Computational and Structural Biotechnology Journal 20 (2022) 5490–5499

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

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

Weverton Gomes da Costa [a,*], Maurício de Oliveira Celeri [b], Ivan de Paiva Barbosa [c], Gabi Nunes Silva [d], Camila Ferreira Azevedo [c], Aluizio Borem [c], Moysés Nascimento [b], Cosme Damião Cruz [a]

[a] Department of General Biology, Bioinformatics Laboratory, Federal University of Viçosa, Viçosa, MG, Brazil
[b] Department of Statistics, Laboratory of Computational Intelligence and Statistical Learning, Federal University of Viçosa – UFV, Viçosa, MG, Brazil
[c] Department of Agronomy, Federal University of Viçosa, Viçosa, MG, Brazil
[d] Department of Mathematics and Statistics, Federal University of Rondônia, Ji-Paraná Campus, RO, Brazil

ARTICLE INFO

Article history:
Received 12 April 2022
Received in revised form 20 September 2022
Accepted 20 September 2022
Available online 24 September 2022

Keywords:
Genome wide selection
Quantitative trait locus
Non-additive effects
Multivariate adaptive regression splines
Genome-enabled prediction

ABSTRACT

Genomic wide selection (GWS) is one contributions 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 (h^2) of 0.3, 0.5 or 0.8 were simulated. Variation in heritability and number of QTL impacts the performance of methods. About quantitative traits (40, 80, 120, 240, and 480 QTLs) was observed highest R2 to Radial Base Network (RBF) and G-BLUP, followed by Random Forest (RF), Bagging (BA), and Boosting (BO). RF and BA also showed better results for traits to h^2 of 0.3 with R^2 values 16.51% and 16.30%, respectively, while MARS methods showed better results for oligogenic traits with R^2 values ranging from 39,12 % to 43,20 % in h^2 of 0.5 and from 59.92% to 78,56% in h^2 of 0.8. Non-additive MARS methods also showed high R2 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.

1. Introduction

Genomic wide selection (GWS), proposed by [1], 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 [2]. Furthermore, GWS made it possible to accelerate the improvement process by half the time in relevant crops, helping to sustain current food demands [3–5].This time reduction allowed breeders to maximize genetic gains per unit of time, in addition to early selection [6,7]. 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 [2,8,9].

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 [2]. 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 [10]. 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 [5,9,11,12].

Among the various methods based on machine learning, Multivariate Adaptive Regression Splines (MARS) is considered one of the most flexible [13], it proves to be more parsimonious and per-

* Corresponding author.
E-mail address: wevertonufv@gmail.com (W.G.d. Costa).

https://doi.org/10.1016/j.csbj.2022.09.029

forms better than artificial neural networks for genomic prediction in some studies [10,14]. MARS produces continuous models that can have multiple partitions, automatically models nonlinearities, and contemplates interactions of predictor variables using adaptively selected spline functions [15–17].

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 [14]. Although several studies have proven the high power of MARS in the evaluation of genomic data in the medical field [14,18–20], 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.

2. Material and methods

2.1. 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 (*P*1 dominant and *P*2 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. Markers were codominant (SNPs - Single Nucleotide Polymorphism), allowing the identification of heterozygous individuals.

2.2. Simulation and constitution of phenotypic values

From the simulated genotypic data of the F2 population, 18 traits with numbers of controlling genes ranging from 8 to 480 and heritability of 0.3, 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 Fig. 1).

Eight QTL controlled for C1, C7 and C13 traits, defined by the central markers of the first eight linkage groups. For traits *C*2 to C6, C8 to C12, and C14 to C18 the QTL were distributed keeping an approximate distance between them, within the first 8 linkage groups (Supplementary Fig. 1).

The total phenotypic values expressed by a given individual for traits C1 to C18 were simulated according with [21–23] 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 [24] $e\ N(0, V_e)$, where $V_e = \left(\left(1 - h^2\right)V_g\right)/h^2$, where V_e is the residual variance, V_g is the genotypic variance, and h^2 the heritability.

2.3. 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 [12,22,25]. 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 [26].

2.4. 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: G-BLUP – G-BLUP; MLP and RBF – NETWORK; DT, BA, BO and RF – TREES; and MARS 1, MARS 2 and MARS 3 – MARS.

2.4.1. Multivariate Adaptive regression Splines (MARS)

The algorithm proposed by [27] Multivariate Adaptive Regression Splines (MARS), considers an expansion in piecewise linear functions, called basis functions (BFs), as follows:

Table 1
Number of controlling loci and heritability (h^2) of the 12 simulated traits (C1 to C18).

h^2	Number of controlling loci					
	8	**40**	**80**	**120**	**240**	**480**
0,3	C1	*C*2	C3	C4	C5	C6
0,5	C7	C8	C9	C10	C11	C12
0,8	C13	C14	C15	C16	C17	C18

$$(x-t)_+ = \begin{cases} x-t, sex > t, \\ 0, ifotherwise. \end{cases}; (t-x)_+ = \begin{cases} t-x, sex < t, \\ 0, ifotherwise. \end{cases}$$

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 used functions from the set $C = \left\{ (X_j - t)_+, \ (t - X_j)_+ \right\}_{t \in \{x_{1j}, \ x_{2j}, \ \ldots, \ x_{Nj}\} \ j=1,2,\ldots,\ p}$ and/or its products. The MARS model, which is a linear combination of the BFs and/or their interactions, is given by [28]:

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

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 [28]:

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

where $\widehat{\beta}_{M+1}$ and $\widehat{\beta}_{M+2}$ 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 [29].

