ANTÔNIO CARLOS DA SILVA JÚNIOR

PREDIÇÃO E IMPORTÂNCIA DE PREDITORES EM ABORDAGENS FUNDAMENTADAS EM INTELIGÊNCIA COMPUTACIONAL E APRENDIZADO DE MÁQUINAS

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

Orientador: Cosme Damião Cruz

VIÇOSA - MINAS GERAIS
2021

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

T

S586p
2021

Silva Júnior, Antônio Carlos da, 1987-
Predição e importância de preditores em abordagens fundamentadas em inteligência computacional e aprendizado de máquinas / Antônio Carlos da Silva Júnior. - Viçosa, MG, 2021.
1 tese eletrônica (138 f.): il. (algumas color.).

Inclui apêndice.
Orientador: Cosme Damião Cruz.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2021.070
Modo de acesso: World Wide Web.

1. Melhoramento genético - Simulação por computador. 2. Rede neurais Computação). 3. Decisão estatística. 4. Determinantes (Matemática). 5. Variáveis (Matemática). I. Universidade Federal de Viçosa. Departamento de Biologia Geral. Programa de Pós-Graduação em Genética e Melhoramento. II. Título.

CDD 22. ed. 576.50285

Bibliotecário(a) responsável: Alice Regina Pinto CRB6 2523

ANTÔNIO CARLOS DA SILVA JÚNIOR

PREDIÇÃO E IMPORTÂNCIA DE PREDITORES EM ABORDAGENS FUNDAMENTADAS EM INTELIGÊNCIA COMPUTACIONAL E APRENDIZADO DE MÁQUINAS

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

APROVADA: 16 de setembro de 2021.

Assentimento:

Antônio Carlos da Silva Júnior
Autor

Cosme Damião Cruz
Orientador

Aos meus familiares e amigos.

AGRADECIMENTOS

Agradeço a Deus e a Nossa Senhora Aparecida por iluminar o meu caminho durante toda esta jornada, dando-me saúde, conhecimento e força para conseguir superar os desafios da vida.

Aos meus pais, Antônio Carlos da Silva e Nilza Helena Gonçalves da Silva, pelo amor, carinho, dedicação e incentivo para que eu conseguisse alcançar mais este objetivo.

A minha irmã Walquíria Gonçalves da Silva pela amizade, compreensão, paciência e incentivo.

À Universidade Federal de Viçosa e ao Programa de Pós-graduação em Genética e Melhoramento, pela oportunidade de cursar a graduação e o mestrado e doutorado.

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

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela concessão da bolsa de estudos.

À Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), pela concessão da bolsa de estudos.

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

Ao professor e amigo Dr. Cosme Damião Cruz, pela excelente orientação, pelos conhecimentos transmitidos, paciência e sobretudo pela amizade. Exemplo de dedicação e competência.

Ao pesquisador da EPAMIG e amigo Dr. Plínio César Soares, pelo apoio durante todo o período de minha formação acadêmica, pela dedicação, incentivo e amizade.

Aos professores e amigo Dr. Leonardo Lopes Bhering e Dr. Moysés Nascimento, pela coorientação e apoio durante a minha formação acadêmica, pela dedicação, incentivo e amizade.

Aos professores dos departamentos de Biologia Geral, Agronomia e Estatística que colaboraram para a minha formação acadêmica.

Aos meus amigos Rafael Said Bhering Cardoso, Guilherme Silva Guerra e Marcello Gomes Filho pela amizade, apoio e compreensão.

As pesquisadoras Dra. Isabela de Castro Sant'Anna, Dra. Michele Jorge Silva e professora Dra. Gabi Nunes da Silva pelas valiosas contribuições e suporte pela colaboração, ensinamentos, disponibilidade, sugestões e amizade.

Aos amigos que fiz durante todo este período, em especial aos amigos do Laboratório de Bioinformática, Vinicius, Alexandre, Iara, Ivan, Renato, Francyse, Ricardo, Rafael, Luciano, Isabela, Gabi, Gabriela, Cynthia, Jeck, Jaquecele, Marciane, Dayane, Cristiano, Luiza, Wender, Hélcio, Laís, Weverton, Ithalo, Jaqueline e Haroldo pelo apoio e paciência.

Aos membros do Laboratório de Inteligência Computacional e Aprendizado Estatístico (LICAE), em especial aos professores Dr. Moyés Nascimento, Dr. Fabyano Fonseca e Silva (*in memoriam*), Dra. Camila Ferreira Azevedo e Dra. Ana Carolina Nascimento pelos ensinamentos, disponibilidade, sugestões e amizade.

Aos funcionários do BIOAGRO pela disponibilidade e amizade, em especial ao senhor Paulo.

Aos secretários do Programa de Pós-graduação em Genética e Melhoramento, Marco Túlio e Odilon. A todos aqueles que colaboraram de alguma forma para este trabalho.

MUITO OBRIGADO!

Tudo posso naquele que me fortalece". Filipenses 4:13"

BIOGRAFIA

ANTÔNIO CARLOS DA SILVA JÚNIOR, filho de Antônio Carlos da Silva e Nilza Helena Gonçalves da Silva, nasceu em 07 de fevereiro de 1987, em Viçosa, no estado de Minas Gerais.

Em fevereiro de 2003, iniciou o ensino médio na Escola Estadual Effie Rolfs, formando-se em dezembro de 2005. Em março de 2010, iniciou o curso de Engenharia Agronômica na Universidade Federal de Viçosa, onde obteve o título em julho de 2015.

Em março de 2016, iniciou o curso de Mestrado no Programa de Genética e Melhoramento na Universidade Federal de Viçosa, submetendo-se à defesa de dissertação em julho de 2017.

Em agosto de 2017, iniciou o curso de Doutorado no Programa de Genética e Melhoramento na Universidade Federal de Viçosa, submetendo-se à defesa de tese em setembro de 2021.

RESUMO

SILVA JÚNIOR, Antônio Carlos, D.Sc., Universidade Federal de Viçosa, setembro de 2021. **Predição e importância de preditores em abordagens fundamentadas em inteligência computacional e aprendizado de máquinas**. Orientador: Cosme Damião Cruz.

O estudo da importância das características permite ao melhorista orientar estratégias para selecionar e acelerar o progresso do melhoramento genético. Embora, a avaliação simultânea de características no programa de melhoramento de plantas forneça uma grande quantidade de informações, identificar qual característica fenotípica é a mais importante é um desafio para o melhorista. Assim, o objetivo deste trabalho foi estimar a melhor abordagem para predição e estabelecer uma rede de melhor poder preditivo via metodologias baseadas em regressão, inteligência artificial e aprendizado de máquinas. A quantificação da importância de variáveis através da rede Percepton Multicamadas (MLP) pode ser obtida através de (i) algoritmo de GARSON (1991) modificado por GOH (1995) (GA), que consiste no particionamento dos pesos de conexão de rede neural para determinar a importância relativa de cada variável de entrada na rede. (ii) Avaliação da importância de variáveis (entrada) através do impacto da desestruturação ou perturbação da informação de uma determinada entrada sobre a estimativa do R^2. Essa importância foi estimada trocando informações ou tornando o valor fenotípico de cada característica constante e verificando as mudanças nas estimativas de R^2. Quando os valores de uma característica são perturbados, o valor de R^2 diminui, indicando que a característica é importante em relação às outras para fins de predição. A importância de variáveis utilizando a rede função de base radial (RBF) foi estimado conforme a MLP. Para aprendizado de máquina foram usadas árvores de decisão, bagging, floresta aleatória e boosting. A qualidade do modelo preditivo foi ajustada determinado com base em R^2, e o MSE foi usado para quantificar a importância das características fenotípicas. A importância da característica explicativa foi determinada estimando o aumento percentual no MSE. No primeiro artigo, avaliou-se a importância de características auxiliares de uma característica principal com base em informações fenotípicas e estrutura genética previamente conhecida usando inteligência computacional e aprendizado de máquina para desenvolver ferramentas preditivas para o melhoramento genético. Foram simulados uma população F_2 representada por 500 indivíduos, obtidos a partir de um cruzamento entre pais homozigotos contrastantes. Os caracteres fenotípicos simulados apresentam com base em médias previamente estabelecidas e estimativas

de herdabilidade (30%, 50% e 80%). As características foram distribuídas em um genoma com 10 grupos de ligação, considerando dois alelos. Foram considerados quatro cenários diferentes. Para a característica principal (PT1), a herdabilidade constitui-se de 50%, e 40 locos de controle foram distribuídos em cinco grupos de ligação. A simulação de outras características de controle fenotípico com a mesma complexidade da característica principal, mas sem qualquer relação genética com ele e sem pleiotropia ou uma ligação fatorial entre os loci de controle. Essas características compartilhavam grande número de locos de controle com a característica principal, mas podiam ser distinguidas pela ação diferencial do ambiente sobre elas, conforme refletido nas estimativas de herdabilidade (30%, 50% e 80%). Os R^2 variaram de 44,0% - 83,0% e 79,0% - 94,0%, para inteligência computacional e aprendizado de máquina, respectivamente. Na rede MLP os R^2 foram 83,03%, 77,89%, 75,49% e 82,14% para os cenários 1, 2, 3 e 4. Pela abordagem GA em todos os cenários, as contribuições relativas de PT5 e PT2 na previsão de PT1 foram quantificadas como maiores e menores, respectivamente. Para a rede RFB a permutação foi eficiente na quantificação da contribuição relativa de PT5 como um fator importante com base na redução na estimativa de R^2 quando a informação foi perturbada e PT2 foi identificado como o traço menos importante. O PT5 foi estimado como o traço fenotípico mais importante em todas as metodologias de aprendizado de máquina e em todos os cenários. As contribuições relativas de características auxiliares em diferentes cenários em programas de melhoramento de plantas podem ser predito com eficiência usando inteligência computacional e aprendizado de máquina. No segundo artigo, o objetivo foi estimar a melhor abordagem para predição e estabelecer uma rede de melhor poder preditivo em arroz irrigado por inundação via tais metodologias. Os experimentos foram realizados nos municípios de Leopoldina, Lambari e Janaúba, estado de Minas Gerais, Brasil. Foram utilizado 75 genótipos de arroz irrigado por inundação. As características avaliadas foram rendimento de grãos, comprimento da panícula e relação comprimento x largura de grãos, que foram utilizadas como variáveis de resposta e outras dez variáveis explicativas. A abordagem de inteligência artificial em Leopoldina proporcionou maior estimativa para as variáveis preditivas PL e GY no procedimento RBF, 83,44% e 78,90%, respectivamente. Em Leopoldina e Lambari, para a variável resposta LGW, obteve estimativa máxima de R^2 de aproximadamente 100% por regressão múltipla e abordagens de inteligência artificial e em Janaúba, de 62%. A contribuição relativa de caracteres auxiliares em arroz por meio de inteligência computacional e aprendizado de máquina mostrou-se eficiente para determinar a importância relativa de variáveis em arroz irrigado por inundação. Os caracteres indicados para auxiliar na tomada de decisão são floração,

número de grãos cheios por panículas e comprimento de panículas para este estudo. No procedimento de boosting, as variáveis que se destacaram foram HP, GL, PL, GP, WG e LGW em todos os ambientes. Os caracteres indicados para auxiliar na tomada de decisão são floração, número de grãos cheios por panículas e comprimento de panículas para este estudo. No terceiro artigo, avaliou-se a contribuição relativa de caracteres auxiliares em aveia branca por meio de tais metodologias. Os experimentos foram conduzidos na região Sul do Brasil. Foram avaliados 78 genótipos de aveia branca avaliados anos de 2008 e 2009. Em cada ano, constitui-se de sem e com fungicida, de forma que foram estabelecidos modelos de predição em quatro conjuntos experimentais. O delineamento foi em blocos casualizados com três repetições. As características avaliadas foram rendimento de grãos que foram utilizadas como variável resposta e dez outras como variáveis explicativas. O procedimento bagging e boosting, verifica-se que as estimativas de R^2 foram superiores a 92.70% e 80%, respectivamente. O R^2 variaram de 30,14% - 96,45% e 10,57% - 94,61%, para inteligência computacional e aprendizado de máquina, respectivamente. Os caracteres indicados para auxiliar na tomada de decisão são estatura de planta, severidade de ferrugem da folha e percentual de acamamento para este estudo. Acredita-se que, com a utilização de procedimento para quantificar a importância de variáveis, as técnicas fundamentadas em inteligência computacional e aprendizado de máquina possam ser facilmente empregadas sem demandar recursos computacionais sofisticados.

Palavras-chave: Rede Neurais Artificiais. Árvore de decisão. Coeficiente de Determinação. Importância de Variáveis.

ABSTRACT

SILVA JÚNIOR, Antônio Carlos, D.Sc., Universidade Federal de Viçosa, September 2021. **Prediction and importance of predictors in approaches based on computational intelligence and machine learning.** Advisor: Cosme Damião Cruz.

The study of the importance of traits allows the breeder to guide strategies to select and accelerate the progress of genetic improvement. Although the simultaneous evaluation of traits in the plant breeding program provides a great deal of information, identifying which phenotypic trait is the most important is a challenge for the breeder. Thus, the objective of this work was to estimate the best prediction approach and establish a network with better predictive power via methodologies based on regression, artificial intelligence, and machine learning. The quantification of the importance of variables through the Multilayer Perception Network (MLP) can be obtained through (i) GARSON's (1991) algorithm modified by GOH (1995) (GA), which consists in the partitioning of the neural network connection weights for determine the relative importance of each input variable in the network. (ii) Evaluation of the importance of variables (input) through the impact of destructuring or disturbing the information of a given input on the estimation of R^2. This importance was estimated by exchanging information or making the phenotypic value of each characteristic constant and checking for changes in the estimates of R^2. When the values of a feature are disturbed, the value of R^2 decreases, indicating that the feature is important over the others for prediction purposes. The importance of variables using the radial basis function network (RBF) was estimated according to the MLP. For machine learning, decision trees, bagging, random forest, and boosting were used. The quality of the predictive model was determined based on R^2, and the MSE was used to quantify the importance of the phenotypic traits. The importance of the explanatory characteristic was determined by estimating the percentage increase in the MSE. In the first manuscript, we assessed the importance of auxiliary traits of a main trait based on phenotypic information and previously known genetic structure using computational intelligence and machine learning to develop predictive tools for genetic improvement. An F_2 population represented by 500 individuals, obtained from a cross between contrasting homozygous parents, was simulated. The simulated phenotypic characters are based on previously established means and heritability estimates (30%, 50%, and 80%). The traits were distributed in a genome with 10 linkage groups, considering two alleles. Four different scenarios were considered. For the main trait (PT1),

heritability was 50%, and 40 control loci were distributed into five linkage groups. The simulation of other phenotypic control traits with the same complexity as the main trait but without any genetic relationship to it and without pleiotropy or a factorial link between the control loci. These traits shared a large number of control loci with the main trait but could be distinguished by the differential action of the environment on them, as reflected in heritability estimates (30%, 50%, and 80%). The R^2 ranged from 44.0% - 83.0% and 79.0% - 94.0%, for computational intelligence and machine learning, respectively. In the MLP network the R^2 were 83.03%, 77.89%, 75.49% and 82.14% for scenarios 1, 2, 3 and 4. By the GA approach in all scenarios, the relative contributions of PT5 and PT2 in the prediction of PT1 were quantified as major and minor, respectively. For the RFB network, permutation was efficient in quantifying the relative contribution of PT5 as an important factor based on the reduction in the estimate of R^2 when the information was perturbed and PT2 was identified as the least important trait. PT5 was estimated as the most important phenotypic trait in all machine learning methodologies and in all scenarios. The relative contributions of auxiliary traits in different scenarios in plant breeding programs can be efficiently predicted using computational intelligence and machine learning. In the second manuscript, the objective was to estimate the best prediction approach and establish a network with better predictive power in flood irrigated rice via such methodologies. The experiments were carried out in the municipalities of Leopoldina, Lambari, and Janaúba, the state of Minas Gerais, Brazil. Seventy-five genotypes of flood irrigated rice were used. The characteristics evaluated were grain yield, panicle length, and grain length x grain width ratio, which were used as response variables and ten other explanatory variables. The artificial intelligence approach in Leopoldina provided a higher estimate for the predictive variables PL and GY in the RBF procedure, 83.44% and 78.90%, respectively. In Leopoldina and Lambari, for the response variable LGW, a maximum estimate of R^2 was approximately 100% by multiple regression and artificial intelligence approaches, and in Janaúba, with a maximum estimate of 62%. The relative contribution of auxiliary characters in rice through computational intelligence and machine learning proved to be efficient to determine the relative importance of variables in flooded rice. Characters indicated to aid in decision making are flowering, number of filled grains per panicle, and panicle length for this study. In the boosting procedure, the variables that stood out were HP, GL, PL, GP, WG, and LGW in all environments. Characters indicated to aid in decision making are flowering, number of filled grains per panicle, and panicle length for this study. In the third manuscript, the relative contribution of auxiliary characters in white oat was evaluated through

such methodologies. The experiments were carried out in southern Brazil. 78 genotypes of white oat evaluated in the years 2008 and 2009 were evaluated. Each year, it consists of without and with fungicide, so that prediction models were established in four experimental sets. The design was in randomized blocks with three replications. The characteristics evaluated were grain yield, which was used as the response variable, and ten others as explanatory variables. In the bagging and boosting procedure, it is verified that the estimates of R^2 were superior to 92.70% and 80%, respectively. The R^2 ranged from 30.14% - 96.45% and 10.57% - 94.61%, for computational intelligence and machine learning, respectively. The characters indicated to assist in decision-making are plant height, leaf rust severity, and lodging percentage for this study. It is believed that, with the use of a procedure to quantify the importance of variables, techniques based on computational intelligence and machine learning can be easily employed without demanding sophisticated computational resources.

Keywords: Artificial Neural Network. Decision Tree. Determination Coefficient. Importance of Variables.

SUMÁRIO

1. INTRODUÇÃO GERAL

O melhoramento de plantas é uma forma eficaz de aumento de produtividade de culturas. Seu objetivo é o desenvolvimento de variedades de alto rendimento com qualidades específicas de grãos, resistência a estresses abióticos e bióticos, e adaptação superior ao ambiente-alvo (YU et al., 2019; SILVA JUNIOR et al., 2021).

Quantificar a importância de variáveis em programas de melhoramento genético permite obter progresso mais rápido, realizar avaliação fenotípica extensiva do germoplasma, selecionar e predizer características que apresentam baixa herdabilidade e/ou dificuldade de medição. Embora, avaliação simultânea de características forneça ampla variedade de informações, identificar qual variável preditora é mais importante é um desafio para o melhorista. A abordagem tradicional para seleção de variáveis é baseada em regressão linear múltipla. Nela é avaliada a relação entre uma variável resposta com duas ou mais variáveis independentes (SKAWSANG et al., 2019). Entretanto, essa abordagem apresenta limitações quanto à sua capacidade de analisar dados de alta dimensão além da não-captura das relações complexas e multivariadas entre variáveis (PASWAN, 2013; PARMLEY et al., 2019; SKAWSANG et al., 2019).

A aplicação de inteligência computacional pode ser uma alternativa para a seleção de variáveis e tem sido utilizada em estudos de predição (VENTURA et al., 2012; SILVA et al., 2014, 2017; SANT'ANNA et al., 2019), classificação (SANT'ANNA et al., 2015), reconhecimento de padrões (BEUCHER et al., 2019), no processo de tomada de decisão (CARNEIRO et al., 2018, 2019; SILVA JUNIOR et al., 2021) e também para minimizar o número de preditores sem comprometer o desempenho do modelo (PARMLEY et al., 2019). Outra alternativa é o aprendizado de máquinas, que é eficiente para explorar grandes conjuntos de dados e informações contrastantes além de identificar variáveis preditoras de melhor desempenho (PARMLEY et al., 2019; SILVA JUNIOR et al., 2021).

As redes neurais artificiais (RNAs) são modelos não-lineares altamente parametrizados, com conjuntos de unidades de processamento chamados neurônios, que podem ser usados para aproximar a relação entre os sinais de entrada e saída de um sistema complexo (STEFANIAK et al., 2005). As RNAs são poderosas ferramentas de predição se comparadas aos modelos convencionais como por exemplo, regressão linear (PARUELO & TOMASEL, 1997; OLDEN et al., 2002; BECK, 2018). Além disso, reproduzem a importância de cada preditor tornando-o facilmente interpretável (ZHANG et al., 2018). No entanto, a importância de características durante o ajuste da rede é, em geral, negligenciada.

A quantificação da importância de variáveis através da rede Percepton Multicamadas (PMC) pode ser obtida através de (i) algoritmo de GARSON (1991) modificado por GOH (1995), que consiste no particionamento dos pesos de conexão de rede neural para determinar a importância relativa de cada variável de entrada na rede. Este algoritmo descreve a magnitude relativa da importância dos descritores (preditor) por meio da dissecação dos pesos sinápticos da rede neural. (ii) Avaliação da importância de variáveis (entrada) através do impacto da desestruturação ou perturbação da informação de uma determinada entrada sobre a estimativa do coeficiente de determinação.

As Redes Base Radial (RBF) também visam estimar a importância de preditores utilizando a técnica de desestruturação do coeficiente de determinação (SANTOS et al., 2018; YADAV et al., 2018; BEUCHER et al., 2019; SILVA JUNIOR et al., 2021). A rede RBF, comparada a outras redes neurais, possui uma estrutura mais simples e um algoritmo de aprendizado mais rápido (SREEKANTH et al., 2010; BASHEER & HAJMEER, 2000). Essa rede consiste em três camadas, ou seja, a camada de entrada, a camada oculta e a camada de saída (SILVA JUNIOR et al., 2021).

Outras alternativas interessantes para estudos de predição e de importância de variáveis, são as metodologias baseadas no aprendizado de máquinas, como por exemplo, as árvores de decisão (BEUCHER et al., 2019; PARMLEY et al., 2019) e seus refinamentos, como *bagging*, *random forest* e *boosting* (DEGENHARDT et al., 2019; SILVA JUNIOR et al., 2021). Tais metodologias permitem a obtenção de boas predições e a importância das características por meio de medidas baseadas, como por exemplo no índice de Gini e Entropia (HASTIE, 2009). Essas metodologias permitem a quantificação do impacto da desestruturação ou perturbação da informação de uma determinada entrada sobre a estimativa do coeficiente de determinação.

As metodologias baseadas em regressão, inteligência artificial e aprendizado de máquinas têm sido utilizadas com sucesso em estudo de predição. PARMLEY et al., (2019) avaliaram as características fenotípicas de alta dimensão em soja através da abordagem de aprendizado de máquina para predição de rendimento de sementes quanto ao desenvolvimento prescritivo de cultivares para práticas de agrícolas. SKAWSANG et al., (2019) aplicaram tais metodologias para prever a população de pragas de insetos usando fatores climáticos e fenológicos da planta hospedeira. SILVA JUNIOR et al., (2021) utilizaram para a predição de produtividade e verificação da importância de variáveis para a produtividade de grãos na cultura de arroz.

2. REVISÃO DE LITERATURA

1. Inteligência computacional no melhoramento de plantas

A inteligência computacional é a área da ciência da computação que visa simular, em máquinas, a capacidade de solucionar problemas e realizar tarefas, que são uma habilidade da inteligência natural do homem (NORVIG & RUSSELL, 2013). Atualmente, a inteligência computacional tem sido aplicada nas áreas de planejamento autônomo, jogos, reconhecimento de linguagem e resolução de problemas (FERNANDES, 2003).

No melhoramento de plantas, a aplicação de inteligência computacional vem sendo utilizado na seleção de variáveis (SILVA JUNIOR et al., 2021), predição (VENTURA et al., 2012; SILVA et al., 2014, 2017; SANT'ANNA et al., 2019), classificação (SANT'ANNA et al., 2015), reconhecimento de padrões (BEUCHER et al., 2019), e no processo de tomada de decisão (CARNEIRO et al., 2018, 2019). Dentre as diversas técnicas de inteligências computacional, encontram-se as redes neurais artificiais – RNA, que são ferramentas com grande potencial de aplicação no melhoramento genético de planta.

1.1 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 (CRUZ & NASCIMENTO, 2018). Pelo fato das RNAs serem aptas a resolver problemas de cunho gerais, tais como aproximação, classificação, categorização e predição (BRAGA et al*.,* 2007), e ser tolerantes a dados faltantes e não lineares, o que representa progresso para os estudos estatísticos e melhoramento genético (SANT'ANNA, 2014).

a-Perceptron Multicamadas (Multilayer Perceptron)

As redes Perceptron Multicamadas que apresenta uma estrutura de redes neurais que se caracteriza pela existência de camada ocultas (ou intermediárias), sendo fundamentada, em seus processos de aprendizagem, no algoritmo de treinamento *backpropagation,* ou seja, redes múltiplas camadas (BRAGA et al., 2007; HAYKIN, 2001; CRUZ & NASCIMENTO, 2018). Esse novo modelo, foi denominado Perceptron Multicamadas (MLP), como demostrado na Figura 1:

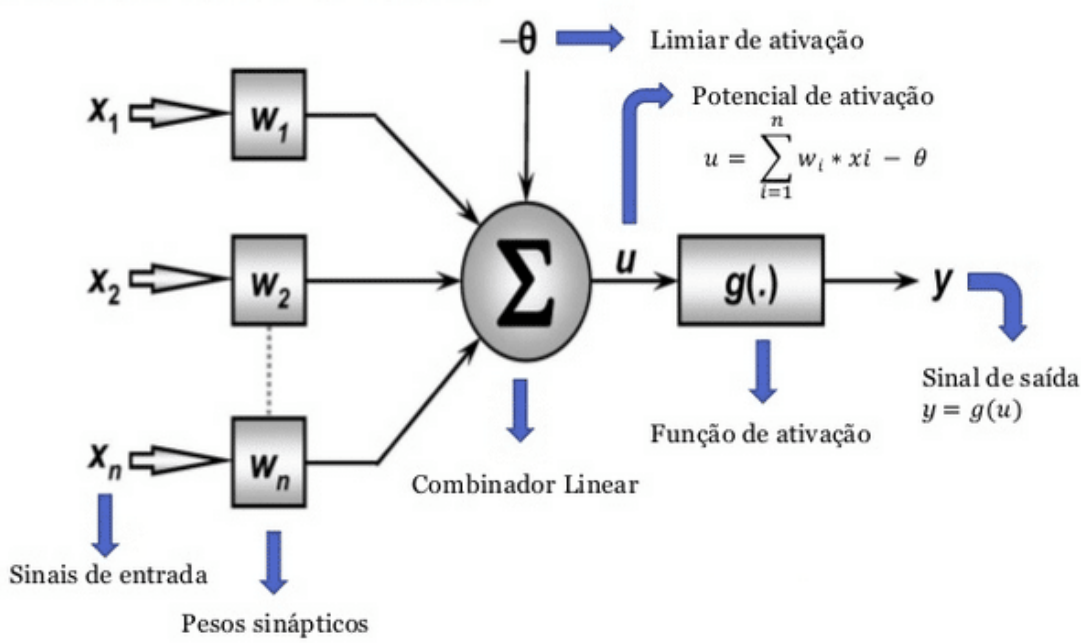


Figura 1: Modelo não linear de neurônio artificial, em que X_1, X_2,..., X_n são as entradas da rede; W_1, W_2,..., W_n são os pesos, ou pesos sinápticos, associados a cada entrada; -Θ é o limiar de ativação (bias); μ é a combinação linear dos sinais de entrada; g (.) é a função de ativação; e y é à saída da rede.

A Figura 2 apresenta uma rede neural perceptron de camada única, na qual m e n representam os números de entradas e os números de neurônios, respectivamente, x_i estão conectados a todos os neurônios em apenas uma camada e y_i é a saída de cada neurônio.

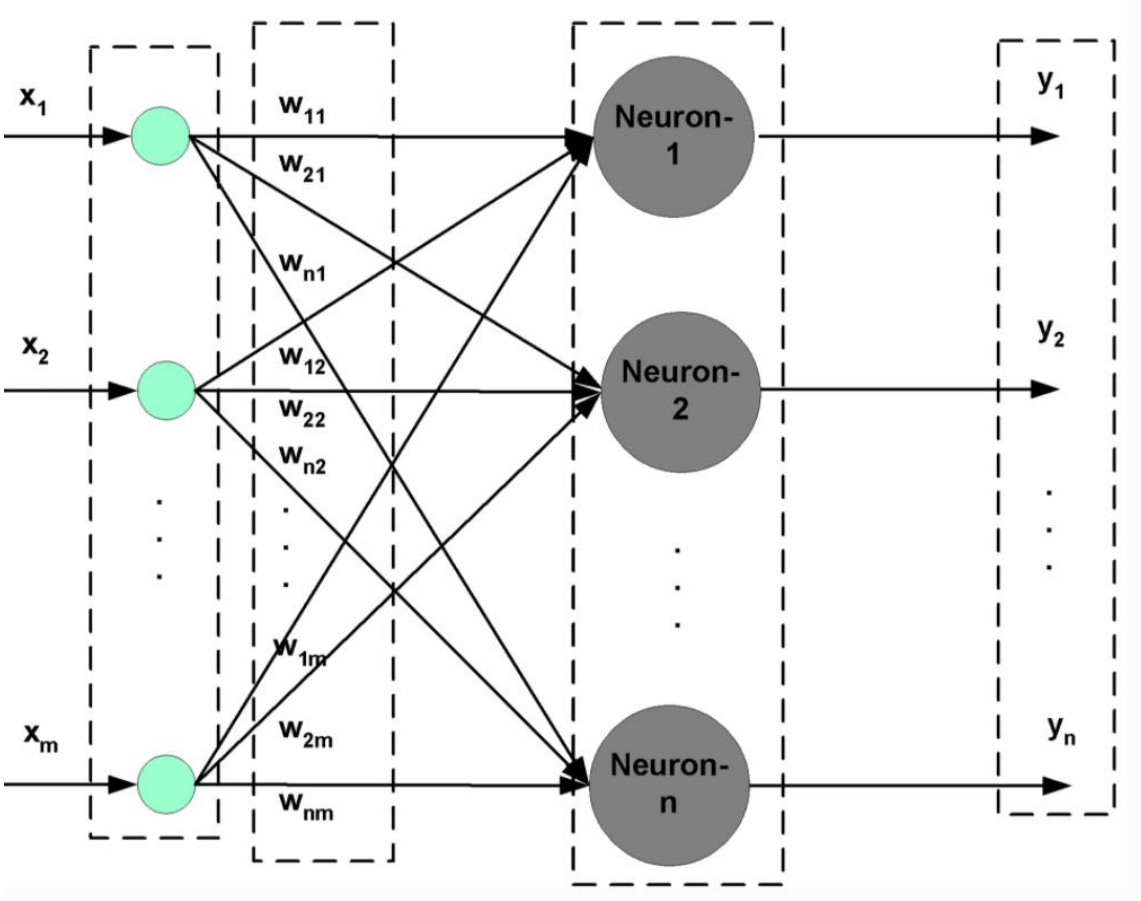


Figura 2. Estrutura da Rede Neural Perceptron. O número de saídas é o mesmo que o número de neurônios intermediário (n). Os sinais de entrada (x_1, x_2... x_m) estão conectados aos neurônios da camada única através de seus pesos correspondentes (w_{11}, w_{12} ... w_{nm}).

Os sinais de entrada (x_1, x_2... x_m) estão conectados aos neurônios da camada única através de seus pesos correspondentes (w_{11}, w_{12} ... w_{nm}), de modo que cada neurônio tenha valor de peso correspondente para cada sinal de entrada, respectivamente.

O sinal de entrada X e o pesos podem ser expresso em uma forma vetorial:

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} \text{ e } \begin{pmatrix} w_{11} & w_{12} & w_{1m} \\ w_{21} & w_{22} & w_{2m} \\ \vdots & \vdots & \vdots \\ w_{n1} & w_{n2} & w_{nm} \end{pmatrix},$$

Onde, n é o número de neurônios e m é o número de entradas.

Função de ativação

A função de ativação incorporada numa estrutura de rede dá a cada neurônio a capacidade de extrair informações não lineares e a potencialidade do aprendizado da rede. O seu papel é determinar a forma e a intensidade de alteração dos valores transmitidos de um neurônio ao outro. Portanto, é importante escolher a função de ativação correta para melhor desempenho da rede. Essas funções podem ser classificadas pelas funções de ativação utilizadas pelos neurônios: em rede homogênea na qual as funções de ativação dos neurônios na rede são iguais e rede heterogênea em que os neurônios da rede utilizam diferentes funções de ativação.

As funções de ativação amplamente utilizadas nas arquiteturas de rede neural são resumidas na Tabela 1. Entre elas, as funções de ativação mais utilizadas são funções lineares e sigmoides.

Tabela 1. Principais funções de ativação usadas nas Rede Neurais Artificias.

Nome	Fórmula	Intervalo
Linear	$f(x) = x$	(-∞, ∞)
Semi-linear	$f(x) = \begin{cases} x, & x > 0 \\ 0, & x \leq 0 \end{cases}$	(0, ∞)
Logística (sigmoidal)	$f(x) = \frac{1}{1 + e^{-ax}}$	(0,1)
Tangente hiperbólica	$f(x) = \tanh\left(\frac{x}{2}\right) = \frac{1 + e^{-x}}{1 + e^{x}}$	(-1,1)
Gaussian	$f(x) = e^{-x^2}$	(0,1)
Exponencial	$f(x) = e^{-x}$	(0, ∞)

Camadas

Nas redes PMC identifica-se três camadas: entrada, saída e intermediária (ou oculta), tendo cada uma delas uma função específica.

A camada de entrada corresponde às informações disponíveis para ser apresentada à estruturas de rede para fins de treinamento. Se houver as conexões certas entre as camadas de entrada e um conjunto suficientemente grande de camadas intermediarias, pode-se sempre

encontrar a representação que irá produzir o mapeamento correto da camada de entrada para a camada de saída, fazendo-se o uso de ajustes de pesos nas camadas intermediárias.

A camada intermediária funciona como extratoras de características contidas no conjunto de dados que são apresentados. Seus pesos traduzem a importância de um agregado de características extraídas dos padrões de entrada e permitem que a rede crie sua própria representação, mais rica e complexa, do problema. A camada de saída recebe os estímulos das camadas intermediária e constrói o padrão que será a resposta.

Em relação ao número de neurônio nas camadas intermediária não há um consenso. Geralmente, o melhor número de neurônio para a resolução de problema na área agrária é definido empiricamente, ou seja, é estabelecida uma determinada arquitetura com número fixo de camada e varia-se, de forma crescente, o número de neurônio, até se encontrar uma solução ótima (CRUZ & NASCIMENTO, 2018).

