NOÉ MITTERHOFER EITERER PONCE DE LEON DA COSTA

REDES NEURAIS REGULARIZADAS NA PREDIÇÃO DE CARACTERÍSTICAS AGRONÔMICAS DE SOJA

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Doctor Scientiae*.

Orientador: Moysés Nascimento

Coorientadoras: Ana Carolina Campana Nascimento
Camila Ferreira Azevedo

VIÇOSA - MINAS GERAIS
2024

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

T
C837r
2024

Costa, Noé Mitterhofer Eiterer Ponce de Leon da, 1998-
Redes neurais regularizadas na predição de características agronômicas de soja / Noé Mitterhofer Eiterer Ponce de Leon da Costa. – Viçosa, MG, 2024.
1 tese eletrônica (54 p.): il. (algumas color.).

Inclui apêndice.
Orientador: Moysés Nascimento.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Estatística, 2024.
Referências bibliográficas: f. 39-49.
DOI: https://doi.org/10.47328/ufvbbt.2024.130
Modo de acesso: World Wide Web.

1. Teoria bayesiana de decisão estatística. 2. Genômica. 3. Redes neurais (Computação). 4. Aprendizado do computador. 5. Controle preditivo. 6. Soja - Melhoramento genético - Métodos estatísticos. I. Nascimento, Moysés, 1979-. II. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. III. Título.

CDD 22. ed. 519.542

Bibliotecário(a) responsável: Bruna Silva CRB-6/2552

NOÉ MITTERHOFER EITERER PONCE DE LEON DA COSTA

REDES NEURAIS REGULARIZADAS NA PREDIÇÃO DE CARACTERÍSTICAS AGRONÔMICAS DE SOJA

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Doctor Scientiae*.

APROVADA: 03/04/2024

Assentimento:

gov.br Documento assinado digitalmente
NOE MITTERHOFER EITERER PONCE DE LEON D.
Data: 04/04/2024 16:02:05-0300
Verifique em https://validar.iti.gov.br

Noé Mitterhofer Eiterer Ponce de Leon da Costa
Autor

gov.br Documento assinado digitalmente
MOYSES NASCIMENTO
Data: 05/04/2024 09:07:12-0300
Verifique em https://validar.iti.gov.br

Moysés Nascimento
Orientador

À minha querida mãe (in memorian),
Marinês Eiterer, que tanto sonhou por esse
título, o meu e o seu sonho está realizado.
Ao meu pai, Felipe André
Ponce de Leon da Costa.
À minha irmã, Flora Mitterhofer Eiterer
Ponce de Leon da Costa.
À minha tia Teresa Cristina Eiterer, por todo
apoio incondicional.

AGRADECIMENTOS

À Universidade Federal de Viçosa, pela oportunidade de realizar minha graduação e pós-graduação.

À minha falecida e querida mãe, pessoa que se manteve forte e corajosa mesmos nos momentos mais difíceis, sempre fez o melhor para sua família. Sinto e sentirei sua falta eternamente em minha vida. Esse momento da minha vida seria ainda mais feliz na sua presença. Ao meu pai, Felipe André Ponce de Leon da Costa. À minha irmã, Flora Mitterhofer Eiterer Ponce de Leon da Costa. À minha tia Teresa pelo apoio que me ajudou a continuar em frente nessa vida.

Ao professor Moysés Nascimento pelos grandes ensinamentos, paciência, apoio, grande amizade, disponibilidade e conselhos. Uma pessoa que sempre irei admirar pela sua humildade e sabedoria. À professora Ana Carolina pela orientação, amizade, disponibilidade e conselhos. Aos professores Eduardo, Filipe e Ithalo pelo aceite em participar da banca e pelas sugestões que contribuem na melhoria desse trabalho.

Aos docentes do Departamento de Estatística da UFV, especialmente os professores Carlos Henrique, Eduardo, Fernando, José Ivo e Paulo Cecon, pessoas que sempre me ajudaram e apoiaram. Ao Júnior por todo auxílio e pela amizade que fizemos na secretaria do DET. Ao prof. Rodrigo Oliveira de Lima por toda sua orientação, ensinamentos e amizade na minha graduação em Agronomia. À profa. Flaviane Ribeiro pela amizade, oportunidades, ensinamentos e conselhos. À todos professores, servidores e estudantes do Departamento de Estatística e Agronomia da UFV que participaram e colaboraram com minha formação. Aos meus amigos do Licae (Laboratório de Inteligência Computacional e Aprendizado Estatístico), que foram minha família no DET e auxiliaram no meu desenvolvimento pessoal.

Eu não acredito em nenhum Deus, mas tenho a minha fé e agradeço pela vida com saúde que tenho.

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

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela concessão da bolsa de estudos.

MUITO OBRIGADO!

A vida é um sopro.

RESUMO

COSTA, Noé Mitterhofer Eiterer Ponce de Leon da, D.Sc., Universidade Federal de Viçosa, abril de 2024. **Redes neurais regularizadas na predição de características agronômicas de soja**. Orientador: Moysés Nascimento. Coorientadoras: Ana Carolina Campana Nascimento e Camila Ferreira Azevedo.

Um dos métodos de aprendizado de máquina utilizado atualmente na Seleção Genômica (SG) são as Redes Neurais Artificiais (RNAs) e, dentre estas, a Rede *Perceptron* de Múltiplas Camadas (PMC). O PMC destaca-se na solução de problemas de classificação ou regressão pelo fato de não exigir um modelo funcional, nem de atender pressuposições e não requerer conhecimento a *priori* sobre o fenômeno em estudo. No entanto, um problema comum nas PMC é o *overfitting*, que se trata de um superajustamento da rede aos dados de treinamento. Nestes casos, o modelo não possui capacidade de generalização fazendo que ele seja menos eficaz nas predições em um novo conjunto de dados ou no conjunto de teste. Para contornar este problema, algumas alternativas são as regularizações L1 e L2, que se baseiam nas regressões em penalizações similares aos métodos *Lasso* e *Ridge*, respectivamente. O objetivo deste estudo foi avaliar a eficiência do uso da regularização em modelos de PMC aplicados na predição genômica. Além disso, os resultados obtidos foram comparados com outros utilizados em predição genômica, tais como o *Perceptron* de Múltiplas Camadas (PMC), Árvore de Decisão (AD), *Random Forest* (RF), *Bagging* (BAG), *Boosting* (BOO) e *Genomic Best Linear Unbiased Prediction* (GBLUP). Os dados são provenientes de 100 genótipos de soja, em um experimento conduzido de setembro a novembro de 2021, no delineamento em blocos ao acaso com três repetições, em que cada parcela foi constituída de uma planta cultivada em um vaso dentro de uma casa de vegetação. Foram avaliadas as características diâmetro de hipocótilo (DH, em milímetros), altura de planta (AP, em centímetros), comprimento total de raiz (CR, em centímetros) e área superficial projetada de raiz (AR, em centímetros quadrados). Na avaliação do PMC regularizado (PMCR), foram utilizados as medidas de capacidade preditiva (CP) e raiz do erro quadrático médio (RMSE) para comparação dos métodos. Em geral, o PMC com regularização L2 melhorou o desempenho em comparação com métodos avaliados em termos de CP e RMSE. Os valores de CP obtidos pelas redes regularizadas L2 foram melhores que todos os métodos avaliados. Especificamente, quando comparado com o segundo o melhor método, os ganhos em termos de CP foram de

6,05%, 25,86%, 32,90% e 0,16% para as características, respectivamente, AP, DH, CR e AR. Já em termos de RMSE, o PMCR apresentou resultados inferiores e desejáveis em 10,81%, 15,00%, 13,62% e 20,41% para as características AP, DH, CR e AR, respectivamente, quando comparado com as redes sem regularização. Quando a comparação é entre todas as metodologias comparativas, o GBLUP obteve o menor valor de RMSE para todas características avaliadas.

Palavras-chave: Capacidade Preditiva. *Machine Learning*. Predição Genômica.

ABSTRACT

COSTA, Noé Mitterhofer Eiterer Ponce de Leon da, D.Sc., Universidade Federal de Viçosa, April 2024. **Regularized neural networks to predict soybean agronomic traits**. Adviser: Moysés Nascimento. Co-advisers: Ana Carolina Campana Nascimento and Camila Ferreira Azevedo.

One of the machine learning methods currently used in Genomic Selection (GS) are Artificial Neural Networks (ANNs) and, among these, the Multilayer Perceptron Network (MLP). MLP stands out in solving classification or regression problems because it does not require a functional model, nor does it meet presuppositions and does not require a priori knowledge about the phenomenon under study. However, a common problem in MLP is overfitting, which is an overfitting of the network to the training data. In these cases, the model does not have generalization capacity, making it less effective in making predictions on a new set of data or on the test set. To overcome this problem, some alternatives are L1 and L2 regularizations, which are based on regressions in penalties similar to the Lasso and Ridge methods, respectively. The objective of this study was to evaluate the efficiency of using regularization in MLP models applied in genomic prediction. Furthermore, the results obtained were compared with others used in genomic prediction, such as Multilayer Perceptron (MLP), Decision Tree (DT), Random Forest (RF), Bagging (BAG), Boosting (BOO) and Genomic Best Linear Unbiased Prediction (GBLUP). The data come from 100 soybean genotypes, in an experiment conducted from September to November 2021, in a randomized block design with three replications, in which each plot consisted of a plant grown in a pot inside a greenhouse. The characteristics of hypocotyl diameter (HD, in millimeters), plant height (PH, in centimeters), total root length (RL, in centimeters) and projected root surface area (SA, in square centimeters) were evaluated. In evaluating the regularized MPL (MPLR), the predictive capacity (PC) and root mean square error (RMSE) measures were used to compare the methods. In general, MPL with L2 regularization improved performance compared to methods evaluated in terms of PC and RMSE. The PC values obtained by the L2 regularized networks were better than all the methods evaluated. Specifically, when compared with the second best method, the gains in terms of PC were 6.05%, 25.86%, 32.90% and 0.16% for the characteristics, respectively, PH, HD, RL and SA. In terms of RMSE, the MPLR presented inferior and desirable results at 10.81%, 15.00%, 13.62% and 20.41% for the characteristics PH, HD, RL and SA, respectively, when compared with

the networks without regularization. When the comparison is between all comparative methodologies, GBLUP obtained the lowest RMSE value for all characteristics evaluated.

Keywords: Predictive Capacity. Machine Learning. Genomic Prediction.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

AD Árvore de Decisão

BAG *Bagging*

BOO *Boosting*

CP Capacidade Preditiva

GBLUP *Genomic Best Linear Unbiased Prediction*

PMC *Perceptron* de Múltiplas Camadas

PMCR *Perceptron* de Múltiplas Camadas com Regularização

PMCR-L1 *Perceptron* de Múltiplas Camadas com Regularização L1

PMCR-L2 *Perceptron* de Múltiplas Camadas com Regularização L2

RAF *Random Forest*

RNA Rede Neural Artificial

RMSE Raiz do Erro Quadrático Médio

RBF Rede Neural com Função de Base Radial

SG Seleção Genômica

QTL *Quantitative Trait Locus*

SUMÁRIO

1. INTRODUÇÃO

Entre as diversas culturas relevantes para o agronegócio brasileiro, a soja [*Glycine max* (L.) Merrill] se destaca, pois possui um uso versátil, sendo utilizada na produção de diversos produtos e subprodutos usados pela agroindústria, indústria química e de alimentos (Shea et al., 2020). Nesse contexto, observa-se um aumento crescente no cultivo de soja em escala mundial (Matei et al., 2018; Shea et al., 2020). Após a consolidação da soja no cerrado, as instituições de pesquisa promoveram significativos avanços, levando o Brasil à posição de um dos principais produtores mundiais do grão. Devido à crescente demanda por alimentos, o melhoramento genético de plantas ganha ainda mais importância na tentativa de aumentar a produtividade agrícola (Fukase; Martin, 2020).

Atualmente, o ganho genético na produtividade da soja está entre 0,5 e 1,8% ao ano nos diferentes países ao redor do mundo, enquanto a média dos três maiores países produtores (Argentina, Estados Unidos e Brasil) este ganho é de 1,1% ao ano (Matei et al., 2018). Este valor é ainda insuficiente para a atender a futura demanda de soja (Matei et al., 2018). O recente avanço no melhoramento genético de plantas está permitindo uma identificação mais rápida desses genótipos considerados superiores nos programas de melhoramento (Torkamaneh et al., 2018). Uma das áreas com avanços recentes e significativos é a seleção genômica (SG) que permite a identificação mais acurada e mais rápida de genótipos superiores (Crossa et al., 2017).