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 [29]. The GCV is calculated as [28]:

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

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 [27], N is the number of datasets used in cross-validation and $\widehat{f}_{\lambda}(x_i)$ 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 "earth" package of R. A preliminary analysis was carried out for the second stopping criterion [30], 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.

2.4.2. 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 $u_a \ N(0, G_a\sigma^2_{u_a})$; $u_d \ N\left(0, G_d\sigma^2_{u_d}\right)$; $u_{epi} \ N\left(0, G_{epi}\sigma^2_{u_{epi}}\right)$ and $\varepsilon \ N(0, I\sigma^2_{\varepsilon})$, 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 [31]:

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

$$S_{ij} = \begin{cases} -2(1-p_j)^2, ifM_{ij} = AA \\ 2p_j(1-p_j), ifM_{ij} = Aa \\ -2p_j^2, ifM_{ij} = aa \end{cases}$$

In this way, we obtain:

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

Where # is the Hadamard product operator.

The mixed model equations for the full model were given by [24]:

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

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

2.4.3. 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 ($f_{xj}(x_i) = x_i$).

2.4.4. 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 hid-

den layer, and an output layer. RBF training is hybrid (supervised and unsupervised) and the input layer information goes through a linear k-means cluster [11]. The hidden layer applies a nonlinear 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 [11]. During the training process, only the weights between the hidden layer and the output layer are modified [9]. 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.

2.4.5. Decision tree

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 $x_1, \cdots, X_m$ 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 [32]:

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

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

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

where: $\widehat{y}_{R_1}$ is the average of the response variable of the training observations belonging to the region $R_1(j,s)$, $\widehat{y}_{R_2}$ 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 traits of each individual.

2.4.6. Bagging

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 [33,34]. Thus, a number B of models are obtained: $\widehat{f}^1(x), \widehat{f}^2(x), \cdots, \widehat{f}^B(x)$. These generated models are used to obtain an average model, given by: $\widehat{f}_{médio}(x) = \frac{1}{B}\sum_{b=1}^{B} \widehat{f}^b(x)$. The number of trees sampled for BA was set at 500 trees.

2.4.7. Random Forest

Random Forest (RF) [35] 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 [25].

2.4.8. Boosting

Boosting (BO) creates trees sequentially using information from previous trees [32]. 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 [36]. 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 [28]:

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

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 $\gamma = \gamma_1, \gamma_2, \cdots, \gamma_m$.

2.5. Efficiency parameters

To evaluate the efficiency of the techniques, the selective accuracy was used, which is measured by the square of the correlation (R2) 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: $R^2 = (cor(\hat{y}, y))^2$ and$RMSE = \sqrt{\frac{\sum(y-y)^2}{n}}$.

2.6. Training and validation

For the training and validation of the techniques used, cross-validation (k-fold) was performed with k = 5 partitions [37]. In each of the five rounds, four of these subsets constituted the training population (80 % – 800 individuals), and the remaining subset constituted the validation population (20–200 individuals). The techniques were compared based on the arithmetic mean of the five performance estimates of the validation sets.

2.7. Computational aspects

Population simulations were performed using the GENES software [38]. The G-BLUP, DT, BA, RF, BO, and MARS methods were performed with the GENES software integrated with the R software [39,40]. The MLP and RBF methods were performed by the GENES software integrated with the MATLAB software [39,41].

3. Results

The phenotypic and genotypes values of the 18 simulated traits are shown in Fig. 1. The low variations can be explained by the 12% coefficient assigned in the simulation and the mean was very close to 100, as defined in the simulation.

