–498. https://doi.org/10.1590/0103-9016-2014-0057

de Sousa, I. C., Nascimento, M., Silva, G. N., Nascimento, A. C. C., Cruz, C. D., Silva, F. F. e, de Almeida, D. P., Pestana, K. N., Azevedo, C. F., Zambolim, L., & Caixeta, E. T. (2021). Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. *Scientia Agricola*, 78(4). https://doi.org/10.1590/1678-992x-2020-0021

Sousa, T. V., Caixeta, E. T., Alkimim, E. R., Oliveira, A. C. B., Pereira, A. A., Sakiyama, N. S., Zambolim, L., & de Resende, M. D. V. (2019). Early selection enabled by the implementation of genomic selection in coffea arabica breeding. *Frontiers in Plant Science*, *9*(January), 1–12. https://doi.org/10.3389/fpls.2018.01934

How to cite this article: Barbosa IdP, da Silva MJ, da Costa WG, de Castro Sant'Anna, I, Nascimento, M, Cruz CD. Genome-enabled prediction through machine learning methods considering different levels of trait complexity. *Crop Science*. 2021;1-13. https://doi.org/10.1002/csc2.20488

5. ARTIGO 2

Adequabilidade de modelos para a seleção genômica de acordo com a complexidade da característica

Resumo

O objetivo deste trabalho foi comparar o desempenho seletivo e preditivo do RR-BLUP (*Random Regression Best Linear Unbiased Predictor*) e métodos de inteligência computacional, MLP (*Multilayer Perceptron*) e RBF (*Radial Basis Function*), e aprendizado de máquina, árvore de regressão com refinamentos BO (*Boosting*), BA (*Bagging*) e RF (*Random Forest*), em populações simuladas para diferentes cenários em relação ao número de QTL, à distribuição dos QTL nos grupos de ligação (GL), ao grau médio de dominância, à herdabilidade e aos diferentes modelos de expressão aditivo e epistaticos. O genoma simulado compreendeu 2010 SNP (S*ingle Nucleotide Polymorphism*) e distribuídos equitativamente em 10 grupos de ligação. Um total de 108 características foram simuladas considerando diferentes valores para os cinco cenários. Os métodos RBF, RF e *Bagging*, de uma maneira, geral apresentaram bons e consistentes resultados de acurácia seletiva (R^2) e acurácia preditiva (REQM). Esses métodos apresentaram resultados iguais ou superiores à média geral dos métodos para todos os cenários avaliados. O aumento no número de QTL, até 88, afetou positivamente os resultados de R^2 e REQM. A presença de efeito de dominância e a baixa herdabilidade provocou impactos negativos para os resultados de R^2 e REQM. Todos os métodos apresentaram uma redução nos valores de R^2 e aumento na REQM quando efeitos não aditivos eram importantes para a característica. Desta forma nenhum método foi excepcionalmente eficiente para a captura deste efeito, em média para os cenários avaliados. A distribuição dos QTL nos GL foi o único cenário que afetou os resultados de R^2 e REQM de modo variado. Neste cenário os melhores resultados de R^2, foram observados obtidos quando os QTL estavam distribuídos em apenas um GL, enquanto que para os resultados de REQM os melhores resultados foram obtidos quando os QTL foram distribuídos em oito GL. Esses resultados mostram que a distribuição dos QTL no genoma pode ser o principal atributo a ser

avaliado quanto ao interesse na seleção ou na predição dos valores genéticos pelo melhorista.

Palavras-chave: Epistasia, inteligência computacional, GWS, redes neurais, árvore de regressão

Introdução

No melhoramento há a busca pela incorporação de novas formas alélicas ou pelo aumento da frequência de alelos favoráveis nas populações com potencial para cultivo de modo a atender às necessidades direta, ou indireta, do produtor, do consumidor e da indústria. De acordo com a *World Resources Institute* (SEARCHINGER *et al*., 2019), a agricultura deve ser capaz de fornecer alimento em quantidade e qualidade, o suficiente para atender ao volume crescente da população que poderá atingir quase 10 bilhões de pessoas até 2050.

Para acompanhar o aumento estimado da demanda por alimentos nos próximos anos, os programas de melhoramento genético de plantas devem ser capazes de aproveitar as maiores taxas de ganho genético possíveis para maximizar sua contribuição no aumento da produtividade agrícola. Nesse contexto, aproveitar o potencial de novas metodologias, se torna uma etapa essencial. Proposta por Meuwissen *et al*. (2001), a seleção genômica ampla (*Genome Wide Selection* - GWS) tornou-se uma metodologia amplamente utilizada em programas de em melhoramento.

A GWS foi proposta para estimar o valor genético genômico (VGG) de indivíduos que ainda não foram fenotipados por meio de informações de marcadores distribuídos em todo o genoma. A maior motivação para a utilização dessa técnica consiste na possibilidade de utilizar genotipagem em grande escala e incorporar informações genômicas no processo de predição, de modo a aumentar a eficiência seletiva, obter ganhos genéticos de forma mais ágil, além da possibilidade de reduzir os custos.

Muitos estudos mostraram que a GWS pode aumentar o ganho de seleção por ano (ALKIMIM *et al*., 2020; BHATTA *et al*., 2020; DEOMANO *et al*., 2020; MASTRODOMENICO *et al*., 2019; MÔRO *et al*., 2019; SMALLWOOD *et al*., 2019). Para isso, modelos estatísticos podem ser usados para auxiliar na seleção de

genótipos superiores e ainda na predição de valores genotípicos de características de interesse. No entanto, os variados modelos estatísticos, apresentam diferentes filosofias e, consequentemente, podem apresentar diferentes resultados. De modo geral, essas metodologias se divergem uma da outra pela forma para lidar com os problemas relacionados à multicolinearidade e dimensionalidade, e ainda a distribuição das variáveis em si, dependendo muitas vezes da pressuposição previa do modelo, de forma que permita a captura dos efeitos complexos, comumente presentes nos dados biológicos, tais como a presença de efeitos de dominância e da epistasia. Nestes casos, é apropriado empregar metodologias que levam em consideração a presença de tais efeitos (FERREIRA *et al.*, 2018). Além disso, a herdabilidade da característica pode afetar drasticamente a acurácia do modelo.

As evidências de efeitos epistáticos relevantes no controle das características das plantas são vastas (DERBYSHIRE *et al.*, 2021; HOLLAND 2001). Por exemplo, Dudley (2008) encontrou a presença de epistasia no óleo, proteína e amido em diferentes cruzamentos de linhagens de milho, e Dudley e Johnson (2010) relataram que adicionar duas interações de locus ao modelo aumenta o poder de previsão. Derbyshire *et al.* (2021) destacam que a inclusão de efeitos genéticos não aditivos tem o potencial de melhorar a precisão da predição para resistência à podridão do caule da esclerotínia em canola, e que esses efeitos devem ser considerados em qualquer aplicação de predição genômica para cultivo de canola para esta característica. Apesar disso, há uma falta de métodos bem estabelecidos para incorporar epistasia na previsão de características complexas em programas de melhoramento de plantas (BERNARDO 2002).

Nos últimos anos, os métodos de aprendizado de máquina (*Machine Learning* – ML) ou de inteligência computacional (IC) têm sido considerados no contexto da previsão genômica. Os métodos ML/IC são modelos não paramétricos que fornecem flexibilidade para se adaptar a associações complicadas entre dados e saída com a capacidade de se adaptar a padrões muito complexos. Os métodos IC fundamentam-se em regras de aprendizagem e mimetizam os princípios do cérebro humano em seu processamento. Métodos ML fazem uso de buscas exaustivas ou heurísticas fundamentadas em amostragens para busca de soluções. O uso de modelos baseados na inteligência computacional e aprendizado de máquina vêm sendo aplicados com sucesso na solução de diversos problemas ligados à genética

(BARBOSA, *et al*., 2021; SOUSA *et al*., 2021; TOMAZ *et al*., 2018). Estas metodologias se diferem das modelagens estocásticas, por não possuírem pressuposições quanto ao modelo, uma vez que seus resultados dependem do aprendizado e não da distribuição das variáveis em si (FERREIRA *et al*., 2018). Assim como as redes neurais artificiais (RNAs), as árvores de regressão e seus refinamentos não requerem suposições sobre o modelo. Além disso, as árvores de regressão permitem a não linearidade dos dados e uma interpretação fácil, pois fornece informações sobre quais atributos são mais importantes para previsão.

Desse modo, o presente trabalho tem por objetivo aplicar métodos de aprendizado de máquinas (ML) e de inteligência computacional (IC) na seleção genômica para avaliar seu potencial em termos de desempenho de predição em comparação com modelos convencionais de predição genômica envolvendo características alta complexibilidade.

Material e métodos

Simulação dos dados

Todas as simulações de dados foram realizadas utilizando o software Genes (CRUZ, 2013). Uma população F2 com tamanho efetivo de 1.000 indivíduos foi simulada. Para isso, um genoma semelhante ao de uma espécie diploide (2n = 2x = 20 cromossomos) de tamanho de 100 centiMorgans (cM) para cada grupo de ligação e compreendeu 2.010 polimorfismos bialélicos de nucleotídeo único (SNPs), codificados com 1 se AA, 0 se Aa e -1 se aa, distribuídos equitativamente entre os grupos de ligação e espaçados de forma equidistante. Para estabelecimento de cada indivíduo F2, um pool de 5000 gametas de cada genitor foi considerado permitindo estabelecer combinações gaméticas aleatórias a partir de ancestrais homozigotos contrastantes (P1 dominante e P2 recessivo). Para a geração de gametas, a porcentagem de recombinação equivalente à distância entre os locos foi de 0,5 cM, proporcionando desequilíbrio de ligação. Apesar de serem dados simulados, é possível que ocorra presença de SNPs com distorção de segregação mendeliana na população F2. Desta forma, para garantir a interpretação adequada dos resultados, foi realizado o teste de segregação pelo qui-quadrado (χ^2) a 5% de probabilidade com correção de Bonferroni para testes múltiplos.