A SG permite identificar e selecionar genótipos superiores por meio de marcadores moleculares, distribuídos ao longo do genoma, em desequilíbrio de ligação com *loci* associados às características quantitativas de interesse (*Quantitative Trait Loci* - QTL) (Meuwissen et al., 2001). O desequilíbrio de ligação consiste na distribuição não aleatória dos alelos nos diferentes *loci*, ou seja, na presença de correlação entre alelos (Qanbari, 2020). A SG permite estimar os efeitos genéticos de marcadores em todo o genoma simultaneamente, combinando dados moleculares e fenotípicos em um conjunto de treinamento para obter estimativas de valores genéticos genômicos dos indivíduos em um conjunto de teste que foi apenas genotipado (Crossa et al., 2017). Assim, o pesquisador pode ajustar um modelo que, uma vez validado, permite a predição do valor genético de indivíduos sem a necessidade de conhecê-los previamente (Crossa et al., 2017).

Os principais métodos de predição genômica podem ser agrupados em métodos paramétricos, quando é necessário estabelecer um modelo funcional entre variável resposta e variáveis explicativas, e não paramétrico quando nenhuma relação funcional entre as variáveis é assumida (James et al., 2013; de Los Campos et al., 2012). Dentro dessa classificação, os métodos RR-BLUP, *Lasso* (*Least Absolute Shrinkage and Selection Operator*), *Elastic Net*, BayesA, BayesB e entre outros são classificados como métodos paramétricos (Resende et al., 2014). Já na segunda classe, destacam-se os métodos como redes neurais, RKHS (*Reproducing Kernel Hilbert Spaces*), regressão kernel não paramétrica via modelos aditivos generalizados, árvores de decisão e seus refinamentos (Resende et al., 2014; de Los Campos et al., 2012). Também existem métodos que proporcionam redução de dimensionalidade como os mínimos quadrados parciais, componentes principais, componentes esparsos, componentes independentes e análise de fatores (Resende et al., 2014).

Dentre os métodos não paramétricos, as redes neurais artificiais (RNAs) vem se destacando por não exigirem um modelo funcional previamente informado pelo pesquisador, nem a necessidade de satisfazer pressuposições a respeito das distribuições da variável resposta ou dos erros, como em métodos estocásticos. Além disso, não demandam nenhum conhecimento *a priori*. Além destas vantagens, destacam-se ainda a capacidade de capturar relações lineares e não lineares entre as variáveis, bem como uma boa adaptação em cenários que envolvem interações complexas entre genes, entre genes e ambiente ou efeitos epigenéticos (Sant' Anna et al., 2019; Cruz; Nascimento, 2018; Zingaretti et al., 2020; Sousa et al., 2021).

Apesar das vantagens citadas, um problema comum em redes neurais é o *overfitting*, caracterizado pelo ajuste excessivo aos dados de treinamento, tornando o modelo menos eficaz na predição de um novo conjunto de dados (Pérez-Enciso; Zingaretti, 2019). Para reduzir o *overfitting* em redes neurais, alguma das abordagens utilizadas são a *early stopping*, *dropout* e as regularizações. A *early stopping* consiste em avaliar a taxa de erro de validação e encerrar o treinamento da rede quando essa taxa atinge seu mínimo. No entanto, exige que o pesquisador investigue o momento em que o erro de validação alcança esse mínimo, que essa investigação geralmente é feita por um gráfico que plota a taxa de erro em função do número de iterações ou épocas, além da dependência dos dados utilizados para treinar a rede (Svozil et al., 1997). O *dropout* consiste no treinamento de várias redes em que neurônios são

aleatoriamente retirados com o objetivo de evitar o treinamento excessivo da rede (Srivastava et al., 2014). A rede neural final é construída a partir da média das redes reduzidas, selecionando-se os neurônios com as maiores probabilidades de retenção (Srivastava et al., 2014).

Uma abordagem mais elegante do ponto de vista matemático e estatístico é a utilização de penalizações (regularização) no ajuste dos pesos de uma rede neural. As regularizações mais comumente empregadas em redes são do tipo L1, a mesma utilizada na regressão *Lasso* (Tibshirani, 1996), ou do tipo L2, utilizada na regressão em *Ridge* (Hoerl; Kennard, 1970). Essas duas metodologias empregam o procedimento de *shrinkage* (encolhimento), que consiste na adição do parâmetro de penalização (λ) na equação de custo do modelo a ser ajustado (Van Erp et al., 2019). As diferenças entre as regularizações dos tipos L1 e L2 residem na forma como a penalização é adicionada, sendo absoluta para L1 e quadrática para L2, respectivamente (Hoerl; Kennard, 1970; Tibshirani, 1996). Essas penalizações são ainda pouco exploradas na área de métodos estatísticos aplicados à seleção genômica.

Nusrat e Jang (2018) avaliaram redes neurais artificiais sem regularização e com regularização L1, além dos métodos *batch normalization*, *autoencoder* e *data augmentation* na predição da temperatura em dados reais disponibilizados pelo governo da Coreia do Sul. Os autores concluíram que a regularização L1 apresentou desempenho superior a uma abordagem que utiliza redução de dimensionalizado por meio de uma rede *autoencoder* e similar aos demais métodos. Porém segundo os autores, os problemas de *overfitting* e *underfitting* com seu conjunto de dados são complexos e não foram completamente resolvidos pelos métodos utilizados. Outra constatação dos autores foi a escassez de literatura disponível para a comparação dos métodos de regularização, apesar da importância desse tema no ajuste de redes neurais.

Mehdi et al. (2023) avaliaram o uso de regularização L2 e *dropout* em redes neurais convolucionais para classificação de imagens de algarismos indo-arábicos do conjunto *Modified National Institute of Standards and Technology* (MNIST) (LeCun et al, 2010). Foi observado pelos autores que as redes L2 apresentaram melhores medidas de acurácia do que o *dropout* em redes neurais com até 100 neurônios em sua camada intermediária. Yang et al. (2023) propuseram uma rede neural profunda

para lidar com dados de risco de crédito de alta dimensionalidade, combinando as regularizações L1 e L2. A combinação dos métodos de regularização, segundo os autores, é justificada pelo fato que a norma L1 irá contribuir com a esparsidade e subsequente seleção de variáveis, enquanto a L2 atua na redução dos pesos restantes. O algoritmo proposto mostrou-se superior à rede neural artificial sem regularização, máquina de vetor suporte e regressão logística com base na capacidade preditiva calculada pela área abaixo da curva ROC (Yang et al., 2023). O efeito das regularizações parece estar mais elucidado em redes neurais convolucionais do que no modelo *Perceptron* de Múltiplas Camadas (PMC) pelo maior número de estudos já publicados com esse modelo (Mehdi et al., 2023; Kumar et al., 2021; Zhai et al., 2019).

Apesar do crescente uso das redes neurais em SG de soja (Silva Júnior et al., 2024; Barbosa et al., 2023; Paixão et al., 2023), as redes regularizadas são ainda muito pouco exploradas neste contexto. Os métodos de penalização guardam características interessantes para SG como a esparsidade do *Lasso*, pelo fato de forçar que os pesos se tornem zeros, e a robustez à presença de *outliers* do *Ridge*, devido maior sensibilidade aos *outliers* por ser uma penalização quadrática (Mehdi et al., 2023). Essas características podem ser interessantes para lidar com o crescente volume de dados, tanto fenotípicos, genotípicos e ambientais.

Assim, o presente trabalho tem como objetivo avaliar a eficiência do uso da regularização em modelos de PMC aplicados na predição genômica. Objetiva-se também, comparar o desempenho dos PMC com regularização (PMCR) com os resultados obtidos em métodos comumente utilizados em predição genômica, tais como o *Perceptron* de Múltiplas Camadas (PMC), Árvore de Decisão (AD), *Random Forest* (RF), *Bagging* (BAG), *Boosting* (BOO) e *Genomic Best Linear Unbiased Prediction* (GBLUP).

2. REFERENCIAL TEÓRICO

2.1 Seleção Genômica

A Seleção Genômica (SG) permite identificar e selecionar genótipos superiores por meio de marcadores moleculares, distribuídos ao longo do genoma, que estão em desequilíbrio de ligação com *loci* associados às características quantitativas de interesse (*Quantitative Trait Loci* - QTL) (Meuwissen et al., 2001). O desequilíbrio de ligação consiste na distribuição não aleatória dos alelos nos diferentes *loci*, ou seja, na presença de correlação entre os marcadores e QTLs (Qanbari, 2020). Os avanços na SG deve-se principalmente ao desenvolvimento dos marcadores do tipo SNP (*Single Nucleotide Polymorphism*) devido sua genotipagem em larga escala, sua abundância no genoma, codominância e também pelo seu barateamento (Rafalski, 2002).

A SG permite estimar os efeitos genéticos de marcadores em todo o genoma simultaneamente, combinando dados moleculares e fenotípicos em um conjunto de treinamento para obter estimativas de valores genéticos genômicos dos indivíduos em um conjunto de teste que foi apenas genotipado (Crossa et al., 2017). Assim, o pesquisador pode ajustar um modelo que, uma vez validado, permite a predição do valor genético de indivíduos sem a necessidade de conhecê-los previamente (Crossa et al., 2017).

Existem vários métodos de SG com aplicação na área do melhoramento como o GBLUP, RR-BLUP (*Ridge Regression-Best Linear Unbiased Predicition*), *Lasso* (*Least Absolute Shrinkage and Selection Operator*), BayesA, BayesB, BGLR (*Bayesian Generalized Linear Regression*) e entre outros. A escolha do método a ser utilizado está muito ligada ao tamanho do conjunto de treinamento, arquitetura genética e herdabilidade da característica, mas os métodos GBLUP, RR-BLUP, BayesA e BayesB são atualmente os principais utilizados na predição genômica (de Los Campos et al., 2012).

2.2 GBLUP

O *Genomic Best Linear Unbiased Predictor* (GBLUP), também chamado de BLUP genômico, utiliza a matriz de parentesco estimada a partir das informações de marcadores moleculares para estimar os valores genéticos de cada genótipo (Resende et al., 2012). O GBLUP assume que os efeitos dos marcadores são

aleatórios com distribuição normal e variância constante, além de cada marcador tem contribuição infinitesimal nos valores genéticos da característica. Por ser um método simples e requerer baixo custo computacional, o GBLUP tem se tornado uma abordagem muito utilizada na SG (Gao et al, 2012).

O seguinte modelo linear misto pode ser utilizado para estimar os valores genéticos aditivos no GBLUP (Resende et al., 2012):

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Wm} + \mathbf{e}, \quad (1)$$

em que $\mathbf{y}$ é o vetor de observações, $\mathbf{b}$ é o vetor de efeitos fixos, $\mathbf{m}$ é o vetor de efeitos aleatórios, onde $\mathbf{m} \sim \mathbf{N}(\mathbf{0}, \mathbf{G}\boldsymbol{\sigma}^2_{\mathbf{m}})$, em que $\mathbf{G}$ é a matriz de parentesco genômico, $\boldsymbol{\sigma}^2_{\mathbf{m}}$ é a variância genotípica aditiva, $\mathbf{e}$ é o termo de erro aleatório, $\mathbf{X}$ e $\mathbf{W}$ são as matrizes de incidência dos efeitos $\mathbf{b}$ e $\mathbf{m}$, respectivamente. A matriz G é obtida por $\mathbf{G} = \frac{\mathbf{WW^t}}{\sum_{i=1}^{n} 2p_i q_i}$, em que p_i e q_i são as frequências alélicas do i-ésimo marcador e $\mathbf{W}$ é a matriz de incidência dos marcadores SNPs (Resende et al., 2012).

O GBLUP assume que todos os *loci* explicam iguais quantidades da variação genética, ou seja, é um modelo infinitesimal com grande número de QTLs de pequeno efeito (Resende et al., 2012; de Los Campos et al., 2012). Cada *loco* tem contribuição na explicação da variância genética dada por σ^2_g/n_Q, em que σ^2_g é a variância genética estimada pelo REML a partir dos dados fenotípicos e n_Q é o número de *loci* quando cada *loco* está marcado perfeitamente por apenas um marcador (Resende et al., 2012).

2.3 Redes Neurais Artificiais

Os modelos primitivos de RNA incluem o modelo de McCulloch e Pitts (Mcculloch; Pitts, 1943), o *Perceptron* de Única Camada (Rosenblatt, 1962) e o Adaline (Widrow, 1960). Destaca-se também o trabalho de Hebb (1949) sobre a aprendizagem não supervisionada em neurônios biológicos para avanço dos estudos das redes neurais. A partir desses modelos primitivos surge uma enorme gama de modelos de redes neurais para tratar problemas nos mais diferentes níveis de complexidade e finalidade.

O modelo *Perceptron* de Múltiplas Camadas (PMC) (Rumelhart et al., 1986) é capaz de resolver problemas linearmente separáveis ou não, fundamentada no algoritmo de treinamento *backpropagation* e com alto grau de conectividade entre

seus elementos processadores (Santos et al., 2019; Carmo et al., 2019). O ajuste de uma PMC consiste nas etapas de treinamento e validação.

