WEVERTON GOMES DA COSTA

EFICIÊNCIA DE TÉCNICAS DE MACHINE LEARNING E DE REDES NEURAIS NA PREDIÇÃO GENÔMICA E IDENTIFICAÇÃO DE MARCADORES

Tese apresentada 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

VIÇOSA - MINAS GERAIS
2022

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

T
C837e
2022

Costa, Weverton Gomes da, 1992-
Eficiência de técnicas de machine learning e de redes neurais na predição genômica e identificação de marcadores / Weverton Gomes da Costa. – Viçosa, MG, 2022.
1 tese eletrônica (120 f.): il. (algumas color.).

Texto em português e inglês.
Inclui anexos.
Orientador: Cosme Damião Cruz.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Biologia Geral, 2022.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2022.326
Modo de acesso: World Wide Web.

1. Mapeamento cromossômico. 2. Marcadores genéticos. 3. Aprendizado do computador. 4. Inteligência computacional. 5. Redes neurais (Computação). 6. Epistasia (Genética). I. Cruz, Cosme Damião, 1958-. II. Universidade Federal de Viçosa. Departamento de Biologia Geral. Programa de Pós-Graduação em Genética e Melhoramento. III. Título.

CDD 22. ed. 572.8633

Bibliotecário(a) responsável: Euzebio Luiz Pinto CRB 6/3317

WEVERTON GOMES DA COSTA

EFICIÊNCIA DE TÉCNICAS DE MACHINE LEARNING E DE REDES NEURAIS NA PREDIÇÃO GENÔMICA E IDENTIFICAÇÃO DE MARCADORES

Tese apresentada 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: 17 de maio de 2022.

Assentimento:

Weverton G. da Costa

Weverton Gomes da Costa
Autor

Cosme Damião Cruz
Orientador

À minha avó Jovita Rodrigues Lana (in memorian) que tanto sonhou por esse título, seu sonho foi realizado.

À minha filha Lara Santos Gomes da Costa que amo tanto, por toda alegria, amor, carinho que veio trazer em minha vida.

À minha esposa Fabíola Suelen dos Santos, por sempre me apoiar, pelo companheirismo, compreensão, amor, sabedoria e carisma.

Aos meus Pais Vicente de Paula da Costa e Silvana Helena Gomes por todo ensinamento, educação e pelo exemplo de vida.

Aos meus irmãos Wesley José da Costa e Gisele Iriane Gomes da Costa pela amizade, companheirismos, carinho e conselhos em todos os momentos.

DEDICO E OFEREÇO

AGRADECIMENTOS

Aos meus pais Vicente de Paula da Costa e Silvana Helena Gomes, por serem meu amparo e apoiarem a conseguir essa conquista, sendo influentes à em meus estudos, com todo amor.

Aos meus irmãos Wesley José da Costa e Gisele Iriane Gomes da Costa pela amizade, apoio, companheirismo e auxílio em toda minha vida, sendo exemplos de amor, dedicação e união.

À minha esposa Fabíola por todo apoio, amor, amizade, companheirismo, dedicação e suporte.

À minha tão amada filha Lara, fonte de inspiração e carinho, com o seu sorriso encantador e amor. Papai te ama muito!

Ao meu orientador professor Cosme Damião Cruz, por ser referência para seus alunos e amigos, especialmente pra mim, pelos grandes ensinamentos, apoio, suporte, amizade, momentos de alegria e descontração.

Ao meu coorientador professor Moysés Nascimento, pela grande amizade, auxílio, suporte, disponibilidade, pelos grandes ensinamentos, pelas oportunidades e descontração.

Ao meu coorientador professor Aluízio Borém de Oliveira, pelo aceite em minha orientação durante o mestrado, apoio, influência nos estudos e ensinamentos.

Ao pesquisador Plínio César Soares da EPAMIG, pela amizade, ensinamentos, suporte e oportunidades a mim dadas.

Ao professor Dr. Leonardo Lopes Bhering por todo auxílio, ensino e disposição ao liberar os equipamentos necessários para que o trabalho fosse realizado.

Aos grandes amigos, membros do laboratório de Bioinformática, pela união, companheirismo, ensinamentos e momentos de descontração e alegria e a todos os amigos e familiares que de alguma forma fazem parte da minha história.

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.

MUITO OBRIGADO!

Crê em ti mesmo, age e verá os resultados.

Quando te esforças, a vida também se esforça para te ajudar

Chico Xavier

RESUMO

COSTA, Weverton Gomes da, D.Sc., Universidade Federal de Viçosa, maio de 2022. **Eficiência de técnicas de machine learning e de redes neurais na predição genômica e identificação de marcadores**. Orientador: Cosme Damião Cruz.

A seleção genômica ampla (*Genome Wide Selection* - GWS), utiliza marcadores moleculares distribuídos ao longo de todo o genoma a fim de predizer o mérito genético de plantas e animais. Os métodos de aprendizado de máquina (ML) e redes neurais artificiais (ANN) não são parametrizados e podem desenvolver modelos mais precisos e parcimoniosos para análise de GWS. Com o intuito de avaliar diferentes métodos de ML e ANN para avaliar a predição baseada em GWS, propusemos duas questões a serem respondidas por esse projeto de pesquisa. A primeira é que métodos diferentes proporcionariam previsões diversas de acordo com a complexidade da característica analisada e a segunda seria que a identificação de marcadores associados aos QTLs (*Quantitative Trait Locis*), também dependeria da complexidade da característica e do método analisado. Dois artigos foram desenvolvidos para responder essas questões. No primeiro artigo, o objetivo foi avaliar a precisão geral e a variabilidade do desempenho de predição de métodos baseados em ML (*Decision Tree, Boosting*, *Bagging*, *Random Forest* e MARS - *Multivariate Adaptive Regression Splines*) e ANN (*Multilayer Perceptron*, *Radial Basis Function*) comparadas ao G-BLUP em análises de predição genômica para características simuladas com diferentes números de genes na presença de epistasia e com diferentes graus de herdabilidades. No segundo artigo, o objetivo foi avaliar os métodos na associação de marcadores importantes identificados com as regiões de presença do QTLs, por meio do conjunto de dados simulado, considerando características com diferentes números de genes na presença de epistasia e de diferentes herdabilidade. Uma população F_2 em equilíbrio de Hardy-Weinberg foi simulada, constituída por 1000 indivíduos e 10 grupos de ligação de 200 cM, cada, correspondendo a 4010 SNP (*Single Nucleotide Polymorphism)*. Na predição, o aumento no número de QTL, beneficiou principalmente os métodos de redes neurais e o G-BLUP para R^2 e REQM. Para os demais métodos, nos cenários de 40 QTLs ou mais, o aumento do número de QTLs afetou positivamente os resultados dos parâmetros avaliados. A variação na herdabilidade provocou efeito inverso nos valores de R^2 e REQM. Os métodos MARS não aditivos apresentaram R^2 alto para

caracteres oligogênicas e para características poligênicas com alta herdabilidade e com 240 QTLs ou mais. Com relação a identificação de marcadores associados aos QTLs, a maioria dos métodos apresentaram maior índice de acertos na identificação dos marcadores em cenários com menor número de QTLs e com maior herdabilidade. A MARS 3 e o Boosting apresentaram alta capacidade de identificar os marcadores de importância, considerando as regiões associadas aos QTLs. O maior índice de erros também ocorreu em cenários com menor número de QTLs, mas com menor herdabilidade. A herdabilidade afetou positivamente o índice relativo na identificação dos marcadores associados aos QTLs. Nos cenários de 40 QTLs ou mais, o aumento do número de QTLs também afetou positivamente o índice relativo para a maioria dos métodos. Contudo, os melhores resultados foram encontrados para o cenário com maior herdabilidade e com 8 QTLs. Os métodos MARS 1, MARS 2, Boosting e Bagging foram os mais efetivos na detecção de marcadores importantes ao longo do genoma, principalmente para as características com 8 e 240 QTLs. A variação na herdabilidade e no número de QTLs impactou o desempenho dos métodos tanto para predição quanto para identificação dos marcadores associados a QTLs. Assim, a distribuição dos QTL nos grupos de ligação pode ser o principal atributo a ser avaliado na predição dos valores genéticos e identificação de marcas associadas à QTLs, quando o experimento é bem conduzido a fim de se obter um maior valor para a herdabilidade. Os métodos de ML e de ANN demonstraram alto potencial para predição de valores genéticos em caracteres com efeitos dominantes e epistáticos. Já para a identificação de marcadores associados às regiões de presença de QTLs, os métodos de aprendizado de máquinas são mais eficientes. O uso de diferentes métodos estatísticos, redes neurais e aprendizado de máquina resultou em diferentes consequências influenciadas pela complexidade e particularidade das características analisadas. Portanto, recomenda-se que ao avaliar a predição de valores genéticos e a importância de marcadores, o uso de múltiplas abordagens seja utilizado, a fim de escolher o melhor método a ser utilizado.

Palavras-chave**:** Inteligência artificial. Seleção Genômica ampla. Importância de variáveis. Característica Quantitativa.

ABSTRACT

COSTA, Weverton Gomes da, D.Sc., Universidade Federal de Viçosa, May, 2022. **Efficiency of machine learning and neural networks in genomic prediction and identification of markers**. Adviser: Cosme Damião Cruz.

Genomic wide selection (GWS) uses molecular markers distributed throughout the genome in order to predict the genetic merit of plants and animals. Machine learning (ML) and artificial neural networks (ANN) methods are not parameterized and can develop more accurate and parsimonious models for GWS analysis. In order to evaluate different ML and ANN methods to evaluate prediction based on GWS, we proposed two questions to be answered by this research project. The first is that different methods would provide different predictions according to the complexity of the analyzed traits and the second would be that the identification of markers associated with QTLs (Quantitative Trait Locus), would also depend on the complexity of the trait and the analyzed method. Two articles were developed to answer these questions. In the first article, the objective was to evaluate the general accuracy and the variability of the prediction performance of methods based on ML (Decision Tree, Boosting, Bagging, Random Forest, and MARS - Multivariate Adaptive Regression Splines) and ANN (Multilayer Perceptron, Radial Basis). Function) compared to G-BLUP in genomic prediction analyses for simulated traits with different numbers of genes in the presence of epistasis and with different degrees of heritability. In the second article, the objective was to evaluate the methods in the association of important markers identified with the regions of the presence of QTLs, through the simulated dataset, considering traits with different numbers of genes in the presence of epistasis and heritability different. An F2 population in Hardy-Weinberg equilibrium was simulated, consisting of 1000 individuals and 10 linkage groups of 200 cM each, corresponding to 4010 SNP (Single Nucleotide Polymorphism). For prediction, the increase in the number of QTLs mainly benefited the neural network methods and the G-BLUP for R^2 and REQM. For the other methods, in the scenarios of 40 QTLs or more, the increase in the number of QTLs positively affected the results of the evaluated parameters. The variation in heritability caused an inverse effect on the values of R^2 and REQM. Non-additive MARS methods showed high R^2 for oligogenic traits and for polygenic traits with high heritability and 240 QTLs or more. Regarding the identification of markers associated with QTLs, most methods showed a higher rate of success in identifying markers in scenarios with fewer

QTLs and higher heritability. MARS 3 and Boosting showed a high ability to identify important markers, considering the regions associated with QTLs. The highest error rate also occurred in scenarios with fewer QTLs, but with lower heritability. Heritability positively affected the relative index in the identification of markers associated with QTLs. In scenarios of 40 QTLs or more, increasing the number of QTLs also positively affected the relative index for most methods. However, the best results were found for the scenario with the highest heritability and with 8 QTLs. The MARS 1, MARS 2, Boosting and Bagging methods were the most effective in detecting important markers along the genome, mainly for traits with 8 and 240 QTLs. The variation in heritability and in the number of QTLs impacted the performance of the methods for both prediction and identification of markers associated with QTLs. Thus, the distribution of QTLs in linkage groups can be the main attribute to be evaluated in the prediction of breeding values and identification of marks associated with QTLs, when the experiment is well conducted in order to obtain a greater value for heritability. The ML and ANN methods showed high potential for predicting genetic values in traits with dominant and epistatic effects. As for the identification of markers associated with regions of the presence of QTLs, machine learning methods are more efficient. The use of different statistical methods, neural networks, and machine learning resulted in different consequences influenced by the complexity and particularity of the analyzed traits. Therefore, it is recommended that when evaluating the prediction of breeding values and the importance of markers, the use of multiple approaches is used, in order to choose the best method to be used.

Keywords: Artificial intelligence. Genomic wide selection. Variables importance. Quantitative Trait Loci.

SUMÁRIO

1. INTRODUÇÃO GERAL

A agricultura deverá fornecer alimentos suficientes para atender à expansão da população global, que poderá atingir 10 bilhões de habitantes até 2050. Assim, a demanda por alimentos de origem vegetal deverá aumentar em mais de 50% até meados deste século (SEARCHINGER et al., 2019). Entretanto, para garantir a sustentabilidade da segurança alimentar futura, esse aumento terá que contornar vários desafios, que podem afetar negativamente a produção global de alimentos, como mudanças climáticas, crescimento populacional, insegurança alimentar e desnutrição, escassez de recursos, degradação de ecossistemas e perda da biodiversidade (BERRY et al., 2015; BORSELLINO; SCHIMMENTI; EL BILALI, 2020; CAPONE et al., 2014; JHA et al., 2020; KITETU; KO, 2020).

Pesquisas multidisciplinares devem ser realizadas com o intuito de agregar o conhecimento para conseguir suprir essa demanda em um curto espaço de tempo. O melhoramento de plantas é uma das principais áreas que deve ser utilizada para solucionar esses desafios. O melhoramento genético de plantas é como um processo industrial em que as matérias-primas (germoplasma) são inseridas em um sistema de fabricação (desenvolvimento de linhagem, híbrido ou clone) para que produtos (cultivares) sejam comercializados e distribuídos após rigorosos testes e controle de qualidade (ensaios multi-ambientais) (BERNARDO, 2020). Nesse contexto, o melhoramento genético visa aumentar a frequência de alelos e/ou obter combinações genotípicas favoráveis (SILVA et al., 2014), a fim atender às necessidades direta, ou indireta, do produtor, do consumidor e da indústria.

Um problema a ser contornado pelos melhoristas é a grande dificuldade quando se trabalha com características quantitativas, ou seja, aquelas controladas por muitos genes. Identificar fatores genéticos para características complexas é um dos principais objetivos dos melhoristas de plantas. Essas características complexas são difíceis de caracterizar, porque dentre os vários genes que as influenciem, ainda possa existir uma possível interação entre eles, o que chamamos de epistasia. Além disso, fatores ambientais devem ser considerados como relevantes, uma vez que o efeito ambiental também pode participar de interações com variantes genéticas (YORK; EAVES, 2001). Dessa forma, a reprodutibilidade dessas características complexas ao nível de seleção é dificultada, reduzindo o ganho de seleção a cada ciclo de melhoramento.

Para solucionar tais problemas é fundamental o conhecimento sobre genética quantitativa, grande aliada do melhoramento genético.

Os princípios quantitativos da genética levaram ao estabelecimento de novos delineamentos genéticos e orientaram o aperfeiçoamento de métodos de melhoramento para características quantitativas ou contínuas (DUDLEY; MOLL, 1969). Com o avanço na área molecular, foram agregados os estudos de mapeamento permitindo identificar locos controladores de características quantitativas (*Quantitative Trait Loci* - QTL). Esses estudos ajudaram a elucidar a natureza da variação quantitativa das caracterítiscas (BERNARDO, 2020; KEARSEY; FARQUHAR, 1998). Nesse contexto, as tecnologias moleculares surgiram como um conjunto de poderosas ferramentas para análise e manipulação genética.

Com marcadores moleculares disponíveis é possível auxiliar o pesquisador a entender melhor os mecanismos biológicos existentes do desenvolvimento da planta diretamente em nível de DNA. Assim, tornando o processo de seleção e as atividades mais eficientes (ACQUAAH, 2012). Os dados fenotípicos gerados rotineiramente em programas de melhoramento genético agregados a informações de marcadores SNP, utilizados para prever os valores genotípicos (BERNARDO, 2020; SANT' ANNA et al., 2018), acelera o processo de escolha dos melhores indivíduos, e, consequentemente, aumenta o ganho de seleção.

A seleção genômica ampla (*Genome Wide Selection* - GWS), proposta por MEUWISSEN et al. 2001, utiliza marcadores moleculares, em especial os codominantes, distribuídos ao longo de todo o genoma a fim de predizer o mérito genético de cada indivíduo. A GWS é caracterizada por utilizar milhares de marcadores SNPs (*Single Nucleotide Polymorphisms*) que cobrem amplamente o genoma e que permite incorporar informações moleculares na predição do mérito genético dos indivíduos (ALKIMIM et al., 2020; GODDARD; HAYES, 2007; RESENDE et al., 2012; SOUSA et al., 2021; VAN EENENNAAM; YOUNG, 2014; ZHAO; METTE; REIF, 2015). Tal abordagem aumenta a acurácia na predição dos valores genéticos e possibilita acelerar o processo de melhoramento (DESTA; ORTIZ, 2014; LI et al., 2018; PEIXOTO et al., 2017; SINGH; SINGH, 2015; WONG; BERNARDO, 2008).

No entanto, alguns desafios estatísticos são enfrentados pelos métodos baseados em GWS. Alguns deles são problemas relacionados com a multicolinearidade e dimensionalidade dos dados. Vários métodos capazes de lidar

com esses problemas têm sido utilizados, como a melhor previsão imparcial linear genômica (*Genomic Best Linear Unbiased Prediction* - GBLUP), melhor previsão imparcial linear (*Random Regression Best Linear Unbiased Prediction* - RR-BLUP) e modelos Bayesianos (Bayes A, Bayes B,..., Lasso). Essas abordagens são adotadas como referências no estudo de GWS, mas a presença de fatores complicadores, tais como epistasia e dominância, dificultam a utilização desses modelos, visto que os mesmos requerem que tais efeitos sejam estabelecidos à priori pelo pesquisador. Essas abordagens podem ser consideradas como clássicas para prever o valor genotípico, uma vez que já foram descritas há duas décadas (BERNARDO, 2020).

Métodos baseados em inteligência computacional já demonstraram elevado potencial para análises baseadas em GWS. As metodologias de redes neurais artificias (*Radial Basis Function* – RBF e *Multilayer Perceptron* - MLP) (AZODI et al., 2019; FERREIRA et al., 2018; GLÓRIA et al., 2016; LI et al., 2018; SANT'ANNA, 2018) e árvore de decisão e seus refinamentos (ALVES et al., 2020; BARBOSA et al., 2021; GHAFOURI-KESBI et al., 2017a; SOUSA et al., 2021), mostraram ser igualmente eficientes ou até melhores na predição de valores genéticos.

Diferentemente dos métodos usuais de predição genômica, as metodologias baseadas em inteligência computacional não requerem suposições a priori sobre a relação entre as entradas (marcadores) e a saída (valores fenotípicos). Isso permite grande flexibilidade para lidar com diferentes tipos de características com controle gênico com efeitos aditivos, dominantes e epistáticos (LI et al., 2018). As metodologias baseadas em inteligência computacional não possuem pressuposições quanto ao modelo, já que seus resultados dependem do processo de aprendizado e não da distribuição das variáveis em si. Essas características dos métodos de aprendizado de máquinas possibilitam captar fatores complicadores, como a interações entre genes (epistasia), no modelo de predição e permitem que não haja nenhum pressuposto quanto à distribuição dos valores fenotípicos. Entretanto, a identificação de tais interações permanece difícil por causa de efeitos fracos ou inexistentes de alguns SNPs, um grande número de SNPs considerados ou falta de informações a priori sobre quais SNPs interagem (LIN et al., 2008).

Uma metodologia ainda não tão difundida na GWS, com ênfase no melhoramento genético de plantas, e que se assemelha às técnicas de inteligência computacionais já consolidadas é a splines regressão adaptativa multivariada

(*Multivariate Adaptive Regression Splines* - MARS) (FRIEDMAN, 1991). A MARS modela automaticamente não linearidades e interações entre as variáveis de entrada (TAYLAN; WEBER, 2019; ZHENG et al., 2020), e apresenta o modelo ajustado ao final do processo, diferentemente de modelos baseados em inteligência artificial. Neste sentido, a MARS além de possivelmente aumentar a acurácia de predição comparadas aquelas provenientes de metodologias tradicionais de GWS (GBLUP, RR-BLUP, Bayesiana), possibilita avaliar como os efeitos são determinados nas características que apresentem diferentes tipos de interações, permitindo ao pesquisador obter informações sobre a arquitetura genética da característica. O MARS é considerado ser mais flexível em comparação com a Árvore de decisão e a tradicional regressão linear (COOK; ZEE; RIDKER, 2004), e apresentou desempenho melhor que a redes neurais artificial para identificar interações entre genes (LIN et al., 2008). Além disso, o MARS utiliza polinômios no teste de interações entre genes (YORK; EAVES; VAN DEN OORD, 2006), possibilitando a identificação de não apenas um, mas de vários genes que posam estar controlando uma característica.

Diante do exposto, os objetivos desse estudo foram: (i) avaliar a precisão geral e a variabilidade do desempenho de predição de métodos baseados em aprendizado de máquinas, incluindo a MARS, e redes neurais em análises de predição genômica; e (ii) avaliar os métodos na associação de marcadores importantes identificados com as regiões de presença do QTLs para características simuladas para diferentes números de genes na presença de dominância e epistasia e com diferentes graus de herdabilidades.

2. REVISÃO DE LITERATURA

2.1. Predição

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; MENDES; CARVALHO, 2006). Essas características complexas são difíceis de se caracterizar porque, dentre os vários genes que as influenciem, ainda pode existir uma possível interação entre eles (epistasia). Além disso, fatores perturbadores ambientais devem ser considerados

como relevantes, uma vez que o efeito ambiental também pode participar de interações com os efeitos variantes genéticos (YORK; EAVES, 2001).

O conhecimento da natureza e da magnitude dos efeitos genéticos que controlam determinado caráter apresentam grande importância na seleção e na predição do comportamento de indivíduos em populações segregantes e gerações híbridas (CRUZ; REGAZZI; CARNEIRO, 2012). Conhecer a magnitude dos efeitos que controlam o caráter possibilita predizer qual será o ganho com a seleção (CRUZ, 2012).

Apesar das dificuldades em se trabalhar com características controladas por muitos genes, deve-se ter em mente que estes, individualmente, seguem as mesmas leis básicas da genética. Dessa forma, exibem os mesmos tipos de ação gênica conhecidos: ação aditiva, ação de dominância e epistática. Fisher (1941) descreveu a variação devido aos efeitos médios de dois alelos em um loco em um diploide como a variância aditiva (σ_A^2) e a parte não transmissível dos efeitos que incluíam desvios de dominância como a variância de dominância (σ_D^2) (BERNARDO, 2020). Por fim, Fisher (1941) reconheceu a existência de efeitos devido a interações de alelos em locais diferentes (epistasia - σ_E^2) (BERNARDO, 2020).

A maior ou menor influência de uma ou outra ação na herança de um caráter terá forte impacto sobre o efeito da seleção realizada para esse caráter (BUENO; MENDES; CARVALHO, 2006). As variâncias aditivas e não aditivas, as correlações e as herdabilidades são os parâmetros 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 (V_g) na variação fenotípica (V_f), ou seja, $h^2 = \frac{V_g}{V_f}$. Quando se decompõe a variância genotípica, o componente aditivo (σ_A^2ou 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. Segundo Cruz et al. (2014), a herdabilidade no sentido restrito pode ser expressa como: $h_r^2 = \frac{V_a}{V_f}$.

De posse da herdabilidade e a variância aditiva (σ_A^2) é possível 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 é o Ganho de Seleção; i é a Intensidade de seleção = $\mathrm{DS}/\hat{\sigma}_f$; h_r é a Acurácia de seleção no sentido restrito; σ_A é o Desvio padrão aditivo; I_g é o Intervalo de geração.

As diferentes estratégias em um programa de melhoramento vão afetar os parâmetros que compõe a Equação 1. Quanto mais acuradas forem as estimativas e maior a intensidade de seleção maior será o ganho de seleção. No entanto, quanto 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. Dessa forma, buscar metodologias que irão predizer de forma mais acurada o valor genético do indivíduo e estimativas de parâmetros é de extrema importância no programa de melhoramento genético.

2.2. Predição auxiliada por marcadores

Marcadores moleculares ligados a diferentes características de importância econômica permite a seleção indireta de características desejáveis nas gerações segregantes precoces. Essa estratégia tem potencial para reduzir tempo, recursos e energias necessários não só para desenvolver grandes populações segregantes por várias gerações, como também para estimar parâmetros usados na seleção indireta (CAIXETA et al., 2016).

Muitas vezes a natureza da característica de interesse dificulta a prática da seleção direta, devido à avaliação dessa característica ser de baixa acurácia no processo (CRUZ, 2012). A escolha do método de seleção que possibilite o desenvolvimento de genótipos superiores é uma das etapas mais importantes na busca de novos cultivares em um programa de melhoramento genético (ENTRINGER et al., 2014). Com o desenvolvimento de cultivares superiores, cada vez mais difícil se torna a identificação de novos genótipos superiores (CRUZ; SALGADO; BHERING, 2013). Assim, uma alternativa, é a realização de uma seleção indireta com base em características de mais fácil medição e menor influência ambiental, que estejam associadas à característica de interesse (MOURA et al., 2013).

A seleção indireta é uma metodologia que apresenta vantagem quando a herdabilidade do caráter auxiliar e a correlação deste com a característica de interesse

são elevados (BERED; BARBOSA NETO; CARVALHO, 1997). Dessa forma, os estudos de correlações fornecem importantes informações, como a possibilidade de identificar associações entre um caráter quantitativo de difícil ganho de seleção por meio de outras características correlacionadas e de maior herdabilidade ou de mais fácil mensuração (CRUZ, 2012). Pela existência de efeitos indiretos, na qual a correlação entre uma variável qualquer está associada à variável quantitativa devido ao efeito de uma terceira variável, deve-se buscar o uso de metodologias mais eficientes que podem ser utilizadas para seleção indireta.

A partir do trabalho de 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. 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; SALGADO; BHERING, 2013). No entanto, os fortes efeitos dos genes determinantes de marcadores fenotípicos podem 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. Adicionalmente, o ambiente pode modificar a expressão dos marcadores fenotípicos (BERED; BARBOSA NETO; CARVALHO, 1997; PATERSON; TANKSLEY; SORRELLS, 1991).

Com o rápido aperfeiçoamento das técnicas moleculares, fundamentadas na amplificação de fragmentos de DNA, grandes avanços têm sido alcançados na área dos marcadores de DNA (CRUZ; SALGADO; BHERING, 2013). Com marcadores moleculares disponíveis é possível auxiliar o pesquisador a entender melhor os mecanismos biológicos existentes do desenvolvimento da planta diretamente em nível de DNA. A vantagem dos marcadores moleculares sobre os dados fenotípicos está, também, na possibilidade de comparar genótipos mesmo que sejam amostrados em diferentes ambientes, tipos de tecido ou estágio de desenvolvimento (CAIXETA et al., 2016).

A partir do mapeamento de características em populações segregantes é possível realizar estudos de segregação genética e testes estatísticos para a detecção

de associações. O mapeamento de características qualitativas pode ser realizado por meio dos testes de segregação, levando em consideração a frequência de recombinantes observada. Já o mapeamento de QTL está baseado em procedimentos estatísticos. Sendo a denominação QTL utilizada para nomear as regiões cromossômicas que contêm genes (ou locos) que controlam esses caracteres quantitativos (FALCONER; MACKAY, 1996).

Os estudos de mapeamento identificaram os principais QTLs que foram considerados úteis no melhoramento das espécies cultivadas e ajudaram a elucidar a natureza da variação quantitativa (BERNARDO, 2020; KEARSEY; FARQUHAR, 1998). No entanto, há falta de métodos bem estabelecidos para incorporar epistasia na predição de características complexas em programas de melhoramento de plantas (BERNARDO, 2010; CARENA; HALLAUER; MIRANDA FILHO, 2010; GONZÁLEZ-CAMACHO et al., 2012).

A seleção por marcadores moleculares é uma forma de se realizar a seleção indireta no qual a característica auxiliar irá apresentar 100% de herdabilidade, uma vez que os marcadores moleculares não são influenciados pelo ambiente (BERED; BARBOSA NETO; CARVALHO, 1997). Assim, o ganho de seleção através da seleção pode ser calculado de acordo com a Equação 2:

$$GS_y = \frac{ir_{g\hat{g}}\,\sigma_{Ay}}{I_g} \qquad \text{Equação (2)}$$

em que: GS_y é o Ganho de Seleção em y; i é a Intensidade de seleção; r_gé a Acurácia na estimativa dos valores genéticos genômicos (VGG); σ_{Ay} é o Desvio padrão aditivo da característica principal y; I_g é o Intervalo de geração.

Os marcadores moleculares apresentam amplo potencial de uso no melhoramento de plantas, sendo 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; MENDES; CARVALHO, 2006).

2.3. Seleção Genômica Ampla (GWS)

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 dados de marcadores moleculares e dados fenotípicos que estejam em ligação gênica próxima com alguns QTLs, para a 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 (RESENDE, 2008).

Para o emprego da SAM é necessário o mapeamento de caracteres de interesse agronômico de forma a maximizar a correlação genética (BERED; BARBOSA NETO; CARVALHO, 1997). Para isso, deve-se determinar quantos marcadores/QTLs são necessários para explicar a maior parte da variação genética da característica quantitativa. Entretanto, QTLs de pequenos efeitos normalmente não são detectados, uma vez que além de efeitos de QTLs os caracteres quantitativos ainda podem apresentar efeitos de interações interalélicas (epistasia) e efeitos de interação de genótipos com ambientes (PODLICH; WINKLER; COOPER, 2004), resultando na capitação de apenas parte da variância genética explicada pelos marcadores/QTLs e, consequentemente, à subestimação desses efeitos (GODDARD; HAYES, 2007). Além disso, 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 (RESENDE, 2008). Dessa forma, a SAM tem sido mais utilizada em 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 e para caracteres de média ou alta herdabilidade (HOLLAND, 2004; HOSPITAL et al., 1997).

Existem diversos outros tipos de marcadores, com diversas especificidades e vantagens que devem ser avaliadas na adoção desses para estudos genéticos. Um dos mais utilizados o SNPs têm permitido maior eficiência na avaliação genética em nível molecular. Entre os vários tipos de variações detectadas na sequência do DNA das plantas, os SNPs têm-se mostrado os mais abundantes (HUANG; HAN, 2014). Atualmente a maioria das espécies de interesse econômico dispõe de número elevado de marcadores passíveis de uso em programas de melhoramento (RESENDE JR. et al., 2013).

A sua alta densidade no genoma, somada ao desenvolvimento de tecnologias de genotipagem, abrem novas possibilidades para a aplicação dos SNPs, como o no melhoramento assistido por marcadores moleculares, na integração de mapas físicos e genéticos e na seleção genômica ampla (CAIXETA et al., 2016). A seleção genômica ampla (*Genome Wide Selection* - GWS) foi proposta por MEUWISSEN; HAYES; GODDARD (2001), ao utilizar alta densidade de marcas e não realizar nenhum teste estatístico para escolha de marcas usadas na predição do mérito genético de indivíduos (FERREIRA et al., 2018). A GWS é caracterizada por utilizar milhares de marcadores SNPs que cobrem amplamente o genoma e que permite incorporar informações moleculares na predição do mérito genético dos indivíduos (ALKIMIM et al., 2020; GODDARD; HAYES, 2007; RESENDE et al., 2012; SOUSA et al., 2021; VAN EENENNAAM et al., 2014; ZHAO; METTE; REIF, 2015).

A GWS tem se mostrado relevante para o melhoramento, principalmente quando comparada a métodos usuais de seleção, como aqueles baseados em índices de seleção e BLUP (*Best linear Unbiased Prediction*), tradicional, por exemplo (FERREIRA et al., 2018). Tal abordagem aumenta a acurácia na predição dos valores genéticos e possibilita acelerar o processo de melhoramento (DESTA; ORTIZ, 2014; LI et al., 2018; PEIXOTO et al., 2017; SINGH; SINGH, 2015; WONG; BERNARDO, 2008). A maior acurácia de seleção e a redução do intervalo entre ciclos são os maiores benefícios da seleção genômica ampla, comparada aos demais métodos tradicionais de seleção. Esses ganhos têm justificado a sua utilização em programas de melhoramento vegetal (FRITSCHE-NETO, 2011; HAYES et al., 2009; RESENDE JR. et al., 2013).

A teoria da GWS baseia-se na cobertura do genoma por um grande número de marcadores, aumentando a probabilidade de que QTLs de interesse estejam em forte desequilíbrio de ligação com os marcadores. O principal intuito dessa abordagem é obter um modelo que prediz o valor genético do indivíduo, mas que não necessariamente determina genes específicos envolvidos no controle do caráter. 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,

permitindo que todos os QTLs, sejam eles de grandes ou pequenos efeitos, estarão em desiquilíbrio de ligação com os marcadores moleculares. Além disso, a GWS pode ser aplicada em toda a população, não se restringindo a uma família específica.

Para estimar o efeito individual de cada marcador, pode-se utilizar regressão linear simples descrita pela Equação 3:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad \text{Equação (3)}$$

em que: y é o valor fenotípico e x representa o valor genotípico, aa, Aa ou AA (-1, 0, 1).

Essa expressão pode ser generalizada para cálculo do efeito referente a todos os marcadores em uma equação multivariada com os efeitos, que são elementos de um vetor β, sendo estimados simultaneamente, assim tem-se a Equação 4:

$$y = \beta_0 + x_1\beta_1 + \cdots + x_m\beta_m + \varepsilon \qquad \text{Equação (4)}$$

em que β_m representa o efeito para cada marcador.

Dessa forma, a partir dos efeitos estimados para cada marcador é possível calcular o Valor Genético Genômico Predito (*Genomic Estimated Breeding Values* - GEBV). A forma matricial do cálculo do GEBV é dada conforme a Equação 5:

$$GEBV = X\hat{\beta} \qquad \text{Equação (5)}$$

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

Apesar de se mostrar uma técnica revolucionária, a GWS impõe desafios estatísticos e computacionais tais como a dimensionalidade do modelo, colinearidade entre marcas e a complexidade das características quantitativas (SANT'ANNA, 2018). Como na maioria das vezes a seleção genômica se aplica a um número maior de marcadores em relação ao número de observações fenotípicas, o uso de métodos baseados em mínimos quadrados ordinários não podem ser utilizado, uma vez que esses métodos requerem que o número de observações seja maior que o número de marcadores. Além disso, à medida que o número de SNPs aumenta e as interações são levadas em consideração, o número de termos necessários em um modelo para descrever estatisticamente todas as interações SNP–SNP possíveis de k maneira de n locos bialélicos aumenta substancialmente (LIN et al., 2008; WADE, 2000).

2.4. Métodos estatísticos utilizados em predições

Para contornar os desafios apresentados em relação à predição, vários

métodos têm sido propostos, na qual se diferem pelo tipo de suposição sobre o modelo genético associado ao caráter quantitativo (SANT'ANNA, 2018). Entre eles, os métodos estatísticos como RR-BLUP, Bayes A e Bayes B (MEUWISSEN; HAYES; GODDARD, 2001) e LASSO Bayesiano (DE LOS CAMPOS et al., 2009; LEGARRA et al., 2008), ou os métodos baseados em inteligência computacional como as Redes Neurais Artificiais (RNAs), Função de Base Radial (RBF) e Perceptron Multicamadas (MPL) (AZODI et al., 2019; GLÓRIA et al., 2016; LI et al., 2018; SANT' ANNA et al., 2018) e as Árvores de decisão e seus refinamentos (Bagging, Random Forest e Boosting) (ALVES et al., 2020; BARBOSA et al., 2021; GHAFOURI-KESBI et al., 2017a; SOUSA et al., 2021).

