DOI: http://dx.doi.org/10.1590/1678-992X-2020-0021

ISSN 1678-992X

SCIENTIA AGRICOLA

Research Article

Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms

Ithalo Coelho de Sousa[1]*, Moysés Nascimento[1], Gabi Nunes Silva[2], Ana Carolina Campana Nascimento[1], Cosme Damião Cruz[3], Fabyano Fonseca e Silva[4], Dênia Pires de Almeida[5], Kátia Nogueira Pestana[6], Camila Ferreira Azevedo[1], Laércio Zambolim[7], Eveline Teixeira Caixeta[8]

[1]Universidade Federal de Viçosa – Depto. de Estatística, Av. Peter Henry Rolfs, s/n – 36570-000 – Viçosa, MG – Brasil.
[2]Universidade Federal de Rondônia – Depto. de Matemática e Estatística, R. Rio Amazonas, 351 – 76900-726 – Ji-Paraná, RO – Brasil.
[3]Universidade Federal de Viçosa – Depto. de Biologia Geral, Av. Peter Henry Rolfs, s/n – 36570-000 – Viçosa, MG – Brasil.
[4]Universidade Federal de Viçosa – Depto. de Zootecnia, Av. Peter Henry Rolfs, s/n – 36570-000 – Viçosa, MG – Brasil.
[5]Universidade Federal de Viçosa/Instituto de Biotecnologia Aplicada à Agropecuária – BioCafé, Av. Peter Henry Rolfs, s/n – 36570-000 – Viçosa, MG – Brasil.
[6]Embrapa Mandioca e Fruticultura, R. Embrapa, s/n – 44380-000 – Cruz das Almas, BA – Brasil.
[7]Universidade Federal de Viçosa – Depto. de Fitopatologia, Av. Peter Henry Rolfs, s/n – 36570-000 – Viçosa, MG – Brasil.
[8]Embrapa Café, Av. w3 Norte (final) – 70770-901 – Brasília, DF – Brasil.
*Corresponding author <ithalo.coelho@gmail.com>

Edited by: Leonado Oliveira Medici

Received January 27, 2020
Accepted April 24, 2020

ABSTRACT: Genomic selection (GS) emphasizes the simultaneous prediction of the genetic effects of thousands of scattered markers over the genome. Several statistical methodologies have been used in GS for the prediction of genetic merit. In general, such methodologies require certain assumptions about the data, such as the normality of the distribution of phenotypic values. To circumvent the non-normality of phenotypic values, the literature suggests the use of Bayesian Generalized Linear Regression (GBLASSO). Another alternative is the models based on machine learning, represented by methodologies such as Artificial Neural Networks (ANN), Decision Trees (DT) and related possible refinements such as Bagging, Random Forest and Boosting. This study aimed to use DT and its refinements for predicting resistance to orange rust in Arabica coffee. Additionally, DT and its refinements were used to identify the importance of markers related to the characteristic of interest. The results were compared with those from GBLASSO and ANN. Data on coffee rust resistance of 245 Arabica coffee plants genotyped for 137 markers were used. The DT refinements presented equal or inferior values of Apparent Error Rate compared to those obtained by DT, GBLASSO, and ANN. Moreover, DT refinements were able to identify important markers for the characteristic of interest. Out of 14 of the most important markers analyzed in each methodology, 9.3 markers on average were in regions of quantitative trait loci (QTLs) related to resistance to disease listed in the literature.

Keywords: *Hemileia vastatrix*, statistical learning, plant breeding, artificial intelligence

Introduction

Coffea arabica, economically speaking the most important coffee species, is responsible for approximately 60 % of Brazil's coffee production in 2017 (ICO, 2018). A number of diseases can affect coffee cultivation, amongst which coffee leaf rust (CLR), caused by the fungus *Hemileia vastatrix*, is the main disease that causes worldwide harm. The symptoms of CLR can be seen on the lower face of the leaf surface, in the form of large orange spore masses, leading to premature leaf fall and can reduce crop yield by up to 35 % (Talinhas, et al., 2016). To minimize the damage caused by the disease, resistant cultivars have been developed through coffee genetic breeding programs and in order to increase the efficiency and accuracy in the selection of improved coffee crops, molecular tools have been incorporated into breeding programs (Barka et al., 2017).

Genomic Selection (GS) is used to accelerate the breeding process. The models used in GS for predicting GEBV, for example, RR-BLUP (Meuwissen et al., 2001) and BLASSO (Park and Casella, 2008), are based on the assumption of the normality of phenotypic values. In order to overcome this limitation, Pérez and Campos (2014) proposed the use of Bayesian Generalized Linear Regression (BGLR), allowing for the use of GS for continuous and discrete models. Although useful, the presence of complicating factors such as epistasis and dominance make it difficult to use the usual models of GS, once their effects have been established as the prior in the model.

Another approach to problem prediction is the use of machine learning algorithms, such as Artificial Neural Network (ANN) (Adetiba and Olugbara, 2015), Decision Trees (DT) and related possible refinements such as Bagging, Random Forest and Boosting (González-Recio and Forni, 2011; Ogutu et al., 2011). These algorithms make no assumptions about the model. This feature of statistical learning allows for the capture of complicated factors such as epistasis and dominance in prediction models since it is not necessary to know a

priori if the data have these effects and do not require any assumptions about the distribution of phenotypic values.

