ITHALO COELHO DE SOUSA

PREDIÇÃO GENÔMICA DA RESISTÊNCIA À FERRUGEM ALARANJADA EM CAFÉ ARÁBICA VIA ALGORITMOS DE APRENDIZAGEM DE MÁQUINA

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

VIÇOSA
MINAS GERAIS – BRASIL
2018

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
S725p
2018

Sousa, Ithalo Coelho de, 1990-
Predição genômica da resistência à ferrugem alaranjada em café arábica via algoritmos de aprendizagem de máquina / Ithalo Coelho de Sousa. – Viçosa, MG, 2018.
ix, 31 f. : il. (algumas color.) ; 29 cm.

Orientador: Moysés Nascimento.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f. 24-31.

1. Algorítimos genéticos. 2. Aprendizado do computador. 3. Decisão estatística. 4. Redes neurais. 5. Boosting (Algorítimo). 6. Ensacamento . I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 519.62

ITHALO COELHO DE SOUSA

PREDIÇÃO GENÔMICA DA RESISTÊNCIA À FERRUGEM ALARANJADA EM CAFÉ ARÁBICA VIA ALGORITMOS DE APRENDIZAGEM DE MÁQUINA

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

APROVADA: 26 de fevereiro de 2018.

Fabyano Fonseca e Silva

Cosme Damião Cruz
(Coorientador)

Moysés Nascimento
(Orientador)

DEDICO

À minha família.

AGRADECIMENTOS

Agradeço primeiramente a DEUS por ser meu refúgio em todos os momentos, principalmente nos mais difíceis.

Aos meus pais, Joselita Coelho Barros e Francisco Barbosa de Sousa Neto, pelo carinho, amizade, amor e por não medirem esforços para me dar a melhor educação possível.

À minha irmã Thalyta, pela companhia, conselhos, amizade, apoio e por se manter próxima a mim mesmo morando longe.

A toda a minha família, incluindo tios, primos e avós, que servem de exemplo para que eu continue determinado e serem compreensivos nos momentos em que não me faço presente, e por saber que estão sempre na torcida e fazem muita falta no meu cotidiano.

Aos meus grandes amigos que considero como irmãos André, Pedro, Franklin, Kaio, Eduardo e Filipe pela amizade e parceria.

Aos meus amigos do EJC por todos os conselhos.

Ao meu orientador Moysés Nascimento pela confiança, ensinamentos, paciência, preocupação e pelo incentivo dado. Agradeço também por ser um grande amigo, além de um exemplo como pessoa e como profissional.

Ao Biocafé pela cooperação e por disponibilizar os dados para a realização deste trabalho.

Aos meus coorientadores Ana Carolina Campana Nascimento e Cosme Damião Cruz por contribuírem diretamente no meu aprendizado e pelas sugestões nos trabalhos até aqui realizados.

Aos professores e funcionários do departamento de estatística da Universidade Federal de Viçosa, que sempre se empenharam e se mostraram acessíveis, dispostos a compartilharem conhecimento e suporte para com os alunos.

Aos membros da banca, Prof. Doutor Moysés Nascimento, Prof. Doutor Cosme Damião Cruz e Prof. Doutor Fabyano Fonseca e Silva por aceitarem o convite e por estarem dispostos a dar suas contribuições para este trabalho.

Aos meus mestres e amigos da Universidade Federal do Piauí, por me ensinarem os primeiros conceitos da estatística.

Aos amigos do LICAE e do laboratório de Bioinformática, pela amizade, ensinamentos e momentos de descontração que tornam o dia a dia mais fácil.

À Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Estatística Aplicada e Biometria pela oportunidade.

À CAPES, pela concessão da bolsa de estudos.

A todos os que contribuíram direta ou indiretamente para a concretização deste trabalho.

BIOGRAFIA

ITHALO COELHO DE SOUSA, filho de Joselita Coelho Barros e Francisco Barbosa de Sousa Neto, nasceu em Teresina, Piauí, em 27 de abril de 1990.

Em março de 2011, ingressou no curso de Bacharelado em Estatística na Universidade Federal do Piauí, Teresina – PI, fez graduação sanduíche entre agosto de 2013 e dezembro de 2014, nas universidades Morgan State University e Southern Illinois University of Carbondale, graduando-se em fevereiro de 2016.

Em março do mesmo ano, iniciou o curso de mestrado do Programa de Pós-Graduação em Estatística Aplicada e Biometria na Universidade Federal de Viçosa, submetendo-se à defesa da dissertação em 26 de fevereiro de 2018.

SUMÁRIO

RESUMO

SOUSA, Ithalo Coelho de, M.Sc., Universidade Federal de Viçosa, fevereiro de 2018. **Predição genômica da resistência à ferrugem alaranjada em café arábica via algoritmos de aprendizagem de máquina.** Orientador: Moysés Nascimento. Coorientadores: Ana Carolina Campana Nascimento e Cosme Damião Cruz.

A seleção genômica (SG) foi proposta como uma forma de aumentar a eficiência e acelerar o melhoramento genético. A SG enfatiza a predição simultânea dos efeitos genéticos de milhares de marcadores dispersos em todo o genoma de um organismo. Algumas metodologias estatísticas têm sido utilizadas em SG para a predição do mérito genético, como por exemplo a *Ridge Regression Best Linear Unbiased Prediction* (RR-BLUP), *Bayesian Lasso* (BLASSO). Porém tais metodologias exigem algumas pressuposições a respeito dos dados tais como normalidade da distribuição dos valores fenotípicos. Além disto, a presença de fatores complicadores tais como epistasia e dominância atrapalham a utilização destes modelos, uma vez que exigem que tais efeitos sejam estabelecidos à priori pelo pesquisador. Visando contornar a não normalidade dos valores fenotípicos a literatura sugere o uso dos modelos lineares generalizados sob o enfoque bayesiano (BGLR). Outra alternativa são os modelos baseados em aprendizagem de máquina (AM), representados por metodologias tais como Redes Neurais (RNA), Árvores de Decisão (AD) e seus possíveis refinamentos (*Bagging, Random Forest* e *Boosting*) as quais podem incorporar a epistasia e a dominância no modelo além de não exigirem pressuposições quanto ao modelo e a distribuição dos valores fenotípicos. Diante disso, o objetivo deste trabalho foi utilizar AD e seus refinamentos *Bagging, Random Forest* e *Boosting* para predição da resistência a ferrugem alaranjada no café arábica. Além disso, AD e seus refinamentos foram utilizadas para identificar a importância dos marcadores relacionados a característica de interesse. Os resultados foram comparados com aqueles provenientes do GBLASSO (Lasso Bayesiano Generalizado) e RNA. Foram utilizados dados da resistência a ferrugem do café de 245 plantas derivadas do cruzamento do Híbrido de Timor e do Catuaí Amarelo, genotipados para 137 marcadores. A AD e seus refinamentos obtiveram resultados satisfatórios, visto que apresentaram valores iguais ou inferiores de Taxa de Erro Aparente comparados com aqueles obtidos pelo GBLASSO e RNA. Ademais, os refinamentos da AD demonstraram ser capazes de identificar marcadores importantes para característica de interesse, visto que dentre os 10 marcadores mais importantes analisados em cada metodologia, 3-4

marcadores estavam próximos a QTL's relacionados a resistência a doença listados na literatura. Por fim, a AD e seus refinamentos mostraram um melhor desempenho em relação ao GBLASSO e a RNA quanto ao custo computacional.

ABSTRACT

SOUSA, Ithalo Coelho de, M.Sc., Universidade Federal de Viçosa, February, 2018. **Genomic prediction of leaf rust resistance to arabica coffee using machine learning algorithms.** Advisor: Moysés Nascimento. Co-advisors: Ana Carolina Campana Nascimento and Cosme Damião Cruz.

Genomic selection (GS) has been proposed as a way to increase efficiency and accelerate genetic improvement. GS emphasizes the simultaneous prediction of the genetic effects of thousands of scattered markers throughout an organism's genome. Some statistical methodologies have been used in GS for the prediction of genetic merit, such as Ridge Regression Best Linear Unbiased Prediction (RR-BLUP), Bayesian Lasso (BLASSO). However such methodologies require some assumptions about the data such as normality of the distribution of phenotypic values. In addition, the presence of complicating factors such as epistasis and dominance hinder the use of these models, since they require that such effects be established a priori by the researcher. In order to avoid the non-normality of phenotypic values, the literature suggests the use of Bayesian Generalized Linear Regression (BGLR). Another alternative is the models based on machine learning, represented by methodologies such as Artificial Neural Networks (ANN), Decision Trees (DT) and their possible refinements such as Bagging, Random Forest and Boosting, which can incorporate epistasis and dominance in the model, besides not requiring assumptions about the model and the distribution of phenotypic values. The aim of this work was to use DT and its refinements Bagging, Random Forest and Boosting for prediction of resistance to orange rust in arabica coffee. In addition, 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 (Generalized Bayesian Lasso) and ANN. Data from the coffee rust resistance of 245 plants derived from the hybrid of the Timor Hybrid and the Yellow Catuaí, genotyped for 137 markers were used. The DT and its refinements obtained satisfactory results, since they presented equal or inferior values of Apparent Error Rate compared to those obtained by GBLASSO and RNA. In addition, DT refinements seem to be able to identify important markers for characteristic of interest, since among the 10 most important markers analyzed in each methodology, 3-4 markers were close to QTLs related to resistance to disease listed in the literature. Finally, the Decision Tree and its refinements showed a better performance in relation to the GBLASSO and RNA regarding computational cost.

1. INTRODUÇÃO

O café é uma das bebidas mais consumidas do mundo, sendo a cultura comercial mais importante e a segunda commodity internacional mais valiosa, atrás somente do petróleo (Kewscience, 2017). O Brasil destaca-se como maior produtor e exportador de café do mundo, tendo exportado cerca de 32,9 milhões de sacas no ano safra 2016/2017 resultando na receita cambial de US$5,64 bilhões, sendo a espécie arábica responsável por 87,84% das exportações (EMBRAPA, 2017).