Na questão de estabelecimento do número de neurônio por camada, deve-se ter o cuidado de não utilizar unidade demais, o que pode levar a sérios problemas na fase de treinamento. Nesse caso, a rede, em vez de aprender, memoriza os dados disponibilizados (*overfitting*), decorando o padrão específico de entrada e da saída, que geralmente incluir um valor verdadeiro acrescido de certo ruído. Por outro lado, o uso de número muito pequeno de neurônio por camada, poderá demandar da rede tempo em excesso tentando buscar uma representação ótima e, como no caso anterior, proporcionar baixa eficácia em validação com a solução encontrada no treinamento.

Arquitetura de rede

A arquitetura de rede pode ser descrita como (números de neurônios na camada de entrada - camada intermediaria - saída); por exemplo, (5-8-1) estrutura significa 5 neurônios na camada de entrada, 8 neurônios ocultos na camada intermediária. Considerando uma estrutura com: 5-8-8-1; significa duas camadas ocultas com 8 neurônios ocultos, respectivamente, e 1 neurônio na camada de saída.

b- Rede Função de Base Radial- RBF

A rede de função de base radial consiste em uma estrutura mais simples que as redes de multiplica camada, uma vez que constitui uma camada de entrada, apenas uma camada

intermediária e de saída de neurônio, que é alimentada para frente, procedimento conhecido de *feedforward*. Com apenas uma camada intermediária na rede já é possível calcular uma função arbitrária qualquer a partir de dados fornecidos (HECHT-NIELSEN, 1989).

Em relação a topologia da RBF que é composta por função de ativação de base radial em sua camada intermediária. Geralmente, essas funções retornam valores cada vez menores à medida que a distância entre o ponto observado e centro da função aumentam. Dentre as funções de ativação disponíveis para redes RBF, destacam-se as multiquadráticas, multiquadráticas inversas e as funções gaussianas. A camada de saída geralmente adota a funções lineares (PARK & SANDBERG, 1991).

A descrição da função de ativação gaussiana, encontra-se na Equação 1.

$$g(u) = e^{\frac{-(u-c)^2}{2\sigma^2}} \quad (1)$$

em que, *c* é o centro da função gaussiana, σ^2 é variância da função gaussiana e *u* é o potencial de ativação.

A função de ativação linear é definida na Equação 2.

$$y_{ri} = g\left(x_0 w_0 + \sum_{j=1}^{q} f_{xj}\,(x_i) w_j\right) \quad (2)$$

em que, x_i é a i-ésima entrada; w_j é o j-ésimo peso sináptico; f_{xj} é a função de ativação da camada oculta associada à entrada x_i ponderada por seu respectivo peso.

Na Figura 3, é apresentado o esquema da rede RBF *feedforward*. A entradas X_1 a X_n na camada de entrada se referem a característica fenotípicas. Uma camada oculta com raio variando de 1 a r e número de neurônios variando de 1 a n. Na saída, a rede retorna o vetor de valores preditos (y).

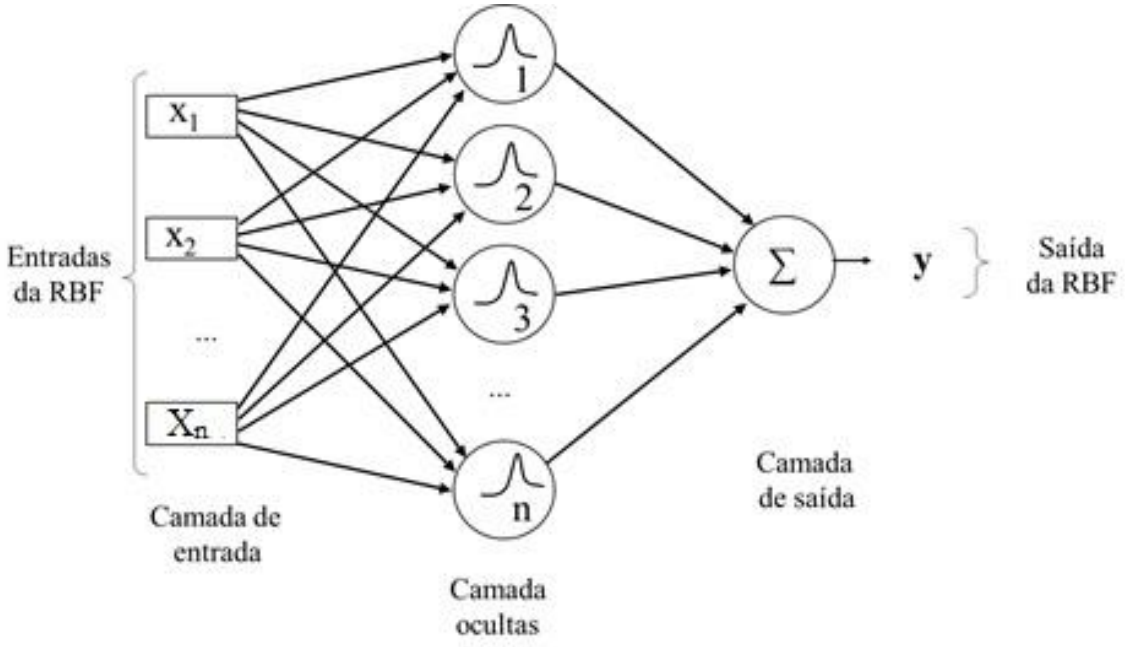


Figura 3: Esquema da rede RBF *feedforward*. Entradas X_1 a X_n na camada de entrada de característica fenotípicas. Uma camada oculta com raio variando de 1 a r e número de neurônios variando de 1 a n. Na saída, a rede retorna o vetor de valores preditos (y).

1.2 Aprendizado de máquina no melhoramento genético de planta

O aprendizado de máquina, também conhecido como aprendizagem estatística, é subcampo de inteligência artificial dedicado ao estudo de algoritmos para previsão e inferência (MOROTA et al., 2018; LIAKOS et al., 2018). Entretanto, na prática, o aprendizado de máquina visa aprender ou escolher conjunto de modelos que podem melhor prever dados não observados. O processo de predição do fenótipo a partir de conjunto de genótipos em que temos conjunto de dados compostos por pares de fenótipos e genótipos correspondente, é conhecido com aprendizado supervisionado.

A escolha do modelo com boa capacidade de predição no aprendizado supervisionado, começamos dividindo o conjunto de dados em dois conjuntos, treinamento e teste, onde este último está relacionado com conjunto de dados não disponível (MOROTA et al., 2018). Entretanto, a seleção de modelo, utilizam as informações exclusivamente do conjunto de dados de treinamento.

Árvore de decisão

A árvore de decisão (AD) é uma metodologia que particiona o espaço preditor em sub-regiões através de alguns critérios, para cada sub-região formada é atribuído 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 é dito externo (folha) quando não ocorre mais divisões dos indivíduos pertencentes a este nó (Figura 4).

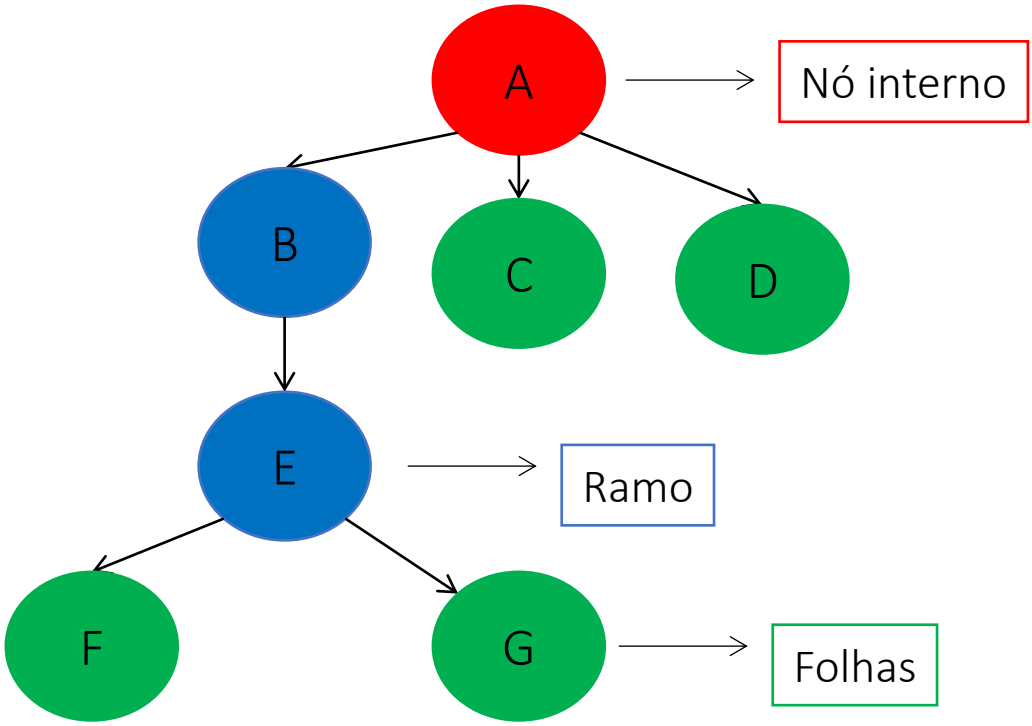


Figura 4. Árvore de decisão, com o nó interno (vermelho), ramos (azul) e folhas (verde).

A AD pode ser classificada como árvore de regressão quando a variável resposta é do tipo quantitativa (distribuição contínua), já quando a variável dependente assume valores qualitativos (distribuição discreta), a AD é classificada como árvore de classificação.

a- Árvores de regressão

A construção da árvore de regressão visa construir regiões R_1, R_2,....., R_M, em que minimiza a Soma de Quadrados dos Resíduos, descrito abaixo:

$$\sum_{m=1}^{M} \sum_{i \epsilon R_m} (y_i - \hat{y}_{R_m})^2, \quad (3)$$

em que, $\hat{y}_{R_m}$: média da variável resposta das observações de treinamento pertencente a m-ésima região.

O custo computacional é muito alto sendo inviável considerar cada partição possível do espaço em M regiões para obter o menor erro quadrático médio. Para contornar o custo computacional, (JAMES et al., 2013) recomendam procedimento baseado em divisões binárias recursivas, em que o objetivo é obter a variável X_p e o ponto s, que dívida o espaço em duas regiões, como:

$$R_1(p,s) = \{X|X_p \leq s\}\ e\ R_2(p,s) = \{X|X_p > s\} \quad (4)$$

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

Quando árvores de decisão são construídas, muitas das arestas ou sub-árvores podem refletir ruídos ou erros. Enquanto, uma árvore muito grande pode sofre *overfitting* (super-ajuste) dos dados, uma árvore pequena pode não capturar uma boa estrutura. Para detectar e excluir essas arestas e sub-árvores, são utilizados métodos de poda (*pruning*) da árvore, cujo objetivo é melhorar a taxa de acerto do modelo para novos exemplos, os quais não foram utilizados no conjunto de treinamento (HAN, 2001).

Uma abordagem para a escolha do tamanho da árvore seria construir uma árvore até que nenhuma região obtenha mais que 5 indivíduos e, em seguida, podá-la usando o custo complexidade da poda (HASTIE et al., 2009). Assim, em uma segunda etapa, é realizada a poda com o objetivo de tornar a árvore de regressão menor e menos complexa, de modo a diminuir a variância deste estimador. Nessa etapa, cada nó é retirado, um por vez, observando-se como o erro de predição varia no conjunto de validação e, posteriormente, baseando-se nas observações, é decidido quais nós permaneceram na árvore *(*HASTIE et al., 2009).

Geralmente, uma única árvore não possui boa precisão preditiva quando comparada com outras abordagens (SOUSA et al., 2020). Alguns refinamentos com o intuito de melhorar a performance do modelo de árvore de decisão são apresentados na literatura. O pior desempenho da Árvore de Decisão quando comparado com seus refinamentos pode ser explicado porque essa metodologia sofre alta variação em termos de predição (JAMES et al., 2013). HASTIE et al. (2009) enfatizaram que a baixa precisão preditiva da Árvore de Decisão pode ser melhorada pelo uso de métodos como, *Bootstrap Aggregation (Bagging), Random Forest* e *Boosting* (BREIMAN, 2001). Essas estratégias combinam múltiplas Árvores de Decisão para reduzir a variabilidade (SOUSA et al., 2020).

b- Árvores de classificação

A árvore de classificação tem como objetivo 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:

$$TEA = 1 - MAX_k(\hat{p}_{mk}) \quad (5)$$

- Índice Gini:

$$G = \sum_{k=1}^{k} \hat{p}_{mk}(1 - \hat{p}_{mk}) \quad (6)$$

- Deviance:

$$D = - \sum_{k=1}^{k} \hat{p}_{mk} log \hat{p}_{mk} \quad (7)$$

em que, $\hat{p}_{mk}$: representa a proporção de observações na m-ésima região pertencentes a k-ésima classe.

Em relação a construção da árvore de classificação é indicado utilizar o índice Gini ou Deviance, uma vez que 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 (SOUSA et al., 2020). Para evitar o super-ajustamento 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).

As RNAs apesar de apresentarem eficiência satisfatória, demandam muito recurso computacional. Por outro lado, as Árvores de Decisão e seus refinamentos (*Boosting*, *Bagging*, *Random Forest*) demandam menos recurso computacional (JAMES et al., 2013; SOUSA et al., 2020). Além disso, como as RNAs, as Árvores de Decisão e seus refinamentos não necessitam de pressuposições sobre o modelo (SOUSA et al., 2020; SILVA JUNIOR et al., 2021). Entretanto, tais metodologias apresentam boa performance preditiva (JAMES et al., 2013), permitindo a não-linearidade dos dados e também são de fácil interpretação (PRASAD et al., 2006), por fornecerem as informações sobre quais atributos são mais importantes para predição ou classificação (EBRAHIMI et al., 2011; BEIKI et al., 2012; HOSSEINZADEH et al., 2012).

c- Refinamento da árvore de decisão

1- Bagging

O problema apresentado pela árvore de decisão é a alta variabilidade entre os resultados obtidos, uma vez que utilizar parte de banco de dados para construirmos uma árvore e em seguida utilizarmos a outra parte do mesmo banco de dados para construir a segunda árvore. Sua construção obtém duas árvores com estruturas diferentes. Uma forma de contornar esse problema, é obter várias amostras de uma mesma população, construir várias árvores e em seguida obter a média/moda dos valores preditos.

A obtenção de vários conjuntos de treinamento de uma população não é uma tarefa fácil. Uma alternativa é a utilização de *Bootstrap Aggregation* (*Bagging*) (BREIMAN, 2001) é um método que aplica a técnica de *bootstrap*. Essa técnica consiste em obter B amostras com reposição da amostragem disponível, obtendo assim B modelos $\hat{f}^1(x), \hat{f}^2(x), \dots, \hat{f}^B(x)$ (EFRON, 1992). A amostragem é feita com a substituição dos dados originais e a formação de novos conjuntos de dados. Os novos conjuntos de dados podem ter uma fração das colunas e das linhas, que geralmente são hiperparâmetros em um modelo. Assim, utilizam-se os modelos gerados para obter uma média, e reduzir a variabilidade obtida nas árvores de decisão (BREIMAN, 2001). Essa média desses modelos irá ser o modelo final, e é dada por:

$$\hat{f}_{médio}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(x) \qquad (8)$$

A técnica de *Bagging* consiste em reduzir a variância das previsões, que combina o resultado de vários classificadores, modelados em diferentes sub-amostras do mesmo conjunto de dados (BREIMAN, 2001). Dessa forma, proporciona frações de linha e coluna menores que 1, em que auxilia na montagem de modelos robustos, menos propensos a *overfitting*. A quantidade de árvores utilizadas no *Bagging* não é um parâmetro que irá resultar num super-ajustamento do modelo, na prática é utilizado uma quantidade onde o erro tenha estabilizado (JAMES et al*.,* 2013).

2- Random Forest

Sob o ponto de vista em utiliza todas as variáveis em cada partição no *Bagging*, as predições obtidas nas Árvores de Decisão estarão altamente correlacionadas, uma vez que as árvores criadas terão estruturas semelhantes. Além disso, está sujeito a quase sempre uma mesma variável esteja no topo da árvore (HASTIE et al., 2009; JAMES et al., 2013). A média de valores altamente correlacionados, não resultada numa grande redução da variância, como ocorre quando é feita com valores não correlacionados (JAMES et al., 2013). Entretanto, ao

melhorar a estimativa de acurácia na classificação dos indivíduos, HO (1995) propôs o *Random Forest* (RF). O *Random Forest* é um método de aprendizagem de máquina versátil, e capaz de executar tarefas de regressão e classificação (CART). Essa metodologia também aplica métodos de redução dimensional, trata valores faltantes, valores anómalos (‘*outliers*’) e outras etapas essenciais da exploração de dados. É um tipo de método de aprendizado onde um grupo de modelos fracos é combinado para formar um modelo mais forte (HO, 1995).

O *Random Forest* apresenta o mesmo princípio de *Bagging*, uma vez que o conjunto de dados alteram as observações, e a quantidade de variáveis preditoras (m<p) utilizadas em cada partição. Diante disso, *Random Forest* obtêm-se os valores preditos mais independentes, uma vez que proporciona redução da variabilidade encontrada nas árvores de decisão. HASTIE et al. (2009) sugerem que o número de variáveis preditora utilizadas em cada partição seja $m = \sqrt{p}$ para árvore de classificação e $m = p/3$ para árvores de regressão. Assim, as predições das árvores se tornam menos correlacionadas e ainda, corrige o fato de que apenas uma variável esteja sempre no topo da árvore.

O processo de modelagem do algoritmo RF é o seguinte: adotar uma técnica de amostragem de bootstrap para extrair informação dos dados de treinamento em relação a N estimadores do conjunto de dados original. O conjunto de treinamento é cerca de $\frac{2}{3}$ do tamanho do conjunto de dados original. A amostra de bootstrap da floresta aleatória durante o processo de treinamento terá cerca de $\frac{1}{3}$ dos dados não extraídos. Essa parte dos dados é chamada de dados fora do saco. O passo seguinte é criar uma árvore de regressão para cada conjunto de treinamento bootstrap.

Um total de árvores de regressão de N estimadores são construídas para formar uma "floresta", mas essas árvores de regressão não são podadas. No processo de crescimento de cada árvore, todos os atributos ideais não são selecionados como o nó interno para ramificação, alternativamente, o atributo ideal é selecionado a partir dos atributos de profundidade Max selecionados aleatoriamente para ramificação. Assim, o algoritmo RF aumenta a diferença entre os modelos de regressão construindo diferentes conjuntos de treinamento, melhorando assim a capacidade de predição de extrapolação do modelo de regressão combinada. Através do treinamento do modelo n-tempo, é obtida uma sequência de modelo de regressão, que é usada para formar um sistema modelo de multirregressão (Forest). E então, coletar os resultados de predição da árvore de regressão dos estimadores N, adotar uma estratégia média simples para calcular o valor da nova amostra. A fórmula de decisão de regressão final é a seguinte: onde representa o modelo combinado de regressão, é um modelo único de regressão de árvores de

decisão, K é o número de árvores de regressão (N estimadores). O processo de modelagem aleatória do algoritmo da floresta é mostrado na Figura 5.

$$\{t_1(x), t_2(x), \dots, k(x)\}$$

$$\hat{f}_{rf}^{k}(x) = \frac{1}{k}\sum_{k=1}^{k} t_i(x)$$

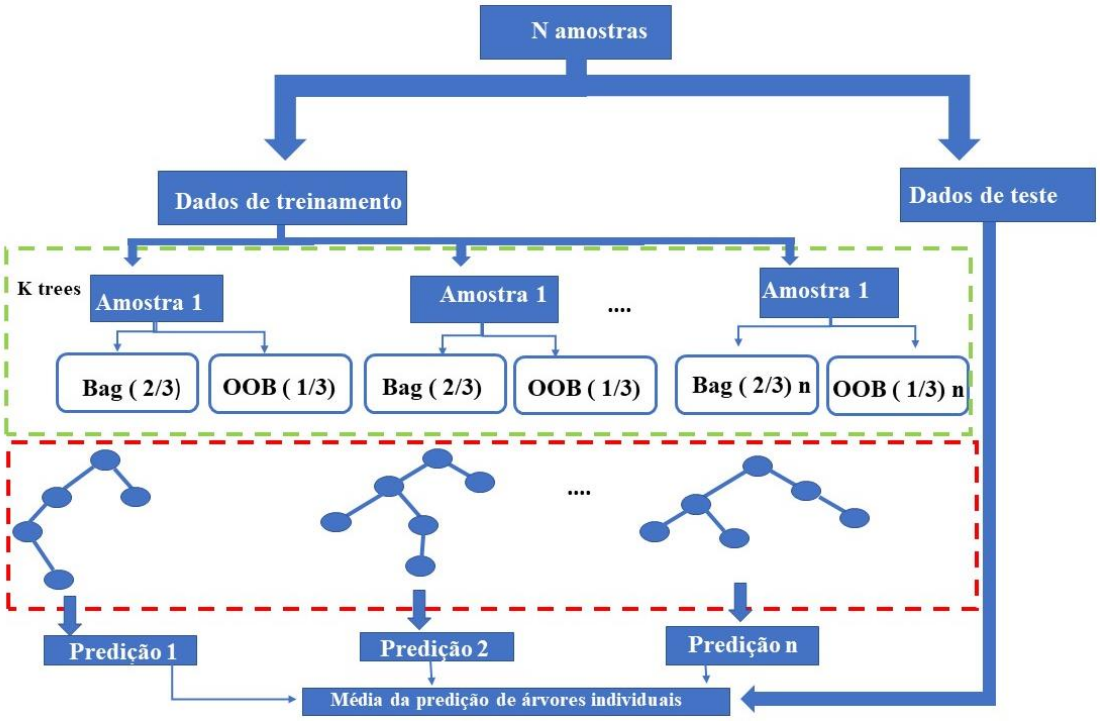


Figura 5. Fluxo de regressão da floresta aleatória.

3- Boosting

O *Boosting* é uma metodologia utilizada para aperfeiçoar a performance obtida por uma única árvore. Ao contrário do *Bagging* que cria múltiplas árvores independentes, o *Boosting* cria árvores sequencialmente utilizando-se de informação prévia da árvore anterior (SOUSA et al., 2020). Essa metodologia ajusta grande número de árvore de decisão, , $\hat{f}^1, \hat{f}^2, \dots, \hat{f}^B$, para o resíduo atual (FREUND & SCHAPIRE, 1999), ao invés de ajustar modelo para a variável resposta Y.

O processo de aprendizagens desta metodologia é considerado lento, uma vez que necessita de vários modelos (B). Entretanto, é necessário ter cuidado para não criar *overfitting* do modelo. Assim, no *Boosting* é utilizada a validação cruzada para se escolher o número de

árvores que será construída, isso reduz a possibilidade de *overfitting*, uma vez que todos os indivíduos participarão do conjunto de validação (BENGIO & GRANDVALET, 2004).

O classificador do *Boosting*, encontra-se descrito na Equação a seguir:

$$\mathrm{H(x)} = \sum_t \propto_t h_t(x), \tag{9}$$

em que, visa minimizar a função de perda através da otimização do escalar $\propto_t$ (importância atribuída a $\mathrm{h_t(x)}$) e do classificador individual $\mathrm{h_t(x)}$ (árvore de decisão individual) a cada iteração t (FREUND & SCHAPIRE, 1999).

Os classificadores individuais $\mathrm{h_t(x)}$ possuem poder classificatório baixo, porém quando utilizados conjuntamente H(x), apresentam bons resultados (MARTINS et al., 2009; SOUZA et al., 2020).

1.3 Metodologias para predição e verificação de importância de características

a- Regressão Múltipla

A regressão múltipla de *Stepwise* é o método de seleção de variáveis que visa explicar a relação entre conjunto de variáveis independentes (x) e em relação a uma variável dependente (y). A descrição do modelo de regressão é apresentada na Equação 10:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{10}$$

Em que, y é a variável resposta, x_1 a x_k são as variáveis explicativas, β_0 representa o intercepto, β_1 e β_k são as inclinações relacionada entre y e x_1 a x_k, e ε erro residual.

O coeficiente de determinação (R^2) visa estimar o quanto da variável independente é explicada pela variação total da variável dependente. Assim, tem-se:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{11}$$

Em que, y: sãos os valores reais e $\hat{y}$ preditos.

b- Inteligência computacional para importância de variáveis

Perceptron Multicamadas- PMC

A importância de variáveis através da rede PMC pode ser utilizadas duas técnicas. A primeira fundamentada no algoritmo de GARSON (1991) modificado por GOH (1995) que consiste no particionamento dos pesos de conexão de rede neural para determinar a importância relativa de cada variável de entrada dentro da rede. Este algoritmo descreve a magnitude relativa da importância dos descritores (preditor) em sua conexão com variáveis de resultado por meio da dissecação dos pesos sinápticos da rede neural.

A Equação da importância relativa de variáveis é igual:

$$IR = WV \tag{12}$$

Matricialmente, tem-se:

$$\begin{pmatrix} IR_1 \\ IR_2 \\ \vdots \\ IR_n \end{pmatrix} = \begin{pmatrix} w_{11} & w_{12} & w_{1m} \\ w_{21} & w_{22} & w_{2m} \\ \vdots & \vdots & \vdots \\ w_{n1} & w_{n2} & w_{nm} \end{pmatrix} \dots \begin{pmatrix} w_{11} & w_{12} & w_{1m} \\ w_{21} & w_{22} & w_{2m} \\ \vdots & \vdots & \vdots \\ w_{n1} & w_{n2} & w_{nm} \end{pmatrix} \dots \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$$

Em que, w_n representa o peso do neurônio de entrada em n-ésimo neurônio; IR: importância relativa da variável no n-ésimo neurônio; v_n é o peso do neurônio intermediário para a saída no n-ésimo neurônio e cada matriz n x m demonstra a camada intermediária.

Resumindo, tem-se:

$$IR = \begin{pmatrix} IR_1 \\ IR_2 \\ \vdots \\ IR_k \end{pmatrix} = \left(W_{N_1 E}^{1}\right)' \left(W_{N_2 N_1}^{2}\right)' \dots \left(W_{N_{c-1} y}^{c}\right)'$$

Em que, W_x^c representa a matriz de pesos do neurônio da camada c, considerando N_j neurônios e N_{j-1} entradas; E é o primeiro neurônio que inicia de entradas; y refere-se a camada de saída desejada e IR: importância relativa da variável.

Estes pesos que conectam aos neurônios em uma RNA são parcialmente análogos aos coeficientes em um modelo linear generalizado (OLDEN et al., 2004). Os efeitos combinados dos pesos representam a importância relativa dos preditores em suas associações com a variável resposta. Os pesos correspondem a influência relativa das informações que são processadas na rede de forma que as variáveis de entrada que não são relevantes em sua correlação com uma variável de resposta sejam suprimidas pelos pesos. Por outro lado, o efeito oposto é visto para

pesos atribuídos a variáveis explicativas que têm associações fortes e positivas com uma variável de resposta.

O método proposto por GARSON 1991 modificado por GOH 1995 identifica a importância relativa das variáveis explicativas para variáveis de resposta específicas em uma rede neural supervisionada. A importância relativa (ou força de associação) de uma variável explicativa específica para uma variável de resposta específica pode ser determinada identificando todas as conexões ponderadas entre os nós de interesse.

Na Figura 4, é demostrado uma rede PMC com três neurônios de entrada (1, 2 e 3), dois neurônios ocultos (A e B) e um neurônio de saída (o) (3-2-1). A importância relativa de variáveis através das redes neurais artificiais pelo algoritmo de GARSON (1991) modificado por GOH (1995), será descrito a seguir.

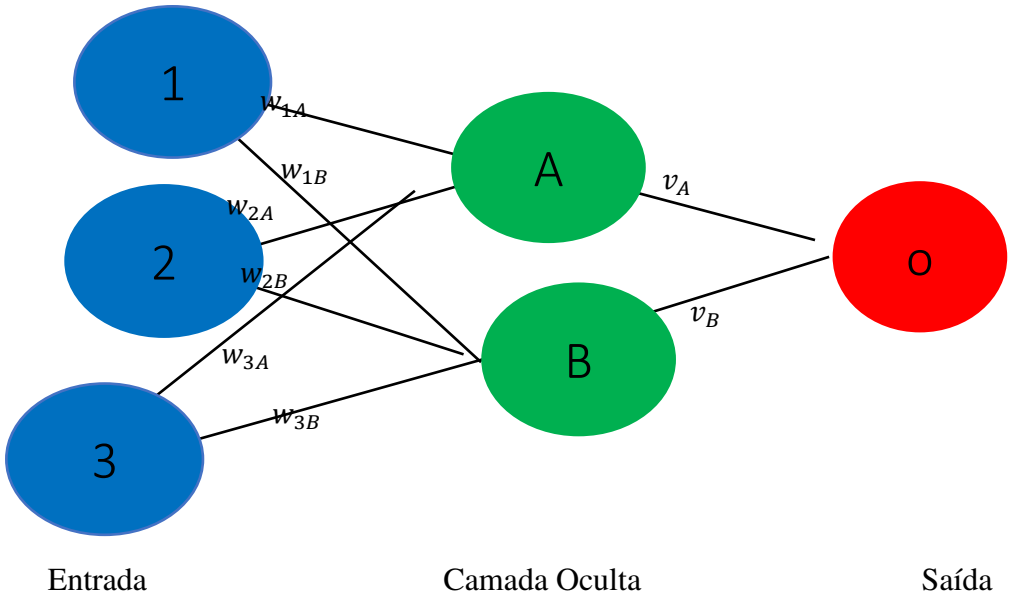


Figura 4. Algoritmo de Garson (1991) modificado por Goh (1995) para particionar e quantificar pesos de conexão de rede neural. Cálculos de amostra mostrados para três neurônios de entrada (1, 2 e 3), dois neurônios ocultos (A e B) e um neurônio de saída (o).

A matriz contendo os pesos de conexão de neurônios das camadas entrada-oculta- saída, encontra-se na Tabela 2. Nesta Tabela, será estimado a contribuição de cada neurônio de entrada na saída através de cada neurônio oculto, e o produto das conexões entrada-oculto e oculta-saída.

Tabela 2. Matriz contendo os pesos de conexão de neurônios das camadas entrada-oculta-saída.

Entrada	Oculta	
	A	B
1	w_{1A}= - 2.65	w_{1B}= - 1.25
2	w_{2A}= 0.15	w_{2B}= - 0.95
3	w_{3A}= - 0.70	w_{3B}= -0.40
Saída	v_A= 1.10	v_B= -0.40

$$C_{1A} = w_{1A}\, v_A = \text{-2.65 x 1.10 = -2.91}$$

$$C_{1A} = w_{2A}\, v_A = \text{0.15 x 1.10 = 0.165}$$

$$C_{3A} = w_{3A}\, v_A = \text{-0.70 x 1.10 = -0.77}$$

$$C_{1B} = w_{1B}\, v_B = \text{-1.25 x (- 0.40) = 0.50}$$

$$C_{1B} = w_{2B}\, v_B = \text{- 0.95 x (-0.40) = 0.38}$$

$$C_{3B} = w_{3B}\, v_B = \text{-0.40 x (- 0.40) = 0.16}$$

A contribuição relativa de cada neurônio é estimada pelo módulo de cada contribuição da entrada na camada oculta sobre o módulo do somatório de todas as contribuições, conforme descrito:

$$r_{A1} = \frac{|C_{1A}|}{|C_{1A} + C_{2A} + C_{3A}|} = \frac{|-2.91|}{(|-2.91| + |0.165| + |-0.77|)} = 0.76$$

$$r_{A2} = \frac{|C_{2A}|}{|C_{1A} + C_{2A} + C_{3A}|} = \frac{|0.165|}{(|-2.91| + |0.165| + |-0.77|)} = 0.04$$

$$r_{A3} = \frac{|C_{3A}|}{|C_{1A} + C_{2A} + C_{3A}|} = \frac{|0.77|}{(|-2.91| + |0.165| + |-0.77|)} = 0.20$$

$$r_{B1} = \frac{|C_{1B}|}{|C_{1B} + C_{2B} + C_{3B}|} = \frac{|0.50|}{(|0.50| + |0.38| + |0.16|)} = 0.48$$

$$r_{B2} = \frac{|C_{2B}|}{|C_{1B} + C_{2B} + C_{3B}|} = \frac{|0.38|}{(|0.50| + |0.38| + |0.16|)} = 0.37$$

$$r_{B3} = \frac{|C_{3B}|}{|C_{1B} + C_{2B} + C_{3B}|} = \frac{|0.16|}{(|0.50| + |0.38| + |0.16|)} = 0.15$$

Na Tabela 3, representa a contribuição relativa de cada neurônio e a soma das contribuições relativas de cada entrada. Através da soma das contribuições relativas é possível

estimar a porcentagem da importância relativa de cada variável. A estimativa da importância relativa de uma entrada é estimada pela soma da contribuição relativa desta entrada sobre as demais somatório das contribuições relativas.

Tabela 3. Contribuição relativa de cada neurônio e soma das contribuições relativas.

Entrada	Oculta		Soma
	A	B	
1	0.76	0.48	S_1= 1.24
2	0.04	0.37	S_2= 0.41
3	0.20	0.15	S_3= 0.35

$$RI_1 = \frac{S_1}{S_1 + S_2 + S_3} x\ 100 = \frac{1.24}{1.24 + 0.41 + 0.35} x\ 100 = 62.0\ \%$$

$$RI_2 = \frac{S_2}{S_1 + S_2 + S_3} x\ 100 = \frac{0.41}{1.24 + 0.41 + 0.35} x\ 100 = 20.50\ \%$$

$$RI_3 = \frac{S_3}{S_1 + S_2 + S_3} x\ 100 = \frac{0.35}{1.24 + 0.41 + 0.35} x\ 100 = 77.70\%$$

c- Importância relativa de variável pelo coeficiente de determinação- R^2

A importância de variáveis (entradas) através do impacto da desestruturação ou perturbação da informação de uma determinada entrada sobre a estimativa do coeficiente de determinação. Este impacto provoca redução da variância explicada pelo modelo (R^2). Permutando uma característica quebra a associação entre o preditor e a resposta, e proporciona a redução do R^2 geral do modelo. A magnitude da redução em R^2 quando uma característica preditora é permutada reflete a força da associação entre essas características preditora e a resposta. Consequentemente, estimativa do R^2 diminui tem-se indicativo de que a variável é importante em relação às demais, para fins de predição com a rede já estabelecida.