The aim of this paper was to use the DT and its refinements (Bagging, Random Forest and Boosting) for predicting the rust resistance of Arabica coffee. Furthermore, the estimates obtained were compared with those obtained by BGLR and ANN. Finally, the importance of markers to rust resistance was evaluated.

Materials and Methods

Genotypes

The population evaluated consisted of two progenitors, the Timor Hybrid UFV 443-03 (resistant to rust), cultivar Catuai amarelo IAC 64 (UFV 2148-57) (susceptible to rust), the hybrid F_1 and 245 F_2 plants. The UFV 443-03 is an important rust resistant source used in breeding programs, and Pestana et al. (2015) identified at least two major genes associated with resistance coffee to three isolates of *H. vastatrix*.

Inoculation of plants

The experiments were carried out in Viçosa, MG, Brazil (20°45′37″ S, 42°52′4″ W, altitude of 648 m). The 245 F_2 plants were inoculated with uredospore of pathotype 001 of *H. vastatrix*. The inoculation was conducted according to the methodology described by Capucho et al. (2009). The evaluation of symptoms (phenotyping) was performed during May, June and Aug 2009, comprising the first, second and third repetition, respectively. The score scale described by Tamayo et al. (1995) was used. The highest score obtained in the tree repetition was used for the next analysis. Plants that received score one (no symptoms) and score two (small chlorotic injuries) were considered resistant. If attributed a score of three, four, five or six, the phenotypes were considered susceptible. Score three corresponded to a plant that contained large injuries without sporulation, score four to large chlorotic injuries with small sporulation occupying less than 25 % of the area, score five to injuries with sporulation occupying from 25 to 50 % of the area and score six to injuries with sporulation occupying more than 50 % of the area (Figure 1).

Genotyping of plants

Genotyping was carried out in the years 2010, 2011 and 2012, with 137 markers (74 AFLP, 58 SSR, 4 RAPD, and 1 specific primer) (Pestana et al., 2015). The marker data for each individual were coded for analyses of genomic selection. For dominant markers linked in the coupling phase to a resistant allele of the progenitor Timor Hybrid UFV 443-03, the code –1 and 1 were attributed to the presence and absence of the band, respectively. For dominant markers in repulsion (allele from the susceptible progenitor Catuai amarelo UFV 2148-57) 1 and –1 were also assigned to the presence and absence of the band, respectively. The codominant markers were coded with 0 for heterozygote, –1 for bands from the resistant progenitor and 1 for bands from the susceptible progenitor (Silva et al., 2017). The genotype data quality control used is described in Pestana et al. (2015).

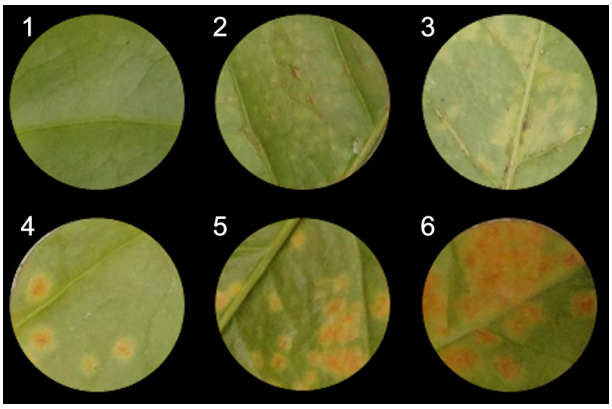

Figure 1 – Scale of notes for the evaluation of coffee resistance to *H. vastatrix*.

In genetic mapping studies, the distance between loci pairs can be divided into four classes, tightly linked (< 1 cM), moderately linked (1-10 cM), loosely linked (11-20 cM) and unlinked (> 20 cM) (Remington et al., 2001; Maccaferri et al., 2005; Jun et al., 2008). Based on Pestana et al. (2015) map, we considered the markers with a distance less than 10 cM (tightly and moderately linked) as being linked to the QTL region.

Classification tree (CT) and its refinements

To construct a classification tree, the objective is to obtain regions R_1, R_2, ..., R_M that minimize the Gini index as given by James et al. (2013):

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}),$$

where $\hat{p}_{mk}$ represents the proportions of observations in the m^{th} region belonging to the k^{th} class. The Gini index decreased according to the tree growth that was produced by recursive binary splitting. To avoid the model over fitting, it is recommended that first, no region have more than five observations and second, prune the tree using the cost complexity pruning given by $R_\alpha(T) = R(T) + \alpha|T|$ where $R(T)$ is the error rate, $|T|$ the number of regions and α the tuning parameter (Hastie et al., 2009). Generally a single tree does not have good predictive accuracy when compared with other approaches. In order to increase the predictive performance of the model bootstrap aggregation (bagging), random forest and boosting should be used.

The bootstrap aggregation (bagging) consists of obtaining B samples with replacement (size equal to N) from the data set, thus obtaining B models ($\hat{f}^1(x), \hat{f}^2(x),..., \hat{f}^B(x)$) that will be used as individual classifiers.