A escolha dos métodos para a predição dos efeitos de marcadores também pode afetar a acurácia dos valores 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 (SANT' ANNA et al., 2018).

A abordagem mais simples para modelar o efeito dos marcadores como um efeito aleatório é o uso do BLUP, por meio de uma regressão ridge ou aleatória (RR-BLUP). Essa metodologia, proposta por (MEUWISSEN; HAYES; GODDARD, 2001), torna-se boa alternativa quando muitos QTLs controlam a característica de interesse e nenhum deles é de grande efeito (RESENDE JR. et al., 2013). O modelo de RR-BLUP pode ser descrito pela Equação 6:

$$y = Xg + e \qquad \text{Equação (6)}$$

em que $g \sim N(0, I\sigma_g^2)$.

Os efeitos dos marcadores podem ser estimados através da Equação 7:

$$g = (X'R^{-1}X + I\sigma_g^2)^{-1}X'R^{-1}y \qquad \text{Equação (7)}$$

em que R é uma matriz diagonal que pode conter pesos relativos associados às acurácias dos valores fenotípicos desregressados (y) utilizados na predição. Em geral quando essa informação não é disponível, a matriz diagonal R é tida como $R = I\sigma_e^2$, e a equação pode ser simplificada: $g = (X'X + I\lambda)^{-1}X'y$, em que $\lambda = \sigma_e^2/\sigma_g^2$ é constantes para todos os marcadores (RESENDE JR. et al., 2013).

O método RR-BLUP é equivalente à substituição da matriz de parentesco (matriz A) pela matriz de parentesco genômico nas equações de modelos mistos

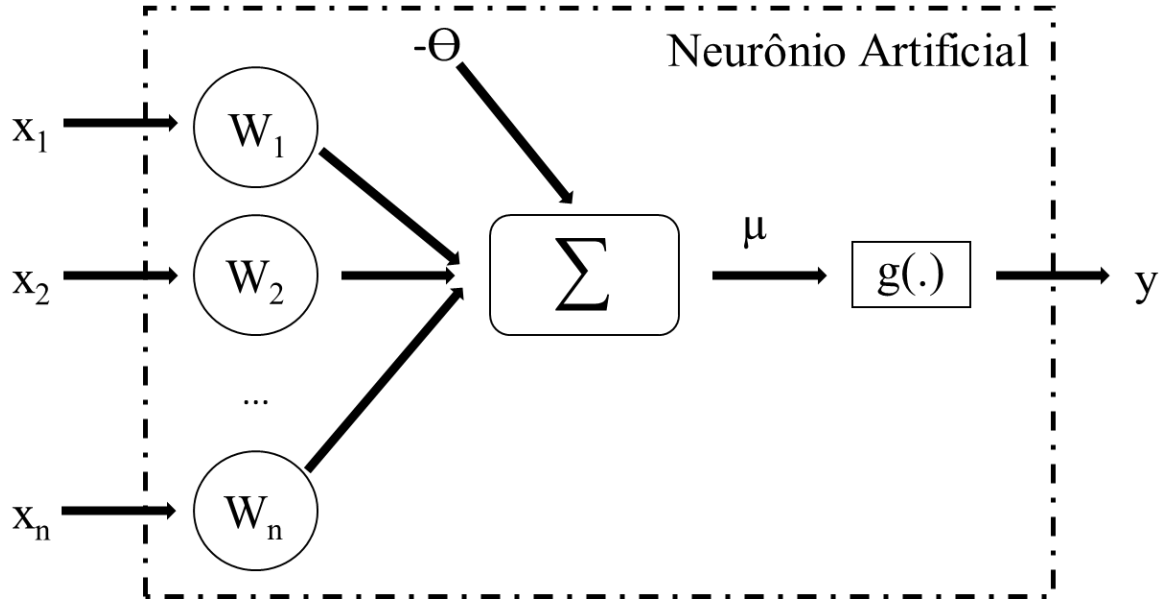
convencionalmente utilizados em análises quantitativas (G-BLUP) (HABIER et al., 2011). A diferença dos dois métodos é que no RR-BLUP estima-se o efeito individual de cada marcador, que é subsequentemente utilizado em conjunto para gerar o valor genético genômico. No caso do uso de G-BLUP, o termo predito é diretamente o valor genético genômico, via parentesco obtido pelos dados genômicos (RESENDE JR. et al., 2013).

2.5. Métodos de inteligência computacional e aprendizado de máquinas utilizados em predições

A inteligência artificial computacional é a área da ciência que estuda a teoria e a aplicação de algumas técnicas inspiradas na natureza: as redes neurais artificiais (RNAs), a lógica nebulosa e a computação evolucionária (TOMAZ et al., 2018). A inteligência artificial é uma área da ciência da computação que visa simular, em máquinas, a capacidade de solucionar problemas e realizar tarefas, que são uma habilidade da inteligência natural do homem (FERNANDES, 2003). Uma vez que a inteligência está diretamente relacionada à capacidade de aprendizado e de realização de uma determinada tarefa, uma das áreas da inteligência artificial de grande importância é o aprendizado de máquinas (COPPIN, 2010).

As RNAs 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, 2008; MCCULLOCH; PITTS, 1943) (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 X1, X2,..., Xn são as entradas da rede; W1, W2,..., Wn 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 à saída da rede.



FONTE: BARBOSA, 2021.

As RNAs são técnicas computacionais baseadas em modelos matemáticos apresentam funcionamento inspirado no cérebro humano, adquirindo conhecimento através da experiência (CRUZ et al., 2018). Pelo fato de as RNAs serem aptas a resolver problemas de cunho geral, tais como aproximação, classificação, categorização e predição, a gama de áreas onde estas podem ser aplicadas é bastante extensa (BRAGA; CARVALHO; LUDEMIR, 2007).

Há séculos que os princípios da inteligência artificial são utilizados em equipamentos utilizados para marcar o tempo e simular comportamento de animais (BITTENCOURT, 2006). A sua utilização nos programas de melhoramento genético é recente, sendo sua grande vantagem à estrutura não linear e processamento paralelo, o que permite captar propriedades mais complexas dos dados (CRUZ et al., 2018). Também destaca a sua capacidade de não requerer informação detalhada sobre os processos físicos do sistema a ser modelado (SUDHEER; GOSAIN; RAMASASTRI, 2003).

As metodologias baseadas em inteligência computacional possuem várias propriedades que as tornam potencialmente mais atraente para a GWS (BARBOSA et al., 2021; CRUZ; NASCIMENTO, 2018; GLÓRIA et al., 2016; SANT' ANNA et al.,

2018; SILVA et al., 2014; SOUSA et al., 2021): (I) o número de marcadores pode exceder o número das observações; (II) todos os marcadores, incluindo aqueles com efeitos fracos, marcadores altamente correlacionados e interagentes têm uma chance de contribuir para o ajuste do modelo; (III) interações complexas entre marcadores podem ser facilmente acomodadas; (IV) podem executar classificação e regressão simples e complexas com precisão; (V) geralmente, requerem modulação fina dos parâmetros e a parametrização padrão apresenta bom desempenho; e (VI) não fazem pressuposições distributivas sobre as variáveis preditoras.

Quando as RNAs são aplicadas à GWS, a camada de entrada é cada marcador com 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 et al., 2018).

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).

As RNAs caracterizam-se pela sua arquitetura e pelo ajustamento de seus pesos às conexões durante o processo de aprendizado (SANT'ANNA, 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 *backward*) 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. Nesses casos, a codificação proveniente da entrada não é suficiente para programar o mapeamento de saída. Isso pode levar a um problema onde os padrões de entrada similares

resultam em padrões de saída similares, o que leva o sistema à incapacidade de aprender importantes mapeamentos, incluindo aqueles não linearmente separáveis (CRUZ et al., 2018). Como resultado, padrões de entrada com estruturas similares, fornecidos do ambiente externo, que levem a saídas diferentes não são possíveis de ser mapeados por redes sem representações internas, isto é, sem camadas intermediárias (CRUZ et al., 2018).

2.5.1. Redes Perceptron Multicamadas (MLP)

A Rede Perceptron Multicamadas ou *Multilayer Perceptron* (MLP) apresenta uma estrutura de redes neurais que se caracteriza pela existência de uma ou mais camadas ocultas de neurônios, fundamentado no processo de aprendizagem denominado de *backpropagation* (CRUZ; NASCIMENTO, 2018). A rede MLP pode apresentar diversos neurônios na camada de saída, cada um desses irá representar uma das saídas do processo a ser mapeado. Assim, se um processo consistir de m saídas a MLP terá m neurônios em sua última camada neural.

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; 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 números excessivos de neurônios ou de camadas intermediárias. Números excessivos desses parâmetros podem ocasionar *overfitting* (sobreajuste do modelo). Em vez de aprender, a rede memoriza os dados e decora o padrão específico de entrada e da saída (CRUZ; NASCIMENTO, 2018) e dados fora dos padrões de entrada não irão ser alocados no padrão correto de saída, aumentando o erro da análise.

O algoritmo *backpropagation* (CHAN; FALLSIDE, 1987) utilizado no treinamento da rede MLP é baseado no cálculo do erro ocorrido na camada de saída da rede neural. Esse algoritmo recalcula o valor dos pesos do vetor w da última camada de neurônios e atualiza os pesos w das camadas de trás para frente, ou seja, da última até atingir a camada de entrada, realizando a retropropagação do erro obtido pela rede. 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 generalizada (Equação 8) até que se retorne à primeira camada oculta da rede (etapa também conhecida como fase de atualização de pesos ou retropropagação).

$$\Delta w_{(t)} = \alpha \Delta w_{(t-1)} + \eta \delta_{(t)} y_{(t)} \qquad \text{Equação (8)}$$

em que α é a constante de *momentum* com $0 < \alpha < 1$; 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 $(t - 1)$.

2.5.2. Redes de funções de base radial (RBF)

As Redes de Funções de Base Radial (RBF), diferentemente MPL, possui apenas uma camada oculta, sendo formada por neurônios com função de ativação de base radial local (HAYKIN, 2008). 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. 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; CARVALHO; LUDEMIR, 2007). Em uma topologia básica, a rede de base radial 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 gaussiana). E, por fim, a camada de saída aplica uma transformação linear no espaço oculto fornecendo uma saída para a rede (HAYKIN, 2008).

Assim como para a MLP, nas redes RBF a escolha da arquitetura a ser utilizada também fica a cargo do pesquisador, assim como 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; 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 que compõem os neurônios da camada intermediária. Essa etapa é realizada com o auxílio de métodos de agrupamento de otimização provenientes da estatística multivariada, como, por exemplo, o método de *k-means*, que tem por objetivo particionar os indivíduos em K grupos, minimizando a distância dos elementos a um conjunto (CRUZ; NASCIMENTO, 2018). O número de grupos corresponde ao número de neurônios que compõem a camada intermediária pode ser definido com base na experiência do pesquisador, tomando como base resultados prévios de uma análise de agrupamento por meio de técnicas hierárquicas ou pelo uso de validação cruzada. 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 MLP, ou seja, supervisionado (HAYKIN, 2008; SILVA; SPATTI; FLAUZINO, 2010).

2.5.3. Árvore de regressão e seus refinamentos

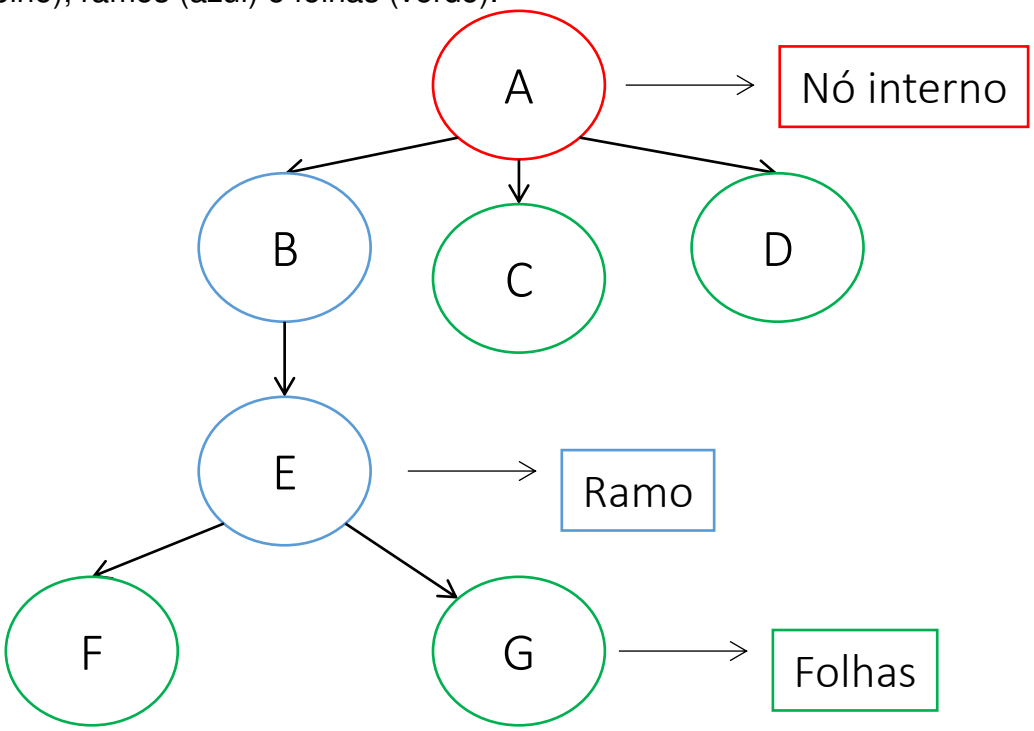
As RNAs apesar de apresentarem eficiência satisfatória, demandam muito recurso computacional. Além disso, a determinação de quais marcadores é relevante para a característica em estudo não é uma tarefa trivial, visto que exige o conhecimento e combinação dos valores dos pesos dos neurônios nas camadas ocultas obtidos após o treinamento da rede. Outra questão é o fato das RNAs não apresentarem solução única (SILVA; SPATTI; FLAUZINO, 2010), ou seja, para o mesmo problema é possível a obtenção de diferentes modelos baseados em RNAs.

Comparada com as RNAs, as árvores de regressão e seus possíveis refinamentos (*Boosting*, *Bagging*, *Random Forest*) demandam menos recurso computacional e apresentam a importância dos marcadores (entrada) de maneira fácil e direta (JAMES et al., 2013; SOUSA et al., 2021).

Assim como as RNAs, as Árvores de Decisão e seus refinamentos não necessitam de pressuposições sobre o modelo (SOUSA et al., 2021). Além do mais, tais metodologias apresentam boa performance preditiva (JAMES et al., 2013), permitindo a não-linearidade dos dados e também são de fácil interpretação (PRASAD; IVERSON; LIAW, 2006), por fornecerem as informações sobre quais atributos são mais importantes para previsão ou classificação (BEIKI; SABOOR; EBRAHIMI, 2012; EBRAHIMI et al., 2011; HOSSEINZADEH et al., 2012).

A Árvore de Decisão (AD) é uma metodologia que particiona o espaço preditor em sub-regiões através de alguns critérios, para cara sub-região formada é atribuído um valor que será utilizado como valor predito para os novos indivíduos que serão alocados a essas sub-regiões (JAMES et al., 2013). A estrutura da AD é composta pelos nós internos, ramos e nós externos/folhas (Figura 2). O nó é dito interno quando os dados contidos neste nó são divididos de acordo com um critério de divisão, formando assim dois novos grupos de dados, sendo esses novos grupos ligados ao grupo antigo pelos ramos, já o nó é dito externo (folha) quando não ocorrem mais divisões dos indivíduos pertencentes a este nó. O aprendizado indutivo de árvores de decisão é geralmente dividido em aprendizado supervisionado e não-supervisionado, embora o aprendizado semi-supervisionado também tem sido considerado ao longo dos últimos anos (CHAPELLE; SCHÖLKOPF; ZIEN, 2006).

Figura 2 - Exemplo de particionamento de uma árvore de decisão, com o nó interno (vermelho), ramos (azul) e folhas (verde).



FONTE: O autor.

De acordo com TAN; STEINBACH; KUMAR (2006), a AD pode ser utilizada para os seguintes propósitos: modelagem descritiva (classificação) e modelagem preditiva (regressão). AD pode ser classificada como árvore de regressão quando a variável resposta é do tipo quantitativa (contínua) ou árvore de classificação quando a variável dependente assume valores qualitativos (categóricos). A AD tem como objetivo subdividir diversas vezes o conjunto de observações de tal forma que os subgrupos formados subsequentes sejam cada vez mais homogêneos (SOUSA et al., 2021). Entretanto, existem diferentes tipos de critérios de seleção, sendo essa uma das variações entre os diversos algoritmos de indução de AD. Esses critérios são definidos em termos da distribuição de classe dos exemplos antes e depois da divisão (TAN; STEINBACH; KUMAR, 2006), como descrito na Equação 9:

$$P(T) = \sum_{m=1}^{M} \sum_{k \in R_m} (y_k - \hat{y}_{R_m})^2 \qquad \text{Equação (9)}$$

em que: $\hat{y}_R$ é a média da variável resposta das observações de treinamento pertencente a m-éssima região. e y_k é o valor verdadeiro da característica de cada indivíduo dentro do grupo k.

Porém, o custo computacional é muito alto sendo inviável considerar cada

partição possível do espaço em M regiões para obter o menor erro quadrático médio. Para contornar o custo computacional, (JAMES et al., 2013) recomendam um procedimento baseado em divisões binárias recursivas, na qual o objetivo é obter a variável X_p e o ponto s, que divida o espaço em duas regiões, como a Equação 10:

$$R_1(p,s) = \{X | X_p \leq s\} \; e \; R_2(p,s) = \{X | X_p > s\} \qquad \text{Equação (10)}$$

em que o ponto s divida a p-ésima variável em duas regiões que obtenha o menor erro quadrático médio, por fim utilizamos a variável que obteve o menor erro quadrático médio para a primeira divisão, em seguida repetimos o processo para cada região gerada.

Quando AD são construídas, muitas das arestas ou sub-árvores podem refletir ruídos ou erros. Enquanto uma árvore muito grande pode sofrer *overfitting* dos dados, uma árvore pequena pode não capturar uma boa estrutura. Para detectar e excluir essas arestas e sub-árvores, são utilizados métodos de poda (*pruning*) da árvore, cujo objetivo é melhorar a taxa de acerto do modelo para novos exemplos, os quais não foram utilizados no conjunto de treinamento (HAN, 2001). Uma abordagem para a escolha do tamanho da árvore seria construir uma árvore até que nenhuma região obtenha mais que 5 indivíduos e, em seguida, podá-la usando o custo complexidade da poda (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Em uma segunda etapa, é realizada a poda com o objetivo de tornar a AD menor e menos complexa, de modo a diminuir a variância deste estimador. Nessa etapa do processo, cada nó é retirado, um por vez, observando-se como o erro de predição varia no conjunto de validação e, posteriormente, baseando-se nas observações, é decidido quais nós permaneceram na árvore (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Geralmente, uma única árvore não possui boa precisão preditiva quando comparada com outras abordagens (SOUSA et al., 2021). Alguns refinamentos com o intuito de melhorar a performance do modelo de AD são apresentados na literatura. O pior desempenho da AD quando comparado com seus refinamentos pode ser explicado porque essa metodologia sofre alta variação em termos de previsão (JAMES et al., 2013). Hastie *et al.* (2009) enfatizaram que a baixa precisão preditiva da Árvore de Decisão pode ser melhorada pelo uso de métodos de *ensemble*, como *Bootstrap Aggregation (Bagging), Random Forest* e *Boosting* (BOEHMKE; GREENWELL, 2019).

2.5.3.1. Bagging

Um dos problemas apresentados pela AD é a grande variabilidade entre os resultados obtidos, ou seja, se utilizarmos uma parte de um banco de dados para construirmos uma árvore e em seguida utilizarmos a outra parte do mesmo banco de dados para construir uma segunda árvore, iremos obter duas árvores com estruturas diferentes (SOUSA et al., 2021). Para contornar esse problema, o ideal seria obter várias amostras de uma mesma população, construir várias árvores e em seguida obter a média/moda dos valores preditos (JAMES et al., 2013).

Como não é uma tarefa fácil obter vários conjuntos de treinamento de uma população, o *Bootstrap Aggregation* (*Bagging*) (BOEHMKE; GREENWELL, 2019) é um método que aplica a técnica de *bootstrap*. O *bootstrap* consiste em obter B amostras com reposição da amostragem disponível, obtendo assim B modelos $\hat{f}^1(x), \hat{f}^2(x), \ldots, \hat{f}^B(x)$ (EFRON, 1992). A amostragem é feita com a substituição dos dados originais e a formação de novos conjuntos de dados, que podem ter uma fração das colunas e das linhas, por fim, utilizam-se os modelos gerados para obter uma média, e assim diminuir a variabilidade obtida nas AD (BOEHMKE; GREENWELL, 2019). Essa média desses modelos irá ser o modelo final, e é dada pela Equação 11:

$$\hat{f}_{médio}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(x) \qquad \text{Equação (11)}$$

Dessa forma, o *Bagging* é uma técnica usada para reduzir a variância das previsões, que combina o resultado de vários classificadores, modelados em diferentes sub-amostras do mesmo conjunto de dados (BOEHMKE; GREENWELL, 2019). Tomando frações de linha e coluna menores que 1 ajuda na montagem de modelos robustos, menos propensos a *overfitting*. A quantidade de árvores utilizadas no *Bagging* não é um parâmetro que irá resultar num super-ajustamento do modelo, na pratica é utilizado uma quantidade onde o erro tenha estabilizado (JAMES et al., 2013).

2.5.3.2. Random Forest

Devido ao fato de sempre utilizarmos todas as variáveis em cada partição no *Bagging*, as predições obtidas estarão altamente correlacionados uma vez que as árvores criadas terão estruturas semelhantes, além disso, está sujeito a uma mesma

variável esteja sempre no topo da árvore (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAMES et al., 2013). A média de valores altamente correlacionados, não resulta na grande redução da variância, como ocorre quando é feita com valores não correlacionados (JAMES et al., 2013). Para melhorar a acurácia na classificação dos indivíduos, Ho (1995) propôs o *Random Forest* (RF). O *Random Forest* é um método de ML versátil e capaz de executar tarefas de regressão e de classificação. Essa metodologia também aplica métodos de redução dimensional, trata valores faltantes, valores anómalos (‘*outliers’*) e outras etapas essenciais da exploração de dados. É um tipo de método de aprendizado onde um grupo de modelos fracos é combinado para formar um modelo mais forte (HO, 1995).

O *Random Forest* segue a mesma ideia do *Bagging*, no entanto, além do conjunto de observações, altera também o número de variáveis preditoras ($m < p$) utilizadas em cada partição. No *Random Forest* obtêm-se os valores preditos mais independentes, o que gera redução da variabilidade. HASTIE; TIBSHIRANI; FRIEDMAN (2009) sugerem que o número de variáveis preditora utilizadas em cada partição seja $m = \sqrt{p}$ para árvore de classificação e $m = p/3$ para árvores de regressão. Assim, as predições das árvores se tornam menos correlacionadas e ainda, corrige o fato de que apenas uma variável esteja sempre no topo da árvore (SOUSA et al., 2021).

2.5.3.3. Boosting

O termo *Boosting* refere-se a uma família de algoritmos que converte uma aprendizagem fraca (também conhecido como base de aprendizagem) em uma aprendizagem forte (SOUSA et al., 2021). Para converter a aprendizagem fraca em aprendizagem forte, a previsão de cada aprendizagem fraca é combinada por métodos que utilizam à média e média ponderada e/ou que consideram a previsão que apresentar mais “votos” (JAMES et al., 2013). Para encontrar uma regra fraca, um processo interativo é utilizado, onde aplicam-se algoritmos de base de aprendizagem com uma distribuição diferente, gerando uma nova regra de previsão fraca, que após muitas iterações, o *Bossting* combina essas regras fracas em uma única regra de predição forte (MARTINS et al., 2009).

Ao contrário do *Bagging* e *Random Forest* que cria múltiplas árvores

independentes, o *Boosting* cria árvores sequencialmente utilizando-se de informação prévia da árvore anterior. Ao invés de ajustar um modelo para a variável resposta Y, o *Boosting* ajusta um grande número de árvores de decisão, $\hat{f}^1, \hat{f}^2, \ldots, \hat{f}^B$, para o resíduo atual (FREUND; SCHAPIRE, 1999). Nessa metodologia a aprendizagem é lenta, necessitando assim que o número de modelos (B) seja grande. Entretanto, é necessário ter cuidado para criar um *overfitting* do modelo. Assim, no *Boosting* é utilizada a validação cruzada para se escolher o número de árvores que será construída, isso diminuiu a possibilidade de *overfitting*, uma vez que todos os indivíduos participarão do conjunto de validação (BENGIO; GRANDVALET, 2004).

2.5.4. Splines de Regressão Adaptativa Multivariada

A *Splines* de Regressão Adaptativa Multivariada (*Multivariate Adaptive Regression Splines* - MARS) (FRIEDMAN, 1991) fornece uma abordagem conveniente para capturar o aspecto de não-linearidade da regressão polinomial, avaliando pontos de corte (nós). Um *spline* é uma curva definida matematicamente por dois ou mais pontos de controle, onde os pontos de controle que ficam na curva são chamados de nós e os demais pontos definem a tangente à curva em seus respectivos nós (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Como as redes neurais, a MARS usa recursos substitutos em vez dos preditores originais, além disso, é um procedimento adaptativo para regressão e é adequado para problemas de alta dimensão (ou seja, um grande número de entradas) (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). No entanto, enquanto as redes neurais são baseadas em combinações lineares dos preditores, a MARS cria duas versões contrastadas de um preditor para entrar no modelo. Os recursos substitutos na MARS geralmente são uma função de apenas um ou dois preditores por vez. A natureza dos recursos da MARS divide o preditor em dois grupos e modela as relações lineares entre o preditor e o resultado em cada grupo, especificando um dado ponto de corte para um preditor, criando dois novos recursos , que são as funções de “dobradiça” (KUHN; JOHNSON, 2013).

A MARS é um procedimento para ajustar a regressão não linear adaptativa que usa funções básicas por partes para definir relações entre uma variável de resposta e algum conjunto de preditores (FRIEDMAN, 1991). As funções de base são definidas

em pares, usando um nó ou valor de uma variável que define um ponto de inflexão ao longo do intervalo de um preditor (LEATHWICK; ELITH; HASTIE, 2006). De acordo com Hastie *et al.* (2009), a primeira etapa (*forward*) seleciona todos os pontos de separação possíveis e determina as funções de base; a segunda (*backward*) fase elimina as funções de base redundantes de uma das seguintes formas:

$$(x - t)_+ = \begin{cases} x - t, \text{se } x > t \\ 0, \text{caso contrário} \end{cases};$$

$$(t - x)_+ = \begin{cases} t - x, \text{se } x < t \\ 0, \text{caso contrário} \end{cases}.$$

Estas duas funções de base apresentadas são ditas um par reflexivo. Considerando o conjunto dos pontos observados para a variável $X_j, \Omega = \{x_{1j}, x_{2j}, \dots, x_{rj}\}$, obtemos uma coleção de funções base $\mathcal{C} = \{(x - t)_+, (t - x)_+\}_{\substack{t \in \Omega \\ j=1,2,\dots,p}}$.

Mais de um nó (isto é, mais de um par de funções básicas) pode ser especificado para uma variável preditora, permitindo que relações complexas não lineares sejam ajustadas. Alternativamente, as funções básicas podem ser consideradas como uma nova matriz preditora, na qual uma ou mais colunas que são funções básicas substituem cada variável preditora nos dados originais (LEATHWICK; ELITH; HASTIE, 2006).

Ao ajustar um modelo MARS, os nós são escolhidos automaticamente de maneira progressiva. Os nós candidatos podem ser colocados em qualquer posição dentro do intervalo de cada variável preditora para definir um par de funções básicas. Em cada etapa, o modelo seleciona o nó e seu par correspondente de funções básicas que dão a maior diminuição na soma de quadrados do resíduo (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). A seleção de nós continua até que algum tamanho máximo do modelo seja atingido, após o qual é aplicado um procedimento de remoção para trás (*backward*), no qual as funções básicas que contribuem menos para o ajuste do modelo são removidas progressivamente. Nesse estágio, uma variável preditora pode ser removida completamente do modelo se nenhuma de suas funções básicas contribuírem significativamente para o desempenho preditivo (LEATHWICK; ELITH; HASTIE, 2006).

A sequência de modelos gerados a partir desse processo é então avaliada usando validação cruzada generalizada e o modelo com o melhor ajuste preditivo é selecionado. Entretanto, um problema comum quando se utiliza a MARS é a facilidade

de apresentar *overfits* (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Nesse sentido, a fase *backward* é realizada para resolver este problema eliminando termos que causam menor incremento no erro quadrático do resíduo.

A MARS foi desenvolvida para automatizar todos os aspectos da modelagem de regressão, lidando com grandes quantidades de variáveis aparentemente não relacionadas (YORK; EAVES, 2001). Segundo Lin *et al.* (2008), a MARS é uma ferramenta automatizada e flexível de mineração de dados que combina as vantagens do particionamento recursivo e ajuste de *splines*. A MARS é um método de regressão não paramétrico baseado numa partição dos dados de treinamento e modelado em funções lineares (ZHENG et al., 2019) e como as RNAs e ADs, também não faz nenhuma pressuposição sobre a distribuição das variáveis preditoras (MOTSINGER; RITCHIE; REIF, 2007).

Essa abordagem combina as forças das ADs e o ajuste das *splines*, substituindo as funções de etapa normalmente associadas às árvores de regressão por funções de base linear por partes (LEATHWICK; ELITH; HASTIE, 2006). Isso permite a modelagem de relacionamentos complexos entre uma variável de resposta e seus preditores. Dessa forma, diferentemente de modelos baseados em inteligência artificial, como as RNAs e ADs, a MARS modela automaticamente não linearidades e interações entre as variáveis de entrada (TAYLAN; WEBER, 2019; ZHENG et al., 2019) e apresenta ao final do processo, um modelo ajustado, ou seja, apresenta como os efeitos são determinados na característica em estudo. Neste sentido, a MARS possivelmente aumenta a acurácia de predição para características que apresentem diferentes tipos de efeitos não aditivos complexos, permitindo ao pesquisador obter informações sobre a arquitetura genética da característica.

Segundo York e Eaves (2001), especificamente, a MARS tem quatro vantagens principais: (1) capacidade de separar preditores relevantes de irrelevantes de um grande número de variáveis independentes; (2) qualquer variável preditora não linear incluída no modelo é automaticamente transformada (na forma de funções básicas de *spline*) em relação à variável de resultado; (3) todas as interações possíveis são testadas em uma ordem definida pelo usuário e certas combinações de interação variável também podem ser proibidas; (4) a validação cruzada especificada pelo usuário protege contra o ajuste excessivo do modelo. O resultado é um algoritmo não paramétrico confiável que não requer conhecimento a priori da forma do

relacionamento entre o resultado e as variáveis preditoras (York; Eaves, 2001).

Uma das grandes vantagens de se utilizar a MARS é sua ação local, explicando o fenômeno em pedaços do intervalo e sendo zero nas demais partes. Outro fator importante da MARS é seu modelo hierárquico, isto é, para se possuir partes com interações entre duas ou mais funções base, é necessário que o modelo compreenda pelo menos uma das funções base em sua construção (YORK; EAVES; VAN DEN OORD, 2006).

A MARS fornece recursos úteis para superar as limitações na exploração de interações SNP – SNP. Segundo Lin *et al.* (2008) alguns desses recursos são: seleção flexível de grupos de referência (blocos gênicos), combinação automática de marcadores e detecção automática de padrões de interação. Além disso, esses autores relatam que a MARS pode incluir vários termos (efeitos principais e interações) em um modelo simultaneamente, e as interações genéticas podem ser avaliadas após o ajuste. Assim, a MARS além de detectar qual SNP está envolvido em uma interação, o modo de herança (dominante, recessivo e aditivo) para os SNPs também pode ser determinado automaticamente (Lin *et al.*, 2008). Dessa forma, o número de parâmetros na modelagem é reduzido drasticamente. Com isso, a MARS não precisa de um longo processo de treinamento e, portanto, pode economizar muito tempo de construção do modelo, especialmente quando o conjunto de dados é enorme (LEE; CHEN, 2005). Como já mencionado, nos estudos de GWS a dimensionalidade é um dos grandes problemas enfrentados nas análises, assim esses recursos da MARS se mostram úteis.

Além desses aspectos, os modelos MARS oferecem vantagens ao trabalhar com efeito aditivo, bem como nos efeitos de interação. Assim, eles podem revelar-se muito úteis em análises exploratórias de interações gene-gene e gene-ambiente (COOK; ZEE; RIDKER, 2004; MOTSINGER; RITCHIE; REIF, 2007). A MARS é mais poderosa do que o ajuste da curva de mínimos quadrados usando polinômios no teste de interações gene-ambiente (YORK; EAVES; VAN DEN OORD, 2006). Isso demonstra que essa metodologia pode ser eficiente em estudos de GWS, permitindo ao pesquisador conhecer os efeitos de uma característica cujo controle gênico é dado por interações entre genes (epistasia).

3. REFERÊNCIAS BIBLIOGRÁFICAS

ACQUAAH, George. **Priciples of Plant Genetics Breeding**. Blackwell: Oxford, 2012. v. 53 DOI: 10.1017/CBO9781107415324.004.

ALKIMIM, Emilly Ruas; CAIXETA, Eveline Teixeira; SOUSA, Tiago Vieira; RESENDE, Marcos Deon Vilela; SILVA, Felipe Lopes; SAKIYAMA, Ney Sussumu; ZAMBOLIM, Laércio. Selective efficiency of genome-wide selection in Coffea canephora breeding. **Tree Genetics and Genomes**, *[S. l.]*, v. 16, n. 3, 2020. DOI: 10.1007/s11295-020-01433-3.

ALTINOK, Gulsah; KARAGOZ, Pinar; BATMAZ, Inci. Learning to rank by using multivariate adaptive regression splines and conic multivariate adaptive regression splines. **Computational Intelligence**, *[S. l.]*, n. April, p. 1–38, 2020. DOI: 10.1111/coin.12413.

ALVES, Anderson Antonio Carvalho; COSTA, Rebeka Magalhães; BRESOLIN, Tiago; FERNANDES JÚNIOR, Gerardo Alves; ESPIGOLAN, Rafael; RIBEIRO, André Mauric Frossard; CARVALHEIRO, Roberto; ALBUQUERQUE, Lucia Galvão. Genome-wide prediction for complex traits under the presence of dominance effects in simulated populations using GBLUP and machine learning methods. **American Society of Animal Science**, *[S. l.]*, p. 1–34, 2020.

AZODI, Christina B.; BOLGER, Emily; MCCARREN, Andrew; ROANTREE, Mark; DE LOS CAMPOS, Gustavo; SHIU, Shin Han. Benchmarking parametric and machine learning models for genomic prediction of complex traits. **G3: Genes, Genomes, Genetics**, *[S. l.]*, v. 9, n. 11, p. 3691–3702, 2019. DOI: 10.1534/g3.119.400498.

BARBOSA, Ivan Paiva; SILVA, Michele Jorge; COSTA, Weverton Gomes; SANT'ANNA, Isabela Castro; NASCIMENTO, Moysés; CRUZ, Cosme Damião. Genome-enabled prediction through machine learning methods considering different levels of trait complexity. **Crop Science**, *[S. l.]*, v. 61, n. 3, p. 1890–1902, 2021. DOI: 10.1002/csc2.20488.

BEIKI, Amir H.; SABOOR, Saba; EBRAHIMI, Mansour. A New Avenue for Classification and Prediction of Olive Cultivars Using Supervised and Unsupervised Algorithms. **PLOS ONE**, *[S. l.]*, v. 7, n. 9, p. 1–9, 2012. DOI: 10.1371/journal.pone.0044164.