Na etapa de treinamento, é utilizada uma função de custo que avalia a diferença entre os valores preditos e os valores observados (Silva et al., 2016). Geralmente, a rede selecionada é aquela que minimiza essa diferença, que também pode ser chamada de função de custo. Essa função de custo pode contar com a adição de um termo de penalização para, por exemplo, penalizar pesos maiores, evitando ajustes excessivos (Mehdi et al, 2023). A etapa de validação é importante para a escolha da arquitetura da PMC, já que é possível obter redes com diferentes números de neurônios por camada ou diferentes números de camadas. Assim, a validação possibilita a obtenção de medidas de qualidade do ajuste e inferir sobre a melhor arquitetura de rede para a resolução do problema proposto (Silva et al., 2016; Costa et al., 2021).

No que diz respeito ao tipo de conexão entre os neurônios, existem redes do tipo *feedforward* e *feedback*. Nas redes do tipo *feedforward*, o fluxo de informação ocorre da entrada para a saída (Braga et al., 2007). Já nas redes do tipo *feedback*, existe também uma retroalimentação de informações de saída para a entrada da rede. As redes PMCs são do tipo *feedforward*, inexistindo qualquer tipo de retroalimentação de valores pela camada de saída ou intermediárias (Silva et al., 2016).

As Redes Neurais de Função de Base Radial (RBF) são parecidas com as PMCs, mas com a diferença do emprego de uma função de ativação em que a imagem cresce ou decresce em relação a um ponto central (Sant' Anna et al., 2021). Funções gaussianas são comumente empregadas como função de ativação em RBFs (Braga et al., 2007). As Redes Neurais Convolucionais (CNN) formam um grupo de redes com alto custo computacional utilizadas em problemas com informações visuais, são redes organizadas em camadas convolucionais, camadas de *pooling* e camadas totalmente conectadas (Mehdi et al., 2023; Kumar et al., 2021; Zhai et al., 2019). As Redes Neurais Recorrentes (RNA-R) são caracterizadas pela presença de conexões neurais do tipo *feedback* e *feedforward*, o que confere à rede a capacidade de adquirir memória adequada para lidar com dados temporais ou sequenciais. As RNA-Rs mais comumente utilizadas são o modelo de *Hopfield*, máquina de Boltzmann e *Long Short-Term Memory* (LSTM) (Joya et al., 2002; Ackley et al., 1985; Tian et al., 2018). As Redes Neurais Multimodais (RNA-MM) são um conjunto de redes que são treinadas

coletivamente com diferentes tipos de dados, como dados genômicos e climáticos, mas com o objetivo de capturar e modelar uma mesma variável resposta (Khan et al., 2023). Outros modelos de redes neurais comumente vistas na literatura são *Generative Neural Network* (Wang et al., 2019), *Encoder-Decoder Network* (Kroner et al., 2020), *Transformers* (Moutik et al., 2023) e entre outros.

2.4 Modelo *Perceptron* de Múltiplas Camadas

O PMC pode ter uma ou mais camadas intermediárias, os neurônios das camadas intermediárias e da saída tem funções semelhantes, o valor de entrada da função de ativação é o produto interno da entrada de vetores e pesos, e separa padrões nos valores de entrada com hiperplanos (Figura 1) (Cruz; Nascimento, 2018). Problemas linearmente separáveis são resolvidos com RNA de uma única camada com função de ativação linear, enquanto uma PMC com função de ativação não linear com duas camadas pode ser utilizada em problemas linearmente separáveis ou não, com alto grau de complexidade (Cruz; Nascimento, 2018).

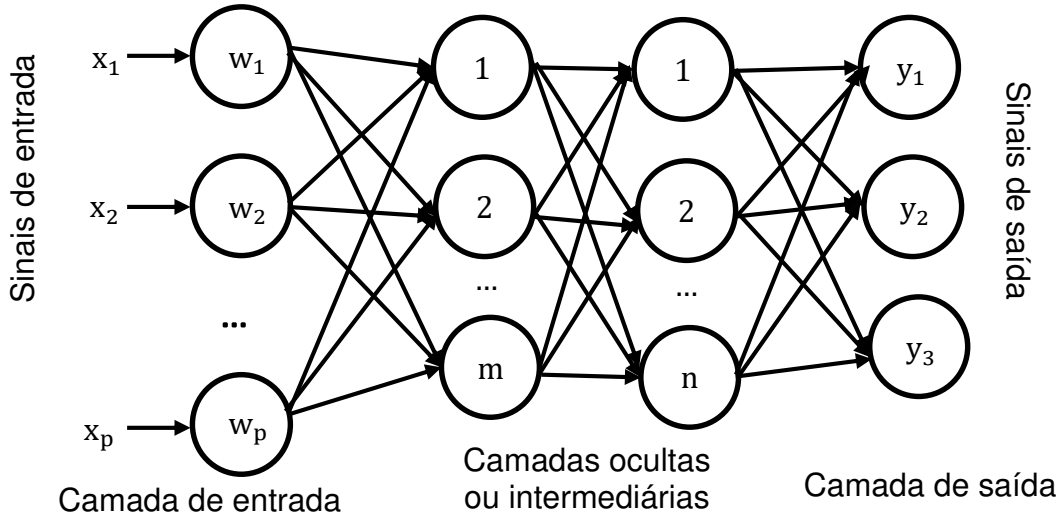

Figura 1. Uma representação esquemática de uma PMC com p variáveis na camada de entrada, duas camadas ocultas com m e n neurônios, respectivamente, na primeira e segunda camadas, e 3 neurônios na camada de saída.

A arquitetura de uma PMC é caracterizada pelo número de camadas, pelas conexões entre as camadas, pela quantidade de neurônios em cada camada e pelo tipo de conexão entre os neurônios (Figura 1). Um PMC pode ser dividido em camadas de entrada, oculta ou intermediária e de saída. A camada de entrada recebe as

informações provenientes dos dados, enquanto o processamento dessas informações ocorre principalmente na camada oculta ou intermediária, a partir de neurônios que extraem características do conjunto de dados (Cruz; Nascimento, 2018). As camadas ocultas ou intermediárias recebem os sinais de saída da camada anterior e, a resposta final é gerada na última camada.

É possível ajustar redes com diferentes números de camadas e diferentes números de neurônios em cada camada (Cruz; Nascimento, 2018). A escolha da arquitetura de rede com maior capacidade preditiva pode ser feita empiricamente, porém, a forma mais utilizada para dimensionamento da rede é por tentativa e erro (Cruz; Nascimento, 2018; Costa et al., 2021). Cada neurônio da camada de saída representa um possível resultado do processo a ser modelado. Em problemas de predição, o número de neurônios da última camada pode ser igual ou menor que o número de neurônios da primeira camada (Braga et al., 2007).

A primeira camada oculta em um PMC gera valores através da expressão (Gianola et al., 2011):

$$t_i^{(s)} = b + g\left(\sum_{j=1}^{n} w_k f_k(w_k' p_{ij})\right) + e_i, \tag{2}$$

em que p_{ij} é o valor da i-ésima observação $(i = 1,2,\ldots,n)$ na j-ésima variável explicativa $(j = 1,2,\ldots,n)$, w_k' e w_k e são os pesos do k-ésimo neurônio $(k = 1,2,\ldots,S)$ na j-ésima variável explicativa da camada de entrada e da primeira camada oculta, respectivamente, $f(.)$ e $g(.)$ são funções de ativação da camada de entrada e da primeira camada oculta, respectivamente, b é um intercepto geral e e_i é o termo de erro aleatório (Gianola et al., 2011). Os valores $t_i^{(s)}$ são utilizados na segunda camada para ajuste dos $y_i^{(s)}$, valores que seguem como co-variáveis para a camada de saída. Segundo Gianola et al. (2011), o PMC com duas camadas pode ser representado algebricamente pela equação:

$$t_i = b + g\left[\sum_{k=1}^{s} w_k g_k\left(b_k + \sum_{j=1}^{n} a_{ij} u_j^{**[k]}\right)\right] + e_i, \tag{3}$$

em que a_{ij} é o valor da i-ésima observação $(i = 1,2,\ldots,n)$ na j-ésima variável explicativa $(j = 1,2,\ldots,n)$, $u_j^{**[k]}$ é o valor do peso do k-ésimo neurônio $(k = 1,2,\ldots,S)$ na j-ésima variável explicativa na segunda camada oculta, b_k é o intercepeto do k-

ésimo neurônio, g_k é a função de ativação do k-ésimo neurônio, w_k é o peso associado à segunda camada incluindo o intercepto do k-ésimo neurônio, $g(.)$ é uma segunda função de ativação, b é um intercepto geral e e_i é o termo de erro aleatório (Gianola et al., 2011).

A função de ativação f_s pode ser linear ou não linear, mas deve ser necessariamente monotonicamente crescente. Algumas funções de ativação lineares comumente utilizadas são a função linear, *Rectified Linear Unit* (ReLU) e *Leaky ReLU* (Tabela 1) e entre outras (Montesinos-López et al., 2021; Negus et al., 2024). Funções de ativação não lineares, como a tangente hiperbólica e exponencial, conferem maior flexibilidade ao PMC e são as mais utilizadas em estudos de predição (Tabela 1) (Montesinos-López et al., 2021; Negus et al., 2024). A função *softmax* é muito utilizada em estudos de classificação (Tabela 1) (Montesinos-López et al., 2021; Negus et al., 2024). O intervalo da imagem da função de ativação deve corresponder ao intervalo de valores da variável em que se deseja modelar (Cruz; Nascimento, 2018).

A aproximação de uma função contínua pode ser alcançada com o uso de uma única camada oculta, enquanto duas camadas ocultas são suficientes para aproximar qualquer função matemática complexa (Cybenko, 1988). Três ou mais camadas ocultas são empregadas apenas em problemas complexos como, série temporal e visão computacional, e são chamadas de redes neurais profundas com maior custo computacional (Emmert-Streib et al., 2020; Negus et al., 2024).

Tabela 1. Funções de ativação utilizadas comumente em redes neurais em estudos de predição ou classificação.

Classificação	Nome	Função
Funções lineares	Linear	$f(x) = x$
	ReLU	$f(x) = \max(0, x)$
	Leaky ReLU	$f(x) = \begin{cases} x, \text{se } x > 0 \\ \alpha x, \text{para outros valores de } x \end{cases}$
Funções não lineares	*Softmax*	$f(x_i) = \frac{\exp(x_i)}{1 + \sum_{c=1}^{C} \exp(x_C)}, i = 1, \ldots, C$
	Tangente hiperbólica	$f(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)}$
	Exponencial	$f(x) = \exp(x)$

2.5 Métodos de regularização

Na regularização L1, baseada na regressão *Lasso*, deseja-se minimizar a seguinte função de custo modificada (Rognvaldsson, 2002):

$$\mathrm{E}(\mathbf{y}, \mathbf{x}, \mathbf{W}) = \mathrm{L}(\mathbf{y}, \mathbf{x}, \mathbf{W}) + \lambda_1 \|\mathbf{W}\|_1, \quad (4)$$

em que $\mathrm{L}(\mathbf{y}, \mathbf{x}, \mathbf{W})$ é uma função de custo qualquer, $\mathbf{y}$ é o vetor da variável resposta, $\mathbf{x}$ é o vetor da variável explicativa, $\mathbf{W}$ são os pesos associados a cada variável explicativa e λ_1 é um parâmetro de penalização (Rognvaldsson, 2002). A regularização *Lasso* reduz a complexidade do modelo ao penalizar alguns pesos, forçando-os a se tornarem zeros (Negus et al., 2024). Assim, variáveis de pouca importância são descartadas do modelo.

Já a regularização L2, baseada na regressão *Ridge*, deseja-se minimizar a seguinte função custo modificada (Rognvaldsson, 2002):

$$\mathrm{E}(\mathbf{y}, \mathbf{x}, \mathbf{W}) = \mathrm{L}(\mathbf{y}, \mathbf{x}, \mathbf{W}) + \lambda_2 \|\mathbf{W}\|_2^2, \quad (5)$$

em que λ_2 é um parâmetro de penalização (Rognvaldsson, 2002). Essa regularização atua reduzindo os valores dos pesos, mas não necessariamente os torna zeros (Negus et al., 2024).

Na maioria das vezes, a regularização é descrita como uma forma de reduzir o *overfitting* e aumentar a precisão do modelo. Contudo, benefícios adicionais podem ser vislumbrados. Redes não-regularizadas ocasionalmente ficam “presas” em mínimos locais da função de custo e gerando diferentes resultados em diferentes execuções. Já as redes regularizadas podem fornecer resultados mais facilmente replicáveis. Em geral, é preferível utilizar modelos mais complexos com regularizações do que modelos simples sem regularizações (Aggarwal, 2018).