A new individual will be classified in the most common class among the predictions of the B individual classifiers. The random forest (RF) follows the same idea of bagging, but uses a smaller number of predictive variables in each split. According to James et al. (2013), RF results in a process of "decorrelating" the generated trees, improving even more the accuracy of predictions. Finally, unlike bagging that creates independent trees, boosting creates a tree sequentially using information from past trees. The boosting classifier $H(x) = \Sigma_t \alpha_t h_t(x)$ that seeks to minimize functional loss through the optimization of the scalar α_t (importance assigned to $h_t(x)$) and of the individual classifier $h_t(x)$ in each iteration t (Freund and Schapire, 1999). The individual classifiers $h_t(x)$ have low classification power, but when used with the $H(x)$ ensemble presented good results (Martins et al., 2009).

Artificial Neural Network (ANN)

The ANN architecture used in this work has a hidden layer only and uses the backpropagation as a learning algorithm (Rumelhart and McClelland, 1986). The network structure considers 137 markers as inputs, a hidden layer, and the outputs that predict resistance or susceptibility of the leaf to rust which is shown in Figure 1.

The neurons W_m are generated by a linear combination of the input variables M_p (markers). Finally the output variable Y_k is generated as a function of linear combination of the neurons W_m as follows:

$$W_m = \sigma(\alpha_{0m} + \alpha_m^T X), m = 1,2,...M$$

$$T_k = \beta_{0k} + \beta_k^T W, k = 1,...,K$$

$$Y_k = g_k(T),\ k = 1, ..., K$$

in which $W = (W_1, W_2, ..., W_m)$, $T = T_1, T_2, ..., T_k$.

The activation function used in ANN was the sigmoid, $\sigma(\upsilon) = 1/(1 + e^{-\upsilon})$, and the output function the softmax,

$$g_k(T) = \frac{e^T k}{\sum_l^K e^T l}.$$

The weights $(\alpha_{0m}, \alpha_m;\ m = 1, 2, ..., M)$ and $(\beta_{0k}, \beta_k;\ k = 1, 2, ..., K)$ are unknown parameters of the network responsible for adjustments to the ANN model to training set. As a measure of adjustment we use the cross-entropy $R(\theta) = -\sum_{i=1}^N \sum_{k=1}^K y_{ik} \log f_k(x_i)$ with backpropagation responsible for its minimization. The numbers of neurons in the hidden layer were chosen based on considering a maximum error of 15 %, at most, for the validation test.

Bayesian Generalized Linear Regression

The genomic selection method based on Bayesian generalized linear regression (Pérez and de los Campos, 2014) was also used for predicting GEBV. The model is given by:

$$\hat{Y} = \mu + X_1\beta_1 + ... + X_j\beta_j + \mu_1 + ... + \mu_q$$

where μ is the intercept, X_i the predictor matrices, $X_j = \{x_{ijk}\}$, β_{jk} the vectors of effects associated with the columns of X_j and $\mu_q = \{\mu_{q1} ..., \mu_{qn}\}$ the vectors of random effects. In this study, since the phenotypic values present a categorical distribution (rust resistance), the probit link was used (Pérez and Campos, 2014) where the probability of each category is linked to the linear predictor according to the following link function:

$$P(y_i = k) = \Phi(\eta_i - \gamma_k) - \Phi(\eta_i - \gamma_{k-1})$$

where $\Phi(.)$ is the standard normal cumulative distribution, η_i the linear predictor and γ_k the threshold parameters, where $\gamma_0 = -\infty$, $\gamma_k \geq \gamma_{k-1}$, $\gamma_k = \infty$.

Training and validation sets

The data set was divided into two parts: training set and validation set. The training set was kept with the same individuals for modeling all the methodologies, composed of 70 % of each class (172 observations), taken at random, while the remaining 30 % (73 observations) were used in the validation set. In the literature, the percentages used in the training set vary between 60 and 90 % as seen in Gianola et al. (2011) and González-Camacho et al. (2012).

Marker selection

The most important markers are those that have a greater influence on the studied trait, whereby all individuals in the construction of the models are used to select such markers with greater precision. In GBLASSO (Generalized Bayesian Lasso) the markers with the highest regression coefficients in absolute values were defined as the most important markers. In the methodologies of Classification Tree and Prune, the most important markers are the ones that were used in the split of the first nodes. In bagging and random forest, we assumed as the most important markers those that on average influenced more in the reduction of the Gini index. In boosting, the most important markers are those that have more relevance in separating the observations of one class from the others.

In ANN, the acquiring of the most important markers is through the construction of a new ANN after canceling the effect of each marker individually. The most important marker will be the one that, after its annulment, presents a higher apparent error rate - APER (Silva et al., 2017). Among the 100 ANN used in the genomic selection process the network used to determine the importance of the markers was the one that presented a lower APER.

We selected the 10 % most important markers (14 first markers) in each methodology and compared with results taken from the literature (Pestana et al., 2015), to verify if the methodologies were able to indicate markers that are associated with the studied trait.

Comparison of methodologies

To compare the methodologies, we repeated the entire process 100 times. We used the average computational cost and the APER confidence interval obtained through the validation sets. We also estimated the accuracy of each methodology, given by

$$r_{\hat{y}g} = \frac{r_{\hat{y}y}}{h},$$

where $r_{\hat{y}g}$ is an accuracy estimator proposed by Legarra et al. (2008) and Hayes et al. (2009), $r_{\hat{y}y}$ the predictive ability represented by the phi correlation (Warrens, 2008) between the phenotype y and the predicted genomic breeding values $\hat{y}$, g the true breeding value and h^2 the heritability (Legarra et al., 2008). The heritability for coffee resistance to *H. vastatrix* pathotype 001 considering the phenotypic data was 0.50 (Pestana et al., 2015).