A importância relativa da variável pela permutação do R^2 será descrita na Equação a seguir:

$$pVR_{x_i} = R^2_{obs} - \bar{R}^2_{perm,x_i} \quad (13)$$

Em que, R^2_{obs} é o R^2 do modelo RNA ajustado às variáveis preditor e resposta observadas; R^2_{perm,x_i} é o R^2 do modelo RNA ajustado ao conjunto de dados modificado onde x_i é permutado; $\bar{R}^2_{perm,x_i}$: é o valor médio de R^2_{perm,x_i} após m-ésima permutação do conjuntos de dados.

Rede Função de Base Radial – RBF

A rede função de base radial se caracteriza por contar com apenas uma camada oculta e fazer uso de função de ativação gaussiana (CRUZ & NASCIMENTO, 2018). A eficiência da predição é medida pelo coeficiente de determinação e a importância relativa de cada entrada estimada pela técnica de desestruturação da informação de cada variável explicativa, conforme já descrito para PMC.

d- Aprendizado de Máquinas para importância de variáveis

A quantificação da importância de variáveis através abordagem de aprendizado de máquinas utilizadas a árvore de decisão e os seus refinamentos, *random forest*, *bagging* e *boosting*. O coeficiente de determinação mede a qualidade do ajuste de modelo preditivo e informações do erro quadrático mínimo (MSE) são empregados para quantificar a importância de variáveis. O erro quadrático mínimo foi estimado como descrito na Equação a seguir:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (14)$$

Em que, y_i e $\hat{y}_i$ correspondem ao valor observado e predito da observação no genótipo *i*, respectivamente, e *n* o número total de observações (variável).

Nestas técnicas, a importância da variável explicativa é feita a quantificação da diminuição média da precisão de predição, que consiste na estimativa da porcentagem de incremento de erro quadrático mínimos (IMSE), que é construído quando permutamos os valores de cada variável do conjunto de dados, e comparando com a predição do conjunto de dados originais não-permutados da variável. Em analogia a análise de regressão, é o aumento médio dos quadrados dos resíduos do conjunto de dados quando a variável é permutada (LI & ZHAN, 2019). Valores mais altos de IMSE representam importância da variável mais alta.

A importância da variável pela permutação do $IMSE$ será descrita na Equação a seguir:

$$IV_{x_i} = MSE_{perm,x_i} - MSE_{nperm} \quad (15)$$

Em que, MSE_{perm,x_i} é a permutação dos valores de cada variável do conjunto de dados onde x_i é permutado; MSE_{nperm}: valores da estimativa dos dados originais não-permutados da variável.

3. REFERÊNCIAS BIBIOGRÁFICAS

Beck M. NeuralNetTools: Visualization and Analysis Tools for Neural Networks**.** 2018; R package version 1.5.2. http://dx.doi.org/10.18637/jss.v085.i11.

Beucher A, Møller AB, Greve MH. Artificial neural networks and decision tree classification for predicting soil drainage classes in Denmark, Geoderma. 2019; 352:351-359. http://dx.doi.org/10.1016/j.geoderma.2017.11.004.

Carneiro ART, Sanglard DA, Azevedo AM, Souza TLPO, Pereira HS & Melo LC. 2019. Fuzzy logic in automation for interpretation of adaptability and stability in plant breeding studies. Scientia Agricola, 76:123-129. https://doi.org10.1590/1678-992x-2017-0207.

Carneiro VQ, Prado AL, Cruz CD, Carneiro PCS, Nascimento M, Carneiro JES. 2018. Fuzzy control systems for decision-making in cultivars recommendation. Acta Scientiarum. Agronomy 40: 1-8. http://dx.doi.org/10.4025/actasciagron.v40i1.39314.

Cruz CD, Nascimento M. Inteligência Computacional aplicada ao melhoramento genético. 1st ed. Vicosa: Editora UFV, 2018. 414p.

Garson GD. Interpreting neural network connection weights. Artificial Intelligence Expert. 1991; 6:46-51.

Goh ATC. Back-propagation neural networks for modeling complex systems. Artificial Intelligence in Engineering. 1995; 9:143-51. http://dx.doi.org/10.1016/0954-1810(94)00011-S.

Li, L, & Zha, Y. (2019). Estimating monthly average temperature by remote sensing in China. Advances in Space Research. 63(8), 2345-2357. https://doi.org/10.1016/j.asr.2018.12.039.

Olden JD, Jackson DA. “Illuminating the “Black Box”: A Randomization Approach for Understanding Variable Contributions in Artifical Neural Networks.” Ecological Modelling. 2002; 154:135–150. http://dx.doi.org/10.1016/s0304-3800(02)00064-9.

Paliwal M and Kumar UA. Assessing the contribution of variables in feed forward neural network. Applied Soft Computing. 2011;11: 3690-3696.

Parmley KA, Higgins RH, Ganapathysubramanian B et al. 2019. Machine Learning Approach for Prescriptive Plant Breeding. Sci Rep 9, 17132. http://dx.doi.org/10.1038/s41598-019-53451-4.

Paruelo JM, Tomasel F. "Prediction of Functional Characteristics of Ecosystems: A Comparison of Artificial Neural Networks and Regression Models." Ecological Modelling. 1997; 98: 173–186. http://dx.doi.org/10.1016/s0304-3800(96)01913-8.

Paswan RP and Begum SA. Regression and Neural Networks Models for Prediction of Crop Production. Int. J. Sci. Eng. Res.2013; 4:11.

Sant'Anna IC, Ferreira RADC, Nascimento M, Carneiro VQ, Silva GN, Cruz CD, Oliveira MS, Chagas FEO. Multigenerational prediction of genetic values using genome-enabled prediction. PLoS One. 2019. 14, e0210531. http://dx.doi.org/10.1371/journal.pone.0210531.

Sant'Anna IC, Tomaz RS, Silva GN, Nascimento M, Bhering LL, Cruz CD. Superiority of artificial neural networks for a genetic classification procedure. Genet. Mol. Res. 2015, 14, 9898–9906.

Silva GN, Tomaz RS, Sant'Anna IC, Carneiro VQ, Cruz CD, Nascimento M. Evaluation of the efficiency of artificial neural networks for genetic value prediction. Genet. Mol. Res. 2016, 15, 1–11.

Silva GN, Tomaz RS, Sant'anna IC, Nascimento M, Bhering LL and Cruz CD. Neural networks for predicting breeding values and genetic gains. Scientia Agricola. 2014; 71, 494-498. http://dx.doi.org/10.1590/0103-9016-2014-0057.

Silva Junior AC, Sant'Anna IC, Silva GN, Cruz CD, Nascimento M, Lopes LB, Soares PC. Computational intelligence and machine learning to study the importance of characteristics in flood-irrigated rice. 2021 a. Acta Scientiarum-Agronomy (*in prelo*).

Silva Júnior AC, Silva MJ, Cruz CD, Santanna IC, Silva GN, Nascimento M, Azevedo C F. Prediction of the importance of auxiliary traits using computational intelligence and machine learning: A simulation study. PLoS One, 21, p. EMID:a920715476, 2021 b.

Skawsang S, Nagai M, Nitin K and Soni P. Predicting Rice Pest Population Occurrence with Satellite-Derived Crop Phenology, Ground Meteorological Observation, and Machine

Learning: A Case Study for the Central Plain of Thailand. Appl. Sci. 2019; 9:4846. http://dx.doi.org/10.3390/app9224846.

Skawsang S, Nagai M, Nitin K and Soni P. Predicting Rice Pest Population Occurrence with Satellite-Derived Crop Phenology, Ground Meteorological Observation, and Machine Learning: A Case Study for the Central Plain of Thailand. Appl. Sci. 2019; 9:4846. http://dx.doi.org/10.3390/app9224846.

Sousa IC, Nascimento M, Silva GN, Nascimento ACC, Cruz CD, Fonseca F, Almeida DP, Pestana KN, Azevedo CF, Zambolim L and Caixeita ET. Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. Scientia Agricola 2020; 78: 1–8. https://doi.org/10.1590/1678-992x-2020-0021.

Sreekanth S, Ramaswamy HS, Sablani SS, Prasher SO. A neural network approach for evaluation of surface heat transfer coefficient. J. Food Proc. Preserv. 2010; 23: 329-348.

Stefaniak B, Cholewiński W, Tarkowska A. Algorithms of Artificial Neural Networks - Practical application in medical science. Polski Merkuriusz Lekarski. 2005; 19:819-22.

Tan K, Li E, Du Q, Du P. An efficient semi-supervised classification approach for hyperspectral imagery. ISPRS Journal of Photogrammetry and Remote Sensing. 2014; 97:36–45. http://dx.doi.org/10.1016/j.isprsjprs.2014.08.003.

Ventura RV, Silva MA, Medeiros TH, Dionello NL, Madalena FE, Fridrich AB, Valente BD, Santos GG, Freitas LS, Wenceslau RR, Felipe VPS & Corrêa GSS. Use of artificial neural networks in breeding values prediction for weight at 205 days in Tabapuã beef cattle. Arquivo Brasileiro de Medicina Veterinária e Zootecnia. 2012; 64:411-418. http://dx.doi.org/10.1590/S0102-09352012000200022.

Yu H, Campbell MT, Zhang Q, Walia H, and Morota G. Genomic Bayesian confirmatory factor analysis and Bayesian network to characterize a wide spectrum of rice phenotypes. G3: Genes, Genomes, Genetics. 2019; 9:1975-1986. http://dx.doi.org/10.1101/435792.

CAPÍTULO 1

Prediction of the importance of auxiliary traits using computational intelligence and machine learning: A simulation study

(*in prelo: PLOS ONE: doi*:*10.1371/journal.pone.0257213*)

VIÇOSA - MINAS GERAIS

2021

RESUMO

SILVA JÚNIOR, Antônio Carlos, D.Sc., Universidade Federal de Viçosa, setembro de 2021. **Predição da importância de características auxiliares usando inteligência computacional e aprendizado de máquina: Um estudo de simulação**. Orientador: Cosme Damião Cruz.

O presente estudo avaliou a importância de características auxiliares de uma característica principal com base em informações fenotípicas e estrutura genética previamente conhecida usando inteligência computacional e aprendizado de máquina para desenvolver ferramentas preditivas para o melhoramento genético. Foram simulados uma população F_2 representada por 500 indivíduos, obtidos a partir de um cruzamento entre pais homozigotos contrastantes. Os caracteres fenotípicos simulados apresentam com base em médias previamente estabelecidas e estimativas de herdabilidade (30%, 50% e 80%). As características foram distribuídas em um genoma com 10 grupos de ligação, considerando dois alelos por marcador. Foram considerados quatro cenários diferentes. Para a característica principal, a herdabilidade constitui-se de 50%, e 40 locos de controle foram distribuídos em cinco grupos de ligação. A simulação de outras características de controle fenotípico com a mesma complexidade da característica principal, mas sem qualquer relação genética com ele e sem pleiotropia ou uma ligação fatorial entre os loci de controle. Essas características compartilhavam um grande número de locos de controle com a característica principal, mas podiam ser distinguidas pela ação diferencial do ambiente sobre elas, conforme refletido nas estimativas de herdabilidade (30%, 50% e 80%). Os coeficientes de determinação foram considerados para avaliar as metodologias propostas. As metodologias utilizadas são regressão múltipla, inteligência computacional e aprendizado de máquina para prever a importância das características testadas. A inteligência computacional e o aprendizado de máquina mostram superiores na extração de informações não lineares das entradas do modelo e na quantificação das contribuições relativas de características fenotípicas. Os valores de R^2 variaram de 44,0%-83,0% e 79,0%-94,0%, para inteligência computacional e aprendizado de máquina, respectivamente. Em conclusão, as contribuições relativas de características auxiliares em diferentes cenários em programas de melhoramento de plantas podem ser predito com eficiência usando inteligência computacional e aprendizado de máquina.

Palavras-chave: Caracteres fenotípicos; controle genético; redes neurais artificiais; coeficiente de determinação.

ABSTRACT

SILVA JÚNIOR, Antônio Carlos, D.Sc., Universidade Federal de Viçosa, September 2021. **Prediction of the importance of auxiliary traits using computational intelligence and machine learning: A simulation study**. Advisor: Cosme Damião Cruz.

The present study evaluated the importance of auxiliary traits of a principal trait based on phenotypic information and previously known genetic structure using computational intelligence and machine learning to develop predictive tools for plant breeding. Data of an F_2 population represented by 500 individuals, obtained from a cross between contrasting homozygous parents, were simulated. Phenotypic traits were simulated based on previously established means and heritability estimates (30%, 50%, and 80%); traits were distributed in a genome with 10 linkage groups, considering two alleles per marker. Four different scenarios were considered. For the principal trait, heritability was 50%, and 40 control loci were distributed in five linkage groups. Another phenotypic control trait with the same complexity as the principal trait but without any genetic relationship with it and without pleiotropy or a factorial link between the control loci for both traits was simulated. These traits shared a large number of control loci with the principal trait, but could be distinguished by the differential action of the environment on them, as reflected in heritability estimates (30%, 50%, and 80%). The coefficient of determination were considered to evaluate the proposed methodologies. Multiple regression, computational intelligence, and machine learning were used to predict the importance of the tested traits. Computational intelligence and machine learning were superior in extracting nonlinear information from model inputs and quantifying the relative contributions of phenotypic traits. The R^2 values ranged from 44.0% - 83.0% and 79.0% - 94.0%, for computational intelligence and machine learning, respectively. In conclusion, the relative contributions of auxiliary traits in different scenarios in plant breeding programs can be efficiently predicted using computational intelligence and machine learning.

Keywords: Phenotypic traits; genetic control; artificial neural networks; coefficient of determination.

1. INTRODUCTION

Plant breeding is effective in increasing the productivity of crops. Its main objective is to increase the frequency of good alleles in plant populations so that superior crops are developed, with high productivity, resistance to diseases and pests, tolerance to abiotic stresses, and superior adaptation to environments (YU et al., 2019). Quantifying the importance of traits in the culture allows the breeder to guide strategies for selection and acceleration of progress via indirect selection, to perform an extensive phenotypic evaluation of the germplasm, and to predict the future performance of those characteristics that present low heritability and/or difficulty in measurement.

One way to increase the efficiency of selection is through select for secondary traits that are easy to measure, have high heritability, and are highly correlated with the principal trait (BÄNZIGER et al., 2000). Indirect selection through a secondary trait may be more efficient than direct selection if h^2 is higher for the secondary trait than for the primary trait and if the genetic correlation between the primary and secondary trait is sufficiently high (ZIYOMO & BERNARDO, 2013).

Although the simultaneous evaluation of traits in the plant breeding program provides a large amount of information, identifying which phenotypic predictive trait is the most important is a challenge for the breeder. The traditional methodology for selecting phenotypic traits is based on multiple linear regression, for example, mixed models that are many used in plant breeding programs (KHAKI et al., 2020). Its principles evaluate the relationship between a response phenotypic characteristic with two or more independent phenotypic characteristics (SKAWSANG et al., 2019). However, this methodology has limitations regarding its ability to analyze high-dimensional data, in addition to not capturing complex and multivariate relationships between phenotypic traits (PASWAN, 2013; MONTESINOS et al., 2017; PARMLEY et al., 2019; SKAWSANG et al., 2019).

The application of computational intelligence can be an alternative to predict complex traits from auxiliary traits (VENTURA et al., 2012; SILVA et al., 2014, 2017; SANT'ANNA et al., 2019). Artificial neural networks (ANN) are highly parameterized nonlinear models, with sets of processing units called neurons, which can be used to approximate the relationship between the input and output signals of a complex system (STEFANIAK et al., 2005). The ANN's are powerful prediction tools when compared to conventional models such as linear regression (PARUELO & TOMASEL 1997; OLDEN et al., 2002; BECK, 2018). Also, they reproduce the importance of each predictive trait making it easily interpretable (ZHANG et al.,

2018). Although presenting the high potential for prediction, in general, ANN's are neglected for studies on the importance of traits.

Multilayer Perceptrons (MLP), and radial basis function (RBF) networks are the most commonly used ANN's. The MLP classifier is a typical architecture of ANN with at least a hidden layer and an output layer, where both layers have nonlinear and differentiable transfer functions (SANTOS et al., 2018; YADAV et al., 2018; BEUCHER et al., 2019; KECMAN, 2001). The RBF network, compared to other neural networks, has a simpler structure and a faster learning algorithm (SREEKANTH et al., 2010; BASHEER & HAJMEER, 2000). The network consists of three layers, i.e, the input layer, the hidden layer, and the output layer. A large amount of nonlinear information is accepted by the input layer and then transmitted through the hidden layer. Finally, the output result is produced from the output layer (SREEKANTH et al., 2010).

Another interesting alternative for predicting and quantifying the importance of auxiliary traits are machine learning-based methodologies, for example, decision trees (BEUCHER et al., 2019; PARMLEY et al., 2019) and their refinements, such as bagging, random forest, and boosting (DEGENHARDT et al. 2019). Such methodologies allow obtaining predictions of the importance of traits using measures based on the Gini, and Entropy index, for example (HASTIE, 2009). Computational intelligence, machine learning, and multiple regression techniques have been shown to be efficient in the study of prediction in various agricultural crops, for example, in soybeans, in which the phenotypic characteristics were studied for seed yield prediction regarding row spacing and seeding density (PARMLEY et al., 2019); for rice, SKAWSANG et al. (2019) applied methodologies that aim to compare and predict the pest population using climatic and phenological factors of the host plant.

Given the above, this work aimed to study the importance of auxiliary traits from a principal trait, considering a set of phenotypic information, with previously known genetic structure, using computational intelligence and machine learning methodologies, for the development of predictive tools useful to plant breeding programs.

2. MATERIAL AND METHODS

1- Dataset

A set of simulated data of an F_2 population represented by 500 individuals, derived from a cross between contrasting homozygous parents, was used.

2- Phenotyping

Eleven phenotypic traits (PTs) were simulated using previously established means and h^2 estimates. The h^2 values used were 30%, 50%, and 80%, respectively (Table 1). The traits were established by the action of 40 allele loci based on 1,000 markers in 10 linkage groups (LGs) with differential additive effects. Previous simulation studies used fewer than 20 quantitative trait loci (QTLs) (GIANOLA et al., 2011; LONG et al., 2010, 2011); therefore, we explored any number of QTLs in the present study. The weights of importance of the loci on the total genotypic variability of the traits were established from a uniform distribution.

PT1 was used as the principal trait of prediction with 50% heritability and 40 control loci of the trait distributed in LGs 1, 2, 3, 4, and 5 (Table 1). Ten auxiliary traits with known genetic control loci were also considered. PT2 was simulated assuming the same complexity as PT1 but without any genetic relationship to PT1 and without pleiotropy or a factorial link between the control loci of PT1 and PT2. Thus, we hypothesized that PT2 is the least important trait for prediction.

Table 1. Description of phenotypic traits (PT) in relation to heritability (h^2) and the distribution of linkage groups (LG).

PT	h^2	LG1	LG2	LG3	LG4	LG5	LG6	LG7	LG8	LG9	LG10
1	0.5	8	8	8	8	8	-	-	-	-	-
2	0.5	-	-	-	-	-	8	8	8	8	8
3	0.3	4	4	4	4	4	4	4	4	4	4
4	0.5	4	4	4	4	4	4	4	4	4	4
5	0.8	4	4	4	4	4	4	4	4	4	4
6	0.3	4	4	4	-	-	8	8	4	4	4
7	0.5	4	4	4	-	-	8	8	4	4	4
8	0.8	4	4	4	-	-	8	8	4	4	4
9	0.3	4	-	-	-	-	8	8	8	8	4
10	0.5	4	-	-	-	-	8	8	8	8	4
11	0.8	4	-	-	-	-	8	8	8	8	4

Phenotypic traits (PT) were simulated with previously established mean and heritability (h^2). The h^2 values used were: 30, 50 and 80%. The traits were established by the action of 40 locus alleles, from 1000 markers in 10 linkage groups (LG), with differential additive effect.

PT3, PT4, and PT5 in different scenarios represented other important traits (Table 1); they shared a large number of controlling loci with PT1 but could be distinguished by the differential action of the environment on them, as reflected in the estimates of heritability. Similarly, subsets of auxiliary traits, namely PT6, PT7, and PT8, as well as PT9, PT10, and PT11, with decreasing importance, were tested. The following statistical model was used:

$$Y_i = \mu + G_i + \varepsilon_i \quad (1)$$

where Y_i is the simulated observation of the trait of the i^{th} individual, μ is the general average of the trait, whose value is specified by the breeder, G_i is the effect associated with the i^{th} individual, with $G_i \sim N(0, \sigma_g^2)$; ε_i is the random error, where $\varepsilon_i \sim N(0, \sigma^2)$, and $\sigma^2 = (1 - h^2)\sigma_g^2/h^2$.

The total phenotypic value in the epistatic model, expressed by an individual belonging to the population, was estimated using the following equation:

$$Y_i = \mu + \sum_{j=1}^{40} p_j \alpha_j + \sum_{j=1}^{39} p_j \alpha_j \alpha_{j+1} + \varepsilon_i \quad (2)$$

here, $\alpha_j = a_i + d_i$ and $\frac{a_i}{d_i}$ = gmd, where $\mu + a_j$, $\mu + d_j$, and $(\mu - a_j)$ are the genotypic values associated with classes AA, Aa, and aa, assumed to be equal to 1, 0, and -1, respectively, when coded; p_j is the contribution of the locus j to the expression of the trait, with a uniform distribution; and d_i is average degree of dominance of each trait (d_i = 0.5).

3- Genotyping

A total of 1,000 codominant molecular markers, with two alleles per marker, were used. These markers were distributed in a genome established by 10 LGs, reflecting a diploid species with 2n = 2x = 20. Each LG was 100 centimorgans; thus, 100 markers were evenly spaced. All the markers are described in Table 2.

Table 2. Location of markers and loci controlling of traits.

LG	Markers	Loci controlling
1	1-100	10 20 30 40 50 60 70 80
2	101-200	110 120 130 140 150 160 170 180
3	201-300	210 220 230 240 250 260 270 280
4	301-400	310 320 330 340 350 360 370 380
5	401-500	410 420 430 440 450 460 470 480
6	501-600	510 520 530 540 550 560 570 580
7	601-700	610 620 630 640 650 660 670 680
8	701-800	710 720 730 740 750 760 770 780
9	801-900	810 820 830 840 850 860 870 880
10	900-1000	910 920 930 940 950 960 970 980

LG: linkage groups. For each link group, eight loci controlling were used.

Four different scenarios were considered to predict phenotypic traits (PT1). These scenarios differed in terms of the four loci controlling quantitative traits PT3 to PT11 (Table 2). Scenarios 1 (10 20 30 40) and 2 (50 60 70 80) represented the first four and the last four loci

controlling the quantitative traits, respectively (Table 2). Scenario 3 (i.e., 10 20 70 80) represented the first two and last two loci controlling quantitative traits. Finally, scenario 4 (30 40 50 60) represented the central loci controlling the quantitative traits (i.e., the first and last two loci were excluded).

4- Indirect selection gain through principal trait

Indirect selection gain through principal traits was estimated as described by Falconer (1960) using the following equation:

$$GS_{y(x)} = ih_x r_g s_{gy}$$

Where $GS_{y(x)}$is the indirect selection gain in y (x), which is selection intensity (0.9659); x is the principal trait under selection; h_x is the square root of heritability; $r_g = \frac{h_x}{h_y}$ is the absolute value of the estimated genetic correlation (estimated from genetic covariances) between the principal traits x and y; and s_{gy} is the phenotypic standard deviation.

5- Prediction of the importance of phenotypic traits

To predict PT1 and determine the importance of other traits (PT2 to PT11), multiple regression, computational intelligence, and machine learning were used.

Stepwise multiple regression

Stepwise multiple regression selects a predictor trait at the expense of the coefficient of determination (R^2) between the dependent and independent traits (AMIRI et al., 2015). For prediction using stepwise multiple regression (GHANI, 2010), PT1 was used as the principal trait and the others as auxiliary traits. The R^2 values were used to verify the extent to which the independent traits explained the total variation in the dependent trait. The following model was used to predict PT1:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \quad (3)$$

Where y is the response variable (PT1); x_1 to x_k(PT2 to PT11) are the explanatory variables; β_0 represents the intercept; β_1 and β_k are the coefficients associated with the variables x_1 to x_k; and ε is the residual effect.

R^2 was calculated using the following equation:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (4)$$

where the actual values are indicated by y and predicted values by $\hat{y}$.

Pearson's correlation analysis was used to evaluate the relationship between PT1 and other traits (MUKAKA, 2012). The first and second groups constituted the phenotypic traits PT1 and PT2, respectively. According to the number of shared control loci with the principal trait, the third group was composed of PT3, PT4, and PT5; the fourth group comprised PT6, PT7, and PT8; and the fifth group comprised PT9, PT10, and PT11.

Computational intelligence

MLP and RBF neural networks were used for information processing and prediction of the importance of phenotypic traits, as described below.

MLP

The MLP networks are characterized by having at least one intermediate (hidden) layer located between the input and output layers. For the best efficiency of this network, before training, the data were normalized by an interval between − -1 and 1. The maximum number of training periods was set to 5,000, and the minimum mean square error (MSE), to stop processing the network, was set at 1.0×10^{-3}. All trained networks included one neuron in the output layer and a single hidden layer with 30 neurons. The *sigmoid tangent* activation function was used in the hidden layer, and Bayesian regulation backpropagation was used *as the training algorithm*.

To quantify the importance of phenotypic traits using the MLP network, two methodologies were used, namely Garson's algorithm (1991) and modified by Goh (1995). In this approach, the neural network connection weights are partitioned to determine the relative importance of each input variable in the network. This function was implemented using the method described by Goh (1995). In this method, the relative importance of each variable was determined as the absolute magnitude. For each input node, all weights connecting an input through the hidden layer to the response variable are identified to return a list of all weights specific to each input variable. In the second methodology, the importance of traits (inputs) was evaluated through the impact of de-structuring or disturbance of information on a given input on the R^2 estimates. This importance was estimated by exchanging information on or making the phenotypic value of each trait constant and verifying the changes in the R^2 estimates. When

the values of a trait are disturbed, the value of R^2 decreases, indicating that the trait is important relative to the others for the purpose of prediction.

The relative importance of the variable, measured by the reduction in R^2 and obtained by permuting its values, was quantified using the following equation:

$$pVR_{x_i} = R^2_{obs} - \bar{R}^2_{perm,x_i} \quad (5)$$

where R^2_{obs} is the R^2 of the ANN topology obtained using the original predictor variables and established by the square of the correlation between the predicted and observed values, and R^2_{perm,x_i} is the R^2 of the same topology as the ANN and was obtained from a dataset in which the values of x_i were altered by the permutation procedure.

RBF

RBF network is characterized by having only one hidden layer and using a Gaussian activation function. The structure of the RBF that best predicted PT1 was established with 10 to 30 neurons (increased by 2 at each processing step) and a radius of 5 to 15 (increased by 0.5, at each processing step). The efficiency of prediction was measured based on R^2, and the relative importance of each trait was measured by de-structuring the information on each explanatory phenotypic trait, as described for MLP above.

Machine learning

Machine learning is one of the techniques used in artificial intelligence, and it allows for the detection of patterns in large datasets and the development of predictive models (MITCHELL, 1997). Learning algorithms based on decision trees are considered one of the most efficient and most used methods of supervised learning (MINGERS, 1989) to build predictive models of high precision, stability, and ease of interpretation. To predict PT1 and determine the importance of phenotypic traits through machine learning, decision trees with bagging, random forest, and boosting refinements were used. The quality of the predictive model fit was determined based on R^2, and the MSE was used to quantify the importance of phenotypic traits.

The importance of the explanatory trait was determined by estimating the percent increase in MSE (%IMSE). %IMSE was derived for each predictor variable from the difference in MSE between the predictive measure based on the original dataset and that based on a permuted dataset, where the predictor in question was randomized (NICODEMUS et al., 2010). To improve the predictive efficiency for the importance of traits, 5,000 trees were generated.

The relative importance of the variable (IV) was obtained by permuting its values and quantified using the following equation:

$$IV_{x_i} = MSE_{perm,x_i} - MSE, \quad (6)$$

where MSE_{perm,x_i} is the mean quadratic error of the methodology obtained from the use of a dataset in which the x_i values were changed by the permutation procedure. MSE is the mean square error of the same methodology obtained from the original predictor variables and is a function of the square of the deviations between the predicted and observed values.

6- Training and validation sets

The dataset was divided into two parts: a training and validation set. The training set included the same individuals for modeling using all methodologies and was composed of 80% individuals in each class, selected at random. The remaining 20% of individuals constituted the validation set. In previous studies, 60% to 90% of individuals constituted the training set (GIANOLA et al., 2011; GONZÁLEZ-CAMACHO et al., 2012). For the training and validation of the algorithms used, cross-validation (k-fold) was performed with k = 10 partitions (BENGIO &GRANDVALET, 2004).

Data simulation and analysis were performed using the R package NeuralNetTools (BECK, 2018) and Genes (CRUZ, 2016).

3. RESULTS AND DISCUSSION

1- Summary of key findings

The R^2 estimates for all methodologies, using the explanatory traits for PT1, are shown in Table 3. Based on these results, the methodologies used were compared and defined, proving the most efficient approach to PT1 prediction. Higher R^2 values indicate that the principal trait of prediction is more adaptable than the other explanatory phenotypic traits (ROY et al., 2008, 2015; SILVA JUNIOR et al., 2021).

Table 3. Maximum estimate of the coefficient of determination (R^2) for all methodologies using the explanatory traits for phenotypic trait 1 (PT1).

	CI		**ML**				**MR**
Scenarios	MLP (NN)	RBF	DT	FR	BA	BO	Stepwise
1	83.02 (30)	54.42	51.00	94.40	94.64	82.12	41.03
2	77.89 (29)	48.51	49.24	93.82	93.83	79.74	33.88
3	75.49 (29)	44.04	43.66	93.99	93.89	79.86	34.82
4	82.14 (25)	47.06	45.75	93.49	93.32	80.01	38.16

CI: Computational intelligence; ML: Machine learning; RM: Multiple regression; MLP: Multilayer perceptron; RBF: Radial basis function; DT: Decision tree; FR: Forest random; BA: Bagging; BO: Boosting. NN: number of neurons in the hidden layer.

Stepwise multiple regression provided the lowest estimate of R^2 (Table 3), indicating the existence of non-linear associations among the explanatory phenotypic traits not considered in the model; in the present study, this result can be attributed to the epistatic action between the control loci of each trait. In multiple linear regression, the absolute value of the *t-statistic* is commonly used as a measure of variable importance. Computational intelligence and machine learning were superior in extracting nonlinear information from model inputs (Table 3). However, in all scenarios, R^2 was lower for computational intelligence than for machine learning (Table 3).

The R^2 values were 83.03%, 77.89%, 75.49%, and 82.14% for scenarios 1, 2, 3, and 4, respectively, in the MLP network with only one neuron in the output layer and a single hidden layer (Table 3). The differences in results obtained with different methodologies indicate that the scenarios influenced the estimation of R^2 and, consequently, the prediction of PT1. Similar results have been reported in studies predicting corn and soybean yield based on climatic conditions using ANNs (R^2= 0.77 for corn, and 0.81 for soybean) and multiple linear regression (R^2= 0.42, maize and 0.46 for soybean, respectively) (KAUL et al., 2005); moreover, the development of linear regression models was time-consuming, and ANN models were superior in terms of accurately predicting corn and soybean yields under typical climatic conditions. SILVA et al. (2014) applied ANNs to simulated traits with 40% and 70% heritability for predicting genetic values and gains and found more efficient selection using ANNs than using maximum likelihood (genotypic mean). In addition, several studies have used this parameter (R^2) to verify the effectiveness of methodologies for the prediction or classification of

simulated populations (SILVA et al., 2014, 2016; SANT'ANNA et al., 2015, 2020; SILVA JUNIOR et al., 2021).

Twenty-five neurons were established for the MLP network, and an R^2 of 82.14% was obtained for scenario 4. For scenarios 2 and 3, 29 neurons were established, and R^2 values of 77.89% and 75.49% were obtained, respectively. In scenario 1, 30 neurons were established, and a maximum R^2 value of 83.02% was obtained (Table 3). High (>90%) R^2 estimates for all analyses were obtained using machine learning with random forest and bagging refinements (Table 3). With boosting, these estimates were >80%, indicating an efficient estimate of R^2. Decision trees were not superior to the remaining machine learning methods, and the R^2 values with decision trees were similar to those with RBF and stepwise multiple regression.

Thus, machine learning is indeed more efficient for the selection of phenotypic traits because it can deal with reduced or redundant information on phenotypic traits (QUINLAN, 1996). COSTA et al. (2020) assessed the importance of variables by bagging, random forest, boosting, decision tree, PML, and RBF and reported that PML and RBF achieved better results. After verifying the efficiency of different computational intelligence and machine learning methodologies in predicting PT1, we sought to identify the explanatory phenotypic traits that should be prioritized and established as auxiliary traits for indirect selection, as described below.

2- Importance of phenotypic traits by Pearson's correlation

Pearson's correlation coefficients were calculated between PT1 and other phenotypic traits in the four scenarios (Table 4).

Table 4. Pearson's correlation coefficient between the phenotypic trait 1 (PT1) and the other traits in the four scenarios.

PT	1	2	3	4
2	0.08^{ns}	0.10^{*}	-0.08^{ns}	-0.05^{ns}
3	0.07^{ns}	0.38^{**}	0.34^{**}	0.27^{**}
4	0.43^{**}	0.30^{**}	0.38^{**}	0.47^{**}
5	0.47^{**}	0.43^{**}	0.39^{**}	0.47^{**}
6	0.14^{**}	0.05^{ns}	0.05^{ns}	0.04^{ns}
7	0.25^{**}	0.19^{**}	0.26^{**}	0.14^{**}
8	0.12^{**}	0.24^{**}	0.15^{**}	0.23^{**}
9	0.03^{ns}	-0.13^{**}	0.06^{ns}	0.21^{**}
10	0.03^{ns}	-0.04^{ns}	0.03^{ns}	0.12^{**}
11	0.04^{ns}	0.03^{ns}	0.02^{ns}	0.08^{ns}

**, * significant at 1% and 5% probability of error, by t-test. ns: non-significant. PT: Phenotypic Traits. The numbers 1 and 2 are the scenarios that represent the first four and the last four controlling loci of quantitative traits, respectively. Numbers 3 and 4 correspond to scenarios with two first and two last, and the first two and last two loci controlling, respectively.