A regularização L1 apresenta como principal vantagem a esparsidade, caracterizada por forçar que alguns pesos se tornem zeros, promovendo uma seleção de variáveis e tornando o modelo mais parcimonioso (Mehdi et al., 2023). Já a regularização L2 tem solução única, pode ser computacionalmente mais eficiente e mais estável, devido ao fato do seu termo adicional ($\lambda_2 \|\mathbf{W}\|_2^2$) ser diferenciável e contínuo, respectivamente (Mehdi et al., 2023). A penalização L1 não tem solução analítica, justificando sua menor eficiência computacional, além de também não possuir solução única (Mehdi et al., 2023). A regularização L2 é mais robusta na presença de *outliers* do que a L1, uma vez que é uma penalização quadrática, tornando-a consequentemente mais sensível aos valores discrepantes (Mehdi et al.,

2023). Os valores de λ_1 e λ_2 são informados previamente pelo pesquisador para ajuste do PMCR e de seus pesos. Aqueles valores de λ_1 e λ_2 que retornam as melhores redes em termos de capacidade preditiva é obtido por tentativa e erro dentro de um intervalo a ser considerado pelo pesquisador.

3. MATERIAIS E MÉTODOS

3.1 Dados reais

O experimento foi conduzido de setembro a novembro em 2021, instalado em delineamento de blocos ao acaso com 100 genótipos e três repetições, em que cada parcela foi constituída de uma planta cultivada em um vaso dentro de uma casa de vegetação. Os tratos culturais empregados seguiram as recomendações técnicas para a cultura da soja (Silva et al., 2022). Foram avaliadas características fenotípicas relacionadas da parte aérea e da parte radicular das plantas de forma destrutiva para cada parcela após 20 dias em estresse hídrico que foi iniciado no estágio vegetativo V2. Na parte aérea, foram mensuradas o diâmetro de hipocótilo em milímetros (DH), com auxílio de um paquímetro digital e a altura de planta (AP) em centímetros. Para avaliação das características de parte radicular, as raízes foram lavadas para remoção de materiais do substrato, e posteriormente armazenadas em uma solução de álcool 70% e armazenadas a uma temperatura de 4ºC. Foi utilizado o *software* WinRHIZOPro® (Winrhizo, 2021) para realizar a medição do comprimento total das raízes em centímetros (CR) e da área superficial projetada (AR) em centímetros quadrados. A produtividade de grãos não foi estudada pelo fato de que a avaliação da parte radicular gerou a destruição da planta.

A genotipagem dos indivíduos foi feita utilizando a plataforma iScan Illumina (Illumina, Inc., San Diego, CA, USA), por meio do chip “BARC-Soy SNP6k” na Deoxi Biotecnologia Ltda®, em Araçatuba, SP. Os 5403 marcadores do tipo SNP obtidos pela genotipagem foram submetidos ao controle de qualidade com uma filtragem baseada no critério de MAF (*Minor allele frequency*). Assim, aqueles marcadores com MAF < 0,05 foram descartados. Após esse controle de qualidade, foram obtidos 3957 marcadores SNPs que seguiram nas análises.

3.2 Análise dos dados fenotípicos

Inicialmente, foi feita uma correção dos dados coletados para o efeito experimental de bloco utilizando-se um modelo linear aleatório. O modelo estatístico utilizado foi (Resende et al., 2012):

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Xb} + \mathbf{Zg} + \boldsymbol{\varepsilon}, \quad (6)$$

em que $\mathbf{y}$ é o vetor de fenótipos observados, $\boldsymbol{\mu}$ representa o vetor de média geral, $\mathbf{b}$ representa o vetor de efeito de bloco, $\mathbf{g}$ representa o vetor de efeito de genótipos e $\boldsymbol{\varepsilon}$

representa o vetor de efeitos residuais. Onde, $\mathbf{b} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma_{\mathrm{b}}^2)$, $\mathbf{g} \sim \mathrm{N}\left(\mathbf{0}, \mathbf{I}\sigma_{\mathrm{g}}^2\right)$, $\boldsymbol{\varepsilon} \sim \mathrm{N}\left(\mathbf{0}, \mathbf{I}\sigma_{\varepsilon}^2\right)$, em que $\mathbf{I}$ representa a matriz identidade, σ_{b}^2, σ_{g}^2 e σ_{ε}^2 representam os componentes de variância de blocos, genótipos e residual, respectivamente. As matrizes $\mathbf{X}$ e $\mathbf{Z}$ são as matrizes de incidência dos efeitos aleatórios de bloco e genótipo, respectivamente. Os efeitos aleatórios genéticos foram estimados via Método da Máxima Verossimilhança Restrita (REML/BLUP, *Restricted Maximum Likelihood*/*Best Linear Unbiased Predictor*) e utilizados nas análises seguintes.

A herdabilidade (h^2) é a proporção da variância fenotípica que tem causas genéticas e é um parâmetro que influencia diretamente na capacidade preditiva da predição genômica (Dordevic et al., 2019). A h^2 foi calculada em seu sentindo amplo através da seguinte expressão (Dordevic et al., 2019; Ravelombola et al., 2020):

$$\mathrm{h}^2 = \frac{\widehat{\sigma}_{\mathrm{G}}^2}{\widehat{\sigma}_{\mathrm{G}}^2 + \frac{\widehat{\sigma}_{\mathrm{e}}^2}{\mathrm{k}}}, \tag{7}$$

em que $\widehat{\sigma}_{\mathrm{G}}^2$ é a variância genotípica estimada, $\widehat{\sigma}_{\mathrm{e}}^2$ é a variância residual estimada e k é o número de repetições (Dordevic et al., 2019; Ravelombola et al., 2020).

O coeficiente de variação amostral $[\mathrm{CV}(\%)]$ foi calculador por (Resende et al., 2014):

$$\mathrm{CV}(\%) = \frac{\mathrm{S_X}}{\overline{\mathrm{X}}} \cdot 100\%, \tag{8}$$

em que $\mathrm{S_X}$ é o desvio padrão e $\overline{\mathrm{X}}$ é a média geral da característica.

3.3 Métodos de predição genômica

Os PMCR foram ajustados com parâmetros de regularização λ_1 variando de 5 até 40 espaçados igualmente em 5 unidades, enquanto para λ_2 considerou-se valores variando entre 5 e 30, espaçados da mesma forma. Para a implementação do PMC e PMCR utilizou-se a função neuralnetwork do pacote ANN2 (Lammers, 2020) do *software* R (R Core Team, 2024).

Os resultados dos PMCR foram comparados com aqueles obtidos por redes neurais sem regularização (PMC), Árvore de Decisão (AD), *Random Forest* (RF), *Bagging* (BAG), *Boosting* (BOO) e *Genomic Best Linear Unbiased Prediction* (GBLUP). O GBLUP destaca-se como um dos métodos de predição mais utilizados em SG devido sua simplicidade, capacidade de aplicação em modelos mistos e baixa exigência de recurso computacional (Lima et al., 2019). Esse método foi

implementado usando a função mmer do pacote sommer (Covarrubias-Pazaran, 2016) do *software* R (R Core Team, 2024). AD, RF, BAG e BOO são metodologias de aprendizado de máquina com boas capacidades preditivas que não necessitam de pressuposições sobre o modelo já que os seus resultados dependem apenas do processo de aprendizado (James et al., 2013). RF e BAG foram implementados a partir da função randomForest do pacote randomForest (Liaw; Wiener, 2002) do software R (R Core Team, 2024). AD e BOO foram implementadas a partir da funções gbm e rpart, respectivamente, dos pacotes gbm (Greg et al., 2024) e rpart (Therneau et al., 2023) do *software* R (R Core Team, 2024).

Neste trabalho, foi realizado o ajuste de redes com duas camadas, variando o número de neurônios em cada camada de 1 a 20. Cada arquitetura foi repetida 10 vezes, em que cada repetição foi gerada após 500 iterações, para fins de cálculo do erro padrão da média das medidas de qualidade do ajuste. A arquitetura de rede que apresentou a melhor capacidade preditiva foi a selecionada e seguiu para a discussão neste trabalho, procedimento também utilizado ou sugerido por outros autores (Cruz; Nascimento, 2018; Costa et al., 2021).

3.4 Metodologias comparativas

Algumas das metodologias comparativas podem ser agrupadas em AD e seus refinamentos BAG e RF. A estrutura de uma AD consiste em nós internos, ramos e folhas. Entretanto, a AD sofre grande variação, de modo que duas árvores construídas a partir de conjuntos de treinamentos diferentes podem apresentar grandes diferenças em seus valores preditos (James et al., 2013). Visando contornar esse problema, o refinamento *Bagging* (BAG) consiste em obter um grande número de amostras dos dados disponíveis para treinamento, construir árvores para cada amostra e obter a média ou moda dos valores preditos (James et al., 2013). Portanto, o BAG reduz a variância apresentada pela AD ao combinar o resultado de vários modelos independentes de um mesmo conjunto de dados.

Entretanto, o BAG utiliza todas as variáveis, resultando em estruturas de árvores altamente correlacionadas (Breiman et al., 1984). Surge então o *Random Forest* (RF), uma abordagem refinada do BAG, modificando o conjunto de observações e o conjunto de variáveis que foram utilizadas na fase de treinamento de modelos independentes. O número de variáveis preditoras utilizadas em RF para

estudos de regressão é $m = p/3$, enquanto em classificação é $m = \sqrt{p}$, em que p é a quantidade total de variáveis disponíveis (Hastie et al., 2009).

No *Boosting* (BOO), as árvores são ajustadas sequencialmente, utilizando informações prévias da árvore anterior, e todas as observações e variáveis são empregadas no processo (Breiman, 1996). Nessa metodologia, é construído um grande número de modelos com o objetivo de minimizar os resíduos, tornando o processo de aprendizado mais lento e aumentando a possibilidade de ocorrer o *overfitting*. Assim, a validação cruzada deve ser utilizada para escolher a quantidade de árvores, a fim de evitar o *overfitting*.

O GBLUP é oriundo do BLUP tradicional, incorporando a utilização da matriz de parentesco genômico obtida através da informação de marcadores (Resende et al., 2012). O modelo de seleção genômica utilizado para estimar os valores genéticos foi (Resende et al., 2012):

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zu} + \boldsymbol{\varepsilon}, \quad (9)$$

em que $\mathbf{y}$ é o vetor de fenótipos observados, $\mathbf{1}$ é um vetor com todos os elementos iguais a um e de mesma dimensão de $\mathbf{y}$, μ é o vetor de média geral, $\mathbf{u}$ é o vetor de efeitos aleatórios dos marcadores, com $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G}\sigma_u^2)$, em que σ_u^2 é a variância genotípica aditiva e $\mathbf{G}$ é a matriz de covariância entre os genótipos, $\boldsymbol{\varepsilon}$ representa o vetor de efeitos residuais, com $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma_\varepsilon^2)$, em que $\mathbf{I}$ representa a matriz identidade e σ_ε^2 representa a variância residual. A matriz $\mathbf{Z}$ é a matriz de incidência do efeito aleatório de genótipo (Resende et al., 2012).

Todas as metodologias foram submetidas a uma validação através do método *hold-out*, em que o conjunto de dados foi divido em treinamento e teste contendo, respectivamente, 80% e 20% do conjunto de dados original (James et al., 2013). Assim, em cada repetição, 80 genótipos foram utilizados para treinamento dos métodos e enquanto os 20 restantes foram utilizados para teste. Cada ajuste de modelo foi repetido 10 vezes, permitindo o cálculo da média e do erro padrão das medidas de qualidade do ajuste. Para a comparação entre as metodologias considerou-se a correlação entre os valores genéticos do conjunto de preditos e os observados no conjunto de teste, assim como a raiz do erro quadrático médio (RMSE). A correlação representa a capacidade preditiva (CP) da metodologia (Barreto et al., 2024). E, a raiz do erro quadrático médio (RMSE) foi utilizada para comparar o erro de predição dos métodos. A RMSE é calculada por:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}, \tag{10}$$

em que n é o número total de observações, Y_i é o i-ésimo valor observado do conjunto de teste e $\widehat{Y}_i$ é o i-ésimo valor predito pelo método. Toda a implementação do trabalho foi pelo *software* R (R Core Team, 2024).

4. RESULTADOS E DISCUSSÃO

A altura de planta (AP, em centímetros) apresentou amplitude de 4,60 a 26,70 cm, com média de 12,43 cm e um coeficiente de variação de 21,80% (Tabela 1). O menor coeficiente de variação foi do DH (16,33%), que apresentou amplitude de 1,10 a 3,30 cm com média em 1,97 cm. O maior valor de CV foi de CR (39,06%), que apresentou amplitude de 106,32 a 2205,48 cm, com média em 796,07 cm. As características DH, AP, CR e AR apresentaram herdabilidade de 0,32; 0,50; 0,25 e 0,35; respectivamente (Tabela 1). Os valores de herdabilidade encontrados são similares aos apresentados em outros trabalhos para essas mesmas características (Wang et al, 2022; Conte et al., 2020).

