Computational intelligence to study the importance of predictors in white oat (Avena sativa L.)

Antônio Carlos Silva Júnior (antonio.silva.c.junior@gmail.com)
Federal University of Viçosa

Isabela Castro Sant'Anna
Instituto Agronômico (IAC)

Michele Siqueira
Federal University of Viçosa

Leonardo Lopes Bhering
Federal University of Viçosa

Moysés Nascimento
Federal University of Viçosa

Ivan Ricardo Carvalho
Universidade Regional do Noroeste do Estado do Rio Grande do Sul

José Antônio Gonzalez Silva
Universidade Regional do Noroeste do Estado do Rio Grande do Sul

Cosme Damião Cruz
Federal University of Viçosa

Article

Keywords: Avena sativa L, multiple regression, computational intelligence, machine learning

Posted Date: July 12th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1734548/v1

Abstract

The objective of this work was to estimate the best approach for prediction 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. Were evaluated without and with fungicide, established prediction models in four experimental sets. The characteristics evaluated were grain yield, which was used as a response variable, and ten others as explanatory variables. Assessing the importance of variables 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 adjusted based on R^2 was used to quantify the importance of the phenotypic trait. The characters indicated to assist in decision-making are plant height, leaf rust severity, and lodging percentage. 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 more elevated than the others.

1. Introduction

White oats (*Avena sativa* L.) are of great agricultural importance worldwide. Brazil is the fifth-largest producer globally and has shown a substantial increase in the area cultivated with white oats in the last ten years [1]. This crop can be used to produce grain, forage, and straw in a no-tillage system [2].

Estimating the importance of predictor variables in breeding programs allows for faster progress, selecting, and predicting traits with low heritability and/or measurement difficulty [3, 4]. Although simultaneous assessment of characteristics provides a wide variety of information, identifying which predictor variable is more critical is challenging for the breeder [5]. The estimation of the importance of variables can be performed by artificial neural networks (ANNs) through algorithms such as Goh (1995)[6], who proposed a modification in the Garson (1991) algorithm[7], which consists of partitioning the neural network connection weights to determine the relative importance of each input variable in the network.

Regression, artificial intelligence, and machine learning-based methodologies have been successfully used in prediction studies. [5] evaluated the high-dimensional phenotypic traits in soybean through the machine learning approach to predict seed yield for the prescriptive development of cultivars for agricultural practices. [8] applied such methodologies to predict the insect pest population using host plant climatic and phenological factors. [4] used these methodologies to predict grain yield, grain length-width ratio, and panicle length in flood-irrigated rice. [3] evaluated the importance of auxiliary traits of the main trait based on phenotypic information and previously known genetic structure using computational intelligence and machine learning to develop good predictive tools in breeding programs. However, there are no studies in the literature related to yield prediction and verification of the importance of variables for grain yield in white oat culture.

Given the above, this work aims to: (1) predict grain yield in white oat using methodologies based on regression, artificial intelligence, and machine learning; (2) identify more relevant predictors, considering different prediction approaches in white oat.

2. Material And Methods

2.1. Experimental data

The field experiment was carried out in the experimental area of the Instituto Regional de Desenvolvimento Rural (IRDeR) of the Universidade Regional do Noroeste do Estado do Rio Grande do Sul (UNIJUÍ) located in the municipality of Augusto Pestana - Rio Grande do Sul - Brazil, at coordinates 28 ° 26 '30' 'S and 54 ° 00' 58 " W, altitude 280 m. The collection of plant material complies with relevant institutional, national, and international guidelines and legislation. The soil is classified as typical distroferric Red Latosol. According to the climatic characterization of Köeppen, the region’s climate is of the Cfa type (humid subtropical), with four distinct seasons. The average annual temperature is 19.9°C, and the average annual rainfall is 1774 mm.

Seventy-eight white oat genotypes were evaluated in 2008 and 2009. Each year, they were assessed without and with a fungicide to establish pr in four experimental sets (E1, E2, E3, and E4). The fungicide used was Orkestra, an active ingredient of the pyraclostrobin group (333 g l^{-1}). The design was in randomized blocks with three replications.