PT1 showed a positive and significant correlation with PT5 ($P < 0.01$) in all scenarios, probably because of the large number of shared control loci and high h^2 (80%) of PT5. Meanwhile, the correlation between PT1 and PT2 was not significant ($P > 0.05$) in all scenarios, except in scenario 2, which is consistent with the simulation results and the absence of shared control loci between these two traits. These results were expected because PT2 had no genetic relationship with PT1 due to the absence of pleiotropy or a factorial link between the loci controlling these phenotypic traits.

The phenotypic correlation network is shown in Figure 1. Consistent with the results in Table 4, these correlations were considered stronger relative to those among the groups. Group 2, represented by PT2, was positioned far from the principal trait (PT1) because of the lack of shared loci or genetic relationships. Group 3, represented by PT3, PT4, and PT5, was closer to

PT1 and represented the most important traits for predicting PT1, particularly PT5, which had the highest h^2 value. The placement of groups 4 and 5 relative to PT1 was consistent with the simulation of genetic structures, allowing satisfactory prediction of importance, albeit lower than expected relative to group 3 but higher than expected relative to PT2.

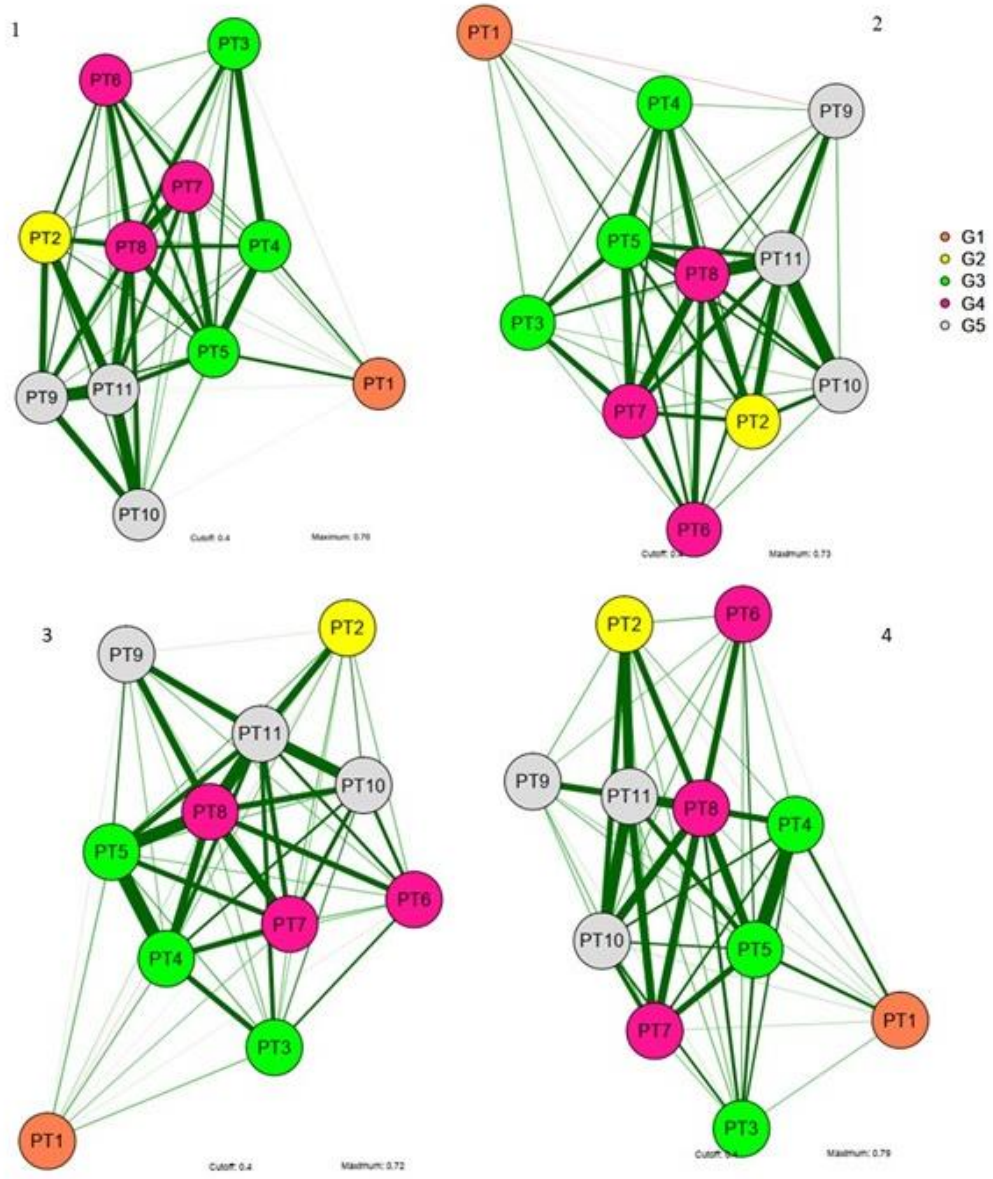


Figure 1. Phenotypic correlation network for the five distinct groups in four different scenarios represented by indicators 1, 2, 3 and, 4. Line width is proportional to the strength of the correlation. Number 1 is the scenario that represents the first four controlling loci; 2 is the scenario that represents the last four loci; 3 is the scenario that represents the first two and the last two loci, and 4 is the scenario that represents the first two and the last two loci. G1: Orange: principal trait of prediction; G2: Yellow: control trait of prediction; G3: Green: phenotypic traits

3, 4 and 5 (PT3, PT4, and PT5); G4: Pink: phenotypic traits 6, 7 and 8 (PT6, PT7, and PT8); G5: Gray: phenotypic traits 9, 10 and 11 (PT9, PT10, and PT11), differentiate as to the number of controlling logic with PT1 and are distinguished by the differential action of the environment.

In breeding programs, understanding the meaning and degree of association between traits plays an important role in the development of selection strategies that facilitate the production of superior genotypes. One of the most used techniques to estimate these associations is the Pearson correlation, which is interpreted as the strength of the linear association between a pair of characteristics (PEARSON, 1920). When more than two characteristics are considered, this measure alone does not show the real meaning and magnitude of the interrelations, making it impossible to determine whether the associations are cause or effect (ALIYU et al., 2000). Therefore, a path analysis is used when there are dependent (interest) characteristics and other characteristics whose interrelationships are of interest to the researcher (OLIVOTO et al., 2017). This technique has been shown to be very useful for revealing cause-and-effect associations and providing support in indirect selection (NARDINO et al., 2016).

The estimates of the correlation coefficients can help improve our understanding of a complex character, such as production, but they do not determine the relative importance of the direct and indirect influences of the other characters on production (OLIVOTO et al., 2017). This is because the correlation between two characteristics measures the association between both, but it does not determine the cause-and-effect relationship between them, which can be determined through the trail analysis (CRUZ & CARNEIRO, 2014). When the correlation between explanatory characteristics increases, the difficulty in assessing their relative importance in predicting the dependent characteristic is greater (BLALOCK, 1963; HOERL & KENNARD, 1981). Machine learning and computational intelligence are approaches that, even in the presence of an association between explanatory characteristics and a high degree of multicollinearity, are able to make appropriate predictions and classifications.

3- Indirect selection through principal traits

Table 5 shows percent indirect selection gains between PT1 and the other traits in the four scenarios. The highest percent indirect gain was achieved with PT5 (h^2 = 80%) in all scenarios. In scenario 2, the highest indirect selection gain was achieved with PT3 and PT4

(3.94 and 2.12, corresponding to h^2= 30% and 50%, respectively). Meanwhile, in scenarios 3 and 4, the highest selection gain was achieved with PT4 (h^2= 50). Therefore, a greater number of shared control loci affected indirect selection gain. The success of indirect selection depends mainly on heritability and genetic correlation between the primary and secondary traits (FALCONE, 1960); therefore, traits closely correlated traits with a greater heritability than the target trait have great potential for indirect selection.

Table 5: Percentage gain indirect selection between the phenotypic trait 1 (PT1) and the other traits in the four scenarios.

PT	1	2	3	4
2	0.06	1.32	**-0.06**	**-1.07**
3	0.05	**3.94+**	0.25	1.97
4	**0.31+**	**2.12+**	0.28+	3.46+
5	**0.34+**	**2.73+**	**0.29+**	**3.12+**
6	0.1	-0.32	0.04	-0.27
7	0.18	1.05	0.19	1.18
8	0.08	1.63	0.11	1.26
9	0.02	-1.4	0.04	2.51
10	0.02	-0.42	0.02	0.76
11	0.03	0.06	0.02	0.32

PT: Phenotypic trait; +: minor importance in the prediction of PT1. Number 1 is the scenario that represents the first four controlling loci; 2 is the scenario that represents the last four loci; 3 is the scenario that represents the first two and the last two loci, and 4 is the scenario that represents the first two and the last two loci.

4- Importance of phenotypic traits by computational intelligence

MLP

The results of quantification of the importance of phenotypic traits using MLP after the permutation of traits and assignment of a zero value to the input trait are shown in Table 6. If R^2 shows a great reduction after disturbing the values of a trait, the phenotypic trait is important relative to others for prediction. The relative importance of phenotypic traits based on the reduction in R^2 (Table 6), independent of the heritability of each phenotypic trait and LGs, differed across scenarios. Permutation was efficient in quantifying the importance of PT5, which was the most important trait in all scenarios except scenario 4, in which PT4 was the most important trait (Table 6). The R^2 values were 83.02%, 77.89%, 75.49%, and 82.14% in

scenarios 1, 2, 3, and 4, respectively (Table 3). Furthermore, permutation was efficient in demonstrating PT2 as the least important trait in scenarios 3 and 4, which was expected based on the number of markers influencing the principal trait (PT1).

Table 6. Estimation of the coefficient of determination (R^2) for the prediction of the phenotypic trait 1 (PT1) using the Multilayer Perceptron (MLP).

	Zero				**Permutation**			
PT	1	2	3	4	1	2	3	4
2	34.13	16.28	17.71	6.69	45.54	36.33	43.89*	44.14*
3	11.33	0.81	11.23	1.47	36.14	29.23	38.78	35.15
4	10.32	3.62	2.72	0.13	19.45	32.84	41.32	12.03+
5	11.93	1.22+	3.30	0.76	17.83+	18.72+	19.09+	16.29
6	14.52	17.54*	14.86	0.03+	50.35*	31.42	33.34	21.11
7	8.34	3.80	23.53*	1.87	42.83	37.08	37.47	24.82
8	0.05+	10.99	1.41+	0.19	22.30	39.62	32.28	29.26
9	22.52	24	13.1	15.44*	31.26	56.68*	38.01	41.98
10	36.77*	1.83	7.11	14.21	46.71	30.00	30.96	34.12
11	24.94	7.99	7.24	0.06	34.48	32.46	38.42	21.34

PT: Phenotypic trait; +, *: Auxiliary traits of major and minor importance in the prediction of PT1, respectively. Number 1 is the scenario that represents the first four controlling loci; 2 is the scenario that represents the last four loci; 3 is the scenario that represents the first two and the last two loci, and 4 is the scenario that represents the first two and the last two loci.

Given the great complexity of interpreting the MLP network to quantify the relative importance of traits, an alternative is to use Garson's algorithm (1991) modified by Goh (1995). This algorithm partitions the ANN connection weights to determine the relative importance of each input trait. The weights associated with the neurons in an ANN are partially analogous to the coefficients in a generalized linear model (OLDEN et al., 2004). The combined effects of the weights represent the relative importance of the predictors for predicting the response variable. The weights correspond to the relative influence of the information that is processed in the network such that the input variables that are not relevant are suppressed by the weights.

The method proposed by Garson (1991), modified by Goh (1995), identifies the relative importance of explanatory variables toward specific response variables in a supervised neural network. The relative importance (or strength of association) of a specific explanatory variable

to a specific response variable can be determined by identifying all the weighted connections between the nodes of interest. This algorithm has the ability to deal with degrees of association and multicollinearity between explanatory characteristics and has been shown to be efficient in quantifying the importance of phenotypic characteristics in studies with known genetic structures.

The percent relative contributions of the 10 phenotypic traits relative to PT1 in the four scenarios estimated by this method are listed in Table 7. In all scenarios, the relative contributions of PT5 and PT2 in predicting PT1 were quantified as major and minor, respectively. The result for PT2 was expected because of the lack of shared LGs controlling PT1. Thus, PT5 presented the highest number of major pleiotropic markers, in addition to the minor markers, but with major heritability.

Table 7. Percentage of relative contribution by Garson´s algorithm (1991) modified by Goh (1995) method for 10 phenotypic traits in relation to PT1, in four scenarios.

PT	**Scenarios**			
	1	2	3	4
2	**6.12***	5.24*	8.28*	8.77*
3	9.26	8.65	10.63	9.27
4	7.89	11.13	9.47	10.26
5	**13.11+**	**12.04+**	**11.41+**	**12.73+**
6	9.47	10.16	9.00	9.11
7	9.87	10.49	8.98	9.79
8	11.14	11.96	10.64	10.16
9	10.81	9.85	11.03	10.35
10	10.77	9.96	10.43	11.33
11	11.55	10.53	**11.12**	11.21

PT: Phenotypic trait; +, *: Auxiliary traits of major and minor importance in the prediction of PT1, respectively. Number 1 is the scenario that represents the first four controlling loci; 2 is the scenario that represents the last four loci; 3 is the scenario that represents the first two and the last two loci, and 4 is the scenario that represents the first two and the last two loci.

ANNs most often exhibit satisfactory performance compared with other machine learning-based predictive algorithms (SANTOS et al., 2018). The MLP network has been widely used in predictive processes (GEDEON et al., 1995; SANTOS et al., 2018), and its

success has already been demonstrated in several studies, which mathematically showed that networks even with only a single hidden layer with different numbers of neurons work very well (DE OÑA et al., 2014; SANTOS et al., 2018).

RBF

The results of quantification of the importance of phenotypic traits using RBF after the permutation of the traits and assignment of a zero value to the input trait are shown in Table 8.

Table 8. Estimation of the coefficient of determination (R^2) for the prediction of the phenotypic trait 1 (PT1) using the radial basis function (RBF).

	Zero				**Permutation**			
PT	1	2	3	4	1	2	3	4
2	37.15*	9.38	27.92	37.99*	47.96	32.27	33.27	37.82
3	12.63	23.66	15.44	25.94	27.35	37.66	24.24	37.01
4	19.39	23.39	21.65	18.91	24.96	24.40	36.66	32.15
5	8.40	8.60+	9.73+	19.91	9.31+	16.19+	14.46+	16.86+
6	32.37	20.21	15.79	30.83	44.45	29.01	33.78	38.95
7	28.17	19.46	33.69	19.94	44.39	40.89*	31.02	33.63
8	29.26	13.03	11.02	9.79+	43.74	38.73	33.11	39.69
9	36.10	36.20*	22.62	28.53	33.47	36.65	39.25	28.79
10	39.41	26.63	32.27	19.12	50.79*	39.77	39.40*	40.38*
11	2.01+	30.15	37.32*	25.23	29.40	40.14	34.84	30.49

PT: Phenotypic trait; +, *: Auxiliary traits of major and minor importance in the prediction of PT1, respectively. Number 1 is the scenario that represents the first four controlling loci; 2 is the scenario that represents the last four loci; 3 is the scenario that represents the first two and the last two loci, and 4 is the scenario that represents the first two and the last two loci.

The relative importance of phenotypic traits based on the reduction in R^2 (Table 8), independent of the heritability of each phenotypic trait and LGs, differed across scenarios. The most important traits obtained with the permutation of traits and assignment of a zero value to the input trait were not consistent. As mentioned earlier, permutation was efficient in quantifying the relative contribution of PT5 as a major based on the reduction in R^2 estimate when the information was disturbed (Table 7). This efficiency could be extended to PT3, PT4, and PT5, which shared the same number of markers as PT1 and differed only in terms of heritability. Using this strategy, PT2 was identified as the least important trait.

Radial basis function networks have the ability to learn from the data used in their training and provide a unique solution. They are also comparatively faster than perception-type ANNs (GONZALES-CAMACHO et al., 2012). In addition, they have a good ability to handle interactions compared to semi-parametric and linear regressions (SANT'ANNA et al., 2020). SANT'ANNA et al. (2020) applied RBF in studies using simulated characteristics with 30% and 60% heredity for variable selection. The authors identified greater efficiency in the selection using RBF when the scenario involved epistatic interactions in the gene control of the studied characters. GONZÁLEZ-CAMACHO et al. (2012) observed that it is possible to improve prediction in non-parametric models when the selection includes markers that are not directly related to the characteristics of interest.

The results obtained corroborate the expectation about the RBF in quantifying and revealing the importance of the characteristics using the strategy of causing disturbances based on the permutations or fixation of the phenotypic values of the input variables.

Our study demonstrates the ability of RNA to quantify the importance of phenotypic characteristics with known genetic structures. Techniques showing the impact of the disruption or disturbance to the information of a given entry on the estimation of the determination coefficient and partitioning of the ANN connection weights have been presented. These techniques were effective in quantifying the true importance of phenotypic characteristics.

5- Importance of phenotypic traits by machine learning

The importance of phenotypic traits using machine learning and its refinements (bagging, random forest, and boosting) in four different scenarios are shown in Table 9. The %IMSE values were calculated, with the highest value representing the most important phenotypic trait.

Table 9. The average estimate of the relative contribution of explanatory phenotypic traits to predict PT1 using a machine learning method in four scenarios.

	Bagging				Random Forest				Boosting			
PT	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
2	8.35*	10.63	16.5	13.14	7.32	13.79	15.71	15.94	1.08*	3.2	11.18	11.42
3	20.5	25.36	18.23	12.7	22.51	25.27	18.39	13.81	11.92	19.22	16.58	9.88
4	29.58	14.72	19.22	23.92	28.7	14.69	20.33	24.85	26.69	6.73	13.04	27.57
5	**34.13+**	**40.45+**	**24.8+**	**29.36+**	**33.9+**	**34.78+**	**25.56+**	**26.98+**	**33.82+**	**31.31+**	**23.28+**	**24.67+**
6	10.67	16.48	7.18	9.17	9.98	16.4	7.82*	10.05	3.23	8.89	5.85	5.94
7	12.22	9.45*	12.42	7.75	13.65	8.51*	13.46	8.38	3.89	2.39*	9.35	1.69
8	12.28	11.99	9.34	5.41	12.96	11.89	9.31	6.03	3.64	3.86	3.88	1.68*
9	13	14.24	7.55*	11.57	13.63	14.99	8.59	14.11	4.71	7.7	1.82*	6.82
10	4.03	15.59	11.64	4.00*	3.82*	15.92	13.05	5.35*	3.2	10.68	4.96	2.62
11	15.43	11.61	18.6	14.48	15.9	12.62	15.03	16.39	7.82	6.04	10.07	7.7

PT: Phenotypic trait; +, *: Auxiliary traits of major and minor importance in the prediction of PT1, respectively. Number 1 is the scenario that represents the first four controlling loci; 2 is the scenario that represents the last four loci; 3 is the scenario that represents the first two and the last two loci, and 4 is the scenario that represents the first two and the last two loci.

PT5 was estimated as the most important phenotypic trait in all machine learning methodologies and in all scenarios. This result is consistent with that of the computational intelligence methods (Tables 6-8). Although machine learning is an efficient tool for quantifying the relative importance of traits, it does not make any assumptions regarding the distribution of explanatory variables and is robust in terms of quantity, redundancy, and environmental influence (TAN et al., 2014; BEUCHER et al., 2019). In addition, random forest and boosting do not require an inheritance specification model and can account for non-additive effects without increasing the number of covariates in the model or computation time (GONZÁLEZ-RECIO et al., 2011). Random forest and bagging show good predictive performance in practice; they work well for multi-dimensional problems and can be used with multi-class output, categorical predictors, and imbalanced problems (GREGORUTTI et al., 2017). Satisfactory results of variable selection using the random forest algorithm in the presence of correlated predictors have been reported for GREGORUTTI et al. (2017).

The discriminatory power, redundancy, precision, and complexity can influence the indices or statistics used to quantify the importance of auxiliary traits in the prediction of a principal trait. Thus, the selection of a prediction method and index reflecting the true importance of each auxiliary trait is imperative. In the present study, we propose some genetic

structures that can allow us to estimate the efficiency of procedures using computational intelligence and machine learning to quantify the relative contribution of phenotypic traits in different scenarios for the implementation of these techniques in other studies. The lack of information on the implementation of computational intelligence and machine learning in breeding programs to predict phenotypic traits is a challenge for breeders. However, with the relevant recent advances in biotechnology and high-throughput phenotyping, more information can be obtained for the identification of genetically superior individuals.

Breeding for desired traits in crops has long been a time-consuming, labor-intensive, and expensive process. Breeders study generations of plants and identify and modify desired genetic traits, as they assess how traits are expressed in offspring (PALIWAL & KUMAR, 2011; FERREIRA et al., 2015). The application of computational intelligence and machine learning to identify optimal suites of observable characteristics (phenotypes) can enable informed decisions and achieve outcomes of great relevance in breeding programs. In addition, these methodologies can help predict genetic traits with the best performance under different agricultural management practices.

Methodologies based on machine learning and computational intelligence do not depend on stochastic information and tend to be more efficient, whereas conventional methodologies depend on the normality of the distribution of phenotypic traits. Moreover, machine learning and computational intelligence methodologies make no assumptions regarding the model and can capture complex factors, such as epistasis and dominance, in prediction models. In machine learning, *a priori* knowledge of prediction is not required if the data produce these effects, and no assumptions are made regarding the distribution of phenotypic values (SOUSA et al., 2020). Machine learning algorithms have the advantage of modeling data in a non-linear and non-parametric manner (OSCO et al., 2020). Unlike many traditional statistical methods, these algorithms are built with the advantage of dealing with noisy, complex, and heterogeneous data (OSCO et al., 2019; SHAH et al., 2019; FU et al., 2019). BARBOSA et al. (2021) reported that machine learning methods are powerful tools for predicting genetic values with epistatic genetic control in traits with different degrees of heritability and different numbers of controlling genes. The results obtained in the present study can be used to select genotypes and test them in the field. Thus, the proposed model can be validated in practice.

4. CONCLUSION

Computational intelligence and machine learning can efficiently predict the relative contributions of auxiliary traits in different scenarios in plant breeding programs. PT5 was identified as the most important predictor of PT1.

5. REFERENCES

Aliyu L, MK Ahmed, and MD Magaji. Correlation and multiple regression analysis between morphological characters and components of yield in pepper (*Capsicum annuum* L.). Crop Res. 2000; 19:318–323.

Barbosa IP, Silva MJ, Costa WG, Sant'Anna IC, Nascimento M, Cruz CD. Genome-enabled prediction through machine learning methods considering different levels of trait complexity. Crop Science. 2021. https://doi.org/10.1002/csc2.20488.

Basheer IA, Hajmeer M. Artificial neural networks: Fundamentals, computing, design, and application. J Microbiol Methods 2000; 43:3-31.

Beck M. NeuralNetTools: Visualization and Analysis Tools for Neural Networks. 2018; R package version 1.5.2. http://dx.doi.org/10.18637/jss.v085.i11.

Bengio Y, Grandvalet Y. No unbiased estimator of the variance of k-fold cross-validation. J. Mach. Learn Res. 2004; 5:1089-1105.

Beucher A, Møller AB, Greve MH. Artificial neural networks and decision tree classification for predicting soil drainage classes in Denmark. Geoderma 2019; 352:351-359. http://dx.doi.org/10.1016/j.geoderma.2017.11.004.

Blalock HM. Correlated independent variables: The problem of multicollinearity. 1963; Soc. Forces 42: 233– 237. http://dx.doi.org/10.1093/sf/42.2.233.

Chen S, Cowan CF and Grant PM. Orthogonal least squares learning algorithm for radial basis function networks. IEEE Trans Neural Netw Learn Syst. 1991; 2(2):302-309.

Costa WGD, Barbosa IP, de Souza JE, Cruz CD, Nascimento M, de Oliveira ACB. Machine learning and statistics to qualify environments through multi-traits in Coffea arabica. PLoS One. 2021 Jan 12;16(1):e0245298. http://dx.doi.org/10.1371/journal.pone.0245298.

Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos G, Burgueño J, Camacho-González JM, Pérez-Elizalde S, Beyene Y, Dreisigacker S. Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. 2017;22(11):961-975.

Cruz CD and Carneiro PCS. Modelos biométricos aplicados ao melhoramento genético. V.2. Viçosa: Editora UFV, 2014. p. 668.

Cruz CD, Nascimento M. Inteligência Computacional aplicada ao melhoramento genético. 1st ed. Vicosa: Editora UFV; 2018.

Cruz, CD. Genes Software – extended and integrated with the R, Matlab and Selegen. Acta Scientiarum 2016; 38:547-552. http://dx.doi.org/10.4025/actasciagron.v38i4.32629.

De Oña J, Garrido C. Extracting the contribution of independent variables in neural network models: a new approach to handle instability. Neural Comput Appl. 2014; 25:859-869.

Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics data sets. Brief Bioinform 2019; 20:492-503. http://dx.doi.org/10.1093/bib/bbx124.

Falconer DS. Introduction to quantitative genetics. Oliver e Boyd: Edimburgo, Reino Unido; Londres, Reino Unido, 1960.

Ferreira MG, Azevedo AM, Siman LI, Silva GH, Carneiro CS, Alves FM, Delazari FT, Silva DJH, Nick C. Automation in accession classification of *Brazilian Capsicum* germplasm through artificial neural networks. Scientia Agricola 2017; 74(4). http://dx.doi.org/10.1590/1678-992X-2015-0451.

Fu P, Meacham-Hensold K, Guan K, Bernacchi CJ. Hyperspectral leaf reflectance as proxy for photosynthetic capacities: An ensemble approach based on multiple machine learning algorithms. Front Plant Sci. 2019; 10.

Garson GD. Interpreting neural network connection weights. Artificial Intelligence Expert. 1991; 6:46-51.

Gedeon TD, Wong PM, Harris D. Balancing bias and variance: network topology and pattern set reduction techniques. Berlin, Heidelberg: Springer Berlin Heidelberg. 1995.

Gianola D, Okut H, Weigel KA, Rosa GJM. Predicting complex quantitative traits with neural

networks: a case study with Jersey cows and wheat. BMC Genetics. 2011; 12:87.

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

Gianola, D and De Los Campos, G. Inferring genetic values for quantitative traits non-parametrically. Genetics Research, 2008; 90(6), pp.525-540.

Glória LS, Cruz CD, Vieira RAM, Resende MDV, Lopes PS, Siqueira OHD and Silva FF. Accessing marker effects and heritability estimates from genome prediction by Bayesian regularized neural networks. Livestock Science. 2016; 191, pp.91-96.

Goh ATC. Back-propagation neural networks for modeling complex systems. Artificial Intelligence in Engineering. 1995; 9:143-51. http://dx.doi.org/10.1016/0954-1810(94)00011-S.

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

González-Camacho JM, Campos G, Pérez P, Gianola D, Cairns JE, Mahuku G, Babu R and Crossa J. Genome-enabled prediction of genetic values using radial basis function neural networks. Theoretical and Applied Genetics. 2012; 125(4): 759-771.

González-Camacho JM, Crossa J, Pérez-Rodríguez P, Ornella L and Gianola D. Genome-enabled prediction using probabilistic neural network classifiers. BMC genomics. 2016; 17(1), p.208.

González-Recio O, Rosa GJ and Gianola D. Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. Livestock Science. 2014; 166:217-231.

González-Recio O, Forni S. Prediction across the genome of discrete traits using Bayesian regressions and machine learning. Genet Sel Evol. 2011; 43:7. https://doi.org/10.1186/1297-9686-43-7.

Gregorutti B, Michel B, Saint-Pierre P. Correlation and variable importance in random forests. Stat Comput 2017; 27:659-678. https://doi.org/10.1007/s11222-016-9646-1.

Hassanzadeh Z, Ghavami R, Kompany-Zareh M. Radial basis function neural networks based on the projection pursuit and principal component analysis approaches: QSAR analysis of fullerene[C60]-based HIV-1 PR inhibitors. Medicinal Chemistry Research. 2015; 25: 19-29. http://dx.doi.org/10.1007/s00044-015-1466-x.

Hastie T, Tibshirani R, Friedman J. The Elements of Statiscal Learning Data Mining, Inference, and Prediction, 2nd ed. New York, NY: Springer. 2009; p. 745.

Hoerl AE and Kennard RW. Ridge regression—1980: Advances, algorithms, and applications. Am. J. Math. Manage. Sci. 1981; 1: 5– 83. http://dx.doi.org/10.1080/01966324.1981.10737061.

Howard R, Carriquiry AL, Beavis WD. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. Genetics. 2014; 4:6.

Jardin PD. Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy. Neurocomputing 73. 2010; 2047–2060.

Kaul M, Hill RL, Walthall C. Artificial neural networks for corn and soybean yield prediction. Agric Syst 2005; 85:1-18.

Kecman V. Learning and Soft Computing. Cambridge, London, England, MA: MIT Press. 2001.

Khaki S, Khalilzadeh Z, Wang L. Predicting yield performance of parents in plant breeding: A neural collaborative filtering approach. PLoS ONE 2020; 15(5): e0233382. https://doi.org/ 10.1371/journal.pone.0233382.

Long N, Gianola D, Rosa GJ and Weigel KA. Marker-assisted prediction of non-additive genetic values. Genetica. 2011; 139(7): 843-854.

Long N, Gianola D, Rosa GJ, Weigel KA, Kranis A and Gonzalez-Recio O. Radial basis function regression methods for predicting quantitative traits using SNP markers. Genetics research 2010; 92(3), pp.209-225.

Mingers J. An empirical comparison of pruning methods for decision tree induction. Machine Learning 1989; 4:227-243.

Mitchell TM. Machine Learning. WCB – McGraw-Hill, Boston, MA. 1997.

Mukaka MM. Statistics Corner: A guide to appropriate use of correlation coefficient in medical research. Malawi Med J. 2012; 24:69-71.

Nicodemus KK, Malley JD, Strobl C, Ziegler A. The behaviour of random forest permutation-based variable importance measures under predictor correlation. BMC Bioinformatics. 2010; 11:110. http://dx.doi.org/10.1186/1471-2105-11-110.

Olden JD, Jackson DA. "Illuminating the "Black Box": A randomization approach for understanding variable contributions in artifical neural networks." Ecological Modelling. 2002; 154:135–150. http://dx.doi.org/10.1016/s0304-3800(02)00064-9.

Olivoto T, Souza VQ, Nardino M, Carvalho IR, Ferrari M, Pelegrin AJ, Szareski VJ, Schmidt D. Multicollinearity in path analysis: a simple method to reduce its effects. Agronomy Journal 2017; 109: 131-142.

Osco LP, Ramos APM, Moriya EAS, Bavaresco LG, Lima BC, Estrabis N, Pereira DR, Creste JE, Marcato Junior J, Gonçalves WN et al. Modeling hyperspectral response of water-stress induced lettuce plants using artificial neural networks. Remote Sens. 2019; 11:2797.

Osco LP, Ramos APM, Pinheiro MMF, Moriya EAS, Imai NN, Estrabis N, Lanczyk F, Araujo FF, Liesenberg V, Jorge LAC, Li J, Ma L, Gonçalves WN, Junior JM and Creste JE. A machine learning framework to predict nutrient content in valencia-orange leaf hyperspectral measurement. Remote Sens. 2020; 12:906. http://dx.doi.org/10.3390/rs12060906.

Paliwal M, Kumar UA. Assessing the contribution of variables in feed forward neural network. Applied Soft Computing. 2011; 11:3690-3696.

Parmley KA, Higgins RH, Ganapathysubramanian B et al. 2019. Machine learning approach for prescriptive plant breeding. Sci Rep 9, 17132. http://dx.doi.org/10.1038/s41598-019-53451-4.

Paruelo JM, Tomasel F. "Prediction of functional characteristics of ecosystems: a comparison of artificial neural networks and regression models." Ecological Modelling. 1997; 98: 173–186. http://dx.doi.org/10.1016/s0304-3800(96)01913-8.

Paswan RP and Begum SA. Regression and neural networks models for prediction of crop production. Int. J. Sci. Eng. Res.2013; 4:11.

Pearson K. Notes on the history of correlation. Biometrika 1920; 13: 25– 45. . http://dx.doi.org/ 10.1093/biomet/13.1.25.

Pérez-Rodríguez P, Gianola D, González-Camacho JM, Crossa J, Manès Y and Dreisigacker, S. Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3: Genes, Genomes, Genetics, 2012; 2(12): 1595-1605.

Quinlan JR. Learning decision tree classifiers ACM Comput. Surv. 1996; 28:71-72.

Sant'Anna IC, Ferreira RADC, Nascimento M, Carneiro VQ, Silva GN, Cruz CD, Oliveira MS, Chagas FEO. Multigenerational prediction of genetic values using genome-enabled prediction. PLoS ONE. 2019. 14, e0210531. http://dx.doi.org/10.1371/journal.pone.0210531.

Sant'Anna IC, Tomaz RS, Silva GN, Nascimento M, Bhering LL, Cruz CD. Superiority of artificial neural networks for a genetic classification procedure. Genet. Mol. Res. 2015, 14, 9898–9906.

Sant'Anna IC, Silva GN, Nascimento M, Cruz C D. Subset selection of markers for the genome-enabled prediction of genetic values using radial basis function neural networks. Acta Scientiarum-Agronomy2020; 43: e46307. https://doi.org/10.4025/actasciagron.v43i1.46307

Santos RP, Dean DL, Weaver JM and Hovanski Y. Identifying the relative importance of predictive variables in artificial neural networks based on data produced through a discrete event simulation of a manufacturing environment. Journal International Journal of Modelling and Simulation.2018; 39:234-245. http://dx.doi.org/ 10.1080/02286203.2018.1558736.

Shah SH, Angel Y, Houborg R, Ali S, McCabe MF. A random forest machine learning approach for the retrieval of leaf chlorophyll content in wheat. Remote Sens. 2019, 11, 920.

Silva GN, Nascimento M, Sant'Anna IC, Cruz CD, Caixeta ET, Carneiro PCS, Rosado RDS, Pestana KN, Almeida DP, Oliveira MS. Artificial neural networks compared with Bayesian generalized linear regression for leaf rust resistance prediction in Arabica

coffee. Pesquisa Agropecuaria Brasileira. 2017; 52:186-193. http://dx.doi.org/10.1590/s0100-204x2017000300009.

Silva GN, Tomaz RS, Sant'Anna IC, Carneiro VQ, Cruz CD, Nascimento M. Evaluation of the efficiency of artificial neural networks for genetic value prediction. Genet. Mol. Res. 2016; 15, 1–11.