BENGIO, Yoshua; GRANDVALET, Yves. No Unbiased Estimator of the Variance of K-Fold Cross-Validation. **Journal of Machine Learning Research**, *[S. l.]*, v. 5, p. 1089–1105, 2004. DOI: 10.1016/S0006-291X(03)00224-9.

BERED, Fernanda; BARBOSA NETO, José Fernandes; CARVALHO, Fernando Irajá Felix. Marcadores moleculares e sua aplicação no melhoramento genético de plantas. **Ciência Rural**, *[S. l.]*, v. 27, n. 3, p. 513–520, 1997. Disponível em: https://www.lume.ufrgs.br/bitstream/handle/10183/22565/000211786.pdf?sequence=1.

BERNARDO, R. **Breeding for quantitative traits in plants**. 3rd edn. ed. Minnesota: Stemma Press, Woodbury, 2010.

BERNARDO, Rex. Reinventing quantitative genetics for plant breeding: something old, something new, something borrowed, something BLUE. **Hereditary**, *[S. l.]*, p. 11, 2020. DOI: 10.1038/s41437-020-0312-1. Disponível em: http://dx.doi.org/10.1038/s41437-020-0312-1.

BERRY, Elliot M.; DERNINI, Sandro; BURLINGAME, Barbara; MEYBECK, Alexandre; CONFORTI, Piero. Food security and sustainability: Can one exist without the other? **Public Health Nutrition**, *[S. l.]*, v. 18, n. 13, p. 2293–2302, 2015. DOI: 10.1017/S136898001500021X.

BITTENCOURT, G. **Inteligência artificial: ferramentas e teorias**. 3. ed. ed. Florianópolis: Editora UFSC, 2006.

BOEHMKE, Brad; GREENWELL, Brandon. Random Forests. *In*: **Hands-On Machine Learning with R**. [s.l.] : Chapman and Hall/CRC, 2019. v. 45p. 203–219. DOI: 10.1201/9780367816377-11. Disponível em: https://www.taylorfrancis.com/books/9781000730197/chapters/10.1201/9780367816377-11.

BORSELLINO, Valeria; SCHIMMENTI, Emanuele; EL BILALI, Hamid. Agri-food markets towards sustainable patterns. **Sustainability (Switzerland)**, *[S. l.]*, v. 12, n. 6, 2020. DOI: 10.3390/su12062193.

BRAGA, A. P.; CARVALHO, A. C. P. L. ..; LUDEMIR, T. B. **Redes neurais artificiais: teoria e aplicações**. 2.ed. ed. Rio de Janeiro: LTC, 2007.
BREIMAN, Leo. Bagging Predictors. **Machine Learning**, *[S. l.]*, v. 24, n. 421, p. 123–140, 1996. DOI: 10.1007/BF00058655.

BUENO, L. C. S.; MENDES, A. N. G.; CARVALHO, S. P. **Melhoramento Genético de Plantas: Princípios e Procedimentos**. 2. ed. ed. Lavras: UFLA, 2006.

CAIXETA, Eveline Teixeira; OLIVEIRA, Antonio Carlos Baião; BRITO, Giovani Greigh; SAKIYAMA, Ney Sussumu. Tipos de Marcadores Moleculares. *In*: BORÉM, Aluízio; CAIXETA, Eveline Teixeira (org.). **Marcadores Moleculares**. Viçosa, MG: Editora UFV, 2016. p. 385.

CAPONE, Roberto; EL BILALI, Hamid; DEBS, Philipp; CARDONE, Gianluigi; DRIOUECH, Noureddin. Food System Sustainability and Food Security: Connecting the Dots. **Journal of Food Security**, *[S. l.]*, v. 2, n. 1, p. 13–22, 2014. DOI: 10.12691/jfs-2-1-2. Disponível em: http://pubs.sciepub.com/jfs/2/1/2/jfs-2-1-2.pdf.

CARENA, Marcelo J.; HALLAUER, Arnel R.; MIRANDA FILHO, J. B. **Quantitative Genetics in Maize Breeding**. New York, NY: Springer New York, 2010. v. 53 DOI: 10.1007/978-1-4419-0766-0. Disponível em: http://link.springer.com/10.1007/978-1-4419-0766-0.

CHAN, L. W.; FALLSIDE, F. An adaptive training algorithm for back propagation networks. **Computer Speech and Language**, *[S. l.]*, v. 2, n. 3–4, p. 205–218, 1987. DOI: 10.1016/0885-2308(87)90009-X.

CHANG, Cheng Ding; WANG, Chien Chih; JIANG, Bernard C. Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors. **Expert Systems with Applications**, *[S. l.]*, v. 38, n. 5, p. 5507–5513, 2011. DOI: 10.1016/j.eswa.2010.10.086. Disponível em: http://dx.doi.org/10.1016/j.eswa.2010.10.086.

CHAPELLE, Olivier; SCHÖLKOPF, Bernhard; ZIEN, Alexander. **Semi-Supervised Learning**. London, England: Massachusetts Institute of Technology Press., 2006. DOI: 10.1109/ICPR.2018.8546327.

COOK, Nancy R.; ZEE, Robert Y. L.; RIDKER, Paul M. Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. **Statistics in Medicine**, *[S. l.]*, v. 23, n. 9, p. 1439–1453, 2004. DOI: 10.1002/sim.1749.

COPPIN, B. **Inteligência Artificial**. Rio de Janeiro, RJ: LTC, 2010.

COSTA, Weverton Gomes; BARBOSA, Ivan de Paiva; SOUZA, Jacqueline Enequio; CRUZ, Cosme Damião; NASCIMENTO, Moysés; OLIVEIRA, Antonio Carlos Baião. Machine learning and statistics to qualify environments through multi-traits in Coffea arabica. **PLOS ONE**, *[S. l.]*, v. 16, n. 1, p. 1–21, 2021. DOI: 10.1371/journal.pone.0245298. Disponível em: https://dx.plos.org/10.1371/journal.pone.0245298.

COSTER, Albart; BASTIAANSEN, John W. M.; CALUS, Mario P. L.; VAN ARENDONK, Johan A. M.; BOVENHUIS, Henk. Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. **Genetics Selection Evolution**, *[S. l.]*, v. 42, n. 1, p. 1–11, 2010. DOI: 10.1186/1297-9686-42-9.

COUTINHO, Alisson Esdras; NEDER, Diogo Gonçalves; SILVA, Mairykon Coêlho; ARCELINO, Eliane Cristina; BRITO, Silvan Gomes; CARVALHO FILHO, José Luiz Sandes. Prediction of phenotypic and genotypic values by BLUP/GWS and neural networks. **Revista Caatinga**, *[S. l.]*, v. 31, n. 3, p. 532–540, 2018. DOI: 10.1590/1983-21252018v31n301rc. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1983-21252018000300532&lng=en&tlng=en.

CRUZ, Cosme Damião. **Princípios de genética quantitativa**. 2. ed. ed. Viçosa, MG: Editora UFV, 2012.

CRUZ, Cosme Damião. GENES - Software para análise de dados em estatística experimental e em genética quantitativa. **Acta Scientiarum - Agronomy**, *[S. l.]*, v. 35, n. 3, p. 271–276, 2013. DOI: 10.4025/actasciagron.v35i3.21251.

CRUZ, Cosme Damião. Genes software – extended and integrated with the R, Matlab and Selegen. **Acta Scientiarum - Agronomy**, *[S. l.]*, v. 38, n. 4, p. 547–552, 2016. DOI: 10.4025/actasciagron.v38i4.32629.

CRUZ, Cosme Damião; NASCIMENTO, Moysés. **Inteligência Computacional aplicada ao melhoramento genético**. Viçosa, MG: Editora UFV, 2018.

CRUZ, Cosme Damião; NASCIMENTO, Moysés; SILVA, Gabi Nunes; ROSADO, Renato Dominiciano Silva. RNA - Perceptron Multicamadas. *In*: CRUZ, Cosme Damião; NASCIMENTO, Moysés (org.). **Inteligência computacional aplicada ao melhoramento genético**. Viçosa, MG, Brazil: Editora UFV, 2018. p. 151–189.

CRUZ, Cosme Damião; REGAZZI, Adair José; CARNEIRO, Pedro Crescêncio Souza. **Modelos biométricos aplicados ao melhoramento genético**. Viçosa: UFV, 2012.

CRUZ, Cosme Damião; REGAZZI, Adair José; CARNEIRO, Pedro Crescêncio Souza; REGAZZI, Adair José. **Modelos biométricos aplicados ao melhoramento genético**. 3. ed. v2 ed. Viçosa, MG: UFV, 2014.

CRUZ, Cosme Damião; SALGADO, Caio César; BHERING, Leonardo Lopes. **Genômica Aplicada**. Visconde do Rio Branco, MG: Suprema, 2013.

DE LOS CAMPOS, Gustavo; NAYA, Hugo; GIANOLA, Daniel; CROSSA, José; LEGARRA, Andrés; MANFREDI, Eduardo; WEIGEL, Kent; COTES, José Miguel. Predicting quantitative traits with regression models for dense molecular markers and pedigree. **Genetics**, *[S. l.]*, v. 182, n. 1, p. 375–385, 2009. DOI: 10.1534/genetics.109.101501.

DE VEAUX, Richard D.; UNGAR, Lyle H. Multicollinearity: A tale of two nonparametric regressions. *[S. l.]*, p. 393–402, 1994. DOI: 10.1007/978-1-4612-2660-4_40.

DESTA, Zeratsion Abera; ORTIZ, Rodomiro. Genomic selection: Genome-wide prediction in plant improvement. **Trends in Plant Science**, *[S. l.]*, v. 19, n. 9, p. 592–601, 2014. DOI: 10.1016/j.tplants.2014.05.006. Disponível em: http://dx.doi.org/10.1016/j.tplants.2014.05.006.

DIAZ-URIARTE, Ramón. GeneSrF and varSelRF: A web-based tool and R package for gene selection and classification using random forest. **BMC Bioinformatics**, *[S. l.]*, v. 8, p. 1–7, 2007. DOI: 10.1186/1471-2105-8-328.

DUDLEY, J. W.; MOLL, R. H. Interpretation and Use of Estimates of Heritability and Genetic Variances in Plant Breeding 1 . **Crop Science**, *[S. l.]*, v. 9, n. 3, p. 257–262, 1969. DOI: 10.2135/cropsci1969.0011183x000900030001x.

EBRAHIMI, Mansour; LAKIZADEH, Amir; AGHA-GOLZADEH, Parisa; EBRAHIMIE, Esmaeil; EBRAHIMI, Mahdi. Prediction of thermostability from amino acid attributes by combination of clustering with attribute weighting: A new vista in engineering enzymes. **PLoS ONE**, *[S. l.]*, v. 6, n. 8, 2011. DOI: 10.1371/journal.pone.0023146.

EFRON, Bradley. Bootstrap Methods: Another Look at the Jackknife. *In*: KOTZ, Samuel; JOHNSON, Norman L. (org.). **Breakthroughs in Statistics Volume II Methodology and Distribution**. New York, NY: Springer Series in Statistics (Perspectives in Statistics), 1992. p. 569–595.

ENTRINGER, Geovana Cremonini; CESAR, Julio; VETTORAZZI, Fiorio; PEREIRA, Messias Gonzaga. Correlação e análise de trilha para componentes de produção de milho superdoce. **Revista Ceres**, *[S. l.]*, v. 61, n. 3, p. 356–361, 2014. DOI: http://dx.doi.org/10.1590/S0034-737X2014000300009.

EVERINGHAM, Y. L.; SEXTON, J. An introduction to Multivariate Adaptive Regression Splines for the cane industry. **33rd Annual Conference of the Australian Society of Sugar Cane Technologists 2011, ASSCT 2011**, *[S. l.]*, n. December 2015, p. 255–268, 2011.

FALCONER, S. D.; MACKAY, T. F. C. **Introduction to quantitative genetics**. [s.l.] : Edinburgh, Addison Wesley Longman, 1996.

FERNANDES, Anita Maria da Rocha. **Inteligência Artificial - Noções Gerais**. [s.l.] : Visual Books, 2003.

FERNANDO, R. L.; CHENG, H.; SUN, X.; GARRICK, D. J. A comparison of identity-by-descent and identity-by-state matrices that are used for genetic evaluation and estimation of variance components. **Journal of Animal Breeding and Genetics**, *[S. l.]*, v. 134, n. 3, p. 213–223, 2017. DOI: 10.1111/jbg.12275.

FERREIRA, R. A. D. C.; SILVA, Gabi Nunes; GLÓRIA, Leonardo Siqueira; SANT'ANNA, Isabela Castro; RODRIGUES, Haroldo Silva; SILVA, Fabyano Fonseca; CRUZ, Cosme Damião. RNA - Aplicação em Estudos de Seleção Genômica Ampla. *In*: CRUZ, Cosme Damião; NASCIMENTO, Moysés (org.). **Inteligência Computacional Aplicado ao Melhoramento Genético**. Viçosa, MG: Editora UFV, 2018. p. 414.

FISHER, R. A. Average excess and average effect of a gene substitution. **Ann Eugen**, *[S. l.]*, v. 11, p. 53–63, 1941.

FREUND, Y.; SCHAPIRE, Robert E. A brief introduction to boosting. **International Joint Conference on Artificial Intelligence**, *[S. l.]*, v. 2, n. 5, p. 1401–1406, 1999.

FRIEDMAN, Jerome H. Multivariate Adaptative regression Splines. **The Annals of Statistics**, *[S. l.]*, v. 19, n. 1, p. 1–141, 1991. Disponível em: http://projecteuclid.org/euclid.aop/1176996548.

FRITSCHE-NETO, R. **Seleção genômica ampla e novos métodos de melhoramento do milho**. 2011. Universidade Federal de Viçosa, Viçosa, *[S. l.]*, 2011.

FULEKY, Peter. **Macroeconomic Forecasting in the Era of Big Data**. [s.l: s.n.]. v. 52 Disponível em: http://link.springer.com/10.1007/978-3-030-31150-6.

GHAFOURI-KESBI, Farhad; RAHIMI-MIANJI, Ghodratollah; HONARVAR, Mahmood; NEJATI-JAVAREMI, Ardeshir. Predictive ability of Random Forests, Boosting, Support Vector Machines and Genomic Best Linear Unbiased Prediction in different scenarios of genomic evaluation. **Animal Production Science**, *[S. l.]*, v. 57, n. 2, p. 229, 2017. a. DOI: 10.1071/AN15538. Disponível em: http://www.publish.csiro.au/?paper=AN15538.

GHAFOURI-KESBI, Farhad; RAHIMI-MIANJI, Ghodratollah; HONARVAR, Mahmood; NEJATI-JAVAREMI, Ardeshir. Predictive ability of Random Forests, Boosting, Support Vector Machines and Genomic Best Linear Unbiased Prediction in different scenarios of genomic evaluation. **Animal Production Science**, *[S. l.]*, v. 57, n. 2, p. 229–236, 2017. b. DOI: 10.1071/AN15538.

GLÓRIA, Leonardo Siqueira; CRUZ, Cosme Damião; VIEIRA, Ricardo Augusto Mendonça; DE RESENDE, Marcos Deon Vilela; LOPES, Paulo Sávio; DE SIQUEIRA, Otávio H. G. B. Dia.; FONSECA E SILVA, Fabyano. Accessing marker effects and heritability estimates from genome prediction by Bayesian regularized neural networks. **Livestock Science**, *[S. l.]*, v. 191, p. 91–96, 2016. DOI: 10.1016/j.livsci.2016.07.015. Disponível em: http://dx.doi.org/10.1016/j.livsci.2016.07.015.

GODDARD, M. E.; HAYES, B. J. Genomic selection. **J. Anim. Breed. Genet.**, *[S. l.]*, v. 124, p. 323–330, 2007. DOI: 10.1007/978-81-322-2316-0_10.

GONZÁLEZ-CAMACHO, J. M.; DE LOS CAMPOS, G.; PÉREZ, P.; GIANOLA, D.; CAIRNS, J. E.; MAHUKU, G.; BABU, R.; CROSSA, J. Genome-enabled prediction of genetic values using radial basis function neural networks. **Theoretical and Applied Genetics**, *[S. l.]*, v. 125, n. 4, p. 759–771, 2012. DOI: 10.1007/s00122-012-1868-9.

HABIER, David; FERNANDO, Rohan L.; KIZILKAYA, Kadir; GARRICK, Dorian J. Extension of the bayesian alphabet for genomic selection. **BMC Bioinformatics**, *[S. l.]*, v. 12, 2011. DOI: 10.1186/1471-2105-12-186.

HAN, J. **Data Mining: Concepts and Techniques**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The elements of statistical learning: Data mining, inference, and prediction**. 2. ed. ed. New York, NY, USA: Springer, 2009. DOI: 10.1007/978-1-4419-9863-7_941.

HAYES, B. J.; BOWMAN, P. J.; CHAMBERLAIN, A. J.; GODDARD, M. E. Invited review: Genomic selection in dairy cattle: Progress and challenges. **Journal of Dairy Science**, *[S. l.]*, v. 92, n. 2, p. 433–443, 2009. DOI: 10.3168/jds.2008-1646. Disponível em: http://dx.doi.org/10.3168/jds.2008-1646.

HAYKIN, S. **Neural Networks and Learnig Machines**. 3. ed. ed. New York: Prentice Hall, 2008.

HAYKIN, S. S. **Redes Neurais: Princípios e Práticas**. 2 ed. ed. Porto Alegre: Bookman, 2001.

HO, Tin Kam. Random Decision Forests. **Proceedings of 3rd International Conference on Document Analysis and Recognition**, *[S. l.]*, p. 278–282, 1995. DOI: 10.1109/ICDAR.1995.598994. Disponível em: https://ieeexplore.ieee.org/abstract/document/598994/.

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

HOSPITAL, F.; MOREAU, L.; LACOUDRE, F.; CHARCOSSET, A.; GALLAIS, A. More on the efficiency of marker-assisted selection. **Theoretical and Applied Genetics**, *[S. l.]*, v. 95, n. 8, p. 1181–1189, 1997. DOI: 10.1007/s001220050679.

HOSSEINZADEH, Faezeh; EBRAHIMI, Mansour; GOLIAEI, Bahram; SHAMABADI, Narges. Classification of lung cancer tumors based on structural and physicochemical properties of proteins by bioinformatics models. **PLoS ONE**, *[S. l.]*, v. 7, n. 7, 2012. DOI: 10.1371/journal.pone.0040017.

HUANG, Xuehui; HAN, Bin. Natural Variations and Genome-Wide Association Studies in Crop Plants. **Annual Review of Plant Biology**, *[S. l.]*, v. 65, n. 1, p. 531–551, 2014. DOI: 10.1146/annurev-arplant-050213-035715.

JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert. **An Introduction to Statistical Learning with Applications in R**. 1. ed. ed. New York, NY, USA: Springer, 2013. DOI: 10.1007/978-1-4614-7138-7_8.

JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert. An Introduction to Statistical Learning. *In*: **Springer Texts in Statistics**. [s.l: s.n.]. p. 612. DOI: 10.1007/978-1-0716-1418-1_1. Disponível em: https://link.springer.com/10.1007/978-1-0716-1418-1_1.

JHA, S.; TRIPATHI, S. K.; SINGH, R.; DIKSHIT, A.; PANDEY, A. Global Scenario of Natural Products for Sustainable Agriculture. *In*: **Natural Bioactive Products in Sustainable Agriculture**. Singapore: Springer, 2020. p. 1–14.

KEARSEY, M. J.; FARQUHAR, A. G. L. QTL analysis in plants; where are we now? **Heredity**, *[S. l.]*, v. 80, n. 2, p. 137–142, 1998. DOI: 10.1038/sj.hdy.6885001.

KITETU, Geoffrey Musyoki; KO, Jong-Hwan. Climate Change on Agriculture in 2050 : A CGE Approach. *In*: 23RD ANNUAL CONFERENCE ON GLOBAL ECONOMIC ANALYSIS (VIRTUAL CONFERENCE). 2020, Purdue University, West Lafayette. **Anais** [...]. Purdue University, West Lafayette: Global Trade Analysis Project (GTAP), 2020. p. 1–23. Disponível em: https://www.gtap.agecon.purdue.edu/resources/res_display.asp?RecordID=6065.

KUHN, Max; JOHNSON, Kjell. **Applied predictive modeling**. New York, NY, USA: Springer Science+Business Media LLC, 2013. DOI: 10.1007/978-1-4614-6849-3.

LEATHWICK, J. R.; ELITH, J.; HASTIE, T. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. **Ecological Modelling**, *[S. l.]*, v. 199, n. 2, p. 188–196, 2006. DOI: 10.1016/j.ecolmodel.2006.05.022. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0304380006002572.

LEE, Tian Shyug; CHEN, I. Fei. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. **Expert Systems with Applications**, *[S. l.]*, v. 28, n. 4, p. 743–752, 2005. DOI: 10.1016/j.eswa.2004.12.031.

LEGARRA, Andres. Comparing estimates of genetic variance across different relationship models. **Theoretical Population Biology**, *[S. l.]*, v. 107, p. 26–30, 2016. DOI: 10.1016/j.tpb.2015.08.005. Disponível em: http://dx.doi.org/10.1016/j.tpb.2015.08.005.

LEGARRA, Andrés; ROBERT-GRANIÉ, Christèle; MANFREDI, Eduardo; ELSEN, Jean Michel. Performance of genomic selection in mice. **Genetics**, *[S. l.]*, v. 180, n. 1, p. 611–618, 2008. DOI: 10.1534/genetics.108.088575.

LI, Bo; ZHANG, Nanxi; WANG, You Gan; GEORGE, Andrew W.; REVERTER, Antonio; LI, Yutao. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. **Frontiers in Genetics**, *[S. l.]*, v. 9, n. JUL, p. 1–20, 2018. DOI: 10.3389/fgene.2018.00237.

LIEW, Bernard X. W.; PEOLSSON, Anneli; RUGAMER, David; WIBAULT, Johanna; LÖFGREN, Hakan; DEDERING, Asa; ZSIGMOND, Peter; FALLA, Deborah. Clinical predictive modelling of post-surgical recovery in individuals with cervical radiculopathy: a machine learning approach. **Scientific Reports**, *[S. l.]*, v. 10, n. 1, p. 1–10, 2020. DOI: 10.1038/s41598-020-73740-7. Disponível em: https://doi.org/10.1038/s41598-020-73740-7.

LIN, Hui Yi; WANG, Wenquan; LIU, Yung Hsin; SOONG, Seng Jaw; YORK, Timothy P.; MYERS, Leann; HU, Jennifer J. Comparison of multivariate adaptive regression splines and logistic regression in detecting SNP-SNP interactions and their application in prostate cancer. **Journal of Human Genetics**, *[S. l.]*, v. 53, n. 9, p. 802–811, 2008. DOI: 10.1007/s10038-008-0313-z.

MA, Wenlong; QIU, Zhixu; SONG, Jie; CHENG, Qian; MA, Chuang. DeepGS: Predicting phenotypes from genotypes using Deep Learning. **bioRxiv**, *[S. l.]*, 2017. DOI: 10.1101/241414.

MARTINS, Ricardo; PINA, Pedro; MARQUES, Jorge S.; SILVEIRA, Margarida. Crater detection by a boosting approach. **IEEE Geoscience and Remote Sensing Letters**, *[S. l.]*, v. 6, n. 1, p. 127–131, 2009. DOI: 10.1109/LGRS.2008.2006004.

MATHEW, Boby; LÉON, Jens; SILLANPÄÄ, Mikko J. A novel linkage-disequilibrium corrected genomic relationship matrix for SNP-heritability estimation and genomic

prediction. **Heredity**, *[S. l.]*, v. 120, n. 4, p. 356–368, 2018. DOI: 10.1038/s41437-017-0023-4. Disponível em: http://dx.doi.org/10.1038/s41437-017-0023-4.

MATLAB. **Natick, Massachusetts: The MathWorks Inc.**Natick, MassachusettsThe MathWorks Inc., , 2019.

MCCULLOCH, WARREN S.; PITTS, WALTER. A logical calculus of the ideas immanent in nervous activity. **Bulletin of Mathematical Biophysics**, *[S. l.]*, v. 5, p. 115–133, 1943.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, *[S. l.]*, v. 157, n. 4, p. 1819–1829, 2001.

MILBORROW, Stephen. Notes on the earth package. *[S. l.]*, p. 1–68, 2019. Disponível em: http://www.milbo.org/doc/earth-notes.pdf%0Ahttps://cran.r-project.org/web/packages/earth/index.html.

MOTSINGER, Alison A.; RITCHIE, Marylyn D.; REIF, David M. Novel methods for detecting epistasis in pharmacogenomics studies. **Pharmacogenomics**, *[S. l.]*, v. 8, n. 9, p. 1229–1241, 2007. DOI: 10.2217/14622416.8.9.1229.

MOURA, Ernandes Guedes; PAMPLONA, Andrezza Kellen Alves; BALESTRE, Marcio. Functional models in genome-wide selection. **Plos One**, *[S. l.]*, v. 14, n. 10, p. e0222699, 2019. DOI: 10.1371/journal.pone.0222699. Disponível em: https://dx.plos.org/10.1371/journal.pone.0222699.

MOURA, Monique Maculan; CARNEIRO, Pedro Crescêncio Souza; CARNEIRO, José Eustáquio de Souza; CRUZ, Cosme Damião. Potencial de caracteres na avaliação da arquitetura de plantas de feijão. **Pesquisa Agropecuaria Brasileira**, *[S. l.]*, v. 48, n. 4, p. 417–425, 2013. DOI: 10.1590/S0100-204X2013000400010.

PARK, J.; SANDBERG, I. W. **Universal approximation using radial basis function networks**. 3. ed. ed. [s.l.] : Neural Comput., 1991.

PATERSON, Andrew H.; TANKSLEY, Steven D.; SORRELLS, Mark E. DNA Markers in Plant Improvement. **Advances in Agronomy**, *[S. l.]*, v. 46, p. 39–90, 1991. DOI: 10.1016/S0065-2113(08)60578-7.

PEIXOTO, Leonardo Azevedo; LAVIOLA, Bruno Galvêas; ALVES, Alexandre Alonso; ROSADO, Tatiana Barbosa; BHERING, Leonardo Lopes. Breeding Jatropha curcas by genomic selection: A pilot assessment of the accuracy of predictive models. **PLoS ONE**, *[S. l.]*, v. 12, n. 3, p. 1–16, 2017. DOI: 10.1371/journal.pone.0173368.

PODLICH, Dean W.; WINKLER, Christopher R.; COOPER, Mark. Mapping As You Go. **Crop Science**, *[S. l.]*, v. 44, n. 5, p. 1560–1571, 2004. DOI: 10.2135/cropsci2004.1560. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.2135/cropsci2004.1560.

PRASAD, Anantha M.; IVERSON, Louis R.; LIAW, Andy. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. **Ecosystems**, *[S. l.]*, v. 9, n. 2, p. 181–199, 2006. DOI: 10.1007/s10021-005-0054-1.

R CORE TEAM; COMPUTING, R. Foundation for Statistical; TEAM, R. Core. **R: A Language and Environment for Statistical Computing**. 2020. Disponível em: https://www.r-project.org/. Acesso em: 1 jul. 2020.

RESENDE JR., Marcio F. R.; ALVEZ, Alexandre Alonso; SÁNCHES, Calor Felipe Barrera; RESENDE, Marcos Deon Vilela; CRUZ, Cosme Damião. Seleção Genômica Ampla. *In*: CRUZ, Cosme Damião; SALGADO, Caio Césio; BHERING, Leonardo Lopes (org.). **Genômica Aplicada**. Visconde do Rio Branco, MG: Suprema, 2013. p. 424.

RESENDE, Marcos D. V. 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 Phytologist**, *[S. l.]*, v. 194, n. 1, p. 116–128, 2012. DOI: 10.1111/j.1469-8137.2011.04038.x.

RESENDE, Marcos Deon Vilela. **Genômica quantitativa e seleção no melhoramento de plantas perenes e animais**. [s.l.] : Colombo: Embrapa Florestas, 2008.

SANT' ANNA, Isabela de Castro; SILVA, Gabi Nunes; NASCIMENTO, Moysés; CRUZ, Cosme Damião. Subset selection of markers for genome-enabled prediction of genetic values using radial basis function neural networks: Genome-enabled prediction of genetic values using radial basis function neural networks. **bioRxiv**, *[S. l.]*, v. 1, n. 2, p. 1–8, 2018. DOI: 10.1101/490474.

SANT'ANNA, Isabela Castro. **Redes neurais artificiais para predição genômica na presença de interações epistáticas**. 2018. Viçosa: Universidade Federal de Viçosa, *[S. l.]*, 2018. DOI: 10.20961/ge.v4i1.19180.

SANT'ANNA, Isabela Castro; NASCIMENTO, Moyses; SILVA, Gabi Nunes; CRUZ, Cosme Damião; AZEVEDO, Camila Ferreira; GLORIA, Leonardo Siqueira; FONSECA E SILVA, Fabyano. Genome-enabled prediction of genetic values for using radial basis function neural networks. **Functional Plant Breeding Journal**, *[S. l.]*, v. 1, n. 2, p. 1–8, 2020. DOI: 10.35418/2526-4117/v1n2a1. Disponível em: http://fpbjournal.com/fpbj/index.php/fpbj/article/view/57/17.

SANT'ANNA, Isabela de Castro; GOUVÊA, Ligia Regina Lima; MARTINS, Maria Alice; SCALOPPI JUNIOR, Erivaldo José; DE FREITAS, Rogério Soares; GONÇALVES, Paulo de Souza. Genetic diversity associated with natural rubber quality in elite genotypes of the rubber tree. **Scientific Reports**, *[S. l.]*, v. 11, n. 1, p. 1–10, 2021. DOI: 10.1038/s41598-020-80110-w. Disponível em: https://doi.org/10.1038/s41598-020-80110-w.

SAX, K. The Association of Size Differences with Seed-Coat Pattern and Pigmentation in Phaseolus Vulgaris. **Genetics**, *[S. l.]*, v. 8, n. 6, p. 552–560, 1923. Disponível em:

http://www.ncbi.nlm.nih.gov/pubmed/17246026%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1200765.

SCHNABLE, Patrick S.; SPRINGER, Nathan M. Progress toward understanding heterosis in crop plants. **Annual Review of Plant Biology**, *[S. l.]*, v. 64, p. 71–88, 2013. DOI: 10.1146/annurev-arplant-042110-103827.

SEARCHINGER, Tim; WAITE, Richard; HANSON, Craig; RANGANATHAN, Janet. Creating a Sustainable Food Future. **World Resources Report**, *[S. l.]*, v. 1, n. July, p. 558, 2019. Disponível em: www.SustainableFoodFuture.org.

SHAO, Yuehjen E.; HOU, Chia Ding; CHIU, Chih Chou. Hybrid intelligent modeling schemes for heart disease classification. **Applied Soft Computing Journal**, *[S. l.]*, v. 14, n. PART A, p. 47–52, 2014. DOI: 10.1016/j.asoc.2013.09.020. Disponível em: http://dx.doi.org/10.1016/j.asoc.2013.09.020.

SILVA, Gabi Nunes; TOMAZ, Rafael Simões; SANT'ANNA, Isabela Castro; NASCIMENTO, Moysés; BHERING, Leonardo Lopes; CRUZ, Cosme Damião. Neural networks for predicting breeding values and genetic gains. **Scientia Agricola**, *[S. l.]*, v. 71, n. 6, p. 494–498, 2014. DOI: 10.1590/0103-9016-2014-0057.

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

SINGH, B. D.; SINGH, A. K. **Marker-assisted plant breeding: Principles and practices**. [s.l: s.n.]. DOI: 10.1007/978-81-322-2316-0.

SOLANO MEZA, Johanna Karina; ORJUELA YEPES, David; RODRIGO-ILARRI, Javier; CASSIRAGA, Eduardo. Predictive analysis of urban waste generation for the city of Bogotá, Colombia, through the implementation of decision trees-based machine learning, support vector machines and artificial neural networks. **Heliyon**, *[S. l.]*, v. 5, n. 11, p. e02810, 2019. DOI: 10.1016/j.heliyon.2019.e02810. Disponível em: https://doi.org/10.1016/j.heliyon.2019.e02810.

SOUSA, Ithalo Coelho et al. Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. **Scientia Agricola**, *[S. l.]*, v. 78, n. 4, p. 1–8, 2021. DOI: 10.1590/1678-992x-2020-0021. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-90162021000401102&tlng=en.

SOUSA, Tiago Vieira; CAIXETA, Eveline Teixeira; ALKIMIM, Emilly Ruas; OLIVEIRA, Antonio Carlos Baião; PEREIRA, Antonio Alves; SAKIYAMA, Ney Sussumu; ZAMBOLIM, Laércio; RESENDE, Marcos Deon Vilela. Early Selection Enabled by the Implementation of Genomic Selection in Coffea arabica Breeding. **Frontiers in Plant Science**, *[S. l.]*, v. 9, n. January, p. 1–12, 2019. DOI: 10.3389/fpls.2018.01934. Disponível em: https://www.frontiersin.org/article/10.3389/fpls.2018.01934/full.

SPEED, Doug; HEMANI, Gibran; JOHNSON, Michael R.; BALDING, David J. Improved heritability estimation from genome-wide SNPs. **American Journal of**

Human Genetics, *[S. l.]*, v. 91, n. 6, p. 1011–1021, 2012. DOI: 10.1016/j.ajhg.2012.10.010. Disponível em: http://dx.doi.org/10.1016/j.ajhg.2012.10.010.

SUDHEER, K. P.; GOSAIN, A. K.; RAMASASTRI, K. S. Estimating actual evapotranspiration from limited climatic data using neural computing technique. **Journal of Irrigation and Drainage Engineering**, *[S. l.]*, v. 129, n. 3, p. 214–218, 2003. DOI: 10.1061/(ASCE)0733-9437(2003)129:3(214).

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introduction to Data Mining**. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2006. DOI: 10.1016/0022-4405(81)90007-8.

TANG, Jie et al. Application of Machine-Learning Models to Predict Tacrolimus Stable Dose in Renal Transplant Recipients. **Scientific Reports**, *[S. l.]*, v. 7, n. February, 2017. DOI: 10.1038/srep42192.

TAYLAN, Pakize; WEBER, Gerhard Wilhelm. CG-Lasso Estimator for Multivariate Adaptive Regression Spline. *In*: TAS, Kenan; BALEANU, Dumitru; MACHADO, J. A. Tenreiro (org.). **Mathematical Methods in Engineering: Aplications in Dynamics of Complex Systems**. [s.l.] : Springer International Publishing AG, 2019. p. 121–136. DOI: 10.1007/978-3-319-90972-1_9.

TOMAZ, Rafael Simões; ALVEZ, Daniel Pedrosa; NASCIMENTO, Moysés; CRUZ, Cosme Damião. Inteligência Computacional. *In*: CRUZ, Cosme Damião; NASCIMENTO, Moysés (org.). **Inteligência Computacional Aplicado ao Melhoramento Genético**. Viçosa, MG: Editora UFV, 2018. p. 414.

TONG, Hao; NIKOLOSKI, Zoran. Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. **Journal of Plant Physiology**, *[S. l.]*, v. 257, p. 153354, 2021. DOI: 10.1016/j.jplph.2020.153354. Disponível em: https://doi.org/10.1016/j.jplph.2020.153354.

VAN EENENNAAM, A. L.; YOUNG, A. E. Prevalence and impacts of genetically engineered feedstuffs on livestock populations. **Journal of Animal Science**, *[S. l.]*, v. 92, n. 10, p. 4255–4278, 2014. DOI: 10.2527/jas2014-8124.

VAN EENENNAAM, Alison L.; WEIGEL, Kent A.; YOUNG, Amy E.; CLEVELAND, Matthew A.; DEKKERS, Jack C. M. Applied Animal Genomics: Results from the Field. **Annual Review of Animal Biosciences**, *[S. l.]*, v. 2, n. 1, p. 105–139, 2014. DOI: 10.1146/annurev-animal-022513-114119.