Algumas doenças podem afetar a cafeicultura, sendo a ferrugem alaranjada causada pelo fungo *Hemileia vastatrix* a principal em abrangência e danos no Brasil e no mundo. Para minimizar os prejuízos causados pela doença são utilizados cultivares resistentes as quais podem ser obtidas por meio de programas de melhoramento com informação do sequenciamento do genoma (Barka et al., 2017) e identificação de marcadores moleculares associados a doença (Alkimim et al., 2017). Outra abordagem interessante que auxilia na obtenção de cultivares resistentes é a Seleção Genômica (SG). A SG, proposta por Meuwissen et al. (2001), é uma variante da seleção assistida por marcadores na qual marcadores cobrindo todo o genoma são utilizados para predizer o mérito genético dos indivíduos visando a seleção de materiais com desempenho desejado de acordo com o caráter em estudo.

Em geral, os modelos estatísticos utilizados em SG para a predição do mérito genético, por exemplo *Ridge Regression Best Linear Unbiased Prediction* (RR-BLUP) e *Bayesian Lasso* (BLASSO), são baseados na pressuposição de normalidade dos valores fenotípicos. Portanto, o uso de tais modelos para valores fenotípicos caracterizados por variáveis categóricas ou binárias, como por exemplo resistência à ferrugem não são adequados. Visando contornar esse limitação, Pérez e de los Campos (2014) propuseram a utilização de modelos lineares generalizados sob o enfoque bayesiano (BGLR) estendendo assim a seleção genômica a modelos contínuos e discretos. Apesar de úteis, a presença de fatores complicadores tais como epistasia e dominância dificultam a utilização dos modelos usuais de SG, visto que os mesmos requerem que tais efeitos sejam estabelecidos à priori pelo pesquisador.

Outra abordagem interessante para problemas de predição é o uso de algoritmos de aprendizado de máquina, como por exemplo as Redes Neurais Artificiais (RNA), Árvores de Decisão (AD) e seus possíveis refinamentos tais como *Bagging*, *Random*

Forest e *Boosting*. Tais algoritmos não possuem pressuposições quanto ao modelo, já que seus resultados dependem do processo de aprendizado e não da distribuição das variáveis em si. Essa característica dos métodos de aprendizado estatístico possibilitam captar fatores complicadores tais como epistasia e dominância no modelo de predição e permitem que não haja nenhum pressuposto quanto a distribuição dos valores fenotípicos.

A utilização de Redes Neurais em SG pode ser vista em Silva et al. (2017) em que os autores propuseram o uso de Redes Neurais para predição de resistência à ferrugem em café arábica (*Coffea arabica*) obtendo resultados satisfatórios quanto a taxa de classificação incorreta comparado com o ajuste de modelos lineares generalizados sob o enfoque bayesiano. Glória et al., (2016) avaliaram a performance preditiva de uma Rede Neural e obtiveram os efeitos dos SNPs e valores de herdabilidades por duas diferentes estratégias (importância relativa e contribuição relativa). Apesar de interessante, as redes demandam muito recurso computacional e a determinação de quais marcadores são relevantes para o caráter em estudo não é uma tarefa trivial, visto que demanda o conhecimento e combinação dos valores dos pesos de neurônios em camadas ocultas obtidos após o treinamento da rede. Outra questão é o fato da rede (Perceptron Multi Camadas) não apresentar solução única (Silva et al., 2010), ou seja, para o mesmo problema é possível a obtenção de diferentes modelos baseados em Redes Neurais. Comparada com as RNA, as AD e seus possíveis refinamentos demandam menos recurso computacional e apresentam a importância dos marcadores de maneira fácil e direta. Além disso, da mesma forma que as RNA, tais metodologias apresentam boa performance preditiva (James et al., 2013).

Diante do exposto, o objetivo deste trabalho foi utilizar árvore de decisão e seus refinamentos *Bagging*, *Random Forest* e *Boosting* para predição da resistência à ferrugem do café arábica. Além disso, os resultados obtidos foram comparados com aqueles advindos do uso de modelos lineares generalizados sob o enfoque bayesiano e RNA do tipo PMC com uma camada oculta. Finalmente, a importância dos marcadores relacionados a resistência à ferrugem foi avaliada e descrita por meio das metodologias utilizadas neste estudo.

2. REFERENCIAL TEÓRICO

2.1. Café Arábica e ferrugem alaranjada

O Brasil é responsável por 70% das exportações mundiais de café, sendo o *Coffea arabica* a espécie mais produzida, representando 76% da produção nacional em 2017 (CONAB, 2017). Vários fatores podem influenciar a produtividade do café, como por exemplo a utilização dos insumos, avanços tecnológicos, fatores climáticos e biológicos e o potencial genotípico do material plantado. Os programas de melhoramento genético contribuem de forma decisiva para o desenvolvimento da cultura, proporcionando, por meio de cruzamentos, ganhos genéticos para características de interesse.

A cafeicultura brasileira originou pela introdução de apenas três plantas de café da variedade *Typica* espécie *C. arabica* e embora existam muitas cultivares de *C. arabica*, todas as cultivares conhecidas da espécie são derivada de *C. arabica* variedade *Typica* e *C. arabica* variedade Bourbon, as quais são consideradas as primeiras variedades de café. Consequentemente, os cafés apresentam uma base genética extremamente estreita (Anthony et al., 2002; Carvalho, 2008). Devido a estreita base genética do *C. arabica*, tem sido explorado com sucesso os genes de *C. canephora* para a obtenção de cultivares com resistência à ferrugem, sendo o hibrido de Timor um dos materiais genéticos mais utilizados nesse contexto (Bettencourt e Fazuolli, 2008).

O híbrido de Timor é um híbrido natural entre as espécies *Coffea arabica* e *Coffea canephora*, sendo um híbrido de grande importância nos programas de melhoramento visando à resistência a *Hemileia vastatrix*, agente causador da ferrugem alaranjada do cafeeiro (Setotawi et al.; 2010). A ferrugem é a doença de maior potencial destrutivo na cultura de café, causada pelo fungo *Hemileia vastatrix* Berk & Br, podendo causar danos de até 50% na produtividade.

Até o momento, foram identificadas 45 raças fisiológicas de *H. vastatrix* (Várzea; Marques, 2005), destas 15 já foram identificadas no Brasil (Zambolim et al., 2005; Cabral et al., 2009). Em cafeeiro, nove genes conferem resistência de diversas raças de *H. vastatrix*, sendo estes genes dominantes/recessivos e aparentemente independentes (Mayne, 1936; Noronha-Wagner; Bettencourt, 1967; bettencourt; Noronha-Wagner, 1971; Bettencourt e Rodrigues, 1988).

Tanto no *C. arabica*, quanto no *C. canephora* existem quatro gentes que conferem resistência a algumas raças de *H. vastatrix*, sendo admitido a possível existência de outros genes ainda não identificados. O Hibrido de Timor, resultante do cruzamento entre *C. arabica* e *C. canephora*, possui fenótipo de *C. arabica*, é tetraploide, autofertil (Rijo, 1974, Carvalho et al., 1989) e tem facilidade de cruzar com as cultivares de *C. arabica*, favorecendo a transferência de sua resistência (Carvalho et. al., 1989). Acessos derivados de Hibrido de Timor possuem cinco genes dominantes que condicionam os espectros de resistência.

2.2. Seleção Genômica Ampla

A Seleção Genômica Ampla (SGA), ou *Genome Wide Selecion (GWS)* foi proposta por Meuwissen et al. (2001) com o objetivo de utilizar informações diretas do DNA na seleção e predição do mérito genético, de forma a permitir alta eficiência seletiva, grande rapidez na obtenção de ganhos genéticos com a seleção de baixo custo em comparação com a seleção tradicional baseada em dados fenotípicos (RESENDE, 2012). A SGA é fundamentada em marcadores moleculares, que têm efeitos estimados a partir de uma metodologia específica e aplicados para predição do mérito genético. Com a SGA também, é possível selecionar indivíduos na fase inicial da vida, pois identifica através de marcadores moleculares, os alelos associados a uma determinada característica de interesse, permitindo a seleção precoce.

Atualmente, os marcadores moleculares mais usuais são os SNP's (polimorfismo de um único nucleotídeo, ou *single nucleotide polymorphisms),* pois são a forma mais abundante de variação do DNA nos genomas, sendo preferidos em relação a outros marcadores genéticos devido a sua baixa taxa de mutação e facilidade de genotipagem, aliados ao baixo custo (RESENDE, 2012). Esses marcadores se distribuem em todo o genoma em grande quantidade, o que facilita a detecção de desequilíbrios de ligação que provavelmente estão associados a uma determinada característica de interesse.

Dentre as diversas metodologias aplicadas em seleção genômica, aquelas baseadas em inferência bayesiana apresentam destaque e são utilizadas em diversos estudos (MEUWISSEN et al., 2001). No trabalho realizado por Meuwissen et al. (2001) foram propostos os métodos bayesianos Bayes A e Bayes B e comparados com o BLUP (*Best Linear Unbiased Predictor*) (HENDERSON, 1974) e LS (*Least Squares*) (LANDE e THOMPSON, 1990), obtendo-se melhores resultados em termos de acurácia. Fan et al.

(2011) também utilizaram metodologias bayesianas para análise de associação genômica, obtendo resultados satisfatórios.

Visando a análise de caráter em que os valores fenotípicos apresentam tanto distribuições contínuas quanto discretas, Péres e de los Campos (2014) propuseram os modelos lineares generalizados sob o enfoque bayesiano (BGLR). Outras abordagens que possibilitam a análises de valores fenotípicos que possuem tanto distribuições contínuas quanto discretas sem fazer nenhum pressuposto sobre as mesmas são as técnicas de aprendizado de máquina, como por exemplo, Redes Neurais, Árvores de Decisão e seus possíveis refinamentos (*Bagging*, *Random Forest* e *Boosting*). A seguir serão apresentadas tais metodologias.

2.3. Árvore de Decisão e seus possíveis refinamentos

A árvore de decisão (AD) é uma metodologia que particiona o espaço preditor em sub-regiões através de alguns critérios, para cara sub-região formada é atribuído um valor que será utilizado como valor predito para os novos indivíduos que serão alocados a essas sub-regiões. A estrutura da AD é composta pelos nós internos, ramos e nós externos/folhas. O nó é dito interno, quando os dados contidos neste nó são divididos de acordo com um critério de divisão, formando assim dois novos grupos de dados, sendo estes novos grupos ligados ao grupo antigo pelos ramos, já o nó e dito externo (folha) quando não ocorre mais divisões dos indivíduos pertencentes a este nó. A AD pode ser classificada como árvore de regressão quando a variável resposta é do tipo quantitativa, já quando a variável dependente assume valores qualitativos, a AD é classificada como árvore de classificação.