Silva GN, Tomaz RS, Sant'anna IC, Nascimento M, Bhering LL and Cruz CD. Neural networks for predicting breeding values and genetic gains. Scientia Agricola. 2014; 71, 494-498. http://dx.doi.org/10.1590/0103-9016-2014-0057.

Silva Junior AC, Sant'Anna IC, Silva GN, Cruz CD, Nascimento M, Lopes LB, Soares PC. Computational intelligence and machine learning to study the importance of characteristics in flood-irrigated rice. 2021. Acta Scientiarum-Agronomy (*in prelo*).

Skawsang S, Nagai M, Nitin K and Soni P. Predicting rice pest population occurrence with satellite-derived crop phenology, ground meteorological observation, and machine learning: A case study for the Central Plain of Thailand. Appl. Sci. 2019; 9:4846. http://dx.doi.org/10.3390/app9224846.

Sousa IC, Nascimento M, Silva GN, Nascimento ACC, Cruz CD, Fonseca F, Almeida DP, Pestana KN, Azevedo CF, Zambolim L and Caixeita ET. Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. Scientia Agricola 2020; 78: 1–8. https://doi.org/10.1590/1678-992x-2020-0021.

Sreekanth S, Ramaswamy HS, Sablani SS, Prasher SO. A neural network approach for evaluation of surface heat transfer coefficient. J. Food Proc. Preserv. 2010; 23: 329-348.

Stefaniak B, Cholewiński W, Tarkowska A. Algorithms of Artificial Neural Networks - Practical application in medical science. Polski Merkuriusz Lekarski. 2005; 19:819-22.

Strobl C, Boulesteix AL, Kneib T, Augustin T and Zeileis A. "Conditional Variable Importance for Random Forests," BMC Bioinformatics 2008; 9(307). doi:10.1186/1471-2105-9-307 http://www.biomedcentral.com/1471-2105/9/307.

Tan K, Li E, Du Q, Du P. An efficient semi-supervised classification approach for hyperspectral imagery. ISPRS Journal of Photogrammetry and Remote Sensing. 2014; 97:36–45. http://dx.doi.org/10.1016/j.isprsjprs.2014.08.003.

Tolo¸si L, Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics 27, 2011; 1986–1994.

Ventura RV, Silva MA, Medeiros TH, Dionello NL, Madalena FE, Fridrich AB, Valente BD, Santos GG, Freitas LS, Wenceslau RR, Felipe VPS & Corrêa GSS. Use of artificial neural networks in breeding values prediction for weight at 205 days in Tabapuã beef cattle. Arquivo Brasileiro de Medicina Veterinária e Zootecnia. 2012; 64:411-418. http://dx.doi.org/10.1590/S0102-09352012000200022.

Yu H, Campbell MT, Zhang Q, Walia H, and Morota G. Genomic Bayesian confirmatory factor analysis and Bayesian network to characterize a wide spectrum of rice phenotypes. G3: Genes, Genomes, Genetics. 2019; 9:1975-1986. http://dx.doi.org/10.1101/435792.

CAPÍTULO 2

Inteligência computacional e aprendizado de máquina para estudar a importância das características do arroz irrigado por inundação

(in prelo: Acta Scientiarum-Agronomy*)*

VIÇOSA - MINAS GERAIS
2021

RESUMO

SILVA JÚNIOR, Antônio Carlos, D.Sc., Universidade Federal de Viçosa, setembro de 2021. **Inteligência computacional e aprendizado de máquina para estudar a importância das características do arroz irrigado por inundação**. Orientador: Cosme Damião Cruz.

O estudo da importância das características permite ao melhorista orientar estratégias para selecionar e acelerar o progresso do melhoramento genético. Embora a avaliação simultânea de características no programa de melhoramento de plantas forneça uma grande quantidade de informações, identificar qual característica fenotípica é a mais importante é um desafio para o melhorista. Assim, o objetivo deste trabalho foi estimar a melhor abordagem para predição e estabelecer uma rede de melhor poder preditivo em arroz irrigado por inundação via metodologias baseadas em regressão, inteligência artificial e aprendizado de máquinas. Foram utilizados a regressão múltipla, inteligência computacional e aprendizado de máquina para prever a importância das características. A inteligência computacional e o aprendizado de máquina se destacaram por sua capacidade de extrair informações não lineares das entradas do modelo. A contribuição relativa de caracteres auxiliares em arroz por meio de inteligência computacional e aprendizado de máquina mostrou-se eficiente para determinar a importância relativa de variáveis em arroz irrigado por inundação. Os caracteres indicados para auxiliar na tomada de decisão são floração, número de grãos cheios por panículas e comprimento de panículas para este estudo. A rede com apenas uma camada oculta com 15 neurônios foi eficiente para determinar a importância relativa de variáveis em arroz irrigado por inundação.

Palavras-chave: *Oryza sativa* L; regressão múltipla; Inteligência computacional; aprendizado de máquina.

ABSTRACT

SILVA JÚNIOR, Antônio Carlos, D.Sc., Universidade Federal de Viçosa, September 2021. **Computational intelligence and machine learning to study the importance of characteristics in flood-irrigated rice**. Advisor: Cosme Damião Cruz.

The study of traits in crops enables breeders to guide strategies for selecting and accelerating the progress of genetic breeding. Although the simultaneous evaluation of characteristics in the plant breeding programme provides large quantities of information, identifying which phenotypic characteristic is the most important is a challenge facing breeders. Thus, this work aims to quantify the best approaches for prediction and establish a network of better predictive power in flood-irrigated rice via methodologies based on regression, artificial intelligence, and machine learning. Multiple regression, computational intelligence, and machine learning were used to predict the importance of the characteristics. Computational intelligence and machine learning were notable for their ability to extract nonlinear information from model inputs. The relative contribution of auxiliary characteristics in rice through computational intelligence and machine learning proved to be efficient in determining the relative importance of variables in flood-irrigated rice. The characteristics indicated to assist in decision making are flowering, number of grains filled by panicles and length of panicles for this study. The network with only one hidden layer with 15 neurons was observed to be efficient in determining the relative importance of variables in flooded rice.

Keywords: *Oryza sativa* L; multiple regression; computational intelligence; machine learning.

1. INTRODUCTION

Rice (*Oryza sativa* L.) is one of the most important crops in the world and is considered one of the main annual crops in Brazil (Streck et al., 2018). It is estimated that, by 2050, global rice production will need to increase by 60 to 110% to supply the population's demand (Ray et al., 2013; Espe et al., 2018; OECD/FAO, 2019; USDA-ERS, 2019).

In general, productivity prediction is performed using multiple linear regression. Although interesting, multiple regression models have some limitations, such as the size of the sample data. Specifically, when the observation number is less than the number of parameters, it is not possible to obtain the estimates using the usual estimation methods. Additionally, such models do not allow the adjustment of complex nonlinear relationships possibly existing in some data sets. Artificial neural networks (ANNs) provide an interesting alternative because they can capture nonlinear relationships between predictors and responses (GIANOLA et al., 2011; SKAWSANG et al., 2019) and ignore assumptions in the data sets.

The application of artificial intelligence, such as ANN, allows the capture of nonlinear effects among the data set and has been used in studies of prediction in plant breeding (SILVA et al., 2014 and 2017; SANT'ANNA et al., 2019). However, although ANNs are powerful predictive tools compared to conventional models, such as multiple linear regression (PARUELO & TOMASEL, 1997; OLDEN et al., 2002; BECK, 2018), they have the limitation of neglecting to quantify the importance of the variables.

Quantifying the importance of variables for prediction in breeding programmes allows for faster progress, selecting and predicting characteristics that have low heritability and/or measurement difficulty. Although simultaneous evaluation of characteristics provides a wide variety of information, identifying which predictor variable is most important is a challenge for breeders (PARMLEY et al., 2019). The quantification of the importance of variables can be performed by ANNs through algorithms such as GOH (1995), who proposed a modification in Garson's algorithm (1991) that consists of partitioning the neural network connection weights to determine the relative importance of each variable entering the network.

Other interesting alternatives for studies of the prediction and importance of variables are methodologies based on machine learning, such as decision trees (BEUCHER et al., 2019; PARMLEY et al., 2019) and their refinements, such as *bagging* (DEGENHARDT et al., 2019), *random forest*, and *boosting* (DEGENHARDT et al., 2019). Such methodologies allow good predictions and the importance of the characteristics to be obtained through measures based, for example, in the index of Gini and Entropy (HASTIE, 2009). These methodologies enable

the quantification of the impact of the disruption or disturbance of the input information on the estimate of the determination coefficient.

Methodologies based on regression, artificial intelligence, and machine learning have been used successfully in a prediction study. PARMLEY et al. (2019) evaluated the phenotypic characteristics of high dimensionality soybeans through a machine learning approach to predict seed yield regarding the prescriptive development of cultivars for agricultural practices. SKAWSANG et al. (2019) applied such methodologies to predict the population of insect pests using climatic and phenological factors of the host plant. However, there are no studies in the literature related to yield prediction and verification of the importance of variables for grain yield in rice culture. Unlike the methods of regression, artificial intelligence and machine learning do not make any prior assumptions about the data structure, in which it captures linear and nonlinear dependencies between the predictor and the response variables, making it a suitable tool for the researcher.

Given the above, this work aims to (1) predict grain yield, grain length and width ratio, and panicle length in flood-irrigated rice through regression, artificial intelligence, and machine learning methodologies; (2) quantify the best approaches to prediction; and (3) establish a network of better predictive power in flood-irrigated rice.

2. MATERIAL AND METHODS

1- Description of the experiment

The experiments were carried out in the State of Minas Gerais, Brazil, in the experimental fields of the Agricultural Research Corporation of Minas Gerais (EPAMIG), in the city of Leopoldina (21° 31' 48.01"S, 42° 38' 24" W), Lambari (21° 58'11.24" S, 45° 20' 59.6" W), and Janaúba (15° 48' 77" S, 43° 17' 59.09" W). Seventy-five genotypes of flood-irrigated rice belonging to the flood-irrigated rice breeding programme were evaluated in the agricultural year 2012/2013. The design was randomized blocks with three replications.

The evaluated characteristics were grain yield (GY, kg ha^{-1}), panicle length (LP, cm), and grain length-to-width ratio (LGW), which were used as response variables and the others as explanatory variables (inputs), that is, plant height (HP, cm), flowering (FL, days), lodging (LO), number of full grains per panicle (GP), percentage of full grains (FG, %), tillering (TI), length (GL, mm), width (GW, mm) and thickness (GT, mm) of grains, and weight of 100 grains

(WG, g). They were used to compose artificial neural networks of genotypes of flood-irrigated rice in the State of Minas Gerais.

2- Methodologies for predicting and verifying the importance of characteristics

a- Multiple Regression

Multiple regression, through the stepwise strategy (GHANI, 2010), was used to predict the variable responses to grain yield, panicle length, and grain length-to-width ratio as a function of the other measured variables and was considered to be explanatory.

The adopted model is represented by Equation 1.0:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon, \quad (1.0)$$

where y is the response variable (grain yield, panicle length or grain length-to-width ratio), x_1 a x_k are the explanatory variables, β_0 represents the intercept, β_1 e β_k are the linear coefficients associated with x_1 a x_k, and ε residual effect.

The estimate of the coefficient of determination R^2 was used to verify how much of the independent variable is explained by the total variation of the dependent variable.

The description of R^2 is found in Equation 2.0:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}, \quad (2.0)$$

where y: the observed values, and $\hat{y}$ predicted.

b- Artificial intelligence

For better network efficiency, before training and validation, the data were normalized in the range between -1 and 1. The training data set, in each location, was established by 2/3 of the phenotypic information, using the strategy of aggregating information from two of the three repetitions for training and the information from the other repetition used as a validation set. In this cross-validation strategy, individuals from each repetition participated at least once in the validation data set in cross-validation (k-fold) k = 3 partitions.

Perceptron Multilayer - PMC

The maximum number of training seasons was set at 5,000; the mean square error (MSE), as a criterion to stop processing the network, was defined as 1.0×10^{-3}. All trained networks

had a neuron in the output layer and a single hidden layer, with 15 neurons. The sigmoid tangent activation function was used in the hidden layer, and the training algorithm was *Bayesian regulation backpropagation*. To quantify the efficiency of the prediction R^2.

Importance of variables

To quantify the importance of variables through the PMC network, two techniques were used. The first is based on the GARSON (1991) algorithm modified by GOH (1995) (AG), which consists of partitioning the neural network connection weights to determine the relative importance of each input variable within the network. This algorithm describes the relative magnitude of the importance of the descriptors (predictor) in their connection with outcome variables through the dissection of synaptic weights from the neural network. In the second technique, the importance of variables (inputs) is assessed through the impact of the disruption or disturbance of the information of a given input on the estimation of the determination coefficient. Thus, this importance is estimated by exchanging information or by making constant the phenotypic values shown for each variable and verifying changes in the estimates of the R^2. When we disturb the values of a variable and R^2 decreases, there is an indication that the input variable is important about the others for purposes of prediction with the network already established.

Radial Base Function Network – RBF

The radial base function network is characterized by having only one hidden layer and making use of the Gaussian activation function (CRUZ & NASCIMENTO, 2018). The structure of the RBF to better predict grain yield, panicle length, and grain length-to-width ratio was established with 10 to 30 neurons (increased by 2, with each processing), and the radius established between 5 and 15 increased by 0.5. The efficiency of the prediction was measured by the R^2, and the relative importance of each entry was measured by the technique of destroying the information of each explanatory variable, as already described for PMC.

c- Machine Learning

To predict grain yield, panicle length, and grain length-to-width ratio and quantify the importance of variables through a machine learning approach, a decision tree and its refinements were used, *random forest*, *bagging,* and *boosting*. The R^2 measured the quality of the predictive model fit, and information from the minimum quadratic error (MSE) was used to quantify the importance of variables in flood-irrigated rice crops. The minimum square error was estimated as described in Equation 3.0 below:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad (3.0)$$

where y_i and $\hat{y}_i$ correspond to the observed and predicted values of observation in genotype *i*, respectively, and *n* is the total number of observations (variable, depending on the environment analysed).

In these techniques, the importance of the explanatory variable is the quantification of the mean decrease in the prediction precision, which consists of the estimate of the percentage of increment of minimum square error (IMSE), which is constructed when we exchange the values of each variable of the data set and are compared with the prediction of the original unchanged data set for the variable. Analogous to the regression analysis, it is the average increase of the squares of the residuals of the data set when the variable is exchanged (LI&ZHAN, 2019). Higher values of IMSE represent the importance of the highest variable. For better efficiency of the prediction estimate of the importance of variables, 5,000 trees were generated.

The analyses were performed with the aid of R software using the NeuralNetTools (BECK, 2018) and Genes (CRUZ, 2016) packages, which use an interface with MATLAB software (MATLAB, 2011).

3. RESULTS AND DISCUSSION

1- Prediction by different approaches

The estimate of R^2 for all methodologies using the explanatory variables to predict grain yield (GY), panicle length (PL), and grain length and width ratio (LGW) in flood-irrigated rice is shown in Figure 1. Based on Figure 1, it is possible to compare and define the variables that proved to be most efficient for the prediction of GY, PL, and LGW. Higher values of this estimate indicate that the target prediction variable has a better adjustment than the other explanatory variables (ROY et al., 2008; HASSANZADEH et al., 2015). Among the methodologies used in this study, it was found that multiple regression showed a lower estimate of R^2 (Figure 1) for the same variable, indicating the existence of nonlinear associations between the explanatory variables not considered in the model. Artificial intelligence and machine learning methodologies, in turn, stood out for their ability to extract nonlinear information from model inputs (PARMLEY et al., 2019; SKAWSANG et al., 2019), as seen in Figure 1. Other authors have already highlighted the abilities of neural networks to better

capture nonlinear relationships when compared to conventional methodologies (SILVA et al., 2014; SANT'ANNA et al., 2016).

The results obtained by different approaches show that there was a discrepancy between the maximum estimate of R^2 for all predictive variables in the same environments (Figure 1). The artificial intelligence approach in the Leopoldina environment provided a higher estimate for the predictive variables PL and GY in the RBF procedure, 83.44% and 78.90%, respectively. The GY response variable had the best estimate of R^2 in the Lambari and Janaúba environments in the PMC network with only one neuron in the output layer and a single hidden layer (Figure 1). In the Leopoldina and Lambari environments, for the LGW response variable, a maximum estimate of R^2 was approximately 100% by multiple regression and artificial intelligence approaches. On the other hand, it is variable in Janaúba, with a maximum estimate of 62%. The differences in the results obtained in these analyses indicate that the environment influences the estimation of R^2 and, consequently, the cause and effect relationships between the response variable and the set of explanatory variables.

Machine learning approaches proved to be more efficient than other approaches (Figure 1). There was a low estimate of R^2 for the predictive variable GY in the Janaúba environment in the *random forest* procedure, which corresponds to 18.57%. This result is inferior to all the approaches used in this study. In this same environment, but for bagging procedures, the estimate of R^2 was 94.76%. High estimates of R^2 (above 80%) were obtained using machine learning methodologies by the procedures *bagging* and *boosting* for all predictive variables (Figure 1). The decision tree (AD) and *random forest* methodologies did not stand out from the other machine learning procedures (Figure 1). SOUSA et al. (2020) emphasized that the AD's low predictive accuracy can be improved using ensemble methods such as *bagging*, *random forest*, and *boosting*. These strategies combine multiple AD to reduce the variability.

Random forests and *bagging* these methods have good predictive performances in practice; they work well for high-dimensional problems and can be used with multiclass output, categorical predictors, and imbalanced problems (GREGORUTTI et al., 2017). This author had satisfactory result variable selection with the *random forests* algorithm in the presence of correlated predictors.

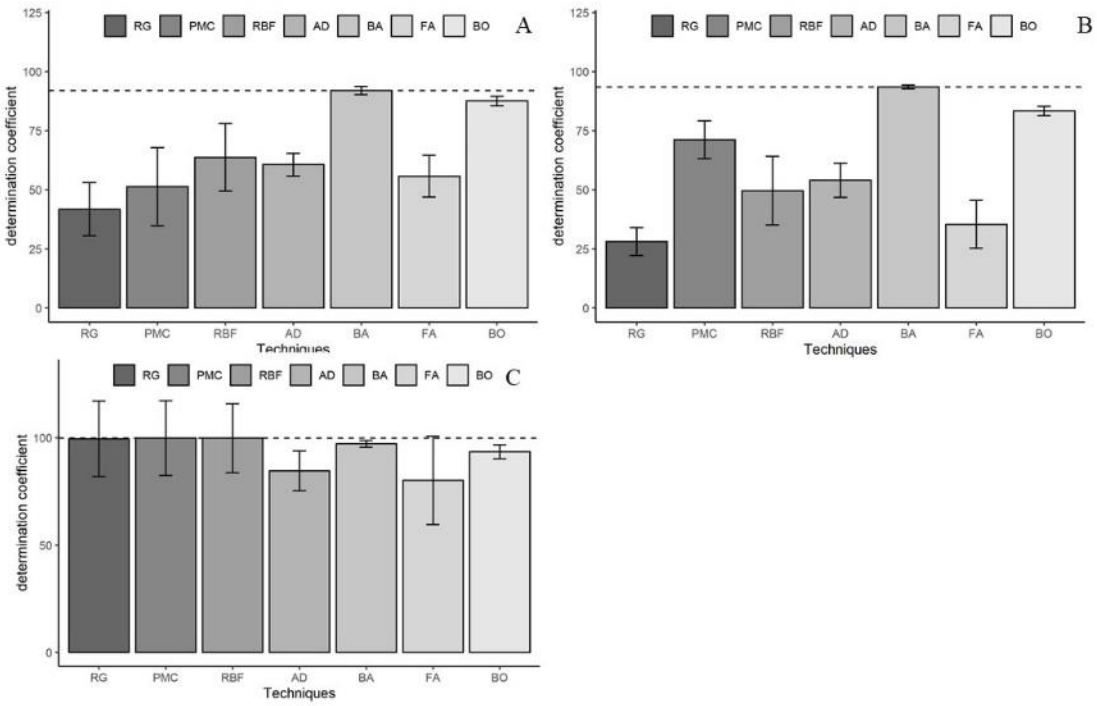


Figure 1: Maximum estimate of the coefficient of determination in three environments to predict grain yield (GY), panicle length (PL), and grain length and width ratio in flood-irrigated rice (LGW). A: panicle length; B: grain yield; C: grain length-to-width ratio; RG: multiple regression; PMC: multilayer perceptron; RBR: radial base network; AD: decision tree; FA: *random forest*; BA: *bagging*; BO: *boosting*.

When the variables are correlated, the simple correlation coefficient produces incomplete information. This is because a high correlation between two variables may have resulted from a third or a group of variables over another variable. Traditional methods, as well as path analysis, decompose into direct and indirect effects on the main variable, and logistic regression becomes unstable in the presence of high correlations. Multicollinearity is caused by the high correlation between the variables, which provides a problem of lack of adjustment of the model that affects the estimates of the parameters. In the literature, the ability of RNA to circumvent the problem of multicollinearity has already been highlighted (CRUZ & NASCIMENTO, 2018). These authors presented an application in which a response variable is predicted through five explanatory variables. By including a sixth explanatory variable, which would assume the same values as the fifth variable, it did not affect the accuracy of the ANN - Adaline in any way. However, they reinforce that in the classic multiple linear regression approach, there would be no solution, since there would be two columns, in the prediction matrix X, linearly dependent, so that the established multicollinearity would lead to an X'X matrix without a common inverse.

The efficiency of ANNs in prediction problems, given their ability to extract relevant information from large data sets (CHAGAS et al., 2013) and generalize relatively inaccurate information (PORWAL et al., 2003), was very well expressed by the results obtained (Figure 1). The same can be seen for methodologies based on machine learning, which are capable of handling more reduced or redundant information in the input variables (QUINLAN, 1996). However, another study as important as the prediction and which is often not carried out is the identification, among the explanatory variables, those of greater importance despite constituting important information in the process of understanding the adjusted model and decision making about dimensionality reduction in future studies (BEUCHER et al., 2019). Thus, after the prediction analysis, the quantification of the importance of variables was performed using artificial intelligence and machine learning methods to identify, among the set of explanatory variables, those that should be prioritized and identified as auxiliary characteristics in indirect responses to selection.

2- Importance of variables in prediction by the artificial intelligence approach

For ease of interpretation, we will denote R^2 the quality of prediction of the methodology and R^{2*} this same quality of adjustment after the disturbance in the explanatory variable.

Multilayer Perceptron (PMC)

Neural networks tend to perform well when compared to other predictive algorithms based on machine learning (SANTOS et al., 2018). These algorithms are capable of learning from linear and nonlinear relationships in the data (SOMERS et al., 2009; HADDOUCHE et al., 2018). It can also measure and incorporate direct effects and effects of interaction between variables in predictive models (TSANG et al., 2017).

The PMC network is widely used in the predictive process (GEDEON et al., 1995; SANTOS et al., 2018) since the success of this network has already been shown in several research groups that have shown mathematically that, with only a single hidden layer, this network works very well with different numbers of neurons in the hidden layer (DE OÑA et al., 2014; SANTOS et al., 2018).

The importance of the variables was quantified by assigning a zero value to the phenotypic information related to each variable to observe what changes would occur in the values of the R^{2*}. The results of the PMC network are shown in Table 1. It is important to note

that, in this table, reductions in the values of R^{2*} after assigning zero value to the phenotypic information referring to each variable, they are indicative that this variable is important about the others for purposes of prediction with the network already established.

Table 1. Estimates of the coefficient of determination, provided by the use of the PMC, to predict grain yield, panicle length and grain length and width after disturbance (zero value assignment) in the explanatory variable values.

	PL				GY				LGW		
Input	E1	E2	E3	Input	E1	E2	E3	Input	E1	E2	E3
LO	48.65	51.12	36.67	LO	**8.08**	24.54	**5.33**	LO	98.47	99.98	63.03
HP	**9.07**	47.62	47.92	HP	**0.04**	48.22	**7.97**	HP	99.98	99.98	62.70
GL	**36.37**	**41.86**	**6.77**	GL	**0.52**	16.01	**7.54**	GL	**37.58**	**46.56**	**20.09**
GT	**37.37**	47.66	28.78	PL	**6.43**	34.78	15.29	PL	99.97	99.97	63.00
FL	46.40	46.01	32.09	GT	**7.42**	22.20	13.83	GT	99.95	99.97	61.22
GW	**37.38**	51.31	**18.46**	FL	12.03	**7.34**	12.96	FL	99.86	99.96	58.88
GP	46.53	51.07	32.97	GW	16.30	21.61	**2.12**	GW	**39.64**	**36.55**	**43.32**
WG	47.95	51.14	20.28	GP	10.72	57.00	**5.27**	GP	99.98	99.98	62.96
TI	47.90	50.02	24.57	WG	17.64	17.01	9.68	WG	99.98	99.97	62.76
FG	45.27	50.26	**3.73**	TI	**1.70**	39.68	**4.35**	TI	99.99	99.97	62.30
GY	41.68	49.80	21.73	FG	10.81	27.50	11.17	FG	99.98	99.97	63.01
LGW	40.83	**17.43**	31.29	LGW	21.32	13.31	26.62	GY	99.97	99.97	63.00

LO: lodging, HP: height (cm), GL: grain length (mm), PL: panicle length (cm), GT: grain thickness (mm), FL: flowering (days), GW: grain width (mm), GP: number of filled grains per panicle, WG: weight of 100 grains (g), TI: tillering, FG: percentages of filled grains and LGW: length-to-width ratio of grains, GY: grain yield, Environment, E1: Leopoldina, E2: Lambari, E3: Janaúba.

The results in Table 1 show great discrepancies in the R^{2*} when comparing the environments with each other, which makes interpretation difficult. For the response variable LGW, it was efficient to quantify grain length and width due to the reduction in the estimate of R^2 as a result of the strategy of assigning a zero value to phenotypic information. It should be remembered that such changes must be seen concerning the values of the R^2 of prediction, which was approximately 100% in the environments of Leopoldina and Lambari, and Janaúba was 63% (Figure 1). For Leopoldina, when zeroing the variables, for example, HP, GL, and TI, the R^{2*} values of these variables were 0.04, 0.52, and 1.70, respectively (Figure 1). This result shows that these variables are important in predicting GY because the disturbance in their values has led to a considerable reduction in the quality of the adjustment. In Lambari, the variable that presented the highest contribution was FL. Independent of the predictive variable

in PMC, with only one neuron in the output layer and a single hidden layer, they agreed to point out that the most important variables were grain width and length, given the significant falls in the values of the estimate of R^{2*} observed when zeroing the variable.

To overcome the difficulties faced when adopting PMC networks to study the importance of variables, an alternative is to use the AG algorithm, which takes into account the partitioning of the RNA connection weights to determine the relative importance of each input variable within the network. The weights that connect neurons in an ANN are partially analogous to the coefficients in a generalized linear model (BECK, 2018) so that the combined effects of weights in the model's predictions represent the relative importance of predictors in their associations with the variable of the predictor. The large number of adjustable weights in an artificial neural network makes it very flexible in modelling nonlinear effects but imposes challenges for its interpretation. In this algorithm, the numbers of neurons were used to obtain the maximum estimate of R^{2*} for a better estimate of the relative contribution of variables.

The percentages of the relative contribution estimated by the GA method are described in Table 2. In this Table, for the GY response variable, the results were consistent in pointing plant height (HP), flowering (FL), and the number of full grains per panicle (GP) in terms of relative contribution. For the variable response PL, the variable with the greatest relative contribution was grain yield (GY) in the environments of Leopoldina and Lambari; however, in Janaúba, the variable that stood out was the length and width of grains. Regarding the explanatory variable LGW, the percentages of the relative contribution revealed that the variables grain length and grain width had the largest relative contribution. This result was expected since the length and width of grain variables are determinants of LGW. The results indicate that the GA approaches are efficient in quantifying the importance of variables in studies involving PMC neural networks.

Table 2. Percentages of the relative contribution estimated by the method of Garson (1991) modified by Goh (1995) of 12 variables to predict grain yield, panicle length, and grain length and width ratio in flood-irrigated rice in three environments in the State of Minas Gerais.

	GY				PL				LGW		
PV	E1	E2	E3	PV	E1	E2	E3	PV	E1	E2	E3
LO	6.50	6.00	5.57	LO	8.48	6.94	7.63	LO	6.94	7.74	7.97
HP	**11.00**	**16.0**	**15.12**	HP	8.30	**8.99**	8.19	HP	7.81	8.16	7.14
GL	8.90	6.10	5.96	GL	6.89	8.00	7.60	**GL**	**9.84**	**9.26**	**9.23**
PL	8.26	8.51	8.12	GT	**8.52**	8.40	8.15	PL	8.65	8.42	8.54
GT	6.80	6.02	6.00	FL	7.58	8.22	8.20	**GT**	**8.77**	**9.19**	**10.48**
FL	**13.00**	**12.8**	**13.55**	GW	8.30	8.64	8.20	FL	7.49	8.50	8.90
GW	6.60	5.67	6.79	GP	8.15	8.44	8.30	GW	8.00	8.60	9.29
GP	**11.20**	**14.10**	**13.02**	WG	8.10	8.00	**8.80**	GP	8.71	8.70	8.06
WG	7.42	6.60	6.80	TI	8.04	8.30	7.89	WG	8.27	7.99	8.68
TI	6.82	5.50	5.85	FG	8.75	7.89	**9.48**	TI	8.92	8.12	5.14
FG	7.00	6.50	6.55	GY	**10.00**	**9.23**	8.50	FG	7.99	7.92	7.62
LGW	6.50	6.20	6.67	LGW	**8.89**	8.95	**9.06**	GY	8.61	7.40	8.95

PV: predictive variable, LO: lodging, HP: height (cm), GL: grain length (mm), PL: panicle length (cm), GT: grain thickness (mm), FL: flowering (days), GW: grain width (mm), GP: number of filled grains per panicle, WG: weight of 100 grains (g), TI: tillering, FG: percentages of filled grains and LGW: length-to-width ratio of grains, GY: grain yield. Environment E1: Leopoldina, E2: Lambari, E3: Janaúba.

Radial Base Network (RBF)

The quantification of the importance of flood-irrigated rice characters by assigning a zero value to the information of an input variable after the RBF was established was performed and is described in Table 3. In this table, the values are used after causing disturbances in the input variables with the action of assigning zero value of the variable in each explanatory variable. When using this strategy of zeroing the value of the variable, drastic reductions in the values of R^{2*} were observed for the most important length (GL) and grain width (GW) variables when the target prediction variable was LGW. For other response variables, this result was very discrepant in quantifying the true importance of variables. When the explanatory variable was GY, in Janaúba, the variables that suffered the greatest reduction in R^{2*} were flowering - R^{2*} = 23.80 and weight of 100 grains (WG) - R^{2*} = 19.91; in Leopoldina, plant height variables were observed (HP) - R^{2*}= 21.26, grain width (GW) - R^{2*}= 24.83 and weight of 100 grains (WG) = 24.25; and in Lambari, the most important variable using this approach was flowering (FL) - R^{2*}= 28.43.

For the variable response PL, we observed changes in the values of R^{2*} in Leopoldina and Lambari for the variable flowering (FL) - R^{2*}= 47.77 and R^{2*}= 46.76, respectively. In Leopoldina, the percentages of full grains (FG) - R^{2*}= 25.51 also showed a drastic reduction in R^{2*}. In Lambari, lower estimates of R^{2*} were obtained for the variable weight of 100 grains (WG) - R^{2*}= 45.60. For Janaúba, the results show that the most important variables using the RBF were grain width (GW) - R^{2*}= 19.76 and weight of 100 grains (WG) - R^{2*} = 23.11.

Table 3. Estimates of the coefficient for determining the grain yield prediction, panicle length, and grain length-to-width ratio using the RBF assigning zero value to the genotype information.

	PL				GY				LGW		
PV	1	2	3	PV	1	2	3	PV	1	2	3
LO	83.20	62.80	43.61	LO	**21.26**	45.58	41.26	LO	99.91	99.84	65.72
HP	53.72	61.65	45.73	HP	70.88	47.39	49.52	HP	99.91	99.92	64.65
GL	70.57	60.72	44.23	GL	29.42	46.46	43.32	GL	**40.93**	**43.11**	**29.84**
GT	61.09	61.93	27.36	PL	38.12	40.79	36.92	PL	99.91	99.90	63.07
FL	**47.77**	**46.76**	41.69	GT	42.21	45.22	36.63	GT	99.83	99.89	62.68
GW	71.73	50.95	**19.76**	FL	30.67	**28.43**	**23.80**	FL	99.73	99.53	65.56
GP	59.41	62.16	29.75	GW	**24.83**	44.75	34.95	GW	**44.98**	**43.19**	**46.46**
WG	64.21	**45.60**	**23.11**	GP	40.96	44.99	34.98	GP	99.88	99.89	63.33
TI	67.71	63.58	26.15	WG	**24.25**	45.43	**19.91**	WG	99.89	99.74	62.52
FG	**25.51**	59.63	27.70	TI	25.58	45.95	35.04	TI	99.89	99.83	63.96
GY	54.38	56.28	31.91	FG	26.60	46.19	35.61	FG	99.84	99.78	63.69
LGW	71.29	60.59	44.11	LGW	31.42	44.27	29.73	GY	99.89	99.82	62.63

PV: predictive variable, LO: lodging, HP: height (cm), GL: grain length (mm), PL: panicle length (cm), GT: grain thickness (mm), FL: flowering (days), GW: grain width (mm), GP: number of filled grains per panicle, WG: weight of 100 grains (g), TI: tillering, FG: percentages of filled grains and LGW: length-to-width ratio of grains, GY: grain yield. Environment E1: Leopoldina, E2: Lambari, E3: Janaúba.

Therefore, there is a certain agreement between the results found by the two computational intelligence methodologies of PMC networks and RBF networks.

3- Importance of variables in prediction by the machine learning approach

Table 4 shows the averages of the relative contributions of the explanatory variables for predicting grain yield, panicle length, and grain length-to-width ratio by estimating the percentage of minimum square error increment (IMSE), which is constructed by exchanging the values of each variable in the data set and comparing it with the prediction of the original

unix exchange data set for the variable. In this case, unlike the strategy used for the computational intelligence methodologies of PMC and RBF networks, for which lower values of R^{2*} indicated greater importance of that variable for the model, in the machine learning approach, the importance of the explanatory variable is related to the estimate of the average decrease in the accuracy of the model through IMSE so that the higher this estimate the greater the importance of the variable.