Tabela 2. Mínimo, média, máximo, coeficiente de variação (CV) e herdabilidade no sentido amplo (h^2) para diâmetro de hipocótilo (DH, em milímetros), altura de planta (AP, em centímetros), comprimento total de raiz (CR, em centímetros) e área superficial projetada de raiz (AR, em centímetros quadrados).

Característica	Mínimo	Média	Máximo	CV(%)	h^2
DH	1,01	1,97	3,30	16,33	0,32
AP	4,60	12,43	26,70	21,80	0,50
CR	106,32	796,07	2205,48	39,06	0,25
AR	29,56	133,83	357,28	31,84	0,35

A Figura 2 apresenta os resultados obtidos de capacidade preditiva média (CP) para as metodologias comparativas e para os PMCR com diferentes valores de regularização. O GBLUP apresentou alto valor de CP (0,41) para AP, característica controlada por muitos QTLs, em que a suposição do método que cada marcador tem efeito infinitesimal no controle da característica faz maior sentido (de Los Campos et al., 2012). Esse resultado é esperado pois tal característica é controlada por 239 QTLs, ou seja, muito genes (soybase.org; Chen et al., 2021). As redes PMC e PMCR-L2 também apresentaram resultados simulares ao GBLUP quanto a CP. Já para as outras características (DH, AR e CR), o GBLUP apresentou valores de CP inferiores ao PMC, PMCR, RF e AD. O BAG em AR e CR, e BOO em AR e CR também apresentaram valores de CP maiores que GBLUP. Esses resultados também são corroborados pelo número de QTL que controlam tais caraterísticas. Especificamente, para a característica área superficial de raiz e comprimento total de raiz são

observados, respectivamente, 13 e 9 QTLs (soybase.org; Chen et al., 2021). Já para características relacionadas hipocótilo em soja são controlados, geralmente, por até 6 QTLs (soybase.org). Meuwissen e Goddard (2010) observaram método que não possuem como pressuposto o modelo infinitesimal, como por exemplo, o BayesB apresentam desempenho superior ao GBLUP para características com uma quantidade baixa de QTLs na predição em dados simulados de humanos. Wolc et al. (2016) também observaram que a acurácia do GBLUP foi inferior daquela obtida com BayesB e BayesC na presença de marcadores com grandes efeitos no controle da característica de peso de ovos em galinhas. No contexto de aprendizado de máquina esse pressuposto não é assumido a priori por nenhuma metodologia visto que as mesma utilizam os dados como principal informação para o seu ajuste.

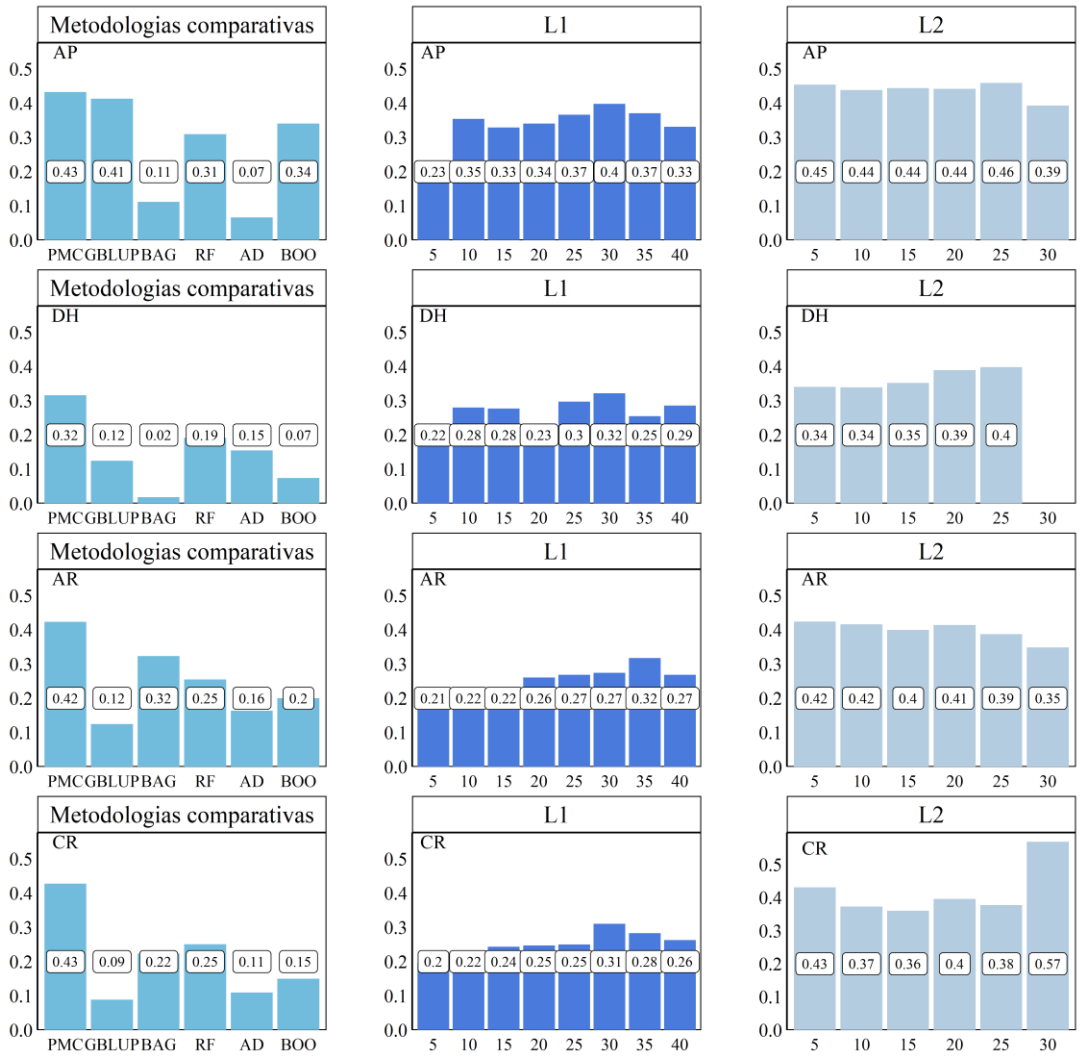

Figura 2. Representação da capacidade preditiva (CP) média na predição de características agronômicas em soja com uso das metodologias AD, BOO, BAG, RF,

PMC e GBLUP no primeiro quadro. Na primeira até a quarta linha da figura são apresentados os resultados para DH, AP, CR, e AR, respectivamente. No segundo e terceiro gráficos de cada linha são exibidos os resultados correspondentes às redes neurais regularizadas L1 e L2, respectivamente. Os diferentes parâmetros de regularização utilizados são representados no eixo das abscissas.

Dentre as metodologias comparativas, o PMC apresentaram os maiores valores para todas as características. Comparando com o segundo maior valor de CP dentre as metodologias comparativas, o PMC foi superior em 4,80%, 65,43%, 70,93% e 30,96%. Sousa et al. (2021) também observaram que PMC obtiveram o maior valor de acurácia frente aos métodos BOO, RF, BAG, AD, AD com poda e *Generalized Bayesian Lasso* na predição da resistência de *Coffea arabica* a ferrugem. Costa et al. (2021) verificou que o PMC juntamente com as redes neurais com função de base radial (RBF) obtiveram menor taxa de erro aparente frente aos métodos BAG, RF, BOO, AD, função discriminante de Fisher e função discriminante de Anderson na classificação de ambientes com base em várias características agronômicas de *Coffea arabica*.

Outras metodologias comparativas que se destacaram foi RF que superou o GBLUP nos valores de CP para as características DH, AR e CR em 53,22%, 183,51% e 104,27%, respectivamente. O BAG obteve o segundo e terceiro maior valor de CP para AR e CR, respectivamente, superando o GBLUP em 153,06% e 159,39, respectivamente. Silva Júnior et al. (2024) observaram resultados similares ou superiores do desempenho de RF frente aos métodos de PMC, BAG, BOO e RBF na predição de produtividade de grãos em soja.

Dentre as redes com regularização, a L1 apresentou incrementos na CP com o aumento do valor de λ_1, até, em geral, atingir o valor máximo de CP em $\lambda_1 = 30$ (ou $\lambda_1 = 35$ para a característica AR). Apesar desses incrementos, a regularização L1 não alcançou os resultados de CP do PMCR-L2 para todas características e do PMC para AP, AR e CR. Apenas na característica DH que PMCR-L1 (com $\lambda_1 = 30$) apresentou CP maior que PMC ($\mathrm{CP}_{\mathrm{PMC}} = 0{,}3162$ e $\mathrm{CP}_{\mathrm{PMCR-L1}} = 0{,}3215$). Considerando $\lambda_1 = 30$, a CP do PMCR-L1 foi inferior aos resultados obtidos com PMC em 7,97%, -1,64%, 27,50% e 24,98% para AP, DH, CR e AP, respectivamente (Figura 2). Portanto, o uso da penalização L1 não trouxe melhorias de CP para as características avaliadas neste

trabalho. Entretanto, os PMCR-L1 obtiveram valores maiores que AD para todas as características e valores similares ou superiores para os refinamentos BAG e RF (Figura 2).

O melhor desempenho em CP do PMCR-L1 comparado ao PMC foi em DH, característica controlada por poucos QTLs, apesar que AR e CR também são controladas por poucos QTLs e o PMCR-L1 apresentou valores bem abaixo de CP comparado ao PMC. Nessa mesma comparação, o segundo melhor desempenho em CP do PMCR-L1 foi em AP, característica controlada por muitos QLTs.

Na regularização L2 foi obtido o maior valor de CP para todas as características. Independentemente do valor de λ_2, verificou-se que o PMCR-L2 apresentou valor de CP superior em relação ao obtido pelo AD e seus refinamentos para todas as características avaliadas. Os valores de CP do PMCR-L2 foram superiores em relação ao PMC em 6,05%; 25,86%; 32,90% e 0,16% para as características AP, DH, CR e AR, respectivamente. O bom desempenho da PMCR-L2 pode ser pelo fato de que todos os marcadores moleculares são utilizados na predição da característica, uma vez que a regressão *Ridge* reduz os pesos ajustados, mas não os forçam a se tornarem zeros (Negus et al., 2024).

Não houve um padrão de CP ao comparar os diferentes valores de λ_2 nas características avaliadas. Em AP, característica controlada por muitos QTLs, foi observado valores de CP bem próximos para os diferentes valores de λ_2 (Figura 2). Já para DH e AR, características controladas por poucos genes, foram observados valores crescentes e decrescentes de CP ao aumentar λ_2, respectivamente (Figura 2). Finalmente, em CR foi observado uma falta de padrão e um valor discrepante de CP para $\lambda_2 = 30$ (Figura 2). Para $\lambda_2 \geq 35$ foi observado que as redes retornavam o mesmo único valor predito para os diferentes genótipos, retornando uma correlação NA pelo fato da inexistência de variância entre os valores preditos. Esse fato também ocorreu para $\lambda_2 = 30$ no caso de DH (Figura 2). Esse fato pode ser explicado por uma forte penalização para esses valores de λ_2 e a rede retornou apenas o seu valor de intercepto naquela característica.

Em uma análise mais geral, a penalização L2 apresentou valores de CP superiores aos apresentados pela L1 em todas as características. Considerando o maior valor de CP obtido com L1 e com L2 em cada característica, independente do valor de λ, a superioridade da L2 foi de 15,24%; 23,79%; 83,32% e 33,52% para AP,

DH, CR e AP, respectivamente (Figura 2). Os valores inferiores de CP observados na regularização L1 comparado com PMC ou PMCR-L2 pode ser justificada pela grande redução do número de marcadores proporcionada pelo *Lasso* (Hastie et al., 2009). O *Lasso* admite até n pesos diferentes de zero, o que pode ser uma restrição muito forte na predição genômica, uma vez que na maioria das vezes são utilizadas informações de muitos marcadores ($p \gg n$) (de Los Campos et al., 2012). Assim, a esparsidade do *Lasso* não foi uma vantagem na predição genômica das características avaliadas neste trabalho.