The characteristics were grain yield (GY, Kg ha^{-1}) which were used as the response variable, and the others as explanatory variables (inputs), that is, mass edition models of a thousand grains (MTG, grams); hectoliter weight (HW, kg ha^{-1}); days between emergence and maturation (DEM, day); lodging percentage (LP, in percentage, where 1% bedded little and 100% bedded down completely); days from emergence to flowering (DEF, day); days from flowering to maturity (DFM, day); plant height (PH, cm); leaf rust severity (LRS); stem rust severity (SRS); leaf spots (LS). They were used to compose artificial neural networks of white oat genotypes.

2.2. Methodologies for predicting and verifying the importance of characteristics

2.2.1. Multiple Regression

Stepwise multiple regression is the variable selection method, which aims to explain the relationship between a set of independent variables and a dependent variable. The coefficient of determination (R^2) aims to estimate how much of the independent variable is explained by the total variation of the dependent variable [3, 4].

2.2.2. Computational intelligence for the importance of variables

2.2.2.1. Multilayer Perceptron - PMC

The importance of predictors through the PMC network was quantified using two techniques. The first, based on Garson's (1991)[7] algorithm modified by Goh (1995)[6], consists of partitioning the neural network connection weights to determine the relative importance of each input variable within the network [3, 4].

The equation of the relative importance of variables is equal to

$$IR = WV$$

1

The matricial model is shown as follows

$$IR = \begin{pmatrix} IR_1 \\ IR_2 \\ : \\ IR_k \end{pmatrix} = \left(W^1_{N_1E}\right)^{'} \left(W^2_{N_2N_1}\right)^{'} \ldots \left(W^c_{N_{c-1y}}\right)^{'}$$

,

where, W^c_x represents the matrix of weights of the layer c neuron, considering N_j neurons and N_{j-1} inputs; E is the first neuron that starts from inputs; y refers to the desired output layer and IR: relative importance of the variable.

After the network is established, the importance of variables (inputs) can also be obtained, considering the impact of destructuring or disturbing the information of a given input on the estimation of the coefficient of determination [3, 4].

The relative importance of the variable by the permutation of R^2 is described in the following equation:

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

2

where, R^2_{obs} is the R^2 of the RNA model adjusted to the observed predictor and response variables; $R^2_{perm,x_i}$is the R^2 of the ANN model fitted to the modified dataset where x_i is permuted; $\overline{R}^{2}_{perm,x_i}$: is the average value of R^2_{perm,x_i} after mth permutation of the datasets.

After some criteria used on the best topology, the following PMC network structures were adopted: (a) topology 1: 10-11-1: ten inputs, 11 hidden neurons in the middle layer and one neuron in the output layer; (b) topology 2: 10-11-11-1: ten inputs and two hidden layers with 11 neurons in the middle layers and one neuron in the output layer; (c) topology 3: 10-11-11-11-1: ten inputs, and three hidden layers with 11 neurons in the middle layers and one neuron in the output layer; (d) topology 4: 10-3-4-11-1: ten inputs, and three hidden layers with three, four and 11 neurons in the middle layers and one neuron in the output layer.

2.2.2.2. Radial Base Function Network – RBF

The prediction efficiency is measured by the coefficient of determination and the relative importance of each input estimated by the technique of destructuring the information of each explanatory variable, as already described for PMC.

2.2.3. Machine Learning for the importance of variables

To quantify the importance of variables through a machine learning approach, the decision tree and its refinements, random forest, bagging, and boosting were used [3, 4].

The importance of variable IV is described in the following Equation:

$$IV_{x_i} = MSE_{perm,x_i} - MSE_{nperm}$$

3

where, MSE_{perm,x_i} is the permutation of the values of each variable in the dataset where x_i is swapped; MSE_{nperm}: values of the estimate of the original non-permuted variable data.

2.2.4. Importance of variables, in reduced models, in the prediction of grain yield

The biometric technique that led to the best GY prediction results and information regarding the importance of predictors was considered.

2.3. Training and Validation Sets

The training set included the same individuals for modeling using all methodologies and was composed of 67% of the individuals, which corresponds to 2/3 of the randomly selected individuals. The remaining 33% (1/3) of the individuals constituted the validation set. In previous studies, 60–90% of individuals constituted the training set [9]. All analyses were performed using the GENES software in integration with the Matlab software [10, 11].

3. Results

3.1. Prediction of grain yield by different approaches

The estimate of the coefficient of determination, for all methodologies using the ten defining agronomic characteristics in the prediction of grain yield (GY) in white oats is shown in Table 1.