Table 4. The average estimate of the relative contributions of the explanatory variables for predicting grain yield, panicle length, and grain length-to-grain ratio in flood-irrigated rice continues using a machine learning approach in three environments in Minas Gerais.

	PL								
	BA			FA			BO		
PV	E1	E2	E3	E1	E2	E3	E1	E2	E3
LO	0	-1.13	0	0	-1.34	0	0	2.31	0
HP	10.89	7.65	-0.05	11.37	7.43	-1.27	13.24	7.96	6.31
GL	2.84	0.28	-0.04	2.97	0.62	0.5	8.33	8.2	11.51
GT	2.37	1.02	0.97	1.46	1.74	1.41	9.14	7.93	8.94
FL	-0.96	6.69	-1.13	-2.01	5.59	-2.4	3.83	4.66	1.25
GW	3.32	12.01	1.76	2.99	10.38	1.78	7.57	18.62	12.3
GP	7.68	0.8	-0.04	5.82	0.67	0.29	11.37	9.28	15.98
WG	1.71	0.87	1.23	0.61	1.74	2.35	5.07	9.82	9.49
TI	-0.16	-0.16	-0.45	0.35	1.17	-0.84	1.59	0	3.65
FG	4.12	-1.75	0.35	3.58	-1.27	-0.42	8.52	6.41	9.09
GY	-0.38	5.21	2.39	0.74	5.08	3.2	13.55	13.76	10.68
LGW	5.56	7.24	-1.37	5.85	7.15	-0.53	15.88	10.79	11.73

	GY								
	BA			FA			BO		
PV	E1	E2	E3	E1	E2	E3	E1	E2	E3
HP	2.52	2.63	-0.68	2.98	2.02	-1.27	9.93	7.98	11.2
LO	0	-0.87	0	0	-0.19	0	0	5.21	0
GL	2.39	-0.92	1.19	3.68	0.1	0.14	11.97	12.54	14.05
PL	1.07	10.33	-0.27	-0.22	10.32	-1.23	8.2	14.8	13.52
GT	0.43	-2.42	-0.47	1.49	-1.19	-2.61	7.52	6.21	7.91
FL	-2.5	5.51	-0.72	-1.72	6.67	-1.73	5.6	5.39	1.71
GW	-0.9	1.46	-3.84	-0.63	1.16	-2.25	10.91	4.59	8.28
GP	2.34	-0.41	-2.72	1.76	2.43	-3.34	13.4	15.09	10.39
WG	-0.25	1.87	1.24	2.15	1.36	1.64	10.95	9.14	11.59
TI	-0.33	0.5	-0.96	-1.98	0.6	-1.06	2.26	0.76	2.23
FG	0.64	-0.32	0.41	-1.14	1.02	1.25	5.22	7.57	8.3
LGW	0.37	1.53	-0.31	0.39	2.52	0.78	9.61	7.41	9.15

	LGW								
	BA			FA			BO		
PV	E1	E2	E3	E1	E2	E3	E1	E2	E3
LO	0	1.62	0	0	-0.7	0	0	1.75	0
HP	-0.13	-0.16	1.06	0.3	0.85	0.92	3.76	4.37	6.92
GL	18.99	18.32	18.37	19.3	17.73	20.26	25.51	27.33	22.62
PL	8.56	11.82	-0.89	8.53	12.01	-1.68	8.79	14.57	10.29
GT	1.87	2.43	-0.36	1.14	2.62	0.98	4.67	3.77	6.05
FL	3.95	3.67	1.84	1.34	3.58	0.93	2.81	1.45	0.83
GW	19.65	17.23	11.32	19.28	18.28	9.88	20.41	20.37	16.46
GP	1.58	0.46	0.81	3.59	1.82	-1.16	10.36	7.26	9.44
WG	9.52	0.21	-0.14	7.61	0.85	-0.76	9.34	6.51	8.42
TI	-0.94	-1.07	-0.72	-1.33	-0.82	1.01	1.66	0	2.81
FG	-0.22	2.22	0.28	-1.28	0.16	2.83	3.83	4.54	5.22
LGW	-1.35	3.01	1.37	-0.72	2.67	0.67	7.56	6.18	10.52

PV: predictive variable, LO: lodging, HP: height (cm), GL: grain length (mm), PL: panicle length (cm), GT: grain thickness (mm), FL: flowering (days), GW: grain width (mm), GP: number of filled grains per panicle, WG: weight of 100 grains (g), TI: tillering, FG: percentages of filled grains and LGW: length and width ratio of grains, GY: grain yield. Environment E1: Leopoldina, E2: Lambari, E3: Janaúba, FA: random forest, BA: bagging, BO: boosting

Based on Table 4, the variables that obtained the highest estimate in all machine learning methodologies were length (GL) and grain width when the prediction target variable was grain length and width ratio (LGW) in all environments. For this same response variable, another variable that had a high IMSE estimate was panicle length (PL) in Leopoldina and Lambari, and Janaúba did not consider this variable to be the most important due to the low estimate of the IMSE percentage. On the other hand, the weight variables of 100 grains (WG) and the number of full grains per panicle (GP) proved to be efficient in quantifying the prediction of LGW by *boosting*. This procedure proved to be more consistent in predicting variables compared to the others.

The variable that obtained the highest IMSE estimate when PL was the target prediction variable was plant height (HP) for Leopoldina and Lambari. On the other hand, this variable in Janaúba was not highlighted in predicting PL. In Leopoldina, another variable that stood out in predicting PL was the number of grains filled per panicle (GP) for all machine learning approaches. When using the explanatory variable PL, the variable GY presented the highest IMSE in Janaúba for procedure *bagging*. Regarding the procedure *boosting* and about the same predictive variable, the results show discrepancies. On the other hand, this procedure was more consistent in predicting the variable. In this procedure, to quantify the importance of a variable using PL as a predictive target, the variables GP, GY, and LGW stood out in Leopoldina. In Lambari, other variables showed better performance in predicting PL, for example, GW, GY, and LGW, and in Janaúba, they were PL, GW, GP, GY, and LGW.

When the target prediction variable was GY, in Leopoldina, the variables that obtained an estimate of the high IMSE percentage were plant height (HP) and grain length (GL) in all machine learning procedures. On the other hand, in Lambari, the variable that stood out was panicle length (PL). In this environment, another variable that showed better predictive performance when GY was used as the main variable was flowering (FL) in *bagging* and *random forest.* In the *boosting* procedure, the variables that stood out were HP, GL, PL, GP, WG, and LGW in all environments.

The literature has highlighted machine learning techniques as efficient tools in quantifying the relative importance of variables, in view of simplicity, the nonuse of assumptions about the distribution of explanatory variables, and their robustness to quantity, redundancy, and environmental influences (TAN et al., 2014; BEUCHER et al., 2019). On the other hand, we verify this premise for the regression method. *Random forests* and *bagging* these methods have good predictive performances in practice; they work well for high-dimensional

problems and can be used with multiclass output, categorical predictors, and imbalanced problems (GREGORUTTI et al., 2017). This author had satisfactory result variable selection with the *random forests* algorithm in the presence of correlated predictors.

Grain yield is a trait controlled by several genes and is therefore a quantitative inheritance (FREITAS et al., 2007). Therefore, grain yield depends on the interaction of several yield components, for example, numbers of spikelets and grains per panicle, mass of a thousand grains, spike fertility index and panicle length, which are controlled by genetic factors, and environmental factors. The length of the panicle, the number of spikelets per panicle, the fertility of the spikelets, and the mass of a thousand grains directly affect grain yield (EVANS, 1977). Thus, knowledge of these relationships can help breeders select new cultivars, which can increase the productivity and quality of grains and decrease the cost of production and the environmental impact.

The longer the flowering period in the rice culture, the more photoassimilates are produced and translocated to the grains, and consequently, an increase in grain yield. However, late-cycle cultivars tend to be more productive about the early cycle since they obtain an increase in the amount of photoassimilates that are translocated to the grains. According to NTANOS & KOUTROUBAS (2002), productivity in rice has been justified by differences in the dynamics of the distribution of assimilates between organs during plant growth and development. From the results of these studies, it was found that the production of dry matter and the translocation of photoassimilates contributed significantly to the development of grains in different cultivars and, consequently, a direct relationship with grain yield.

Grain dimensions are the main determinants of grain weight and one of the three components (number of panicles per plant, number of grains per panicle, and weight of grains) of grain yield; therefore, they are important characteristics that affect yield in rice. In plant breeding applications, grain size is generally assessed by the weight of the grain, which is positively correlated with various characteristics, including the length, width, and thickness of the grain (FAN et al., 2006). These characteristics also influence acceptability for consumers, and therefore, the size/shape of the rice grain is an important preferential target characteristic for breeders (HUANG et al., 2012; ANACLETO et al., 2015). Cultivars of the short and long types are highly preferred by many consumers in Japan, South Korea, and North China, while consumers in India, the USA, and other countries in South and Southeast Asia prefer long and medium grains (MISRA et al., 2017).

Methodologies based on machine learning and computational intelligence do not depend on stochastic information and tend to be more efficient. These methodologies make no assumptions about the model but capture complex factors such as epistasis and dominance in prediction models. It is not necessary to know if the data have these effects and do not require any assumptions about the distribution of phenotypic values (SOUSA et al., 2020). Machine learning algorithms have the advantage of modelling data in a nonlinear and a nonparametric manner (OSCO et al., 2020). Unlike many traditional statistical methods, these algorithms are built with the advantage of dealing with noisy, complex, and heterogeneous data (OSCO et al., 2019).

In this study, we compare different approaches to quantifying the importance of variables to identify relevant predictive variables within a regression problem. Additionally, we included in our comparison a traditional method that aims to find a small subset of important variables with ideal forecasting performance in flood-irrigated rice.

It is noteworthy that the 13 characteristics used in this study are laborious to obtain, and their evaluation can be costly if there are a greater number of genotypes to be evaluated. In this context, the study of the most important characteristics in prediction is necessary, since it is possible to reduce physical effort, cost, labour, and time in experimentation (PALIWAL & KUMAR, 2011; FERREIRA et al., 2015).

Predicting the importance of flood-irrigated rice characteristics is of paramount importance for breeding programmes, as it directs genotype selection more practically, in addition to serving as a theoretical and practical framework in support of new recommendation cultivars. In practical terms, these results are consistent.

Therefore, our study presents the performance of some methodologies to evaluate the relative contributions of each variable through computational intelligence and machine learning in flood-irrigated rice culture. An approach to quantify the effect of explanatory variables on genetic improvement has successfully identified the true importance of each variable, including those that exhibit strong and weak correlations with the main variables, which in our case are grain yield, length of panicle and grain length-to-width ratio.

Researchers can now identify the individual and interactive contributions of the predictor variables to the rice crop using artificial intelligence and machine learning.

4. CONCLUSION

Computational intelligence and machine learning methodologies were able to quantify the importance of explanatory variables in the prediction of grain yield in rice, grain length and width ratio, and panicle length. In addition to artificial intelligence and machine learning, it is able to handle more reduced or redundant information in the input variables.

The characteristics able to assist in decision making are flowering, number of grains filled by panicles, and panicle length.

The network with only one hidden layer with 15 neurons was efficient in determining the relative importance of variables in flooded rice.

5. REFERENCE

Amiri, S.S., Mottahedi, M., Asadi, S. (2015). Using multiple regression analysis to develop energy consumption indicators for commercial buildings in the US. Energy Build., 109, 209–216. https://doi.org/10.1016/j.enbuild.2015.09.073.

Anacleto, R. *et al*. (2015). Prospects of breeding high-quality rice using post-genomic tools. Theor. Appl. Genet. 128, 1449–1466.

Beck, M. (2018). NeuralNetTools: Visualization and Analysis Tools for Neural Networks**.** R package version 1.5.2. https://CRAN.R-project.org/package=NeuralNetTools. http://dx.doi.org/10.18637 / jss.v085.i11.

Beucher, A., Møller, A.B., Greve, M.H. (2019). Artificial neural networks and decision tree classification for predicting soil drainage classes in Denmark. Geoderma. 352, 351-359. https://doi.org/10.1016/j.geoderma.2017.11.004.

Cruz, C.D. (2016). Genes Software – extended and integrated with the R, Matlab and Selegen. Acta Scientiarum. 38(4), 547-552, http://dx.doi.org/10.4025/actasciagron.v38i4.32629.

Cruz, C.D., & Nascimento, M. (2018). Inteligência computacional aplicada ao melhoramento genético. Viçosa- MG: Editora UFV, 414p.

De Oña, J., Garrido, C. (2014). Extracting the contribution of independent variables in neural network models: a new approach to handle instability. Neural Comput Appl. 25(3–4):859–869.

Degenhardt, F. Seifert, S., Szymczak, S. (2019). Evaluation of variable selection methods for random forests and omics data sets, Briefings in Bioinformatics,20(2), 492–503, https://doi.org/10.1093/bib/bbx124.

Evans, L.E, Bhatt, G.M. (1977). Influence of seed size, protein content and cultivar on early seedling vigor in rice. Canadian Journal of Plant Science; Otawa.57, 929-935.

Fan, C., Xing, Y., Mao, H. *et al.*(2006). *GS3*, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein. Theor Appl Genet 112**,** 1164–1171. https://doi.org/10.1007/s00122-006-0218-1.

Ferreira, M.G., Azevedo, A.M., Siman, L.I., Silva, G.H., Carneiro, C.S., Alves, F.M., Delazari, F.T., Silva, D.J.H., Nick, C. (2017). Automation in accession classification of *Brazilian Capsicum* germplasm through artificial neural networks. Scientia Agricola. http://dx.doi.org/10.1590/1678-992X-2015-0451.

França, M.G.C., Rossiello, R.O.P., Zonta, E., Araújo, A.P., Ramos, F.T. (1999). Desenvolvimento radicular e influxo de nitrogênio em duas cultivares de arroz. Pesquisa Agropecuária Brasileira, 34, 1845-1853. http://dx.doi.org/10.1590/S0100-204X1999001000011.

Freitas, J.G., Cantarella, H., Salomon, M.V., Malovolta, V.M.A., Castro, L.H.S.M., Gallo, P.B., Azzini, L.E. (2007). Produtividade de cultivares de arroz irrigado resultante da aplicação de doses de nitrogênio. Bragantia, Campinas, 66(2), 317-325. http://dx.doi.org/10.1590/S0006-87052007000200016.

Garson, G.D. (1991). Interpreting neural network connection weights. Artificial Intelligence Expert 6:46-51.

Gedeon, T.D., Wong, P.M., Harris, D. (1995). editors. Balancing bias and variance: network topology and pattern set reduction techniques. Berlin, Heidelberg: Springer Berlin Heidelberg.

Ghani, I.M.M., Ahmad, S. (2010). Stepwise Multiple Regression Method to Forecast Fish Landing. *Procedia—Soc. Behav. Sci. 8.* https://doi.org/10.1016/j.sbspro.2010.12.076.

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 Genet 12, 87. https://doi.org/10.1186/1471-2156-12-87.

Goh, A.T.C. (1995). Back-propagation neural networks for modeling complex systems. Artificial Intelligence in Engineering. 9:143-51. https://doi.org/10.1016/0954-1810(94)00011-S.

Gregorutti, B., Michel, B., Saint-Pierre, P. (2017). Correlation and variable importance in random forests. Stat Comput. 27:659–678. https://doi.org/10.1007/s11222-016-9646-1.

Haddouche, R., Chetate, B., Said Boumedine, M. (2018). Neural network ARX model for gas conditioning tower. International Journal of Modeling and Simulation.1–12.

Hassanzadeh, Z., Ghavami, R., Kompany-Zareh, M. (2015). Radial basis function neural networks based on the projection pursuit and principal component analysis approaches: QSAR analysis of fullerene[C60]-based HIV-1 PR inhibitors. Medicinal Chemistry Research. https://doi.org/10.1007/s00044-015-1466-x.

Hastie, T., Tibshirani, R., Friedman, J. (2009). The Elements of Statiscal Learning Data Mining, Inference, and Prediction, 2nd ed. New York, NY: Springer, 745p.

Huang, X. *et al*. (2012a) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nature Genet. 44 , 32–39, https://doi.org/10.1038/ng.1018 .

Li, L., & Zha, Y. (2019). Estimating monthly average temperature by remote sensing in China. Advances in Space Research. 63(8), 2345-2357. https://doi.org/10.1016/j.asr.2018.12.039.

Misra, G., Badoni, S., Anacleto, R. *et al*. (2017). Whole genome sequencing-based association study to unravel genetic architecture of cooked grain width and length traits in rice. Sci Rep 7, 12478. https://doi.org/10.1038/s41598-017-12778-6.

Ntanos, D.A., &Koutroubas, S.D. (2002). Dry matter and Naccumulation and translocation for Indica and Japonica riceunder Mediterranean conditions. Field Crops Research, 74, 93-101. https://doi.org/10.1016/S0378-4290(01)00203-9.

Olden, J.D., Jackson, D.A. (2002). “Illuminating the “Black Box”: A Randomization Approach for Understanding Variable Contributions in Artifical Neural Networks.” Ecological Modelling, 154(1–2), 135–150. https://doi.org/10.1016/s0304-3800(02)00064-9.

Olden, J.D., Joy, M.K., Death, R.G. (2004). “An Accurate Comparison of Methods for Quantifying Variable Importance in Artificial Neural Networks Using Simulated Data.” Ecological Modelling, 178(3–4),389–397. https://doi.org/10.1016/j.ecolmodel.2004.03.013.

Osco, L.P., Ramos, A.P.M., Moriya, E.A.S., Bavaresco, L.G., Lima, B.C., Estrabis, N., Pereira, D.R., Creste, J.E., Marcato Junior, J., Gonçalves, W.N. (2019). Modeling hyperspectral response of water-stress induced lettuce plants using artificial neural networks. Remote Sens. 11: 2797.

Osco, L.P., Ramos, A.P.M., Pinheiro, M.M.F., Moriya, E.A.S., Imai, N.N., Estrabis, N., Lanczyk, F., Araujo, F.F., Liesenberg, V., Jorge, L.A.C., Li. J., Ma, L., Gonçalves, W.N., Junior, J.M. and Creste, J.E. (2020). A Machine Learning Framework to Predict Nutrient Content in Valencia-Orange Leaf Hyperspectral Measurement. Remote Sens. 12: 906. http://dx.doi.org/10.3390/rs12060906.

Parmley, K.A., Higgins, R.H., Ganapathysubramanian, B. Sarkar, S & Singh, A. K. (2019). Machine Learning Approach for Prescriptive Plant Breeding. Sci Rep 9, 17132. https://doi.org/10.1038/s41598-019-53451-4.

Paruelo, J.M., Tomasel, F. (1997). “Prediction of Functional Characteristics of Ecosystems: A Comparison of Artificial Neural Networks and Regression Models.” Ecological Modelling, 98(2–3), 173–186. https://doi.org/10.1016/s0304-3800(96)01913-8.

Paswan, R.P., Begum, S.A. (2013). Regression and Neural Networks Models for Prediction of Crop Production. Int. J. Sci. Eng. Res., 4, 11.

Porwal, A., Carranza, E.J.M., Hale, M. (2003). Artificial neural networks for mineral potential mapping; a case study from Aravalli Province, Western India Nat. Resour. Res., 12 (3) 155-171. https://doi.org/10.1023/A:1025171803637.

Quinlan, J.R. (1996). Learning decision tree classifiers ACM Comput. Surv., 28 (1). 71-72.

Roy, P.P., Roy, K. (2008). On some aspects of variable selection for partial least squares regression models. QSAR Comb Sci 27:302–313. doi:10.1002/qsar.200710043.

Sant'Anna, I.C., Ferreira, R.A.D.C., Nascimento, M., Carneiro, V.Q., Silva, G.N., Cruz, C.D., Oliveira, M.S., Chagas, F.E.O. (2019). Multigenerational prediction of genetic values using genome-enabled prediction. PLoS One, 14, e0210531. https://doi.org/10.1371/journal.pone.0210531.

Santos, A.B., Costa, J.D. (1995). Comportamento de variedades de arroz de sequeiro em diferentes populações de plantas, com e sem irrigação suplementar. Scientia Agricola, 52, 1-8. http://dx.doi.org/10.1590/S0103-90161995000100002.

Santos, I.G., Carneiro, V.Q., Silva Junior, A.C., Cruz, C.D. et al. (2019). Self-organizing maps in the study of genetic diversity among irrigated rice genotypes. *Acta Sci. Agron*. 41. https://doi.org/10.4025/actasciagron.v41i1.39803.

Santos, R. P, Dean, D. L., Weaver, J. M. & Hovanski, Y. (2018). Identifying the relative importance of predictive variables in artificial neural networks based on data produced through a discrete event simulation of a manufacturing environment. Journal International Journal of Modelling and Simulation. https://doi.org/10.1080/02286203.2018.1558736.

Shah, S.H., Angel, Y., Houborg, R., Ali, S., McCabe, M.F. (2019). A random forest machine learning approach for the retrieval of leaf chlorophyll content in wheat. Remote Sens. 11: 920.

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 Agropecuaria Brasileira, v. 52, p. 186-193. http://dx.doi.org/10.1590/s0100-204x2017000300009.

Silva, G.N., Tomaz, R.S., Sant'anna, I.C., Nascimento, M., Bhering, L.L., Cruz, C.D. (2014). Neural networks for predicting breeding values and genetic gains. Scientia Agricola, 71, 494-498. http://dx.doi.org/10.1590/0103-9016-2014-0057.

Skawsang, S. Nagai, M. Nitin, K. and Soni, P. (2019). Predicting Rice Pest Population Occurrence with Satellite-Derived Crop Phenology, Ground Meteorological Observation, and Machine Learning: A Case Study for the Central Plain of Thailand. Appl. Sci., 9(22), 4846; https://doi.org/10.3390/app9224846.

Somers, M.J., Casal, J.C. (2009). Using artificial neural networks to model nonlinearity: the case of the job satisfaction—job performance relationship. Organizational Res Methods. 12(3):403–417.

Sousa, I.C., Nascimento, M., Silva, G.N., Nascimento, A.C.C., Cruz, C.D., Fonseca, F., Almeida, D.P., Pestana, K.N., Azevedo, C.F., Zambolim, L. and Caixeta, E.T. (2020). Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. Scientia Agricola 78: 1–8. http://dx.doi.org/10.1590/1678-992x-2020-0021.

Stefaniak, B., Cholewiński, W., Tarkowska, A. (2005). Algorithms of Artificial Neural Networks - Practical application in medical science. Polski Merkuriusz Lekarski.19:819-22.

Tsang, M., Cheng, D., Liu, Y. (2017). Detecting statistical interactions from neural network weights. arXiv preprint arXiv:170504977.

Ventura, R.V., Silva, M.A., Medeiros, T.H., Dionello, N.L., Madalena, F.E., Fridrich, A.B., Valente, B.D., Santos, G.G., Freitas, L.S., Wenceslau, R.R., Felipe, V.P.S., Corrêa, G.S.S. (2012). Use of artificial neural networks in breeding values prediction for weight at 205 days in Tabapuã beef cattle. Arquivo Brasileiro de Medicina Veterinária e Zootecnia, 64, 411-418. http://dx.doi.org/10.1590/S0102-09352012000200022.

Yu, H., Campbell, M.T., Zhang, Q., Walia, H., and Morota, G. (2019). Genomic Bayesian confirmatory factor analysis and Bayesian network to characterize a wide spectrum of rice phenotypes. G3: Genes, Genomes, Genetics. 9:1975-1986. http://dx.doi.org/10.1101/435792.

CAPÍTULO 3

Inteligência computacional para estudar a importância de preditores em aveia branca (*Avena sativa* L.)

VIÇOSA - MINAS GERAIS
2021

RESUMO

SILVA JÚNIOR, Antônio Carlos, D.Sc., Universidade Federal de Viçosa, setembro de 2021. **Inteligência computacional para estudar a importância de preditores em aveia branca (*Avena sativa* L.).** Orientador: Cosme Damião Cruz.

O objetivo deste trabalho foi estimar a melhor abordagem para predição e estabelecer uma rede de melhor poder preditivo em aveia branca via metodologias baseadas em regressão, inteligência artificial e aprendizado de máquinas. Setenta e oito genótipos de aveia branca foram avaliados nos anos de 2008 e 2009. Foram avaliados sem e com fungicida, estabelecidos modelos de predição em quatro conjuntos experimentais. O delineamento foi em blocos casualizados com três repetições. As características avaliadas foram rendimento de grãos que foram utilizadas como variável de resposta e dez outras como variáveis explicativas. A quantificação da importância de variáveis através da rede Percepton Multicamadas (MLP) pode ser obtida através de (i) algoritmo de GARSON (1991) modificado por GOH (1995), que consiste no particionamento dos pesos de conexão de rede neural para determinar a importância relativa de cada variável de entrada na rede. (ii) Avaliação da importância de variáveis (entrada) através do impacto da desestruturação ou perturbação da informação de uma determinada entrada sobre a estimativa do R^2. Essa importância foi estimada trocando informações ou tornando o valor fenotípico de cada característica constante e verificando as mudanças nas estimativas de R^2. Quando os valores de uma característica são perturbados, o valor de R^2 diminui, indicando que a característica é importante em relação às outras para fins de predição. A importância de variáveis utilizando a rede função de base radial foi estimado conforme a MLP. Para aprendizado de máquina foram usadas árvores de decisão, bagging, floresta aleatória e boosting. A qualidade do modelo preditivo foi ajustada determinado com base em R^2, e o MSE foi usado para quantificar a importância da característica fenotípica. A inteligência computacional e aprendizado de máquina mostrou-se eficiente e permitiu determinar a importância relativa de variáveis preditoras em aveia branca. Os caracteres indicados para auxiliar na tomada de decisão são estatura de planta, severidade de ferrugem da folha e percentual de acamamento para este estudo. Os R^2 variaram de 30,14% - 96,45% e 10,57% - 94,61%, para inteligência computacional e aprendizado de máquina, respectivamente. A técnica *bagging* apresentou estimativa elevada do coeficiente de determinação superior as demais.

Palavras-chave: *Avena sativa* L; regressão múltipla; Inteligência computacional; aprendizado de máquina.

ABSTRACT

SILVA JÚNIOR, Antônio Carlos, D.Sc., Universidade Federal de Viçosa, September 2021. **Computational intelligence to study the importance of predictors in white oat (*Avena sativa* L.).** Advisor: Cosme Damião Cruz.

The objective of this work was to estimate the best prediction approach and establish a network with better predictive power in white oat using methodologies based on regression, artificial intelligence and machine learning. Seventy-eight white oat genotypes were evaluated in 2008 and 2009. Prediction models were established in four experimental sets without and with fungicide. The design was in randomized blocks with three replications. The characteristics evaluated were grain yield, which were used as a response variable and ten others as explanatory variables. The quantification of the importance of variables through the Multilayer Percepton Network (MLP) can be obtained through (i) GARSON's (1991) algorithm modified by GOH (1995), which consists in the partitioning of the neural network connection weights to determine the importance relative of each input variable in the network. (ii) Evaluation of the importance of variables (input) through the impact of destructuring or disturbing the information of a given input on the estimation of R^2. This importance was estimated by exchanging information or making the phenotypic value of each characteristic constant and checking for changes in the estimates of R^2. When the values of a feature are disturbed, the value of R^2 decreases, indicating that the feature is important over the others for prediction purposes. The importance of variables using the radial basis function network was estimated according to the MLP. For machine learning, decision trees, bagging, random forest and boosting were used. The quality of the predictive model was determined based on R^2, and the MSE was used to quantify the importance of the phenotypic trait. Computational intelligence and machine learning proved to be efficient and allowed to determine the relative importance of predictor variables in white oat. The characters indicated to assist in decision making are plant height, leaf rust severity and lodging percentage for this study. The R^2 ranged from 30.14% - 96.45% and 10.57% - 94.61%, for computational intelligence and machine learning, respectively. The bagging technique showed a high estimate of the coefficient of determination higher than the others.

Keywords: *Avena sativa* L.); multiple regression; computational intelligence; machine learning

1. INTRODUÇÃO

A aveia branca (*Avena sativa* L.) tem grande importância agrícola em todo o mundo. O Brasil é o quinto maior produtor mundial e apresentou aumento substancial na área cultivada com aveia branca nos últimos 10 anos (CONAB, 2021). Esta cultura pode ser utilizada para a produção de grãos, forragem e palha em sistema de plantio direto.

Estimar a importância de variáveis preditoras em programas de melhoramento genético permite obter progresso mais rápido, selecionar e predizer características que apresentam baixa herdabilidade e/ou dificuldade de medição (SILVA JÚNIOR et al., 2021 a, b). Embora, avaliação simultânea de características forneça ampla variedade de informações, identificar qual variável preditora é mais importante é um desafio para o melhorista (PARMLEY et al., 2019). A estimação da importância de variáveis pode ser realizada pelas redes neurais artificiais (RNAs) por meio de algoritmos como de GOH (1995) que propôs uma modificação no algoritmo de GARSON (1991) que consiste no particionamento dos pesos de conexão de rede neural para determinar a importância relativa de cada variável de entrada na rede.

As metodologias baseadas em regressão, inteligência artificial e aprendizado de máquinas têm sido utilizadas com sucesso em estudo de predição. PARMLEY et al., (2019) avaliaram as características fenotípicas de alta dimensão em soja através da abordagem de aprendizado de máquina para predição de rendimento de sementes quanto ao desenvolvimento prescritivo de cultivares para práticas de agrícolas. SKAWSANG et al., (2019) aplicaram tais metodologias para prever a população de pragas de insetos usando fatores climáticos e fenológicos da planta hospedeira. SILVA JÚNIOR et al., (2021 a) utilizaram essas metodologias para prever rendimento de grãos, relação entre comprimento e largura de grãos e comprimento da panícula em arroz irrigado por inundação. SILVA JÚNIOR et al. (2021 b) avaliaram a importância de características auxiliares de uma característica principal com base em informações fenotípicas e estrutura genética previamente conhecida usando inteligência computacional e aprendizado de máquina para desenvolver ferramentas preditivas úteis em programas de melhoramento genético. Entretanto, na literatura não há estudos relacionados à predição de produtividade e verificação da importância de variáveis para a rendimento de grãos na cultura da aveia branca.

Diante do exposto, esse trabalho tem por objetivos: (1) predição de rendimento de grãos em aveia branca via metodologias baseadas em regressão, inteligência artificial e aprendizado de máquinas; (2) identificar preditores mais relevantes, considerando diferentes abordagens de predição em aveia branca.

2. MATERIAL E MÉTODOS

1- Descrição do experimento

O experimento de campo foi realizado na área experimental do Instituto Regional de Desenvolvimento Rural (IRDeR) da Universidade Regional do Noroeste do Estado do Rio Grande do Sul (UNIJUÍ) localizada no município de Augusto Pestana - RS, em coordenadas 28 ° 26 '30' 'S e 54 ° 00' 58 " W, altitude 280 m. Setenta e oito genótipos de aveia branca foram avaliados nos anos de 2008 e 2009. Em cada ano, foram avaliados sem e com fungicida, de forma que foram estabelecidos modelos de predição em quatro conjuntos experimentais (E1, E2, E3 e E4). O fungicida utilizado foi Orkestra, princípio ativo do grupo das piraclostrobina (333 g l^{-1}). O delineamento foi em blocos casualizados com três repetições. O solo é classificado como Latossolo Vermelho distroférrico típico. De acordo com a caracterização climática de Köeppen, o clima da região é do tipo Cfa (subtropical úmido), com quatro estações distintas. A temperatura média anual é de 19,9 ° C e a precipitação média anual de 1774 mm.

A características foi rendimento de grãos (RG, em Kg ha^{-1}) que foram utilizadas como variável resposta e as demais como variáveis explicativas (*inputs*), ou seja, massa de mil grãos (MMG, em gramas); peso hectolitro (PH, em Kg ha^{-1}); dias entre a emergência a maturação (DEM, em dia); percentual de acamamento (ACAM, em porcentagem, onde 1% acamou pouco e 100% acamou totalmente); dias da emergência a floração (DEF, em dia); dias de floração a maturação (DFM, em dia); estatura da planta (EST, em cm); severidade de ferrugem da folha (FFO); severidade de ferrugem do colmo (FCO); manchas foliares (MF). Elas foram usadas para compor redes neurais artificiais de genótipos de aveia branca.

2. Procedimento biométrico

2.1 Metodologias para predição e verificação de importância de características

Regressão Múltipla

A regressão múltipla de *Stepwise* é o método de seleção de variáveis, que visa explicar a relação entre conjunto de variáveis independentes e em relação a uma variável dependente. O coeficiente de determinação (R^2) visa estimar o quanto da variável independente é explicada pela variação total da variável dependente (SILVA JÚNIOR et al., 2021 a, b).

2.2 Inteligência computacional para importância de variáveis

a- *Perceptron Multicamadas- PMC*

A importância de preditores por meio da rede PMC foi quantificada pelo emprego de duas técnicas. A primeira, fundamentada no algoritmo de GARSON (1991) modificado por GOH (1995), que consiste no particionamento dos pesos de conexão de rede neural para determinar a importância relativa de cada variável de entrada dentro da rede (SILVA JÚNIOR et al. 2021 a, b).

A equação da importância relativa de variáveis é igual a

$$IR = WV \tag{1}$$

Matricialmente, tem-se:

$$IR = \begin{pmatrix} IR_1 \\ IR_2 \\ \vdots \\ IR_k \end{pmatrix} = \left(W^1_{N_1 E}\right)' \left(W^2_{N_2 N_1}\right)' \ldots \left(W^c_{N_{c-1 y}}\right)',$$

Em que, W_x^c representa a matriz de pesos do neurônio da camada c, considerando N_j neurônios e N_{j-1} entradas; E é o primeiro neurônio que inicia de entradas; y refere-se a camada de saída desejada e IR: importância relativa da variável.

A importância de variáveis (entradas) também pode ser obtida, após a rede ser estabelecida, considerando o impacto da desestruturação ou perturbação da informação de uma determinada entrada sobre a estimativa do coeficiente de determinação (SILVA JÚNIOR et al., 2021 a, b).

A importância relativa da variável pela permutação do R^2 é descrita na equação a seguir:

$$VR_{x_i} = R^2_{obs} - \bar{R}^2_{perm,x_i} \tag{2}$$

Em que, R^2_{obs} é o R^2 do modelo RNA ajustado às variáveis preditoras e resposta observadas; R^2_{perm,x_i} é o R^2 do modelo RNA ajustado ao conjunto de dados modificado onde x_i é permutado; $\bar{R}^2_{perm,x_i}$: é o valor médio de R^2_{perm,x_i} após m-ésima permutação do conjuntos de dados.