2.3.1. Árvore de Regressão

Para cada região formada na árvore é atribuído um valor, que será utilizado para predizer o valor da variável resposta de um novo indivíduo, sendo este valor a média de todos os indivíduos pertencente a região utilizados na construção da respectiva árvore.

Para a construção da árvore de regressão, o objetivo é construir regiões R_1, R_2, ..., R_M que minimiza a Soma de Quadrados dos Resíduos dado por:

$$\sum_{m=1}^{M} \sum_{i \in R_m} \left(y_i - \hat{y}_{R_m}\right)^2,$$

em que $\hat{y}_{R_m}$ é a média da variável resposta das observações de treinamento pertencente a m-éssima região. Porém, o custo computacional é muito alto sendo inviável considerar cada partição possível do espaço em M regiões para se obter a menor soma de quadrados dos resíduos (SQR). Para contornar o custo computacional, a literatura propõe realizar um procedimento baseado em divisões binárias recursivas, na qual objetiva-se, obter a variável X_p e o ponto *s*, que divida o espaço em duas regiões

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

tal que o ponto *s* divida a p-ésima variável em duas regiões que obtenha a menor soma de quadrados dos resíduos, por fim utilizamos a variável que obteve o menor SQR para a primeira divisão, em seguida repetimos o processo para cada região gerada.

Enquanto uma árvore muito grande pode se super ajustar aos dados, uma árvore pequena pode não capturar uma boa estrutura, ambos casos obtendo maus resultados quando utilizada a árvore construída para a predição de novos indivíduos não utilizados no treinamento da árvore. Uma abordagem para a escolha do tamanho da árvore seria construir uma árvore até que nenhuma região obtenha mais que 5 indivíduos e, em seguida, podá-la usando o custo complexidade da poda (Hastie et al., 2009).

O custo complexidade da poda consiste em analisar sub árvores (T_α) que possam ser obtidas através da árvore original (T_0). Para cada sub árvore, minimizamos o critério do custo de complexidade que é baseado em dois parâmetros, a taxa de erro *R(T)* e o tamanho da árvore em termos de folhas $|T|$, dado por:

$$R_\alpha(T) = R(T) + \alpha|T|$$

Sendo o parâmetro de ajuste $\alpha \geq 0$, responsável pelo equilíbrio entre o tamanho da árvore e a qualidade do ajuste dos dados de treinamento.

2.3.2. Árvore de Classificação

Para a construção da árvore de classificação, o objetivo é obter regiões R_1, R_2, ..., R_M que minimizam um dos 3 critérios apresentados a seguir (James et al., 2013):

- Taxa de Erro Aparente:

$$TEA = 1 - \max_k(\hat{p}_{mk}),$$

- Índice Gini:

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

- *Deviance*:

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} log \hat{p}_{mk}$$

em que $\hat{p}_{mk}$ representa a proporção de observações na m-ésima região pertencentes a k-ésima classe. Na construção da árvore de classificação é indicado utilizar o índice Gini ou o *Deviance,* pois estes são mais sensíveis para analisar a pureza do nó. Os índices diminuem de acordo com o crescimento da árvore que ocorre através da divisão binária recursiva. Para evitar o superajustamento do modelo, indica-se que nenhuma região obtenha mais que 5 indivíduos e em seguida podá-la usando qualquer um dos critérios como guia no custo complexidade da poda (Hastie et al., 2009), sendo a TEA o mais utilizado se a maior acurácia for o objetivo.

2.3.3. Bagging

Árvores de decisão possuem o problema de ter alta variabilidade, ou seja, se utilizarmos uma parte de um banco de dados para construirmos uma árvore e em seguida utilizarmos a outra parte do mesmo banco de dados para construir uma segunda árvore, iremos obter duas árvores com estruturas diferentes. Para contornar esse problema, o ideal seria obter várias amostras de uma mesma população, construir várias árvores e em seguida obter a média/moda dos valores preditos. Como não é uma tarefa fácil obter vários conjuntos de treinamento de uma população, é aplicada a técnica de *bootstrap*, que consiste em obter B amostras com reposição da amostragem disponível, obtendo assim B modelos $\hat{f}^1(x)$, $\hat{f}^2(x)$,..., $\hat{f}^B(x)$, por fim utilizamos os modelos gerados para obter uma média, dado por:

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

e assim diminuir a variabilidade obtida nas árvores de decisão.

A quantidade de árvores utilizadas no *Bagging* não é uma parâmetro que irá resultar num superajustamento do modelo, na pratica é utilizado uma quantidade onde o erro tenha estabilizado (James, 2013). Para analisar o erro, não é necessário utilizar validação cruzada ou utilizar dados de teste. No processo de *bootstrap*, em média 1/3 dos dados não são utilizados em cada modelo criado, sendo estes chamado de observações OOB (*out-of-bag*). Assim, ao final do *bagging* teremos que cada indivíduo estará presente em média em 2/3 dos modelos criados, estes sendo utilizados para obter o valor predito do indivíduo, e sendo testado nos 1/3 dos modelos restantes, onde será analisado a taxa de erro do teste.

2.3.4. Random Forest

Devido ao fato de sempre utilizarmos todas as variáveis em cada partição no *Bagging*, as predições obtidos nas AD estarão altamente correlacionados uma vez que as AD terão estruturas semelhantes. A média de valores altamente correlacionados, não resultada numa grande redução da variância, como ocorre quando é feita com valores não correlacionados. Ho (1995) visando melhorar a acurácia na classificação dos indivíduos propôs o *Random Forest* (RF), o qual segue a mesma ideia do *Bagging*, alterando somente o número de variáveis preditoras (m<p) utilizadas em cada partição, obtendo os valores preditos mais independentes, ocasionando assim na redução da variabilidade encontrada nas árvores de decisão. Hastie, et al., (2009) sugere que o número de variáveis preditoras utilizadas em cada partição seja dada da seguinte forma, para árvore de classificação $m = \sqrt{p}$ e para árvores de regressão $m = p/3$.

2.3.5. Boosting

Outra metodologia utilizada para melhorar a performance obtida por uma única árvore, é o *Boosting*. Ao contrário do *Bagging* que cria múltiplas árvores independentes, o *Boosting* cria árvores sequencialmente utilizando-se de informação prévia da árvore

anterior. Ao invés de ajustar um modelo para a variável resposta Y, nós ajustamos um grande número de árvores de decisão, $\hat{f}^1, \ldots, \hat{f}^B$, para o resíduo atual. Nessa metodologia a aprendizagem é lenta, necessitando assim que *B* seja grande, porém é necessário ter cuidado para não super ajustar o modelo, sendo utilizada a validação cruzada para se escolher o número de árvores que será construída.

2.4. Redes Neurais Artificiais

As Redes Neurais Artificiais (RNA) funcionam conceitualmente de forma similar ao cérebro humano, tentando reconhecer regularidades e padrões de dados, e são capazes de aprender com a experiência e fazer generalizações baseadas no seu conhecimento previamente acumulado. Embora biologicamente inspiradas, elas encontraram aplicações em diferentes áreas científicas.

O princípio central da teoria de redes neurais está no fato de que fornecendo exemplos (população de treinamento) do relacionamento entre variáveis de entrada X e um alvo T, a rede neural irá capturar a relação entre as variáveis, podendo generalizar essas informações para novos casos (população de validação) (Mackay, 1994).

A arquitetura de uma rede neural é definida pelo número de camadas (camada única ou múltiplas camadas), pelas conexões entre camadas, pelo número de neurônios em cada camada, pelo tipo de conexão entre eles (*feedforward* ou *feedback*) e pelo algoritmo de aprendizado (Haykin, 2001). Na Figura 1 esta ilustrada o modelo básico de um neurônio artificial.

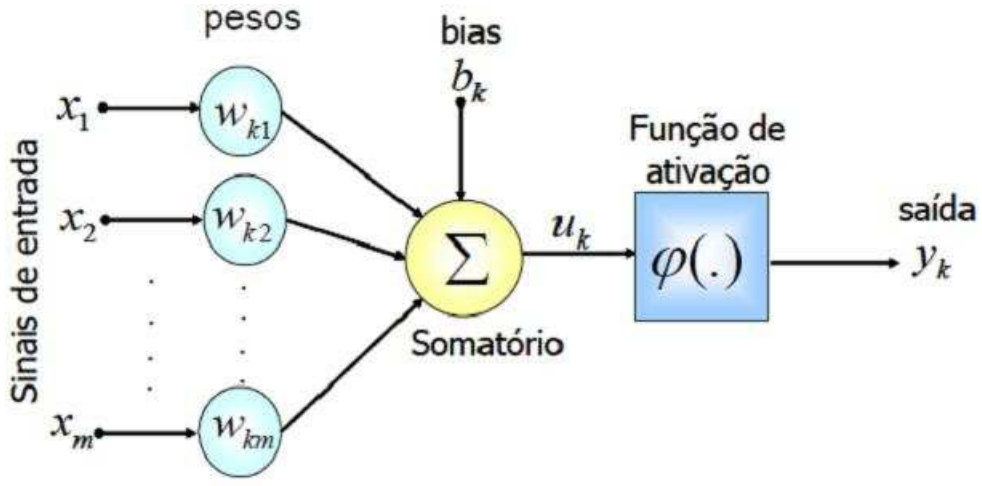

Figura 1- Modelo não linear de um neurônio artificial (Adaptado de Haykin, 2001). Nesta figura x_1, x_2, ..., x_m representam os valores dos marcadores; W_{k1}, W_{k2}, ..., W_{km} os pesos

sinápticos associados a cada entrada; b_k é o termo bias; u_k é a combinação linear dos sinais de entrada; φ(.) é a função de ativação e y_k é a saída do neurônio.

Matematicamente podemos definir a saída de cada neurônio como:

$$y_k = \varphi\left(\sum_{j=1}^{m} w_{kj} x_j + b_k\right),$$

em que:

- w_{kj} são os pesos sinápticos associados a cada entrada;
- x_j são as entradas da rede;
- b_k é o termo bias;
- $\varphi(.)$ é a função de ativação;
- y_k é a saída de neurônio.