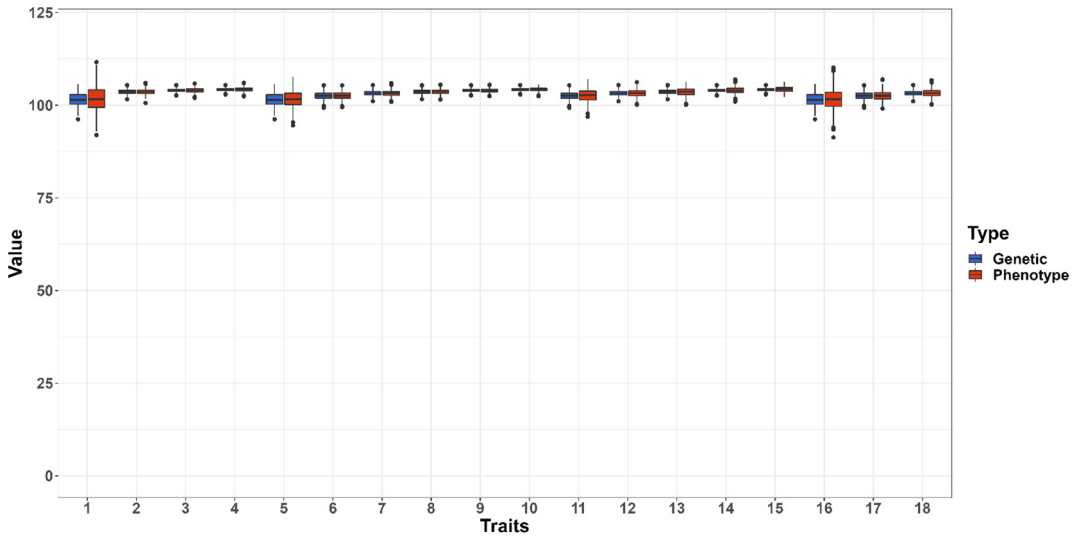

Fig. 1. Boxplot of the genetic and phenotypic values of the 18 simulated traits, considering a coefficient of variation equal to 12% and mean to 100. The specification of each characteristic is represented in Table 1.

The selective accuracy (R2) of the prediction of breeding values for all methods was higher in scenarios with higher heritability (Fig. 1). On the other hand, the variation in the number of QTL showed that the methods have diversity among the results obtained, indicating that the number of QTL of the traits 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.

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 (Fig. 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, except for MARS and DT, 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 R2 and always above the general average (red line) for the traits for both heritability scenarios (Fig. 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 R2 values (above 60 %), and considering the standard error, values lower only than G-BLUP (Fig. 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 (Fig. 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 (Fig. 1). In addition, these methods presented values of R^2 greater than the overall mean of R^2 in all scenarios (Fig. 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 (Fig. 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.

For scenarios with 8 QTLs, the trees (Fig. 3) showed better results. It was also observed that the higher the increase in the number of QTLs, the lower the difference between the methods for RMSE. In the largest QTLs scenarios, DT had higher RMSE values in most scenarios.

The RBF method presented very similar RMSE values when compared to those obtained through G-BLUP for all scenarios (Fig. 2). From 40 QTL, 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 %.

4. 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 [20]. Furthermore, in hybrid populations, non-additive effects, i.e., dominance, and epistasis, are highly relevant and should also be considered [42]. In this sense, methods that deal with high dimensionality and take into account possible interactions between predictor variables have important traits. Many recent studies have been applied to GWS and have shown that machine learning and neural network methods can perform