WADE, M. J. **Epistasis and evolutionary process**. New York: Oxford University Press, 2000.

WANG, Jiabo; ZHOU, Zhengkui; ZHANG, Zhe; LI, Hui; LIU, Di; ZHANG, Qin; BRADBURY, Peter J.; BUCKLER, Edward S.; ZHANG, Zhiwu. **Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traitsHeredity**, 2018. DOI: 10.1038/s41437-018-0075-0.

WONG, C. K.; BERNARDO, R. Genomewide selection in oil palm: Increasing selection gain per unit time and cost with small populations. **Theoretical and Applied Genetics**, *[S. l.]*, v. 116, n. 6, p. 815–824, 2008. DOI: 10.1007/s00122-008-0715-5.

YABE, Shiori; HARA, Takashi; UENO, Mariko; ENOKI, Hiroyuki; KIMURA, Tatsuro; NISHIMURA, Satoru; YASUI, Yasuo; OHSAWA, Ryo; IWATA, Hiroyoshi. Potential of genomic selection in mass selection breeding of an allogamous crop: An empirical study to increase yield of common buckwheat. **Frontiers in Plant Science**, *[S. l.]*, v. 9, n. March, p. 1–12, 2018. DOI: 10.3389/fpls.2018.00276.

YORK, Timothy P.; EAVES, Lindon J. Common Disease Analysis Using Multivariate Adaptive Regression Splines (MARS): Genetic Analysis Workshop 12 Simulated Sequence Data. **Genetic Epidemiology**, *[S. l.]*, v. 21, n. S1, p. S649–S654, 2001. DOI: 10.1002/gepi.2001.21.s1.s649.

YORK, Timothy P.; EAVES, Lindon J.; VAN DEN OORD, Edwin J. C. G. Multivariate adaptive regression splines: A powerful method for detecting disease-risk relationship differences among subgroups. **Statistics in Medicine**, *[S. l.]*, v. 25, n. 8, p. 1355–1367, 2006. DOI: 10.1002/sim.2292.

ZHANG, Haohao; YIN, Lilin; WANG, Meiyue; YUAN, Xiaohui; LIU, Xiaolei. Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. **Frontiers in Genetics**, *[S. l.]*, v. 10, n. MAR, p. 1–10, 2019. DOI: 10.3389/fgene.2019.00189.

ZHANG, Wengang; GOH, Anthony T. C. Multivariate adaptive regression splines and neural network models for prediction of pile drivability. **Geoscience Frontiers**, *[S. l.]*, v. 7, n. 1, p. 45–52, 2016. DOI: 10.1016/j.gsf.2014.10.003. Disponível em: http://dx.doi.org/10.1016/j.gsf.2014.10.003.

ZHAO, Yusheng; METTE, Michael F.; REIF, Jochen C. Genomic selection in hybrid breeding. **Plant Breeding**, *[S. l.]*, v. 134, n. 1, p. 1–10, 2015. DOI: 10.1111/pbr.12231.

ZHENG, Gang; YANG, Pengbo; ZHOU, Haizuo; ZENG, Chaofeng; YANG, Xinyu; HE, Xiaopei; YU, Xiaoxuan. Evaluation of the earthquake induced uplift displacement of tunnels using multivariate adaptive regression splines. **Computers and Geotechnics**, *[S. l.]*, v. 113, n. May, p. 103099, 2019. DOI: 10.1016/j.compgeo.2019.103099. Disponível em: https://doi.org/10.1016/j.compgeo.2019.103099.

ZHENG, Gang; ZHANG, Wenbin; ZHOU, Haizuo; YANG, Pengbo. Multivariate adaptive regression splines model for prediction of the liquefaction-induced settlement of shallow foundations. **Soil Dynamics and Earthquake Engineering**, *[S. l.]*, v. 132, n. August 2019, p. 106097, 2020. DOI: 10.1016/j.soildyn.2020.106097. Disponível em: https://doi.org/10.1016/j.soildyn.2020.106097.

ZINGARETTI, Laura M.; GEZAN, Salvador Alejandro; FERRÃO, Luis Felipe V.; OSORIO, Luis F.; MONFORT, Amparo; MUÑOZ, Patricio R.; WHITAKER, Vance M.; PÉREZ-ENCISO, Miguel. Exploring Deep Learning for Complex Trait Genomic

Prediction in Polyploid Outcrossing Species. **Frontiers in Plant Science**, *[S. l.]*, v. 11, n. February, p. 1–14, 2020. DOI: 10.3389/fpls.2020.00025. Disponível em: https://www.frontiersin.org/article/10.3389/fpls.2020.00025/full.

4. ARTIGO 1

Genomic prediction through machine learning and neural networks for traits with epistasis

Abstract

Genome wide selection (GWS) is one contribution of molecular genetics to breeding. Machine learning (ML) and artificial neural networks (ANN) methods are non-parameterized and can develop more accurate and parsimonious models for GWS analysis. Multivariate Adaptive Regression Splines (MARS) is considered one of the most flexible ML methods, automatically modeling nonlinearities and interactions of the predictor variables. This study aimed to evaluate and compare methods based on ANN, ML, including MARS, and G-BLUP through GWS. An F2 population formed by 1000 individuals and genotyped for 4010 SNP markers and twelve traits from a model considering epistatic effect, with QTL numbers ranging from eight to 480 and heritability of 0.5 or 0.8 were simulated. Variation in heritability and number of QTL impacts the performance of methods. MARS methods showed better results for oligogenic traits. For quantitative traits it was observed the highest R^2 to Radial Base Network (RBF), followed by G-BLUP, Random Forest (RF), Bagging (BA), and Boosting (BO). Non-additive MARS methods also showed high R^2 for traits with high heritability and 240 QTLs or more. ANN and ML methods are powerful tools to predict genetic values in traits with epistatic effect, for different degrees of heritability and QTL numbers.

Keywords: Genome wide selection; Quantitative Trait Locus; non-additive effects; Multivariate Adaptive Regression Splines; Genome-enabled prediction.

Introduction

Genome wide selection (GWS), proposed by MEUWISSEN; HAYES; GODDARD (2001), has become one of the main contributions of molecular genetics to breeding. The GWS approach increased the accuracy in the prediction of breeding values, making the selection of elite genotypes more efficient and accurate (TONG; NIKOLOSKI, 2021). Furthermore, GWS made it possible to accelerate the

improvement process by half the time in relevant crops, helping to sustain current food demands (LI et al., 2018; PEIXOTO et al., 2017; SINGH; SINGH, 2015).This time reduction allowed breeders to maximize genetic gains per unit of time, in addition to early selection (SOUSA et al., 2019; YABE et al., 2018). All these benefits are due to the direct use of DNA information in the selection of individuals, associating marker information with phenotypic information, reducing the time and resources allocated to the development of a new cultivar (ALKIMIM et al., 2020; SANT'ANNA et al., 2020; TONG; NIKOLOSKI, 2021).

Genome-based prediction is influenced by several factors, such as the predictive ability of the methods, the complexity of the trait's genetic architecture due to non-additive effects (dominance and epistasis), number of phenotypic observations and markers used TONG e NIKOLOSKI (2021). Increasingly, researchers are turning to machine learning and neural network techniques, which have built-in predictor selection capabilities and are unparameterized to develop more accurate and parsimonious models (LIEW et al., 2020). Furthermore, some of these methods allow identifying interactions between markers. This property allows great flexibility to deal with different types of traits with gene control with additive, dominant and epistatic effects (CRUZ; NASCIMENTO, 2018; LI et al., 2018; SANT'ANNA et al., 2020; SOUSA et al., 2021).

Among the various methods based on machine learning, Multivariate Adaptive Regression Splines (MARS) is considered one of the most flexible (COOK; ZEE; RIDKER, 2004), it is parsimonious and performs better than artificial neural networks for genomic prediction in some studies (LIEW et al., 2020; LIN et al., 2008). MARS produces continuous models that can have multiple partitions, automatically models nonlinearities, and contemplates interactions of predictor variables using adaptively selected spline functions (ALTINOK; KARAGOZ; BATMAZ, 2020; TAYLAN; WEBER, 2019; ZHENG et al., 2020).

In the genetic context, MARS can be able to adjust the genetic architecture of the trait and can also detect interactions, such as epistasis, and can be used to define the type of trait control. Thus, the inheritance mode of the markers and their interactions can also be determined automatically, therefore, the number of parameters in the modeling can be drastically reduced (LIN et al., 2008). Although several studies have proven the high power of MARS in the evaluation of genomic data

in the medical field (CHANG; WANG; JIANG, 2011; LIN et al., 2008; TANG et al., 2017; YORK; EAVES; VAN DEN OORD, 2006), it is not known whether more advanced machine learning, such as MARS, offer superior performance over traditional statistical methods for genetic improvement. In this sense, the objectives of this study were: (i) to evaluate the general accuracy and the variability of the prediction performance of methods based on machine learning, including MARS, and neural networks in genomic prediction analyzes for simulated traits for different numbers of genes in the presence of dominance and epistasis and with different degrees of heritability and (ii) to compare the results obtained with G-BLUP in different scenarios.

Material and methods

Simulation of population genome

To simulate the data, an F2 population of a diploid species ($2n = 2x = 20$) with an effective size of 1000 individuals was taken as reference. The genotypic constitution of each individual was established considering the information in the genome and the random union of gametes from the parents, assuming a gametic pool of 5000 reproductive units, per parent, in each fertilization. The population was generated using divergent parental lines, i.e., contrasting homozygous parents (P1 dominant and P2 recessive), with a genome established considering 10 linkage groups with a size of 200 cM each. To provide linkage disequilibrium between markers, the percentage of recombination was equivalent to a distance between loci of 0.5 cM. The genome was generated with a saturation level of 401 equidistantly spaced molecular markers in each linkage group, resulting in a total of 4010 molecular markers in the genome. Marks were codominant (SNPs - Single Nucleotide Polymorphism), allowing the identification of heterozygous individuals.

Simulation and constitution of phenotypic values

From the simulated genotypic data of the F2 population, 12 traits with numbers of controlling genes ranging from 8 to 480 and heritability of 0.5 or 0.8 were simulated (Table 1). The controlling genes (QTL - Quantitative Trait Locus) were distributed equally among the first 8 linkage groups (Supplementary Figure 1).

Table 1: Number of controlling loci and heritability (h^2) of the 12 simulated traits (C1 to C12).

h^2	Número de locos controladores da característica					
	8	**40**	**80**	**120**	**240**	**480**
0,5	C1	C2	C3	C4	C5	C6
0,8	C7	C8	C9	C10	C11	C12

Fonte: O autor.

Eight QTL controlled for C1 and C7 traits, defined by the central markers of the first eight linkage groups. For traits C2 to C6 and C8 to C12 the QTL were distributed keeping an approximate distance between them, within the first 8 linkage groups (Supplementary Figure 1).

The total phenotypic values expressed by a given individual for traits C1 to C12 were simulated considering the mean equal to 100 and coefficient of variation equal to 12%, with a dominance level (d_i) equal to 0.5 and by a model with epistatic effect according to the following equation:

$$Y_{ij} = \mu + \sum_{j} \alpha_{ji} + \sum_{j} \alpha_{ji}\alpha_{ji+1} + e_i$$

where Y_i is the phenotypic value for observation i; μ is the general mean; α_j is the effect of the favorable allele at locus j of individual i, that is, it assumes the values $u + a_i$, $u + d_i$ and $u - a_i$ for the genotypic values associated with classes AA, Aa and aa, respectively, with u being the mean between the dominant homozygote (AA) and the recessive homozygote (aa). Classes were identified by coding 1, 0 or -1, respectively; $\alpha_{ji}\alpha_{ji+1}$ represents the interaction between favorable alleles at different loci. The variance structure of the residues was given by $e \sim N(0, V_e)$, where $V_e = ((1 - h^2)V_g)/h^2$, where V_e is the residual variance, V_g is the genotypic variance, and h^2 the heritability.

Prediction of breeding values

The genomic breeding values (GEBVs) were predicted using methods based on statistical approaches, represented by G-BLUP, on neural network approaches, represented by the Multilayer Perceptron Network (MLP) and Radial Basis Function

Network (RBF) and on learning approaches from Multivariate machine Adaptive Regression Splines (MARS), Decision Tree (DT), Boosting (BO), Bagging (BA) and Random Forest (RF).

Neural network approaches are often treated as machine learning (BARBOSA et al., 2021; COSTA et al., 2021; SOUSA et al., 2021). However, each approach has its specificity and here they will be considered as different approaches. As neural networks work like the human brain, they are composed of neurons organized in layers that capture all available information to generate a decision-making criterion, they differ from machine learning methods, which model the limitations of data separation with based on the learning decision rules on the input characteristics of the model (SOLANO MEZA et al., 2019).

Data analysis

For all methods, the input data was a matrix of molecular markers, represented by the genotypic values encoded in -1, 0 and 1, simulated for 4010 markers and 1000 individuals. The methods returned in the output a vector with the GEBV for each individual. For comparison, the methods were grouped according to their respective learning approach: GBLUP – GBLUP; MLP and RBF –Network; DT, BA, BO and RF – Trees; and MARS 1, MARS 2 and MARS 3 – MARS.

Multivariate Adaptive Regression Splines (MARS)

The algorithm proposed by (FRIEDMAN, 1991) Multivariate Adaptive Regression Splines (MARS), considers an expansion in piecewise linear functions, called basis functions (BFs), as follows:

$$(x-t)_+ = \begin{cases} x-t, & se\ x > t, \\ 0, & if\ otherwise. \end{cases} ; \quad (t-x)_+ = \begin{cases} t-x, & se\ x < t, \\ 0, & if\ otherwise. \end{cases}$$

Each function is a piecewise linear spline, with a node at the value t. These two BFs are called a reflexive pair. MARS forms reflexive pairs for each input (marker) X_j, with nodes at each observed value x_{ij} of that input. The model building strategy is like a progressive linear regression, but instead of using the original inputs, we use functions from the set. $C = \{(X_j - t)_+, (t - X_j)_+\}_{t \in \{x_{1j},\ x_{2j},\ldots,\ x_{Nj}\}\ j=1,2,\ldots,p}$ and/or its

products. The MARS model, which is a linear combination of the BFs and/or their interactions, is given by (HASTIE; TIBSHIRANI; FRIEDMAN, 2009):

$$f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X)$$

where β_0 is the regression constant, β_m with m = 1, 2, ..., M, are the regression coefficients, and $h_m(X)$ is a function in C, or a product of two or more functions.

The estimation process of the parameters β_0 and β_m is based on the minimization of the residual sum of squares. First, the forward phase is performed on the training data, initially starting to build the model only with the constant function $h_0(X) = 1$, and all functions in the C set are candidate functions. At each subsequent step, the base pair that produces the maximum reduction in training error is added. Considering a model with basic M functions, the next pair to be added to the model is (HASTIE; TIBSHIRANI; FRIEDMAN, 2009):

$$\hat{\beta}_{M+1} h_l(X)(X_j - t)_+ + \hat{\beta}_{M+2} h_l(X)(t - X_j)_+, h_l \in M$$

where $\hat{\beta}_{M+1}$ and $\hat{\beta}_{M+2}$ are coefficients estimated by the least square method, together with all other $M + 1$ coefficients in the model. This process of adding BFs continues until the model reaches a predetermined maximum number, often leading to a purposefully oversized model (ZHANG; GOH, 2016).

The backward phase improves the model by removing the least significant terms until finding the best submodel. The model subsets are compared using the generalized cross-validation (GCV) method. The GCV is the root-mean-square residual error divided by a penalty that depends on the complexity of the model (ZHANG; GOH, 2016). The GCV is calculated as (HASTIE; TIBSHIRANI; FRIEDMAN, 2009):

$$GCV(\lambda) = \frac{\frac{1}{N}\sum_{i=1}^{N}[y_i - \hat{f}_\lambda(x_i)]^2}{\left[1 - \frac{C(M)}{N}\right]^2}$$

where M is the effective number of model parameters, $C(M)$ is a cost function for each basis function included in the developed submodel, which by default is adopted by default value of 3 (FRIEDMAN, 1991), N is the number of datasets used in cross-validation and $\hat{f}_\lambda(x_i)$ denotes the predicted MARS values.

To identify the possible interaction between the QTLs, MARS models with

degrees equal to 1, 2, and 3 were used, with the model with degree 1 considered an additive model and the others non-additive, which allow interactions between markers. For the stopping criterion of the forward phase, the maximum number of terms in the adopted model was equal to 200, as the default of the "land" package of R. A preliminary analysis was carried out for the second stopping criterion (MILBORROW, 2019), in which incrementing a term in the model would change the coefficient of determination from less than 0.001 (default) to 0.05, choosing the best model that presented the highest selective accuracy (R^2) for the validation set.

Genomic BLUP (G-BLUP)

An epistatic model, including dominance and additive effects, for the REML/G-BLUP method was used according to the following expression:

$$y = Xb + Zu_a + Zu_d + Zu_{epi} + \varepsilon$$

where y is the vector of phenotypic observations; b is the vector of fixed effects (in this study, the general mean) with incidence matrix X; u_a, u_d e u_{epi} are vectors of genetic values of additive, dominant and epistatic effects, respectively; Z is the incidence matrix for these vectors; and ε is the random error vector. The variance structure was given by $u_a \sim N(0, G_a\sigma^2_{u_a})$; $u_d \sim N(0, G_d\sigma^2_{u_d})$; $u_{epi} \sim N(0, G_{epi}\sigma^2_{u_{epi}})$ and $\varepsilon \sim N(0, I\sigma^2_{\varepsilon})$, where G_a, G_d and G_{epi} are the genomic relationship matrices for the additive, dominant and epistatic effects, respectively , and $\sigma^2_{u_a}$, $\sigma^2_{u_d}$ and $\sigma^2_{u_{epi}}$ are the additive, dominance and epistatic variances, respectively.

For the construction of the genomic relationship matrices (W and S) used in the model, M_{ij} was considered to be the incidence of the number of alleles of brand j of individual i and p_j the frequency of the dominant allele A in brand j. In this way, the W and S matrices were given by (ZHANG et al., 2019):

$$W_{ij} = \begin{cases} 2 - 2p_j, & if\ M_{ij} = AA \\ 1 - 2p_j, & if\ M_{ij} = Aa \\ 0 - 2p_j, & if\ M_{ij} = aa \end{cases}, \quad and$$

$$S_{ij} = \begin{cases} -2\left(1 - p_j\right)^2, & if\ M_{ij} = AA \\ 2p_j\left(1 - p_j\right), & if M_{ij} = Aa \\ -2p_j^2, & if\ M_{ij} = aa \end{cases}$$

In this way, we obtain:

$$G_a = \frac{WW'}{\sum_{j=1}^{n} 2p_j(1-p_j)}; G_d = \frac{SS'}{\sum_{j=1}^{n}[2p_j(1-p_j)]^2}; G_{epi} = G_a \# G_a$$

Where # is the Hadamard product operator.

The mixed model equations for the full model were given by (Resende et al. 2014):

$$\begin{bmatrix} X'X & X'Z & X'Z & X'Z \\ Z'X & Z'Z + {G_a}^{-1}\lambda_1 & X'Z & X'Z \\ Z'X & Z'Z & Z'Z + {G_d}^{-1}\lambda_2 & X'Z \\ Z'X & Z'Z & X'Z & Z'Z + {G_{epi}}^{-1}\lambda_3 \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u}_a \\ \hat{u}_d \\ \hat{u}_i \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ Z'y \\ Z'y \end{bmatrix}$$

were $\lambda_1 = \frac{\sigma_\varepsilon^2}{\sigma_{u_a}^2}$, $\lambda_2 = \frac{\sigma_\varepsilon^2}{\sigma_{u_d}^2}$ and $\lambda_3 = \frac{\sigma_\varepsilon^2}{\sigma_{u_{epi}}^2}$ and the variances were estimated by the Restricted Maximum Likelihood Method (REML).

Multilayer Perceptron Neural Network (MLP)

The Levenberg-Marquardt backpropagation training algorithm was used for the Multilayer Perceptron Neural Network (MLP). Preliminary tests were performed with different architectures, being represented by 1 layer and the number of neurons varying from 5 to 15, to choose the best topology to be used. The linear activation function (purelin) was used.

The linear function for the nth neuron of the output layer of an MLP was represented by:

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

where: p is a linear activation function, x_0 is the bias term of the nth neuron, x_i is the i-th input, w_j is the synaptic weights to be adjusted and $f_{xj}(x_i)$ is the value coming from the layer hidden for each input i, assigned to an activation function. The activation function used in this work was the linear one ($f_{xj}(x_i) = x_i$).

Neural Network Radial Base Function (RBF)

The Radial Base Function Neural Network (RBF) uses a feedforward

architecture. This model also consists of an input layer, a hidden layer, and an output layer. RBF training is hybrid (supervised and unsupervised) and the input layer information goes through a linear k-means cluster (CRUZ; NASCIMENTO, 2018). The hidden layer applies a non-linear transformation of the input space to a high-dimensional hidden space with a Gaussian function. The output layer applies a transformation to the hidden space, providing an output vector for the network. The RBF optimization training included: the weights between the hidden layer and the output layer, the activation function, the center of activation functions, the distribution of the center of activation functions, and the number of hidden neurons (CRUZ; NASCIMENTO, 2018). During the training process, only the weights between the hidden layer and the output layer are modified (SANT'ANNA et al., 2020). To select the best RBF architecture, according to the MLP, preliminary tests were carried out. The number of neurons ranged from 5 to 50 and radius size from 30 to 50. The mean square error was set to 0.05.

The linear function for the nth neuron of the output layer of an RBF was represented by:

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

where: g is a linear function, x_0 is the bias term of the nth neuron, x_i is the i-th input, w_j is the synaptic weights to be adjusted, and $f_{xj}(x_i)$ is the value coming from the hidden layer for each input i, assigned to the Gaussian activation function, which is given by the equation: $e^{\frac{-(u-c)^2}{2\sigma^2}}$, where c is the center of the Gaussian function, σ^2 is the variance of the Gaussian function and i is the value of the individual's input, which represents the activation potential of the clustering phase.

Decision Tree (DT)

The decision tree structure was based on a regression tree, created from the search for the tree that would lead to the data partition until the formation of homogeneous groups was obtained. To perform recursive binary division, first is the marker X_j and the cutoff point s so that the division of the predictor space into the

regions $\{x|x_j < s\}$ e $\{x|x_j \geq s\}$ leads to the greatest possible reduction in RSS. That is, we consider all markers $x_1, \ldots, X_m$ and all possible values of the cutoff s for each of the markers, and then choose the marker and cutpoint so that the resulting tree has the smallest RSS. The equation that reflects the binary division is (JAMES et al., 2021):

$$R_1(j,s) = \{X|X_j < s\} \, e \, R_2(j,s) = \{X|X_j \geq s\},$$

and then we look for the value of J and S that minimize the equation:

$$\sum_{i:x_i \in R_1(j,s)} \left(y_i - \hat{y}_{R_1}\right)^2 + \sum_{i:x_i \in R_2(j,s)} \left(y_i - \hat{y}_{R_2}\right)^2$$

where: $\hat{y}_{R_1}$ is the average of the response variable of the training observations belonging to the region $R_1(j,s)$, $\hat{y}_{R_2}$ is the average of the response variable of the training observations belonging to the region $R_2(j,s)$ and y_i is the true value of the characteristic of each individual.

Bagging (BA)

The Bagging (BA) method creates several similar datasets by resampling (bootstrapping) to obtain an average of several regression trees that are performed without pruning for each dataset (BREIMAN, 1996; PRASAD; IVERSON; LIAW, 2006). Thus, a number B of models are obtained: $\hat{f}^1(x), \hat{f}^2(x), \ldots, \hat{f}^B(x)$. These generated models are used to obtain an average model, given by: $\hat{f}_{médio}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(x)$. The number of trees sampled for BA was set at 500 trees.

Random Forest (RF)

Random Forest (RF) (BOEHMKE; GREENWELL, 2019) is similar to BA in that bootstrap samples are used to build multiple trees, the difference being that each tree is established with a random subset of predictors. The number of predictors used to find the best split at each node is a subset that was chosen by $m = \frac{v}{3}$, with v being the total number of predictors. The number of trees for the RF was set at 500. For the RF, the trees grow to their maximum size without pruning, and the aggregation is done by averaging the trees (COSTA et al., 2021).

Boosting (BO)

Boosting (BO) creates trees sequentially using information from previous trees (JAMES et al., 2021). In this sense, BO is an approach repeatedly trained on the same sample so that at each iteration, a measure of prediction error is calculated for each marker, and in the next iteration, markers with higher errors receive greater weight in the model training. The prediction is performed by weighting the results of the set of all regression trees (GHAFOURI-KESBI et al., 2017). The number of trees sampled was 500, with a learning rate of 0.01 and a depth of 2. The following model was used to adjust the BO (HASTIE; TIBSHIRANI; FRIEDMAN, 2009):

$$f(x) = \sum_{m=1}^{M} \beta_m b(x; \gamma_m)$$

Where β_m, m = 1, 2, ..., M are the coefficients of base expansion and $b(x; \gamma_m)$ are simple functions of the multivariate argument x, with a set of parameters $\gamma = \gamma_1, \gamma_2, \cdots, \gamma_m$.

Efficiency parameters

To evaluate the efficiency of the techniques, the selective accuracy was used, which is measured by the square of the correlation (R^2) between the estimated values - GEBVs ($\hat{y}$) and the real values (y), and the root means square error ($RMSE$), which expresses the predictive accuracy. The selective and predictive accuracies were given respectively by the following equations $R^2 = \left(cor(\hat{y}, y)\right)^2$ and $RMSE = \sqrt{\frac{\sum(\hat{y}-y)^2}{n}}$.

Training and validation

For training and validation, k-fold cross-validation was performed, considering k = 5. Each fold corresponds to the part of the population that is treated as a validation population. The remaining folds (k-1) were used as a cross-validation training population. In the end, the mean and the standard error among the 5 iterations for the validation set of each parameter were estimated.

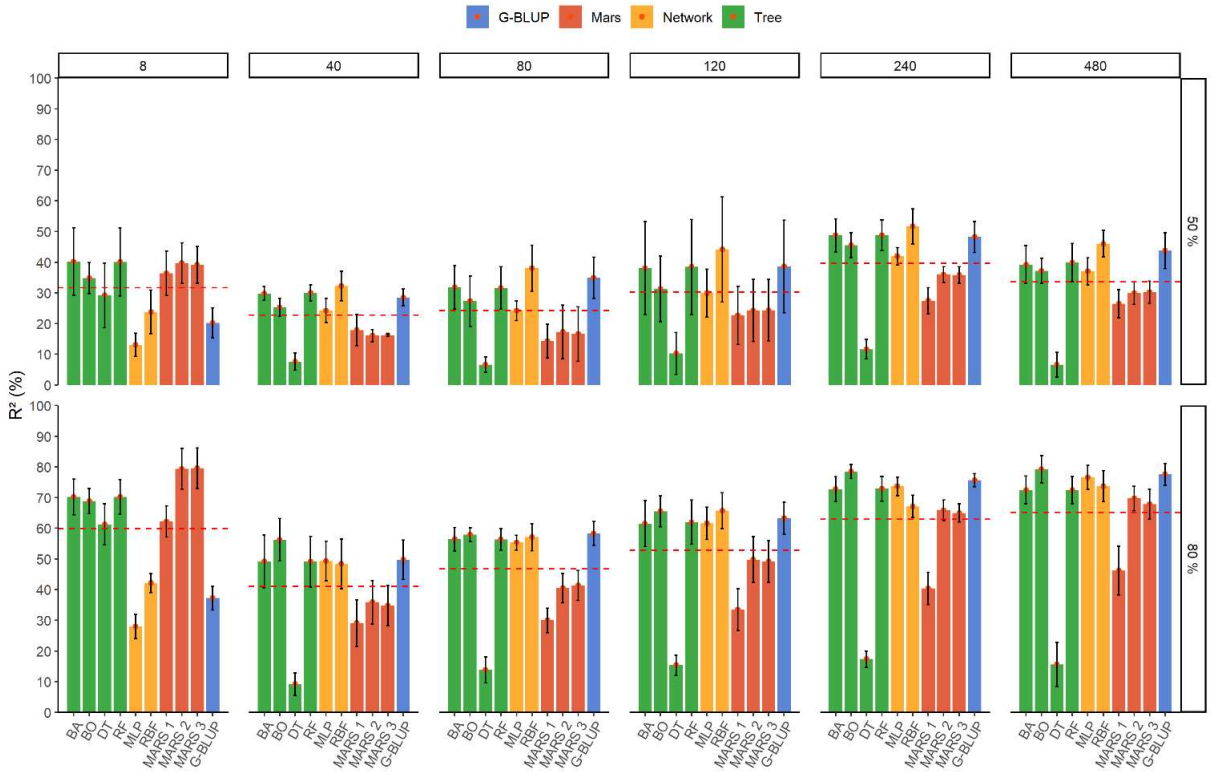
Computational aspects

Population simulations were performed using the GENES software (CRUZ, 2013). The G-BLUP, DT, BA, RF, BO, and MARS methods were performed with the GENES software integrated with the R software (CRUZ, 2016; R CORE TEAM; COMPUTING; TEAM, 2020). The MLP and RBF methods were performed by the GENES software integrated with the MATLAB software (CRUZ, 2016; MATLAB, 2019).

Results

The selective accuracy (R^2) of the prediction of breeding values for all methods was higher in scenarios with higher heritability (Figure 1). On the other hand, the variation in the number of QTL showed that the methods produce diverse results, indicating that the number of QTL of the trait directly influences the prediction of GEBVs according to the method used and that the increase in the number of QTL is harmful to the MARS approach and the DT method, while for the other methods the increase in the number of QTL reflects an improvement in R^2.

Figure 1 - Average results of selective accuracy (R^2) as a function of the number of genes and heritability for the families of the methods: Trees [Bagging (BA), Boosting (BO), Decision Tree (DT); and Random Forest (RF)]; Network (Multilayer Perceptron Network (MLP) and Radial Base Function Network (RBF) MARS (MARS 1, 2 and 3); and G-BLUP. The red dashed line refers to the overall mean value of the selective accuracy (R^2) between all methods for comparison purposes.



FONTE: O autor.

For both heritability scenarios, the methods based on machine learning, MARS and Trees, presented higher values of R^2 for the traits with the lowest number of QTL, when compared to the other methods (Figure 1). The effect of the interaction between markers was even more evident for higher heritability (80%), resulting in higher R^2 values for the non-additive MARS models (MARS 2 and 3).

From the scenarios with 40 QTL, an increase was observed for the values of R^2 as the number of QTL increased, reaching values close to those of the real genetic variation when the trait presents 480 QTL. In these scenarios, the G-BLUP and RBF methods, followed by RF and BA, presented the highest values for R^2 and always above the general average (red line) for the traits for both heritability scenarios (Figure 1). For scenarios with 80% heritability and 40 or more QTL, the MLP and BO methods

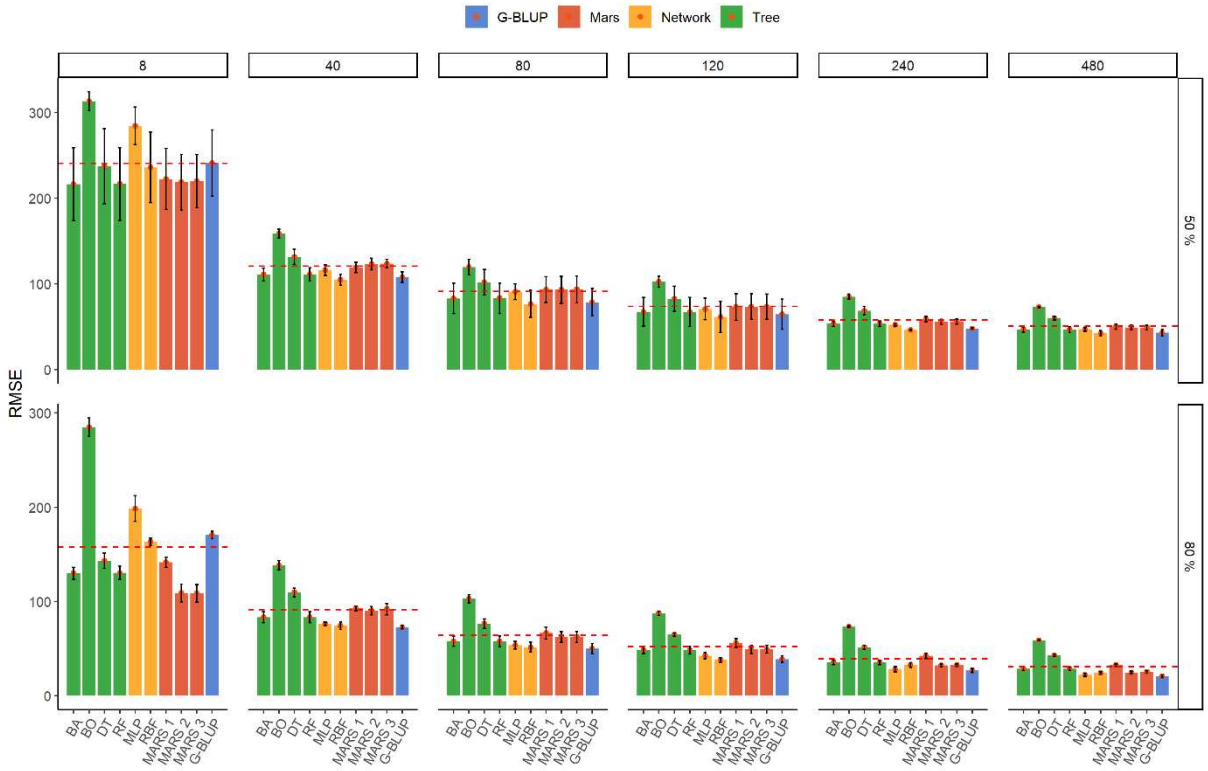
also deserve to be highlighted.

Despite presenting lower values for R^2 compared to other methods, the predictive power of MARS methods for traits with many QTLs cannot be neglected, especially when there is a very high number of QTLs, such as 240 and 480 genes, the non-additive MARS methods (MARS 2 and 3) showed high R^2 values (above 65%), and considering the standard error, values lower only than G-BLUP (Figure 1). It is worth mentioning that for these scenarios, MARS had high predictive potential, explaining almost all the genetic variations of these traits.

Methods based on MARS and regression trees did not obtain a linear response as a function of increasing the number of QTL. On the other hand, both methods based on neural networks and G-BLUP showed a substantial improvement the higher the QTL number (Figure 1). With the exception of DT, which presented lower R^2 values in almost all scenarios, the tree-based methods presented R^2 values close to the simulated heritability, mainly for scenarios with 240 QTL (Figure 1). In addition, these methods presented values of R^2 greater than the overall mean of R^2 in all scenarios (Figure 1). The BO method presented the best result when the scenarios were of greater heritability and 40 or more QTL. BO was also the method that showed the greatest sensitivity to heritability and showed a substantial improvement in results in higher heritability scenarios.

The predictive accuracy results ($REQM$), referring to the error in the prediction of the GEBVs of the individuals, were always smaller according to the increase in the number of QTL, that is, the greater the number of QTL, the lower the error in the prediction of the GEBVs of the individuals, regardless of the method used (Figure 2). In this case, the impact caused by the increase in the number of QTLs on the prediction error of GEBVs is greater than the change in heritability and is inversely proportional. This result was possible due to the fixed number of markers, providing a greater proportion of direct effects of markers on traits in relation to those poorly correlated with the phenotype and without direct effect.

Figure 2 - Average results of predictive accuracy ($RMSE$) as a function of the number of genes and heritability for the families of the methods: Trees [Bagging (BA), Boosting (BO), Regression Tree (DT); and Random Forest (RF)]; Network (Multilayer Perceptron Network (MLP) and Radial Base Function Network (RBF) MARS (MARS 1, 2 and 3) and G-BLUP. The red dashed line refers to the overall mean value of predictive accuracy ($RMSE$) between all methods for comparison purposes.