Table 1
Mean of the maximum estimate of the coefficient of determination for the training set, in four environments corresponding to the data set of experiments with and with fungicide in two agricultural years, to predict the grain yield in white oat (*Avena sativa* L.).

Approach	Technique	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**
Conventional	RM	61.02	46.07	20.67	32.72

AI: Artificial Intelligence; AM: Machine Learning; RM: Multiple Regression; PMC: Multilayer Perceptron; PMC: Multilayer Perceptron; PMC-1: Multilayer Perceptron with (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: Radial Base Network; DT: Decision Tree; RF: Random Forest; BA: Bagging; BO: boosting. E: environments. E1 and E3: no fungicide; E2 and E4: with fungicide.

Based on Table 1, it is possible to compare the approach that is more efficient for the prediction of GY. Higher values of R^2indicate that the prediction target variable has a better fit considering the ten explanatory variables used as predictors in this analysis [3, 4]. Among the methodologies used in this study, it was found that multiple regression presented a lower estimate of R^2, indicating the existence of non-linear 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 non-linear information from model inputs [5, 8], as seen in Table 1. Other authors have already highlighted the abilities of neural networks [12, 13] and machine learning [3, 4, 14] to better capture non-linear relationships when compared to conventional methodologies.

The results obtained by different approaches show that there was a discrepancy between the maximum estimate of $R^2$for the predictive variable in the same environments (Table 1). This discrepancy in the estimate of $R^2$was also reported by [3, 4]. It is noteworthy that the differences in results obtained in these analyzes are indicative that the environment influences the estimate ofR^2 and, consequently, the choice of the best prediction model for the response variable.

The machine learning approach proved to be more efficient compared to the other approaches (Table 1). There was a low estimate of R^2maximum in the random forest procedure, for all environments. On the other hand, this procedure was superior to the multiple regression approach for the same environment, except the environment without fungicide (E3), which corresponds to 10.57%. The low estimate of R^2maximum in the random forest procedure was also demonstrated in flood-irrigated rice [4] and on simulated data with different heritability [3]. This procedure involves the steps of randomly resampling the set of explanatory variables, and building several decision trees that will constitute a random forest that will allow the prediction and estimation of scores that will lead to the evaluation of the importance of predictors in a process repeated several times.

Regarding the environments and the bagging procedure, it appears that the estimates of R^2were higher than 92.70 %,making this approach the best highlight for use in the analyzed data sets. High estimates (with reference to values around 80%) of R^2were also obtained using machine learning methodologies by boosting procedures, in addition to bagging, for all prediction data sets (Table 1). [3, 4] showed that the machine learning approaches for the bagging and boosting procedures were more consistent in obtaining a higher overall mean estimate of R^2, about predictive variables. The decision tree (DT) and random forest methodology did not stand out from other machine learning procedures (Table 1).

Artificial intelligence approaches based on RBF provided adjustments whose R^2were greater than 70% in all environments (Table 1). In this procedure, the highest estimate R^2maximum was 90.12% (± 5.79) and the lowest 73.75% (± 1.67), which corresponds to environments E1 and E2, respectively. [4] found an estimate of maximum R^2 ranging from 48–99% in different environments for the flood-irrigated rice crop. For simulated data with the different genetic structures the maximum estimate of R^2 ranging from 44–54% [3] and [15] obtained results of R^2 consistent for different genetic structures. [16] evaluated bean cultivars and obtained an estimate of R^2 for the characteristics days to first flower and flowering days of 94.10% and 94.40%, respectively. This procedure has a good ability to handle complex interactions compared to semiparametric and linear regressions [15, 17]. Generally, RBF is quick to learn from the data used as training information and provides a unique solution compared to perceptron ANNs [9, 15, 17].

Radial basis function networks have a good ability to handle interactions compared to semiparametric and linear regressions [15]. [15] applied the RBF in studies using simulated traits with 30% and 60% heredity for variable selection. The authors identified greater efficiency in the selection by the RBF when the scenario involved epistatic interactions in the gene control of the studied characters. [9] observed that it is possible to improve prediction in nonparametric models when the selection includes markers that are not directly related to the characteristics of interest. [4] applied RBF to predict grain yield, grain length-width ratio, and panicle length in flood-flooded rice. These authors argue that RBF has a high performance in predicting the importance of variables. [3] evaluated the importance of auxiliary traits of the main trait based on phenotypic information and previously known genetic structure using RBF and demonstrated the efficiency of this network to quantify the importance of variables.