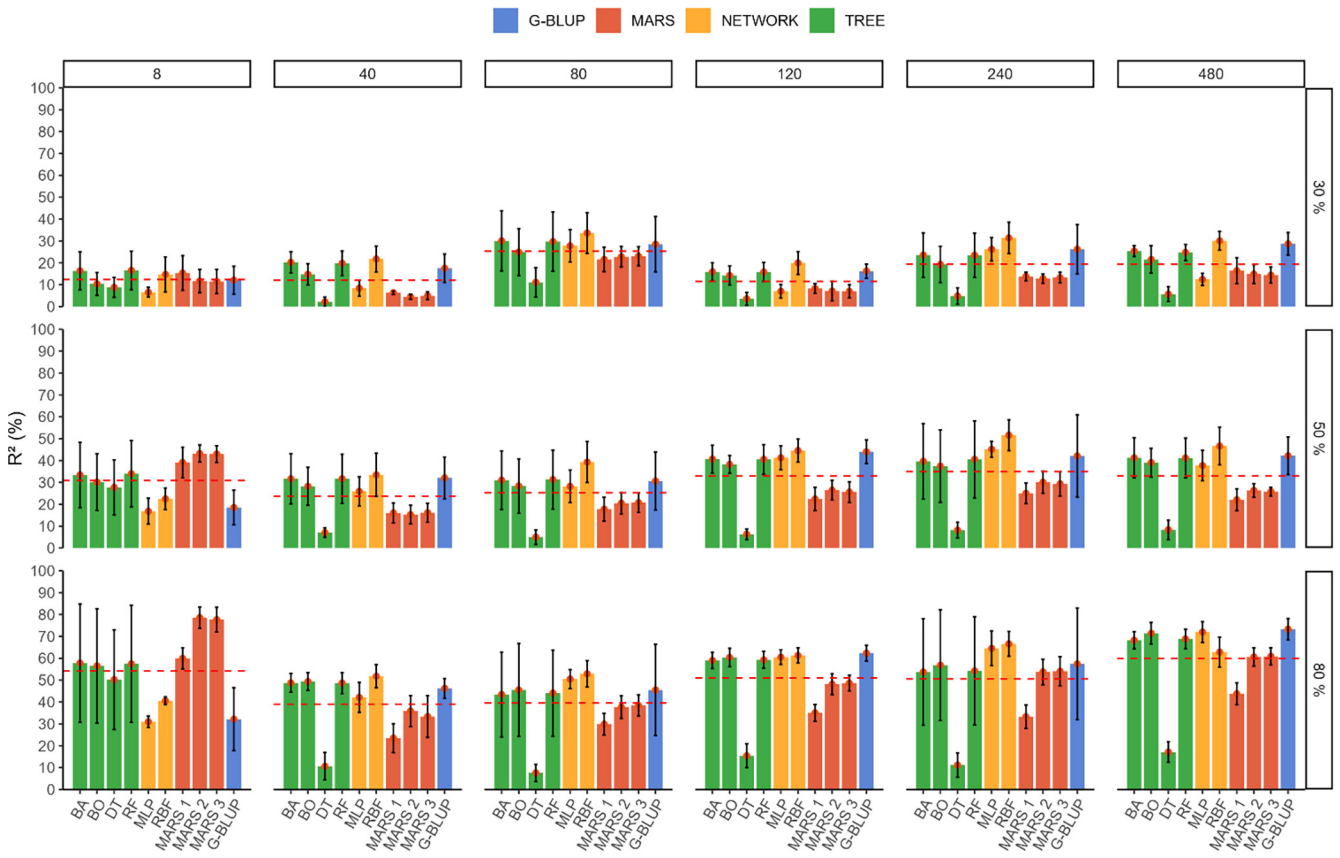

Fig. 2. Average results of selective accuracy (R^2) as a function of the number of genes and heritability for the families of the methods: Trees [Bagging (BA), Boosting (BO), Decision Tree (DT); and Random Forest (RF)]; Network (Multilayer Perceptron Network (MLP) and Radial Base Function Network (RBF) 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. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

better or similarly in predicting genotypic data phenotypes compared to statistical methods [9,12,22,43–46].

However, there is still a gap to be filled on which methods may be preferable 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 were able to demonstrate the strong effect of heritability and increase in QTL number on R2 and RMSE values. Several studies have shown that there is a favorable effect of heritability on selective accuracy [9,22,36,47,48], as also obtained in this study. This is justified by the greater genetic variation in higher heritability's and, consequently, less environmental effect, contributing to more accurate predictions of marker effects [22].

The results showed that there was a reduction in the RMSE in scenarios with a greater number of QTL, and that, in the same way, there was also a reduction in the RMSE in the scenarios with greater heritability, however in a smaller proportion. Results similar to those obtained in this study were observed by [36], in scenarios with heritability ranging from 0.1 to 0.5 and 100 QTL, and [22] 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 [22,36,49].

As MARS can simultaneously include multiple terms (additive and epistatic effects) in a model [50] genetic interactions can be better evaluated. Apparently, this fact could lead to a better prediction for GWS, since, in this way, it would be possible to reduce the residual variance of the model, by capturing information that before was isolated only residual component, such as the effects of the interaction between markers. However, as [14] explain, some patterns of interaction tend to be less pronounced to be detected in features with a high number of QTL. 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 [12], 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 [51].

These results show that MARS is an 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 [22]. 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