FONTE: O autor.

As obtained for R^2, MARS-based methods showed better results of RMSE for traits with the lowest number of QTL for both heritability's (Figure 2). For these scenarios, it was also observed that the MARS models showed greater differences from the other methods when the genetic effects were greater than the environmental effects, that is, greater heritability.

The RBF method presented very similar RMSE values when compared to those obtained through G-BLUP for all scenarios (Figure 2). On 40 QTLs, these methods presented the lowest values for RMSE. Similar values of these methods were obtained by the non-additive MARS (MARS 2 and MARS 3) for the scenarios with 240 and 480 QTL and heritability of 80%. On the other hand, the BO method presented the highest RMSE values in all scenarios, followed by the MLP method for scenarios with few QTL (8) and DT in the other scenarios.

Discussion

The inclusion of a greater number of marker variables in a predictive model can be useful to obtain a better performance, but it can lead to the addition of redundant information and make it difficult to apply in practice (TANG et al., 2017). Furthermore, in segregating populations, non-additive effects, i.e., dominance, and epistasis, are highly relevant and should also be considered (SCHNABLE; SPRINGER, 2013). Thus, methods that deal with high dimensionality and take into account possible interactions between predictor variables have important aspects. Many recent studies have been applied to GWS and have shown that machine learning and neural network methods can perform better or similarly in predicting genotypic data phenotypes compared to statistical methods (BARBOSA et al., 2021; MA et al., 2017; SANT'ANNA et al., 2020; SHAO; HOU; CHIU, 2014; SILVA et al., 2014; SOUSA et al., 2021; ZINGARETTI et al., 2020).

However, there is still a gap to be filled on which method is better to perform the prediction when it comes to different degrees of heritability and QTL number, including when considering epistatic effects. The variability in the results for the different methods suggests that any method is prone to produce a differentiated result under some type of data perturbations. The results obtained demonstrated the strong effect of heritability and increase in QTL number on R^2 and RMSE values. Several studies have shown that there is a favorable effect of heritability on selective accuracy (BARBOSA et al., 2021; COUTINHO et al., 2018; GHAFOURI-KESBI et al., 2017; MOURA; PAMPLONA; BALESTRE, 2019; SANT'ANNA et al., 2020), as also obtained in this study. This is justified by the greater genetic variation in higher heritability and, consequently, less environmental effect, contributing to more accurate predictions of marker effects (BARBOSA et al., 2021).

The results showed a reduction in the RMSE in scenarios with a greater number of QTL, and that there was also a reduction in the RMSE in the scenarios with high heritability, however in a smaller proportion. Results similar to those obtained in this study were observed by GHAFOURI-KESBI et al. (2017), in scenarios with heritability ranging from 0.1 to 0.5 and 100 QTL, and BARBOSA et al. (2021) evaluating scenarios with QTL numbers ranging from 2 to 88 and heritability from 0.3 to 0.8. The reduction in RMSE due to the increase in the number of QTLs may have occurred due to the

lower influence of the multiplicative effect between the additive and dominant effects that characterize epistatic effects in more complex traits (BARBOSA et al., 2021; COSTER et al., 2010; GHAFOURI-KESBI et al., 2017).

As MARS can simultaneously include multiple terms (additive and epistatic effects) in a model (EVERINGHAM; SEXTON, 2011) genetic interactions can be better evaluated. Apparently, this fact could lead to a better prediction for GWS, since it would be possible to reduce the residual variance of the model, by capturing information that before was isolated only in the residual component, such as the effects of the interaction between markers. However, as LIN et al. (2008) explain, some interaction patterns tend to be less pronounced in traits with a high number of QTLs, making it difficult to identify these interactions. Thus, methods based on recursive partitioning, such as MARS and Trees, benefit from situations in which the predictor variables can be partitioned into well-defined regions (SOUSA et al., 2021), as is the case with features with lower QTL numbers (oligogenic). This is because traits controlled by few genes have well-defined phenotypic classes and suffer little or no environmental influence (CRUZ, 2012).

These results suggest that MARS is good alternative to be used, especially when it is easier to identify groups of individuals based on the population genome. Due to the identification of markers and/or interactions between markers of greater effect, MARS proved to be more efficient when the multiplicative effects of the controlling genes (epistasis) may be more important, since, for traits with lower QTL numbers, the multiplicative effects control genes (epistasis) may be of greater magnitude in proportional terms, as the individual effect of each gene is greater than in traits controlled by a greater number of QTL (BARBOSA et al., 2021). This is a direct result of its modeling philosophy, which tries to approximate a (possibly higher-order) function with a set of basic functions that are locally lower-order, so it has more power and flexibility to model relationships that are almost additive or involve interactions in at most a few variables (FRIEDMAN, 1991).

The frame is different when there is a large number of predictor variables with high correlation and the response characteristic has high genetic (dominance and epistasis effect) and environmental noise. When the trait involves a greater number of QTL, there is a greater chance that a marker explains in genetic terms the same variation of another marker, in addition to providing a greater action of the

environmental effect, thus impairing the prediction efficiency. The excess of markers associated with a reduced number of genotypic observations can also lead to multicollinearity problems (SOUSA et al., 2021). As FRIEDMAN (1991) points out, MARS is not particularly robust against correlated inputs and relies heavily on data to infer the process model, in these cases, MARS loses explanatory power. Thus, analyses should use an optimal set of informative SNPs, according to expectations regarding the number of QTLs, to adopt the best analytical strategy, maximizing predictive accuracy estimates Another method susceptible to multicollinearity vulnerability is DT. If two predictors are highly collinear, MARS or DT has to make an arbitrary knot or split selection that minimizes the residual sum of squares, this can profoundly affect all subsequent selections and final predictions (DE VEAUX; UNGAR, 1994).

Also, more generally, recursive partitioning methods have difficulties when the dominant interactions involve a small fraction of the total number of variables, so one cannot discern whether the approximation function approximates a simple one, such as linear or additive, or if it involves complex interactions between variables (FRIEDMAN, 1991). This explains why MARS did not perform well in genomic regions where strong genetic interactions are present, such as for traits from 40 to 120 QTL. However, for scenarios where these effects are diluted, high QTL number (240 to 480), MARS has high predictive potential and lower model variance. Thus, it is notable that the excellent results obtained for the non-additive MARS show that this approach should be considered for GWS.

The greater number of genotypic classes in scenarios with a greater number of QTL reduces the representativeness of each genotypic combination in the training set and overparameterization of the model (BARBOSA et al., 2021). In this context, it was these restrictions that led these algorithms to not present such satisfactory results when the trait is polygenic, mainly DT. Low DT efficiency was also founded by SOUSA et al. (2021) to predict the genetic values for rust incidence in *Coffea arabica* and by BARBOSA et al. (2021) for simulated features with epistatic effects with 16 or more QTL.

The approaches based on decision trees (BA, BO, and RF) showed excellent results regarding the accuracy of the GEBVS prediction for traits with many QTL. A differential of the BA and RF approaches is the resampling of the original data in sub-

samples (bootstrap) to perform the prediction according to a number of determined trees. This resampling of data brings concrete benefits for prediction in these cases, allowing for the easy evaluation of poorly predicted samples and possible discrepancies (DIAZ-URIARTE, 2007). BA analyzes its main effects on variance and can make forecasting more robust by decreasing the variance lead time and RF not only combines a large number of decision trees to reduce forecast variance like BA but also decreases dependency between decision trees by projecting random features to obtain a much smaller prediction error (FULEKY, 2020). As a result, these methods perform better to maximize prediction in a target population, suggesting that bootstrapping can be performed by other methods to achieve better prediction results.

As it is a gradient-enhancing algorithm that has a learning rate, the BO method combines individually weak predictors to produce a strong classifier (JAMES et al., 2021), thus allowing a better prediction of the genetic effects of individuals, as observed in this study. However, BO was the method with the worst result for RMSE, explained due to the intensive use of input values (SNPs) to build the trees, thus reducing the bias of the model estimates, but at the same time increasing the variance, as additional parameters are being adjusted to produce a better fit of the data.

Neural network approaches, as used in this study, MLP and RBF, apparently, are not affected by correlated inputs. The MLP and RBF methods are defended for being efficient in capturing nonlinear effects, in this case, provided by interallelic interactions (BARBOSA et al., 2021). However, both RBF and MLP were harmed by the excess of ineffective markers, showing lower performance compared to other methods. Similar results were also found by BARBOSA et al. (2021) and SANT'ANNA et al. (2021). However, neural networks were efficient in predicting traits with many QTL, especially when the phenotypic value of the trait was mostly due to genetic value (heritability = 80%).

G-BLUP considers the interaction between marker pairs and relies on the LD between SNPs and QTL, moreover, when QTL are in strong LD and the use of unweighted genomic relationship matrix in G-BLUP can cause upward bias in the heritability estimate (FERNANDO et al., 2017; LEGARRA, 2016; MATHEW; LÉON; SILLANPÄÄ, 2018; SPEED et al., 2012). However, if only a few markers are important, the technique is hampered by this bias, as confirmed in this study. On the other hand, the G-BLU P was highlighted in the performance of the prediction of the GEBVs,

presenting very similar results to the Family Network methods for the characteristics with more QTL. Some markers are more informative for some traits than others, this increase in the amount of information using the genomic matrix G (genomic relationship matrix) can sometimes lead to better and more accurate estimates and predictions (SOUSA et al., 2021). Similar results to those found in this study were found in BARBOSA et al. (2021) and WANG et al. (2018). These results are justified for G-BLUP due to the principle that genomic predictions are based on the relatedness derived from all markers (WANG et al., 2018), so when more markers have a genetic effect the prediction accuracy increases. However, divergent results were found by SANT'ANNA et al. (2020). These authors observed that G-BLUP presented better results when compared to RBF in any heritability scenario.

Although MARS performs the selection of SNPs, eliminating a large number of markers, the performance of this method showed a greater difference for the RBF, MLP, and G-BLUP methods, which consider all markers, and BA, RF, and BO in the scenarios between 40 to 280 QTL. This can be explained by the fact that the F2 population has a high rate of linkage disequilibrium (LD), due to the recombination process. This LD can then cause false-positive signals for some loci, which have no connection with the studied trait in question (MATHEW; LÉON; SILLANPÄÄ, 2018). So, the SNPs closest to a QTL are not sampled often enough and the QTL signal may be captured by more distant SNPs, consequently, the signal from a QTL to MARS may be blurred compared to other methods.

The main limitation of the additive MARS is that the model is constrained to be additive. With many variables, important interactions can be missed. On the other hand, as the model is additive, we can examine the effect of each marker on the prediction of GEBVs individually. Furthermore, the model can be represented in a way that separately identifies additive contributions and those associated with different multivariate interactions, being useful for future studies applied to Genomic Wide Association Studies (GWAS). MARS also has several ways of tuning that can improve the predictability of the traits, for example the change in gamma, where the model becomes more flexible to detect close variables for inclusion in the model for the forward phase.

In general, as the number of QTLs increases, the total genetic variation is expected to be divided among the QTLs, which can reduce the efficiency of methods

to estimate small QTL effects and lead to a loss of precision (GHAFOURI-KESBI et al., 2017; RESENDE et al., 2012). This is confirmed only for traits that present stronger effects between interactions in the same linkage group, such as for traits with 40 QTL, since, as they have a smaller number of QTL in a single linkage group, the expression of interactions between these QTL is stronger. On the other hand, the increase in efficiency for a greater number of QTL can be attributed to the excess of marks with null effects, which can impair the accuracy of the methods (BARBOSA et al., 2021; SOUSA et al., 2021).

Each technique has its specificity and must be evaluated in a wide set of data so that the decision on which method to use is correctly made (TONG; NIKOLOSKI, 2021). It is rare that more than one technique is used when performing GWS analyses, but these results align with the view that evaluating multiple methods is a useful strategy to ensure that uncertainty in data is considered from multiple angles.

Conclusions

MARS ability to simplify complex relationships is quite pertinent to GWS, as most traits of interest in plant breeding are affected by complex interactions of biological, environmental, and management conditions.

Non-additive MARS is better for predicting breeding values than additive MARS in the scenarios evaluated. The additive and non-additive MARS methods showed superior results in the prediction of genetic values in characters with dominant and epistatic effects for scenarios with eight QTL in relation to G-BLUP methods, neural networks, and other machine learning methods.

The use of different statistical methods, neural networks, and machine learning, such as MARS, to estimate genetic values resulted in different consequences influenced by the complexity and particularity of the analyzed traits. Therefore, it is recommended that when evaluating the prediction of genetic values, the use of multiple approaches is used, in order to choose the best method.

Acknowledgment

The authors are grateful for the financial support of the Coordination for the

Improvement of Higher Education Personnel - CAPES financial code 001, and the National Council for Scientific and Technological Development - CNPq.

References

ALKIMIM, Emilly Ruas; CAIXETA, Eveline Teixeira; SOUSA, Tiago Vieira; RESENDE, Marcos Deon Vilela; SILVA, Felipe Lopes; SAKIYAMA, Ney Sussumu; ZAMBOLIM, Laércio. Selective efficiency of genome-wide selection in Coffea canephora breeding. **Tree Genetics and Genomes**, *[S. l.]*, v. 16, n. 3, 2020. DOI: 10.1007/s11295-020-01433-3.

ALTINOK, Gulsah; KARAGOZ, Pinar; BATMAZ, Inci. Learning to rank by using multivariate adaptive regression splines and conic multivariate adaptive regression splines. **Computational Intelligence**, *[S. l.]*, n. April, p. 1–38, 2020. DOI: 10.1111/coin.12413.

BARBOSA, Ivan Paiva; SILVA, Michele Jorge; COSTA, Weverton Gomes; SANT'ANNA, Isabela Castro; NASCIMENTO, Moysés; CRUZ, Cosme Damião. Genome-enabled prediction through machine learning methods considering different levels of trait complexity. **Crop Science**, *[S. l.]*, v. 61, n. 3, p. 1890–1902, 2021. DOI: 10.1002/csc2.20488.

BOEHMKE, Brad; GREENWELL, Brandon. Random Forests. *In*: **Hands-On Machine Learning with R**. [s.l.] : Chapman and Hall/CRC, 2019. v. 45p. 203–219. DOI: 10.1201/9780367816377-11. Disponível em: https://www.taylorfrancis.com/books/9781000730197/chapters/10.1201/9780367816377-11.

BREIMAN, Leo. Bagging Predictors. **Machine Learning**, *[S. l.]*, v. 24, n. 421, p. 123–140, 1996. DOI: 10.1007/BF00058655.

CHANG, Cheng Ding; WANG, Chien Chih; JIANG, Bernard C. Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors. **Expert Systems with Applications**, *[S. l.]*, v. 38, n. 5, p. 5507–5513, 2011. DOI: 10.1016/j.eswa.2010.10.086. Disponível em: http://dx.doi.org/10.1016/j.eswa.2010.10.086.

COOK, Nancy R.; ZEE, Robert Y. L.; RIDKER, Paul M. Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. **Statistics in Medicine**, *[S. l.]*, v. 23, n. 9, p. 1439–1453, 2004. DOI: 10.1002/sim.1749.

COSTA, Weverton Gomes; BARBOSA, Ivan de Paiva; SOUZA, Jacqueline Enequio; CRUZ, Cosme Damião; NASCIMENTO, Moysés; OLIVEIRA, Antonio Carlos Baião. Machine learning and statistics to qualify environments through multi-traits in Coffea arabica. **PLOS ONE**, *[S. l.]*, v. 16, n. 1, p. 1–21, 2021. DOI:

10.1371/journal.pone.0245298. Disponível em: https://dx.plos.org/10.1371/journal.pone.0245298.

COSTER, Albart; BASTIAANSEN, John W. M.; CALUS, Mario P. L.; VAN ARENDONK, Johan A. M.; BOVENHUIS, Henk. Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. **Genetics Selection Evolution**, *[S. l.]*, v. 42, n. 1, p. 1–11, 2010. DOI: 10.1186/1297-9686-42-9.

COUTINHO, Alisson Esdras; NEDER, Diogo Gonçalves; SILVA, Mairykon Coêlho; ARCELINO, Eliane Cristina; BRITO, Silvan Gomes; CARVALHO FILHO, José Luiz Sandes. Prediction of phenotypic and genotypic values by BLUP/GWS and neural networks. **Revista Caatinga**, *[S. l.]*, v. 31, n. 3, p. 532–540, 2018. DOI: 10.1590/1983-21252018v31n301rc. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1983-21252018000300532&lng=en&tlng=en.

CRUZ, Cosme Damião. **Princípios de genética quantitativa**. 2. ed. ed. Viçosa, MG: Editora UFV, 2012.

CRUZ, Cosme Damião. GENES - Software para análise de dados em estatística experimental e em genética quantitativa. **Acta Scientiarum - Agronomy**, *[S. l.]*, v. 35, n. 3, p. 271–276, 2013. DOI: 10.4025/actasciagron.v35i3.21251.

CRUZ, Cosme Damião. Genes software – extended and integrated with the R, Matlab and Selegen. **Acta Scientiarum - Agronomy**, *[S. l.]*, v. 38, n. 4, p. 547–552, 2016. DOI: 10.4025/actasciagron.v38i4.32629.

CRUZ, Cosme Damião; NASCIMENTO, Moysés. **Inteligência Computacional aplicada ao melhoramento genético**. Viçosa, MG: Editora UFV, 2018.

DE VEAUX, Richard D.; UNGAR, Lyle H. Multicollinearity: A tale of two nonparametric regressions. *[S. l.]*, p. 393–402, 1994. DOI: 10.1007/978-1-4612-2660-4_40.

DIAZ-URIARTE, Ramón. GeneSrF and varSelRF: A web-based tool and R package for gene selection and classification using random forest. **BMC Bioinformatics**, *[S. l.]*, v. 8, p. 1–7, 2007. DOI: 10.1186/1471-2105-8-328.

EVERINGHAM, Y. L.; SEXTON, J. An introduction to Multivariate Adaptive Regression Splines for the cane industry. **33rd Annual Conference of the Australian Society of Sugar Cane Technologists 2011, ASSCT 2011**, *[S. l.]*, n. December 2015, p. 255–268, 2011.

FERNANDO, R. L.; CHENG, H.; SUN, X.; GARRICK, D. J. A comparison of identity-by-descent and identity-by-state matrices that are used for genetic evaluation and estimation of variance components. **Journal of Animal Breeding and Genetics**, *[S. l.]*, v. 134, n. 3, p. 213–223, 2017. DOI: 10.1111/jbg.12275.

FRIEDMAN, Jerome H. Multivariate Adaptative regression Splines. **The Annals of Statistics**, *[S. l.]*, v. 19, n. 1, p. 1–141, 1991. Disponível em: http://projecteuclid.org/euclid.aop/1176996548.

FULEKY, Peter. **Macroeconomic Forecasting in the Era of Big Data**. [s.l: s.n.]. v. 52 Disponível em: http://link.springer.com/10.1007/978-3-030-31150-6.

GHAFOURI-KESBI, Farhad; RAHIMI-MIANJI, Ghodratollah; HONARVAR, Mahmood; NEJATI-JAVAREMI, Ardeshir. Predictive ability of Random Forests, Boosting, Support Vector Machines and Genomic Best Linear Unbiased Prediction in different scenarios of genomic evaluation. **Animal Production Science**, *[S. l.]*, v. 57, n. 2, p. 229–236, 2017. DOI: 10.1071/AN15538.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The elements of statistical learning: Data mining, inference, and prediction**. 2. ed. ed. New York, NY, USA: Springer, 2009. DOI: 10.1007/978-1-4419-9863-7_941.

JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert. An Introduction to Statistical Learning. *In*: **Springer Texts in Statistics**. [s.l: s.n.]. p. 612. DOI: 10.1007/978-1-0716-1418-1_1. Disponível em: https://link.springer.com/10.1007/978-1-0716-1418-1_1.

LEGARRA, Andres. Comparing estimates of genetic variance across different relationship models. **Theoretical Population Biology**, *[S. l.]*, v. 107, p. 26–30, 2016. DOI: 10.1016/j.tpb.2015.08.005. Disponível em: http://dx.doi.org/10.1016/j.tpb.2015.08.005.

LI, Bo; ZHANG, Nanxi; WANG, You Gan; GEORGE, Andrew W.; REVERTER, Antonio; LI, Yutao. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. **Frontiers in Genetics**, *[S. l.]*, v. 9, n. JUL, p. 1–20, 2018. DOI: 10.3389/fgene.2018.00237.

LIEW, Bernard X. W.; PEOLSSON, Anneli; RUGAMER, David; WIBAULT, Johanna; LÖFGREN, Hakan; DEDERING, Asa; ZSIGMOND, Peter; FALLA, Deborah. Clinical predictive modelling of post-surgical recovery in individuals with cervical radiculopathy: a machine learning approach. **Scientific Reports**, *[S. l.]*, v. 10, n. 1, p. 1–10, 2020. DOI: 10.1038/s41598-020-73740-7. Disponível em: https://doi.org/10.1038/s41598-020-73740-7.

LIN, Hui Yi; WANG, Wenquan; LIU, Yung Hsin; SOONG, Seng Jaw; YORK, Timothy P.; MYERS, Leann; HU, Jennifer J. Comparison of multivariate adaptive regression splines and logistic regression in detecting SNP-SNP interactions and their application in prostate cancer. **Journal of Human Genetics**, *[S. l.]*, v. 53, n. 9, p. 802–811, 2008. DOI: 10.1007/s10038-008-0313-z.

MA, Wenlong; QIU, Zhixu; SONG, Jie; CHENG, Qian; MA, Chuang. DeepGS: Predicting phenotypes from genotypes using Deep Learning. **bioRxiv**, *[S. l.]*, 2017. DOI: 10.1101/241414.

MATHEW, Boby; LÉON, Jens; SILLANPÄÄ, Mikko J. A novel linkage-disequilibrium corrected genomic relationship matrix for SNP-heritability estimation and genomic prediction. **Heredity**, *[S. l.]*, v. 120, n. 4, p. 356–368, 2018. DOI: 10.1038/s41437-017-0023-4. Disponível em: http://dx.doi.org/10.1038/s41437-017-0023-4.

MATLAB. **Natick, Massachusetts: The MathWorks Inc.**Natick, MassachusettsThe MathWorks Inc., , 2019.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, *[S. l.]*, v. 157, n. 4, p. 1819–1829, 2001.

MILBORROW, Stephen. Notes on the earth package. *[S. l.]*, p. 1–68, 2019. Disponível em: http://www.milbo.org/doc/earth-notes.pdf%0Ahttps://cran.r-project.org/web/packages/earth/index.html.

MOURA, Ernandes Guedes; PAMPLONA, Andrezza Kellen Alves; BALESTRE, Marcio. Functional models in genome-wide selection. **Plos One**, *[S. l.]*, v. 14, n. 10, p. e0222699, 2019. DOI: 10.1371/journal.pone.0222699. Disponível em: https://dx.plos.org/10.1371/journal.pone.0222699.

PEIXOTO, Leonardo Azevedo; LAVIOLA, Bruno Galvêas; ALVES, Alexandre Alonso; ROSADO, Tatiana Barbosa; BHERING, Leonardo Lopes. Breeding Jatropha curcas by genomic selection: A pilot assessment of the accuracy of predictive models. **PLoS ONE**, *[S. l.]*, v. 12, n. 3, p. 1–16, 2017. DOI: 10.1371/journal.pone.0173368.

PRASAD, Anantha M.; IVERSON, Louis R.; LIAW, Andy. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. **Ecosystems**, *[S. l.]*, v. 9, n. 2, p. 181–199, 2006. DOI: 10.1007/s10021-005-0054-1.

R CORE TEAM; COMPUTING, R. Foundation for Statistical; TEAM, R. Core. **R: A Language and Environment for Statistical Computing**. 2020. Disponível em: https://www.r-project.org/. Acesso em: 1 jul. 2020.

RESENDE, Marcos D. V. 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 Phytologist**, *[S. l.]*, v. 194, n. 1, p. 116–128, 2012. DOI: 10.1111/j.1469-8137.2011.04038.x.

SANT'ANNA, Isabela Castro; NASCIMENTO, Moyses; SILVA, Gabi Nunes; CRUZ, Cosme Damião; AZEVEDO, Camila Ferreira; GLORIA, Leonardo Siqueira; FONSECA E SILVA, Fabyano. Genome-enabled prediction of genetic values for using radial basis function neural networks. **Functional Plant Breeding Journal**, *[S. l.]*, v. 1, n. 2, p. 1–8, 2020. DOI: 10.35418/2526-4117/v1n2a1. Disponível em: http://fpbjournal.com/fpbj/index.php/fpbj/article/view/57/17.

SANT'ANNA, Isabela de Castro; GOUVÊA, Ligia Regina Lima; MARTINS, Maria Alice; SCALOPPI JUNIOR, Erivaldo José; DE FREITAS, Rogério Soares; GONÇALVES, Paulo de Souza. Genetic diversity associated with natural rubber quality in elite

genotypes of the rubber tree. **Scientific Reports**, *[S. l.]*, v. 11, n. 1, p. 1–10, 2021. DOI: 10.1038/s41598-020-80110-w. Disponível em: https://doi.org/10.1038/s41598-020-80110-w.

SCHNABLE, Patrick S.; SPRINGER, Nathan M. Progress toward understanding heterosis in crop plants. **Annual Review of Plant Biology**, *[S. l.]*, v. 64, p. 71–88, 2013. DOI: 10.1146/annurev-arplant-042110-103827.

SHAO, Yuehjen E.; HOU, Chia Ding; CHIU, Chih Chou. Hybrid intelligent modeling schemes for heart disease classification. **Applied Soft Computing Journal**, *[S. l.]*, v. 14, n. PART A, p. 47–52, 2014. DOI: 10.1016/j.asoc.2013.09.020. Disponível em: http://dx.doi.org/10.1016/j.asoc.2013.09.020.

SILVA, Gabi Nunes; TOMAZ, Rafael Simões; SANT'ANNA, Isabela Castro; NASCIMENTO, Moysés; BHERING, Leonardo Lopes; CRUZ, Cosme Damião. Neural networks for predicting breeding values and genetic gains. **Scientia Agricola**, *[S. l.]*, v. 71, n. 6, p. 494–498, 2014. DOI: 10.1590/0103-9016-2014-0057.
SINGH, B. D.; SINGH, A. K. **Marker-assisted plant breeding: Principles and practices**. [s.l: s.n.]. DOI: 10.1007/978-81-322-2316-0.

SOLANO MEZA, Johanna Karina; ORJUELA YEPES, David; RODRIGO-ILARRI, Javier; CASSIRAGA, Eduardo. Predictive analysis of urban waste generation for the city of Bogotá, Colombia, through the implementation of decision trees-based machine learning, support vector machines and artificial neural networks. **Heliyon**, *[S. l.]*, v. 5, n. 11, p. e02810, 2019. DOI: 10.1016/j.heliyon.2019.e02810. Disponível em: https://doi.org/10.1016/j.heliyon.2019.e02810.

SOUSA, Ithalo Coelho et al. Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. **Scientia Agricola**, *[S. l.]*, v. 78, n. 4, p. 1–8, 2021. DOI: 10.1590/1678-992x-2020-0021. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-90162021000401102&tlng=en.

SOUSA, Tiago Vieira; CAIXETA, Eveline Teixeira; ALKIMIM, Emilly Ruas; OLIVEIRA, Antonio Carlos Baião; PEREIRA, Antonio Alves; SAKIYAMA, Ney Sussumu; ZAMBOLIM, Laércio; RESENDE, Marcos Deon Vilela. Early Selection Enabled by the Implementation of Genomic Selection in Coffea arabica Breeding. **Frontiers in Plant Science**, *[S. l.]*, v. 9, n. January, p. 1–12, 2019. DOI: 10.3389/fpls.2018.01934. Disponível em: https://www.frontiersin.org/article/10.3389/fpls.2018.01934/full.

SPEED, Doug; HEMANI, Gibran; JOHNSON, Michael R.; BALDING, David J. Improved heritability estimation from genome-wide SNPs. **American Journal of Human Genetics**, *[S. l.]*, v. 91, n. 6, p. 1011–1021, 2012. DOI: 10.1016/j.ajhg.2012.10.010. Disponível em: http://dx.doi.org/10.1016/j.ajhg.2012.10.010.

TANG, Jie et al. Application of Machine-Learning Models to Predict Tacrolimus Stable Dose in Renal Transplant Recipients. **Scientific Reports**, *[S. l.]*, v. 7, n. February, 2017. DOI: 10.1038/srep42192.

TAYLAN, Pakize; WEBER, Gerhard Wilhelm. CG-Lasso Estimator for Multivariate Adaptive Regression Spline. *In*: TAS, Kenan; BALEANU, Dumitru; MACHADO, J. A. Tenreiro (org.). **Mathematical Methods in Engineering: Aplications in Dynamics of Complex Systems**. [s.l.] : Springer International Publishing AG, 2019. p. 121–136. DOI: 10.1007/978-3-319-90972-1_9.

TONG, Hao; NIKOLOSKI, Zoran. Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. **Journal of Plant Physiology**, *[S. l.]*, v. 257, p. 153354, 2021. DOI: 10.1016/j.jplph.2020.153354. Disponível em: https://doi.org/10.1016/j.jplph.2020.153354.

WANG, Jiabo; ZHOU, Zhengkui; ZHANG, Zhe; LI, Hui; LIU, Di; ZHANG, Qin; BRADBURY, Peter J.; BUCKLER, Edward S.; ZHANG, Zhiwu. **Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traitsHeredity**, 2018. DOI: 10.1038/s41437-018-0075-0.

YABE, Shiori; HARA, Takashi; UENO, Mariko; ENOKI, Hiroyuki; KIMURA, Tatsuro; NISHIMURA, Satoru; YASUI, Yasuo; OHSAWA, Ryo; IWATA, Hiroyoshi. Potential of genomic selection in mass selection breeding of an allogamous crop: An empirical study to increase yield of common buckwheat. **Frontiers in Plant Science**, *[S. l.]*, v. 9, n. March, p. 1–12, 2018. DOI: 10.3389/fpls.2018.00276.

YORK, Timothy P.; EAVES, Lindon J.; VAN DEN OORD, Edwin J. C. G. Multivariate adaptive regression splines: A powerful method for detecting disease-risk relationship differences among subgroups. **Statistics in Medicine**, *[S. l.]*, v. 25, n. 8, p. 1355–1367, 2006. DOI: 10.1002/sim.2292.

ZHANG, Haohao; YIN, Lilin; WANG, Meiyue; YUAN, Xiaohui; LIU, Xiaolei. Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. **Frontiers in Genetics**, *[S. l.]*, v. 10, n. MAR, p. 1–10, 2019. DOI: 10.3389/fgene.2019.00189.

ZHANG, Wengang; GOH, Anthony T. C. Multivariate adaptive regression splines and neural network models for prediction of pile drivability. **Geoscience Frontiers**, *[S. l.]*, v. 7, n. 1, p. 45–52, 2016. DOI: 10.1016/j.gsf.2014.10.003. Disponível em: http://dx.doi.org/10.1016/j.gsf.2014.10.003.

ZHENG, Gang; ZHANG, Wenbin; ZHOU, Haizuo; YANG, Pengbo. Multivariate adaptive regression splines model for prediction of the liquefaction-induced settlement of shallow foundations. **Soil Dynamics and Earthquake Engineering**, *[S. l.]*, v. 132, n. August 2019, p. 106097, 2020. DOI: 10.1016/j.soildyn.2020.106097. Disponível em: https://doi.org/10.1016/j.soildyn.2020.106097.

ZINGARETTI, Laura M.; GEZAN, Salvador Alejandro; FERRÃO, Luis Felipe V.; OSORIO, Luis F.; MONFORT, Amparo; MUÑOZ, Patricio R.; WHITAKER, Vance M.; PÉREZ-ENCISO, Miguel. Exploring Deep Learning for Complex Trait Genomic Prediction in Polyploid Outcrossing Species. **Frontiers in Plant Science**, *[S. l.]*, v. 11,

n. February, p. 1–14, 2020. DOI: 10.3389/fpls.2020.00025. Disponível em: https://www.frontiersin.org/article/10.3389/fpls.2020.00025/full.

5. ARTIGO 2

Importância de marcadores para detecção de QTLs por métodos de aprendizado de máquinas

Resumo

A identificação de marcadores associados a locus de características quantitativas (*Quantitative Trait Locus* – QTLs) auxilia pesquisadores em estratégias de predição de genoma, de estudos de associação e de importância de covariáveis. Métodos paramétricos (G-BLUP) e alguns métodos de aprendizado de máquina (ML) e redes neurais artificiais (ANN) já foram utilizados para esse fim. No entanto, ainda não existe um consenso sobre a estimação da importância dos marcadores por esses métodos em características com efeitos não-aditivos. Esse trabalho teve com objetivo avaliar os métodos de ML na associação de marcadores com as regiões de presença do QTLs. Uma população F2 foi simulada, considerando características com diferentes estimativas de herdabilidade e números de genes na presença de epistasia. A maioria dos métodos apresentaram maior índice de acertos na identificação dos marcadores em cenários com menor número de QTLs e com maior herdabilidade. A MARS 3 e o Boosting apresentaram alta capacidade de identificar os marcadores de importância. O maior índice de erros também ocorreu em cenários com menor número de QTLs, mas com menor herdabilidade. De modo geral, nos cenários de 40 ou mais QTLs, o aumento no número de QTLs também afetou positivamente o índice relativo para a maioria dos métodos. Contudo, os melhores resultados foram encontrados para o cenário com maior herdabilidade e com 8 QTLs. Os métodos MARS 1, MARS 2, Boosting e Bagging foram os mais efetivos na detecção de marcadores importantes, principalmente para as características com 8 e 240 QTLs. O uso de diferentes métodos resultou em diferentes consequências influenciadas pela complexidade e particularidade das características analisadas. Portanto, recomenda-se que ao avaliar a importância de marcadores, o uso de múltiplas abordagens seja utilizado, a fim de escolher o melhor método a ser utilizado.

Palavras-chaves: *Quantitative Trait Locus*. Efeitos não-aditivos. Epistasia. MARS. Estudos de Associação Genômica.

Introdução

A identificação de importantes regiões cromossômicas subjacentes a importantes características agronômicas é fundamental para entender melhor a arquitetura genética da característica, descrever a variabilidade genética e capturar os melhores marcadores informativos populacionais. Estudos de associação de loci de características quantitativas (*Quantitative Trait Locis* - QTLs) revelam que existe impacto de regiões genômicas específicas no desempenho de genótipos para uma determinada característica.

Avaliar o conjunto mais discriminador de marcadores dentro dessas regiões genômicas previamente associadas foi descrito por autores em melhoramento animal (SCHIAVO et al., 2020), que implantaram várias abordagens estatísticas para esse fim. Existem várias publicações nas quais métodos paramétricos (RR-BLUP, G-BLUP) e métodos não paramétricos de aprendizado de máquina (ML) e redes neurais artificiais (ANN) foram aplicados para selecionar um importante subconjunto de marcadores para estratégias de predição de genoma, estudos de associação e importância de covariáveis (BERMINGHAM et al., 2015; DE LOS CAMPOS et al., 2013; HOWARD; CARRIQUIRY; BEAVIS, 2014; JACQUIN; CAO; AHMADI, 2016; LI et al., 2018; LIANG; KELEMEN, 2008; OKSER et al. , 2014; WALDMANN, 2016).

Modelos paramétricos que incluem efeitos epistáticos tornam-se inviáveis devido a muitos parâmetros a serem estimados (esforço computacional), e esta é a grande vantagem do ML sobre a modelagem paramétrica. O ML e as ANNs, provaram ser úteis para prever o desempenho do genótipo usando dados SNP simulados com interações aditivas e epistáticas entre alelos, principalmente quando a arquitetura genética subjacente era controlada por epistasia (BARBOSA et al., 2021; HOWARD; CARRIQUIRY; BEAVIS, 2014, LIN et al. 2008).