We used Cohen's Kappa coefficient proposed by Cohen (1960) to analyze the agreement between the methodologies both in classification of the individuals (using the training set) and in the identification of markers (the 10 most important in each methodology). The coefficient of Cohen's Kappa is given by:

$$kappa \frac{NAO - NAEC}{NOA - NAEC}$$

where *NAO* is the degree of agreement observed, *NAEC* the degree of agreement expected by chance and *NOA* the number of observations analyzed (Resende et al., 2014).

Computational aspects

Data analysis was carried out on a computer with a 3.40GHz core i7 processor and 16GB of RAM using the R software 3.40 program. Prediction was facilitated through, the nnet function, part of the nnet package in ANN which was chosen in recognition of the limitation of APER to a maximum of 15 % or a maximum of 5000 iterations in the validation set. The BGLR function, belonging to the BGLR package, was used to estimate the Bayesian generalized models, from 100,000 iterations with the first 20,000 observations being discarded (burn-in) and values were saved at every 10 observations (thinning). To construct the classification tree we used the tree function belonging to the tree package. As part of the randomForest package the randomForest function was used to construct the model of the Bagging and the Random Forest. Finally, the gbm function from the gbm package was used to construct the boosting.

Results

The average apparent error rate values obtained by adjusting the models under study (DT, decision tree with prune – DTP, Bagging, Random Forest, Boosting, ANN, GBLASSO) ranged from 19.5 % to 24.9 %. Specifically, the lower average APER was obtained by taking into account the adjustment of boosting (19.5 %), which does not present a significant difference when compared with those obtained by adjusting ANN with one hidden layer (19.6 %), bagging (19.7 %), random forest (20.4 %) and DTP (21.1 %). According to the 95 % confidence interval, most of the methodologies, except the DT (24.9 %), presented better results (lower APER values) when compared to those obtained by GBLASSO (22.7 %) (Figure 2). Average accuracy ranged from 30.6 % to 60.7 %. Overall all the methodologies outperformed GBLASSO, since their accuracy values were higher than those presented by GBLASSO, according to the 95 % confidence interval.

When comparing the computational cost of the techniques to GBLASSO (traditional GWS methodology), only the ANN require more computational time, being 9.20 times slower. The DT was the model with the lowest computational cost, being 2850 times faster than GBLASSO.

According to Landis and Koch (1977), a Cohen's Kappa coefficient greater than 0.4 can be considered from moderate to an excellent agreement estimate between the methodologies. In general, the mean Cohen's Kappa coefficient presented positive and high values in the classification of genotypes. Comparing the techniques to GBLASSO, the methodology with the lowest percentage agreement was DT (70.3 %), while RF was the most similar methodology as pertains to the classification provided by GBLASSO (88.9 %) (Table 1). Considering all the methodologies evaluated, the highest percentage agreement was observed in the results obtained through RF and boosting (93.4 %).

Among the fourteen most important markers indicated by each methodology (Table 1), the highest percent of Cohen's Kappa agreement (76.1 %) was between GBLASSO and boosting, which presented eleven markers in common. The ANN and DT presented the lowest percentage agreement (4.5 %), presenting two markers in common.

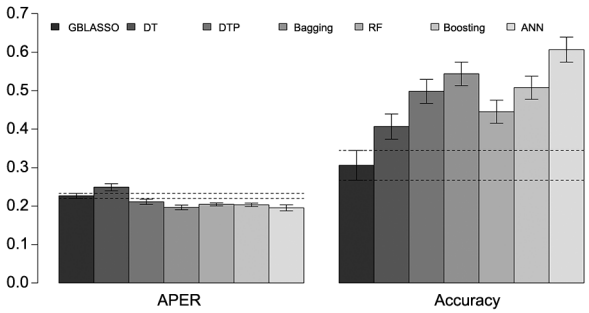

Figure 2 – Apparent error rate (APER) and accuracy at 95 % confidence interval (CI) obtained by each adjusted model. The dotted lines highlight the limits of CI obtained by GBLASSO. ANN = Artificial Neural Network; GBLASSO = Generalized Bayesian Lasso; DT = Decision Tree; DTP = Decision Tree with Prune; RF = Random Forest.

Table 1 – Average time in seconds (diagonal) with the standard error in parentheses, average of Cohen's Kappa coefficient between the classifications (above the diagonal) and Cohen's Kappa coefficient between the most important markers of each methodology (below the diagonal).

Models	ANN	GBLASSO	DT	DTP	Bagging	RF	Boosting
ANN	**355.92 (22.10)**	76.4 %	69.0 %	73.5 %	78.7 %	77.8 %	81.1 %
GBLASSO	44.3 %	**38.68 (5.55)**	70.3 %	77.2 %	83.1 %	88.9 %	86.4 %
DT	4.5 %	20.4 %	**0.01 (0.01)**	83.1 %	78.3 %	73.5 %	73.7 %
DTP	0.0 %*	23.0 %	23.0 %	**0.14 (0.02)**	85.7 %	81.0 %	81.7 %
Bagging	28.4 %	44.3 %	20.4 %	23.0 %	**4.10 (0.22)**	89.4 %	89.1 %
RF	28.4 %	52.3 %	12.5 %	23.0 %	68.2 %	**3.21 (0.12)**	93.4 %
Boosting	44.3 %	76.1 %	20.4 %	23.0 %	68.2 %	68.2 %	**13.63 (0.33)**

ANN = artificial neural network; GBLASSO = generalized Bayesian Lasso; DT = decision tree; DTP = decision tree with prune; RF = random forest; *Coefficient lower than zero.