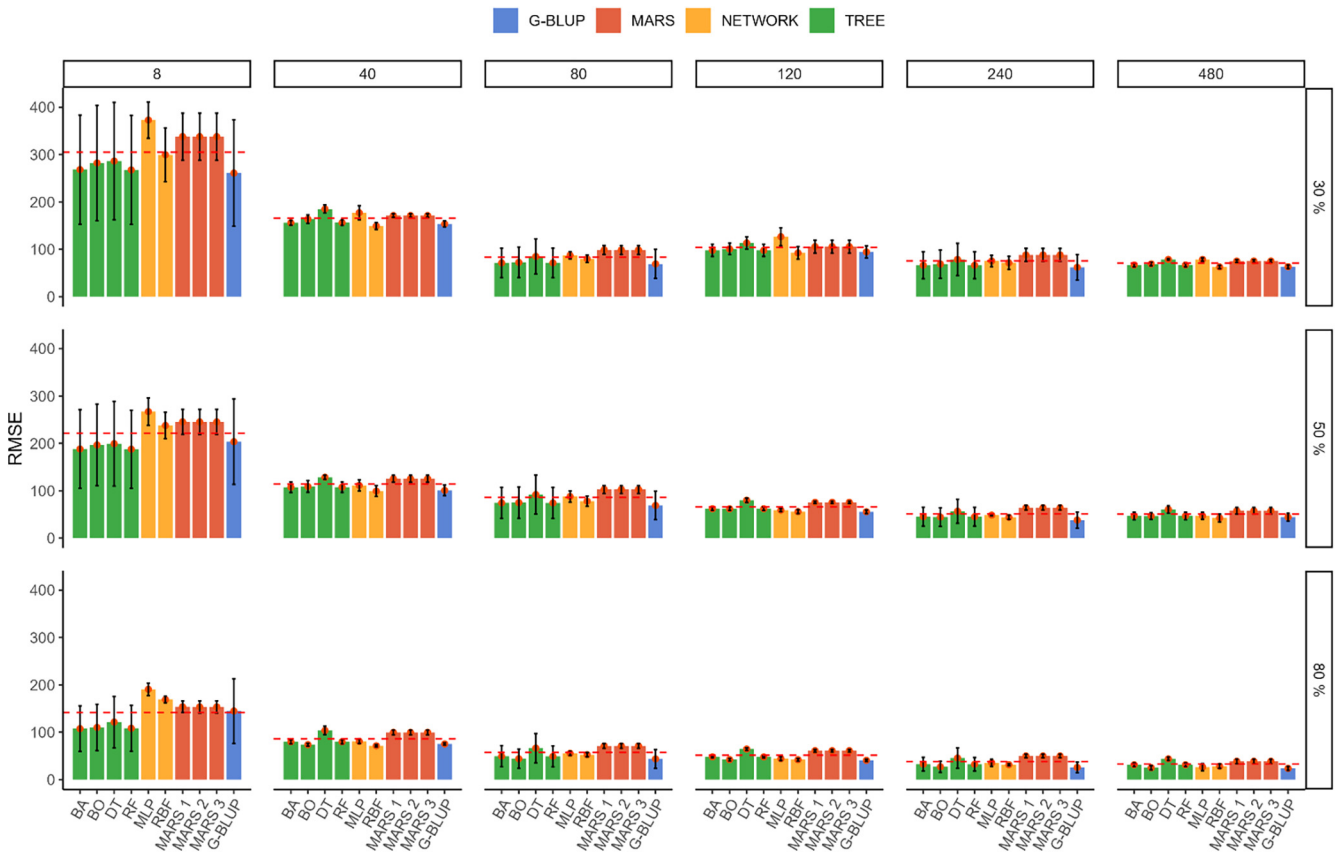

Fig. 3. Average results of predictive accuracy (RMSE) as a function of the number of genes and heritability for the families of the methods: Trees [Bagging (BA), Boosting (BO), Regression Tree (DT); and Random Forest (RF)]; Network (Multilayer Perceptron Network (MLP) and Radial Base Function Network (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. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

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 [27].

The frame is different when there is a large number of predictor variables with high correlation and response trait 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 [12]. As [27] 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 [12,22]. In addition to MARS, 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 [52].

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 [27]. 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 (250 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 [22]. 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 observed by [12] to predict the genetic values for rust incidence in *Coffea arabica* and by [22] 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 [53]. 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 [54]. 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 [32], thus allowing a better prediction of the genetic effects of individuals, as observed in this study.

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 [22]. Both RBF and MLP were harmed by the excess of ineffective markers, showing lower performance compared to other methods, as also found by [22] and [55]. 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.

G-BLUP considers the interaction between marker pairs and relies on the DL between SNPs and QTL, moreover, when QTL are in strong LD and the use of an unweighted genomic relationship matrix in G-BLUP can cause upward bias in the heritability estimate [56–59]. 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-BLUP was highlighted in the performance of the prediction of the GEBVs, presenting very similar results to the Family Network methods for the traits with more QTL. Results similar to those found in this study were found in [22,60]. However, 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 [12]. These results corroborate other studies where the GBLUP precision increased for characters with a high proportion of non-additive variation and when with increasing heritability [61–63] and justified due to G-BLUP principle that genomic predictions are based on the relatedness derived from all markers [60], so when more markers have a genetic effect the prediction accuracy increases.

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 and 280 QTL. This can be explained by the fact that the F2 population has a high rate of linkage disequilibrium (LD), due to the combination process. This LD can then cause false-positive signals for some loci, which have no connection with the studied trait in question [59]. 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. Alternatively, use of hybrid modeling schemes (combination of two or more methodologies) including MARS had been previously very effective with initial data clustering using c-means or principal components [64–66], it could be more important for diverse population for genomic predictions. For example, studies such as the one by [43] proposed hybrid smart modeling schemes for heart disease classification using combined MARS-ANN. This would be a viable alternative to improve predictability and decrease the effect of multicollinearity using MARS on genomic data, which is worthy of further investigation and deserves further research.

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 improvement that can improve the predictability of the traits and that must be tested, mainly 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 [36,67]. 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 markers with null effects, which can impair the accuracy of the methods [12,22].

Each technique has its specificity and must be evaluated in a wide set of data so that the decision on which method to base is correctly made [2]. 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.

5. 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 to be used.

CRediT authorship contribution statement

Weverton Gomes da Costa: Conceptualization, Methodology, Formal analysis, Data curation, Investigation, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Mauricio de Oliveira Celeri:** Methodology, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Ivan de Paiva Barbosa:** Conceptualization, Methodology, Visualization, Writing – original draft. **Gabi Nunes Silva:** Visualization, Writing – original draft. **Camila Ferreira Azevedo:** Visualization, Writing – original draf. **Aluizio Borem:** Visualization, Writing – original draft. **Moysés Nascimento:** Investigation , Visualization, Writing – original draft, Project administration, Supervision, Writing - Review & Editin. **Cosme Damião Cruz:** Investigation, Visualization, Writing – original draft, Project administration, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.09.029.