Os pesos são os parâmetros ajustáveis que mudam e se adaptam à medida que o conjunto de treinamento é apresentado à rede. Assim, o processo de aprendizado supervisionado em uma RNA com pesos, resulta em sucessivos ajustes dos pesos sinápticos, de tal forma que a saída da rede seja a mais próxima possível da resposta desejada. O modelo neural também inclui um termo chamado de "bias", aplicado externamente, simbolizado por b_k. O b_k tem o efeito do acréscimo ou decréscimo da função de ativação na entrada da rede, dependendo se é positiva ou negativa, respectivamente. Um neurônio biológico dispara quando a soma dos impulsos que ele recebe ultrapassa o seu limiar de excitação (*threshold*) ele é tomado fora do corpo do neurônio e conectado usando uma entrada adicional.

As funções de ativação fornecem o valor da saída de um neurônio, elas limitam a saída do neurônio, geralmente nos intervalos de $[0,1]$ ou $[-1,1]$. As principais funções de ativação são a Função Limiar (Degrau), Função "Sigmoidal" ou "Sigmóide", Função Signum, Função Tangente Hiperbólica.

Na Figura 2 é apresentada uma arquitetura de rede neural composta por uma camada de entrada, duas camada intermediárias e uma camada de saída que é responsável por retornar os valores preditos da variável de interesse, sendo possível retornar uma ou mais variáveis respostas. Essa figura representa o modelo da rede Perceptron multicamadas.

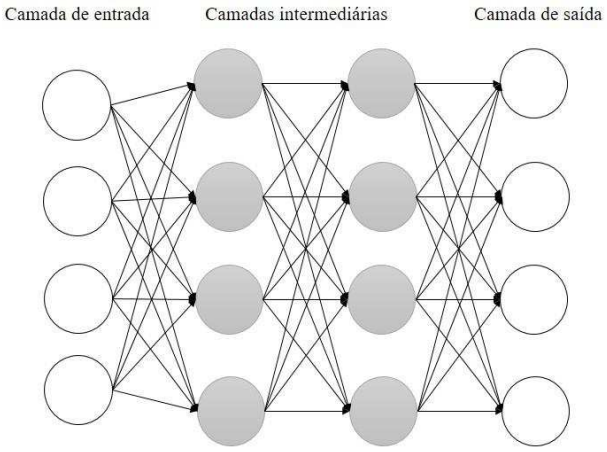

Figura 2- Representação dos três tipos de camadas existentes em redes neurais de múltiplas camadas, característica do modelo Perceptron Múltiplas Camadas.

2.5. Modelos Lineares Generalizados sob o enfoque Bayesiano

Os modelos lineares generalizados sob o enfoque Bayesiano (BGLR), proposto por Pérez e de los Campos (2014) para SG, são utilizados tanto para caracteres quantitativos como para qualitativos (binário ou multinominal). Quando o caráter for contínuo ($y_i; i = 1, \dots, n$) a equação é dada por $y_i = \boldsymbol{n}_i + \varepsilon_i$, em que $\boldsymbol{n}_i$ é um preditor linear e ε_i são resíduos independentes com distribuição $N(0, w_i^2\sigma_\varepsilon^2)$, sendo $w_i's$ pesos atribuídos. Em notação matricial o modelo é dado por $\boldsymbol{y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon}$, em que $\boldsymbol{y} = \{y_1, \dots, y_n\}$, $\boldsymbol{\eta} = \{\eta_1, \dots, \eta_n\}$ e $\boldsymbol{\varepsilon} = \{\varepsilon_1, \dots, \varepsilon_n\}$.

O preditor linear é estruturado da seguinte forma:

$$\boldsymbol{\eta} = 1\mu + \sum_{j=1}^{J} \boldsymbol{X}_j\boldsymbol{\beta}_j + \sum_{l=1}^{L} \boldsymbol{u}_l,$$

em que:

- μ é um intercepto;
- $\boldsymbol{X}_j$ é a matriz dos preditores;
- $\boldsymbol{\beta}_j$ são os vetores que contém os efeitos associados às colunas de $\boldsymbol{X}_j$;
- $\boldsymbol{u}_l = \{\mu_{l1}, \dots, \mu_{ln}\}$ são vetores de efeitos aleatórios.

A distribuição condicional dos dados é expresso por:

$$p(y|\boldsymbol{\theta}) = \prod_{i=1}^{n} N\left(y_i|\mu + \sum_{j=1}^{J}\sum_{k=1}^{K_j} x_{ijk}\beta_{jk} + \sum_{l=1}^{L} \mu_{li}\,, \sigma_\varepsilon^2 w_i^2\right)$$

em que $\boldsymbol{\theta}$ representa valores desconhecidos como o intercepto, coeficientes de regressão, variância residual e outros efeitos aleatórios. A priori admite a seguinte fatoração $p(\boldsymbol{\theta}) = p(\mu)p(\sigma_\varepsilon^2)\prod_{j=1}^{J} p(\boldsymbol{\beta}_j)\prod_{l=1}^{L} p(\boldsymbol{u}_l)$. O intercepto assume uma priori flat e a variância residual assume uma qui-quadrada invertida escalonada com densidade $p(\sigma_\varepsilon^2) = \chi^{-2}(\sigma_\varepsilon^2|S_\varepsilon, gl_\varepsilon)$ com graus de liberdade $gl_\varepsilon(>0)$ e o parâmetro escalar $S_\varepsilon(>0)$.

Os coeficientes da regressão $\{\beta_{jk}\}$ podem assumir tanto uma priori informativa como não informativa. Quando assume-se uma priori não informativa para β_{jk}, os efeitos fixos são estimados baseados somente na informação contida na verossimilhança. Para os coeficientes que assumem uma priori informativa, a escolha da priori tem um papel importante na determinação do tipo de encolhimento dos efeitos induzidos. Como exemplo de priori informativa temos a Gaussiana que induz encolhimento da estimativa similar ao da regressão *ridge*. A exponencial dupla (priori utilizada no modelo BLASSO) e a t-escalonada (priori utilizada no modelo Bayes A) que possuem distribuição mais concentradas no ponto zero e caudas mais densas do que na distribuição normal.

Para os vetores de efeitos aleatórios $\boldsymbol{\mu}_l$ são atribuídos prioris de distribuição normal multivariada com média zero e matriz de covariância $Cov(\boldsymbol{u}_l, \boldsymbol{u}_l') = \boldsymbol{K}_l\sigma_{ul}^2$ em que $\boldsymbol{K}_l$ é uma matriz simétrica positiva semi-definida e σ_{ul}^2 é o parâmetro de variância com priori $\sigma_{ul}^2 \sim \chi^{-2}(gl_l, S_l)$.

Quando o caráter é qualitativo, é utilizada a ligação probit (Pérez e de los Campos, 2014). A probabilidade de cada categoria é ligada ao preditor linear de acordo com a seguinte função de ligação:

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

onde $\Phi(.)$ é a função acumulada da distribuição normal padrão, η_i é o preditor linear e γ_k são os parâmetros do limiar, com $\gamma_0 = -\infty, \gamma_k \geq \gamma_{k-1}, \gamma_k = \infty$.

2.6. Coeficiente de Coincidência de *Cohen's Kappa*

O coeficiente de *Cohen's Kappa* proposto por Cohen (1960) com o objetivo de verificar a concordância entre duas metodologias diferentes possui a vantagem de considerar o grau de concordância dado o acaso. O coeficiente é dado por:

$$Kappa = \frac{Nº\ concordâncias\ observadas - Nº\ concordâncias\ esperadas\ ao\ acaso}{Nº\ observações\ analisada - Nº\ concordâncias\ esperadas\ ao\ acaso}$$

Por exemplo, se selecionarmos 30 (20%) indivíduos em cada uma das metodologias dentre um grupo de 150 indivíduos, e tenha sido observado uma concordância de 15 indivíduos selecionados pelos dois métodos. O número de concordâncias observadas ao acaso é 6, dado por 20% do número de indivíduos selecionados. Neste exemplo temos então que o coeficiente de *Kappa* é igual a 0,375.

Ornella et al., (2014) utilizaram este coeficiente para estimar a acurácia preditiva de diferentes algoritmos de classificação na seleção genômica. Landis e Koch (1977) propuseram a seguinte classificação para o índice *Cohen's Kappa* (Tabela 1), sendo este considerado bom quando obter valores acima de 0,4.

Tabela 1. Classificação do índice *Cohen's Kappa*.

Índice de Kappa	Classificação
$\hat{k} \leq 0{,}2$	Ruim
$0{,}2 < \hat{k} \leq 0{,}4$	Razoável
$0{,}4 < \hat{k} \leq 0{,}6$	Bom
$0{,}6 < \hat{k} \leq 0{,}8$	Muito bom

$0{,}8 < \hat{k} \leq 1{,}0$	Excelente

3. MATERIAL E MÉTODOS

O experimento foi efetuado na casa de vegetação e no laboratório de biotecnologia do Cafeeiro (Biocafé), estabelecido no instituto de biotecnologia aplicada à agropecuária (Bioagro), da Universidade Federal de Viçosa (UFV). Foram utilizados duas populações como progenitores, o Híbrido de Timor UFV 443-03 (resistente a ferrugem) e a Catuaí Amarelo 64 UFV 2148-57 (suscetível a ferrugem), gerando um híbrido F1 e desta população foi feita a autofecundação para se obter as 245 plantas utilizadas para a construção do mapa genético e identificação dos marcadores moleculares ligados aos genes/QTL envolvidos na resistência. A fenotipagem foi realizada em Maio, Julho e Agosto de 2009 sendo a 1ª, 2ª e 3ª repetição respectivamente. A genotipagem foi realizada nos anos 2010, 2011 e 2012, sendo esta realizada com 137 marcadores (74 AFLP, 58 SSR, 4 RAPD e 1 primer especifico) (Pestana et al., 2015).

Os dados dos marcadores para cada indivíduo foram codificados para análises de seleção genômica. Para os marcadores dominantes em aproximação (alelo proveniente do progenitor resistente Híbrido de Timor UFV 443-03) foram atribuídos -1 para a presença e 1 para a ausência da banda. Nos marcadores dominantes em repulsão (alelo proveniente do progenitor suscetível Catuaí Amarelo UFV 2148-57) as codificações foram 1 e -1 para a presença e ausência da banda, respectivamente. Os marcadores codominantes foram codificados como 0 para heterozigoto, -1 para bandas provenientes do progenitor resistente e 1 para bandas provenientes do progenitor suscetível.