Table 2 – Fourteen most important markers indicated by each methodology.

ANN	GBLASSO		DT	DTP	Bagging		RF		Boosting	
M	M	Importance	M	M	M	Importance	M	Importance	M	Importance
2	**43**	0.037619	43	43	**97**	0.129982	**43**	0.046065	**97**	0.13016
24	**73**	0.032772	97	97	**43**	0.083155	**97**	0.042635	**43**	0.093231
25	**61**	0.031536	7	-	**61**	0.043129	**61**	0.039375	**61**	0.066988
63	**97**	0.030817	86	-	**47**	0.030975	**64**	0.024595	**47**	0.037109
67	**11**	0.029281	84	-	**73**	0.030622	**12**	0.02182	**55**	0.031592
77	**12**	0.024199	34	-	**12**	0.023706	**11**	0.018815	**12**	0.029706
125	**29**	0.022077	73	-	**55**	0.020561	**21**	0.017176	**29**	0.027303
26	**24**	0.02205	94	-	115	0.020464	59	0.015569	**11**	0.024407
29	**55**	0.021031	118	-	19	0.0204	**29**	0.014864	**128**	0.020311
55	**128**	0.020407	52	-	101	0.014927	**47**	0.013759	**73**	0.020216
61	67	0.019127	74	-	**29**	0.014906	**55**	0.01373	**64**	0.01853
64	**21**	0.017907	61	-	**13**	0.013909	82	0.013447	**24**	0.017379
73	**107**	0.017566	20	-	**64**	0.013803	**85**	0.013072	**85**	0.01679
11	68	0.017554	23	-	**85**	0.012589	**13**	0.012153	**107**	0.016095

Markers in bold are those which appeared two or more times among GBLASSO, Bagging, RF and Boosting. ANN = artificial neural network; GBLASSO = generalized Bayesian Lasso; DT = decision tree; DTP = decision tree with prune; RF = random forest; M = marker; underlined marker, marker in a QTL region according to Pestana et al., 2015. The importance obtained for the markers is calculated using different criteria in each methodology.

To obtain marker candidates to support the selection process, we first considered those indicated as most important for at least four methods. The approaches that have more markers in common were GBLASSO, Bagging, Random Forest and Boosting. Next, we analyzed four selected approaches and chose the markers which appeared two or more times selecting these as the most important markers (those in bold in Table 2). The markers 11, 12, 13, 21, 24, 29, 43, 47, 55, 61, 64, 73, 85, 97, 107 and 128 were considered jointly most important in at least two approaches. The candidate markers can be used directly for selection resistant cultivars. When using the candidate markers individually to classify the phenotype, the APER varied between 24.1 % and 49.4 % (Table 3). The marker with the lower APER (97) is one of the best markers according to Table 2.

The joint importance of these markers was also quantified, showing that in GBLASSO they are responsible for 30.7 % of the variability of the leaf rust resistance. The Bagging, RF and Boosting are responsible for 41.7 %, 27.8 % and 55.0 %, respectively (Table 2).

Table 3 – Apparent error rate, considering only the candidate markers individually.

Marker	APER	Marker	APER
11*	0.3959	55*	0.2857
12*	0.2531	61*	0.2531
13**	0.4939	64*	0.2694
21*	0.2531	73*	0.4531
24**	0.4898	85*	0.2490
29*	0.4898	97*	0.2408
43*	0.2490	107*	0.3184
47*	0.4857	128*	0.4571
Average APER	0.3523		

APER = apparent error rate; *Presence of the allele of the resistance progenitor; **Presence of the allele of the susceptible progenitor.

We compared our statistical approach (Table 2) to the QTL mapping of Pestana et al. (2015) QTL. Both studies used the same population and markers. According to the Pestana et al. (2015) map, four QTLs associated with coffee resistance to *H. vastatrix* pathotype 001 were allocated in a tree linkage group (LG), LG 2 (2 QTL),

LG 3 and LG 10. Several markers coincided with our approach and the markers flanking the QTL. In Table 2, the underlined markers are those that are in any of the three chromosomes with QTL.

Discussion

Machine learning algorithms such as those based on Artificial Neural Network (ANN), Decision Tree (DT) and its refinements Bagging, Random Forest (RF) and Boosting were used and tested to predict the genetic resistance of coffee to rust. The results obtained were compared with those coming from the traditional approach to GS studies in which the trait does not present a normal distribution, the Bayesian generalized linear regression (Figure 2). Finally, the sixteen most important markers were selected according to each methodology, that is, those markers that exerted greater influence in rust resistance (Table 2).

The use of methods based on machine learning (ML) for predicting genetic resistance to leaf rust in *Coffea arabica* were efficient, since all models (except DT) presented lower APER values and higher accuracy values when compared with the results obtained by GBLASSO (Figure 2). Furthermore, the accuracy values provided by bagging and ANN were higher than the estimated heritability of the trait ($h^2 = 0.50$) presented by Pestana et al. (2015).

The worst performance of DT when compared with its refinements can be attributed to this methodology which suffered from high variance in terms of prediction (James et al., 2013). Hastie et al. (2009) emphasized that the DT's low predictive accuracy can be improved by the use of ensemble methods such as bagging, random forest and boosting (Breiman, 2001). These strategies combine multiple DT to reduce the variability.