References

[1] Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics 2001;157:1819–29.
[2] Tong H, Nikoloski Z. Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. J Plant Physiol 2021;257:. https://doi.org/10.1016/j.jplph.2020.153354153354.
[3] Singh BD, Singh AK. Marker-assisted plant breeding: Principles and practices. 2015. 10.1007/978-81-322-2316-0.
[4] Peixoto LA, Laviola BG, Alves AA, Rosado TB, Bhering LL. Breeding Jatropha curcas by genomic selection: A pilot assessment of the accuracy of predictive models. PLoS ONE 2017;12:1–16. https://doi.org/10.1371/journal.pone.0173368.
[5] Li B, Zhang N, Wang YG, George AW, Reverter A, Li Y. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. Front Genet 2018;9:1–20. https://doi.org/10.3389/fgene.2018.00237.
[6] Yabe S, Hara T, Ueno M, Enoki H, Kimura T, Nishimura S, et al. Potential of genomic selection in mass selection breeding of an allogamous crop: An empirical study to increase yield of common buckwheat. Front Plant Sci 2018;9:1–12. https://doi.org/10.3389/fpls.2018.00276.
[7] Sousa TV, Caixeta ET, Alkimim ER, Oliveira ACB, Pereira AA, Sakiyama NS, et al. Early Selection Enabled by the Implementation of Genomic Selection in Coffea arabica Breeding. Front Plant Sci 2019;9:1–12. https://doi.org/10.3389/fpls.2018.01934.
[8] Alkimim ER, Caixeta ET, Sousa TV, Resende MDV, Silva FL, Sakiyama NS, et al. Selective efficiency of genome-wide selection in Coffea canephora breeding. Tree Genet Genomes 2020;16. https://doi.org/10.1007/s11295-020-01433-3.
[9] Sant'Anna IC, Nascimento M, Silva GN, Cruz CD, Azevedo CF, Gloria LS, et al. Genome-enabled prediction of genetic values for using radial basis function neural networks. Funct Plant Breed J 2020;1:1–8. 10.35418/2526-4117/v1n2a1.
[10] Liew BXW, Peolsson A, Rugamer D, Wibault J, Löfgren H, Dedering A, et al. Clinical predictive modelling of post-surgical recovery in individuals with cervical radiculopathy: a machine learning approach. Sci Rep 2020;10:1–10. https://doi.org/10.1038/s41598-020-73740-7.
[11] Cruz CD, Nascimento M. Inteligência Computacional aplicada ao melhoramento genético. Viçosa, MG: Editora UFV; 2018.
[12] Sousa IC, Nascimento M, Silva GN, Nascimento ACC, Cruz CD, Fonseca e Silva F, et al. Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. Sci Agric 2021;78:1–8. https://doi.org/10.1590/1678-992x-2020-0021.
[13] Cook NR, Zee RYL, Ridker PM. Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. Stat Med 2004;23:1439–53. https://doi.org/10.1002/sim.1749.
[14] Lin HY, Wang W, Liu YH, Soong SJ, York TP, Myers L, et al. Comparison of multivariate adaptive regression splines and logistic regression in detecting SNP-SNP interactions and their application in prostate cancer. J Hum Genet 2008;53:802–11. https://doi.org/10.1007/s10038-008-0313-z.
[15] Taylan P, Weber GW. CG-Lasso Estimator for Multivariate Adaptive Regression Spline. In: Tas K, Baleanu D, Machado JAT, editors. Math. Methods Eng. Apl. Dyn. Complex Syst., Springer International Publishing AG; 2019, p. 121–36. 10.1007/978-3-319-90972-1_9.
[16] Altinok G, Karagoz P, Batmaz I. Learning to rank by using multivariate adaptive regression splines and conic multivariate adaptive regression splines. Comput Intell 2020:1–38. https://doi.org/10.1111/coin.12413.
[17] Zheng G, Zhang W, Zhou H, Yang P. Multivariate adaptive regression splines model for prediction of the liquefaction-induced settlement of shallow foundations. Soil Dyn Earthq Eng 2020;132:. https://doi.org/10.1016/j.soildyn.2020.106097106097.
[18] York TP, Eaves LJ, van den Oord EJCG. Multivariate adaptive regression splines: A powerful method for detecting disease-risk relationship differences among subgroups. Stat Med 2006;25:1355–67. https://doi.org/10.1002/sim.2292.
[19] Chang CD, Wang CC, Jiang BC. Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors. Expert Syst Appl 2011;38:5507–13. https://doi.org/10.1016/j.eswa.2010.10.086.
[20] Tang J, Liu R, Zhang YL, Liu MZ, Hu YF, Shao MJ, et al. Application of Machine-Learning Models to Predict Tacrolimus Stable Dose in Renal Transplant Recipients. Sci Rep 2017;7. https://doi.org/10.1038/srep42192.
[21] Cruz CD. Programa GENES – Análise Multivariada e Simulação. Viçosa, MG, Brazil: Editora UFV; 2006.
[22] Barbosa IP, Silva MJ, Costa WG, Sant'Anna IC, Nascimento M, Cruz CD. Genome-enabled prediction through machine learning methods considering different levels of trait complexity. Crop Sci 2021;61:1890–902. https://doi.org/10.1002/csc2.20488.