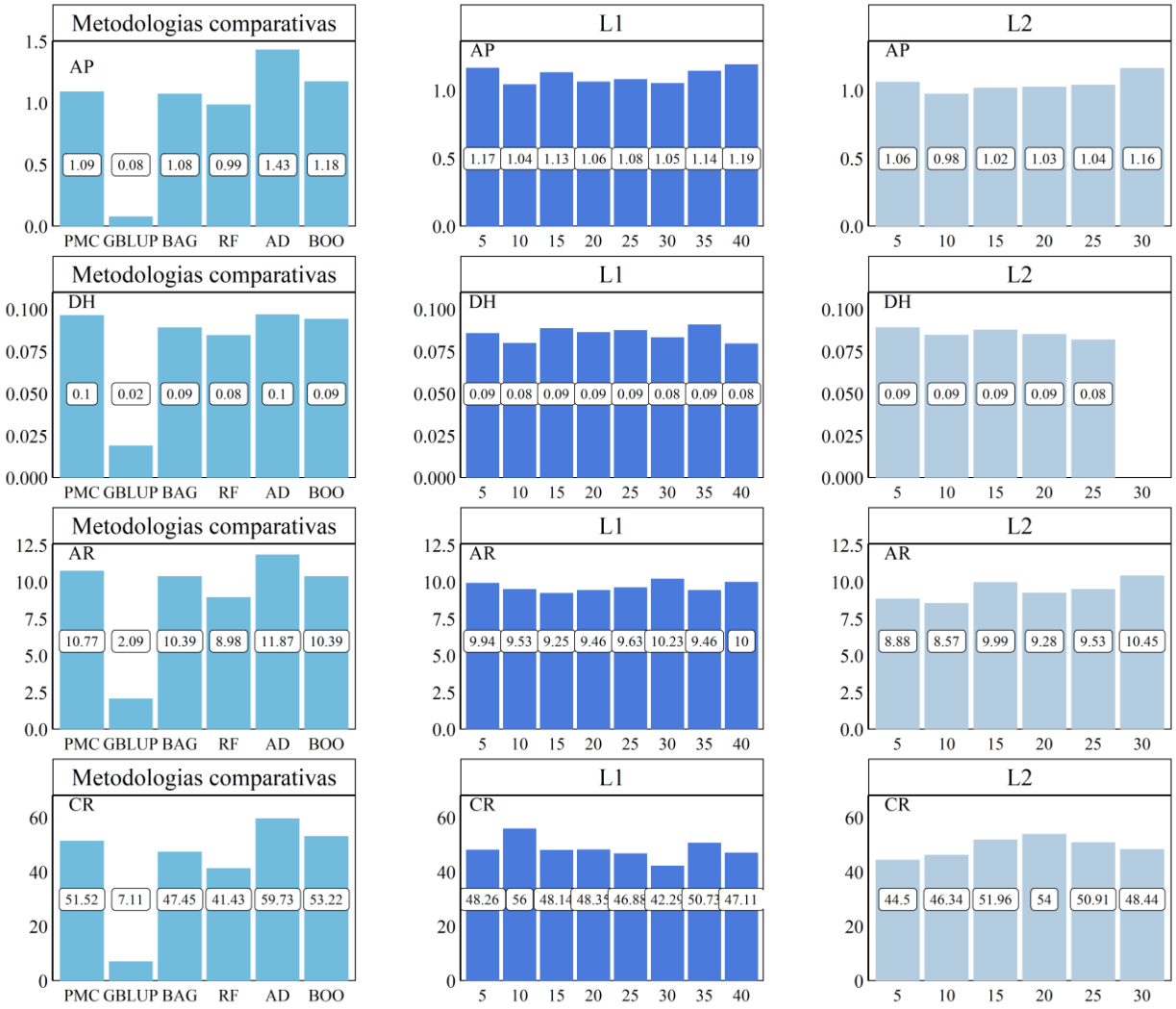

Figura 3. Representação da raiz do erro quadrático (RMSE) médio na predição de características agronômicas em soja com uso das metodologias AD, BOO, BAG, RF, PMC e GBLUP no primeiro quadro. Na primeira até a quarta linha da figura são apresentados os resultados para DH, AP, CR, e AR, respectivamente. No segundo e

terceiro gráficos de cada linha são exibidos os resultados correspondentes às redes neurais regularizadas L1 e L2, respectivamente. Os diferentes parâmetros de regularização utilizados são representados no eixo das abscissas.

Em relação ao RMSE, uma métrica que avalia o erro de predição, a AD e o GBLUP apresentaram o maior e menor valor de RMSE, respectivamente, para todas as características (Figura 3). O PMCR-L2 apresentou resultados inferiores e desejáveis em 10,81%, 15,00%, 13,62% e 20,41% para as características AP, DH, CR e AR, respectivamente, quando comparado com as redes sem regularização. Quando a comparação é entre todas as metodologias comparativas, o GBLUP obteve o menor valor de RMSE para todas características avaliadas. A variável CR apresentou uma amplitude em RMSE de 52,62 centímetros, a maior amplitude de RMSE considerando todas as características. O pequeno valor de RMSE do GBLUP e a similaridade entre todos os outros métodos avaliados com valores maiores de RMSE, revela um elevada e discrepante menor erro de predição do GBLUP. A maioria das redes regularizadas apresentaram menores valores de RMSE comparado às redes sem regularização. Nos PMCR- L2 foram observados valores de RMSE ligeiramente inferiores à L1 para a maioria dos valores de penalização avaliados (Figura 3).

Tabela 3. Arquitetura da rede selecionada com o maior valor de CP para predição de diâmetro de hipocótilo (DH, em milímetros), altura de planta (AP, em centímetros), comprimento total de raiz (CR, em centímetros) e área superficial projetada de raiz (AR, em centímetros quadrados). Foram avaliadas todas as redes com duas camadas variando de 1 a 20 neurônios por camada.

	Característica							
	DH		AP		CR		AR	
	Camada							
Tipo de rede	1°	2°	1°	2°	1°	2°	1°	2°
Sem regularização	5	4	15	20	10	6	11	17
$\lambda_1 = 5$	3	3	16	6	15	6	11	8
$\lambda_1 = 10$	17	7	8	16	18	14	12	9
$\lambda_1 = 15$	9	20	13	13	4	18	18	4

$\lambda_1 = 20$	10	6	19	18	13	3	19	14
$\lambda_1 = 25$	16	14	18	12	8	20	18	20
$\lambda_1 = 30$	19	14	20	20	11	15	17	20
$\lambda_1 = 35$	17	13	15	14	5	11	16	16
$\lambda_1 = 40$	13	11	17	19	13	8	18	14
$\lambda_2 = 5$	4	13	6	6	8	20	1	7
$\lambda_2 = 10$	13	18	2	3	16	20	19	14
$\lambda_2 = 15$	13	15	18	19	10	10	13	16
$\lambda_2 = 20$	2	14	5	17	1	15	18	2
$\lambda_2 = 25$	7	14	5	7	4	3	9	3
$\lambda_2 = 30$	NA	NA	19	11	18	9	17	17

As redes selecionadas com regularização L1 ou L2 apresentam similaridade quanto à complexidade de sua arquitetura e nenhuma diferença discrepante das redes selecionadas sem regularização (Tabela 3). Dentre todas os PMCR-L1 selecionados, 14 das 32 ou 43,75% das redes apresentaram menos de 10 neurônios em pelo menos uma das camadas ocultas, valor similar ao observado para os PMCR-L2 (14 das 31 ou 45,61%). Portanto, os PMCR selecionados apresentam, em sua grande maioria, certa complexidade em sua arquitetura. A variação do valor do parâmetro de regularização não promoveu nenhuma variação na arquitetura da rede selecionada com maior valor de CP (Tabela 3).

5. CONCLUSÕES

Entre as redes regularizadas, aquelas do tipo L2 apresentaram os melhores valores de CP para todas as características avaliadas. Conclui-se que as redes com regularização L2 apresentaram resultados similares ou superiores de CP comparado às redes sem regularização e demais métodos avaliados para todas as variáveis. Entre as características avaliadas, não houve um padrão entre os valores de λ_2 e CP. Já para PMCR-L1, valores de $\lambda_1 = 30$ para AP, DH e CR retornaram os valores máximos de CP, enquanto que $\lambda_1 = 35$ retornou o valor máximo de CP para AR. Portanto, o comportamento de λ parece ser mais previsível para PMCR-L1. Os métodos de regularização apresentaram, em geral, baixos valores de RMSE, menores que os apresentados por Árvore de Decisão e alguns dos seus refinamentos, além das redes sem regularização em algumas das características.

REFERÊNCIAS

ACKLEY, D.H.; HINTON, G.E.; SEJNOWSKI, T.J. A learning algorithm for Boltzmann machines. **Cognitive Science**, v. 9, n. 1, p. 147-169, 1985. DOI: 10.1016/S0364-0213(85)80012-4.

AGGARWAL, C.C. **Neural networks and deep learning: a textbook**. 1a edição, Springer, New York, 2018. DOI: 10.1007/978-3-319-94463-0_1.

BARBOSA, W.F.; HASHIMOTO, T.P.; MATSUO, E.; NASCIMENTO, A.C.C.; AZEVEDO, C.F.; BEZERRA, A.R.G.; NASCIMENTO, M. Artificial neural networks based on segmented model for adaptability and stability evaluation of soybean genotypes. **Australian Journal of Crop Science**, v. 17, n.9, 735-740, 2023. DOI: 10.21475/ajcs.23.17.09.p3986.

BARRETO, C.A.V.; DIAS, K.O.G.; SOUSA, I.C.; AZEVEDO, C.F.; NASCIMENTO, A.C.C.; GUIMARÃES, L.J.M.; GUIMARÃES, C.T.; PASTINA, M.M.; NASCIMENTO, M. Genomic prediction in multi-environment trials in maize using statistical and machine learning methods. **Scientific Reports**, v. 14, n. 1, 2024. DOI: 10.5061/dryad.ps22r.

BRAGA, A.P.; CARBALHO, A.P.L.F.; LUDERMIR, T.B. **Redes Neurais Artificiais – Teoria e aplicações**. 2a edição, LTV, Rio de Janeiro, 2007.

BREIMAN, L. Heuristics of instability and stabilization in model selection. **The annals of statistics**, v. 24, n. 2, p. 2350-2383, 1996. DOI: 10.1214/aos/1032181158.

BREIMAN, L.; FRIEDMAN, J.H.; OLSHEN, R.A.; STONE, C.J. Classification and Regression Trees. **Wadsworth Int. Group, Belmont**, California, USA, 1984. DOI: 10.1201/9781315139470.

CARMO, D.G.; FARIAS, E.S.; COSTA, T.L.; QUEIROZ, E.A.; NASCIMENTO, M.; PICANCO, M.C. Instar Determination of *Blaptostethus pallescens* (Hemiptera: Anthocoridade) Using Artificial Neural Networks. **Annals of the Entomological Society of America**, v. 1, p. 1-5, 2019. DOI: 10.1093/aesa/saz059.

CHEN, H.; KUMAWAT, G.; YAN, Y.; FAN, B.; XU, D. Mapping and validation of a major QTL for primary root length of soybean seedlings grown in hydroponic conditions. **BMC Genetics**, v. 22, n. 132, 2021. DOI: 10.1186/s12864-021-07445-0.

CONTE, M.V.D.; CARNEIRO, P.C.S.; RESENDE, M.D.V.; SILVA, F.L.; PETERNELLI, L.A. Overcoming collinearity in path analysis of soybean [*Glycine max* (L.) Merr.] grain oil content. PLoS ONE, v. 15, n. 5, 2020. DOI: 10.1371/journal.pone.0233290.

COSTA, W.G.; BARBOSA, I.P.; SOUZA, J.E.; CRUZ, C.D.; NASCIMENTO, M.; OLIVEIRA, A.C.B. Machine learning and statistics to qualify environments through multi-traits in *Coffea arabica*. **PLoS One**, v. 16, 2021. DOI: 10.1371/journal.pone.0245298.

COVARRUBIAS-PAZARAN, G. Genome assisted prediction of quantitative traits using the R package sommer. **PLoS ONE**, v. 11, p. 1-15, 2016. DOI: 10.1371/journal.pone.0156744.

CROSSA, J.; PÉREZ-RODRÍGUEZ, P.; CUEVAS, J.; MONTESINOS-LÓPEZ, O.; JARQUÍN, D.; DE LOS CAMPOS, G.; BURGUEÑO, J.; GONZÁLEZ-CAMACHO, J.M.; PÉREZ-ELIZALDE, S.; BEYENE, Y.; DREISIGACKER, S.; SINGH, R.; ZHANG, X.; GOWDA, M.; ROORKIWAL, M.; RUTKOSKI, J.; VARSHNEY, R.K. Genomic Selection in Plant Breeding: Methods, Model, and Perspectives. **Trends in Plant Science**, v. 22, p. 961-975, 2017. DOI: 10.1016/j.tplants.2017.08.011.

CRUZ, C.D.; NASCIMENTO, M. **Inteligência computacional aplicada ao melhoramento genético**. 1a edição, Editora UFV, Viçosa, 2018.

CYBENKO, G. **Continuous Valued Neural Networks with Two Hidden Layers Are Suficient**. Technical Report. Tufts University, Medford, 1988.