According to the APER and Accuracy values (Figure 2) bagging presented equal or better results than those observed in RF which uses a number of predictors in the split of each node less ($m \approx \sqrt{p}$) than the bagging. This modification in the numbers of predictors that occurs in RF aims to break the correlation between the trees constructed in each iteration in order to increase the predictive capacity of the model (Hastie et al., 2009). According to James et al. (2013), using a restricted number of predictors in RF, will be advantageous if there are many correlated predictor variables. In this study, only 1.6 % of the markers presented a strong correlation between themselves, which explains why RF does not improve results in relation to bagging. Additionally, since rust resistance is an oligogenic trait (Bettencourt and Rodrigues Jr., 1988), the framework of RF, whereby the division of the nodes in the DT is performed using a small random number of markers, it is possible that certain nodes can have chosen only markers that are not associated with the trait, explaining the lower performance of RF compared with bagging.

The methods based on ML were used under the GS focus in many studies, such as that adopted by Ogutu et al. (2011). In this study, the authors compared the predictive ability of RF, boosting and support vector machine (SVM) to predict Genomic Estimated Breeding Values (GEBV) through simulated data, and confirmed this to be the best performance to the detriment of the other two methods. Gianola et al. (2014) evaluated and proved that the GBLUP (Genomic Best Linear Prediction) predictive capacity can be improved by bagging. They verified that the use of bagging in the GBLUP, in addition to improving the predictive performance of the method, made it more robust in relation to the over fitting of the data. Such an approach was also used with success in the studies of Abdollahi-Arpanahi et al. (2015) and Mehrban et al. (2017), in which they proved the efficiency of the use of bagging together with the GBLUP in the prediction of GEBV for chicken and bulls of the Jersey breed respectively. Ornella et al. (2014) used many classification algorithms for the genomic prediction of maize and verified that those methodologies are a promising alternative for GS in the breeding of plants. Different from these studies, the methodologies evaluated in this study were compared with the results of BGLR, which contemplate the qualitative nature of the trait evaluated in modeling. Out of the several methods of machine learning, only ANNs were compared with the BGLR in the presence of a qualitative nature (Silva et al., 2017), as the ANNs are more efficient in predicting rust resistance in *Coffea arabica*.

The high percentage of agreement between the methodologies may indicate that the study trait (leaf rust resistance) does not present complicating factors to modeling such as the presence of great dominance and epistasis, which require the use of more complex models (Table 1). Despite the similarity, it should be emphasized that the models based on ML are flexible and do not depend on an a priori specification adjustment of the model, which makes it easier to contemplate such complicated factors (Silva et al., 2017). In terms of computational cost, the methods based in ML (except ANN) present a great advantage when compared to GBLASSO, since GBLASSO uses Markov chain Monte Carlo methods that require the construction of large chains (Table 1).

Candidate markers have been used to select leaf rust resistance genotypes. Diola et al. (2011) identified molecular markers linked to the SH Gene, which is one of the major genes that confers resistance to coffee rust (Brito et al., 2010). Alkimim et al. (2017) used molecular markers to identify coffee plants carrying the genes SH3 and other SH genes which also confer resistance to coffee rust. Once the phenotype has been considered dichotomous (resistant and susceptible) to fit the machine learning models, the candidate markers can be used directly to select resistant cultivars. For example, considering marker 97 as a candidate, the APER was

equal to 24.1 %. Although interesting, the joint use of markers outperforms the use of markers individually.

According to Pestana et al. (2015), four QTL regions are associated with the rust resistance pathotype 001, two in LG 2, one in LG 3 and one in LG 10. Mean disease severity reduction associated with the presence of these QTLs range from 17.8 % (QTL 3 – LG 3) to 31.0 % (QTL 2 – LG 2), both the QTL 3 with less effect and the QTL 3 with higher effect. These QTLs can be used under a marker-assisted selection (MAS) approach, to the introgression of these loci into new cultivars with durable resistance of *H. vastatrix* (Pestana et al., 2015). Among the selected markers in the ANN methodology, two markers were located close to the QTL of the LG 3, two in the LG 10 and one in the LG 2. The other markers do not link to any LG or are in other LGs with no QTL.

The DT and DTP were also not efficient, as they identified only one or two markers linked to the QTL of LG 10, respectively (Table 2). The DTP is not an interesting strategy for obtaining such information, since the prune process removes several markers from the prediction process. The other methodologies, GBLASSO, bagging, RF and boosting allow for the selection of several markers that are near the four QTLs previously identified. The RF was the methodology that identified more markers linked to the QTLs, six markers (12, 13, 21, 29, 47 and 59) linked to the two QTLs of LG 2 selected, three markers (43, 61 and 64) linked to the QTL of LG 10, and two more markers (55 and 82) linked to the QTL of LG 3. The GBLASSO, bagging and boosting, identified 7, 9 and 8 markers linked to the LG with regions of QTLs, respectively, as we can see in Table 2.

Although marker 97 is in the LG 11 that does not have a QTL region (Pestana et al., 2015), it was identified as an important marker in six methodologies. LG 11 is a small LG and with the saturation of the map it can be linked to a LG with a QTL region. Another hypothesis is the detection of a new QTL region that was not detected by Pestana et al. (2015). Both hypotheses are related to the increase in map saturation. Different from QTL mapping, which requires a higher mapping saturation, methodologies based on GS may help to indicate important regions in the genome which can contain a QTL.