Simulação das características

Foram simuladas 108 características fenotípicas pela combinação de diferentes cenários de herdabilidade, número de QTL, posição dos QTL no genoma, grau médio de dominância e diferentes modelos de interação gênica (Tabela 1). A herdabilidade das características foram de 0,3 ou 0,6; o número de QTL foi de 8, 48 ou 88, distribuídos igualmente e equidistantes dentro do primeiro GL ou distribuídos igualmente e equidistantes dentro dos 8 primeiro GL; o grau médio de dominância foi d=0, d=0,5 ou d=1 e os modelos para distribuição dos efeitos foram o modelo aditivo (Equação 1), epistático (Equação 2) e epistático (Equação 3).

$$\mathrm{Yi} = \mu + \sum_{j=1}^{n} p_j \alpha_j + e_i \qquad \text{(Equação 1)}$$

$$\mathrm{Yi} = \mu + \sum_{j=1}^{n} p_j \alpha_j + \sum_{j=1}^{n} \sum_{j'=1}^{n} p_j \alpha_j \alpha_{j'} + e_i \qquad \text{(Equação 2)}$$

$$\mathrm{Yi} = \mu + \sum_{j=1}^{n} p_j \alpha_j + \sum_{j=1}^{n} p_j \alpha_j \alpha_{j+1} + e_i \qquad \text{(Equação 3)}$$

em que Yi é o valor fenotípico para observação i; μ é uma média geral; p_j é a contribuição do loco j para a manifestação da característica estabelecida por uma distribuição uniforme de modo que cada loco j teve a mesma contribuição para a característica sob consideração; $\alpha_j, \alpha_{j'}\ e\ \alpha_{j+1}$ assumem os valores $u + a_j$, $u + d_j$ e $u - a_j$ para os genótipos associados às classes AA, Aa e aa, respectivamente, com u sendo a média dos homozigotos, a_j é a metade da diferença do valor genotípico entre ambos os homozigotos, e d_j é a diferença entre o valor genotípico do heterozigoto e a média dos homozigotos. A codificação 1, 0 ou -1 identificou as classes AA, Aa e aa, respectivamente. Nas equações 1, 2 e 3, o primeiro somatório da expressão refere-se à contribuição do loco individual por meio de seu efeito aditivo e de dominância. O segundo somatório da equação 2 representa os efeitos multiplicativos correspondentes às interações epistáticas entre pares de locos, o termo $\alpha_j \alpha_{j'}$ é o efeito multiplicativo do alelo favorável no loco j e j'. O segundo somatório da equação 3 representa os efeitos multiplicativos correspondentes às interações epistáticas entre pares de locos, o termo $\alpha_j \alpha_{j+1}$ é o efeito multiplicativo do alelo favorável no loco j e $j + 1$. O e_i é o efeito ambiental, gerado de acordo com a estrutura de variância dos resíduos foi dada por $e_i \sim \mathrm{N}(0, V_e)$, com $V_e = \left[V_g(1 - h^2)\right]/h^2$.

No cenário de 8 QTL, esses foram alocados ao centro dos oito primeiros grupos de ligação, dessa forma, nenhum deles estavam ligados. Já para os cenários de 48 ou 88, foram adicionados mais 5 ou 10 QTL por grupo de ligação, respectivamente. Desta forma, as diferentes configurações simuladas em relação ao número de QTL, possibilitariam também a comparar para os efeitos de QTL não ligados, e de QTL ligados.

Tabela 1 - Características fenotípicas simuladas pela combinação de diferentes cenários de herdabilidade, número de QTL, posição dos QTL no genoma, grau médio de dominância e diferentes modelos de interação gênica (T1 a T108)

GMD[1]	h²	Modelo	Número de locos controladores da característica (QTL)					
			8		48		88	
			NGC*		NGC		NGC	
			1	8	1	8	1	8
0	0,3	1	T1	T19	T37	T55	T73	T91
		2	T2	T20	T38	T56	T74	T92
		3	T3	T21	T39	T57	T75	T93
	0,6	1	T4	T22	T40	T58	T76	T94
		2	T5	T23	T41	T59	T77	T95
		3	T6	T24	T42	T60	T78	T96
0.5	0,3	1	T7	T25	T43	T61	T79	T97
		2	T8	T26	T44	T62	T80	T98
		3	T9	T27	T45	T63	T81	T99
	0,6	1	T10	T28	T46	T64	T82	T100
		2	T11	T29	T47	T65	T83	T101
		3	T12	T30	T48	T66	T84	T102
1	0,3	1	T13	T31	T49	T67	T85	T103
		2	T14	T32	T50	T68	T86	T104
		3	T15	T33	T51	T69	T87	T105
	0,6	1	T16	T34	T52	T70	T88	T106
		2	T17	T35	T53	T71	T89	T107
		3	T18	T36	T54	T72	T90	T108

*NGC - Número de grupos de ligação com QTL controladores da característica
[1] GMD – Grau médio de dominância
Fonte: O autor.

Predição de valores genéticos

A partir dos dados simulados, os efeitos dos marcadores foram estimados, e os valores genômicos (EGBV – *Estimated Genomic Breeding Values*) foram preditos usando os métodos de redes neurais MLP e RBF e de aprendizado de máquina (árvore de regressão com refinamentos BO (*Boosting*), BA (*Bagging*) e RF (*Random Forest*). Os resultados encontrados foram comparados aos obtido pelo método RR-BLUP (*Random Regression Best Linear Unbiased Predictor*).

RR-BLUP

A metodologia RR-BLUP foi utilizada conforme descrição de Meuwissen *et al*. (2001), de acordo com o seguinte modelo:

$$y = Wb + Xm + e$$

em que: *y* é o vetor de observações fenotípicas; *b* é o vetor de efeitos fixos (média geral) com matriz de incidência W (com todas as entradas iguais a 1); *m* é o vetor dos efeitos aleatórios dos marcadores com matriz de incidência X com $m \sim N(0, I\sigma_m^2)$ sendo σ_m^2 a variância do marcador e e refere-se ao vetor de erros aleatórios com $e \sim N(0, I\sigma_e^2)$, sendo σ_e^2 a variância do erro. X é a matriz de incidência composta pelos valores 1, 0 e -1 de acordo com o número de alelos do marcador dos genótipos MM, Mm e mm, respectivamente.

As equações de predição foram modeladas assumindo que todos os locos explicaram quantidades equivalentes da variação genética e, portanto, apresentaram σ_m^2 comum. Assim, a variação genética explicada por cada loco é dada por (σ_g^2/n_Q), em que σ_g^2 é a variação genética total e n_Q é o número de locos controladores da característica e que pode ser inferido por $n_Q = \sum_{i=1}^{n} 2p_i(1-p_i)$ sendo *p* a frequência alélica do marcador.

Rede *Perceptron* Multicamadas

Para a Rede *Perceptron* Multicamadas (MLP) foi utilizado o algoritmo de treinamento *backpropagation*, com regularização bayesiana, que é uma modificação do algoritmo de treinamento Levenberg-Marquardt para produzir redes que generalizem bem e reduz a dificuldade de determinar a arquitetura de rede ideal (DEMUTH e BEALE, 2000). O algoritmo *backpropagation*, utilizado no treinamento das redes MLP consiste em duas fases: o passo onde a rede é alimentada para frente (*forward*) e o passo onde a rede é alimentada para trás (*backward*). Na etapa *forward*, os pesos sinápticos w(p) permanecem inalterados e os sinais funcionais da rede neural são calculados para cada neurônio até que seja produzida a saída desejada na camada de saída (essa etapa também é conhecida como fase de propagação). A etapa *backward*, por sua vez, se inicia na camada de saída da rede, passando os sinais de erro para as camadas anteriores, de modo que os pesos sinápticos sejam

recalculados de acordo com a regra Delta (Equação 4) até que se retorne à primeira camada oculta da rede (etapa também conhecida como fase de atualização de pesos ou retro-propagação).

$$\Delta w_{(t)} = \alpha \Delta w_{(t-1)} + \eta \delta_{(t)} y_{(t)} \qquad \text{Equação (4)}$$

onde

α é a constante de *momentum* com 0 < α < 1;

δ, é o gradiente local;

η, taxa de aprendizagem;

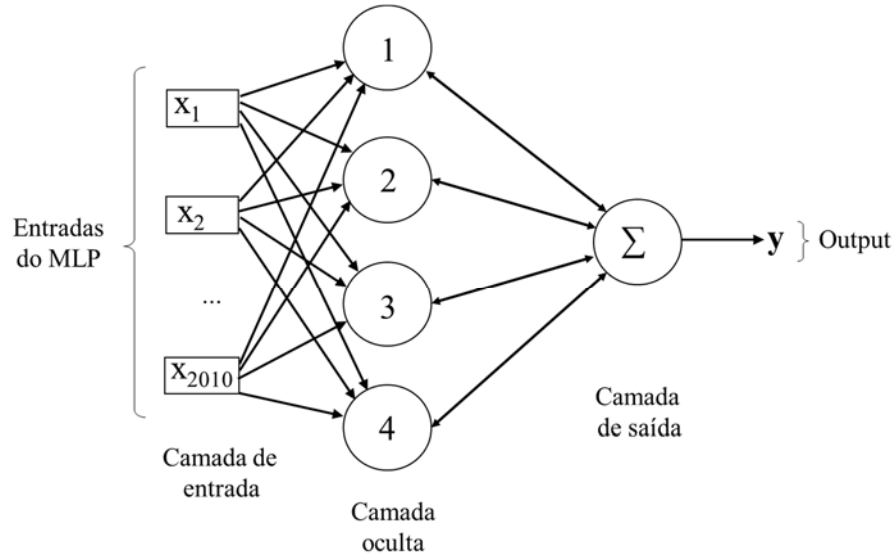
y, a saída da rede;

$\Delta w_{(t)}$ é o erro obtido pela rede neural na iteração t; e

$\Delta w_{(t-1)}$ é o erro obtido pela rede neural na iteração anterior (t − 1).

Foram realizados testes preliminares para definição da arquitetura da rede (dados não apresentados). A arquitetura que apresentou melhores resultados para o conjunto de dados neste estudo, envolveu uma única camada oculta e número de neurônios (n) variando de um a quatro (n = 1, ... ,4). Na camada de entrada, foi fornecida a matriz de marcadores moleculares com 2010 marcadores, para que a camada de saída retornasse ao valor fenotípico predito de cada indivíduo. Na Figura 1 é apresentado o esquema da rede MLP utilizado. A função de ativação utilizada foi a tangente hiperbólica, definida pela equação a seguir: $tanh\, x = \frac{1-e^{-2x}}{1+e^{-2x}}$.