Regarding procedure PMC-1 (10-11-1), the highest estimate ofR^2maximum was observed in E1- 73.25% and the lowest in E3, with an estimate of 30.14%, both environments correspond to the one without fungicide. In the procedure PMC-2 (10-11-11-1) and PMC-3 (10-11-11-11-1) the highest estimates were observed in E1 and E2 and the smallest in E3 and E4, respectively. For the same hidden layer number that corresponds to PMC-3 (10-11-11-11-1) and PMC-4 (10-3-4-11-1). We observed lower estimates of maximum R^2 for the PMC-4 procedure, except the E3 environment. This shows that the number of neurons in the layer influences the estimation of R^2 maximum. [3] argued that the number of neurons influences the estimation of the coefficient of determination.

The PMC network is widely used in the predictive process [3, 4, 18], since the success of this network has already been demonstrated 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 [18].

Thus, machine learning is actually more efficient for selecting phenotypic traits because it can handle reduced or redundant information about phenotypic traits [3]. [19] evaluated the importance of variables by bagging, random forest, boosting, decision tree, PML and RBF and reported that PML and RBF achieved better results. [3, 4] verified that the methodologies of computational intelligence and machine learning in the prediction allowed to identify the explanatory phenotypic characteristics that should be prioritized and established as auxiliary characteristics for the indirect selection.

The efficiency of ANNs in prediction problems given their ability to extract relevant information from large data sets [20] and generalize relatively inaccurate information [21], was very well expressed by the results obtained (Table 1). The same can be seen for methodologies based on machine learning, which are capable of dealing with more reduced or redundant information in the input variables [3, 4]. However, another study as important as prediction and which is often not carried out is the identification of more important predictive variables, which is an important factor in the decision-making process [22]. Thus, after the prediction analyses, analyzes were carried out to quantify the importance of variables through the methods of artificial intelligence and machine learning, in order to identify, among the set of explanatory variables, those that should be prioritized and identified as auxiliary characteristics in indirect responses to selection.

3.2. Linear relationship between predictor and grain yield variables in white oat

The greatest linear associations with GY may be a preliminary indication that the variables, individually, are important in the prediction of GY. In multivariate prediction models, a predictor variable, with high correlation with the response variable, may lose its importance due to its redundancy, considering that, in the model, it may be represented by another associated. Thus, in addition to quantifying the linear relationships between predictor-response, it is important to quantify and appreciate the linear relationships, expressed by linear correlation coefficients, between all predictors in the search for redundancies. In this work, these associations were represented in a correlation network that contains red and green lines that represent negative and positive correlations, respectively, and their width is proportional to the magnitude of the correlations (Fig. 1). Regarding the phenotypic correlation network, they observed that the structure of correlated groups aiming to predict GY. In this network, the similarity between the phenotypic characteristics and the phenotypic correlation patterns is highlighted.

The characteristics that present groups with GY in E1 were MTG, HW and PH that correlated positively, but varying in magnitude, and the negatively correlated was LRS. In relation to E2, the positively correlated characteristics consist of: PH and MTG; and negatively LS and DFM. For E3, which represents no fungicide, the characteristic that is negatively correlated was SRS. Environment 4, the positively correlated group consists of HW and DEF and the negative DEM (Fig. 1).

3.3. Importance of variables in prediction by Artificial Intelligence approach

3.3.1. Multilayer Perceptron (PMC)

Estimates of the coefficient of determination of grain yield prediction by PMC attributing perturbation to the genotypic information are shown in Table 2. These results show large discrepancies in theR^{2^*} in comparing the environments with each other, which makes interpretation difficult. In environments E1 and E4, which correspond to environments without fungicide, the characteristics LP, PH, LRS were efficient in quantifying the response variable GY due to the reduction in the estimate of R^{2^*} as a function of the strategy of attributing disturbance to phenotypic information.

Table 2
Estimates of the coefficient of determination of grain yield prediction in white oat (*Avena sativa* L.), using PMC attributing perturbation to genotypic information.