Na fenotipagem referente à avaliação da manifestação da resistência ou suscetibilidade do genótipo, realizaram-se inoculações em que os uredósporos do patótipo 001 de *H. vastatrix* foram multiplicados e inoculados de acordo com a metodologia descrita por Capucho et al. (2009). A inoculação foi realizada através da metodologia de discos foliares descrita por Eskes (1982), com três replicações. A fenotipagem foi efetuada de acordo com Capucho et al. (2009), seguindo a escala descrita por Tamoyo et al. (1995), baseada na presença ou na ausência de uredósporos. Os fenótipos foram classificados como mais resistentes quando atribuído as notas 1 e 2; em que a nota 1 apresenta ausência de sintomas e a nota 2 contém lesões cloróticas pequenas. Quando atribuído as notas 3, 4 , 5 e 6, os fenótipos são classificados como mais suscetíveis, sendo

atribuída a nota 3 quando apresenta lesões cloróticas grandes sem esporulação, a nota 4 é atribuída quando apresenta lesões cloróticas grandes com poucos uredósporos, ocupando menos de 25% da área foliar, a nota 5 é atribuída quando contém lesões com esporulação ocupando de 25 a 50% da área e recebe nota 6 quando se tem lesões como esporulação ocupando mais de 50% da área com uredósporos. Para predizer a característica, definimos os fenótipos em duas classes, maior resistência (1 = notas 1 e 2) e maior suscetibilidade (0 = notas de 3 a 6).

3.1. Árvore de Classificação (AC) e seus possíveis refinamentos

Para a construção da árvore de classificação, o objetivo é obter regiões R_1, R_2, ..., R_M que minimizam o índice Gine dado por (James et al., 2013):

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

em que $\hat{p}_{mk}$ representa a proporção de observações na m-ésima região pertencentes a k-ésima classe. O índice Gini diminui de acordo com o crescimento da árvore que ocorre através da divisão binária recursiva. Para evitar o superajustamento do modelo, indica-se que nenhuma região obtenha mais que 5 indivíduos e em seguida podá-la usando o *custo complexidade* da poda (Hastie et al., 2009). Visando aumentar a performance preditiva do modelos foram utilizados o *bootstrap aggregation (Bagging), Random Forest* e *Boosting*.

O *bootstrap aggregation (Bagging)* consiste em obter B amostras com reposição (tamanho igual a N) do conjunto de dados, obtendo assim B modelos ($\hat{f}^1(x)$, $\hat{f}^2(x)$,..., $\hat{f}^B(x)$) que serão utilizados como classificadores individuais. Um novo indivíduo será classificado na classe mais comum dentre as predições dos B classificadores individuais. Já o *Random Forest* (RF) segue a mesma ideia do *Bagging*, alterando somente o número de variáveis preditoras (m<p) utilizadas em cada partição. Segundo James et al., (2013) o RF resulta num processo de descorrelação entre as árvores geradas melhorando ainda mais a acurácia das predições. Finalmente, ao contrário do *Bagging* que cria múltiplas árvores independentes, o *Boosting* cria árvores sequencialmente utilizando informações das árvores anteriores. O classificador do *Boosting* possui a forma $\boldsymbol{H}(\boldsymbol{x}) = \sum_t \alpha_t \boldsymbol{h_t}(\boldsymbol{x})$ que busca minimizar uma função de perda *L* através da otimização do escalar α_t (importância atribuída a $\boldsymbol{h_t}(\boldsymbol{x})$) e do classificador individual $\boldsymbol{h_t}(\boldsymbol{x})$ (árvore de decisão

individual) a cada iteração t (Freund e Schapire, 1999). Os classificadores individuais $\boldsymbol{h_t(x)}$ possuem um poder classificatório baixo, porém quando utilizados conjuntamente $\boldsymbol{H(x)}$, apresentam bons resultados (Appel, et al., 2013).

3.2. Rede Neural Artificial (RNA)

A arquitetura da RNA utilizada neste trabalho possui somente uma camada oculta e utiliza o *backpropagation* como algoritmo de aprendizado (Rumelhart et al., 1986). A configuração da rede considerando os 137 marcadores como entradas uma camada oculta e a saída a qual prediz a resistência ou susceptibilidade da folha a ferrugem é apresentada na Figura 3.

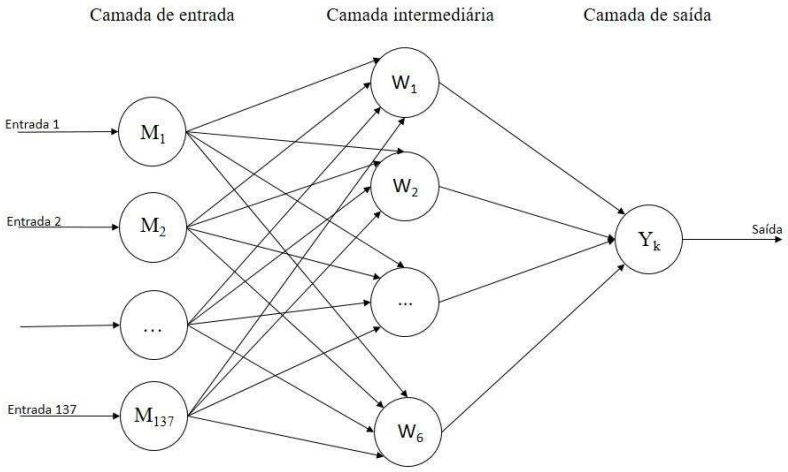

Figura 3. Arquitetura da RNA composta de 137 entradas (marcadores) na camada de entrada, uma camada oculta com os neurônios W_i (i = 1, 2,…,6) e a saída Y_k que prediz a resistência da folha a ferrugem.

Os neurônios W_m são gerados por combinações lineares das variáveis de entrada M_j (marcadores) e por fim a variável de saída Y_k é modelada como uma função de combinações lineares dos neurônios W_m da seguinte maneira:

$$W_m = \sigma(\alpha_{0m} + \alpha_m^T X), m = 1, 2, \dots, 6$$

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

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

em que $W = (W_1, W_2, \dots, W_m),\ T = T_1, T_2, \dots, T_K$.

A função de ativação utilizada na RNA foi a sigmóide, $\sigma(v) = 1/(1 + e^{-v})$ e a função de saída a *softmax*, $g_k(T) = \frac{e^{T_k}}{\sum_l^K e^{T_l}}$. Os pesos $(\alpha_{0m}, \alpha_m; m = 1, 2, \dots, M)$ e $(\beta_{0k}, \beta_k; k = 1, 2, \dots, K)$ são parâmetros desconhecidos da rede responsáveis por ajustar bem o modelo da RNA ao conjunto de treinamento. Como medida de ajuste utilizamos o *cross-entropy* $R(\theta) = -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} log f_k(x_i)$, sendo o *backpropagation* responsável pela sua minimização.

3.3. Modelo Linear Generalizado sob o enfoque bayesiano (BGLR)

O método de seleção genômica baseado em modelos lineares generalizados sob o enfoque bayesiano (de los Campos e Pérez, 2014) também foi utilizada para a predição do mérito genético. O modelo é dado por:

$$Y = \mu + X_1\beta_1 + \cdots + X_j\beta_j + u_1 + \cdots + u_q$$

sendo μ o intercepto, X_i as matrizes dos preditores, $X_p = \{x_{ijk}\}$, β_j os vetores de efeitos associados as colunas de X_j e os $u_q = \{u_{q1}, \dots, u_{qn}\}$ são vetores de efeitos aleatórios. Nesse estudo, visto que os valores fenotípicos apresentam distribuição categórica (resistência da folha à ferrugem), a função de ligação *probit* (*probit link*) foi utilizada (Pérez e de los Campos, 2014) onde a probabilidade de cada categoria é ligada ao preditor linear de acordo com a seguinte função de ligação:

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

em que $\Phi(.)$ é a função acumulada da distribuição normal padrão, η_i é o preditor linear e γ_k são os parâmetros do limiar, com $\gamma_0 = -\infty, \gamma_k \geq \gamma_{k-1}, \gamma_k = \infty$.

3.4. População de treinamento e validação

O conjunto de dados foi particionado em duas partes: conjunto de treinamento e conjunto de validação. Os conjuntos de treinamentos foram mantidos com os mesmos indivíduos para a modelagem de todas metodologias, sendo estes compostos por 70% de cada classe (172 indivíduos), tomadas aleatoriamente, os 30% (73 indivíduos) restantes

foram utilizados no conjunto de validação. Na literatura as porcentagens utilizadas no conjunto de treinamento variam entre 60 e 90% como visto em Gianola et al. (2011) e González-Camacho et al. (2012).

3.5. Importância de marcadores

Para obtenção dos marcadores mais importantes em cada metodologia, utilizamos todos os indivíduos na construção dos modelos. No GBLASSO definiu-se como os marcadores de maior importância aqueles com maiores coeficientes de regressão em valores absolutos. No processo da árvore de classificação e da poda, utilizamos os marcadores escolhidos nos primeiros níveis da árvore. No *Bagging* e *Ranfom Forest* assumimos como os marcadores mais importantes aqueles que em média influenciaram mais na redução do índice Gini. No *Boosting* os marcadores mais importantes são aqueles que possuem uma maior relevância em separar as observações de uma classe das demais.

Na RNA a obtenção dos marcadores mais importantes, ocorre através da construção de novas RNA após anular o efeito de cada marcador individualmente. O marcador mais importante será aquele que após sua retirada apresenta uma maior TEA (Silva et al., 2017). Dentre as 100 redes utilizadas no processo de seleção genômica, a rede utilizada para se obter a importância dos marcadores foi aquela que apresentou uma menor TEA.

Selecionamos os 10 marcadores mais importantes em cada metodologia e comparamos com resultados obtidos na literatura (Pestana et al., 2015), para verificar se as metodologias realmente são capazes de indicar marcadores que estão associados a uma certa característica de interesse.

3.6. Comparação de metodologias

Para a comparação entre as metodologias, dividimos aleatoriamente o conjunto em dois subconjuntos (treinamento e validação) num total de 100 vezes. Utilizamos o custo computacional médio e o intervalo de confiança da Taxa de Erro Aparente (TEA) obtida através dos conjuntos de validações como podemos ver na Figura 4.