Figura 1 - Esquema da rede MLP *Backpropagation*. Entradas X_1 a X_{2010} na camada de entrada se referem aos 2010 marcadores considerados nas análises. Com uma camada oculta e número de neurônios variando de um a quatro. Na saída a rede retorna ao vetor de valores preditos (y)



Fonte: O autor.

Rede de Função de Base Radial

A Rede de Função de Base Radial (RBF) foi utilizada com uma arquitetura do tipo *feedforward*. Essa metodologia, consiste de uma camada de entrada, uma camada oculta e a camada de saída. A camada de entrada conecta à rede ao ambiente (agrupa os dados de entrada em *clusters*). A camada oculta, aplica uma transformação não linear do espaço de entrada para um espaço oculto de alta dimensionalidade (geralmente utilizadas funções de ativação de base radial gaussianas). Por fim, a camada de saída aplica uma transformação no espaço oculto fornecendo uma saída para a rede.

O número de neurônios na camada oculta variou de 10 a 100 neurônios e o tamanho do raio variou na mesma amplitude, de 10 a 100. De modo similar à MLP, na camada de entrada foi utilizada a matriz de marcadores moleculares com 2010 marcadores, de modo que a saída desejada foram os valores fenotípicos simulados. A função de ativação utilizada na camada oculta foi a gaussiana, conforme a equação a seguir:

$$g(u) = e^{\frac{-(u-c)^2}{2\sigma^2}}$$

em que: *c* é o centro da função gaussiana; σ^2 é a variância da função gaussiana e *u* é o potencial de ativação.

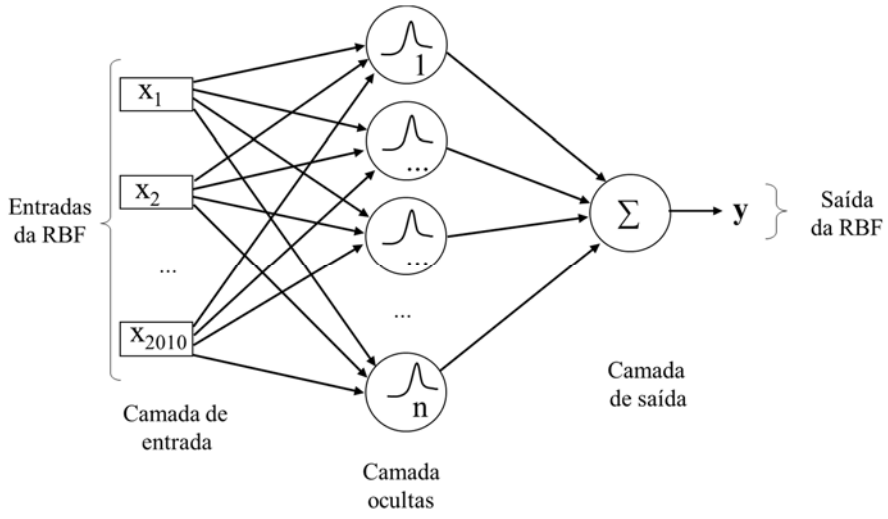
E na camada de saída, foi utilizada a função de ativação do tipo linear, conforme a equação a seguir:

$$y_{ri} = g\left(x_0 w_0 + \sum_{j=1}^{q} f_{xj}(x_i) w_j\right)$$

em que: x_i: i-ésima entrada; w_j: j-ésimo peso sináptico; f_{xj}: função de ativação da camada oculta associada à entrada x_i ponderada por seu respectivo peso.

Na Figura 2 é apresentado o esquema da Rede de Função de Base Radial (RBF) utilizada.

Figura 2 - Esquema de uma rede RBF *feedforward*. Entradas X_1 a X_{2010} na camada de entrada se referem aos 2010 marcadores considerados nas análises. Uma camada oculta com raio variando de 10 a 100 (r = 10, ..., 100) e de neurônios (n) variando de 10 a 100 (n = 10, ... ,100). Na saída a rede retorna o vetor de valores preditos (y)



Fonte: O autor.

Árvore de Regressão e seus refinamentos (*Boosting*, *Bagging* e *Random Forest)*

A árvore de regressão tem como objetivo subdividir diversas vezes o conjunto de observações de tal forma que os subgrupos formados subsequentes sejam cada vez mais homogêneos (BREIMAN *et al.*, 1984). A estrutura da árvore de regressão foi

feita pela busca da árvore que levasse a partição dos dados até a formação de grupos homogêneos. Para isso, avaliou-se o quão razoável foi uma dada árvore *T* através de seu erro quadrático médio, como na equação abaixo:

$$P(T) = \sum_{R} \sum_{k \in R} (y_k - \hat{y}_R)^2$$

em que: $\hat{y}_R$ é o valor predito para a resposta fenotípica da característica e y_k é o valor verdadeiro da característica de cada indivíduo dentro do grupo.

Em uma segunda etapa, foi realizado a poda com o objetivo de tornar a árvore de regressão menor e menos complexa, de modo a diminuir a variância deste estimador. Nessa etapa do processo, cada nó foi retirado, um por vez, observando-se como o erro de predição variou no conjunto de validação. Posteriormente, baseando-se nas observações, foi decidido quais nós permanecem na árvore.

Geralmente, uma única árvore não possui boa precisão preditiva quando comparada com outras abordagens (ROGAN *et al*., 2008). Alguns refinamentos com o intuito de melhorar a performance do modelo de árvore de regressão são apresentados na literatura e apresentam desempenhos superiores (BALTA e TOPAL, 2020; SOUSA *et al*., 2021). Dessa forma, também foi testado a performance preditiva dos modelos *Bagging*, *Random Forest* (RF) e *Boosting*.

Um dos problemas apresentados pela árvore de regressão é a grande variabilidade entre os resultados obtidos. Para contornar esse problema o *Bagging* é um método que aplica a técnica de *bootstrap*, em outras palavras, obtém-se B amostras do conjunto de observações, com reposição, adquirindo assim um número de B modelos $\hat{f}^1(x), \hat{f}^2(x), \ldots \hat{f}^B(x)$ (BREIMAN, 1996). A média aritmética desses modelos irá ser o modelo final.

O RF segue a mesma ideia do *Bagging*, no entanto, além do conjunto de observações, altera também o número de variáveis preditoras ($m = \sqrt{p}$) utilizadas em cada partição. O modelo RF funciona das seguintes maneiras: (1) ele produz subfases dos dados anteriores usando a ferramenta de reamostragem de *bootstrap*, que é equivalente a tamanhos zero no conjunto de dados anterior, (2) ele gera árvores de decisão aplicando as subfases, e (3) em última análise, ele produz a saída ao fundir os resultados da previsão de todas as árvores de decisão (CHEN *et al*., 2019).

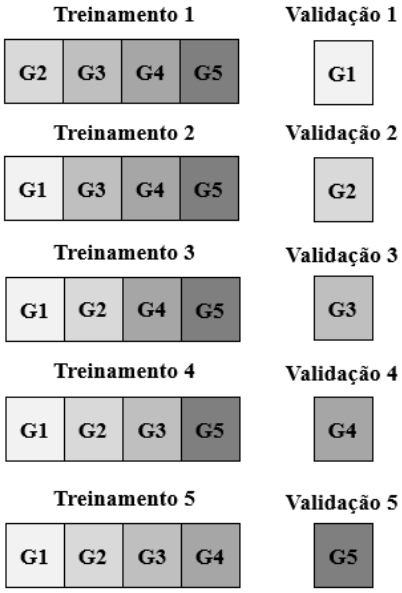
Já o *Boosting* cria árvores sequencialmente utilizando informações das árvores anteriores, ao contrário do *Bagging* que cria múltiplas árvores independentes, e é realizado com um processamento sequencial. O *Boosting* é empregado combinando modelos preditores fracos para produzir melhor precisão preditiva. Os dados incorretos da previsão anterior são classificados como dados "difíceis" e serão usados para o próximo processo de previsão para que o valor de precisão alcance um ponto máximo. Depois que todo o processo de previsão é realizado, todos os modelos são mesclados. O impulso transforma um modelo de preditor fraco em um preditor complexo confiável. As etapas deste processo de aprendizagem são a previsão para regressão, cálculo de erros do resíduo e processo de aprendizagem para processar o resíduo (SYAHRANI, 2019).

Treinamento e validação

Para o treinamento e validação das técnicas utilizadas, foi realizada validação cruzada (*k-fold*) com k = 5 partições (Figura 3). A seleção dos indivíduos foi realizada de forma aleatória para compor os k conjuntos na validação cruzada *k-fold*. Em cada uma das cinco rodadas, quatro desses subconjuntos constituíram a população de treinamento (80% dos indivíduos) e o subconjunto restante constituiu a população de validação (20% dos indivíduos). As técnicas foram comparadas com base na média aritmética e no erro padrão médio das estimativas de desempenho dos conjuntos de validação. Os grupos de treinamento e validação foram os mesmos para todos os métodos avaliados, a fim de evitar a influência da variação de grupos aleatórios nos resultados entre um método e outro.

Figura 3 - Processo de validação cruzada *K-fold*, amostras de treinamento e validação. O número dentro da caixa representa a proporção de indivíduos da população no grupo de teste

	Ind.	Trait1	...	Trait108	Mrk1	...	Mrk2010
G1	1	X_1	...	Y_1	-1	...	0
	...	...	...	...	...	...	...
	200	X_{200}	...	Y_{200}	0	...	1
G2	201	X_{201}	...	Y_{201}	1	...	-1
	...	...	...	...	...	...	...
	400	X_{400}	...	Y_{400}	1	...	-1
G3	401	X_{401}	...	Y_{401}	1	...	-1
	...	...	...	...	...	...	...
	600	X_{600}	...	Y_{600}	-1	...	0
G4	601	X_{601}	...	Y_{601}	-1	...	1
	...	...	...	...	...	...	...
	800	X_{800}	...	Y_{800}	0	...	1
G5	801	X_{801}	...	Y_{801}	1	...	-1
	...	...	...	...	...	...	...
	1000	X_{1000}	...	Y_{1000}	1	...	0



Fonte: O autor.

Comparação da eficiência das metodologias

Para avaliar a eficiência das metodologias foram utilizados os parâmetros de raiz do erro quadrático médio de validação ($REQM$) e o quadrado da correlação de validação (R^2).