DE LOS CAMPOS, G.; HICKEY, J.M.; DAETWYLER, H.D.; PONG-WONG, R.; CALUS, M.P.L. Whole Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. **Genetics**, v. 193, p. 327-345, 2012. DOI: 10.1534/genetics.112.143313.

DORDEVIC, V.; CERAN, M.; MILADINOVIC, J.; BALESEVIC-TUBIC, S.; PETROVIC, K.; MILADINOV, Z.; MARINKOVIC, J. Exploring the performance of genomic prediction models for soybean yield using different validation approaches. **Molecular Breeding**, v. 39, n. 74, 2019. DOI: 10.1007/s11032-019-0983-6.

EMMERT-STREIB, F.; YANG, Z.; FENG, H.; TRIPATHI, S.; DEHMER, M. An introductory review of deep learning for prediction models with big data. **Frontiers in Artificial Intelligence**, v. 3, p. 1-23, 2020. DOI: 10.3389/frai.2020.00004.

FUKASE, E.; MARTIN, W. Economic growth, convergence, and world food demand and supply. **World Development**, v. 132, p. 1-12, 2020. DOI: 10.1016/j.worlddev.2020.104954.

GAO, H.; CHRISTENSEN, O.F.; MADSEN, P.; NIELSEN, U.S.; ZHANG, Y.; LUND, M.S.; SU, G. Comparison on genomic predictions using three GBLUP methods and two single-step blending methods in the Nordic Holstein population. **Genetics Selection Evolution**, v. 44, n. 8, p. 2-8, 2012. DOI: 10.1186/1297-9686-44-8.

GIANOLA, D.; OKUT, H.; WEIGEL, K.A.; ROSA, G.J.M. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. **BMC Genetics**, v. 12, n. 87, 2011. DOI: 10.1186/1471-2156-12-87.

GREG, R.; EDWARS, D.; KRIEGLER, B.; SCHROEDL, S.; SOUTHWORTH, H.; GREENWELL, B.; BOEHMKE, B.; CUNNINGHAM, J. **gbm: Generalized Boosted Regression Models**. 2024, pacote R disponível em: https://CRAN.R-project.org/package=gbm. Acesso em 07 de abril de 2024.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference and prediction**. 2a edição, Springer, New York, 2009. DOI: 10.1007/978-0-387-84858-7.

HEBB, D.O. **The Organization of Behavior**. John Wiley & Sons, New York, 1949, p. 365.

HOERL, A.; KENNARD, R. Ridge regression: Biased Estimation for Nonorthogonal Problems. **Technometrics**, v. 12, n. 1, p. 55-67, 1970. DOI: 10.2307/1267351.

JAMES, G. W.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning with Application in R**. 1a edição, Springer, New York, 2013. DOI: 10.1007/978-1-0716-1418-1. DOI: 10.1007/978-1-0716-1418-1.

JOYA, G.; ATENCIA, M.A.; SANDOVAL, F. Hopfield neural networks for optimization: study of the different dynamics. **Neurocomputing**, v. 43, p. 219-237, 2002. DOI: 10.1016/S0925-2312(01)00337-X.

KRONER, A.; SENDEN, M.; DRIESSENS, K.; GOEBEL, R. Contextual encoder–decoder network for visual saliency prediction. **Neural Networks**, v. 129, p. 261-270, 2020. DOI: 10.1016/j.neunet.2020.05.004.

KHAN, R.A.; FUA, M.; BURBRIDGE, B.; LUO, Y.; WU, F.X. A multi-modal deep neural network for multi-class liver cancer diagnosis. **Neural Networks**, v. 165, p. 553-561, 2023. DOI: 10.1016/j.neunet.2023.06.013.

KUMAR, A.; SHAIKH, A.M.; LI, Y.; BILAL, H.; YIN, B. Pruning filters with L1-norm and capped L1-norm for CNN compression. **Applied Intelligence**, v. 51, p.1152-1160, 2021. DOI: 10.1007/s10489-020-01894-y.

LAMMERS, B. **ANN2: Artificial Neural Networks for Anomaly Detection**. 2020, pacote R disponível em: https://CRAN.R-project.org/package=ANN2. Acesso em 29 de março de 2024.

LECUN, Y.; CORTES, C.; BURGES, C.J. **MNIST handwritten digit database**, 2010. Disponível em: http://yann.lecun.com/exdb/mnist/.

LIAW, A.; WIENER, M. Classification and Regression by randomForest. **R News**, v. 2, n.3, p. 18-22, 2002.

LIMA, L.P.; AZEVEDO, C.F.; RESENDE, M.D.V.; VIANA, J.M.S.; OLIVEIRA, E.J. Triple categorical regression for genomic selection: application to cassava breeding. **Scientia Agricola**, v. 76, p. 368-375, 2019. DOI: 10.1590/1678-992X-2017-0369.

MATEI, G.; WOYANN, L.G.; MILIOLI, A.S.; OLIVEIRA, I.B.; ZDZIARSKI, A.D.; ZANELLA, R.; COELHO, A.S.G.; FINATTO, T.; BENIN, G. Genomic selection in soybean: accuracy and time gain in relation to phenotupic selection. **Molecular Breeding**, v. 38, n. 117, 2018. DOI: 10.1007/s11032-018-0872-4.

MCCULLOCH, W.S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **Bulletin of Mathematical Biphysics**, v. 157, p. 115-133, 1943. DOI: 10.1007/BF02478259.

MEHDI, C.A.; NOUR-EDDINE, J.; MOHAMED, E. Regularization in CNN: A Mathematical Study for L_1, L_2 and Dropout Regularizers. In: **Internation Conference on Advanced Intelligent Systems for Sustainable Development**. Springer, p. 442-450, 2023. DOI: 10.1007/978-3-031-26384-2.

MEUWISSEN, T.H.E.; HAYES, B.J.; GODDARD, M.E. Prediction of total genetic value using genome wide dense marker maps. **Genetics**, v. 157, p. 1819-1829, 2001. DOI: 10.1093/genetics/157.4.1819.

MEUWISSEN, T.; GODDARD, M. Accurate Prediction of Genetic Values for Complex Traits by Whole-Genome Resequencing. **Genetics**, v. 185, p. 623-631, 2010. DOI: 10.1534/genetics.110.116590.

MONTESINOS-LÓPEZ, O.A.; MONTESINOS-LÓPEZ, A.; PÉREZ-RODRÍGUEZ, P.; BARRÓN-LÓPEZ, J.A.; MARTINI, J.W.R.; FAJARDO-FLORES, S.B.; GAYTAN-LUGO, L.S.; SANTANA-MANCILLA, P.C.; CROSSA, J. A review of deep learning applications for genomic selection. **BMC Genomics**, v. 22, n. 19, 2021. DOI: 10.1186/s12864-020-07319-x.

MOUTIK, O.; SEKKAT, H.; TIGANI, S.; CHEHRI, A.; SAADANE, R.; TCHAKOUCHT, T.A.; PAUL, A. Convolutional Neural Networks or Vision Transformers: Who Will Win the Race for Action Recognitions in Visual Data? **Sensors**, v. 23, n. 2, 2023. DOI: 10.3390/s23020734.

NEGUS, K.L.; LI, X.; WELCH, S.M.; YU, J. The role of artificial intelligence in crop improvement. In: Advances in Agronomy. **Academic Press**, v. 184, p. 1-66, 2024. DOI: 10.1016/bs.agron.2023.11.001.

NUSRAT, I.; JANG, S.B. A comparison of regularization techniques in deep neural networks. **Symmetry**, v. 10, p. 1-18, 2018. DOI: 10.3390/sym10110648.

PAIXÃO, J.V.C.C.; MATSUO, E.; SOUSA, I.C.; NASCIMENTO, M.; OLIVEIRA, I.S.; MACEDO, A.F.; SANTANA, G.M. Classification of soybean cultivars by means of artificial neural networks. **Agronomy Science and Biotechnology**, v. 9, p. 1-11, 2023. DOI: 10.33158/ASB.r186.v9.2023.

PÉREZ-ENCISO, M.; ZINGARETTI, L.M. A Guide for Using Deep Learning for Complex Trait Genomic Prediction. **Genes**, v. 10, 2019. DOI: 10.3390/genes10070553.

QANBARI, S. On the Extent of Linkage Disequilibrium in the Genome of Farm Animals. **Frontiers in genetics**, v. 10, 2020. DOI: 10.3389/fgene.2019.01304.

RAFALSKI, A. Applications of single nucleotide polymorphisms in crop genetics. **Current opinion in plant biology**, v. 5, p. 94-100, 2002. DOI: 10.1016/s1369-5266(02)00240-6.

RAVELOMBOLA, W.S.; QIN, J.; SHI, A.; NICE, L.; BAO, Y.; LORENZ, A.; ORF, J.H.; YOUNG, N.D.; CHEN, S. Genome-wide association study and genomic selection for tolerance of soybean biomass to soybean cyst nematode infestation. **PLoS ONE**, v. 15, n. 7, 2020. DOI: 10.1371/journal.pone.0235089.

R CORE TEAM. **R: A language and environment for statistical computing**. Vienna: R Foundation for Statistical Computing, 2024. Disponível em: <https://www.rproject.org/>.

RESENDE, M.D.V; SILVA, F.F.; AZEVEDO, C.F. **Estatística matemática, biométrica e computacional: modelos mistos, multivariados, categóricos e generalizados (REML/BLUP), inferência bayesiana, regressão aleatória, seleção genômica, QTL-GWAS, estatística espacial e temporal, competição, sobrevivência**. 1a edição, Editora UFV, Viçosa, 2014.

RESENDE, M.D.V.; SILVA, F.F.; LOPES, P.S.; AZEVEDO, C.F. **Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial**. Viçosa, Universidade Federal de Viçosa, p. 291, 2012.

ROGNVALDSSON, T.S. A simple trick for estimating the weight decay parameter. In: **Neural networks: Tricks of the trade**. Springer, Berlin, p. 71-92, 2002. DOI: 10.1007/978-3-642-35289-8_6.

ROSENBLATT, F. **Principles of neurodynamics: perceptrons and the theory of brain mechanisms**. 1a edição, Spartan Books, Washington, 1962.

RUMELHART, D.E.; HINTON, G.E.; WILLIAMS, R.J. Learning representations by back-propagating errors. **Nature**, v. 323, p. 533-536, 1986. DOI: 10.1038/323533a0.

SANT' ANNA, I.C.; CABRAL FERREIRA, R.A.D.; NASCIMENTO, M.; SILVA, G.N.; CARNEIRO, V.Q.; CRUZ, C.D.; OLIVEIRA, M.S.; CHAGAS, F.E. Multigenerational

prediction of genetic values using genome-enabled prediction. **PLoS One**, v. 14, 2019. DOI: 10.1371/journal.pone.0210531.

SANT' ANNA, I.C.; SILVA, G.N.; NASCIMENTO, M.; CRUZ, C.D. Subset selection of markers for the genome-enabled prediction of genetic values using radial basis function neural networks. **Acta Scientiarum-Agronomy**, v. 43, 2021. DOI: 10.4025/actasciagron.v43i1.46307.

SANTOS, I.G.; CRUZ, C.D.; NASCIMENTO, M.; FERREIRA, R.P. Selection index as a priori information for using artificial neural networks to classify alfafa genotypes. **Genetics and Molecular Research**, v. 18, p. 18221, 2019. DOI: 10.4238/gmr18221.

SHEA, Z.; SINGER, W.M.; ZHANG, B. Soybean Production, Versatility, and Improvement. **Legume Crops**, v. 1, p. 1-22, 2020. DOI: 10.5772/intechopen.91778.

SILVA, F.; BORÉM, A.; SEDIYAMA, T.; CÂMARA, G. **Soja: Do plantio à colheita**. 2a edição, Oficina de Textos, São Paulo, 2022.

SILVA, I.N.; SPATTI, D.H.; FLAUZINO, R.A. **Redes neurais artificiais para engenharia e ciências aplicadas**. 2a edição, Artiliber, São Paulo, 2016.

SILVA JÚNIOR, A.C.; COSTA, W.G.; GUIMARÃES, A.G.; MOURA, W.M.; BHERING, L.L.; CRUZ, C.D.; BATISTA, R.O.; SANTOS, J.B.; CAMPOS, W.F.; EVARISTO, A.B. Trait prediction through computational intelligence and machine learning applied to soybean (Glycine max) breeding in shaded environments. **bioRxiv**, p. 1-27, 2024. DOI: 10.1101/2024.01.31.578252.

SOUSA, I.C.; NASCIMENTO, M.; SILVA, G.N.; NASCIMENTO, A.C.C.; CRUZ, C.D.; SILVA, F.F.; ALMEIDA, D.P.; PESTANA, K.N.; AZEVEDO, C.F.; ZAMBOLIM, L.; CAIXETA, E.T. Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. **Scientia Agricola**, v. 78, p. 1-8, 2021. DOI: 10.1590/1678-992X-2020-0021.