Utilizamos o coeficiente de coincidência de Cohen (Medida Kappa) para analisar a concordância entre as metodologias tanto na classificação dos indivíduos (utilizando o

conjunto de validação) quanto na identificação dos marcadores (os 10 marcadores mais importantes de cada metodologia). O coeficiente de Cohen's kappa é dado por:

$$kappa = \frac{NCO - NCEA}{NOA - NCEA}$$

sendo NCO o número de concordância observadas, NCEA o número de concordância esperada ao acaso e NOA o número de observações analisada (Resende et al., 2014).

3.7. Aspectos computacionais

As análises dos dados foram realizadas em um computador com processador core i7 3.40GHz e 16 GB de memória RAM e foi utilizado o software R 3.40 (R Core Team, 2017), sendo utilizado para a predição por meio das RNA a função *nnet*, através do pacote *nnet* (Venables e Ripley, 2002). As características da RNA foram escolhidas de acordo com a limitação da TEA de no máximo 15% ou um máximo de 5000 iterações no conjunto de validação. A função *BGLR*, pertencente ao pacote *BGLR* (de los Campos e Pérez, 2016), foi utilizada para estimar os modelos generalizados bayesiano, sendo realizadas 100.000 iterações com *burn-in* de 20.000 e *thin* de 10. Para a construção da árvore de classificação utilizamos a função *tree* pertencente ao pacote *tree* (Ripley, 2016). A função *randomForest* utilizada para construção do modelo do *Bagging* e do *Random Forest* pertence ao pacote *randomForest* (Liaw e Wiener, 2002). Por fim a função *gbm* do pacote *gbm* (Ridgeway et al., 2017) foi utilizada para a utilização do *Boosting.*

4. RESULTADOS

Os valores de Taxa de Erro Aparente (TEA) média obtidas por meio do ajuste dos modelos em estudo (AD, Árvore de Decisão com Poda - ADP, *Bagging, Random Forest, Boosting*, RNA, GBLASSO) variaram de 19,49% até 24,90% (Tabela 2). Especificamente, a menor TEA média foi obtida considerado o ajuste do *Boosting* (19,49%) a qual não apresenta grande diferença quando comparada com aquelas obtidas por meio do ajuste da RNA com uma camada oculta (19,56%), *Bagging* (19,67%), *Random Forest* (20,44%) e ADP (21,12%) (Tabela 2). De acordo com os intervalos de confiança de 95%, todas as metodologias, exceto a AD (24,90%), apresentaram resultados

superiores (menores valores de TEA) quando comparado com aqueles obtidos por meio do ajuste de um GBLASSO (TEA médio de 22,69%) (Figura 4).

Tabela 2. Tempo médio em segundos (diagonal), coeficiente de Cohen's Kappa médio entre as classificações (acima da diagonal) e coeficiente de Cohen's Kappa entre os marcadores mais importantes de cada metodologia (abaixo da diagonal) obtidos pelas metodologias utilizadas nesse estudo.

Modelos	RNA	GBLASSO	AD	ADP	*Bagging*	RF	*Boosting*
RNA	355,92	76,36%	69,00%	73,54%	78,74%	77,77%	81,07%
GBLASSO	24,49%	38,68	70,33%	77,19%	83,06%	88,88%	86,44%
AD	0%*	24,49%	0,01	83,09%	78,27%	73,54%	73,69%
ADP	**	**	**	0,14	85,67%	80,97%	81,66%
Bagging	2,91%	56,85%	24,49%	**	4,10	89,43%	89,10%
RF	2,91%	56,85%	13,70%	**	46,06%	3,21	93,44%
Boosting	13,70%	89,21%	24,49%	**	67,64%	67,64%	13,63

Abreviações: RNA, Rede Neural Artificial; GBLASSO, Lasso Bayesiano Generalizado; AD, Árvode de Decisão; ADP, Árvore de Decisão com Poda; RF, Random Forest. * Coeficiente menor que zero. ** Coeficiente não calculado devido a metodologia identificar poucos marcadores importantes.

Após a classificação dos genótipos de acordo com a predição obtida por cada metodologia, o coeficiente Kappa de Cohen médio entre as mesmas foi estimado e estão apresentados acima da diagonal da Tabela 2. Adicionalmente, a razão entre o tempo gasto para o ajuste/treinamento dos modelos também foram obtidos e são apresentados abaixo da diagonal na Tabela 2.

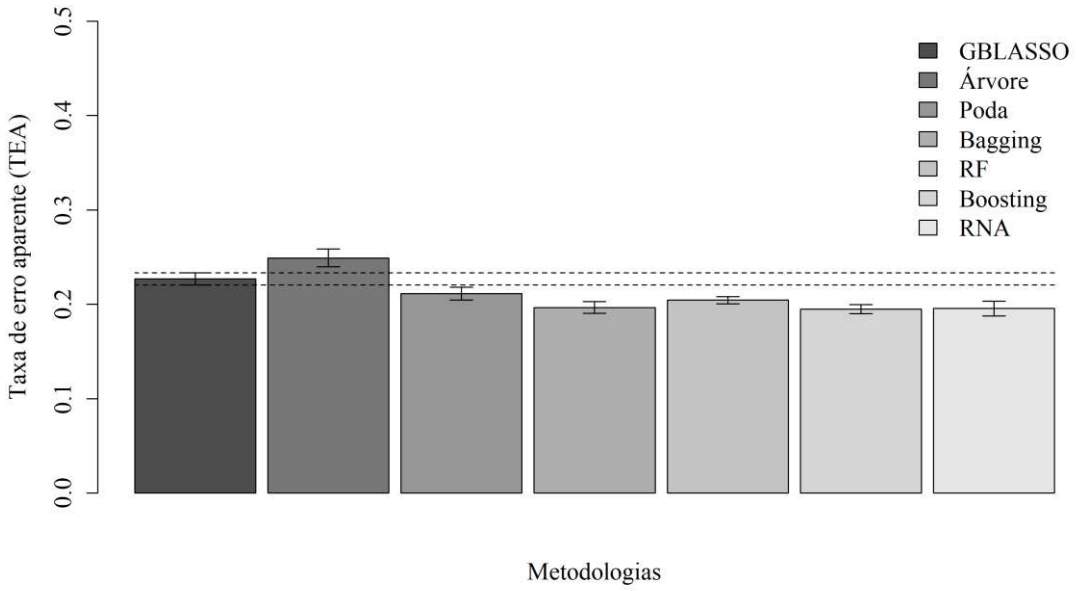

Figura 4: Taxa de erro aparente e intervalos de confiança com 95% obtidos para cada modelo ajustado. As linhas pontilhadas destacam os limites do IC obtidos por meio do GBLASSO. Abreviações: RNA, Rede Neural Artificial; GBLASSO, Lasso Bayesiano Generalizado; AD, Árvore de de Decisão; RF, *Random Forest*.

De acordo com Landis e Koch (1977), um coeficiente *Kappa* de Cohen superior a 0,4 pode ser considerado uma estimativa boa a excelente de concordância entre as metodologias. Em geral, o coeficiente *Kappa* de Cohen médio apresentou valores positivos e altos na classificação dos genótipos. Comparando as técnicas ao GBLASSO, observa-se que a metodologia com menor percentual de concordância foi a AD (70,33%), enquanto que o RF foi a metodologia mais similar quanto a classificação fornecida pelo GBLASSO (88,88%) (Tabela 2). Já considerando todas as metodologias avaliadas, o maior percentual de concordância foi observado entre os resultados obtidos por meio do RF e *Boosting* (93,44%). Em termos da demanda computacional, dentre todos os modelos ajustados, apenas a RNA não demanda menor tempo computacional quando comparada com o ajuste do GBLASSO, exigindo 9,20 vezes mais do custo computacional. A AD foi a que obteve um menor custo computacional quando comparada ao GBLASSO, sendo 2850 vezes mais rápida que a mesma.

Dentre os dez marcadores mais importantes indicados por cada metodologia (Tabela 3) observa-se que o maior percentual de concordância Kappa Cohen (89,21%) foi observado entre GBLASSO e Boosting o qual apresentou 9 marcadores em comum. Já as metodologias RNA e AD apresentaram menor percentual de concordância (0,00%), não tendo apresentado nenhum marcador em comum.

Tabela 3. Dez marcadores mais importantes indicados por cada metodologia.

RNA	GBLASSO	AD	ADP	*Bagging*	RF	*Boosting*
2	**43**	**43**	**43**	97	**43**	97
24	73	97	97	**43**	97	**43**
25	**61**	7	-	**61**	**61**	**61**
26	97	86	-	**47**	**64**	**47**
29	11	84	-	73	**12**	**55**
55	**12**	34	-	**12**	11	**12**
63	29	94	-	**55**	21	29
67	24	118	-	115	59	11
77	**55**	52	-	**19**	29	128
125	128	73	-	101	**47**	73

Os marcadores destacados em negritos são marcadores em região de QTL de acordo com Pestana et al., 2015. Abreviações: RNA, Rede Neural Artificial; GBLASSO, Lasso Bayesiano Generalizado sob enfoque Bayesiano; AD, Árvode de Decisão; ADP, Árvore de Decisão com Poda; RF, Random Forest.

5. DISCUSSÃO

Nesse estudo objetivou-se apresentar e utilizar algoritmos de aprendizado de máquina tais como aqueles baseados em RNA, árvore de decisão (AD) e seus possíveis

refinamentos *Bagging*, *Random Forest* e *Boosting* para predição da resistência à ferrugem do café arábica. Os resultados obtidos foram comparados com aqueles advindos da abordagem tradicional para estudos de SG para situações nas quais as características não apresentam distribuição Normal, os modelos lineares generalizados sob o enfoque bayesiano. Finalmente, por meio dos resultados de cada metodologia foram selecionados os dez marcadores mais importantes quanto a resistência a ferrugem.