A raiz do erro quadrático médio é adotada para expressar a acurácia preditiva dos modelos, pois apresenta a vantagem de apresentar os valores do erro na mesma escala da variável de interesse, e é descrita conforme a seguir:

$$REQM = \sqrt{\frac{\sum(\hat{y}-y)^2}{n}}$$

A acurácia seletiva é medida pelo quadrado da correlação entre os valores estimados ($\hat{y}$) e os valores verdadeiros (y), ou seja, mede o quanto a estimativa obtida é relacionada com o valor real do parâmetro, que em genética quantitativa, expressa a herdabilidade da característica (RESENDE *et al.*, 2010). A acurácia foi dada pela seguinte equação:

$$R^2 = \left(cor(\hat{y}, y)\right)^2$$

A comparação dos resultados para cada cenário, foi realizada a partir da média de todos os resultados que estavam contidos na avaliação de cada nível deste efeito,

desconsiderando qualquer efeito de interação dos resultados de R^2 e REQM entre os diferentes cenários.

Aspectos computacionais

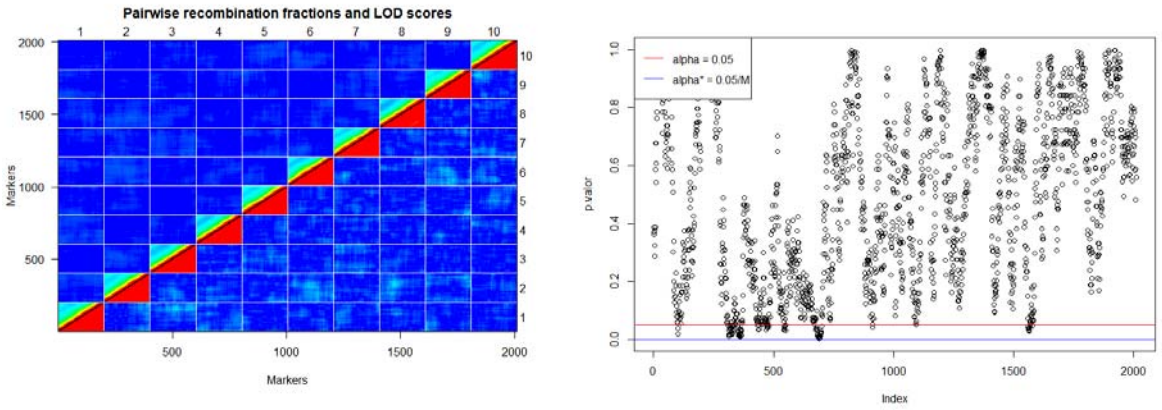
A simulação da população foi realizada por meio do software GENES (CRUZ, 2013). As metodologias de RR-BLUP, árvore de regressão, *Boosting*, *Bagging* e *Random Forest* foram realizadas com auxílio do software Genes em integração com o software R (R CORE TEAM, 2019; CRUZ, 2016). Para realizar as análises baseadas nas metodologias MLP e RBF, foi utilizado o *software* Genes em integração com o software *Matlab* (MATLAB, 2010; CRUZ, 2016).

Resultados

A taxa de recombinação entre locos de 0,5 cM foi suficiente para criar desequilíbrio de ligação e a aleatorização no *pool* gamético foi adequada em garantir a presença de locos polimórficos no final da simulação (Figura 4A). Nenhum SNP mostrou distorções significativas para a segregação Mendeliana esperada para a população simulada, sugerindo que os dados simulados para os testes propostos são adequados e, portanto, nenhum marcador foi removido (Figura 4B).

Figura 4 - *Heatmap* (A) apresenta as estimativas das frequências de recombinação (r) (acima da diagonal) e LOD (abaixo da diagonal) para os marcadores, entre e dentro dos 10 grupos de ligação simulados. O mapa de calor apresenta uma escada de cores azuis para marcadores não ligados (ou seja, r≈0,5 e baixo LOD), e de vermelho, com cores mais quentes para os marcadores ligados (ou seja, r <0,5 e alto LOD). À direita são apresentados os resultados de p-valor para o teste de segregação a nível individual e o limiar ao nível de significância de 0,5 (em vermelho), e ao nível de significância de 0,5/M (em azul). Sendo M = 2010 (número de marcadores)

A. B.



Fonte: O autor.

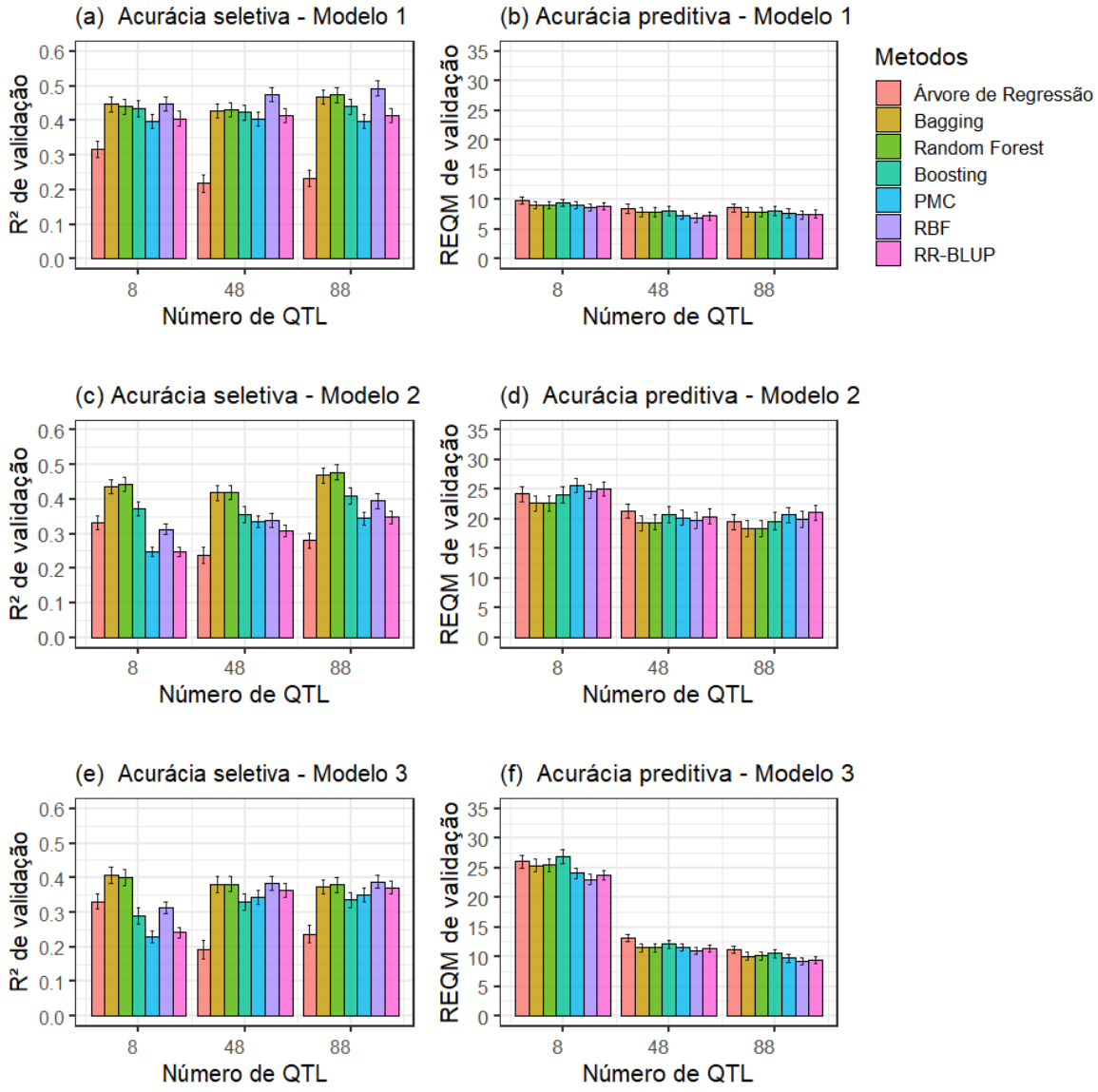
Os resultados de acurácia seletiva (R^2) e acurácia preditiva (REQM) obtidos para os cenários com diferentes números de QTL de acordo com o modelo de simulação estão apresentados na Figura 5.

Para o modelo 1, o método RBF apresentou resultados de R^2 superiores aos observados para o RR-BLUP para os cenários com 8, 48 ou 88 QTL (Figura 5a). Já o BA e o RF foram superiores ao RR-BLUP apenas no cenário com 88 QTL. Para este modelo, a árvore de regressão apresentou resultados de R^2 inferiores aos demais métodos nos três cenários. Para os resultados de REQM, todos os métodos apresentaram resultados semelhantes nos diferentes cenários com 8, 48 e 88 QTL (Figura 5b).

Para o modelo 2 os métodos BA e RF apresentaram resultados superiores aos demais nos cenários com 8, 48 ou 88 QTL (Figura 5c). O PMC foi o único método que não superou os resultados de R^2 observados pelo RR-BLUP para o cenário com 8 QTL. Nos cenários com 48 e 88 QTL a árvore de regressão apresentou resultados de R^2 inferiores aos demais (Figura 5c). Os resultados de REQM foram semelhantes entre os métodos dentro dos cenários de 8, 48 e 88 QTL, no entanto, todos os métodos apresentaram piores resultados no cenário de 8 QTL (Figura 5d).

Já para o modelo 3, a superioridade dos métodos BA e RF só foram observadas no cenário com 8 QTL (Figura 5e). Neste modelo, o PMC ainda foi o único método que não superou os resultados de R^2 observados para o RR-BLUP no cenário com 8 QTL. A árvore de regressão apresentou resultados de R^2 superiores aos observados para o RR-BLUP no cenário com 8 QTL, porém, inferior a todos os métodos nos cenários de 48 e 88 QTL, de forma semelhante ao observado para o modelo 2. Ainda no modelo 3, todos os métodos apresentaram baixos valores de REQM nos cenários de 48 e 88 QTL, com valores próximos aos observados para o modelo 1 (Figuras 5b e 5f). Já no cenário com 8 QTL, os métodos apresentaram valores de REQM semelhantes aos observados para o modelo 2 (Figuras 5d e 5f). Nesse último cenário, o RBF apresentou resultados de REQM inferiores aos observados pelos métodos de árvore de regressão e seus refinamentos BA, RF e BO (Figura 5f).