[23] Sant'Anna IC, Tomaz RS, Silva GN, Nascimento M, Bhering LL, Cruz CD. Superiority of artificial neural networks for a genetic classification procedure. Genet Mol Res 2015;14:9898–906. https://doi.org/10.4238/2015.August.19.24.
[24] Resende MDV, Silva FF, Azevedo CF. Estatística Matemática, Biométrica e Computacional. Viçosa, MG, Brazil: Editora UFV; 2014.
[25] Costa WG, Barbosa I, De P, Souza JE, Cruz CD, Nascimento M, Oliveira ACB. Machine learning and statistics to qualify environments through multi-traits in Coffea arabica. PLoS ONE 2021;16:1–21. https://doi.org/10.1371/journal.pone.0245298.
[26] Solano Meza JK, Orjuela Yepes D, Rodrigo-Ilarri J, Cassiraga E. 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 2019;5:e02810.
[27] Friedman JH. Multivariate Adaptative regression Splines. Ann Stat 1991;19:1–141.
[28] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: Data mining, inference, and prediction. 2. ed. New York, NY, USA: Springer; 2009. 10.1007/978-1-4419-9863-7_941.
[29] Zhang W, Goh ATC. Multivariate adaptive regression splines and neural network models for prediction of pile drivability. Geosci Front 2016;7:45–52. https://doi.org/10.1016/j.gsf.2014.10.003.
[30] Milborrow S. Notes on the earth package; 2019:1–68.
[31] Zhang H, Yin L, Wang M, Yuan X, Liu X. Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. Front Genet 2019;10:1–10. https://doi.org/10.3389/fgene.2019.00189.
[32] James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning. Springer Texts Stat 2021:612. https://doi.org/10.1007/978-1-0716-1418-1_1.
[33] Breiman L. Bagging Predictors. Mach Learn 1996;24:123–40. https://doi.org/10.1007/BF00058655.
[34] Prasad AM, Iverson LR, Liaw A. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. Ecosystems 2006;9:181–99. https://doi.org/10.1007/s10021-005-0054-1.
[35] Boehmke B, Greenwell B. Random Forests. Hands-On Mach. Learn. with R, vol. 45, Chapman and Hall/CRC; 2019, p. 203–19. 10.1201/9780367816377-11.
[36] Ghafouri-Kesbi F, Rahimi-Mianji G, Honarvar M, Nejati-Javaremi A. Predictive ability of Random Forests, Boosting, Support Vector Machines and Genomic Best Linear Unbiased Prediction in different scenarios of genomic evaluation. Anim Prod Sci 2017;57:229–36. https://doi.org/10.1071/AN15538.
[37] Bengio Y, Grandvalet Y. No Unbiased Estimator of the Variance of K-Fold Cross-Validation. J Mach Learn Res 2004;5:1089–105. https://doi.org/10.1016/S0006-291X(03)00224-9.
[38] Cruz CD. GENES - Software para análise de dados em estatística experimental e em genética quantitativa. Acta Sci - Agron 2013;35:271–6. https://doi.org/10.4025/actasciagron.v35i3.21251.
[39] Cruz CD. Genes software – extended and integrated with the R, Matlab and Selegen. Acta Sci - Agron 2016;38:547–52. https://doi.org/10.4025/actasciagron.v38i4.32629.
[40] R Core Team, Computing RF for S, Team RC. R: A Language and Environment for Statistical Computing 2020. https://www.r-project.org/. (accessed July 1, 2020).
[41] MATLAB. Natick, Massachusetts: The MathWorks Inc.; 2019.
[42] Schnable PS, Springer NM. Progress toward understanding heterosis in crop plants. Annu Rev Plant Biol 2013;64:71–88. https://doi.org/10.1146/annurev-arplant-042110-103827.
[43] Shao YE, Hou CD, Chiu CC. Hybrid intelligent modeling schemes for heart disease classification. Appl Soft Comput J 2014;14:47–52. https://doi.org/10.1016/j.asoc.2013.09.020.
[44] Silva GN, Tomaz RS, Sant'Anna IC, Nascimento M, Bhering LL, Cruz CD. Neural networks for predicting breeding values and genetic gains. Sci Agric 2014;71:494–8. 10.1590/0103-9016-2014-0057.
[45] Ma W, Qiu Z, Song J, Cheng Q, Ma C. DeepGS: Predicting phenotypes from genotypes using Deep Learning. BioRxiv 2017. https://doi.org/10.1101/241414.
[46] Zingaretti LM, Gezan SA, Ferrão LFV, Osorio LF, Monfort A, Muñoz PR, et al. Exploring Deep Learning for Complex Trait Genomic Prediction in Polyploid Outcrossing Species. Front Plant Sci 2020;11:1–14. https://doi.org/10.3389/fpls.2020.00025.
[47] Coutinho AE, Neder DG, Silva MC, Arcelino EC, Brito SG, Carvalho Filho JLS. Prediction of phenotypic and genotypic values by BLUP/GWS and neural networks. Rev Caatinga 2018;31:532–40. https://doi.org/10.1590/1983-21252018v31n301rc.
[48] Moura EG, Pamplona AKA, Balestre M. Functional models in genome-wide selection. PLoS ONE 2019;14:e0222699.