O uso de métodos baseados em aprendizado de máquina (AM) para predição da resistência à ferrugem do café arábica foi eficiente, visto que em todos os ajustes, exceto aquele obtido por meio da AD, apresentaram menores valores de TEA quando comparado com aqueles obtidos por meio do ajuste do GBLASSO. O menor desempenho da AD comparado as outras metodologias era esperado visto que tal metodologia apresenta grande variabilidade em termos de predição (James et al., 2013). Hastie et al. (2009) enfatizam que a baixa acurácia preditiva das AD pode ser contornada pela utilização de metodologias tais como *Bagging*, *Random Forest* e *Boosting* (Breiman, 2001) que agregam várias árvores conjuntamente visando reduzir a variabilidade em termos preditivos. Outro resultado interessante é a similaridade dos valores de TEA obtidos por meio do bagging (19,49%) e RF (20,44%). Tal abordagem utiliza um número de preditores inferior ($m \approx \sqrt{p}$) ao bagging. Essa estratégia visa quebrar a correlação entre os modelos construídos em cada iteração de forma a aumentar a capacidade preditiva do modelo (Hastie et al., 2009). Nesse estudo a semelhança entre os valores de TEA é explicado pois apenas 1,61% dos marcadores apresentaram forte correlação entre si, além disso, temos que a resistência a ferrugem é uma caraterística oligogênica, assim a divisões de algumas sub-regiões são realizadas somente com marcadores que não estão associados a resistência (Bettencourt; Rodrigues, 1988). Os marcadores utilizados nesse estudos foram obtidos apenas em regiões nas quais se observaram a presença QTL em estudos prévios nessa população (Pestana et al., 2015), caracterizando assim um mapeamento fino apenas em regiões cromossômicas de interesse, o que explica o número reduzido de marcadores utilizados e o baixo desequilíbrio de ligação entre os mesmos.

Os métodos baseados em AM foram utilizados sob o enfoque de SG em diversos estudos, como por exemplo o realizado por Ogutu et al. (2011). Neste estudo, os autores compararam a habilidade preditiva do RF, *boosting* e máquinas de vetores suporte (MVS) para predizer valores genéticos genômicos (VGG) por meio de dados simulados e verificaram melhor performance do *boosting* em detrimento dos outros dois métodos.

Gianola et al., (2014) avaliaram e comprovaram que a capacidade preditiva do GBLUP (*Genomic Best Binear Unbiased Prediction*) pode ser melhorada por meio do *bagging*. Os autores verificaram que o uso do *bagging* no GBLUP além de melhorar desempenho preditivo do método tornou o mesmo mais robusto em relação ao superajustamento aos dados. Tal abordagem também foi utilizada com sucesso nos estudos de Abdollahi-Arpahi et al. (2015) e Mehrban et al. (2017) nos quais os autores comprovaram a eficiência do uso do *bagging* conjuntamente com o GBLUP na predição de valores genéticos genômicos (VGG) para, respectivamente, frangos e touros da raça Jersey. Já Ornella et al. (2014) utilizaram diversos algoritmos e classificação para a predição genômica em milho e verificaram que tais metodologias são uma alternativa promissora para SG no melhoramento de plantas. Diferentemente desses estudos, os quais apontaram para superioridade dos métodos baseados em aprendizado de máquina em relação a abordagens tradicionais (exemplo, GBLUP, Lasso Bayesiano e Bayes C), devido à natureza qualitativa do caráter avaliado nesse estudos, resistência a doença, as metodologias avaliadas foram comparadas com os resultados provenientes do ajuste de um modelo lineares generalizados sob o enfoque bayesiano o qual contempla essa característica naturalmente na modelagem. Dentre os diversos métodos de aprendizado de máquina, apenas as RNA foram comparadas com o ajuste de modelos lineares generalizados sob o enfoque bayesiano na presença de uma caráter qualitativo (Silva et al., 2017). Nesse estudo, da mesma forma que os demais a metodologia baseada em AM, no caso RNA, obteve resultados superiores em relação aos obtidos por meio do GBLASSO.

O alto percentual de concordância entre as metodologias, pode indicar que o caráter em estudo (resistência a doença) não apresenta fatores complicadores a modelagem tais como a presença de grande dominância e epistasia os quais demandam a utilização de modelos mais complexos. Apesar da similaridade, deve-se ressaltar que os modelos baseados em AM são mais flexíveis e não dependem da especificação a priori do modelo a ser ajustado o que facilita comtemplar tais fatores complicadores (Silva et al., 2017). Em termos de custo computacional, visto que o GBLASSO utiliza métodos de simulação de Monte Carlo via Cadeia de Markov para obter amostras da distribuição marginal dos parâmetros os quais demandam da obtenção de grandes cadeias para o alcance da convergência, os métodos baseados em AM, exceto as RNA, apresentam grande vantagem quando comparados ao enfoque tradicional.

No presente estudo nenhum QTL foi identificado em regiões próximas os dez marcadores indicados pela RNA como os mais importantes. Dentre os dez indicados pelo ajuste do GBLASSO, quatro marcadores (43, 61, 12 e 55) se encontram próximos a regiões nas quais QTL's foram detectados (Pestana et al., 2015). No RF e no boosting, dos dez marcadores, cinco (RF = 43, 61, 64, 12 e 47; *boosting* = 43, 61, 47, 55 e 12) encontram-se em regiões próximas a QTL's. Já no *bagging* seis marcadores (43, 61, 47, 12, 55 e 19) estão localizados próximos a QTL's (Pestana et al., 2015). A ADP não é uma estratégia interessante para obter tais informações, visto que o processo de Poda retira vários marcadores do processo de predição. Esses resultados indicam que quando aplicado os refinamentos (*bagging*, RF e *boosting*) às AD, proporcionam boas alternativas para determinação de marcas e consequentemente regiões importantes para a determinação do caráter em estudo. Por outro lado, o resultado obtido pelo ajuste da RNA indica que a estratégia utilizada nesse estudo para determinar a importância de um marcador por meio dessa técnica não é eficiente. Apesar dos resultados obtidos nesse estudo apontar para um superioridade do *bagging* tanto na predição de VGG e determinação da importância de marcadores, mais estudos com dados simulados e reais com diferentes arquiteturas genéticas e conjuntos de dados de diferentes tamanhos (indivíduos e marcadores) são necessários para confirmar a eficiência de tais metodologias comparadas com aqueles utilizadas tradicionalmente para tais tarefas.

6. CONCLUSÃO

Houve indícios que a RNA utilizada neste trabalho (perceptron multicamadas com uma camada oculta) e os refinamentos da árvore de decisão (Poda, *Bagging*, RF e *Boosting*) mostraram maior eficácia, em termos da TEA, comparada ao método tradicional utilizado (GBLASSO) na predição do caráter em estudo (resistência a ferrugem), exigindo ainda menor custo computacional (exceto a RNA). Os refinamentos da AD foram capazes de detectar marcadores próximo à regiões onde QTL's foram identificados para a característica em estudo.

REFERÊNCIAS BIBLIOGRÁFICAS

ABDOLLAHI-ARPANAHI, R.; MOROTA, G.; VALENTE, B.D.; KRANIS, A.; ROSA, G.J.M.; GIANOLA, D. Assessment of bagging GBLUP for whole-genome prediction of

broiler chicken traits. **Journal of Animal Breeding and Genetics**, v.132, p.218-228, 2015.

ALKIMIM, E.R.; CAIXETA, E.T.; SOUSA, T.V.; PEREIRA, A.A.; OLIVEIRA, A.C.B. de; ZAMBOLIM, L.; SAKIYAMA, N.S. Marker-assisted selection provides arabica coffee with genes from other coffea species targeting on multiple resistance to rust and coffee berry disease. **Molecular Breeding**, v.37:6, p.1-10, 2017.

ANTHONY, F.; COMBES, M.C.; ASTORGA, C.; BERTRAND, B.; GRAZIOSI, G.; LASHERMES, P. The origin of cultivated *Coffea arabica* L. varieties revealed by AFLP and SSR markers. **Theoretical and Applied Genetics**, v.104:5, p.894-900, 2002.

APPEL, R.; FUCHS, T.; DOLLÁR, P.; PERONA, P. Quickly Boosting Decision Trees – Pruning Underachieving Features Early -. In: Journal of Machine Learning Research: Workshop and Conference Proceedings, 30., 2013, Atlanta. **Proceedings.** Atlanta: International Conference on Machine Learning, 2013. V.28. p.594-602.

BARKA, G.D.; CAIXETA, E.T.; ALMEIDA, R.F. de; ALVARENGA, S.M.; ZAMBOLIM, L. Differential expression of molecular rust resistance components have distinctive profiles in Coffea arabica – Hemileia vastatrix interactions. **European Journal of Plant Pathology**, v.149, p.543-561, 2017.

BETTENCOURT, A.J.; NORONHA-WAGNER, M. Genetic factors conditioning resistance of *Coffea arabica* L. to *Hemileia vastatrix* Berk. & Br. **Agronomia Lusitana**, v. 31, p. 285-292, 1971.

BETTENCOURT, A.J.; ROGRIGUES JR., C.J. Principles and practice of coffee breeding for resistance to rust and other diseases. In: CLARKE, R. J.; MACRAE, R., (Eds.) **Coffee**. London: Elsevier Applied Science, 1988. v.4, p.199-234.

BETTENCOUTRT, A.J.; FAZUOLLI, L.C. **Melhoramento genético de *Coffea Arabica L.*:** Transferência de genes de resistência a *Hemileia vastatrix* do Híbrido de Timor para a cultivar Villa Sarchí de *Coffea Arabica.* Campinas: Instituto Agronômico, 2008, 20p.

BREIMAN, L. Random Forests. **Machine Learning**. v.45, p.5-32, 2001.

CABRAL, P.G.C.; MACIEL-ZAMBOLIM, E.; ZAMBOLIM, L.; LELIS, T. de P.; CAPUCHO, A.S.; CAIXETA, E. T. Identification of a new race of *Hemileia vastatrix* in Brazil. **Australasian plant disease notes**, v. 4, p. 129-130, 2009.

CAPUCHO, A.S.; EVELINE, T.C.; ZAMBOLIM, E.M.; ZAMBOLIM, L. Herança da resistência do Híbrido de Timor UFV 443-03 à ferrugem-do-cafeeiro. **Pesq. agropec. bras.**, v.44:3, p.276-282, 2009.

CARVALHO, A.; FAZUOLI, L.C.; COSTA, W.M. Melhoramento do cafeeiro: XLI. Produtividade do Híbrido Timor, de seus derivados e outras fontes de resistência a *Hemilia vastatrix.* **Bragantia**, Campinas, v.48, n.1, p.73-86, 1989.

CARVALHO, C.H.S. **Cultivares de café:** origem, características e recomendações, 1st ed. Brasília, DF: Embrapa Café, 2008, 334p.