Figura 5 – Acurácia seletiva (R^2) e acurácia preditiva (REQM) [erros padrão] obtidas com os métodos de árvore de regressão com refinamentos BO (*Boosting*), BA (*Bagging*) e RF (*Random Forest*), de redes neurais MLP (*Multilayer Perceptron*) e RBF (*Radial Basis Function*) e RR-BLUP (*Random Regression Best Linear Unbiased Predictor*), para características simuladas nos cenários com diferentes números de QTL e de acordo com o modelo de simulação aditivo (modelo 1) e epistáticos (modelos 2 e 3)



Fonte: O autor.

Para os cenários com relação ao número de grupos de ligação (GL) contendo QTL, os resultados de acurácia seletiva (R^2) e acurácia preditiva (REQM) obtidos pelos diferentes métodos estão apresentados de acordo com o modelo de simulação na Figura 6. Para o modelo 1, os métodos BA, RF e RBF apresentaram resultados de R^2

superiores aos demais métodos no cenário onde os QTL foram distribuídos em apenas um GL e no cenário onde os QTL foram distribuídos em oito GL. Apenas o RF e o RBF apresentaram resultados de R^2 superiores aos observados para o RR-BLUP. A árvore de regressão apresentou resultados inferiores de R^2 no cenário onde os QTL foram distribuídos em 8 GL (Figura 6a). Já para os resultados de REQM, a árvore de regressão e o BO foram os únicos que apresentaram resultados inferiores ao do RR-BLUP para o cenário de apenas um GL com QTL e os métodos de árvore de regressão, BA e RF apresentaram resultados inferiores ao do RR-BLUP para o cenário de oito GL com QTL.

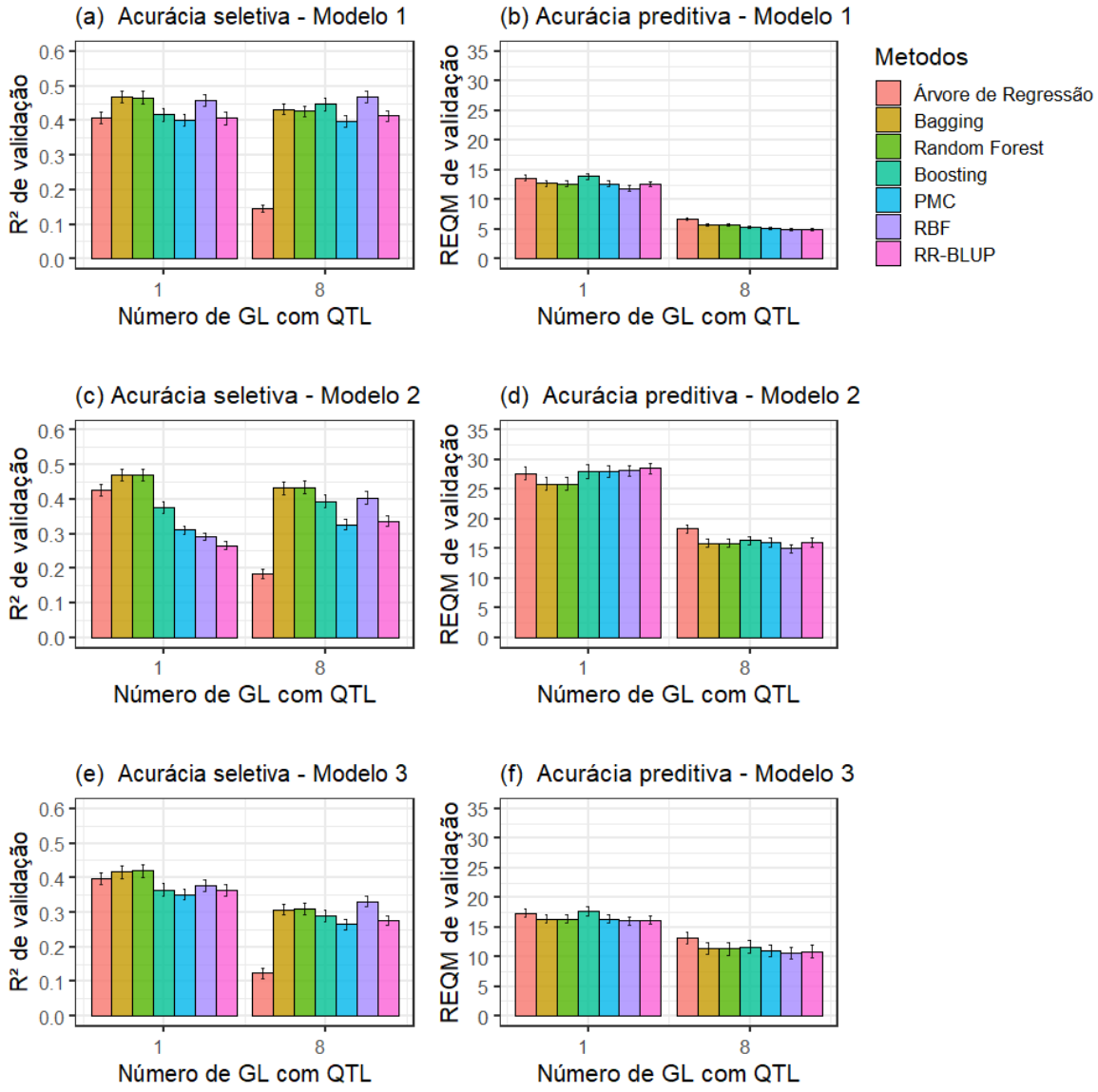
Para o modelo 2, todos os métodos apresentaram resultados de R^2 superiores aos do RR-BLUP quando os QTL estavam distribuídos em apenas um GL, sendo o BA e RF os métodos que apresentaram melhores resultados neste cenário. Já para o cenário de oito GL com QTL apenas os métodos BA, RF, BO e RBF apresentaram resultados superiores aos do RR-BLUP (Figura 6c). A árvore de regressão apresentou resultados superiores de R^2 em relação ao BO, PMC, RBF e RR-BLUP no cenário de um GL com QTL, no entanto, foi o método com os piores resultados para este parâmetro no cenário de oito GL com QTL. Para os resultados de REQM, o BA e RF apresentaram resultados melhores que o RR-BLUP, no cenário de um GL com QTL e a árvore de regressão foi o único método com resultados inferiores para este parâmetro no cenário com oito GL com QTL (Figura 6d).

No modelo 3, apenas os métodos de árvore de regressão, BA e RF apresentaram resultados de R^2 superiores aos observados pelo RR-BLUP para o cenário de um GL com QTL (Figura 6e). Já para o cenário de oito GL com QTL, apenas os métodos BA, RF e RBF foram superiores ao RR-BLUP para os resultados de R^2. A árvore de regressão novamente apresentou resultados de R^2 inferiores aos demais métodos nos cenários de oito GL com QTL. Os métodos apresentaram resultados semelhantes de REQM independente do cenário em relação à distribuição dos QTL nos GL (Figura 6f).

Em todos os três modelos, foram observados piores resultados de REQM, quando os QTL estavam distribuídos em apenas um grupo de ligação (Figura 6). Sendo essa diferença menos acentuada entre os cenários para o modelo 3. Os piores resultados de REQM de forma geral, foram observados para as características

simuladas pelo modelo 2, e os melhores resultados deste parâmetro quando os modelos foram simulados pelo modelo 1.

Figura 6 – Acurácia seletiva (R^2) e acurácia preditiva (REQM) [erros padrão] obtidas com os métodos de árvore de regressão com refinamentos BO (*Boosting*), BA (*Bagging*) e RF (*Random Forest*), de redes neurais MLP (*Multilayer Perceptron*) e RBF (*Radial Basis Function*) e RR-BLUP (*Random Regression Best Linear Unbiased Predictor*), para características simuladas nos cenários com distribuição dos QTL em apenas um grupo de ligação (GL) ou em oito GL e de acordo com o modelo de simulação aditivo (modelo 1) e epistáticos (modelos 2 e 3)



Fonte: O autor.

Para os cenários com relação ao grau médio de dominância, os resultados de acurácia seletiva (R^2) e acurácia preditiva (REQM) obtidos pelos diferentes métodos estão apresentados de acordo com o modelo de simulação na Figura 7.