Após alguns critérios utilizados sobre a melhor topologia, adotou-se as seguintes estruturas de redes PMC: (a) toplogia 1: 10-11-1: dez entrada, 11 neurônios ocultos na camada intermediária e um neurônio na camada de saída; (b) topologia 2: 10-11-11-1: dez entrada e

duas camadas ocultas com 11 neurônios nas camadas intermediárias e um neurônio na camada de saída; (c) topologia 3: 10-11-11-11-1: dez entrada, e três camadas ocultas com 11 neurônios nas camadas intermediárias e um neurônio na camada de saída; (d) topologia 4: 10-3-4-11-1: dez entrada, e três camadas ocultas com 3, 4 e 11 neurônios nas camadas intermediárias e um neurônio na camada de saída.

b- Rede Função de Base Radial – RBF

A eficiência da predição é medida pelo coeficiente de determinação e a importância relativa de cada entrada estimada pela técnica de desestruturação da informação de cada variável explicativa, conforme já descrito para PMC.

2.3. Aprendizado de Máquinas para importância de variáveis

Para quantificar a importância de variáveis através abordagem de aprendizado de máquinas foram utilizadas a árvore de decisão e os seus refinamentos, *random forest*, *bagging* e *boosting* (SILVA JÚNIOR et al., 2021 a, b).

A importância da variável IV é descrito na Equação a seguir:

$$IV_{x_i} = MSE_{perm,x_i} - MSE_{nperm} \quad (3)$$

Em que, MSE_{perm,x_i} é a permutação dos valores de cada variável do conjunto de dados onde x_i é permutado; MSE_{nperm}: valores da estimativa dos dados originais não-permutados da variável.

2.3. Importância de variáveis, em modelos reduzidos, na predição de rendimento de grão

Foi considerada a técnica biométrica que conduziu aos melhores resultados de predição de RG e as informações relativas à importância dos preditores.

2.4. Conjuntos de treinamento e validação

O conjunto de treinamento incluiu os mesmos indivíduos para modelagem utilizando todas as metodologias e foi composto por 67% dos indivíduos, que corresponde a 2/3 dos indivíduos selecionados aleatoriamente. Os 33 % (1/3) restantes dos indivíduos constituíram o

conjunto de validação. Em estudos anteriores, 60% a 90% dos indivíduos constituíam o conjunto de treinamento (GIANOLA et al., 2011; GONZÁLEZ-CAMACHO et al., 2012).

3. RESULTADOS E DISCUSSÃO

Predição do rendimento de grãos por diferentes abordagens

A estimativa do coeficiente de determinação, para todas as metodologias, utilizando as dez características agronômicas explicativas na predição de rendimento de grãos (RG) em aveia branca, encontra-se na Tabela 1.

Tabela 1. Média da estimativa máxima do coeficiente de determinação para conjunto de treinamento, em quatro ambientes que corresponde à conjunto de dados de experimentos sem e com fungicida em dois anos agrícolas, para predizer o rendimento de grãos em aveia branca (*Avena sativa* L.).

Abordagem	Técnica	E1	E2	E3	E4
AM	BO	92.29	86.69	81.23	79.23
	DT	85.37	76.39	61.78	64.65
	BA	**94.61**	**93.89**	**92.70**	**92.98**
	RF	64.91	55.09	10.57	24.48
IA	PMC-1	73.25	71.42	30.14	59.84
	PMC-2	**96.45**	**90.12**	56.72	57.94
	PMC-3	86.13	88.58	61.45	68.62
	PMC-4	75.16	85.32	**87.34**	58.77
	RBF	90.12	73.76	80.72	**76.44**
Convencional	RM	61.02	46.07	20.67	32.72

IA: Inteligência Artificiais; AM: Aprendizados de Máquinas; RM: Regressão Múltipla; PMC: Perceptron Multicamadas; PMC: Multilayer Perceptron; PMC-1: Multilayer Perceptron com (10-11-1); PMC-2: Multilayer Perceptron (10-11-11-1); PMC-3: Multilayer Perceptron (10-11-11-11-1); PMC-4: Multilayer Perceptron (10-3-4-11-1); RBR: Rede de Base Radial; DT: Árvore de Decisão; RF: Floresta Aleatória; BA: Bagging; BO: boosting. E: ambientes. E1 e E3: sem fungicida; E2 e E4: com fungicida.

Com base na Tabela 1, pode-se comparar a abordagem que se mostra mais eficientes para a predição da RG. Valores maiores de R^2 indicam que a variável alvo de predição tem melhor ajuste considerando as dez variáveis explicativas utilizadas como preditoras nesta análise (HASSANZADEH et al., 2015; SILVA JÚNIOR et al., 2021 a, b). Dentre as

metodologias utilizadas neste estudo, constatou-se que a regressão múltipla apresentou menor estimativa de R^2, indicando a existência de associações não-lineares entre as variáveis explicativas não consideradas no modelo. As metodologias de inteligência artificiais e aprendizado de máquinas, por sua vez, destacaram-se pela capacidade que possuem de extrair informações não-lineares a partir de entradas do modelo (PARMLEY et al., 2019; SKAWSANG et al., 2019), tal como constatado na Tabela 1. Outros autores já destacaram as habilidades das redes neurais (SILVA et al., 2014; SANT'ANNA et al., 2016) e aprendizado de máquina (SOUSA et al., 2020; SILVA JUNIOR et al., 2021 a, b) em captar melhor relações não-lineares quando comparadas com metodologias convencionais.

Os resultados obtidos por diferentes abordagens mostram que houve discrepância entre a estimativa máxima de R^2 para a variável preditiva nos mesmos ambientes (Tabela 1). Essa discrepância na estimativa de R^2 também foram reportados por SILVA JUNIOR et al. (2021 a, b). Ressalta-se que as diferenças de resultados obtidas nestas análises são indicativas que o ambiente influência na estimativa de R^2 e, consequentemente, na escolha do melhor modelo de predição da variável resposta.

A abordagem de aprendizado de máquina mostrou ser mais eficiente em relação às demais abordagens (Tabela 1). Houve baixa estimativa de R^2 máximo no procedimento *random forest,* para todos os ambientes. Por outro lado, esse procedimento foi superior à abordagem de regressão múltipla para o mesmo ambiente, com exceção do ambiente sem fungicida (E3), o que corresponde a 10.57 %. A baixa estimativa de R^2 máximo no procedimento *random forest* também foram demonstrados em arroz irrigado por inundação (SILVA JUNIOR et al., 2021 a) e em dados simulados com diferente herdabilidade (SILVA JUNIOR et al., 2021 b). Esse procedimento envolve as etapas de reamostras aleatoriamente do conjunto de variáveis explicativas, e construir várias árvores de decisões que constituirão uma floresta aleatória que permitirá a predição e a estimação de pontuações que conduzirão a avaliação da importância dos preditores em processo repetido várias vezes.

Em relação aos ambientes e ao procedimento *bagging*, verifica-se que as estimativas de R^2 foram superiores a 92.70 % tornando esta abordagem de melhor destaque para uso nos conjuntos de dados analisados. Estimativas altas (tendo como referência valores em torno de 80%) de R^2 também foram obtidas usando metodologias de aprendizado de máquina pelos procedimentos *boosting,* além do *bagging,* para todos os conjuntos de dados de predição (Tabela 1). SILVA JUNIOR et al. (2021 a, b) mostraram que as abordagens de aprendizado de máquinas para os procedimentos *bagging* e *boosting* foram mais consistem em obter maior

estimativa média geral de R^2, em relação variáveis preditivas. A metodologia da árvore de decisão (DT) e *random forest* não se destacaram em relação aos demais procedimentos de aprendizado de máquina (Tabela 1).

As abordagens de inteligência artificiais fundamentadas em RBF proporcionaram ajustes cujo R^2 foram superiores a 70 % em todos os ambientes (Tabela 1). Neste procedimento, a maior estimativa R^2 máxima foi 90.12% (± 5.79) e a menor 73.75% (± 1.67), o que corresponde aos ambientes E1 e E2, respectivamente. SILVA JUNIOR et al. (2021 a) encontraram estimativa de R^2 máxima que variou de 48 % a 99 % em diferentes ambientes para a cultura de arroz irrigado por inundação. Para dados simulados com diferente estrutura genética, a estimativa máxima de R^2 variando de 44 % a 54% (SILVA JUNIOR et al., 2021 b) e SANT'ANNA et al. (2020) obtiveram resultados de R^2 consistente para diferente estrutura genética. ROSADO et al. (2020) avaliaram cultivares de feijão, e obtiveram estimativa de R^2 para a características dias para primeira flor e dias de floração de 94.10% e 94.40%, respectivamente. Esse procedimento apresenta boa capacidade de lidar com interações complexas em comparação com regressões semiparamétricas e lineares (SANT'ANNA et al., 2019, 2020). Geralmente, a RBF tem rápida capacidade de aprender com os dados utilizados como informações de treinamentos e fornecem uma solução única em comparação com as RNAs do tipo perceptron (GONZALES-CAMACHO et al., 2012; SANT'ANNA et al., 2019, 2020).

As redes de função de base radial têm boa capacidade de lidar com interações em comparação com regressões semiparamétricas e lineares (SANT'ANNA et al., 2020). SANT'ANNA et al. (2020) aplicou a RBF em estudos usando características simuladas com 30% e 60% de herdabilidade para seleção de variáveis. Os autores identificaram maior eficiência na seleção pelo RBF quando o cenário envolvia interações epistáticas no controle gênico dos caracteres estudados. GONZÁLEZ-CAMACHO et al. (2012) observaram que é possível melhorar a predição em modelos não paramétricos quando a seleção inclui marcadores que não estão diretamente relacionados às características de interesse. SILVA JUNIOR et al., (2021 a) aplicaram RBF para prever rendimento de grãos, relação entre comprimento e largura de grãos e comprimento da panícula em arroz irrigado por inundação. Esses autores argumentam que a RBF apresenta alto desempenho em predizer a importância de variáveis. SILVA JUNIOR et al., (2021 b) avaliaram a importância de características auxiliares de uma característica principal com base em informações fenotípicas e estrutura genética previamente

conhecida usando RBF e demonstraram a eficiência dessa rede para quantificar a importância de variáveis.

Em relação ao procedimento PMC-1 (10-11-1) a maior estimativa de R^2máxima foi observado no E1- 73.25% e o menor em E3, com estimativa de 30.14%, ambos os ambientes corresponde à sem fungicida. O procedimento PMC-2 (10-11-11-1) e PMC-3 (10-11-11-11-1) as maiores estimativas foram observadas no E1 e E2 e as menores em E3 e E4, respectivamente. Para o mesmo número de camada oculta que corresponde a PMC-3 (10-11-11-11-1) e PMC-4 (10-3-4-11-1). Observamos menores estimativas de R^2 máxima para o procedimento PMC-4, com exceção ao ambiente E3. Isso mostra que o número de neurônio na camada influência na estimativa de R^2 máxima. SILVA JUNIOR et al. (2021 b) argumentaram que o número de neurônio influência na estimativa do coeficiente de determinação.

A rede PMC é amplamente utilizada no processo preditivo (SANTOS et al., 2018; SILVA JUNIOR et al., 2021 a, b), uma vez que o sucesso desta rede já foi demonstrado em vários grupos de pesquisa que mostraram matematicamente que, com apenas uma única camada oculta, essa rede funciona muito bem com diferentes números de neurônios na camada oculta (DEOÑA et al., 2014; SANTOS et al., 2018).

Assim, o aprendizado de máquina é de fato mais eficiente para a seleção de características fenotípicas porque pode lidar com informações reduzidas ou redundantes sobre características fenotípicas (SILVA JUNIOR et al., 2021 b). COSTA et al. (2020) avaliaram a importância das variáveis por *bagging*, *random forest*, *boosting*, árvore de decisão, PML e RBF e relatou que PML e RBF alcançaram melhores resultados. SILVA JUNIOR et al. (2021 a, b) verificaram que as metodologias de inteligência computacional e aprendizado de máquina na predição permitiram identificar as características fenotípica explicativas que deveriam ser priorizados e estabelecidos como características auxiliares para a seleção indireta.

A eficiência das RNAs em problemas de predição dada sua capacidade de extrair informações relevantes de grandes conjuntos de dados (CHAGAS et al., 2013) e generalizar informações relativamente imprecisas (PORWAL et al., 2003), ficou muito bem expressa pelos resultados obtidos (Tabela 1). O mesmo pode ser constatado para as metodologias baseadas em aprendizado de máquinas, sendo estas capazes de lidar com informações mais reduzidas ou redundantes nas variáveis de entrada (SILVA JUNIOR et al. 2021 a, b). No entanto, outro estudo tão importante quanto a predição e que muitas vezes não é realizado, é a identificação de variáveis preditoras mais importantes, que constitui fator importante no processo de tomada de decisão (BEUCHER et al., 2019). Desse modo, após as análises de predição, procedeu-se às

análises de quantificação da importância de variáveis por meio dos métodos de inteligência artificiais e aprendizado de máquinas, a fim de identificar, entre o conjunto de variáveis explicativas, aqueles que devem ser priorizadas e identificadas como características auxiliares em respostas indiretas à seleção.

Relação linear entre variáveis preditoras e de rendimento de grãos em aveia branca

As maiores associações lineares com RG, pode ser indicativo preliminar de que as variáveis, individualmente, são importantes na predição de RG. Em modelos de predição multivariados uma variável preditora, de alta correlação com a variável resposta, pode perder sua importância em razão de sua redundância tendo em vista que, no modelo, possa estar representada por outra associada. Assim, além de quantificar as relações lineares entre preditor-resposta, é importante quantificar e apreciar as relações lineares, expressas pelos coeficientes de correlação linear, entre todos os preditores na busca de redundâncias. Neste trabalho, estas associações foram representadas em uma rede de correlação que contém linhas vermelhas e verdes que representam correlações negativas e positivas, respectivamente e a largura das mesmas é proporcional à magnitude das correlações (Figura 1). Em relação à rede de correlação fenotípica observaram que estrutura da grupos correlacionados visando a predição de RG. Nesta rede, é destacado a semelhança entre as características fenotípicas e os padrões de correlação fenotípicas.

As características que apresentam grupos com RG em E1 foram MMG, PH e EST que correlaciona positivamente, porém variando em magnitude, e a correlacionada negativamente foi FFO. Em relação ao E2, as características correlacionadas positivamente consistem em: EST e MMG; e as negativamente MF e DFM. Para o E3, que representa sem fungicida, a característica que apresenta correlacionado negativamente foi FCO. O ambiente 4, o grupo correlacionado positivo consiste em PH e DEF e a negativa DEM (Figura 1).

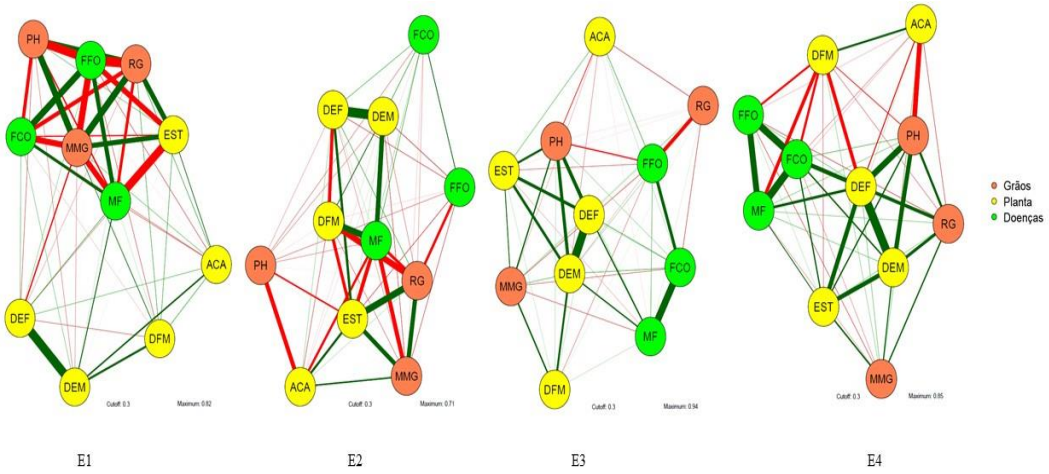


Figura 1. Rede de correlação fenotípica para os três grupos distintos em quatro ambientes que correspondem à sem e com fungicida em dois anos agrícolas, para predizer a rendimento de grãos em aveia branca (*Avena sativa* L.). A largura da linha é proporcional à força da correlação. E1 e E3; E2 e E4 representam os ambientes sem e com fungicida, respectivamente. A cor laranja representa as características de grãos; A cor amarela representa as características de planta e a verde severidade de doenças. MMG = Massa de Mil Grãos em gramas; PH = Peso Hectolitro; DEM = Dias entre a Emergência a Maturação; ACA= Percentual de Acamamento; RG = Rendimento de grãos em quilos por hectare; DEF = Dias da Emergência a Floração; DFM= Dias de Floração a Maturação; EST= Estatura da Planta; FFO= Severidade de Ferrugem da Folha; FCO= Severidade de Ferrugem do Colmo e MF= Manchas Foliares.

Importância de variáveis na predição por abordagem de Inteligência Artificial

Perceptron multicamadas (PMC)

A estimativas do coeficiente de determinação da predição de rendimento de grãos por PMC atribuindo perturbação à informação genotípica, encontra-se na Tabela 2. Esses resultados mostram grandes discrepâncias no R^{2*} na comparação dos ambientes entre si, o que dificulta a interpretação. Nos ambientes E1 e E4, que correspondem aos ambientes sem fungicida, as características ACA, EST, FFO foram a eficiente em quantificar a variável resposta RG devido à redução na estimativa de R^{2*} em função da estratégia de atribuir perturbação à informação fenotípica.

Independentemente do número de neurônio na camada de saída e uma única camada oculta, eles concordaram em apontar as variáveis mais importantes para predizer RG. Esse resultado mostra que essas variáveis são importantes na predição de RG, pois a perturbação em

seus valores levou a uma redução considerável na qualidade do ajuste. No ambiente E2, a característica MMG foi a mais importante em predizer RG.

Houve discrepância no número de neurônio na camada de saída e camada oculta, em apontar que as variáveis mais importantes em E4, em que corresponde ao ambiente com fungicida. Com apenas um neurônio na camada de saída e uma única camada oculta mostraram que DEF e FCO foram a mais importante devido à redução na estimativa de R^{2*}. Com dois neurônios na camada intermediária e uma única camada oculta demostraram que FFO e MF para a variável alvo de predição. Quando utilizamos um neurônio na camada de entrada, e três camadas ocultas com 11 neurônios na camada intermediária e um neurônio na camada de saída as características que demonstram ser a mais importante foram PH e FCO. Por outro lado, com três camadas ocultas com três, quatro e 11 neurônios na camada intermediária as características importantes em predizer a RG foram: FFO, DEF e PH. SILVA JUNIOR et al. (2021 a) reportaram que com apenas um neurônio na camada de saída e uma única camada oculta, eles concordaram em apontar que as variáveis mais importantes eram a largura e o comprimento dos grãos em arroz irrigado, dadas as quedas significativas nos valores da estimativa de R^{2*} observado quando perturbamos as variáveis.

Tabela 2. Estimativas do coeficiente de determinação da predição de rendimento de grãos em aveia branca (*Avena sativa* L.), utilizando a PMC atribuindo perturbação à informação genotípica.

	E1				E2			
Input	TOP1	TOP2	TOP3	TOP4	TOP1	TOP2	TOP3	TOP4
MMG	70.02	87.37	64.93	74.93	**31.92**	**23.19**	**9.73**	**26.47**
PH	71.78	78.42	70.44	72.44	54.25	87.37	86.02	84.30
DEF	76.51	76.36	74.89	64.89	54.68	65.92	48.33	75.16
DFM	75.18	86.87	68.59	78.59	43.67	36.65	70.15	50.05
DEM	76.54	77.17	83.87	73.87	56.49	74.88	75.94	77.60
EST	**61.01**	80.26	**49.89**	**59.89**	53.23	63.91	33.01	55.37
ACA	75.26	**66.07**	**62.90**	**67.90**	46.46	71.41	76.46	68.43
FFO	**52.80**	**33.62**	**10.33**	**8.33**	52.72	73.18	85.34	67.20
FCO	76.59	78.03	71.10	71.10	57.33	80.89	58.86	60.60
MF	75.19	80.32	81.71	71.81	56.85	76.40	74.77	72.44
	E3				E4			
Input	TOP1	TOP2	TOP3	TOP4	TOP1	TOP2	TOP3	TOP4
MMG	32.34	33.29	52.69	65.34	51.85	38.73	50.93	58.67
PH	**21.20**	**12.64**	**26.31**	**42.81**	47.09	52.58	**26.67**	**37.82**
DEF	30.70	54.58	73.53	57.07	**37.84**	45.99	42.69	**34.25**
DFM	30.93	33.23	36.08	50.12	55.95	52.81	56.06	53.65
DEM	32.58	48.50	79.57	68.96	40.57	46.46	31.72	50.97
EST	29.51	35.36	57.87	44.98	50.74	53.91	55.85	56.06
ACA	**18.57**	39.66	**29.95**	51.51	59.52	48.74	57.27	59.37
FFO	**4.48**	**11.46**	**24.87**	**21.62**	44.69	**29.64**	45.38	**39.15**
FCO	24.65	**19.57**	38.52	**9.99**	**39.55**	42.86	**31.70**	40.07
MF	26.36	12.91	49.10	45.01	56.54	**37.20**	45.79	59.26

MMG = Massa de Mil Grãos em gramas; PH = Peso Hectolitro; DEM = Dias entre a Emergência a Maturação; ACA= Percentual de Acamamento; RG = Rendimento de grãos em quilos por hectare; DEF = Dias da Emergência a Floração; DFM= Dias de Floração a Maturação; EST= Estatura da Planta; FFO= Severidade de Ferrugem da Folha; FCO= Severidade de Ferrugem do Colmo e MF= Manchas Foliares; E: ambientes. E1 e E3: sem fungicida; E2 e E4: com fungicida. Topologia- TOP1: Multilayer Perceptron com (10-11-1); TOP2: Multilayer Perceptron (10-11-11-1); TOP3: Multilayer Perceptron (10-11-11-11-1); TOP4: Multilayer Perceptron (10-3-4-11-1); E: ambientes. E1 e E3: sem fungicida; E2 e E4: com fungicida.

A importância das variáveis foi quantificada atribuindo desestruturar à informação genotípica referente a cada variável, de modo a observar-se quais mudanças ocorreriam nos valores do R^2. É importante salientar que, nesta Tabela, reduções nos valores de R^2 após

atribuir desestruturação à informação genotípica referente a cada variável, são indicativos de que esta variável é importante em relação as demais para fins de predição com a rede já estabelecida.

Rede Base Radial (RBF)

A estimação da importância dos caracteres em aveia branca atribuindo perturbação à informação de uma variável de entrada após a RBF ter sido estabelecida e está descrita na Tabela 3. Nesta Tabela, a importância relativa de cada entrada estimada pela técnica de desestruturação da informação de cada variável explicativa. Ao usar esta estratégia, reduções drásticas nos valores de R^{2*} foram observadas para as variáveis mais importantes é FFO para a variável preditiva RG, nos ambientes E1 e E4. Na prática, a intensidade dessa característica reduz progresso genético para o aumento de rendimento de grãos. No ambiente E2, a variável que sofreram a maior redução em R^{2*} foi DMF, com estimativa de 44,47%. Essa característica aumenta o rendimento dos grãos, uma vez que mais fotoassimilados são produzidos e translocados para os grãos. No entanto, cultivares de ciclo tardio tendem a ser mais produtivas em relação ao ciclo inicial, pois obtêm aumento na quantidade de fotoassimilados que são translocados para os grãos (SILVA JUNIOR et al., 2021 a).

Os resultados mostram que a variável mais importante usando a RBF foi MMG, nos ambientes E2, E3 e E4, com estimativa de 58.97%, 47,98% e 40,97%, respectivamente. Na prática, MMG influência no rendimento de grãos em aveia branca, uma vez que quanto maior MMG, consequentemente, maior RG. Isso justifica os resultados desse estudo em aveia branca na predição de RG.

Tabela 3. Estimativas do coeficiente para determinar a predição de rendimento de grãos, em aveia branca (*Avena sativa* L.), usando a RBF atribuindo perturbação à informação genotípica.

Input	E1	E2	E3	E4
MMG	81.04	58.97	**47.98**	40.77
PH	**76.70**	60.73	65.01	53.99
DEF	85.43	68.16	72.11	**47.52**
DFM	84.30	44.37	**52.47**	65.02
DEM	80.99	68.16	74.24	54.61
EST	**73.97**	59.36	**62.75**	72.19
ACA	81.96	68.07	64.71	64.71
FFO	**60.1**3	63.30	71.59	**45.04**
FCO	84.38	70.50	69.10	63.74
MF	88.37	61.23	**54.09**	52.25

MMG = Massa de Mil Grãos em gramas; PH = Peso Hectolitro; DEM = Dias entre a Emergência a Maturação; ACA= Percentual de Acamamento; RG = Rendimento de grãos em quilos por hectare; DEF = Dias da Emergência a Floração; DFM= Dias de Floração a Maturação; EST= Estatura da Planta; FFO= Severidade de Ferrugem da Folha; FCO= Severidade de Ferrugem do Colmo e MF= Manchas Foliares; E: ambientes. E1 e E3: sem fungicida; E2 e E4: com fungicida.

Os resultados obtidos corroboram a expectativa sobre o RBF em quantificar e revelar a importância das características utilizando a estratégia de causar distúrbios a partir das permutações ou fixação dos valores fenotípicos das variáveis de entrada. Nosso estudo demonstra a capacidade do RNA de quantificar a importância das características fenotípicas em aveia branca. Foram apresentadas técnicas que mostram o impacto da interrupção ou perturbação nas informações de uma determinada entrada na estimativa do coeficiente de determinação e partição dos pesos de conexão da RNA. Essas técnicas foram eficazes em estimar a verdadeira importância das características fenotípicas. Portanto, há uma certa concordância entre os resultados encontrados pelas duas metodologias de inteligência computacional de redes PMC e redes RBF.

Importância de variáveis na predição por abordagem Aprendizado de Máquinas

A Tabela 4 apresenta as médias das contribuições relativas das variáveis explicativas para predição de rendimento de grãos por meio da estimativa da porcentagem de incremento de erro quadrático mínimos (IMSE), que é construído permutando os valores de cada variável do conjunto de dados, e comparando com a predição do conjunto de dados originais não-

permutados da variável. Nesse caso, diferentemente da estratégia utilizada para as metodologias de inteligência computacional de redes PMC e RBF, para as quais valores inferiores de R^2 indicavam maior importância daquela variável para o modelo, na abordagem de aprendizado de máquinas a importância da variável explicativa está relacionada com a estimativa da diminuição média da precisão do modelo por meio do IMSE de modo que, quanto maior esta estimativa maior é a importância da variável.

Tabela 4. Estimativa médias das contribuições relativa das variáveis explicativa para predição de rendimento de grãos em aveia branca utilizando abordagem de aprendizado de máquinas, em quatro ambientes que corresponde a sem e com aplicação de fungicida.

	E1			E2			E3			E4		
VA	BA	RF	BO	BA	RF	BO	BA	RF	BO	BA	RF	BO
MMG	**7.58**	**7.94**	**12.37**	**10.47**	**9.84**	**10.89**	1.04	1.53	**4.49**	3.55	3.21	3.75
PH	**10.11**	**10.68**	**15.29**	2.19	2.23	6.57	2.2	1.75	3.51	3.83	4.44	3.93
DEF	3.29	2.42	7.55	6.85	5.79	6.73	**3.58**	**3.5**	**4.60**	**11.46**	**11.49**	**9.18**
DFM	1.59	2.21	2.97	**16.94**	**16.84**	**12.25**	0.8	-0.4	4.22	**5.57**	**4.68**	**4.82**
DEM	3.46	3.1	6.35	6.44	6.06	6.28	2.14	1.86	3.43	**6.12**	**5.95**	**5.17**
EST	**10.74**	**10.3**	**9.65**	**10.01**	**8.29**	**9.32**	-0.93	-0.24	2.72	0.8	-0.45	1.02
ACA	2.83	2.49	5.94	1.36	1.08	2.79	3.04	3.04	2.96	0.36	-0.66	0.89
FFO	**20.87**	**20.1**	**29.59**	**9.27**	**9.29**	**16.05**	**9.91**	**10.91**	**12.29**	**4.19**	**4.58**	**7.02**
FCO	7.32	7.76	5.60	3.09	2.25	3.65	**3.52**	**3.97**	**3.71**	0.8	1.62	2.04
MF	3.11	3.67	4.69	3.62	2.91	3.74	3.22	2.95	3.30	3.99	3.49	3.59

MMG = Massa de Mil Grãos em gramas; PH = Peso Hectolitro; DEM = Dias entre a Emergência a Maturação; ACA= Percentual de Acamamento; RG = Rendimento de grãos em quilos por hectare; DEF = Dias da Emergência a Floração; DFM= Dias de Floração a Maturação; EST= Estatura da Planta; FFO= Severidade de Ferrugem da Folha; FCO= Severidade de Ferrugem do Colmo e MF= Manchas Foliares; FA: *random forest*; BA: *Bagging*; BO: *Boosting*; VA: variável auxiliar; E: ambientes. E1 e E3: sem fungicida; E2 e E4: com fungicida.

Com base na Tabela 4, as variáveis que obtiveram maior estimativa de IMSE em todos as metodologias de aprendizado de máquinas em relação aos ambientes sem fungicidas foram: FFO, PH, EST e MMG; DEF, FCO e FFO, E1 e E3, respectivamente. A variável que apresentou ser mais eficiente nestes ambientes foi a FFO. Isso justifica que essa variável pode ser utilizada no processo de seleção indireta quando a variável alvo de predição é RG. Em relação aos ambientes com fungicidas as variáveis mais importantes foram: MMG, DFM, EST e FFO; DEF, DFM, DEM e FFO, que são representados por E2 e E4, respectivamente. Para esse ambiente

com fungicida as variáveis DFM e FFO mostrou ser eficiente em estimar a predição de rendimento de grãos em aveia branca.

As metodologias *random forest* e *bagging* foram coincidentes em quantificar as mesmas variáveis explicativa. Resultado similar foi reportado por SILVA JUNIOR et al. (2021 a, b). Em relação ao procedimento *boosting* os resultados mostram discrepantes. Por outro lado, este procedimento foi mais consistente na predição de variável. Neste procedimento para estimar a importância de variável utilizando a RG como alvo preditivo, as variáveis: MMG, PH, EST e FFO; MMG DEF e FFO se destacaram no ambiente sem fungicidas, representado por E1 e E3, respectivamente. Em relação ao ambiente com fungicida as variáveis importantes foram: MMG, DFM, EST e FFO; DEF, DFM, DEM e FFO, respectivamente. Ao utilizar o procedimento *boosting* a variável que se destacou em todos os ambientes foi a FFO. Isso justifica que essa variável pode ser utilizada para predição de RG em aveia branca.

A técnica *bagging* envolve a geração de várias séries de treinamento distintas a partir de conjunto de dados original. As predições finais são calculadas pela média de todas as predições geradas. Isso é útil para técnicas de árvores de decisão e rede neural artificial que são sensíveis a pequenas mudanças nos dados de treinamento (Song et al., 2021).

Importância de variáveis, em modelos reduzidos, na predição por abordagem Aprendizado de Máquinas

Neste tópico foi considerada a técnica biométrica que conduziu aos melhores resultados de predição do RG e as informações relativas à importância dos preditores, que foi o *bagging*.

A estimativa média das contribuições relativa das variáveis explicativa para predição de rendimento de grãos em aveia branca utilizando a técnica bagging, após eliminação de variáveis auxiliares de menores contribuições relativas, em quatro ambientes que corresponde a sem e com aplicação de fungicida é demonstrado na Tabela 5. A escolha desta técnica (bagging) baseou-se na estimativa do coeficiente de determinação (Tabela 1) que foi superior a 90% e a eliminação variáveis auxiliares da menores contribuições relativas estabelecido pela Tabela 4.

Tabela 5. Estimativa do coeficiente de determinação para conjunto de treinamento, em quatro ambientes que corresponde à conjunto de dados de experimentos sem e com fungicida em dois anos agrícolas, para predizer o rendimento de grãos em aveia branca (*Avena sativa* L.) por meio da técnica *Bagging*.

Preditores	E1	E2	E3	E4
R^2 (v=10)	**94.61**	**93.89**	**92.70**	**92.98**
Excluído	DFM	ACA	EST	ACA
R^2 (v=9)	94.85	94.34	92.83	93.05
Excluídos	DFM, ACA	ACA, PH	EST, DFM	ACA, EST
R^2 (v=8)	94.26	93.50	92.03	93.11
Excluídos	FCO, MF	FCO, MF	FCO, MF	FCO, MF
R^2 (v=8)	94.95	94.40	91.74	92.84

PH = Peso Hectolitro; ACA= Percentual de Acamamento; DEF = Dias da Emergência a Floração; DFM= Dias de Floração a Maturação; EST= Estatura da Planta; FCO= Severidade de Ferrugem do Colmo e MF= Manchas Foliares; E: ambientes. E1 e E3: sem fungicida; E2 e E4: com fungicida; R^2: coeficiente de determinação.

A importância de preditores por meio da eliminação de variáveis auxiliares de menores contribuições relativas foi quantificada pelo emprego de várias maneiras. A primeira, fundamentada na eliminação de apenas uma das variáveis preditoras (DFM, ACA, EST e ACA), em E1, E2, E3 e E4, respectivamente e, depois, de duas variáveis que menos contribuíram. Por fim, optou-se pela eliminação das variáveis FCO e MF que demonstraram menor estimativa da porcentagem de incremento de erro quadrático mínimos em todos os ambientes.

Após eliminação de variáveis auxiliares de menores contribuições relativas, a estimativa máxima do coeficiente de determinação é similar quando utilizamos todas as variáveis auxiliares para predizer RG (Tabelas 1 e 5).

A literatura tem destacado as técnicas de aprendizado de máquinas como ferramentas eficientes na quantificação da importância relativa de variáveis, tendo em vista a simplicidade, o não uso de suposições sobre a distribuição das variáveis explicativas e também devido a sua robustez em relação à quantidade, redundância e influências ambientais (TAN et al., 2014; BEUCHER et al., 2019; SILVA JUNIOR et al., 2021 a, b). Além disso, tais técnicas não requerem um modelo de especificação de herança e podem ser responsáveis por efeitos não aditivos sem aumentar o número de covariáveis no modelo ou tempo de computação (GONZÁLEZ-RECIO et al., 2011). A técnica *bagging* mostra bom desempenho preditivo na prática; ela funciona bem para problemas multidimensionais e pode ser usado com saída de

várias classes, preditores categóricos e problemas desequilibrados (GREGORUTTI et al., 2017). Resultados satisfatórios da seleção de variáveis usando o algoritmo de *bagging* e *random forest* na presença de preditores correlacionados foram relatados por GREGORUTTI et al. (2017). O poder discriminatório, redundância, precisão e complexidade podem influenciar os índices ou estatísticas usadas para quantificar a importância dos traços auxiliares na predição de uma característica principal.