SOYBASE. Integrating Genetics and Genomics to Advance Soybean Research. Disponível em: https://www.soybase.org/. Acesso em: 29 de março de 2024.

SRIVASTAVA, N.; HINTON, G.; KRIZHEVSKY, A.; SUTSKEVER, I.; SALAKHUTDINOV, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. **Journal of Machine Learning Research**, v. 15, p. 1-30, 2014. DOI: 10.5555/2627435.2670313.

SVOZIL, D.; KVASNICKA, V.; POSPICHAL, J. Introduction to multi-layer feed-forward neural networks. **Chemometrics and intelligent laboratory systems**, v. 39, p. 43-62, 1997. DOI: 10.1016/S0169-7439(97)00061-0.

THERNEAU, T.; ATKINSON, B. **rpart: Recursive Partitioning and Regression Trees**. 2023, pacote disponível em: https://CRAN.R-project.org/package=rpart. Acesso em 07 de abril de 2024.

TIAN, C.; MA, J.; ZHANG, P. A Deep Neural Network Model for Short-Term Load Forecast Based on Long Short-Term Memory Network and Convolutional Neural Network. **Energies**, v. 11, n. 12, 2018. DOI: 10.3390/en11123493.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Royal Statistical Society Series B**, v. 58, p. 267-288, 1996.

TORKAMANEH, D.; BOYLE, B.; BELZILE, F. Efficient genome-wide genotyping strategies and data integration in crop plants. **Theoretical and Applied Genetics**, v. 131, p. 499-511, 2018. DOI: 10.1007/s00122-018-3056-z.

VAN ERP, S.; OBERSKI, D.L.; MULDER, J. Shrinkage priors for Bayesian penalized regression. **Journal of Mathematical Psychology**, v. 89, p. 31-50, 2019. DOI: 10.1016/j.jmp.2018.12.004.

WANG, G.; LEDWOCH, A.; HASANI, R.M.; GROSU, R.; BRINTRUP, A. A generative neural network model for the quality prediction of work in progress products. **Applied Soft Computing**, v. 85, 2019. DOI: 10.1016/j.asoc.2019.105683.

WANG, Z.; HUANG, C.; NIU, Y.; YUNG, W.S.; XIAO, Z.; WONG, F.L.; HUANG, M.; WANG, X.; MAN, C.K.; SZE, C.C.; LIU, A.; WANG, W.; CHEN, Y.; LIU, S.; WU, C.; LIU, L.; HOU, W.; HAN, T.; HAN, T.; LI, M.W.; LAM, H.M. QTL analyses of soybean root system architecture revealed genetic relationships with shoot-related traits. **Theoretical and Applied Genetics**, v. 135, p. 4507-4522, 2022. DOI: 10.1007/s00122-022-04235-4.

WOLC, A.; ARANGO, J.; SETTAR, P.; FULTON, J.E.;O'SULLIVAN, N.P.; DEKKERS, J.C.M.; FERNANDO, R.; GARRICK, D.J. Mixture models detect large effect QTL better than GBLUP and result in more accurate and persistent predictions. **Journal of Animal Science and Biotechnology**, v. 7, n. 7, 2016. DOI: 10.1186/s40104-016-0066-z.

WIDROW, B. An adaptive "Adaline" neuron using chemical memistors. In: **Technical Report No. 1553-2**, Stanford Electronics Laboratories, Stanford University, 1960.

WINRHIZO. 2021. Disponível em: <https://regentinstruments.com/assets/images_winrhizo/WinRHIZO_2021.pdf>.

YANG, M.; LIM, M.K.; QU, Y.; LI, X.; NI, D. Deep neural networks with L1 and L2 regularization for high dimensional corporate credict risck prediction. **Expert Systems with Applications**, v. 213, 2023. DOI: 10.1016/j.eswa.2022.118873.

ZHAI, Y.; DENG, W.; XU, Y.; KE, Q.; GAN, J.; SUN, B.; ZENG, J.; PIURI, V. Robust SAR Automatic Target Recognition Based on Transferred MS-CNN with L^2-Regularization. **Computational Intelligence and Neuroscience**, v. 2019, p. 1-13, 2019. DOI: 10.1155/2019/9140167

ZINGARETTI, L.M.; GEZAN, S.A.; FERRÃO, L.F.V; OSORIO, L.F.; MONFORT, A.; MUNOZ, P.R.; WHITAKER, V.M.; PÉREZ-ENCISO, M. Exploring Deep Learning for Complex Trait Genomic Prediction in Polyploid Outcrossing Species. **Frontiers in Plant Science**, v. 11, p. 1-14, 2020. DOI: 10.3389/fpls.2020.00025.

APÊNDICE – SCRIPT R

```
library(c(ANN2,readxl,caret,ggplot2,dplyr,nlme,sommer))
rm(list=ls())
dat<-read_excel("veg.xlsx",sheet="veg")
dat$Bl<-as.factor(dat$Bl)
dat$Trat<-as.factor(dat$Trat)
dat$gen<-as.factor(dat$genn)
dat$year<-as.factor(dat$year)
##Filtering dataset
dat<-dat %>% filter(year==2)
dat<-dat %>% filter(Trat==2)
dat<-as.data.frame(dat)
dat<-dat[,-12]
## https://biostatmatt.com/archives/2718
dat$Dummy <- factor(1)
dat <- groupedData(sar ~ 1 | Dummy, dat)
m0<-lme(sar~1,random=pdBlocked(list(pdIdent(~0+Bl),pdIdent(~0+gen))),
      data=dat,na.action = na.omit)
m00<-as.matrix(ranef(m0))
m00<-t(m00)
fen<-as.data.frame(m00[4:103,])
colnames(fen)[1]<-"sar"
fen$gen<-rownames(fen)
mrk<-read.table("mrkfilt.txt",h=T)
colnames(mrk)[1]<-"gen"
dados<-full_join(fen,mrk,by="gen")
rm(dat,fen,m0,m00,mrk)
## Redes com e sem regularização
```

```
sample_size = floor(0.8*nrow(dados))
Results <- list()
for (k in 1:length(iterations)){
m<-0
R <- matrix(NA,400, 4) #matriz de todos os resultados
colnames(R) <- c("N_Neuronios_Cam1", "N_Neuronios_Cam2", "Correlação","Erro padrão média","EQM","Erro padrão média")
for(p in 1:20) {
for(s in 1:20) {
n <- n + 1
for (i in 1:10) {
assign(paste0("picked",i),sample(seq_len(nrow(dados)),size = sample_size))
assign(paste0("train",i),subset(get("dados")[get(paste0("picked",i)),]))
assign(paste0("test",i),subset(get("dados")[-get(paste0("picked",i)),]))
assign(paste0("NN",i),neuralnetwork(X = get(paste0("train",i))[,-c(1,2)],
y = get(paste0("train",i))[,1],hidden.layers = c(p,s),
loss.type = "huber",L1=5,
learn.rates = 0.01,activ.functions = "tanh",regression = T,
n.epochs = 500,val.prop = 0))
assign(paste0("ypred",i),predict(get(paste0("NN",i)),newdata = get(paste0("test",i))[,-c(1,2)]))
assign(paste0("ypred",i),get(paste0("ypred",i))$predictions)
assign(paste0("cor",i),cor(as.numeric(get(paste0("ypred",i))),get(paste0("test",i))[,1]))
assign(paste0("rmse",i),sqrt(mean((as.numeric(get(paste0("ypred",i))) - get(paste0("test",i))[,1])^2)))}
cor<-rbind(cor1,cor2,cor3,cor4,cor5,cor6,cor7,cor8,cor9,cor10)
rmse<-rbind(rmse1,rmse2,rmse3,rmse4,rmse5,rmse6,rmse7,rmse8,rmse9,rmse10)
co<-mean(cor,na.rm = T)
```

```
rmse<-mean(rmse,na.rm = T)
R[n,1:4] <- c(p, s, round(co,4), round(rmse,4))}
m <- m + 1}}
########## GBLUP
rownames(dados)<-dados$gen
y.trn<-dados
marca3<-A.mat(as.matrix(dados[,3:3959]))
for (i in 1:10) {
  y.trn<-dados
  vv <- sample(dados$gen,round(nrow(dados)/5))
  y.trn[vv,"sar"]<-NA
  assign(paste0("gblup",i),mmer(sar~1,
  random= ~vsr(gen,Gu=marca3),rcov= ~ units,
  data=y.trn, verbose = TRUE))
  assign(paste0("ans",i),as.data.frame(get(paste0("gblup",i))$U$`u:gen`$sar))
  assign(paste0("cor",i),cor(get(paste0("ans",i))[vv,],dados[vv,"sar"],use="complete"))
  assign(paste0("rmse",i),sqrt(mean((get(paste0("ans",i))[vv,] - dados[vv,"sar"]))^2))}
## Random Forest
library(randomForest)
sample_size = floor(0.8*nrow(dados))
for (i in 1:10) {
  assign(paste0("picked",i),sample(seq_len(nrow(dados)),size = sample_size))
  assign(paste0("train",i),subset(get("dados")[get(paste0("picked",i)),]))
  assign(paste0("test",i),subset(get("dados")[-get(paste0("picked",i)),]))
  assign(paste0("classifier",i),randomForest(sar~.,data=get(paste0("train",i)),
n.trees=500,mtry=3957/3))
  assign(paste0("ypred",i),predict(get(paste0("classifier",i)),
                                   newdata = get(paste0("test",i))[,-1]))
```

```
  assign(paste0("ypred",i),get(paste0("ypred",i)))
  assign(paste0("cor",i),cor(as.numeric(get(paste0("ypred",i))),get(paste0("test",i))[,1]))
  assign(paste0("rmse",i),sqrt(mean((as.numeric(get(paste0("ypred",i))) -
get(paste0("test",i))[,1])^2)))}
## Bagging
sample_size = floor(0.8*nrow(dados))
for (i in 1:10) {
  assign(paste0("picked",i),sample(seq_len(nrow(dados)),size = sample_size))
  assign(paste0("train",i),subset(get("dados")[get(paste0("picked",i)),]))
  assign(paste0("test",i),subset(get("dados")[-get(paste0("picked",i)),]))
  assign(paste0("classifier",i),randomForest(sar~.,data=get(paste0("train",i)),
n.trees=500,mtry=3957))
  assign(paste0("ypred",i),predict(get(paste0("classifier",i)),
                           newdata = get(paste0("test",i))[,-1]))
  assign(paste0("ypred",i),get(paste0("ypred",i)))
  assign(paste0("cor",i),cor(as.numeric(get(paste0("ypred",i))),
                       get(paste0("test",i))[,1]))
  assign(paste0("rmse",i),sqrt(mean((as.numeric(get(paste0("ypred",i))) -
get(paste0("test",i))[,1])^2)))}
## Boosting
library(gbm)
for (i in 1:10) {
  assign(paste0("picked",i),sample(seq_len(nrow(dados)),size = sample_size))
  assign(paste0("train",i),subset(get("dados")[get(paste0("picked",i)),]))
  assign(paste0("test",i),subset(get("dados")[-get(paste0("picked",i)),]))
  assign(paste0("classifier",i),gbm(sar~.,data=get(paste0("train",i)),
    n.trees=500,distribution = "gaussian",shrinkage = 0.1))
  assign(paste0("ypred",i),predict(get(paste0("classifier",i)),
```

```
                        newdata = get(paste0("test",i))[,-1]))
  assign(paste0("ypred",i),get(paste0("ypred",i)))
  assign(paste0("cor",i),cor(as.numeric(get(paste0("ypred",i))),
                        get(paste0("test",i))[,1]))
  assign(paste0("rmse",i),sqrt(mean((as.numeric(get(paste0("ypred",i))) -
get(paste0("test",i))[,1])^2)))}
## Árvore de decisão
library(rpart)
for (i in 1:10) {
  assign(paste0("picked",i),sample(seq_len(nrow(dados)),size = sample_size))
  assign(paste0("train",i),subset(get("dados")[get(paste0("picked",i)),]))
  assign(paste0("test",i),subset(get("dados")[-get(paste0("picked",i)),]))
  assign(paste0("classifier",i),rpart(sar~.,data=get(paste0("train",i)),method="anova"))
  assign(paste0("ypred",i),predict(get(paste0("classifier",i)),
                        newdata = get(paste0("test",i))[,-1]))
  assign(paste0("ypred",i),get(paste0("ypred",i)))
  assign(paste0("cor",i),cor(as.numeric(get(paste0("ypred",i))),
                        get(paste0("test",i))[,1]))
  assign(paste0("rmse",i),sqrt(mean((as.numeric(get(paste0("ypred",i))) -
get(paste0("test",i))[,1])^2)))}
```