These results indicate that when the refinements (bagging, RF and boosting) are applied to the DT, provided good alternatives for the determination of the markers are associated with a study trait, they have a lower computational cost when compared to GBLASSO (Table 1). By contrast, the results obtained by ANN indicate that the strategy used in this study to determine the marker importance through this approach is not efficient. The results obtained from the data of this study show the superiority of bagging and boosting in both the GEBV and the determination of the importance of markers (Table 2).

Conclusions

Evaluations of the APER and the accuracy of the prediction of leaf rust resistance confirmed that all methodologies showed greater effectiveness than GBLASSO (except DT in APER), incurring an even lower computational cost (except ANN). The DT refinements were capable of detecting markers near regions where QTLs were identified for the trait study.

Acknowledgments

We thank the Foundation for Research Support of the state of Minas Gerais State (FAPEMIG), the Coordination for the Improvement of Higher Level Personnel (CAPES), and the Brazilian National Council for Scientific and Technological Development (CNPq) for financial support.

Authors' Contributions

Conceptualization: Sousa, I.C.; Nascimento, M.; Caixeta, E.T. **Data acquisition:** Almeida, D.P.; Pestana, K.N.; Zambolim, L.; Caixeta, E.T. **Data analysis:** Sousa, I.C.; Nascimento, M.; Caixeta, E.T. **Design of methodology:** Sousa, I.C.; Nascimento, M.; Cruz, C.D.; Silva, F.F. **Software development:** Sousa, I.C.; Nascimento, M.; Silva, G.N. **Writing and editing:** Sousa, I.C.; Nascimento, M.; Nascimento, A.C.C.; Cruz, C.D.; Silva, F.F.; Azevedo, C.F.; Caixeta, E.T.

References

Abdollahi-Arpanahi, R.; Morota, G.; Valente, B.D.; Kranis, A.; Rosa, G.J.M.; Gianola, D. 2015. Assessment of bagging GBLUP for whole-genome prediction of broiler chicken traits. Journal of Animal Breeding and Genetics 132: 218-228.

Adetiba, E.; Olugbara, O.O. 2015. Lung cancer prediction using neural network ensemble with histogram of oriented gradient genomic features. The Scientific World Journal 2015: 1-17.

Alkimim, E.R.; Caixeta, E.T.; Sousa, T.V.; Pereira, A.A.; Oliveira, A.C.B.; Zambolim, L.; Sakiyamam, N.S. 2017. Marker-assisted selection provides arabica coffee with genes from other coffea species targeting on multiple resistance to rust and coffee berry disease. Molecular Breeding 37: 1-10.

Barka, G.D.; Caixeta, E.T.; Almeida, R.F.; Alvarenga, S.M.; Zambolim, L. 2017. Differential expression of molecular rust resistance components have distinctive profiles in Coffea arabica – *Hemileia vastatrix* interactions. European Journal of Plant Pathology 149: 543-561.

Bettencourt, A.J.; Rodrigues Jr., C.J. 1988. Principles and practice of coffee breeding for resistance to rust and other diseases. p. 199-234. In: Clarke, R.J.; Macrae, R., eds. Coffee, 4. Elsevier, Barking, UK.

Breiman, L. 2001. Random forests. Machine Learning 45: 5-32.

Brito, G.G.; Caixeta, E.T.; Gallina, A.P.; Zambolim, E.M.; Zambolim, L.; Diola, V.; Loureiro, M.E. 2010. Inheritance of coffee leaf rust resistance and identification of AFLP markers linked to the resistance genex. Euphytica 173: 255-264.

Capucho, A.S.; Caixeta, E.T.; Zambolim, E.M.; Zambolim, L. 2009. Inheritance of coffee leaf rust resistance in Timor Hybrid UFV 443-03. Pesquisa Agropecuária Brasileira 44: 276-282 (in Portuguese, with abstract in English).

Cohen, J.A. 1960. Coefficient of agreement for nominal scales. Educational and Pshychological Measurement 20: 37-46.

Diola, V.; Brito, G.G.; Caixeta, E.T.; Maciel-Zambolim, E.; Sakiyama, N.S.; Loureiro, M.E. 2011. High-density genetic mapping for coffee leaf rust resistance. Tree Genet Genomes 7: 1199-1208.

Freund, Y.; Schapire, R.E. 1999. A short introduction to boosting. Journal of Japanese Society for Artificial Intelligence 14: 771-780.

Gianola, D.; Okut, H.; Weigel, K.A.; Rosa, G.J.M. 2011. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. BMC Genetics 12: 1-14.

Gianola, D.; Weigel, K.A.; Krämer, N.; Stella, A.; Schön, C-C. 2014. Enhancing genome-enabled prediction by bagging genomic BLUP. PlosOne: e91693.

González-Camacho, J.M.; Campos, G.; Pérez, P.; Gianola, D.; Cairns, J.E.; Mahuku, G.; Babu, R.; Crossa, J. 2012. Genome-enabled prediction of genetics values using radial basis function neural networks. Theoretical and Applied Genetics 125: 759-771.

González-Recio, O.; Forni, S. 2011. Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. Genetics Selection Evolution 43: 1-12.