O melhoramento genético para as características desejadas nas diversas safras tem sido um processo demorado, trabalhoso e caro. Os melhoristas estudam gerações de plantas e identificam e modificam as características genéticas desejadas, à medida que avaliam como as características são expressas na prole (PALIWAL & KUMAR 2011; FERREIRA et al., 2015). A aplicação de inteligência computacional e aprendizado de máquina para identificar conjuntos ideais de características observáveis (fenótipos) pode permitir decisões informadas e alcançar resultados de grande relevância em programas de melhoramento. Além disso, essas metodologias podem ajudar a prever características auxiliares com o melhor desempenho sob diferentes práticas de manejo agrícola.

Comparamos diferentes abordagens de seleção ou descarte de variáveis que foram recentemente proposto para identificar variáveis preditivas relevantes dentro de problema regressão. Além disso, nós incluímos em nossa comparação método tradicional que visa encontrar pequeno subconjunto de variáveis importantes com desempenho ideal de predição na cultura de aveia branca. Destaca-se, o fato de que as caracteristicas utilizados neste estudo são trabalhosas de se obter e a sua avaliação pode ser onerosa se existir número maior de genótipos a serem avaliados. Nesse contexto, o estudo das características mais importantes na predição torna-se necessário, uma vez que é possível reduzir o esforço físico, custo, uso de mão-de-obra e tempo na experimentação (PALIWAL & KUMAR, 2011; FERREIRA et al., 2015).

Diante disso, o nosso estudo apresenta o desempenho de algumas metodologias para avaliar contribuições relativa de cada variável através da inteligência computacional e aprendizado de máquinas na cultura de aveia branca. Considera-se, que abordagem para estimar o efeito de variáveis explicativas no melhoramento genético identificou, com sucesso, a verdadeira importância de cada variável, incluindo aquelas que exibem correlações fortes e fracas com as variáveis principais, que no nosso caso é rendimento de grãos.

Metodologias baseadas em aprendizado de máquina e inteligência computacional não dependem de informações estocásticas e tendem a ser mais eficientes, enquanto as metodologias convencionais dependem da normalidade da distribuição das características

fenotípicas. Além disso, as metodologias de aprendizado de máquina e inteligência computacional não fazem suposições sobre o modelo e podem capturar fatores complexos em modelos de predição. No aprendizado de máquina, o conhecimento a priori da previsão não é necessário se os dados produzirem esses efeitos, e nenhuma suposição é feita quanto à distribuição dos valores fenotípicos (SOUSA et al., 2020). Algoritmos de aprendizado de máquina têm a vantagem de modelar dados de maneira não linear e não paramétrica (OSCO et al., 2020). Ao contrário de muitos métodos estatísticos tradicionais, esses algoritmos são construídos com a vantagem de lidar com dados ruidosos, complexos e heterogêneos (OSCO et al., 2019). Os pesquisadores agora têm a capacidade de identificar as contribuições individuais e interativas das variáveis preditoras para a cultura da aveia branca utilizando a inteligência artificiais e aprendizado de máquinas.

4. CONCLUSÃO

As metodologias de inteligência computacional e aprendizado de máquinas conseguiram quantificar a importância de variáveis explicativas na predição do rendimento de grãos em aveia branca. A rede com apenas uma camada oculta foi eficiente para determinar a importância relativa de variáveis em aveia branca.

A técnica *bagging* apresentou estimativa elevada do coeficiente de determinação superior as demais. Modelos mais simples, com exclusão de preditores, se mostram tão eficientes quanto os modelos mais complexos, indicando que quantificação da importância de preditores é importante para minimizar custos, garantindo mesmos níveis de eficiência do modelo preditivo.

5. REFERENCIA

Beucher A, Møller AB, Greve MH. Artificial neural networks and decision tree classification for predicting soil drainage classes in Denmark. Geoderma 2019; 352:351-359. http://dx.doi.org/10.1016/j.geoderma.2017.11.004.

CONAB: Companhia Nacional de Abastecimento, Conab - Página inicial: acesso em 10/06/2021.

Costa WGD, Barbosa IP, de Souza JE, Cruz CD, Nascimento M, de Oliveira ACB. Machine learning and statistics to qualify environments through multi-traits in Coffea arabica.

PLoS One. 2021 Jan 12;16(1):e0245298. http://dx.doi.org/10.1371/journal.pone.0245298.

Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos G, Burgueño J, Camacho-González JM, Pérez-Elizalde S, Beyene Y, Dreisigacker S. Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. 2017;22(11):961-975.

Cruz, CD. Genes Software – extended and integrated with the R, Matlab and Selegen. Acta Scientiarum 2016; 38:547-552. http://dx.doi.org/10.4025/actasciagron.v38i4.32629.

DeOña J, Garrido C. Extracting the contribution of independent variables in neural network models: a new approach to handle instability. Neural Comput Appl. 2014; 25:859-869.

Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics data sets. Brief Bioinform 2019; 20:492-503. http://dx.doi.org/10.1093/bib/bbx124.

Ferreira MG, Azevedo AM, Siman LI, Silva GH, Carneiro CS, Alves FM, Delazari FT, Silva DJH, Nick C. Automation in accession classification of *Brazilian Capsicum* germplasm through artificial neural networks. Scientia Agricola 2017; 74(4). http://dx.doi.org/10.1590/1678-992X-2015-0451.

Garson GD. Interpreting neural network connection weights. Artificial Intelligence Expert. 1991; 6:46-51.

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

Goh ATC. Back-propagation neural networks for modeling complex systems. Artificial Intelligence in Engineering. 1995; 9:143-51. http://dx.doi.org/10.1016/0954-1810(94)00011-S.

González-Camacho JM, Campos G, Pérez P, Gianola D, Cairns JE, Mahuku G, Babu R and Crossa J. Genome-enabled prediction of genetic values using radial basis function neural networks. Theoretical and Applied Genetics. 2012; 125(4): 759-771.

González-Camacho JM, Crossa J, Pérez-Rodríguez P, Ornella L and Gianola D. Genome-enabled prediction using probabilistic neural network classifiers. BMC genomics. 2016; 17(1), p.208.

González-Recio O, Forni S. Prediction across the genome of discrete traits using Bayesian regressions and machine learning. Genet Sel Evol. 2011; 43:7. https://doi.org/10.1186/1297-9686-43-7.

González-Recio O, Rosa GJ and Gianola D. Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. Livestock Science. 2014; 166:217-231.

Gregorutti B, Michel B, Saint-Pierre P. Correlation and variable importance in random forests. Stat Comput 2017; 27:659-678. https://doi.org/10.1007/s11222-016-9646-1.

Hassanzadeh Z, Ghavami R, Kompany-Zareh M. Radial basis function neural networks based on the projection pursuit and principal component analysis approaches: QSAR analysis of fullerene[C60]-based HIV-1 PR inhibitors. Medicinal Chemistry Research. 2015; 25: 19-29. http://dx.doi.org/10.1007/s00044-015-1466-x.

Olivoto T, Souza VQ, Nardino M, Carvalho IR, Ferrari M, Pelegrin AJ, Szareski VJ, Schmidt D. Multicollinearity in path analysis: a simple method to reduce its effects. Agronomy Journal 2017; 109: 131-142.

Osco LP, Ramos APM, Moriya EAS, Bavaresco LG, Lima BC, Estrabis N, Pereira DR, Creste JE, Marcato Junior J, Gonçalves WN et al. Modeling hyperspectral response of water-stress induced lettuce plants using artificial neural networks. Remote Sens. 2019; 11:2797.

Osco LP, Ramos APM, Pinheiro MMF, Moriya EAS, Imai NN, Estrabis N, Lanczyk F, Araujo FF, Liesenberg V, Jorge LAC, Li J, Ma L, Gonçalves WN, Junior JM and Creste JE. A machine learning framework to predict nutrient content in valencia-orange leaf hyperspectral measurement. Remote Sens. 2020; 12:906. http://dx.doi.org/10.3390/rs12060906.

Paliwal M, Kumar UA. Assessing the contribution of variables in feed forward neural network. Applied Soft Computing. 2011; 11:3690-3696.

Parmley KA, Higgins RH, Ganapathysubramanian B et al. 2019. Machine learning approach for prescriptive plant breeding. Sci Rep 9, 17132. http://dx.doi.org/10.1038/s41598-019-53451-4.

Paruelo JM, Tomasel F. "Prediction of functional characteristics of ecosystems: a comparison of artificial neural networks and regression models." Ecological Modelling. 1997; 98: 173–186. http://dx.doi.org/10.1016/s0304-3800(96)01913-8.

Paswan RP and Begum SA. Regression and neural networks models for prediction of crop production. Int. J. Sci. Eng. Res.2013; 4:11.

Pérez-Rodríguez P, Gianola D, González-Camacho JM, Crossa J, Manès Y and Dreisigacker, S. Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3: Genes, Genomes, Genetics, 2012; 2(12): 1595-1605.

Sant'Anna IC, Ferreira RADC, Nascimento M, Carneiro VQ, Silva GN, Cruz CD, Oliveira MS, Chagas FEO. Multigenerational prediction of genetic values using genome-enabled prediction. PLoS ONE. 2019. 14, e0210531. http://dx.doi.org/10.1371/journal.pone.0210531.

Sant'Anna IC, Tomaz RS, Silva GN, Nascimento M, Bhering LL, Cruz CD. Superiority of artificial neural networks for a genetic classification procedure. Genet. Mol. Res. 2015, 14, 9898–9906.

Sant'Anna IC, Tomaz RS, Silva GN, Nascimento M, Bhering LL, Cruz CD. Superiority of artificial neural networks for a genetic classification procedure. Genet. Mol. Res. 2015, 14, 9898–9906.

Sant'Anna IC, Silva GN, Nascimento M, Cruz C D. Subset selection of markers for the genome-enabled prediction of genetic values using radial basis function neural networks. Acta Scientiarum-Agronomy2020; 43: e46307. https://doi.org/10.4025/actasciagron.v43i1.46307

Sant'Anna IC, Silva GN, Nascimento M, Cruz C D. Subset selection of markers for the genome-enabled prediction of genetic values using radial basis function neural networks. Acta Scientiarum-Agronomy2020; 43: e46307. https://doi.org/10.4025/actasciagron.v43i1.46307

Santos RP, Dean DL, Weaver JM and Hovanski Y. Identifying the relative importance of predictive variables in artificial neural networks based on data produced through a discrete event simulation of a manufacturing environment. Journal International Journal of Modelling and Simulation.2018; 39:234-245. http://dx.doi.org/ 10.1080/02286203.2018.1558736.

Silva GN, Nascimento M, Sant'Anna IC, Cruz CD, Caixeta ET, Carneiro PCS, Rosado RDS, Pestana KN, Almeida DP, Oliveira MS. Artificial neural networks compared with Bayesian generalized linear regression for leaf rust resistance prediction in Arabica coffee. Pesquisa Agropecuaria Brasileira. 2017; 52:186-193. http://dx.doi.org/10.1590/s0100-204x2017000300009.

Silva GN, Tomaz RS, Sant'Anna IC, Carneiro VQ, Cruz CD, Nascimento M. Evaluation of the efficiency of artificial neural networks for genetic value prediction. Genet. Mol. Res. 2016; 15, 1–11.

Silva GN, Tomaz RS, Sant'anna IC, Nascimento M, Bhering LL and Cruz CD. Neural networks for predicting breeding values and genetic gains. Scientia Agricola. 2014; 71, 494-498. http://dx.doi.org/10.1590/0103-9016-2014-0057.

Silva Junior AC, Sant'Anna IC, Silva GN, Cruz CD, Nascimento M, Lopes LB, Soares PC. Computational intelligence and machine learning to study the importance of characteristics in flood-irrigated rice. 2021 a. Acta Scientiarum-Agronomy (*in prelo*).

Silva Júnior AC, Silva MJ, Cruz CD, Santanna IC, Silva GN, Nascimento M, Azevedo C F. Prediction of the importance of auxiliary traits using computational intelligence and machine learning: A simulation study. PLoS One, 21, p. EMID:a920715476, 2021 b.

Skawsang S, Nagai M, Nitin K and Soni P. Predicting rice pest population occurrence with satellite-derived crop phenology, ground meteorological observation, and machine learning: A case study for the Central Plain of Thailand. Appl. Sci. 2019; 9:4846. http://dx.doi.org/10.3390/app9224846.

Sousa IC, Nascimento M, Silva GN, Nascimento ACC, Cruz CD, Fonseca F, Almeida DP, Pestana KN, Azevedo CF, Zambolim L and Caixeita ET. Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. Scientia Agricola 2020; 78: 1–8. https://doi.org/10.1590/1678-992x-2020-0021.

Song H, Liu A, Li G, Liu X. Bayesian bootstrap aggregation for tourism demand forecasting. IntJ Tourism Res. 2021;1–14.https://doi.org/10.1002/jtr.245314.

Tan K, Li E, Du Q, Du P. An efficient semi-supervised classification approach for hyperspectral imagery. ISPRS Journal of Photogrammetry and Remote Sensing. 2014; 97:36–45. http://dx.doi.org/10.1016/j.isprsjprs.2014.08.003.

Tolo¸si L, Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics 27, 2011; 1986–1994.

Ventura RV, Silva MA, Medeiros TH, Dionello NL, Madalena FE, Fridrich AB, Valente BD, Santos GG, Freitas LS, Wenceslau RR, Felipe VPS & Corrêa GSS. Use of artificial neural networks in breeding values prediction for weight at 205 days in Tabapuã beef cattle. Arquivo Brasileiro de Medicina Veterinária e Zootecnia. 2012; 64:411-418. http://dx.doi.org/10.1590/S0102-09352012000200022.

Yu H, Campbell MT, Zhang Q, Walia H, and Morota G. Genomic Bayesian confirmatory factor analysis and Bayesian network to characterize a wide spectrum of rice phenotypes. G3: Genes, Genomes, Genetics.

CONCLUSÕES GERAIS

Muitos são os desafios existentes nos programas de melhoramento de plantas, principalmente na avaliação simultânea de características que forneça ampla variedade de informações. Quantificar a importância de variáveis permite obter progresso mais rápido, realizar avaliação fenotípica extensiva do germoplasma, selecionar e predizer características que apresentam baixa herdabilidade e/ou dificuldade de medição. A abordagem tradicional para seleção de variáveis é baseada em regressão linear múltipla. Nela é avaliada a relação entre uma variável resposta com duas ou mais variáveis independentes. Entretanto, essa abordagem apresenta limitações quanto à sua capacidade de analisar dados de alta dimensão além da não-captura das relações complexas e multivariadas entre variáveis.

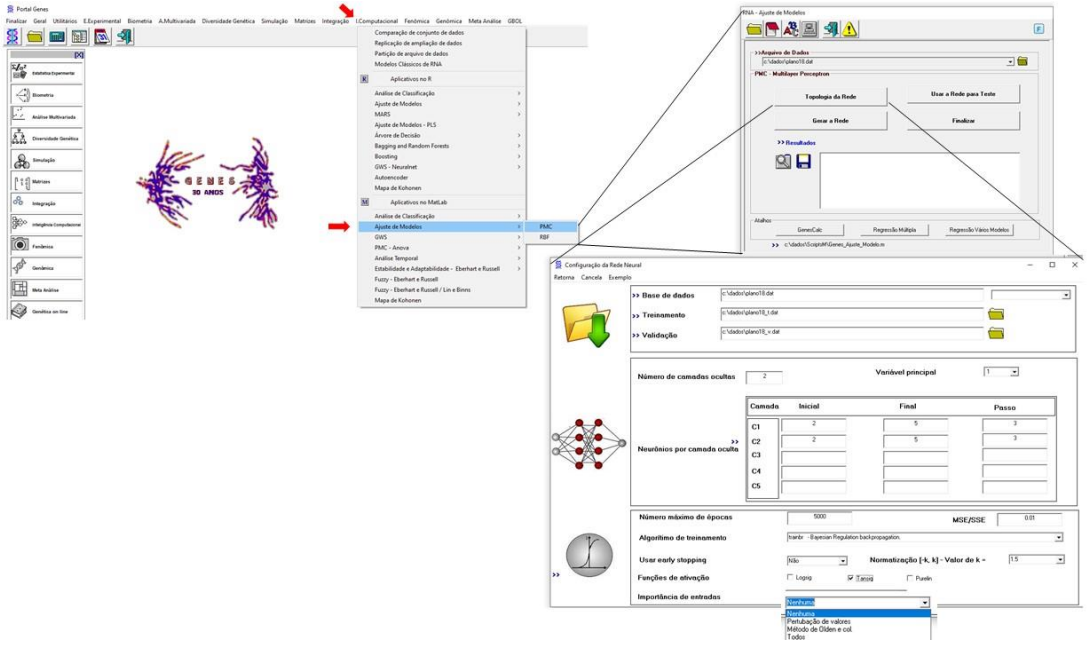
Desta forma, o objetivo deste trabalho no primeiro capítulo foi avaliar a importância de características auxiliares de uma característica principal com base em informações fenotípicas e estrutura genética previamente conhecida usando inteligência computacional e aprendizado de máquina para desenvolver ferramentas preditivas para o melhoramento genético. Para tanto, no segundo e terceiro capítulos, testamos se a utilização da melhor abordagem para predição e estabelecer uma rede de melhor poder preditivo em arroz irrigado por inundação, e aveia branca,

respectivamente, via metodologias baseadas em regressão, inteligência artificial e aprendizado de máquinas. Os resultados das análises evidenciam a possibilidade que o pesquisador tem para quantificar a importância de variáveis e reduzir o número de variáveis explicativas e, também, garantir que em estudos por outras abordagens não enfrente problemas de multicolinearidade e de dimensionalidade, sem que haja perda de informações genéticas relevantes. A contribuição relativa de caracteres auxiliares por meio de inteligência computacional e aprendizado de máquina mostrou-se eficiente para determinar a importância relativa de variáveis. Acredita-se que, com a utilização de procedimento para quantificar a importância de variáveis, as técnicas fundamentadas em inteligência computacional e aprendizado de máquina possam ser facilmente empregadas sem demandar recursos computacionais sofisticados.

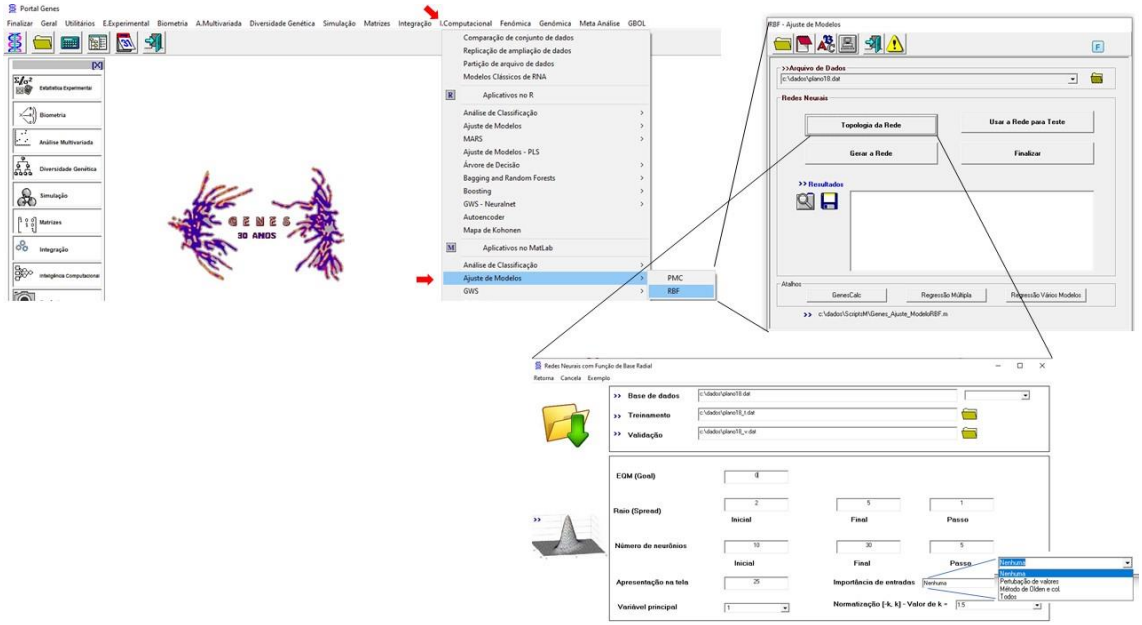
APÊNDICE A

Para realizar o download do software GENES deve-se acessar o site: ftp://ftp.ufv.br/dbg/biodata/. Deve ser estabelecido um arquivo.txt contendo informações fenotípica para as técnicas no software GENES.

Etapas: Clique em inteligência computacional com demonstrado pela seta em vermelha. Em seguinda em *Ajuste de Modelos* e *PMC*. Abrirá uma janela para estabelecer a topologia da rede neural. Na Figura abaixo têm-se:



Etapas: Clique em inteligência computacional com demonstrado pela seta em vermelha. Em seguinda em Ajuste de Modelos e RBF. Abrirá uma janela para estabelecer a topologia da rede neural. Na Figura abaixo, têm-se:




```
######################################################################
################ Garson ###############
setwd("C:\\Predicao")
#library(NeuralNetTools)
#library(nnet)
#library(neuralnet)
#library(RSNNS)
#library(ggplot2)
#library(nnet)
#library(caret)
dados<-read.table("local_1_Garson_total.txt",h=T)
head(dados)
names(dados)
rede<-nnet(Prod ~ Alt + Flo + CP + NGCHPan + Acam + Perf +
Comp + Larg + Esp +  RelCL + P100g + PorGC, method = 'nnet',
data = dados1,linout = TRUE, size = 30)
rede
garson(rede)
garson(rede, bar_plot = FALSE)*100
#Grafico
```

```
#library(tidyverse)
#library(ggplot2)
#########################################################
cd_PMC<-read.table("PMC_grafico_Geral.txt",header = T)
head(cd_PMC)
cd_PMC$Environment<-factor(cd_PMC$Environment,
                           levels =
cd_PMC$Environment[order(c(1:8))])
pv_PMC<- ggplot(cd_PMC, aes(x=Environment, y=TM,
fill=Environment)) +
  geom_bar(stat="identity", color="black",
           position=position_dodge()) +
  geom_errorbar(aes(ymin=TM-TD, ymax=TM+TD), width=.2,
                position=position_dodge(.9))+
  ylim(0,105)+
  ylab("determination coefficient")+
  theme_classic()+
  theme(legend.title = element_blank(),
        legend.position = c(-0.5,-0.95),
        legend.direction = "horizontal",
        legend.justification = "center")+
  scale_fill_grey(start = 0.4, end = 0.9, aesthetics =
"fill")+
  guides(fill = guide_legend(nrow = 1))+
  geom_hline(yintercept = 98.6, linetype = "dashed")
pv_PMC
ggsave(pv_PMC,filename="grafico_PMC.jpeg", units = "cm", dpi =
350)
#Dependencia:  tree
#install.packages('tree', repos='http://cran-r.c3sl.ufpr.br/')
#install.packages('ISLR', repos='http://cran-r.c3sl.ufpr.br/')
# Leitura dos dados
```

```
Xo<-read.table("trabalho$$$.dat",header=T)

Aux<-read.table("auxiliar$$$.dat",header=F)

nnos = Aux[1,1]

perc = Aux[2,1]

AuxBD<-read.table("basedado$$$.dat",header=T)

Y <-Xo[,1]

X <- Xo[,2:ncol(Xo)]

ajuste_tree <- tree(Y ~. ,data = X)

summary(ajuste_tree)

plot(ajuste_tree)

text(ajuste_tree, pretty = 0)

round(predict(ajuste_tree),2)

r<-
cor.test(Y,predict(ajuste_tree),method="pearson",alternative="
two.sided")

rs= cor(Y,predict(ajuste_tree))

r2 = rs*rs

erro = Y- predict(ajuste_tree)

  erro2<-erro^2

soma<-sum(erro2)

nn<-length(erro)

reqmt<-sqrt(soma/nn)

plot(Y,predict(ajuste_tree))

nteste = round(perc*nrow(X))

sorteio =sample(1:nrow(X), nteste)

X.val=X[-sorteio,]

Y.val=Y[-sorteio]

X.trein=X[sorteio,]

Y.trein=Y[sorteio]

set.seed(2)

ajuste_treeT=tree(Y.trein~.,data = X.trein)
```

```
saidaT=predict(ajuste_treeT,X.trein)
rst= cor(Y.trein,saidaT)
rs= cor(Y,predict(ajuste_tree))
r2 = rst*rst
erro<-Y.trein - saidaT
erro2<-erro^2
soma<-sum(erro2)
nn<-length(erro)
reqmt<-sqrt(soma/nn)
saidaVT=predict(ajuste_treeT, newdata=X.val)
rsvt= cor(Y.val,saidaVT)
r2 = rsvt*rsvt
erro<-Y.val - saidaVT
erro2<-erro^2
soma<-sum(erro2)
nn<-length(erro)
reqmv<-sqrt(soma/nn)
plot(saidaVT,Y.val)
abline(0,1)
# validação cruzada
cv_tree <- cv.tree(ajuste_tree)
plot(cv_tree)
# Prume
nnos = Aux[1,1]
melhor_tree <- prune.tree(ajuste_tree, best = nnos)
plot(melhor_tree)
text(melhor_tree, pretty = 0)
round(predict(melhor_tree),2)
r<-
cor.test(Y,predict(melhor_tree),method="pearson",alternative="
two.sided")
```

```
rs= cor(Y,predict(melhor_tree))
r2 = rs*rs
erro<-Y - predict(melhor_tree)
  erro2<-erro^2
soma<-sum(erro2)
nn<-length(erro)
reqmBest<-sqrt(soma/nn)
plot(Y,predict(melhor_tree))

# Dependencia: randomForest
#install.packages('randomForest', repos='http://cran-
r.c3sl.ufpr.br/')
#library(randomForest)
# Leitura dos dados
Xo<-read.table("trabalho$$$.dat",header=T)
Aux<-read.table("auxiliar$$$.dat",header=F)
nvariavel = Aux[1,1]
narvore = Aux[2,1]
perc = Aux[3,1]
perc= perc/100

# Bagging
Y <-Xo[,1]
X <-Xo[,2:ncol(Xo)]
nteste = round(perc*nrow(X))
nval =  nrow(X)-nteste
sorteio =sample(1:nrow(X), nteste)
X.val=X[-sorteio,]
Y.val=Y[-sorteio]
X.trein=X[sorteio,]
Y.trein=Y[sorteio]
```

```
# 1 analise com todos os dados

rf1=randomForest(Y ~. ,data=X,importance=TRUE, mtry = ncol(X),
ntree = narvore)

rf1

saida1 = predict(rf1,X)

eqm = mean((saida1-Y)^2)

round(importance(rf1), 2)

varImpPlot(rf1,main = "Importância -Bagging  Todos os dados")

rr= cor(Y,saida1)

r2 = rr*rr

texto=paste("R²=", round(r2,4))

plot(saida1, Y, main = "Ajuste -Bagging  todos os dados", sub
= texto)

abline(0,1)

#2 analise com dados de treinamento

round(importance(rf1), 2)

rr= cor(Y.trein,saida2)

r2 = rr*rr

texto=paste("R²=", round(r2,4))

plot(saida2, Y.trein,  main = "Ajuste -Bagging  dados de
treinamento", sub = texto)

abline(0,1)

#3 teste de validação

saida3 = predict(rf2,newdata=X.val)

eqm = mean((saida3-Y.val)^2)

rr= cor(Y.val,saida3)

r2 = rr*rr

texto=paste("R²=", round(r2,4))

plot(saida3, Y.val, main = "Ajuste -Bagging  dados de
validação", sub = texto)

abline(0,1)

# Floresta Aleatória
```

```
#1 analise com todos os dados

bg1=randomForest(Y ~. ,data=X,mtry = nvariavel,
importance=TRUE, ntree = narvore)

bg1

saida1a = predict(rf1,data=X)

eqm = mean((saida1a-Y)^2)

round(importance(bg1), 2)

varImpPlot(bg1,main = "Importância -Floresta Aleatória Todos
os dados")

rr= cor(Y,saida1a)

r2 = rr*rr

texto=paste("R²=", round(r2,4))

plot(saida1a, Y, main = "Ajuste -Floresta Aleatória todos os
dados", sub = texto)

abline(0,1)

#2  analise com dados de treinamento

bg2=randomForest(Y.trein ~. ,data=X.trein,mtry =
nvariavel,ntree=narvore)

bg2

saida2a = predict(bg2,X.trein)

eqm = mean((saida2a-Y.trein)^2)

round(importance(bg2), 2)

rr= cor(Y.trein,saida2a)

r2 = rr*rr

texto=paste("R²=", round(r2,4))

plot(saida2a, Y.trein,  main = "Ajuste -Floresta Aleatória
dados de treinamento", sub = texto)

abline(0,1)

#3 teste de validação

saida3a = predict(bg2,newdata=X.val)

eqm = mean((saida3a-Y.val)^2)

rr= cor(Y.val,saida3a)
```

```
r2 = rr*rr

texto=paste("R²=", round(r2,4))

plot(saida3a, Y.val, main = "Ajuste -Floresta Aleatória dados
de validação", sub = texto)

abline(0,1)


Procedimento: Boosting

# Dependencia: gbm

#install.packages('gbm', repos='http://cran-r.c3sl.ufpr.br/')

#library(gbm)

# 2. Leitura dos dados

Xo<-read.table("trabalho$$$.dat",header=T)

Aux<-read.table("auxiliar$$$.dat",header=F)

iteracao = Aux[1,1]

narvore = Aux[2,1]

perc = Aux[3,1]

valork = Aux[4,1]

delta= Aux[5,1]

perc= perc/100

AuxBD<-read.table("basedado$$$.dat",header=T)

Y <-Xo[,1]

X <-Xo[,2:ncol(Xo)]

nteste = round(perc*nrow(X))

sorteio =sample(1:nrow(X), nteste)

X.val=X[-sorteio,]

Y.val=Y[-sorteio]

X.trein=X[sorteio,]

Y.trein=Y[sorteio]

#1 Boosting - todos os dados

boost0=gbm(Y~.
,data=X,distribution="gaussian",n.trees=narvore,interaction.de
pth=iteracao)
```

```
summary(boost0)

#2 Análise gráfica

par(mfrow=c(1,ncol(X.trein)))

for(i in 1:ncol(X.trein))

{

varsel <- i

plot(boost0,i=colnames(X)[varsel], main = "Dados de
Treinamento")

}

saida0=predict(boost0,newdata=X,n.trees=narvore)

head(round(saida0, 2),20)

tail(round(saida0, 2),20)

eqm = mean((saida0 - Y)^2)

par(mfrow=c(1,1))

plot(saida0, Y, main = "Todos os dados: obs. e predito")

abline(0,1)

rr= cor(Y,saida0)

r2 = rr*rr

boost1=gbm(Y.trein ~.
,data=X.trein,distribution="gaussian",n.trees=narvore,interact
ion.depth=iteracao)

boost1

summary(boost1)

# Análise gráfica

par(mfrow=c(1,ncol(X.trein)))

for(i in 1:ncol(X.trein))

{

varsel <- i

plot(boost1,i=colnames(X.trein)[varsel], main = "Dados de
Treinamento")

}

#.3 Eficiência da predição                          ", "\n")
```

```
saida1=predict(boost1,data=X.trein,n.trees=narvore)

head(round(saida1, 2),20)

tail(round(saida1, 2),20)

eqm = mean((saida1 - Y.trein)^2)

par(mfrow=c(1,1))

plot(saida1, Y.trein, main = "Dados de treinamento : obs. e
predito")

abline(0,1)

rr= cor(Y.trein,saida1)

r2 = rr*rr

saida2=predict(boost1,newdata=X.val,n.trees=narvore)

head(round(saida2, 2),20)

tail(round(saida2, 2),20)

eqm = mean((saida2 - Y.val)^2)

par(mfrow=c(1,1))

plot(saida2, Y.val, main = "Validação: obs. e predito")

abline(0,1)

rr= cor(Y.val,saida2)

r2 = rr*rr

# Boosting - Validação cruzada

boost2=gbm(Y.trein ~. ,data=X.trein,distribution="gaussian",
cv.folds = valork,
n.trees=narvore,interaction.depth=iteracao,shrinkage=delta,ver
bose=F)

boost2

#Importância relativa

summary(boost2)

#Eficiência da predição em dados de treinamento

saida3=predict(boost2,newdata=X.trein,n.trees=narvore)

saida2

eqm = mean((saida3 - Y.trein)^2)
```

```
plot(saida3, Y.trein,main = "Validação cruzada- treinamento:
obs. e predito")

abline(0,1)

rr= cor(Y.trein,saida3)

r2 = rr*rr

#Eficiência da predição em dados de validação

saida4=predict(boost2,newdata=X.val,n.trees=narvore)

saida4

eqm = mean((saida4 - Y.val)^2)

plot(saida4, Y.val,main = "Validação cruzada - validação: obs.
e predito")

abline(0,1)

rr= cor(Y.val,saida4)

r2 = rr*rr

#check performance using an out-of-bag estimator

# OOB underestimates the optimal number of iterations

summary(boost2,n.trees=1)          # based on the first tree

best.iter <- gbm.perf(boost2,method="OOB")

summary(boost2,n.trees=best.iter)

# check performance using a 50% heldout test set

#best.iter <- gbm.perf(boost2,method="test")

#print(best.iter)

#summary(boost2,n.trees=best.iter)

best.iter <- gbm.perf(boost2,method="cv")

summary(boost2,n.trees=best.iter)

par(mfrow=c(1,ncol(X.trein)))

for(i in 1:ncol(X.trein))

{

varsel <- i

plot(boost2,varsel,best.iter, main = "Melhor iteração")

}
```

```
p= ncol(X.trein)
np = p*(p-1)/2
par(mfrow=c(1,np))
fim=ncol(X.trein)-1
for(i in 1:fim)
{
ini <- i+1
for(j in ini:ncol(X.trein))
{
varsel1 <- i
varsel2 <- j
plot(boost2,c(varsel1,varsel2),best.iter, main = "Dispersão Xi
Xj após melhor iteração")
}
}
```