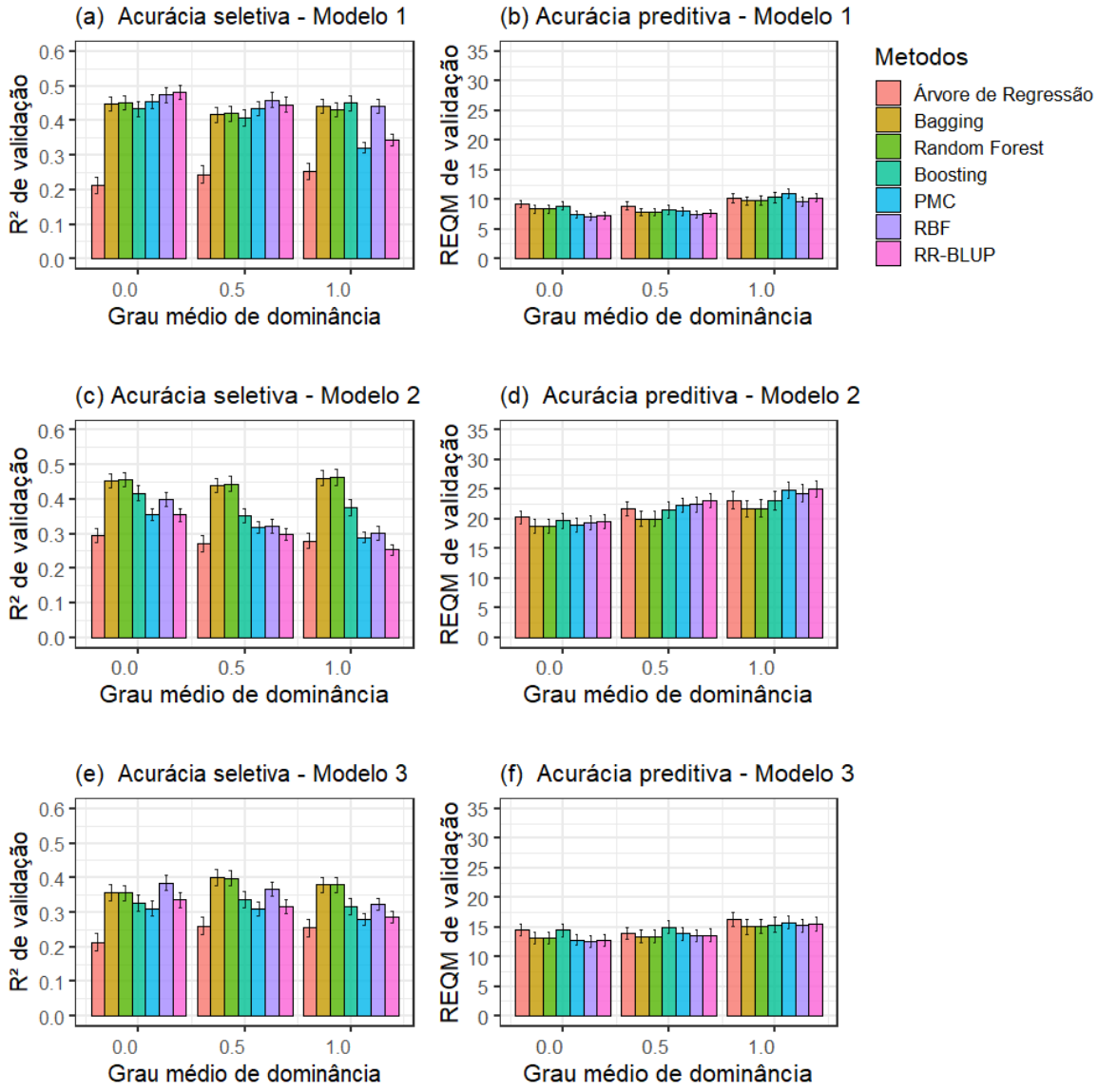
Para o modelo 1 a árvore de regressão apresentou resultados de R^2 inferiores aos demais métodos em todos os cenários de GMD (Figura 7a). Apenas no cenário com GMD igual a 1, os métodos BA, RF, BO e RBF foram superiores ao RR-BLUP para o parâmetro R^2. Neste cenário o PMC apresentou baixos valores de R^2, semelhante ao observado para o RR-BLUP. No cenário com GMD igual a 0, apenas os métodos de árvore de regressão e BO apresentaram resultados de REQM piores que os observados para o RR-BLUP (Figura 7b). Já nos demais cenários, nenhuma diferença foi observada entre os métodos para este parâmetro.

Para o modelo 2 os métodos BA, RF e BO apresentaram resultados de R^2 superiores aos observados para o RR-BLUP para todos os cenários de GMD (Figura 7c). Já o RBF, apresentou resultados de R^2 superiores aos do RR-BLUP apenas nos cenários com GMD igual a 0 ou 1 e o PMC com resultados superiores aos do RR-BLUP, apenas no cenário com GMD igual a 1. Para os resultados de REQM, apenas o BA e o RF apresentaram resultados melhores que os do RR-BLUP nos cenários com GMD de 0,5 ou 1 (Figura 7d).

Já para o modelo 3 o RBF foi o único método que apresentou resultados de R^2 superior aos observados para o RR-BLUP em todos os cenários de GMD (Figura 7e). Os BA e RF apresentaram resultados de R^2 superior ao do RR-BLUP para os cenários de GMD igual a 0,5 ou 1 e a árvore de regressão foi inferior ao RR-BLUP nos cenários com GMD de 0 ou 0,5. Não foram observadas diferenças para os valores de REQM para os diferentes métodos nos diferentes cenários de GMD (Figura 7a).

De forma geral, em todos os modelos, apenas os métodos PMC, RBF e RR-BLUP apresentaram redução para os resultados de R^2 com o aumento do GMD (Figura 7). Porém, para os valores de REQM, em todos os modelos, houve uma tendencia de aumento dos valores de REQM com o aumento do GMD.

Figura 7 – Acurácia seletiva (R^2) e acurácia preditiva (REQM) [erros padrão] obtidas com os métodos de árvore de regressão com refinamentos BO (*Boosting*), BA (*Bagging*) e RF (*Random Forest*), de redes neurais MLP (*Multilayer Perceptron*) e RBF (*Radial Basis Function*) e RR-BLUP (*Random Regression Best Linear Unbiased Predictor*), para características simuladas nos cenários com diferentes graus médios de dominância (d=0, d=0,5 ou d=1) e de acordo com o modelo de simulação aditivo (modelo 1) e epistáticos (modelos 2 e 3)



Fonte: O autor.

Para os cenários com relação à herdabilidade, os resultados de acurácia seletiva (R^2) e acurácia preditiva (REQM) obtidos pelos diferentes métodos estão apresentados de acordo com o modelo de simulação na Figura 8.

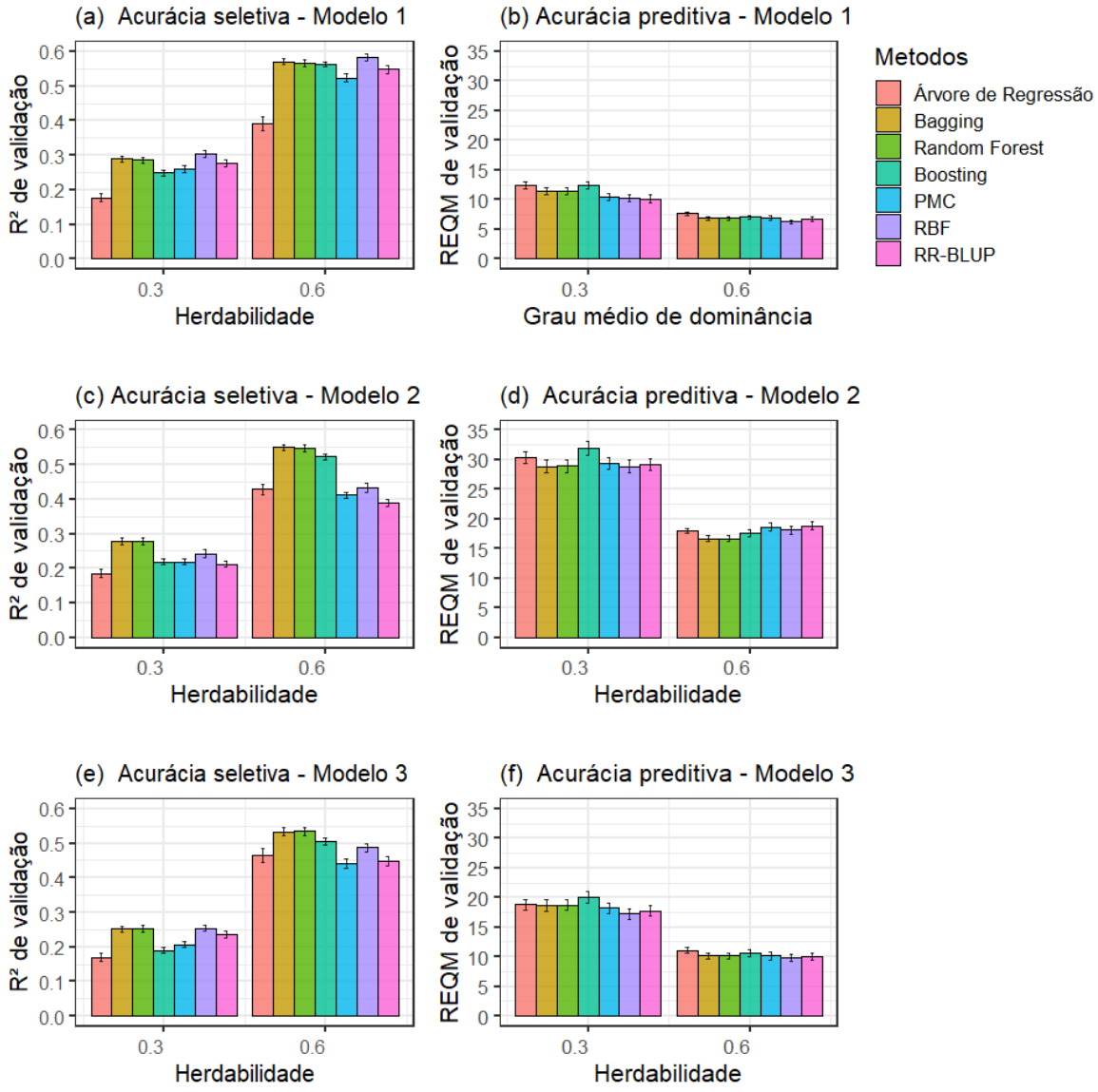
Para o modelo 1 apenas o RBF apresentou resultados de R^2 superiores aos do RR-BLUP para ambas as herdabilidades (Figura 8a). O BA apresentou resultados de R^2 superiores aos do RR-BLUP no cenário de herdabilidade de 0,6. A árvore de regressão apresentou resultados de R^2 inferiores aos demais métodos em ambas as herdabilidades e o BO apresentou resultados inferiores aos do RR-BLUP nos cenários de baixa herdabilidade. A árvore de regressão e o BO também apresentaram resultados inferiores aos do RR-BLUP para o parâmetro REQM, no cenário de baixa herdabilidade (Figura 8b).

Para o modelo 2 o BA, RF e RBF apresentaram resultados de R^2 superiores aos observados para o RR-BLUP em ambos os cenários de herdabilidade (Figura 8c). O BO e o PMC apresentaram resultados superiores aos do RR-BLUP apenas no cenário com herdabilidade de 0,6. O BA e RF também apresentaram melhores resultados de REQM em relação aos observados para o RR-BLUP nos cenários com herdabilidade de 0,6 (Figura 8d). Já o BO apresentou resultados de REQM piores que os do RR-BLUP para o cenário com herdabilidade de 0,3.

Para o modelo 3 os métodos BA, RF, BO e RBF apresentaram resultados de R^2 superiores aos do RR-BLUP apenas no cenário com herdabilidade de 0,6 (Figura 8e). Já no cenário com herdabilidade de 0,3, nenhum método apresentou resultado de R^2 superior ao observado para o RR-BLUP, que por sua vez, foi superior aos resultados observados para os métodos de árvore de regressão, BO e PMC neste cenário. Apenas o BO apresentou resultados piores de REQM em comparação com o RR-BLUP no cenário com herdabilidade de 0,3 (Figura 8f).