Hastie, T.; Tibshirani, R.; Friedman, J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2ed. Springer, New York, NY, USA.

Hayes, B.J.; Bowman, P.J.; Chamberlain, A.J.; Doddard, M.E. 2009. Invited review: genomic selection in dairly cattle; progress and challenges. Journal of Dairy Science 92: 433-443.

International Coffee Organization [ICO]. 2018. Coffee market report. Available at: http://www.ico.org/documents/cy2017-18/cmr-0818-e.pdf [Accessed Oct 20, 2018]

James, G.; Witten, D.; Hastie, T.; Tibshirani, R. 2013. An Introduction to Statistical Learning: with Applications in R. Springer, New York, NY, USA.

Jun, T.H.; Van, K.; Kim, M.Y.; Lee, S.H.; Walker, D.R. 2008. Association analysis using SSR markers to find QTL for seed protein content in soybean. Euphytica 162: 179-191.

Landis, J.R.; Koch, G.G. 1977. The measurement of observer agreement for categorical data. Biometrics 33: 159-174.

Legarra, A.; Robert-Granié, C.; Manfredi, E.; Elsen, J.M. 2008. Performance of genomic selection in mice. Genetics 180: 611-618.

Maccaferri, M.; Sanguineti, M.C.; Noli, E.; Tuberosa, R. 2005. Population structure and long-range linkage disequilibrium in a durum wheat elite collection. Molecular Breeding 15: 271-289.

Martins, R.; Pina, P.; Marques, J.S.; Silveira, M. 2009. Crater detection by a Boosting approach. IEEE Geoscience and Remote Sensing Letters 6: 127-131.

Mehrban, H.; Lee, D.H.; Moradi, M.H.; IlCho, C.; Naserkheil, M.; Ibáñez-Escriche, N. 2017. Predictive performance of genomic selection methods for carcass traits in Hanwoo beef cattle: impacts of the genetic architecture. Genetics Selection Evolution 49: 1-13.

Meuwissen, T.H.E.; Hayes, B.J.; Goddard, M.E. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819-1829.

Ogutu, J.O.; Piepho, H-P.; Schulz-Streeck, T. 2011. A comparison of random forests, boosting and support vector machines for genomic selection. In: Szydlowski, M., ed. BMC Proceedings 5: 11.

Ornella, L.; Pérez, P.; Tapia, E.; González-Camacho, J.M.; Burgueño, J.; Zhang, X.; Singh, S.; Vicente, F.S.; Bonnett, D.; Dreisigacker, S.; Singh, R.; Long, N.; Crossa, J. 2014. Genomic-enabled prediction with classification algorithms. Heredity 112: 616-626.

Park, T.; Casella, G. 2008. The Bayesian Lasso. Journal of the American Statistical Association 103: 681-686.

Pérez, P.; Campos, G. 2014. Genome-wide regression and prediction with the BGLR statistical package. Genetics 198: 483-495.

Pestana, K.N.; Capucho, A.S.; Caixeta, E.T.; Almeida, D.P.; Zambolim, E.M.; Cruz, C.D.; Zambolim, L.; Pereira, A.A.; Oliveira, A.C.B.; Sakiyama, N.S. 2015. Inheritance study and linkage mapping of resistance loci to *Hemileia vastatrix* in Híbrido de Timor UFV 443-03. Tree Genetics & Genomes 11: 1-13.

Remington, D.L.; Thornsberry, J.M.; Matsuoka, Y.; Wilson, L.M.; Whitt, S.R.; Doebley, J.; Kresovich, S.; Goodman, M.M.; Buckler, E.S. 2001. Structure of linkage disequilibrium and phenotypic association in the maize genome. Proceedings of the National Academy of Sciences 98: 11479-11484. https://doi.org/10.1073/pnas.201394398

Resende, M.D.V.; Silva, F.F.; Azevedo, C.F. 2014. Mathematical Statistics, Biometric and Computational: Mixed, Multivariate, Categorical and Generalized (REML/BLUP) Models, Bayesian Inference, Random Regression, Genomic Selection, QTL-GWAS, Spatial and Temporal Statistics, Competition, Survival = 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. UFV, Viçosa, MG, Brazil (in Portuguese).

Rumelhart, D.E.; McClelland, J.L. 1986. Parallel Distributed Processing Explorations in the Microstructure of Cognition: Foundations. MIT Press, Cambridge, MA, USA.

Silva, G.N.; Nascimento, M.; Sant'anna, I.C.; Cruz, C.D.; Caixeta, E.T.; Carneiro, P.C.S.; Rosado, R.D.S.; Pestana, K.N.; Almeida, D.P.; Oliveira, M.S. 2017. Artificial neural networks compared with Bayesian generalized linear regression for leaf rust resistance prediction in Arabica coffee. Pesquisa Agropecuária Brasileira 41: 186-193.

Tamayo, P.J.; Vale, F.X.R.; Zambolim, L.; Chaves, G.M.; Pereira, A.A. 1995. Catimor resistance to rust and virulence of physiological races of *Hemileia vastatrix* Berk & Br = Resistência do Catimor à ferrugem e virulência de raças fisiológicas de *Hemileia vastatrix* Berk & Br. Fitopatologia Brasileira 20: 572-576 (in Portuguese).

Warrens, M.J. 2008. On association coefficients for 2 × 2 tables and properties that do not depend on the marginal distributions. Psychometria 73: 777-789.