[49] Coster A, Bastiaansen JWM, Calus MPL, Van Arendonk JAM, Bovenhuis H. Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. Genet Sel Evol 2010;42:1–11. https://doi.org/10.1186/1297-9686-42-9.
[50] Everingham YL, Sexton J. An introduction to Multivariate Adaptive Regression Splines for the cane industry. 33rd Annu Conf Aust Soc Sugar Cane Technol 2011, ASSCT 2011 2011:255–68.
[51] Cruz CD. Princípios de genética quantitativa. 2. ed. Viçosa, MG: Editora UFV; 2012.
[52] De Veaux RD, Ungar LH. Multicollinearity: A tale of two nonparametric regressions 1994:393–402. 10.1007/978-1-4612-2660-4_40.
[53] Diaz-Uriarte R. GeneSrF and varSelRF: A web-based tool and R package for gene selection and classification using random forest. BMC Bioinf 2007;8:1–7. https://doi.org/10.1186/1471-2105-8-328.
[54] Fuleky P. Macroeconomic Forecasting in the Era of Big Data. vol. 52. 2020.
[55] Sant'Anna I de C, Gouvêa LRL, Martins MA, Scaloppi Junior EJ, de Freitas RS, Gonçalves P de S. Genetic diversity associated with natural rubber quality in elite genotypes of the rubber tree. Sci Rep 2021;11:1–10. 10.1038/s41598-020-80110-w.
[56] Speed D, Hemani G, Johnson MR, Balding DJ. Improved heritability estimation from genome-wide SNPs. Am J Hum Genet 2012;91:1011–21. https://doi.org/10.1016/j.ajhg.2012.10.010.
[57] Legarra A. Comparing estimates of genetic variance across different relationship models. Theor Popul Biol 2016;107:26–30. https://doi.org/10.1016/j.tpb.2015.08.005.
[58] Fernando RL, Cheng H, Sun X, Garrick DJ. A comparison of identity-by-descent and identity-by-state matrices that are used for genetic evaluation and estimation of variance components. J Anim Breed Genet 2017;134:213–23. https://doi.org/10.1111/jbg.12275.
[59] Mathew B, Léon J, Sillanpää MJ. A novel linkage-disequilibrium corrected genomic relationship matrix for SNP-heritability estimation and genomic prediction. Heredity (Edinb) 2018;120:356–68. https://doi.org/10.1038/s41437-017-0023-4.
[60] Wang J, Zhou Z, Zhang Z, Li H, Liu D, Zhang Q, et al. Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traits. Heredity (Edinb) 2018;121:648–62. https://doi.org/10.1038/s41437-018-0075-0.
[61] Dufflocq P, Pérez-Enciso M, Lhorente JP, Yáñez JM. Accuracy of genomic predictions using different imputation error rates in aquaculture breeding programs: A simulation study. Aquaculture 2019;503:225–30. https://doi.org/10.1016/j.aquaculture.2018.12.061.
[62] Pocrnic I, Lourenco DAL, Masuda Y, Misztal I. Accuracy of genomic BLUP when considering a genomic relationship matrix based on the number of the largest eigenvalues: A simulation study. Genet Sel Evol 2019;51:1–10. https://doi.org/10.1186/s12711-019-0516-0.
[63] Liu X, Wang H, Hu X, Li K, Liu Z, Wu Y, et al. Improving Genomic Selection With Quantitative Trait Loci and Nonadditive Effects Revealed by Empirical Evidence in Maize. Front Plant Sci 2019;10. 10.3389/fpls.2019.01129.
[64] De Andrés J, Lorca P, De Cos Juez FJ, Sánchez-Lasheras F. Bankruptcy forecasting: A hybrid approach using fuzzy c-means clustering and multivariate adaptive regression splines (MARS). Expert Syst Appl 2011;38:1866–75. https://doi.org/10.1016/j.eswa.2010.07.117.
[65] Deconinck E, Coomans D, Vander HY. Exploration of linear modelling techniques and their combination with multivariate adaptive regression splines to predict gastro-intestinal absorption of drugs. J Pharm Biomed Anal 2007;43:119–30. https://doi.org/10.1016/j.jpba.2006.06.022.
[66] Nayana BM, Kumar KR, Chesneau C. Wheat Yield Prediction in India Using Principal Component Analysis-Multivariate Adaptive Regression Splines (PCA-MARS). AgriEngineering 2022;4:461–74. https://doi.org/10.3390/agriengineering4020030.
[67] Resende MDV, Resende MFR, Sansaloni CP, Petroli CD, Missiaggia AA, Aguiar AM, et al. Genomic selection for growth and wood quality in Eucalyptus: Capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytol 2012;194:116–28. https://doi.org/10.1111/j.1469-8137.2011.04038.x.