Em todos os três modelos foi observado uma melhora nos resultados de R^2 e REQM quando o cenário passou de 0,3 para 0,6 de herdabilidade (Figura 8). O modelo 2 foi o que proporcionou maiores valores de REQM para todos os métodos.

Figura 8 – Acurácia seletiva (R^2) e acurácia preditiva (REQM) [erros padrão] obtidas com os métodos de árvore de regressão com refinamentos BO (*Boosting*), BA (*Bagging*) e RF (*Random Forest*), de redes neurais MLP (*Multilayer Perceptron*) e RBF (*Radial Basis Function*) e RR-BLUP (*Random Regression Best Linear Unbiased Predictor*), para características simuladas nos cenários com valores de herdabilidade baixos (h^2 = 0,30) ou moderados (h^2 = 0,60) e de acordo com o modelo de simulação aditivo (modelo 1) e epistáticos (modelos 2 e 3)



Fonte: O autor.

Discussão

Nossos resultados, para os diversos cenários e para os três diferentes modelos de ação genotípica, indicaram que os métodos BA, RF e RBF apresentam resultados

tanto de acurácia seletiva (R^2) quanto preditiva (REQM) sempre igual ou superior aos observados para o método RR-BLUP. As técnicas fundamentadas em árvore de regressão, *Bagging* e RF fundamentam-se em encontrar partições nas variáveis preditoras que possibilitariam melhor predição na variável resposta. A técnica RBF também tem, em seu processamento híbrido, uma primeira etapa de agrupamento de valores identificando grupos de indivíduos similares, segundo suas entradas.

Resultado semelhante para o método RF foi observado por Alves *et al*. (2020), onde, na comparação do desempenho preditivo do GBLUP (*Genomic Best Linear Unbiated Predictor*) e métodos de aprendizado de máquina (*Random Forest* - RF, *Support Vector Machine* - SVM e *Artificial Neural Network* - ANN) em populações simuladas, apenas o RF foi capas de capturar implicitamente os sinais de dominância. Já para os resultados de RBF, González-Camacho *et al*. (2012), também observaram resultados superiores na avaliação de linhagens simuladas e reais de milho genotipadas com marcadores de alta densidade e avaliadas para várias combinações de características e ambiente. Estes autores ainda ressaltam o poder dessa técnica para capturar padrões epistáticos. Conforme apontado por Gianola *et al*. (2006), e corroborado por Long *et al*. (2010), os modelos não paramétricos não impõem suposições fortes sobre a relação fenótipo-genótipo e permitem capturar interações entre loci.

Uma ligeira melhora nos resultados de R^2 e REQM com o aumento no número de QTL foi observado, principalmente para os resultados de REQM no modelo 3. Resultados semelhantes foram observados por Barbosa *et al*. (2021), onde os autores destacam que apesar do aumento no valor de acurácia em função do número de QTL, não se espera que essa relação possa ser projetada para valores maiores de QTL de forma linear, uma vez que o valor de acurácia seletiva depende também de uma relação ideal entre o número de marcadores e QTL. Isso se deve ao fato de que, com o aumento do número de QTL, espera-se que a variação genética total seja dividida entre os QTL, assim, a eficiência dos métodos para estimar esses pequenos efeitos de QTL diminuirá, levando a uma perda de precisão (GHAFOURI-KESBI *et al*., 2017; RESENDE *et al*., 2012).

Sob a ação genética com ausência de efeitos epistáticos (modelo 1), as diferenças nos resultados de R^2 e REQM para os diferentes métodos foram pouco acentuadas, com exceção da árvore de regressão que apresentou resultados muito

inferiores na maioria dos cenários. Já para os modelos que incluíam efeitos epistáticos (modelos 2 e 3) o desempenho diferencial dos métodos foram mais acentuados, principalmente para os resultados de R^2. As simulações realizadas a partir do modelo 3, foram configurações mais desafiadoras para os resultados de R^2 obtido para todos os métodos, com resultados de R^2 piores que os observados para os cenários simulados a partir do modelo 2. Apesar do modelo 3 conter menor número de combinações epistáticas, cada combinação pode apresentar efeitos maiores, o que pode ter impactado a análise. Já para os resultados de REQM, foram os cenários simulados a partir do modelo 2 que se mostraram mais desafiadoras, com valores de REQM na maioria dos cenários, piores que os observados para o modelo 3.

Os métodos BA e RF, exibiram bons e semelhantes resultados de R^2 e REQM e se mostraram robustos para os três modelos aditivo e com diferentes níveis de interação epistática. Já o RBF, teve uma redução de eficiência em alguns dos cenários onde o número de combinações epistáticos eram maiores (modelo 2).

Para os cenários onde os QTL foram distribuídos apenas dentro de um GL ou dentro de oito GL para um total de dez em ambos os casos, em média os métodos tiveram menores valores de R^2 quando os QTL estavam distribuídos em mais GL. Esse resultado ocorre porque as técnicas penalizam uma certa proporção de marcadores que apresentam baixo efeito. Com isso, nos cenários onde os QTL estavam distribuídos apenas dentro de um GL, as técnicas poderiam facilmente penalizar SNP presentes em nove dos demais GL que não apresentavam nenhum efeito para a característica, o que pode ser dificultado nos demais cenários, principalmente devido ao elevado desequilíbrio de ligação presente dentro de cada GL. Resultados semelhantes foram observados por Long *et al*. (2011), onde os autores encontraram resultados de acurácia seletiva superiores em cenários onde os QTL estavam vinculados e apresentavam elevado desequilíbrio de ligação.

O grande problema das técnicas de predição é a capacidade de reconhecer os verdadeiros sinais de causa e efeito, reduzindo ruídos ambientais. Técnicas como PMC e *Boosting*, repassar os dados, em diferentes épocas, buscando a melhor aproximação. Os efeitos genéticos e ambientais sobre as técnicas utilizadas podem ser apreciados considerando os valores de acurácia nos cenários de diferente herdabilidade. Assim, verifica-se que a variação na herdabilidade foi o efeito que provocou maior impacto nos valores médios de R^2, e nesses cenários as técnicas BA

e RF se destacaram para baixa e alta herdabilidade. O impacto da herdabilidade sobre os resultados de R^2 foram também destacados em outros trabalhos (ALVES *et al*., 2020; BARBOSA *et al*., 2021). De acordo com Barbosa *et al*., (2021) o *Boosting* não é uma boa estratégia, principalmente para dados com baixa herdabilidade e com maior influência da variância residual, uma vez que é treinada repetidamente na mesma amostra para que a cada iteração, uma medida de erro de previsão seja calculada para cada indivíduo, e na próxima iteração, indivíduos com maiores erros recebam maior peso no treinamento do modelo.

Nos modelos de regressão, como RR-BLUP, os efeitos não lineares devem ser considerados de forma explícita no modelo o que pode resultar em uma maior dimensionalidade na matriz X que, originalmente, já é de alta dimensão computando apenas os efeitos de dose. Técnicas de inteligência computacional são preconizadas como capazes de captar os efeitos de interação, representados neste estudo pela dominância e pela epistasia, por meio de camadas abstratas em sua topologia. Procedimentos de aprendizado de máquina possibilitam, por meio de partições, adequar efeitos de interações epistática de forma similar aos modelos biológicos fundamentados em vias biossintéticas.

Em geral, nossos resultados apontaram que BA, RF e RBF apresentaram resultado iguais ou superiores à média entre todos os métodos, tanto para os resultados de R^2 e REQM. Estes métodos apresentaram resultados superiores aos obtidos com abordagem RR-BLUP em muitos dos cenários, principalmente para os valores de R^2. Esses pontos indicam que esses métodos são abordagens adequadas para predizer valores genéticos genômico e desempenho fenotípico para caracteres complexos na presença de efeitos de dominância e epistasia.

Conclusão

Todas as técnicas sofrem impacto com a adição de efeitos perturbadores, no entanto algumas técnicas podem apresentar melhores resultados de acordo com o cenário em estudo.

O desempenho dos métodos pode ser diferente de acordo com o modelo de simulação para efeitos epistáticos.

As metodologias *Bagging*, *Randon Forest* e RBF são técnicas que apresentam resultados consistentemente eficientes em diferentes cenários, o que as tornam alternativas adequadas para predizer valores genéticos totais e desempenho fenotípico para caracteres complexos na presença de efeitos de dominância e epistasia.

Agradecimento

Os autores agradecem o apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES e do Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq.

Referências:

ABDOLLAHI- ARPANAHI, R.; GIANOLA, D.; PEÑAGARICANO, F. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. **Genet. Sel. Evol**. 52:1-15, 2020.

ALKIMIM, E.R.; CAIXETA, E.T.; SOUSA, T.V.; *et al*. Selective efficiency of genome-wide selection in Coffea canephora breeding. **Tree Genetics & Genomes** 16, 41, 2020.

ALVES, A.A.C.; COSTA, R.M.; BRESOLIN, T.; ALVES, G.; JÚNIOR, F.; ESPIGOLAN, R.; RIBEIRO, A.M.F.; CARVALHEIRO, R.; ALBUQUERQUE, L.G. Genome-wide prediction for complex traits under the presence of dominance effects in simulated populations using GBLUP and machine learning methods. **Journal of Animal Science**, Volume 98, Issue 6, June 2020, skaa179, 2020.

BALTA B.; TOPAL, M. Describing factors affecting birth weight and growth traits in hemsin lambs using decision tree methods. **The Journal of Animal & Plant Sciences**, 30(3): Page: 560-567, 2020.

BARBOSA, I.D.P.; DA SILVA, M.J.; DA COSTA, W.G.; CASTRO SANT'ANNA, I.; NASCIMENTO, M.; CRUZ, C. D. (. Genome-enabled prediction through machine learning methods considering different levels of trait complexity. **Crop Science**, 61(3), 1890– 1902, 2021.

BEAM, A.L.; MOTSINGER-REIF, A.; DOYLE, J. Redes neurais bayesianas para a detecção de epistasia em estudos de associação genética. **BMC Bioinformatics** 15, 368, 2014.

BELLOT, P.; DE LOS CAMPOS, G.; PÉREZ-ENCISO, M. Can Deep Learning Improve Genomic Prediction of Complex Human Traits? **Genetics**. 210:809–819, 2018.