Bedo et al. (2008) realizaram pesquisas de validação de QTL em um procedimento de mapeamento de precisão por meio de métodos de ML usando conjuntos de dados fenotípicos e genotípicos de cevada duplo haploides derivados de F1. A importância de cada grupo de marcadores foi avaliada de acordo com a mudança na variância explicada após a eliminação de cada marcador. Sua estratégia detectou picos mais estreitos de efeitos de QTL, levando a uma identificação alvo de marcadores importantes dentro do intervalo QTL.

A estratégia de avaliação da importância dos marcadores abordada neste estudo difere da apresentada por Bedo et al. (2008) e se baseia principalmente na utilização de um índice relativo (IR) para avaliar o mérito dos métodos tanto na detecção dos marcadores associados à região de presença do QTL, quanto na inferência errônea dos métodos. Ainda são poucos os estudos descritos na literatura sobre a identificação de marcadores moleculares importantes dentro de QTLs que controlam as principais características no melhoramento de plantas. Portanto, o objetivo deste estudo é avaliar a eficiência de procedimentos de identificação de marcadores por meio de aprendizado de máquina, redes neurais artificiais e *Genomic Best linear Unbiased Prediction* (G-BLUP).

Material e métodos

Simulação do genoma da população

Uma população F2 constituída por 1000 indivíduos e 10 grupos de ligação com 200 cM cada foi simulada. A constituição genotípica de cada indivíduo da população foi estabelecida admitindo pool gamético de 5000 gametas, por genitor, em cada fertilização. A simulação dos genótipos de cada indivíduo foi feita em relação a 401 marcadores moleculares codominantes e equidistantes, por grupo de ligação, totalizando 4010 marcadores espaçados entre si a 0.5 cM. Estes foram codificados como 1, 0 e -1 para representar os indivíduos A_iA_i, A_iA_j e A_jA_j, respectivamente.

Simulação de valores genotípicos e fenotípicos de características

Foram simulados valores genotípicos dos indivíduos para 12 características com herança controlada por oito a 240 locos com dois alelos cada. Adotou-se valores de herdabilidade iguais a 0.5 ou 0.8 (Tabela 1).

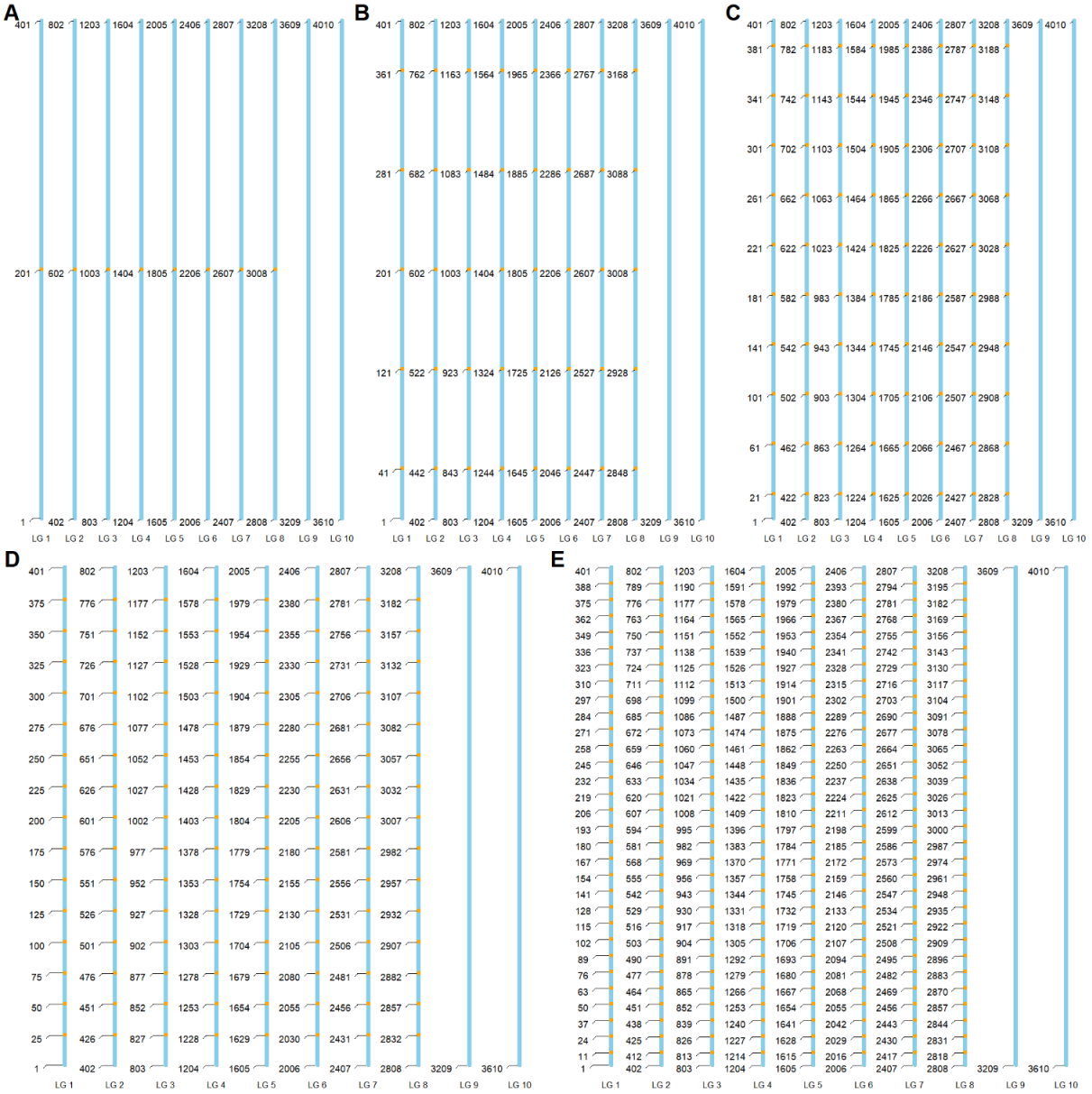
Tabela 1: Número de locos controladores e herdabilidades (h^2) das 10 características simuladas (C1 a C10).

h^2	Número de locos controladores da característica				
	8	**40**	**80**	**120**	**240**
0,5	C1	C2	C3	C4	C5
0,8	C6	C7	C8	C9	C10

Fonte: O autor.

Os locos controladores foram distribuídos igualmente entre os oito primeiros grupos de ligação (Figura 1). Oito QTLs controlaram as características C1 e C5, definidas pelos marcadores centrais dos oito primeiros grupos de ligação (posições 201, 602, 1003, 1404, 1805, 2206, 2607, 3008, respectivamente). Para as características C2 a C5 e C6 a C10 os QTLs foram distribuídos mantendo uma distância aproximada entre eles, dentro dos primeiros 8 grupos de ligação conforme a Figura 1.

Figura 1 - Modelo de distribuição de loci nos grupos de ligação (GL) para as 10 características simuladas. Modelo de distribuição dos QTLs para as características: (A) C1 e C6 com os marcadores centrais nos oito primeiros grupos de ligação; (B) C2 e C7, (C) C3 e C8, (D) C4 e C9, e (E) C5 e C10, com 40, 80, 120 e 240 QTLs, respectivamente, distribuídos equidistantemente dentro dos primeiros oito GL.



Fonte: O autor.

Os valores fenotípicos totais expressos por um determinado indivíduo para as 10 características foram simulados considerando média igual a 100 e coeficiente de variação igual a 12%, com grau médio de dominância igual a 0.5. Foi considerado modelo aditivo-epistático, conforme apresentado a seguir:

$$Y_{ij} = \mu + \sum_{j} \alpha_{ji} + \sum_{j} \alpha_{ji}\alpha_{ji+1} + e_i$$

Em que Y_i é o valor fenotípico para observação i; μ é a média geral; α_j é o efeito do alelo favorável no loco j do indivíduo i, ou seja, assume os valores $u + a_i$, $u + d_i$ e $u - a_i$ para os valores genotípicos associados às classes AA, Aa e aa, respectivamente, sendo u a média entre o homozigoto dominante (AA) e o homozigoto recessivo (aa). As classes foram identificadas pela codificação 1, 0 ou -1, respectivamente; $\alpha_{ji}\alpha_{ji+1}$ representa a interação entre alelos favoráveis em diferentes loci. A estrutura de variância dos resíduos foi dada por $e \sim N(0, V_e)$, onde $V_e = ((1 - h^2)V_g)/h^2$, onde V_e é a variância residual, V_g é a variância genotípica, e h^2 a herdabilidade no sentido amplo.

Importância de marcadores

Para verificar a acurácia na detecção de marcadores foram utilizadas diferentes técnicas de aprendizado de máquina (Multivariate Adaptive Regression Splines - MARS, Decision Tree - DT, Boosting - BO, Bagging - BAG e Random Forest -RF), redes neurais artificiais (Multilayer Perceptron Network - MLP e Radial Basis Function Network - RBF) e Genomic Best Linear Unbiased Prediction (GBLUP).

Devido à alta saturação (4010 marcadores) e proximidade entre os marcadores (0.5 cM) foi definido que a região considerada como a da presença do QTL seria representada por um bloco gênico correspondente a quatros marcadores anteriores e posteriores ao QTL. Como, numa população F2 é esperado que a correlação entre pares de marcadores codificados como sendo -1, 0 e 1 e segregando na proporção 1:2:1 seja de 0.9904 espaçados a 0.5 cM, esta correlação deverá ser de 0.96 se espaçado a 2 cM. Assim, a região correta de detecção do bloco gênico seria abrangente a nove marcadores, sendo o marcador central o correto.

Para um dado marcador identificado pela técnica como importante, esse marcador foi considerado como acerto se estivesse presente na região correta de detecção do QTL. Da mesma forma, se um marcador vizinho a ele também fosse detectado como importante e este também estiver presente na região correta de detecção do QTL, os dois marcadores contabilizaram apenas um acerto.

De modo contrário, para considerar a identificação do marcador pela técnica

como erro, o marcador não deveria estar presente na região correta de detecção do QTL, mesmo que um marcador vizinho a ele estivesse ou não na região correta de detecção do QTL. Caso, um marcador vizinho não estivesse presente na região correta de detecção do QTL, os dois marcadores contabilizaram dois erros. Além disso, a não identificação de um marcador importante na região do bloco gênico também contabilizou como um erro. Para facilitar a visualização, o número de acertos e de erros foram divididos pelo número de genes da respectiva característica avaliada, correspondendo ao índice de acerto e ao índice de erros, respectivamente.

A importância relativa (RI) de cada marcador foi determinada de acordo com a técnica utilizada. As descrições de como foram obtidas as importâncias dos marcadores para cada método foram descritas nos tópicos a seguir. As marcas foram consideradas como importantes para característica ao apresentarem RI maior do que 12.5.

Genomic Best Linear Unbiased Prediction (G-BLUP)

Para o G-BLUP a importância dos efeitos de marcadores foi realizada a partir da estrutura de valores genômicos preditos. O efeito de cada marca foi estimado pela equação (RESENDE et al., 2011):

$$\hat{\beta} = (X'X)^{-1}X'\widehat{GEBV}$$

Em que $\hat{\beta}$ é o valor estimado do efeito do marcador, X é a matriz de incidência dos efeitos fixos da média geral e $\widehat{GEBV}$ é o vetor dos valores genéticos genômicos estimados dos indivíduos, definido pela seguinte equação:

$$\widehat{GEBV}_i = \sum_{j=1}^{N} z_{ji}\hat{u}_{aj} + \sum_{j=1}^{N} z_{ji}\hat{u}_{dj} + \sum_{j=1}^{N} z_{ji}\hat{u}_{ej}$$

Onde $\widehat{GEBV}_i$ é o valor genético genômico estimado do indivíduo i, z_{ji} é a matriz de incidência dos efeitos aleatórios dos vetores dos efeitos aditivos, dominantes e epistáticos, u_a, u_d e u_e, respectivamente.

O seguinte modelo linear misto que inclui efeitos aditivos, de dominância e epistáticos para o método REML/G-BLUP foi utilizado:

$$y = Xb + Zu_a + Zu_d + Zu_{epi} + \varepsilon$$

Onde y é o vetor de observações fenotípicas; b é o vetor de efeitos fixos (média geral)

com matriz de incidência X; u_a, u_d e u_e são vetores de valores genéticos dos efeitos aditivos, dominantes e epistáticos, respectivamente; Z é a matriz de incidência para esses vetores; e ε é o vetor de erros aleatórios. A estrutura de variância foi dada por $u_a \sim N(0, G_a\sigma^2_{u_a})$; $u_d \sim N(0, G_d\sigma^2_{u_d})$; $u_e \sim N(0, G_e\sigma^2_{u_e})$ e o $\varepsilon \sim N(0, I\sigma^2_{\varepsilon})$, em que G_a, G_d e G_e são as matrizes de relacionamento genômico para os efeitos aditivos, dominantes e epistáticos, respectivamente, e $\sigma^2_{u_a}$, $\sigma^2_{u_d}$ e $\sigma^2_{u_e}$ são as variâncias aditiva, de dominância e epistática, respectivamente.

Redes Neurais Artificiais (Perceptron Multicamadas - MLP e Função de Base Radial - RBF)

Para os métodos baseados em redes neurais artificiais (Perceptron Multicamadas - MLP e Função de Base Radial - RBF) a importância dos marcadores foi realizada pela randomização dos valores dos marcadores, entre os indivíduos, na entrada da rede (COSTA et al., 2021). A randomização que resultou em redução da eficiência da rede, implica que o marcador é importante na impressão do resultado final, de acordo com a magnitude da redução da eficiência resultante após sua randomização (COSTA et al., 2021). Para cada marcador foi calculado a taxa de eficiência da rede modicada (AER_{mod}) e a taxa de eficiência da rede completa (AER_{com}). A magnitude da importância do marcador j foi dada pela soma das magnitudes nos conjuntos de teste realizados (($AER_j = \sum_i^k (AER_{mod(j)} - AER_{com})_i$). Para facilitar a interpretação, os valores de AER_j foram transformados em uma escala percentual, representando a importância relativa (*RI*) de cada marcador, conforme a equação abaixo:

$$RI_j(\%) = \frac{\sum_i^k (AER_{mod(j)} - AER_{com})_i}{\sum_j^m \sum_i^k (AER_{mod(j)} - AER_{com})_i} \; x\, 100$$

Em que RI_j é a importância relativa do marcador j em percentagem; $AER_{mod\,(j)}$ é a taxa de eficiência aparente do modelo modificado com a randomização do marcador j; AER_{com} é a taxa de eficiência aparente do modelo completo; i é o número de repetições (conjunto de teste), variando de 1 a k; e j é o número de marcadores, variando de 1 a m.

Para a MLP foi utilizado o algoritmo de treinamento *backpropagation* com

algoritmo de otimização de Levenberg-Marquardt, com topologia de uma camada de entrada, uma camada oculta composta por cinco a 15 neurônios e uma camada de saída. Foi estipulado que o número máximo de iterações seria 500. Cada marcador constituiu uma das entradas, portanto, esse apresentou 4010 entradas. Os dados moleculares, inicialmente simulados como -1, 0 e 1, foram padronizados para o intervalo entre 0 e 1 para garantir maior eficiência computacional. A função de ativação linear, denominada Purelin no Matlab, foi utilizada.

A RBF utiliza uma arquitetura do tipo *feedforward*. Este modelo também consiste em uma camada de entrada, uma camada oculta e uma camada de saída. O treinamento da RBF é híbrido e as informações da camada de entrada passam por um agrupamento linear do tipo *k-means* (CRUZ; NASCIMENTO, 2018). A camada oculta aplica uma transformação não linear do espaço de entrada a um espaço oculto de alta dimensão com uma função gaussiana. A camada de saída aplica uma transformação ao espaço oculto, fornecendo um vetor de saída para a rede. Para selecionar a melhor arquitetura da RBF, conforme a MLP, foram realizados testes preliminares. A arquitetura definida variou de acordo com cada fold, sendo que o número de neurônios variou de 5 a 50 e tamanho de raio de 30 a 50.

O erro quadrático médio (EQM) foi utilizado para avaliar a eficiência do treinamento das redes. Durante o treinamento foi estipulado que o processo de treinamento terminaria se EQM fosse inferior a 0.05 em uma interação qualquer. Caso durante o treinamento não fosse obtido EQM inferior a 0.05, o processo de treinamento terminaria ao atingir as 500 iterações.

Splines de regressão adaptativa múltipla (MARS)

Para a identificação da importância de marcadores para a MARS foi utilizada a função “evimp” do pacote “earth” do software R, baseado na soma de quadrados dos resíduos (RSS). O critério RSS primeiro calcula a diminuição no RSS para cada subconjunto em relação ao subconjunto anterior durante a fase *backward*. Em seguida, para cada variável, ele soma essas diminuições em todos os subconjuntos que incluem a variável. As variáveis que causam diminuições líquidas maiores no RSS são as consideradas mais importantes (MILBORROW, 2019).

O algoritmo proposto por Friedman (1991), Splines de Regressão Adaptativa Multivariada (MARS), considera uma expansão em funções lineares definidas por partes, chamadas de funções de base (BFs). Cada função é um spline linear por partes, com um nó no valor t. Estas duas BFs apresentadas são ditas um par reflexivo. A construção do modelo é como uma regressão linear progressiva, mas em vez de usar as entradas originais, usa-se funções do conjunto $C = \{(X_j - t)_+, (t - X_j)_+\}_{t \in \{x_{1j},\ x_{2j},\ldots,\ x_{Nj}\}\ j=1,2,\ldots,p}$ e/ou seus produtos.

O modelo MARS, que é uma combinação linear das BFs e/ou suas interações, é descrito por (HASTIE; TIBSHIRANI; FRIEDMAN, 2009), como:

$$f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X)$$

em que β_0 é a constante da regressão, β_m com m = 1, 2, ..., M, são os coeficientes da regressão e $h_m(X)$ é uma função em C, ou um produto de duas ou mais funções.

O processo de estimação dos parâmetros β_0 e β_m baseia-se na minimização da soma de quadrados do resíduo. A fase direta (*forward*) é executada nos dados de treinamento começando inicialmente a construção do modelo apenas com a função constante $h_0(X) = 1$, e todas as funções no conjunto C são funções candidatas. Em cada etapa subsequente, o par de base que produz a redução máxima no erro de treinamento é adicionado. Considerando um modelo com funções básicas M, o próximo par a ser adicionado ao modelo é (HASTIE; TIBSHIRANI; FRIEDMAN, 2009):

$$\hat{\beta}_{M+1} h_l(X)(X_j - t)_+ + \hat{\beta}_{M+2} h_l(X)(t - X_j)_+, h_l \in M$$

Onde $\hat{\beta}_{M+1}$ e $\hat{\beta}_{M+2}$ são coeficientes estimados pelo método dos mínimos quadrados, em conjunto com todos os outros coeficientes $M + 1$ no modelo. Esse processo de adição de BFs continua até que o modelo alcance um número máximo predeterminado, geralmente levando a um modelo propositalmente superdimensionado (ZHANG; GOH, 2016).

A fase *backward* melhora o modelo removendo os termos menos significativos até encontrar o melhor submodelo. Os subconjuntos do modelo são comparados usando o método de validação cruzada generalizada (GCV). O GCV é calculado como (HASTIE; TIBSHIRANI; FRIEDMAN, 2009):

$$GCV(\lambda) = \frac{\frac{1}{N}\sum_{i=1}^{N}[y_i - \hat{f}_\lambda(x_i)]^2}{\left[1 - \frac{C(M)}{N}\right]^2}$$

em que M é o número efetivo de parâmetros do modelo, $C(M)$ é uma função de custo para cada função de base incluída no submodelo desenvolvido, N é o número de conjuntos de dados usados na validação cruzada e $\hat{f}_\lambda(x_i)$ denota os valores preditos de MARS.

Para identificar a possível interação existente entre os QTLs foi utilizado modelos MARS com graus igual a 1, 2 e 3, sendo o modelo com grau 1 considerado modelo aditivo e os demais não-aditivos, os quais permitem interações entre marcadores. Para o critério de parada da fase *forward* o número máximo de termos no modelo adotado foi igual a 200, como *default* do pacote "*earth*" do R. Uma análise preliminar foi realizada para o segundo critério de parada (MILBORROW, 2019), na qual o incremento de um termo no modelo alterasse o coeficiente de determinação em menos do que 0.001 (*default*) até 0.05, sendo escolhido o melhor modelo que apresentou maior acurácia seletiva (R^2) para o conjunto de validação.

Árvore de Decisão (DT) e refinamentos (Bagging - BAG, Random Forest - RF) e Boosting - BO)

Para a importância de marcadores para a DT e seus refinamentos (BAG, RF e BO) foi utilizada a função "varImp" do pacote "caret" do software R, adotando como critério a RSS. A relação entre cada preditor e o resultado é avaliada, um modelo linear é ajustado e o valor absoluto do valor t para a inclinação de cada preditor é usado. A estatística R^2 é calculada para este modelo em relação ao modelo apenas com o intercepto. Esse número é retornado como uma medida relativa de importância.

A estrutura da árvore de regressão foi criada a partir da busca pela árvore que levaria à partição dos dados para a formação de grupos homogêneos. A divisão binária recursiva, inicialmente, seleciona o preditor X_j e o ponto de corte s dividindo o espaço preditor nas regiões $\{x|x_j < s\}$ e $\{x|x_j \geq s\}$ levando à maior redução possível no RSS. Essa equação pode ser representada por (JAMES et al., 2021):

$$R_1(j,s) = \{X|X_j < s\} \; e \; R_2(j,s) = \{X|X_j \geq s\},$$

e então procura-se o valor de J e S que minimizam a equação:

$$\sum_{i:x_i \in R_1(j,s)} \left(y_i - \hat{y}_{R_1}\right)^2 + \sum_{i:x_i \in R_2(j,s)} \left(y_i - \hat{y}_{R_2}\right)^2$$

em que: $\hat{y}_{R_1}$ é a média da variável resposta das observações de treinamento pertencente a região $R_1(j,s)$, $\hat{y}_{R_2}$ é a média da variável resposta das observações de treinamento pertencente a região $R_2(j,s)$ e y_i é o valor verdadeiro da característica de cada indivíduo.

O método Bagging (BAG) e Random Forest (RF) são uma extensão da árvore de regressão, que cria vários conjuntos de dados semelhantes por reamostragem (*bootstrapping*), para obter uma média das várias árvores de regressão que são realizadas sem poda para cada conjunto de dados (BREIMAN, 1996; PRASAD; IVERSON; LIAW, 2006). Assim são obtidos um número B de modelos: $\hat{f}^1(x), \hat{f}^2(x), \ldots, \hat{f}^B(x)$. A diferença entre esses métodos é que na RF cada árvore é estabelecida com um subconjunto aleatório de preditores (BOEHMKE; GREENWELL, 2019). O número de preditores usados para localizar a melhor divisão em cada nó é um subconjunto que foi escolhido por $m = \frac{v}{3}$, sendo v o número total de preditores (4010 marcadores). O número de árvores para BAG e RF foi fixado em 500.

O Boosting (BO) cria árvores sequencialmente utilizando informações das árvores anteriores que treina repetidamente na mesma amostra para que a cada iteração, uma medida de erro de previsão seja calculada para cada marca, e na próxima iteração, as marcas com maiores erros recebam maior peso no treinamento do modelo (JAMES et al., 2021). O número de árvores amostradas foi de 500, com uma taxa de aprendizado de 0.01 e profundidade igual a 2.

Treinamento e validação

A técnica de reamostragem de dados, k-fold (BENGIO; GRANDVALET, 2004), foi utilizada para particionar os dados utilizados no ajuste dos modelos, considerando $k = 5$. Os 1000 indivíduos da população F2 foram particionados em dados de treinamento e teste, compostos por 800 e 200 indivíduos, respectivamente. Assim, todos os indivíduos foram representados no teste. Para cada partição foram obtidos os valores de importância de marcadores por meio de cada uma das técnicas. Ao final,

foi utilizada a estimativa média obtida para cada marcador em todas as rodadas.

Comparação de metodologias

Para comparar as técnicas quanto o número de acerto e erros dos marcadores importantes detectados, foi avaliado o índice relativo (IR) baseado na equação:

$$IR = \frac{Acerto}{(Acerto - Erro)} * 100$$

Em que IR é o índice relativo; $Acerto$ é o número de identificação de regiões corretas dos QTLs; e $Erro$ é o número de marcadores identificados como importantes pelo método sem estar na região do QTL e o número de não identificação de marcadores na região do bloco gênico.

Aspectos computacionais

As simulações das populações foram realizadas usando o software GENES (CRUZ, 2013). Os métodos G-BLUP, DT, BAG, RF, BO e MARS foram realizados com o software GENES integrado ao software R (CRUZ, 2016; R CORE TEAM; COMPUTING; TEAM, 2020). Os métodos baseados em redes neurais MLP e RBF foram realizados pelo software GENES integrado ao software MATLAB (CRUZ, 2016; MATLAB, 2019).

Resultados

Ambos os métodos baseados em rede neurais artificiais (MLP e RBF) e o G-BLUP capturaram todas as marcas correspondentes aos QTLs, independente da característica avaliada (Tabela 2). Esses métodos estimaram importâncias relativas acima do limite estabelecido (12.5) para grande número de marcas, inclusive nos grupos de ligação 9 e 10, onde não havia QTL (Figuras suplementares 1 a 10 – ANEXO B). Apesar da identificação de alto percentual de acertos, esses métodos também proporcionam alto número de erros, ou seja, falsas descobertas (erro tipo I) e, por isso, não são úteis para identificação de marcas associadas a QTLs. Assim, os métodos MLP, RBF e GBLUP não foram considerados nas análises subsequentes.

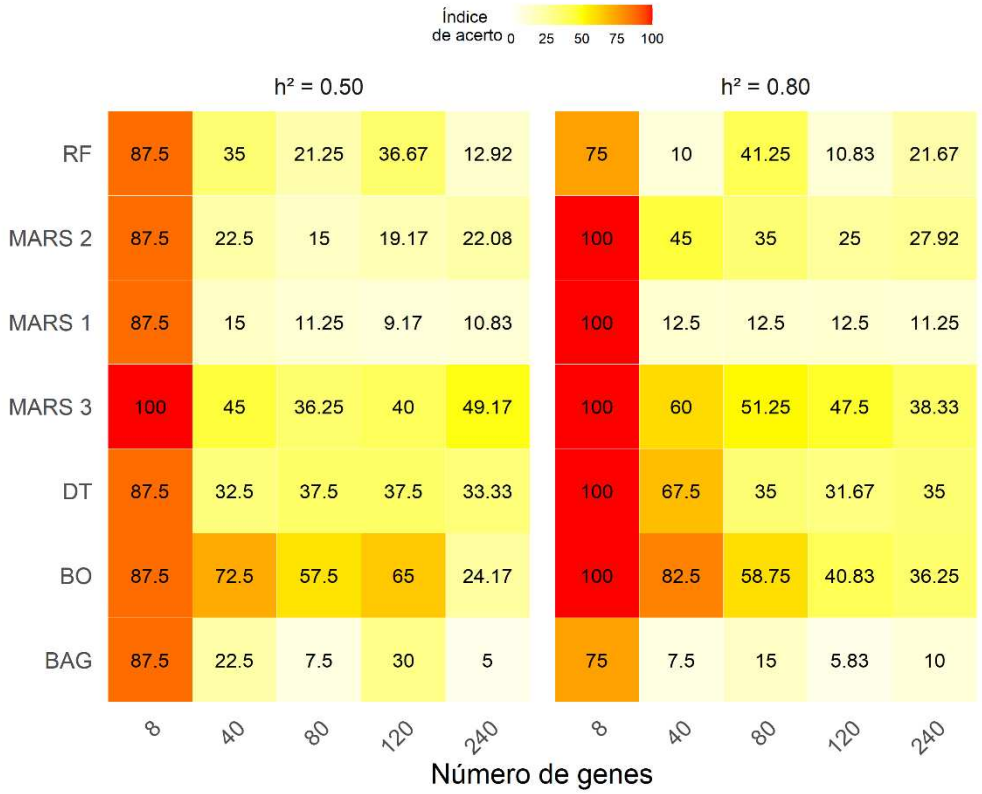
Tabela 2: Número de acerto e erro na identificação de marcadores pelos métodos Bagging (BAG), Boosting (BO), Árvore de Decisão (DT); Random Forest (RF); Perceptron Multicamadas (MLP), Função de Base Radial (RBF); MARS 1, MARS 2, MARS 3 e G-BLUP. *1 a 5 representam variáveis com herdabilidade de 0.5 e número de QTLs de 8, 40, 80, 120 e 240, respectivamente. 6 a 10 representam variáveis com herdabilidade de 0.8 e número de QTLs de 8, 40, 80, 120 e 240 respectivamente.

Variável*	Método	Acerto	Erro	Variável*	Método	Acerto	Erro
1	BAG	7	12	6	BAG	6	1
1	BO	7	8	6	BO	8	1
1	DT	7	91	6	DT	8	14
1	G-BLUP	8	3878	6	G-BLUP	8	3911
1	MARS 3	8	46	6	MARS 3	8	5
1	MARS 1	7	5	6	MARS 1	8	0
1	MARS 2	7	7	6	MARS 2	8	1
1	MLP	8	3519	6	MLP	8	3762
1	RBF	8	3887	6	RBF	8	3612
1	RF	7	20	6	RF	6	9
2	BAG	9	78	7	BAG	3	26
2	BO	29	369	7	BO	33	165
2	DT	13	337	7	DT	27	380
2	G-BLUP	40	3610	7	G-BLUP	40	3625
2	MARS 3	18	243	7	MARS 3	24	264
2	MARS 1	6	31	7	MARS 1	5	37
2	MARS 2	9	74	7	MARS 2	18	86
2	MLP	40	3595	7	MLP	40	3276
2	RBF	40	3643	7	RBF	40	3642
2	RF	14	288	7	RF	4	55
3	BAG	6	29	8	BAG	12	65
3	BO	46	244	8	BO	47	207
3	DT	30	293	8	DT	28	312
3	G-BLUP	80	3262	8	G-BLUP	80	3275
3	MARS 3	29	225	8	MARS 3	41	210
3	MARS 1	9	32	8	MARS 1	10	33
3	MARS 2	12	79	8	MARS 2	28	77
3	MLP	80	3195	8	MLP	80	3107
3	RBF	80	3259	8	RBF	80	3284
3	RF	17	137	8	RF	33	221
4	BAG	36	119	9	BAG	7	11
4	BO	78	313	9	BO	49	89
4	DT	45	229	9	DT	38	238
4	G-BLUP	120	2917	9	G-BLUP	120	2881
4	MARS 3	48	149	9	MARS 3	57	228
4	MARS 1	11	23	9	MARS 1	15	27
4	MARS 2	23	56	9	MARS 2	30	88
4	MLP	120	2805	9	MLP	120	2460
4	RBF	120	2924	9	RBF	120	2925
4	RF	44	249	9	RF	13	41
5	BAG	12	6	10	BAG	24	15
5	BO	58	34	10	BO	87	71
5	DT	80	134	10	DT	84	133
5	G-BLUP	240	1833	10	G-BLUP	240	1825
5	MARS 3	118	107	10	MARS 3	92	74
5	MARS 1	26	8	10	MARS 1	27	14
5	MARS 2	53	24	10	MARS 2	67	36
5	MLP	240	1810	10	MLP	240	1749
5	RBF	240	1849	10	RBF	240	1846
5	RF	31	36	10	RF	52	72

Efeito da herdabilidade sobre a eficiência de técnicas na identificação de marcadores importantes

O aumento da herdabilidade gerou aumento na identificação de marcadores associados aos QTLs em todos cenários ao utilizar a MARS 2 (Tabela 2). Os métodos RF e BAG apresentaram menor índice de acerto na maioria dos cenários quando houve aumento na h^2 (Figura 2).

Figura 2 - Índice de acerto na identificação de marcadores pelos métodos Bagging (BAG), Boosting (BO), Árvore de Decisão (DT); Random Forest (RF); MARS 1, MARS 2 e MARS 3 em função da característica avaliada. A intensidade da cor indica o maior valor do índice de acerto (vermelho) para o menor (branco).



Fonte: O autor.

Para as características oligogênicas (controladas por 8 QTLs), o aumento na h^2 provocou aumento na identificação de QTLs para os métodos MARS 2, MARS 1, DT e BO de 87,5 para 100 no índice de acerto. O MARS 3 identificou todas os blocos gênicos associados aos QTLS em ambos os valores de h^2. Ao aumentar o número de QTLs foi observado que a identificação de marcadores associados aos QTLs variou

de acordo com o método, não obtendo um padrão linear entre a identificação dos marcadores e o aumento da h^2 (Figura 2).

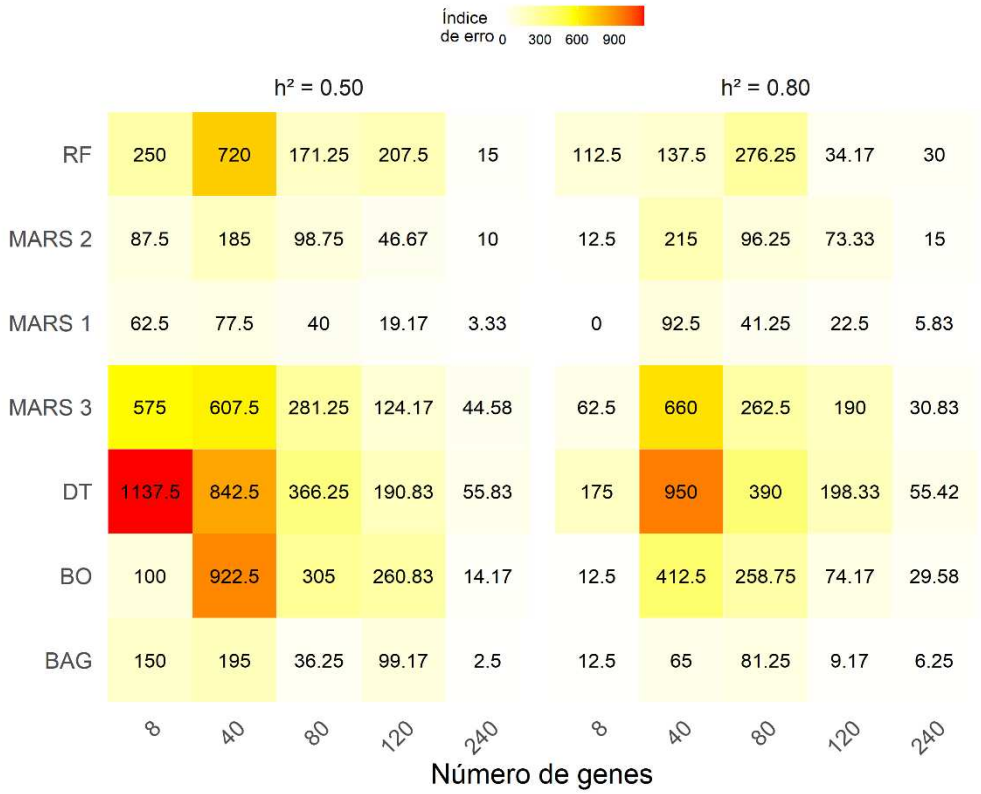
Os métodos MARS 3, DT e BO foram os que apresentaram maior número de acerto ao aumentar a h^2, quando comparado as características com 40 QTLs (Tabela 2). O único método que apresentou mais de 70 de índice de acerto foi o BO em ambos os cenários (Figura 2). Já para as características com 80 QTLs, apenas DT apresentou diminuição no índice de acerto quando houve aumento na h^2, mas em pequena magnitude, de 37,5 (h^2 = 0.5) para 35% (h^2 = 0.8). Do mesmo modo ao cenário de 40 QTLs, os métodos MARS 3 e BO foram os que apresentaram maiores índices de acerto em cenários com 80 QTLs (Figura 2).