COHEN, J. A Coefficient of Agreement for Nominal Scales. **Educational and Pshychological Measurement**, v.20, p.37-46, 1960.

CONAB – Companhia Nacional de Abastecimento. Acompanhamento da safra brasileira de café, v.4-Safra 2017, n.4 – Quarto levantamento, Brasília, p.1-84, dez. 2017. Disponível em: < http://www.conab.gov.br/OlalaCMS/uploads/arquivos/18_01_08_09_06_29_cafe_dezembro.pdf>. Acesso em: 23 Jan. 2018.

DE LOS CAMPOS, G; PÉREZ RODRIGUES, P. **BGLR**: Bayesian Generalized Linear Regression. R package version 1.0.5. Disponível em: < https://cran.r-project.org/web/packages/BGLR/BGLR.pdf >. Acesso em: 17 Jan. 2018.

EMBRAPA. **Agroindústria:** Cafés do Brasil geram receita cambial de US$ 5,64 bilhões no ano safra 2016/2017. Disponível em: < https://www.embrapa.br/busca-de-noticias/-/noticia/25326094/cafes-do-brasil-geram-receita-cambial-de-us-564-bilhoes-no-ano-safra-20162017>. Acesso em: 4 Dez. 2017.

ESKES, A.B. The use of leaf disk inoculations in assessing resistance to coffee leaf rust (*Hemileia vastatrix*). **Netherlands Journal of Plant Pathology**, v.88:4, p.127-141, 1982.

FAN, B.; ONTERU, S.K.; DU, Z.Q.; GARRICK, D.J.; STALDER, K.J.; ROTHSCHILD, M.F. Genome-wide association study identifies Loci for composition and structural soundness traits in pigs. **PlosOne**, v.6(2), p.1-11, 2011.

FREUND, Y.; SCHAPIRE, R.E. A Short Introduction to Boosting. **Journal of Japanese Society for Artificial Intelligence**, v.14(5), p.771-780, 1999.

GIANOLA, D.; OKUT, H.; WEIGEL, K.A.; ROSA, G.J.M. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. **BMC Genetics**, v.12:87, p.1-14, 2011.

GIANOLA, D.; WEIGEL, K.A.; KRÄMER, N.; STELLA, A.; SCHÖN, C.C. Enhancing Genome-Enabled Prediction by Bagging Genomic BLUP. **PlosOne**, v.9(4), p.1-18, 2014.

GLÓRIA, L.S.; CRUZ, C.D.; VIEIRA, R.A.M.; RESENDE, M.D.V. de; LOEPS, P.S.; SIQUEIRA, O.H.G.B.D. de; SILVA, F.F. e. Accessing marker effects and heritability estimates from genome prediction by Bayesian regularized neural networks. Livestock Science. V.191, p.91-96, 2016.

GONZÁLEZ CAMACHO, J.M.; DE LOS CAMPOS, G.; PÉREZ RODRIGUES, P.; GIANOLA, D.; CAIRNS, J.E.; MAHUKU, G.; BABU, R.; CROSSA, J. Genome-enabled prediction of genetics values using radial basis function neural networks. **Theoretical and Applied Genetics**, v.125, p.759-771, 2012.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statiscal Learning** Data Mining, Inference, and Prediction, 2nd ed. New York, NY: Springer, 2009, 745p.

HAYKIN, S.; **Redes Neurais – Princípios e prática.** 2nd ed. Porto Alegre, RS: Bookman, 2001. 900p.

HENDERSON, C.R. **Aplications of linear models in animal breeding**. Guelph, ON: University of Guelph, 1984. 462p.

HO, T.K. Random Decision Forests. In: International Conference on Document Analysis and Recognition, 3., 1995, Quebec. **Proceedings.** Quebec: IEEE, 1995. v.2. p.278-282.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning** with Applications in R. 1st ed. New York, NY: Springer, 2013. 426p.

LANDE, R.; THOMPSON, R. Efficiency of marker-assisted selection in the improvement of quantitative traits. **Genetics**, v.124, p.743-756, 1990.

LANDIS, J.; KOCH, G. The measurement of observer agreement for categorical data. **Biometrics**, v.33, p.159-174, 1977.

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

MACKAY, D.J.C. Bayesian non-linear modelling for the prediction competition. In: ASHRAE Transactions, 1994, Orlando. **Proceedings.** Atlanta: ASHRAE, 1994. V.100(2). p.1053-1062.

MAYNE, W.W. **Annual report of the coffee scientific officer.** [S.I.]: Mysore Coffee Experimental Station, 1936. 21 p. (Bulletin 14).

MEHRBAN, H.; LEE, D.H.; MORADI, M.H.; ILCHO, C.; NASERKHEIL, M.; IBÁÑEZ-ESCRICHE, N. Predictive performance of genomic selection methods for carcass traits in Hanwoo beef cattle: impacts of the genetic architecture. **Genetics Selection Evolution**, v.49(1), p1-13, 2017.

MEUWISSEN, T.H.E.; HAYES, B.J.; GODDARD, M.E. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. **Genetics**, v. 157, p.1819-1829, 2001.

NORONHA-WAGNER, M.; BETTENCOURT, A.J. Genetic study of the resistance of *Coffeea* sp to leaf rust 1. Identification and behavior of four factors conditioning disease reation in *Coffea arabica* to twelve physiologic races of *Hemileia vastatrix.* **Canadian Journal of Botany**, v. 45, p. 2021-31, 1967.

OGUTU, J.O.; PIEPHO, H.P.; SCHULZ-STREECK, T. A comparison of random forests, boosting and support vector machines for genomic selection. In: European workshop on QTL mapping and marker assisted selection (QTL-MAS), 14., Poznan. **Proceedings**. Bydgoszcz: BMC Proceedings, 2011. V.5(3)11. p.1-5

ORNELLA, L.; PÉREZ, P.; TAPIA, E.; GONZÁLEZ-CAMACHO, J.M.; BURGUEÑO, J.; ZHANG, X.; SINGH, S.; VICENTE, F.S.; BONNETT, D.; DREISIGACHER, S.; SINGH, R.; LONG, N.; CORSSA, J. Genomic-enabled prediction with classification algorithms. **Heredity**, v.112, p.616-626, 2014.

PÉREZ RODRIGUES, P.; DE LOS CAMPOS, G. Genome-Wide Regression and Prediction with the BGLR Statistical Package. **Genetics**, v.198, p.483-495, 2014.

PESTANA, K.N.; CAPUCHO, A.S.; CAIXETA, E.T.; ALMEIDA, D.P. de; ZAMBOLIM, E.M.; CRUZ, C.D.; ZAMBOLIM, L.; PEREIRA, A.A.; OLIVEIRA, A.C.B. de; SAKIYAMA, N.S. Inheritance study and linkage mapping of resistance loci to *Hemileia vastatrix* in Híbrido de Timor UFV 443-03. **Tree Genetics & Genomes**, v.11:72, p.1-13, 2015.

KEWSCIENCE. **Plants of the World online**: Coffea arabica L. Disponível em: <http://powo.science.kew.org/taxon/urn:lsid:ipni.org:names:747038-1#bibliography>. Acesso em: 4 Dec. 2017.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2017. https://www.R-project.org/. Acesso em 19 Jan 2018.

RESENDE, M.D.V.; SILVA, F.F.; AZEVEDO, C.F. **Estatística matemática, biométrica e computacional: modelos mistos, multivariados, categóricos e generalizados (REML/BLUP), inferência bayesiana, regressão aleatória, seleção genômica, QTL- GWAS, estatística espacial e temporal, competição, sobrevivência**. Viçosa, MG: UFV, 2014. 881p.

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

RIDGEWAY, G. **gbm**: Generalized Boosted Regression Models. R package version 2.1.3. Disponível em: < https://CRAN.R-project.org/package=gbm >. Acesso em: 17 Jan. 2018.

RIJO, L. Observações cariológicas no cafeeiro Híbrido do Timor. **Portugaliae Acta Biológica**, v.13, p.157-168, 1974.

RIPLEY, B. **tree:** Classification and Regression Trees. R package version 1.0-37. Disponível em: < https://CRAN.R-project.org/package=tree >. Acesso em: 17 Jan. 2018.

RUMELHART, D.E.; McCLELLAND, J.L. **Parallel Distributed Processing** Explorations in the Microstructure of Cognition: Foundations. Cambridge, MA: The MIT Press, 1986. 567p.

SETOTAW, T.A.; CAIXETA, E.T.; PENA, G.F.; ZAMBOLIM, E.M.; PEREIRA, A.A.; SAKIYAMA, N.S. Breeding potential and genetic diversity of "Híbrido do Timor" coffee evaluated by molecular markers. **Crop Breeding and Applied Biotechnology**, v.10, p.298-304, 2010.

SILVA, G.N.; NASCIMENTO, M.; SANT'ANNA, I. de C.; CRUZ, C.D.; CAIXETA, E.T.; CARNEIRO, P.C.S.; ROSADO, R.D.S.; PESTANA, K.N.; ALMEIDA, D.P. de; OLIVEIRA, M. da S. Artificial neural networks compared with Bayesian generalized linear regression for leaf rust resistance prediction in Arabica coffee. **Pesquisa Agropecuária Brasileira**, v.41, p.186-193, 2017.

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

VÁRZEA, V.M.P.; MARQUES, D.V. Population variability of *Hemileia vastatrix* vs. coffee durable resistance. In: ZAMBOLIM, L.; ZAMBOLIM, E> M.; VÁRZEA, V. M. P. (Eds.). **Durable resistance to coffee leaf rust**. Viçosa, MG: DFT/UFV, 2005. P. 53-74.

VENABLES, W.N.; RIPLEY, B.D. **Modern Apllied Statistics with S**. 4th ed. New York: Springer, 2002. 495p.

ZAMBOLIM, L.; MACIEL-ZAMBOLIM, E.; VALE, F.X.R.; PEREIRA, A.A.; SAKIYAMA, N.S.; CAIXETA, E.T. Physiological races of *Hemalaia vastatrix* Berk et Br in Brazil – Physiological variability, current situation and future prospects. In: ZAMBOLIM, L.; ZAMBOLIM, E.M.; VÁRZEA, V.M.P. (Orgs.). **Durable resistance**

to coffee leaf rust. 1. ed. Visconde do Rio Branco, MG: Suprema Gráfica e Editora Ltda., 2005. v. 1, p. 75-98.