BHATTA, M.; GUTIERREZ, L.; CAMMAROTA, L.; CARDOZO, F.; GERMÁN, S.; GÓMEZ-GUERRERO, B.; PARDO, M.F.; LANARO, V.; SAYAS, M.; CASTRO, A.J. Multi-trait genomic prediction model increased the predictive ability for agronomic and malting quality traits in barley (Hordeum vulgare L.). **G3-Genes-Genom Genet**. 2020;10:1113–24, 2020.

BREIMAN, L.; FRIEDMAN, J.H.; OLSHEN, R.A.; STONE, C.J. Classification and Regression Trees, **Wadsworth Int. Group, Belmont**, California, USA, 1984.

BREIMAN, L. **Bagging predictors. Machine Learning**, 24, pp. 123-140, 1996.

CHEN, W.; HONG, H.; LI, S.; SHAHABI, H.; WANG, Y.I.; WANG, X.; AHMAD, B.B. **Flood susceptibility modelling using novel hybrid approach of reduced-error pruning trees with bagging and random subspace ensembles**. J. Hydrol., 575, pp. 864-873, 2019.

CRUZ, C.D. Genes Software – extended and integrated with the R, Matlab and Selegen. Acta **Scientiarum. Agronomy**. Maringá, v. 38, n. 4, p. 547-552, Oct.-Dec., 2016.

CRUZ, C.D.; SALGADO, C.C.; BHERING, L.L. **Genômica aplicada**. Visconde do Rio Branco, MG: Suprema, 424f. 2013.

DEOMANO, E.; JACKSON, P.; WEI, X.; AITKEN, K.; KOTA, R.; PÉREZ-RODRÍGUEZ, P. Genomic prediction of sugar content and cane yield in sugar cane clones in different stages of selection in a breeding program, with and without pedigree information. **Mol Breeding**.; 40:38, 2020.

DEMUTH, H.; BEALE, M. **Neural Network Toolbox User's Guide The Mathworks, Natick**, 2000.

DERBYSHIRE, M.C.; KHENTRY, Y.; SEVERN-ELLIS, A.; *et al*. Modeling first order additive × additive epistasis improves accuracy of genomic prediction for sclerotinia stem rot resistance in canola. **Plant Genome.** 2021.

DUDLEY, John.W. Epistatic interactions in crosses of Illinois high oil × Illinois low oil and of Illinois high protein × Illinois low protein corn strains. **Crop Sci**. 48 59– 68, 2008.

DUDLEY JW, JOHNSON GR (2010) Epistatic models improve between years prediction and prediction of testcross performance in corn. **Crop Sci** 50:763–769.

FERREIRA, R.A.D.C.; SILVA, G.N.; GLÓRIA, L.S.; SANT'ANNA, I.C.; RODRIGUES, H.S.; SOLVA, F.F.; CRUZ, C.D. RNA – Aplicação em estudos de seleção genômica ampla. **In Inteligência computacional aplicada ao melhoramento genético**. Cruz CD e Nascimento M. Viçosa, Ed. UFV, 2018.

GHAFOURI-KESBI, F.; GHOLIZADEH, M.; Genetic and phenotypic aspects of growth rate and efficiency-related traits in sheep. **Small Ruminant Research**, 149, 181–187, 2017.

GIANOLA, D.; FERNANDO, R.; STELLA, A. Genomic-assisted prediction of genetic values with semiparametric procedures. **Genetics** 173:1761–1776, 2006.

GONZALEZ-CAMACHO, J. M.; DE LOS CAMPOS, G.; PE´REZ, P.; GIANOLA, D.; CAIRNS, J. E.; MAHUKU, G. BABU, R.; CROSSA, J. Genome-enabled prediction of genetic values using radial basis function neural networks. **Theor. Appl. Genet**. 125, 759–771, 2012.

HOLLAND James B. Epistasis and plant breeding. **Plant Breed Rev** 21:27–92, 2001.

HOWARD, R.; CARRIQUIRY, A. L.; Beavis, W. D. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. **G3**. 4(6):1027-1046, 2014.

LONG, N.; GIANOLA, D.; ROSA, G.J.M.; WEIGEL, K.A.; KRANIS, A.; GONZALEZ RECIO, O. Radial basis function regression methods for predicting quantitative traits using SNP markers. **Genet Res** 92:209–225, 2010.

LONG, N.; GIANOLA, D.; ROSA, G.J.M. *et al*. Marker-assisted prediction of non-additive genetic values. **Genetica** 139, 843–854, 2011.

MA W.; QIU, Z.; SONG, J.; LI, J.; CHENG, Q.; ZHAI, J.; MA, C. A deep convolutional neural network approach for predicting phenotypes from genotypes. **Planta**. 248:1307–1318, 2018.

MASTRODOMENICO, A.T.; BOHN, M.O.; LIPKA, A.E.; BELOW, F.E. Genomic selection using maize ex-plant variety protection germplasm for the prediction of nitrogen-use traits. **Crop Sci.** 2019;59(2019):212–20.

MATLAB (2010). **Matlab Version 7.10.0.** Natick, Massachusetts: The Math Works Inc.

MEUWISSEN, T.H.E.; HAYES, B.J.; GODDARD, M.E. Prediction of total genetic value using genome wide dense marker maps. **Genetics**, v. 157, p. 1819-1829, 2001.

MONTESINOS-LÓPEZ, OA, MONTESINOS-LÓPEZ, A., PÉREZ-RODRÍGUEZ, P. *et al*. Uma revisão das aplicações de aprendizagem profunda para seleção genômica. **BMC Genomics** 22, 19, 2021.

MÔRO, G.V.; SANTOS, M.F.; DE SOUZA JÚNIOR, C.L. Comparison of genome-wide and phenotypic selection indices in maize. **Euphytica** 215, 76, 2019.

R CORE TEAM. R: **A language and environment for statistical computing**. Vienna: R Foundation for Statistical Computing, 2017. Available at: <https://www.R-project.org/>.

RESENDE, M.D.; RESENDE, M.F. Jr.; SANSALONI, C.P.; *et al*. Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. **New Phytol** 2012; 194: 116–128.

ROGAN, J.; FRANKLIN, J.; STOW, D.; MILLER, J.; WOODCOCK, C.; D. ROBERTS. "Mapping Land-Cover Modifications over LargeAreas:AComparison of Machine Learning Algorithms," Remote Sensing of Environment, 112(5):2272–2283, 2008.

SEARCHINGER, I.; WAITE, R.; HANSON, C.; RANGANATHAN, J. **Creating a Sustainable Food Future: A Menu of Solutions to Feed Nearly 10 Billion People by 2050**. Washington, DC, USA, July 2019. 556p.

SMALLWOOD, C.J.; SAXTON, A.M.; GILLMAN, J.D.; BHANDARI, H.S.; WADL, P.A.; FALLEN, B.D.; HYTEN, D.L.; SONG, Q.; PANTALONE, V.R. Context-specific Genomic Selection Strategies Outperform Phenotypic Selection for Soybean Quantitative Traits in the Progeny Row Stage. **Crop Sci.**;59(1):54–67, 2019.

SOUSA, I.C.; NASCIMENTO, M.; SILVA, G.N.; NASCIMENTO, A.C.C.; CRUZ, C.D.; SILVA, F.F.E.; DE ALMEIDA, D.P.; PESTANA, K.N.; AZEVEDO, C.F.; ZAMBOLIM, L.; CAIXETA, E.T. Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. **Scientia Agricola**, 78(4), 2021.

SYAHRANI, I. M. Comparation Analysis of Ensemble Technique With Boosting(Xgboost) and Bagging (Randomforest) For Classify Splice Junction DNA Sequence Category. **J. Penel. Pos dan Inform.** 9, 27–36, 2019.

TOMAZ, R.S.; ALVES, D.P.; NASCIMENTO, M.; CRUZ, C.D. Inteligência computacional. **In Inteligência computacional aplicada ao melhoramento genético.** Cruz CD e Nascimento M. Viçosa, Ed. UFV, 2018.

WALDMANN, Patrik. Genome-wide prediction using Bayesian additive regression trees. **Genet. Sel. Evol**. 48(42):1-12, 2016.

6. CONCLUSÃO GERAL

O desempenho dos métodos estatísticos, de inteligência computacional e de aprendizado de máquina são diferentes, de acordo com a complexidade das características analisadas.

Todas os métodos sofrem impacto com a adição de efeitos perturbadores, no entanto algumas técnicas podem apresentar melhores resultados de acordo com o cenário em estudo.

Não a uma integral consistência na performance dos métodos de acordo com o cenário esperado para as características, dessa forma uma técnica pode apresentar valores superiores em determinados cenários de complexidade e inferiores para outros.

Apesar disso, os métodos *Bagging*, *Random Forest* e RBF apresentam resultados consistentemente eficientes em diferentes cenários e com resultados sempre iguais ou superiores aos observados pelo RR-BLUP, o que as tornam alternativas adequadas para predizer valores genéticos totais e desempenho fenotípico para caracteres complexos na presença de efeitos de dominância e epistasia.

7. CONSIDERAÇÕES GERAIS

A principal contribuição científica deste trabalho refere-se à ampliação da base de conhecimento relacionada ao comportamento e potencialidade das metodologias baseadas em inteligência computacional e aprendizado de máquinas em relação aos diferentes cenários de complexidade de características fenotípicas, e o desempenhe dessa em comparação com metodologias tradicionalmente aplicadas. Os conhecimentos gerados por este trabalho poderão ser utilizados como literatura para ajudar no desenvolvimento de para cientistas em diversas áreas de estudo, como na genética e estatística. O trabalho pode ajudar, também a reduzir esforços e recursos financeiros com a aplicação de técnicas mais adequadas e eficientes. Esse trabalho fornece ainda, oportunidades para novas perguntas e questionamentos que devem ser resolvidos em pesquisas futuras, a fim de obter um ganho contínuo no conhecimento científico.

A execução deste trabalho, permitiu adquirir conhecimentos para um melhor entendimento da influência dos diferentes cenários de complexidade, modelos de simulação e suas consequências para os resultados de R^2 e REQM. Esses conhecimentos podem fornecer subsídios para a escolha de metodologias mais apropriadas para a seleção e/ou predição de valores genéticos superiores, tornando mais eficientes as atividades dentro de um programa de melhoramento genético.

Espera-se ainda que os resultados possam ser motivadores, no sentido de outros pesquisadores optarem pelo uso da técnica, de forma que possam agregar conhecimento e aumentar a eficiência em seus programas de melhoramento.