O efeito do aumento da h^2 nas características com 120 QTLs foi favorável para todos os métodos baseados em MARS, contudo os índices de acertos para MARS 1 e MARS 2 continuaram baixos quando houve o aumento da h^2 (Figura 2). Por outro lado, a MARS 3 aumentou o índice de acerto de 40% (h^2 = 0.5) para 47.5% (h^2 = 0.8).

Apesar de ter apresentado menor índice de acerto quando comparado ao aumento da h^2 para os cenários com 120 QTLs, o BO apresentou o maior índice de acerto (65) e segundo maior índice de acerto (40,83) para as h^2 de 0.5 e 0.8, respectivamente. O mesmo ocorreu para MARS 3 no cenário com 240 QTLs, que também apresentou redução do índice de erro quando houve aumento na h^2, no entanto foi o método que apresentou os maiores índices de acertos em ambas as h^2.

Os métodos apresentaram maior índice de erros para as características com menor número de genes (Figura 3). Comparando as características com 8 QTLs, todos os métodos apresentaram menor índice de erro ao aumentar a h^2 de 0,5 para 0,8. Ao aumentar o valor da h^2 foi verificado houve aumento no índice de erros para todas as características quando comparadas com o mesmo número de QTLs apenas para a MARS 1 (Figura 3).

Figura 3 - Índice de erro na identificação de marcadores pelos métodos Bagging (BAG), Boosting (BO), Árvore de Decisão (DT); Random Forest (RF), MARS 1, MARS 2 e MARS 3 em função da característica avaliada. A intensidade da cor indica o maior valor do índice de erro (vermelho) para o menor (branco).



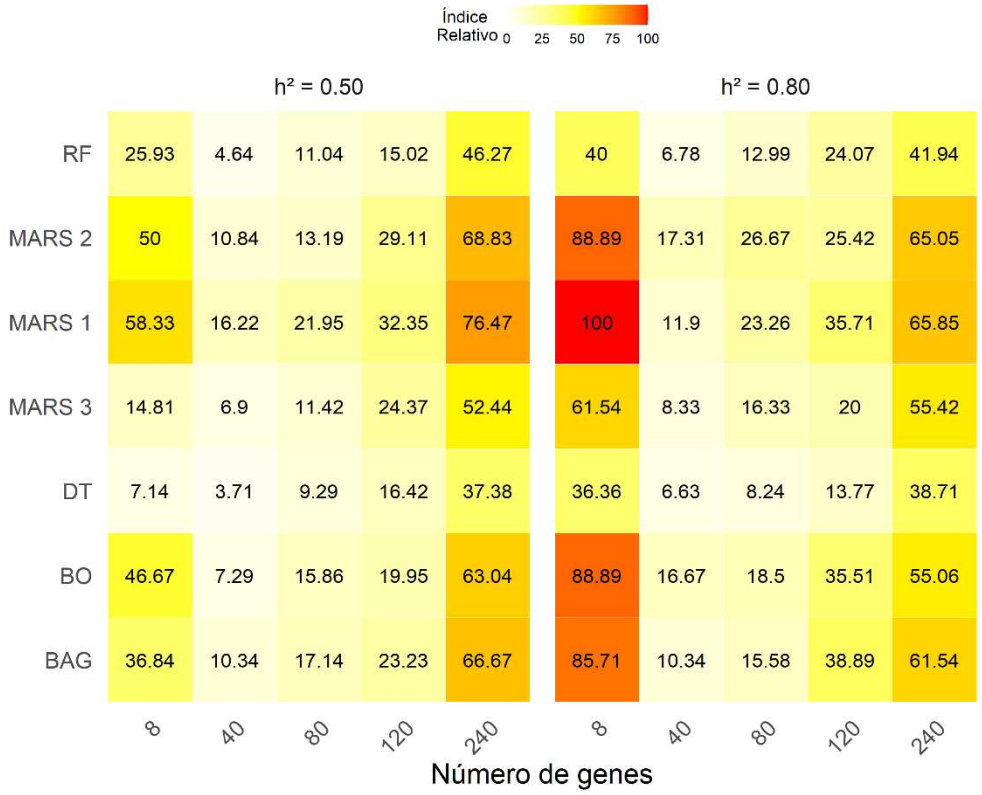
Fonte: O autor.

Os métodos RF e MARS 3 apresentaram comportamentos contrários para o índice de erros, enquanto MARS 3 apresentou aumento no índice de erros ao elevar a h^2 nos cenários de 40 e 120 QTLs, o RF apresentou redução (Figura 3). Já, quando avaliado os cenários de 80 e 240 QTLs, o aumento da h^2 proporcionou aumento no índice de erros para RF, mas, em contrapartida, uma redução para MARS 3 (Figura 3). Do mesmo modo, também foi observado um comportamento contrário para os métodos DT e BAG, enquanto BAG obteve aumento no índice de erros apenas para os cenários com 240 QTLs, DT apresentou redução apenas nesse cenário ao aumentar o valor da h^2 (Figura 3).

Na avaliação do índice relativo, obtido pelo quociente do número de acertos e de erros, foi possível verificar que o aumento da h^2 tem efeito positivo no índice na maioria dos cenários, ou seja, houve melhora no índice relativo ao aumentar a h^2 para a maioria dos métodos (Figura 4). Apenas os métodos MARS 1, BAG e DT nos

cenários de 40, 80 e 120 QTLs, respectivamente, não apresentaram esse padrão (Figura 4).

Figura 4 - Índice relativo na identificação de marcadores pelos métodos Bagging (BAG), Boosting (BO), Árvore de Decisão (DT); Random Forest (RF); MARS 1, MARS 2 e MARS 3 em função da característica avaliada. A intensidade da cor indica o maior valor do índice relativo (vermelho) para o menor (branco).



Fonte: O autor.

Efeito do número de QTLs sobre a eficiência de técnicas na identificação de marcadores importantes

O aumento do número de QTLs sobre a eficiência de técnicas na identificação dos marcadores também não gerou uma resposta linear. A maioria dos métodos apresentou redução no índice de acerto quando o número de QTLs aumentou de 8 para 40 em ambos os cenários de h^2. O BO foi o método que apresentou menor proporção de redução nesses cenários, de 87,5 para 72,5 quando a h^2 era de 0,5 e de 100 para 82,5 para h^2 igual a 0,8 (Figura 2).

Ao comparar os cenários de 40 com o de 80 QTLs, a maioria dos métodos

apresentaram aumento na identificação de marcadores associados aos QTLs (Tabela 2). No entanto, ainda nesses cenários, para a h^2 de 0,5, apenas a DT apresentou maior índice de acerto ao aumentar o número de QTLs, mas em baixa proporção (Figura 2). Em contrapartida, para h^2 = 0.8, o RF apresentou melhora substancial no índice de acerto ao aumentar o número de QTLs, passando de 10, em 40 QTLs, para 41,25 no cenário com 80 QTLs.

Na h^2 de 0.5, considerando o aumento de 80 para 120 QTLs, a maioria dos métodos apresentam aumento no índice de acertos, exceto para a MARS 1 que diminui e DT que manteve o mesmo valor (Figura 2). Já para os cenários de h^2 = 0.5, apenas a MARS 1 manteve o mesmo valor nos cenários de 80 e 120 QTLs, os demais métodos apresentaram redução no índice de acerto (Figura 2).

Ao considerar o aumento do número de QTLs de 120 para 240 foi observado aumento no índice de acerto para o método MARS 2 tanto para ambos os valores de h^2. A MARS 3 teve melhora apenas no cenário de h^2 = 0.5, sendo o método com a maior índice de acerto nesse cenário. Apesar de diminuir o índice de acerto quando comparado o aumento de 120 para 240 QTLs no cenário de h^2 = 0.8, a MARS 3 também foi o método que exibiu o maior índice de acerto de marcadores importantes.

A maioria dos métodos demonstrou aumento no índice de erros quando houve aumento do número de QTLs de 8 para 40 em ambos os cenários de h^2, com exceção de DT para h^2 = 0,5 (Figura 3). Em contrapartida, ao aumentar o número de QTLs de 40 até 240 QTLs, verificou-se que a maioria dos métodos apresentaram redução no valor do índice de erros (Figura 3). As exceções foram os métodos RF e BAG nos cenários com h^2 = 0,5 e do aumento de 80 para 120 QTLs, cujos índices de erro eram de 171,25 e 36,25 e passaram para 207,5 e 99,17, respectivamente. Além desses cenários, esses métodos também apresentaram aumento no índice de erros nos cenários com h^2 = 0,8, considerando o aumento de 40 para 80 QTLs, onde os índices de erro foram de 137,5 e 65 e aumentaram para 276,25 e 81,25, respectivamente (Figura 3).

O desempenho dos métodos mostrou que quando houve aumento do número de 8 para 40 QTLs, observou uma tendência na redução do índice relativo em ambos os cenários de h^2 (Figura 4). De modo contrário, ao elevar o número de 40 até 240 QTLs a tendência foi de aumentar o valor do índice relativo para a maioria dos

métodos. Apenas a MARS 2 para o cenário de h^2 = 0,8, considerando o aumento do número de QTLs de 80 para 120, apresentou uma redução no índice, mas em pequena magnitude (26,67 para 25,42) (Figura 4).

Considerações gerais sobre a abordagem biométrica de melhor desempenho na quantificação da importância de marcadores

Características de maior h^2, C6 a C10, proporcionaram maior índice de acertos (Figura 2). O método DT apresentou índice de acertos alto para características controladas por oito QTLs (87.5 e 100 em C1 e C6, respectivamente) (Figura 3). A eficiência foi reduzida à medida em que o número de QTLs aumentou. Entretanto, o número alto de marcas identificadas erroneamente como importantes foi alto para todas as características (Tabela 2). Os índices de erro variaram de 55.4 para a característica C10 até 1137.5 para C1 (Figura 3).

As abordagens de MARS 1, 2 e 3 apresentaram alto índice de acertos para características governadas por pouco genes (C1 e C8) (Figura 2). Características governadas por 120 ou 240 QTLs (C5, C6, C9 e C10). apresentaram índice de acertos pouco superior na MARS 3 (Figura 2). Em relação ao índice de erros, a MARS 3 teve o pior desempenho se comparada a seus pares, com índice de erros variando entre 30 a 660 (Figura 3). Para a MARS 1 o resultado foi bastante promissor para a característica C6 (governada por 8 genes e h^2 igual a 0.8), pois esta não apresentou erros e todas as marcas de importância foram identificadas (Figura Suplementar 6). Baixos índices de erros também foram identificadas para C5 e C10, com valores iguais a 3.3 e 5.8, respectivamente (Figura 3). Para a MARS 2 também foram observados baixo índice de erros para C5 e C10 (Figura 3).

A eficiência relativa das técnicas, dada pela relação entre número de acertos pelo número total de erros e acertos, destacou os métodos MARS 1, MARS 2, BO e BAG para identificação de marcadores importantes (Figura 4). Essa eficiência se deu principalmente em cenários com 8 QTLs. À medida em que se aumentou o número de QTLs até 120 (C2, C3, C4, C7, C8, C9) a eficiência ficou prejudicada (Figura 4). Entretanto, para características controladas por 240 QTLs, as técnicas conseguiram identificar as marcas importantes de forma superior ao que foi observado para

características controladas por 40, 80 ou 120 QTLs.

Discussão

O objetivo principal de selecionar subconjuntos de marcadores no genoma de qualquer indivíduo ou população é reduzir o custo de genotipagem, mantendo a mesma capacidade preditiva obtida quando se utiliza o conjunto completo de marcadores (LONG et al., 2011). Diversos estudos têm apontado que a seleção de marcadores proporciona confiabilidade razoavelmente alta na predição de valores genéticos (MACCIOTTA et al., 2009; PIMENTEL et al., 2009; SCHULZ-STREECK; OGUTU; PIEPHO, 2015). No entanto, multicolinearidade entre os marcadores e alta dimensionalidade dos dados têm sido um desafio enfrentado pelos pesquisadores (AZEVEDO et al., 2014; CROSSA et al., 2017). Métodos baseados em aprendizado de máquinas e redes neurais apresentam alto potencial para lidar com esses problemas em estudo de dados genômicos (BARBOSA et al., 2021; OKSER et al., 2014; SILVA et al., 2014; SOUSA et al., 2021).

Marcadores SNP comumente apresentam moderada a alta correlação devido, principalmente, ao desequilíbrio de ligação (JACQUIN; CAO; AHMADI, 2016). De modo a contornar esse problema, métodos baseados em redes neurais (MLP e RBF) vem sendo utilizados, apresentando alta eficiência para predição em dados genômicos (BARBOSA et al., 2021; HOWARD; CARRIQUIRY; BEAVIS, 2014). Isso porque consideram que todas as marcas têm importância para predição dos valores genéticos. Para o G-BLUP, considerar todo o conjunto de marcadores como importante é premissa básica. O método considera que as predições genômicas baseiam-se no parentesco derivado de todos os marcadores (WANG et al., 2018). Já para as redes neurais, as técnicas de importância de variáveis são usadas para determinar a influência de cada marcador e sua contribuição para a predição do valor genético (ZHANG et al., 2018). Por outro lado, quando se pretende relacionar marcadores com QTLs importantes para uma característica a eficiência não é a mesma.

Esses métodos não são, portanto, métodos de poda, com o objetivo de minimizar o número de variáveis explicativas, mas procedimentos para estimar a contribuição relativa de cada variável de entrada (GEVREY; DIMOPOULOS; LEK,

2003). No contexto do aprendizado de máquina, um método conhecido como seleção de recursos é comumente implementado para identificar o subconjunto de variantes com maior poder preditivo para o traço fenotípico específico (OKSER et al., 2014). As abordagens baseadas em particionamento recursivo, como as DT e seus refinamentos (BAG, RF e BO) e as MARS, são métodos de aprendizado de máquinas que possuem essa funcionalidade. Por isso, apresentam maior poder de detecção de marcadores com efeito verdadeiro na característica avaliada. No entanto, para RF, WALTERS; LAURIN; LUBKE (2012) demonstraram que as medidas de importância variável são sistematicamente afetadas pelo desequilíbrio de ligação (LD). O presente estudo provocou desequilíbrio de ligação pela proximidade entre marcadores, por isso, o RF foi afetado, não apresentando resultados satisfatórios.

A MARS, se diferencia dos demais métodos de aprendizado de máquinas por permitir a incorporação de interação entre os marcadores. Como os dados simulados nesse estudo foram oriundos de modelo epistático, cujo efeito interativo entre marcadores é importante, era esperado que a MARS 2 e/ou 3 apresentassem melhores desempenhos, devido à capacidade de captar interações entre um maior número de marcadores. Entretanto, a eficiência da MARS 3 somente foi observada quando considerado o número absoluto de acertos isoladamente (Figura 2). A técnica proporcionou alto número de erros de identificação, pois identificou como importantes um número elevado de marcadores fora do intervalo dos QTLs (Figura 3). Já a MARS 2 e MARS 1, foram os métodos que apresentaram maiores valores para o índice relativo (Figura 4), indicando consistência para identificação de marcadores, conforme hipotetizado nesse estudo. Os métodos BO e BAG também merecem ser destacados, uma vez que apresentaram uma eficiência relativa alta (Figura 4).

Além de detectar possíveis interações entre marcadores, a MARS possui a fase *backward* que é exclusiva para seleção de variáveis cujo efeito é medido pela soma de quadrados de resíduos (FRIEDMAN, 1991). Além disso, possui controle do número máximo de marcadores que pode ser adicionado ao modelo. Dessa forma, pode-se tentar diferentes combinações antes de definir modelo que produza o subconjunto de marcadores cujo tamanho esteja próximo do valor desejado ou abaixo de algum valor pré-definido. A MARS e os métodos BO e BAG podem ser usados para seleção de marcadores visando diminuir o tempo computacional do processamento dos dados. Isso devido à menor densidade de marcadores que evita níveis altos de

multicolinearidade no conjunto de dados. Como relatado por Okser *et al.* (2014), não há regra de ouro para a seleção de variáveis, o modelo deve ser selecionado com base nas características dos dados e objetivos da aplicação genética.

Apesar de nenhum método ter apresentado uma eficiência relativa alta para todas características, os benefícios em utilizar os marcadores selecionados para seleção de indivíduos baseados no genoma são altos, visto que esse processo pode acelerar a fase de seleção de indivíduos em programas de melhoramento. Os métodos baseados em aprendizado de máquina demonstraram fornecer meios aprimorados de aprender painéis multilocus de variantes genéticas e suas interações que são mais preditivas de características fenotípicas complexas (OKSER et al., 2014). Além disso, a seleção de SNPs para compor subconjunto representativo do genoma reduz o impacto do LD e também facilitam correção potencial para o efeito do MAF (WALTERS; LAURIN; LUBKE, 2012). A seleção de marcadores com posterior aplicação de métodos com alto poder preditivo, como as redes neurais (BARBOSA et al., 2021; SOUSA et al., 2021), podem ser aliadas para acelerar o processo de seleção de indivíduos superiores.

Conclusões

A distribuição dos QTL nos grupos de ligação pode ser o principal atributo a ser avaliado na predição dos valores genéticos e identificação de marcas associadas à QTLs, quando o experimento é bem conduzido a fim de se obter um maior valor para a herdabilidade.

Os métodos MARS 1, MARS 2, Boosting e Bagging foram os mais efetivos na detecção de marcadores importantes ao longo do genoma, principalmente para as características com 8 e 240 QTLs, considerando o índice relativo. A MARS 3 e o Boosting apresentaram alta capacidade de identificar os marcadores de importância.

Todos esses métodos de aprendizado de máquinas podem ser aplicados a dados de sequenciamento de larga escala para detecção de SNPs associados às regiões onde possivelmente exista QTLs de características com efeitos aditivos e não-aditivos (dominância e epistasia), auxiliando pesquisadores a melhor compreender o genoma de diferentes espécies. Além disso, a seleção de marcadores pode viabilizar o tempo computacional gasto por modelos mais complexos visando a predição de

valores genéticos genômicos.

Agradecimentos

Os autores agradecem o suporte financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES code de financiamento 001.

Referências

AZEVEDO, C. F. et al. Supervised independent component analysis as an alternative method for genomic selection in pigs. **Journal of Animal Breeding and Genetics**, *[S. l.]*, v. 131, n. 6, p. 452–461, 2014. DOI: 10.1111/jbg.12104.

BARBOSA, Ivan Paiva; SILVA, Michele Jorge; COSTA, Weverton Gomes; SANT'ANNA, Isabela Castro; NASCIMENTO, Moysés; CRUZ, Cosme Damião. Genome-enabled prediction through machine learning methods considering different levels of trait complexity. **Crop Science**, *[S. l.]*, v. 61, n. 3, p. 1890–1902, 2021. DOI: 10.1002/csc2.20488.

BEDO, Justin; WENZL, Peter; KOWALCZYK, Adam; KILIAN, Andrzej. by machine learning. *[S. l.]*, v. 18, p. 1–18, 2008. DOI: 10.1186/1471-2156-9-35.

BENGIO, Yoshua; GRANDVALET, Yves. No Unbiased Estimator of the Variance of K-Fold Cross-Validation. **Journal of Machine Learning Research**, *[S. l.]*, v. 5, p. 1089–1105, 2004. DOI: 10.1016/S0006-291X(03)00224-9.

BERMINGHAM, M. L. et al. Application of high-dimensional feature selection: Evaluation for genomic prediction in man. **Scientific Reports**, *[S. l.]*, v. 5, p. 1–12, 2015. DOI: 10.1038/srep10312.

BOEHMKE, Brad; GREENWELL, Brandon. Random Forests. *In*: **Hands-On Machine Learning with R**. [s.l.] : Chapman and Hall/CRC, 2019. v. 45p. 203–219. DOI: 10.1201/9780367816377-11. Disponível em: https://www.taylorfrancis.com/books/9781000730197/chapters/10.1201/9780367816377-11.

BREIMAN, Leo. Bagging Predictors. **Machine Learning**, *[S. l.]*, v. 24, n. 421, p. 123–140, 1996. DOI: 10.1007/BF00058655.

COSTA, Weverton Gomes; BARBOSA, Ivan de Paiva; SOUZA, Jacqueline Enequio; CRUZ, Cosme Damião; NASCIMENTO, Moysés; OLIVEIRA, Antonio Carlos Baião. Machine learning and statistics to qualify environments through multi-traits in Coffea arabica. **PLOS ONE**, *[S. l.]*, v. 16, n. 1, p. 1–21, 2021. DOI:

10.1371/journal.pone.0245298. Disponível em: https://dx.plos.org/10.1371/journal.pone.0245298.

CROSSA, José et al. Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. **Trends in Plant Science**, *[S. l.]*, v. 22, n. 11, p. 961–975, 2017. DOI: 10.1016/j.tplants.2017.08.011. Disponível em: https://doi.org/10.1016/j.tplants.2017.08.011.

CRUZ, Cosme Damião. GENES - Software para análise de dados em estatística experimental e em genética quantitativa. **Acta Scientiarum - Agronomy**, *[S. l.]*, v. 35, n. 3, p. 271–276, 2013. DOI: 10.4025/actasciagron.v35i3.21251.

CRUZ, Cosme Damião. Genes software – extended and integrated with the R, Matlab and Selegen. **Acta Scientiarum - Agronomy**, *[S. l.]*, v. 38, n. 4, p. 547–552, 2016. DOI: 10.4025/actasciagron.v38i4.32629.

CRUZ, Cosme Damião; NASCIMENTO, Moysés. **Inteligência Computacional aplicada ao melhoramento genético**. Viçosa, MG: Editora UFV, 2018.

DE LOS CAMPOS, Gustavo; HICKEY, John M.; PONG-WONG, Ricardo; DAETWYLER, Hans D.; CALUS, Mario P. L. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. **Genetics**, *[S. l.]*, v. 193, n. 2, p. 327–345, 2013. DOI: 10.1534/genetics.112.143313. Disponível em: https://academic.oup.com/genetics/article/193/2/327/6065337.

FALCONER, S. D.; MACKAY, T. F. C. **Introduction to quantitative genetics**. [s.l.] : Edinburgh, Addison Wesley Longman, 1996.

FRIEDMAN, Jerome H. Multivariate Adaptative regression Splines. **The Annals of Statistics**, *[S. l.]*, v. 19, n. 1, p. 1–141, 1991. Disponível em: http://projecteuclid.org/euclid.aop/1176996548.

GEVREY, Muriel; DIMOPOULOS, Ioannis; LEK, Sovan. Review and comparison of methods to study the contribution of variables in artificial neural network models. **Ecological Modelling**, *[S. l.]*, v. 160, n. 3, p. 249–264, 2003. DOI: 10.1016/S0304-3800(02)00257-0.

GHAFOURI-KESBI, Farhad; RAHIMI-MIANJI, Ghodratollah; HONARVAR, Mahmood; NEJATI-JAVAREMI, Ardeshir. Predictive ability of Random Forests, Boosting, Support Vector Machines and Genomic Best Linear Unbiased Prediction in different scenarios of genomic evaluation. **Animal Production Science**, *[S. l.]*, v. 57, n. 2, p. 229, 2017. DOI: 10.1071/AN15538. Disponível em: http://www.publish.csiro.au/?paper=AN15538.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The elements of statistical learning: Data mining, inference, and prediction**. 2. ed. ed. New York, NY, USA: Springer, 2009. DOI: 10.1007/978-1-4419-9863-7_941.

HOWARD, Réka; CARRIQUIRY, Alicia L.; BEAVIS, William D. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. **G3: Genes, Genomes, Genetics**, *[S. l.]*, v. 4, n. 6, p. 1027–1046, 2014. DOI: 10.1534/g3.114.010298.

HULSEGGE, B.; CALUS, M. P. L.; WINDIG, J. J.; HOVING-BOLINK, A. H.; MAURICE-VAN EIJNDHOVEN, M. H. T.; HIEMSTRA, S. J. Selection of SNP from 50K and 777K arrays to predict breed of origin in cattle. **Journal of Animal Science**, *[S. l.]*, v. 91, n. 11, p. 5128–5134, 2013. DOI: 10.2527/jas.2013-6678.

JACQUIN, Laval; CAO, Tuong Vi; AHMADI, Nourollah. A unified and comprehensible view of parametric and kernel methods for genomic prediction with application to rice. **Frontiers in Genetics**, *[S. l.]*, v. 7, n. AUG, p. 1–16, 2016. DOI: 10.3389/fgene.2016.00145.

JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert. An Introduction to Statistical Learning. *In*: **Springer Texts in Statistics**. [s.l: s.n.]. p. 612. DOI: 10.1007/978-1-0716-1418-1_1. Disponível em: https://link.springer.com/10.1007/978-1-0716-1418-1_1.

JOLLIFE, Ian T.; CADIMA, Jorge. **Principal component analysis: A review and recent developmentsPhilosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, 2016. DOI: 10.1098/rsta.2015.0202.

LI, Bo; ZHANG, Nanxi; WANG, You Gan; GEORGE, Andrew W.; REVERTER, Antonio; LI, Yutao. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. **Frontiers in Genetics**, *[S. l.]*, v. 9, n. JUL, p. 1–20, 2018. DOI: 10.3389/fgene.2018.00237.

LIANG, Yulan; KELEMEN, Arpad. Statistical advances and challenges for analyzing correlated high dimensional SNP data in genomic study for complex diseases. **Statistics Surveys**, *[S. l.]*, v. 2, p. 43–60, 2008. DOI: 10.1214/07-SS026.

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**, *[S. l.]*, v. 128, n. 4, p. 247–257, 2011. DOI: 10.1111/j.1439-0388.2011.00917.x.

MACCIOTTA, Nicolò PP; GASPA, Giustino; STERI, Roberto; PIERAMATI, Camillo; CARNIER, Paolo; DIMAURO, Corrado. Pre-selection of most significant SNPS for the estimation of genomic breeding values. **BMC Proceedings**, *[S. l.]*, v. 3, n. S1, p. 1–4, 2009. DOI: 10.1186/1753-6561-3-s1-s14.

MATLAB. **Natick, Massachusetts: The MathWorks Inc.**Natick, MassachusettsThe MathWorks Inc., , 2019.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, *[S. l.]*, v. 157, n. 4, p. 1819–1829, 2001.

MILBORROW, Stephen. Notes on the earth package. *[S. l.]*, p. 1–68, 2019. Disponível em: http://www.milbo.org/doc/earth-notes.pdf%0Ahttps://cran.r-project.org/web/packages/earth/index.html.

OKSER, Sebastian; PAHIKKALA, Tapio; AIROLA, Antti; SALAKOSKI, Tapio; RIPATTI, Samuli; AITTOKALLIO, Tero. Regularized Machine Learning in the Genetic Prediction of Complex Traits. **PLoS Genetics**, *[S. l.]*, v. 10, n. 11, 2014. DOI: 10.1371/journal.pgen.1004754.

PIMENTEL, Eduardo CG; KÖNIG, Sven; SCHENKEL, Flavio S.; SIMIANER, Henner. Comparison of statistical procedures for estimating polygenic effects using dense genome-wide marker data. **BMC Proceedings**, *[S. l.]*, v. 3, n. S1, p. 1–5, 2009. DOI: 10.1186/1753-6561-3-s1-s12.

PRASAD, Anantha M.; IVERSON, Louis R.; LIAW, Andy. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. **Ecosystems**, *[S. l.]*, v. 9, n. 2, p. 181–199, 2006. DOI: 10.1007/s10021-005-0054-1.
R CORE TEAM; COMPUTING, R. Foundation for Statistical; TEAM, R. Core. **R: A Language and Environment for Statistical Computing**. 2020. Disponível em: https://www.r-project.org/. Acesso em: 1 jul. 2020.

RESENDE, Marcos Deon Vilela; SILVA, Fabyano Fonseca e; VIANA, José Marcelo Soriano; PETERNELLI, Luiz Alexandre; RESENDE JÚNIOR, Márcio Fernando Ribeiro; VALLE, Patrício Muños Del. **Métodos estatísticos na seleção genômica ampla**. Colombo - PR: Embrapa Floresta, 2011.

SCHIAVO, G.; BERTOLINI, F.; GALIMBERTI, G.; BOVO, S.; OLIO, S. Dall; COSTA, L. Nanni; GALLO, M.; FONTANESI, L. A machine learning approach for the identification of population- informative markers from high-throughput genotyping data : application to several pig breeds. **Animal, The International Journal of Animal Biosciences**, *[S. l.]*, v. 14, n. 2, p. 223–232, 2020. DOI: 10.1017/S1751731119002167. Disponível em: http://dx.doi.org/10.1017/S1751731119002167.

SCHULZ-STREECK, Torben; OGUTU, Joseph O.; PIEPHO, Hans Peter. Pre-selection of markers for genomic selection. **BMC Proceedings**, *[S. l.]*, v. 5, n. Suppl 3, p. 3–6, 2015. DOI: 10.1186/1753-6561-5-S3-S12.

SILVA, Gabi Nunes; TOMAZ, Rafael Simões; SANT'ANNA, Isabela Castro; NASCIMENTO, Moysés; BHERING, Leonardo Lopes; CRUZ, Cosme Damião. Neural networks for predicting breeding values and genetic gains. **Scientia Agricola**, *[S. l.]*, v. 71, n. 6, p. 494–498, 2014. DOI: 10.1590/0103-9016-2014-0057.

SOUSA, Ithalo Coelho et al. Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. **Scientia Agricola**, *[S. l.]*, v. 78, n. 4, p. 1–

8, 2021. DOI: 10.1590/1678-992x-2020-0021. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-90162021000401102&tlng=en.

WALDMANN, Patrik. Genome-wide prediction using Bayesian additive regression trees. **Genetics Selection Evolution**, *[S. l.]*, v. 48, n. 1, p. 1–12, 2016. DOI: 10.1186/s12711-016-0219-8.

WALTERS, Raymond; LAURIN, Charles; LUBKE, Gitta H. An integrated approach to reduce the impact of minor allele frequency and linkage disequilibrium on variable importance measures for genome-wide data. **Bioinformatics**, *[S. l.]*, v. 28, n. 20, p. 2615–2623, 2012. DOI: 10.1093/bioinformatics/bts483.

WANG, Jiabo; ZHOU, Zhengkui; ZHANG, Zhe; LI, Hui; LIU, Di; ZHANG, Qin; BRADBURY, Peter J.; BUCKLER, Edward S.; ZHANG, Zhiwu. **Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traitsHeredity**, 2018. DOI: 10.1038/s41437-018-0075-0.

WILKINSON, Samantha; ARCHIBALD, Alan L.; HALEY, Chris S.; MEGENS, Hendrik Jan; CROOIJMANS, Richard P. M. A.; GROENEN, Martien A. M.; WIENER, Pamela; OGDEN, Rob. Development of a genetic tool for product regulation in the diverse British pig breed market. **BMC Genomics**, *[S. l.]*, v. 13, n. 1, 2012. DOI: 10.1186/1471-2164-13-580.

WILKINSON, Samantha; WIENER, Pamela; ARCHIBALD, Alan L.; LAW, Andy; SCHNABEL, Robert D.; MCKAY, Stephanie D.; TAYLOR, Jeremy F.; OGDEN, Rob. Evaluation of approaches for identifying population informative markers from high density SNP Chips. **BMC Genetics**, *[S. l.]*, v. 12, 2011. DOI: 10.1186/1471-2156-12-45.

ZHANG, Wengang; GOH, Anthony T. C. Multivariate adaptive regression splines and neural network models for prediction of pile drivability. **Geoscience Frontiers**, *[S. l.]*, v. 7, n. 1, p. 45–52, 2016. DOI: 10.1016/j.gsf.2014.10.003. Disponível em: http://dx.doi.org/10.1016/j.gsf.2014.10.003.

ZHANG, Zhongheng; BECK, Marcus W.; WINKLER, David A.; HUANG, Bin; SIBANDA, Wilbert; GOYAL, Hemant. Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. **Annals of Translational Medicine**, *[S. l.]*, v. 6, n. 11, p. 216–216, 2018. DOI: 10.21037/atm.2018.05.32.

6. CONCLUSÃO GERAL

A distribuição dos QTL nos grupos de ligação pode ser o principal atributo a ser avaliado na predição dos valores genéticos e identificação de marcas associadas à QTLs, visto que quanto maior a herdabilidade melhores são os resultados. Assim, deve-se sempre buscar conduzi bem o experimento a fim de se obter um maior valor para a herdabilidade.

Os métodos de ML e de ANN demonstraram alto potencial para predição de valores genéticos em caracteres com efeitos dominantes e epistáticos. Já para a identificação de marcadores associados às regiões de presença de QTLs, os métodos de aprendizado de máquinas foram mais eficientes.

O uso de diferentes métodos estatísticos, redes neurais artificiais e aprendizado de máquina resultou em diferentes consequências influenciadas pela complexidade e particularidade das características analisadas. Portanto, recomenda-se que ao avaliar a predição de valores genéticos e a importância de marcadores, o uso de múltiplas abordagens seja utilizado, para escolha do melhor método.

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, comparadas a metodologias tradicionalmente aplicadas. Esses conhecimentos podem fornecer subsídios para a escolha de metodologias mais apropriadas para a predição de valores genéticos superiores e detecção de marcadores associados à QTLs em características com efeitos aditivos e não-aditivos, tornando mais eficientes as atividades dentro de um programa de melhoramento genético.

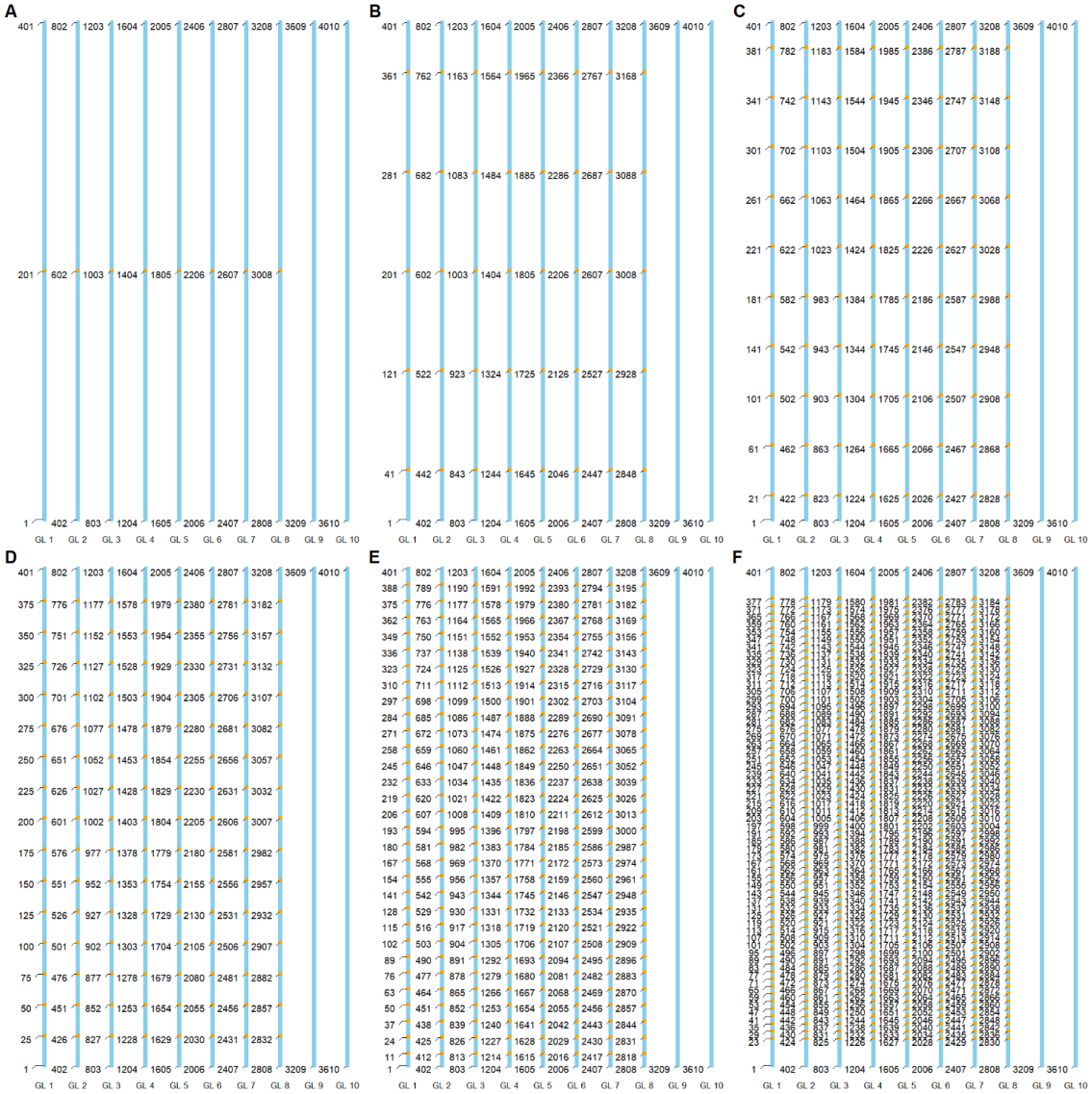
A execução deste trabalho, permitiu adquirir conhecimentos para um melhor entendimento da influência dos diferentes cenários de complexidade, modelos de simulação e suas consequências para os resultados de R^2 e REQM. Além disso, permitiu avaliar diferentes métodos de aprendizado de máquinas para a identificação de marcadores nos diferentes cenários. Espera-se ainda que os resultados possam

ser motivadores, no sentido de outros pesquisadores optarem pelo uso dos métodos, de forma que possam agregar conhecimento e aumentar a eficiência em seus programas de melhoramento.

Os conhecimentos gerados por este trabalho poderão ser utilizados como literatura para auxiliar no desenvolvimento de cientistas em diversas áreas de estudo, como na genética, estatística, bioinformática, melhoramento de plantas, entre outros. 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.

ANEXO A – Figura suplementar artigo 1

Supplementary Figure 3 - Model of distribution of loci in linkage groups for the traits. QTL distribution model for the characteristics: (A) C1 and C7 with the central markers in the first eight linkage groups; (B) C2 and C8, (C) C3 and C9, (D) C4 and C10, (E) C5 and C11 and (F) C6 and C12, with 40, 80, 120, 240 and 480 QTLs, respectively, distributed equidistantly within the first 8 linkage groups. *GL: Linkage group.



ANEXO B – Figura suplementar artigo 2

Figura Suplementar 1 - Importância das marcas em porcentagem para a característica 1 em função dos métodos: Bagging (BA), Boosting (BO), Árvore de Decisão (DT); Random Forest (RF); Rede Perceptron Multicamadas (MLP) e Rede de Função de Base Radial (RBF); MARS 1, MARS 2 e MARS 3; e G-BLUP de acordo com grupo de ligação. Os pontos em vermelho referem-se à identificação de marcas dentro de um QTL com verdadeira importância para característica 1. Os pontos em verde referem-se à não identificação de marcas dentro de um QTL com verdadeira importância para característica 1. Os pontos em laranja referem-se à identificação de marcas fora da região de um QTLs de importância para característica 1. A linha pontilhada vermelha refere-se ao ponte de corte de 12.5 para identificação de importância para as marcas. A faixa em rosê refere-se ao QTL com verdadeira importância para o grupo de ligação da característica 1.

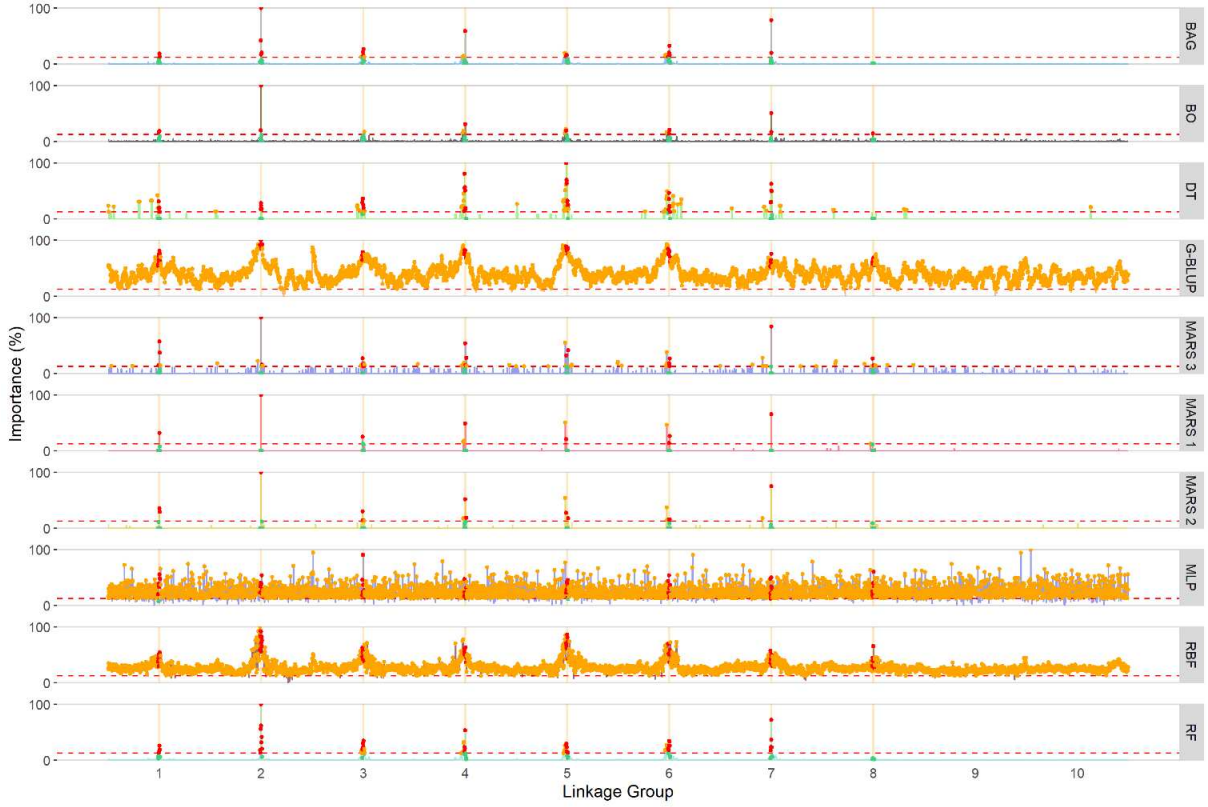


Figura Suplementar 2 - Importância das marcas em porcentagem para a característica 2 em função dos métodos: Bagging (BA), Boosting (BO), Árvore de Decisão (DT); Random Forest (RF); Rede Perceptron Multicamadas (MLP) e Rede de Função de Base Radial (RBF); MARS 1, MARS 2 e MARS 3; e G-BLUP de acordo com grupo de ligação. Os pontos em vermelho referem-se à identificação de marcas dentro de um QTL com verdadeira importância para característica 2. Os pontos em verde referem-se à não identificação de marcas dentro de um QTL com verdadeira importância para característica 2. Os pontos em laranja referem-se à identificação de marcas fora da região de um QTLs de importância para característica 2. A linha pontilhada vermelha refere-se ao ponte de corte de 12.5 para identificação de importância para as marcas. A faixa em rosê refere-se ao QTL com verdadeira importância para o grupo de ligação da característica 2.

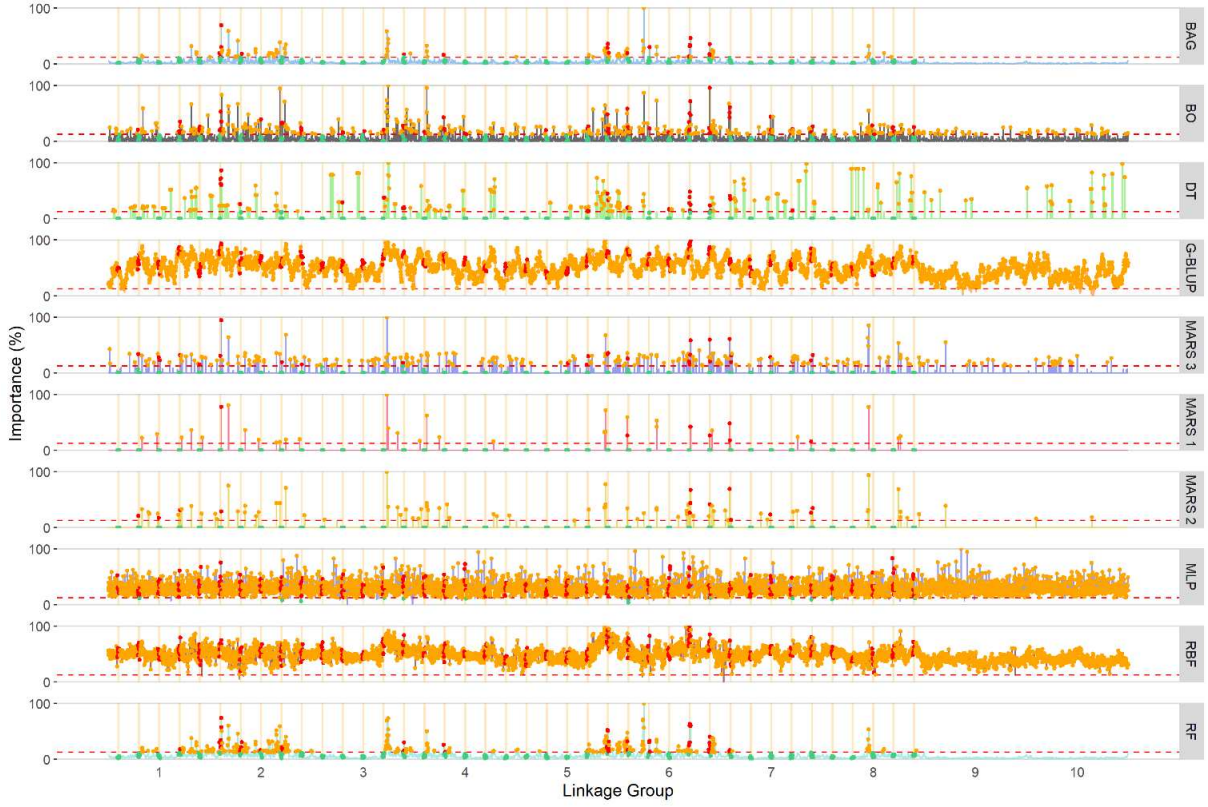


Figura Suplementar 3 - Importância das marcas em porcentagem para a característica 3 em função dos métodos: Bagging (BA), Boosting (BO), Árvore de Decisão (DT); Random Forest (RF); Rede Perceptron Multicamadas (MLP) e Rede de Função de Base Radial (RBF); MARS 1, MARS 2 e MARS 3; e G-BLUP de acordo com grupo de ligação. Os pontos em vermelho referem-se à identificação de marcas dentro de um QTL com verdadeira importância para característica 3. Os pontos em verde referem-se à não identificação de marcas dentro de um QTL com verdadeira importância para característica 3. Os pontos em laranja referem-se à identificação de marcas fora da região de um QTLs de importância para característica 3. A linha pontilhada vermelha refere-se ao ponte de corte de 12.5 para identificação de importância para as marcas. A faixa em rosê refere-se ao QTL com verdadeira importância para o grupo de ligação da característica 3.

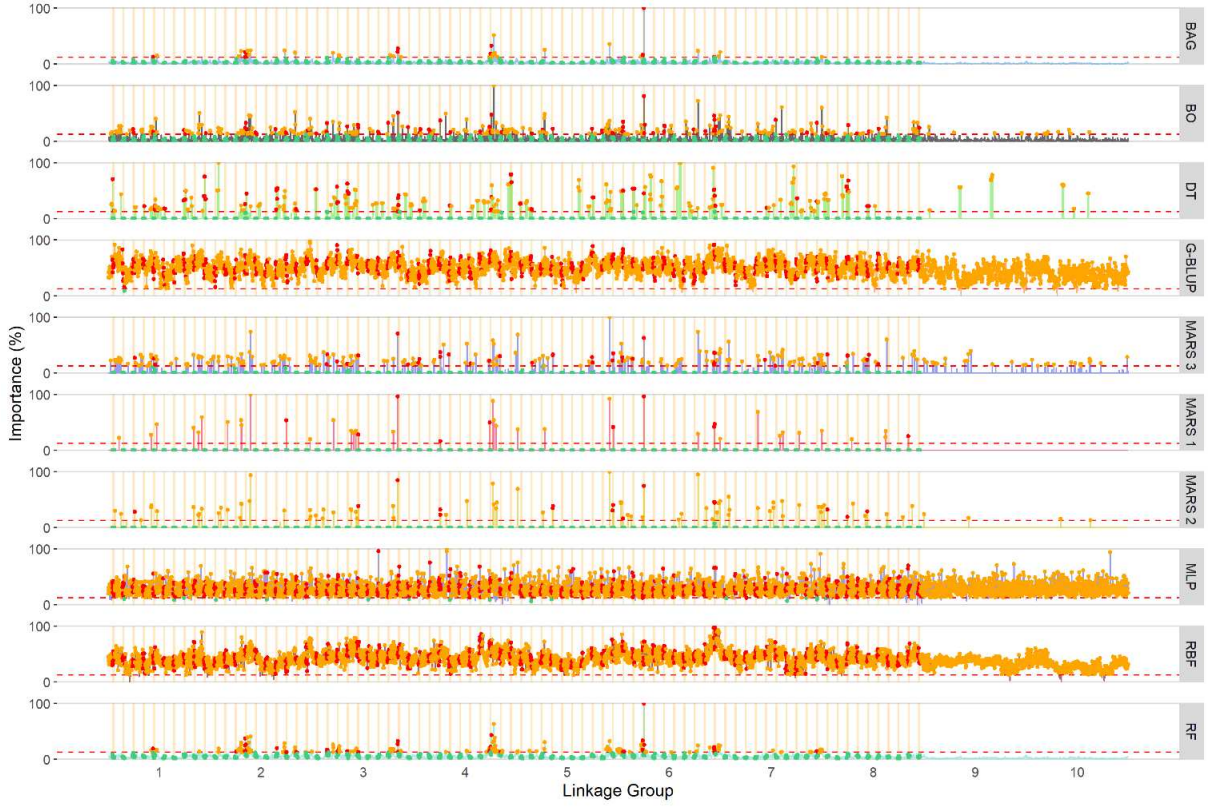


Figura Suplementar 4 - Importância das marcas em porcentagem para a característica 4 em função dos métodos: Bagging (BA), Boosting (BO), Árvore de Decisão (DT); Random Forest (RF); Rede Perceptron Multicamadas (MLP) e Rede de Função de Base Radial (RBF); MARS 1, MARS 2 e MARS 3; e G-BLUP de acordo com grupo de ligação. Os pontos em vermelho referem-se à identificação de marcas dentro de um QTL com verdadeira importância para característica 4. Os pontos em verde referem-se à não identificação de marcas dentro de um QTL com verdadeira importância para característica 4. Os pontos em laranja referem-se à identificação de marcas fora da região de um QTLs de importância para característica 4. A linha pontilhada vermelha refere-se ao ponte de corte de 12.5 para identificação de importância para as marcas. A faixa em rosê refere-se ao QTL com verdadeira importância para o grupo de ligação da característica 4.

Figura Suplementar 5 - Importância das marcas em porcentagem para a característica 5 em função dos métodos: Bagging (BA), Boosting (BO), Árvore de Decisão (DT); Random Forest (RF); Rede Perceptron Multicamadas (MLP) e Rede de Função de Base Radial (RBF); MARS 1, MARS 2 e MARS 3; e G-BLUP de acordo com grupo de ligação. Os pontos em vermelho referem-se à identificação de marcas dentro de um QTL com verdadeira importância para característica 5. Os pontos em verde referem-se à não identificação de marcas dentro de um QTL com verdadeira importância para característica 5. Os pontos em laranja referem-se à identificação de marcas fora da região de um QTLs de importância para característica 5. A linha pontilhada vermelha refere-se ao ponte de corte de 12.5 para identificação de importância para as marcas. A faixa em rosê refere-se ao QTL com verdadeira importância para o grupo de ligação da característica 5.

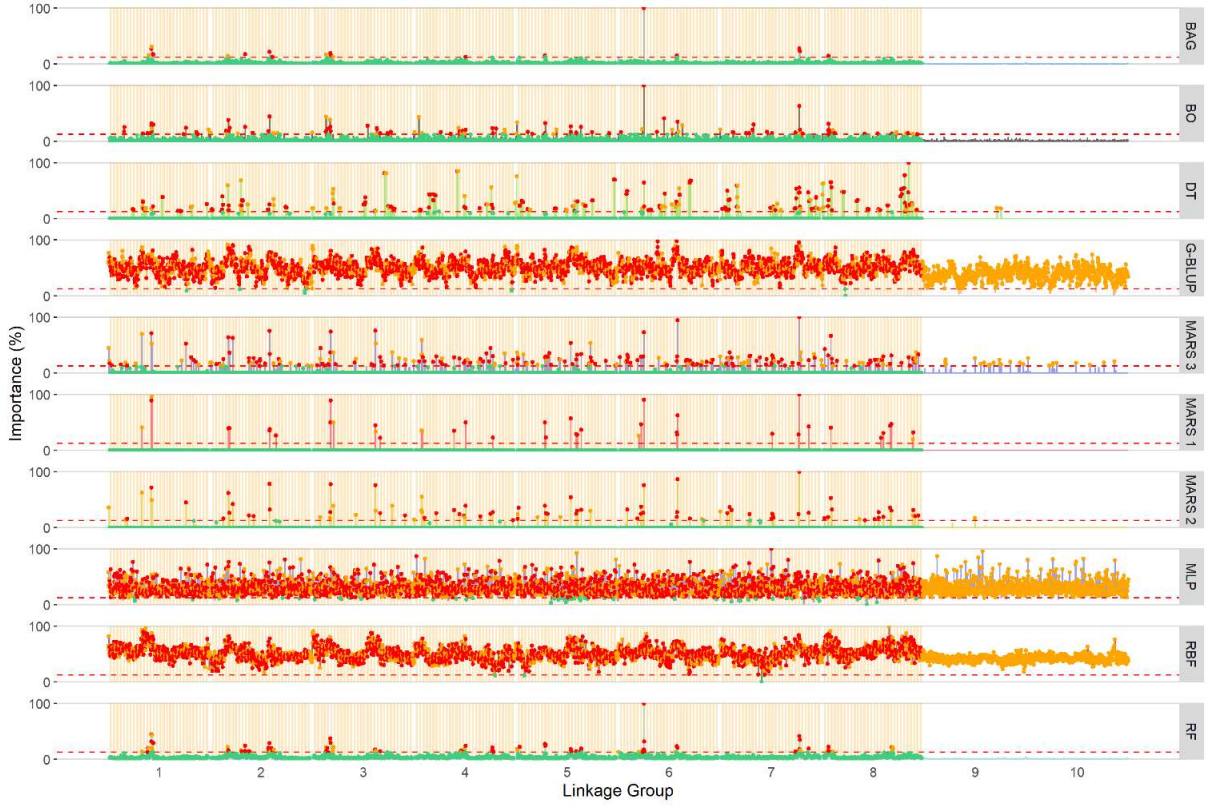


Figura Suplementar 6 - Importância das marcas em porcentagem para a característica 6 em função dos métodos: Bagging (BA), Boosting (BO), Árvore de Decisão (DT); Random Forest (RF); Rede Perceptron Multicamadas (MLP) e Rede de Função de Base Radial (RBF); MARS 1, MARS 2 e MARS 3; e G-BLUP de acordo com grupo de ligação. Os pontos em vermelho referem-se à identificação de marcas dentro de um QTL com verdadeira importância para característica 6. Os pontos em verde referem-se à não identificação de marcas dentro de um QTL com verdadeira importância para característica 6. Os pontos em laranja referem-se à identificação de marcas fora da região de um QTLs de importância para característica 6. A linha pontilhada vermelha refere-se ao ponte de corte de 12.5 para identificação de importância para as marcas. A faixa em rosê refere-se ao QTL com verdadeira importância para o grupo de ligação da característica 6.

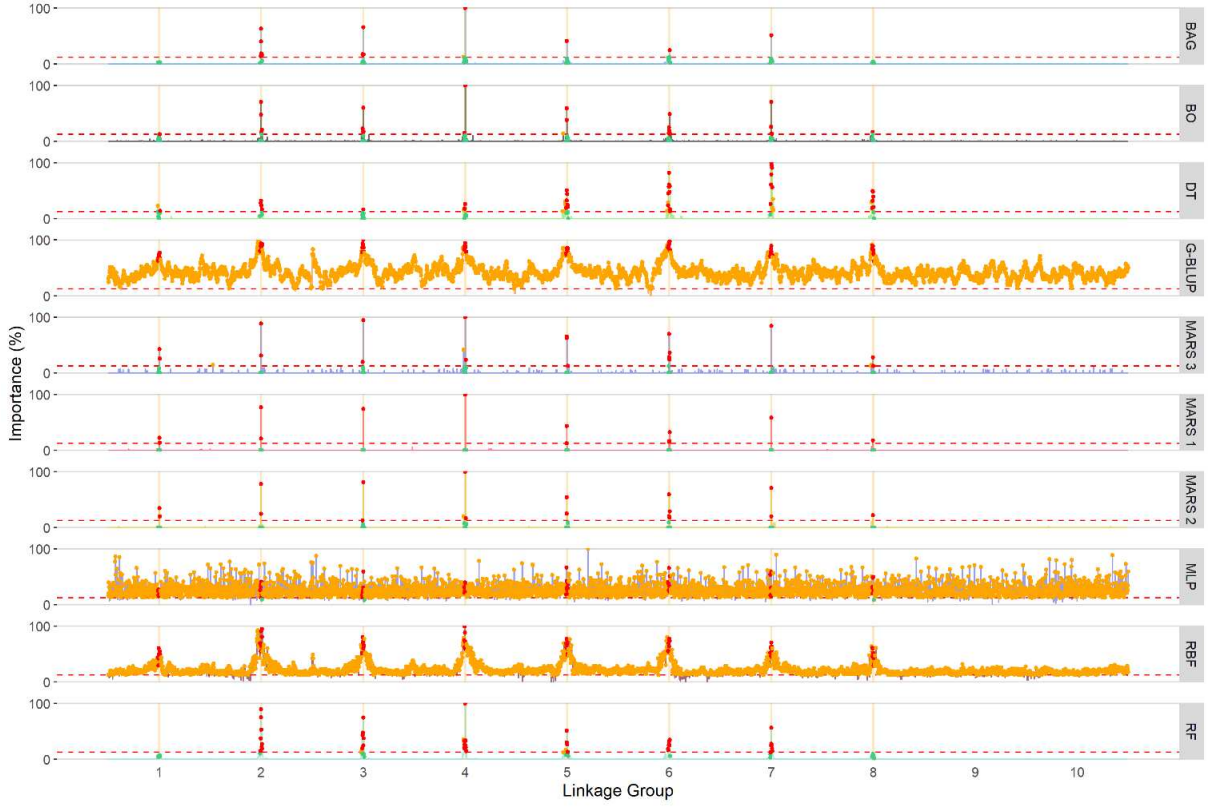


Figura Suplementar 7 - Importância das marcas em porcentagem para a característica 7 em função dos métodos: Bagging (BA), Boosting (BO), Árvore de Decisão (DT); Random Forest (RF); Rede Perceptron Multicamadas (MLP) e Rede de Função de Base Radial (RBF); MARS 1, MARS 2 e MARS 3; e G-BLUP de acordo com grupo de ligação. Os pontos em vermelho referem-se à identificação de marcas dentro de um QTL com verdadeira importância para característica 7. Os pontos em verde referem-se à não identificação de marcas dentro de um QTL com verdadeira importância para característica 7. Os pontos em laranja referem-se à identificação de marcas fora da região de um QTLs de importância para característica 7. A linha pontilhada vermelha refere-se ao ponte de corte de 12.5 para identificação de importância para as marcas. A faixa em rosê refere-se ao QTL com verdadeira importância para o grupo de ligação da característica 7.

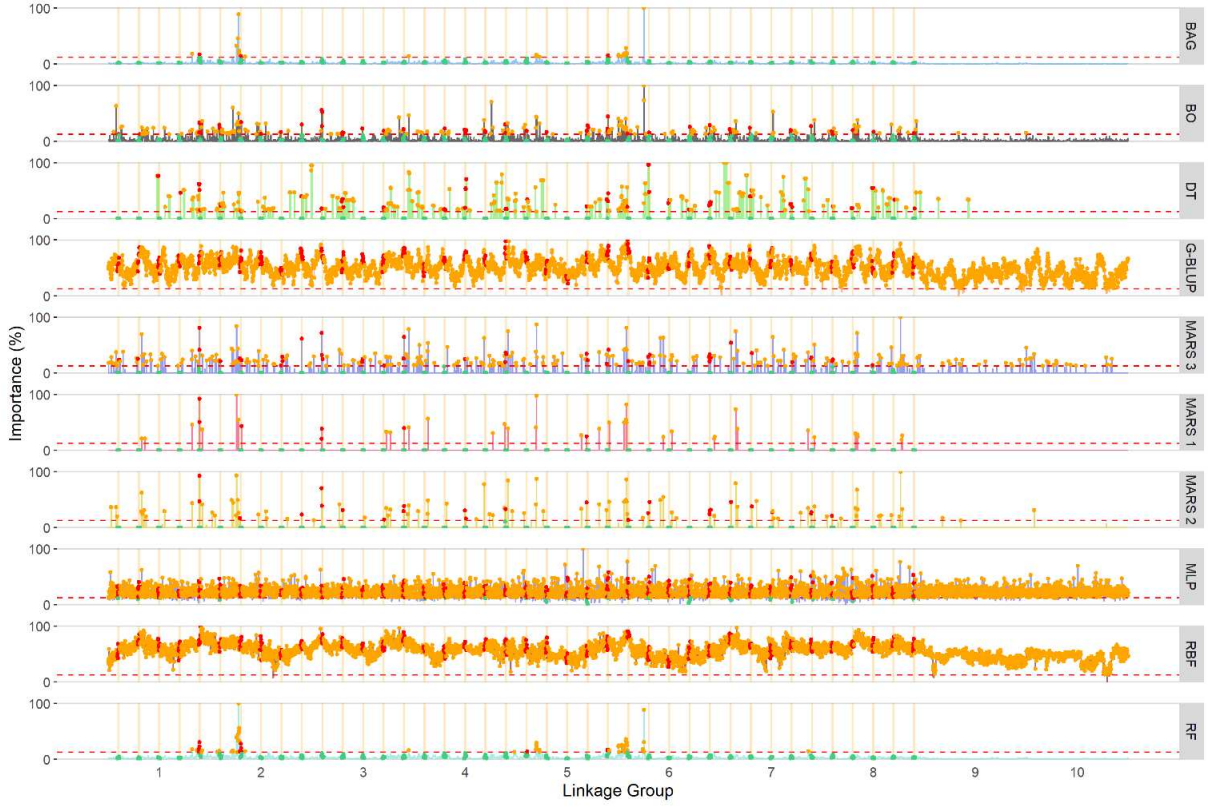


Figura Suplementar 8 - Importância das marcas em porcentagem para a característica 8 em função dos métodos: Bagging (BA), Boosting (BO), Árvore de Decisão (DT); Random Forest (RF); Rede Perceptron Multicamadas (MLP) e Rede de Função de Base Radial (RBF); MARS 1, MARS 2 e MARS 3; e G-BLUP de acordo com grupo de ligação. Os pontos em vermelho referem-se à identificação de marcas dentro de um QTL com verdadeira importância para característica 8. Os pontos em verde referem-se à não identificação de marcas dentro de um QTL com verdadeira importância para característica 8. Os pontos em laranja referem-se à identificação de marcas fora da região de um QTLs de importância para característica 8. A linha pontilhada vermelha refere-se ao ponte de corte de 12.5 para identificação de importância para as marcas. A faixa em rosê refere-se ao QTL com verdadeira importância para o grupo de ligação da característica 8.

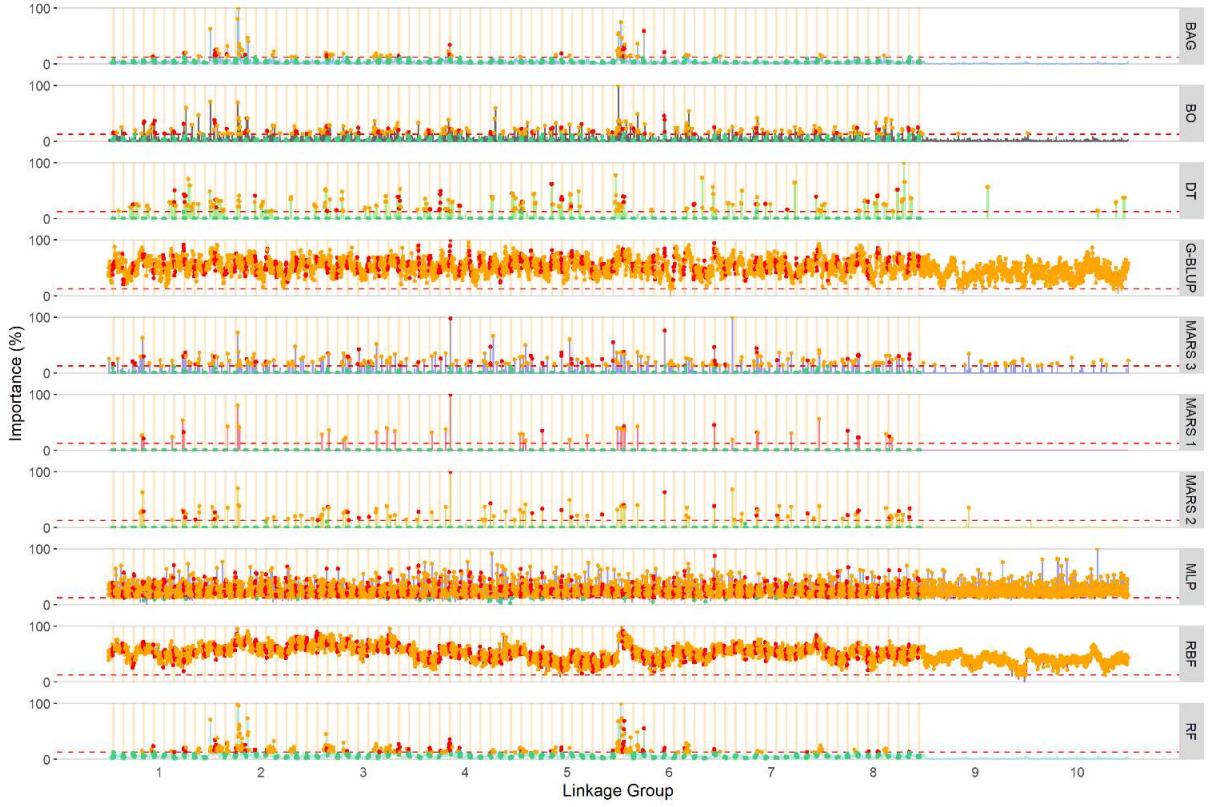


Figura Suplementar 9 - Importância das marcas em porcentagem para a característica 9 em função dos métodos: Bagging (BA), Boosting (BO), Árvore de Decisão (DT); Random Forest (RF); Rede Perceptron Multicamadas (MLP) e Rede de Função de Base Radial (RBF); MARS 1, MARS 2 e MARS 3; e G-BLUP de acordo com grupo de ligação. Os pontos em vermelho referem-se à identificação de marcas dentro de um QTL com verdadeira importância para característica 9. Os pontos em verde referem-se à não identificação de marcas dentro de um QTL com verdadeira importância para característica 9. Os pontos em laranja referem-se à identificação de marcas fora da região de um QTLs de importância para característica 9. A linha pontilhada vermelha refere-se ao ponte de corte de 12.5 para identificação de importância para as marcas. A faixa em rosê refere-se ao QTL com verdadeira importância para o grupo de ligação da característica 9.

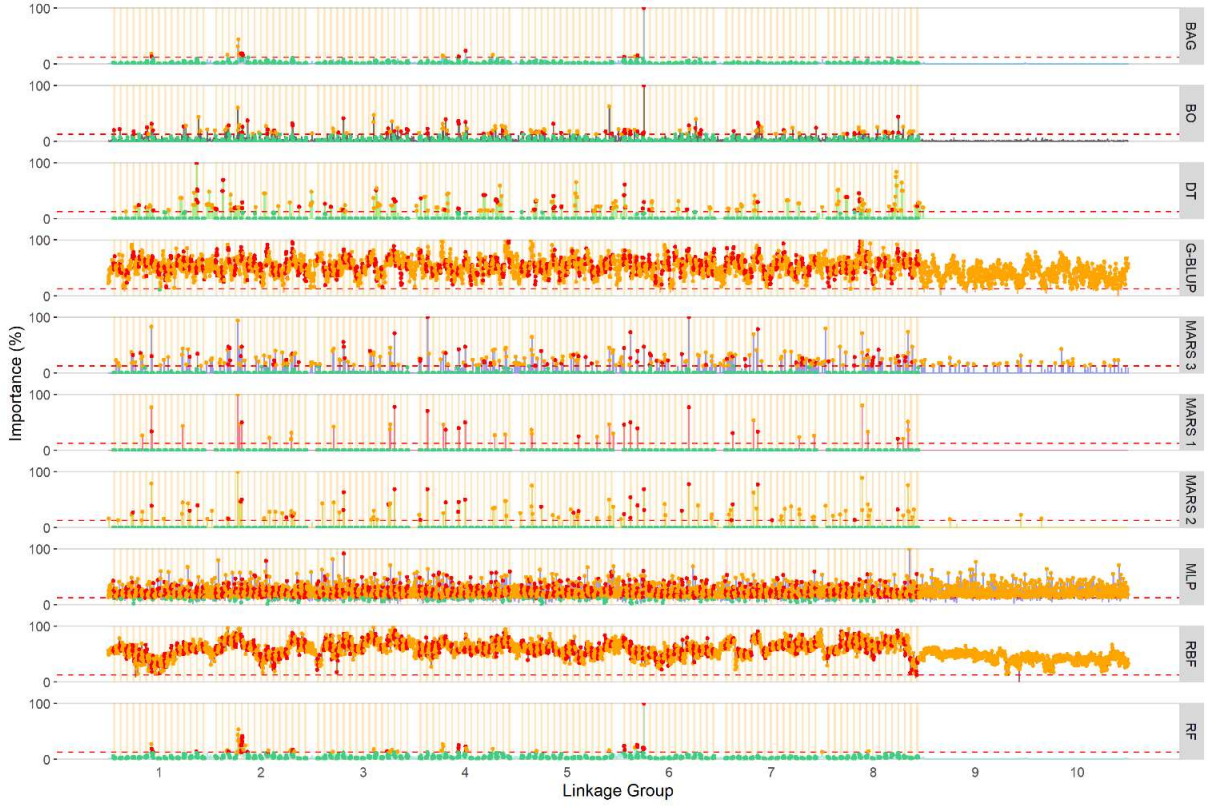


Figura Suplementar 10 - Importância das marcas em porcentagem para a característica 10 em função dos métodos: Bagging (BA), Boosting (BO), Árvore de Decisão (DT); Random Forest (RF); Rede Perceptron Multicamadas (MLP) e Rede de Função de Base Radial (RBF); MARS 1, MARS 2 e MARS 3; e G-BLUP de acordo com grupo de ligação. Os pontos em vermelho referem-se à identificação de marcas dentro de um QTL com verdadeira importância para característica 10. Os pontos em verde referem-se à não identificação de marcas dentro de um QTL com verdadeira importância para característica 10. Os pontos em laranja referem-se à identificação de marcas fora da região de um QTLs de importância para característica 10. A linha pontilhada vermelha refere-se ao ponte de corte de 12.5 para identificação de importância para as marcas. A faixa em rosê refere-se ao QTL com verdadeira importância para o grupo de ligação da característica 10.