	E1				**E2**			
Input	TOP1	TOP2	TOP3	TOP4	TOP1	TOP2	TOP3	TOP4
MTG	70.02	87.37	64.93	74.93	**31.92**	**23.19**	**9.73**	**26.47**
HW	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
PH	**61.01**	80.26	**49.89**	**59.89**	53.23	63.91	33.01	55.37
LP	75.26	**66.07**	**62.90**	**67.90**	46.46	71.41	76.46	68.43
LRS	**52.80**	**33.62**	**10.33**	**8.33**	52.72	73.18	85.34	67.20
SRS	76.59	78.03	71.10	71.10	57.33	80.89	58.86	60.60
LS	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
MTG	32.34	33.29	52.69	65.34	51.85	38.73	50.93	58.67
HW	**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
PH	29.51	35.36	57.87	44.98	50.74	53.91	55.85	56.06
LP	**18.57**	39.66	**29.95**	51.51	59.52	48.74	57.27	59.37
LRS	**4.48**	**11.46**	**24.87**	**21.62**	44.69	**29.64**	45.38	**39.15**
SRS	24.65	**19.57**	38.52	**9.99**	**39.55**	42.86	**31.70**	40.07
LS	26.36	12.91	49.10	45.01	56.54	**37.20**	45.79	59.26

MTG = Thousand Grain Mass in grams; HW = Hectoliter Weight; DEM = Days between Emergency and Maturation; PH = percentage of lodging; GY = Grain yield in kilograms per hectare; DEF = Days from Emergence to Flowering; DFM = Days from Flowering to Maturation; PH = Plant Height; LRS = Leaf Rust Severity; SRS = Stem Rust Severity and LS = Leaf Spots; E: environments. E1 and E3: no fungicide; E2 and E4: with fungicide. Topology- TOP1: Multilayer Perceptron with (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: environments. E1 and E3: no fungicide; E2 and E4: with fungicide.

Regardless of the number of neurons in the output layer and a single hidden layer, they agreed to pinpoint the most important variables to predict GY. This result shows that these variables are important in predicting GY, as the disturbance in their values led to a considerable reduction in the quality of the fit. In the E2 environment, the MTG characteristic was the most important in predicting GY.

There was a discrepancy in the number of neurons in the output layer and hidden layer, pointing out that the most important variables in E4, which correspond to the fungicide environment. With only one neuron in the output layer and a single hidden layer they showed that DEF and SRS were the most important due to the reduction in the estimate ofR^{2^*}. With two neurons in the middle layer and a single hidden layer they demonstrated that LRS and LS for the target prediction variable. When we use a neuron in the input layer, and three hidden layers with 11 neurons in the intermediate layer and one neuron in the output layer, the characteristics that proved to be the most important were HW and SRS. On the other hand, with three hidden layers with three, four and 11 neurons in the intermediate layer, the important characteristics in predicting the GY were: LRS, DEF and HW. [4] reported that 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 in irrigated rice, given the significant drops in estimated values of R^{2^*}observed when we disturb the variables

The importance of the variables was quantified by assigning destructuring to the genotypic information referring to each variable, in order to observe what changes would occur in the values of theR^2. It is important to point out that, in this Table, reductions in the values of R^2 after attributing disruption to the genotypic information referring to each variable, are indicative that this variable is important in relation to the others for purposes of prediction with the already established network.

3.3.2. Radial Base Network (RBF)

The estimation of the importance of characters in white oat attributing disturbance to the information of an input variable after the RBF has been established is described in Table 3. In this Table, the relative importance of each input estimated by the technique of destructuring the information of each variable explanatory. When using this strategy, drastic reductions in the values of R^{2^*}were observed for the most important variables and LRS for the predictive variable GY, in the E1 and E4 environments. In practice, the intensity of this trait reduces genetic progress to increase grain yield. In the E2 environment, the variable that suffered the greatest reduction in R^{2^*} was DMF, with an estimate of 44.47%. This feature increases grain yield, as more photoassimilates are produced and translocated to grains. However, late cycle cultivars tend to be more productive in relation to the initial cycle, as they obtain an increase in the amount of photoassimilates that are translocated to the grains [4].

Table 3
Coefficient estimates for determining grain yield prediction in white oat (*Avena sativa* L.) using the RBF attributing perturbation to genotypic information.

Input	E1	E2	E3	E4
MTG	81.04	58.97	**47.98**	40.77
HW	**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
PH	**73.97**	59.36	**62.75**	72.19
LP	81.96	68.07	64.71	64.71
LRS	**60.1**3	63.30	71.59	**45.04**
SRS	84.38	70.50	69.10	63.74
LS	88.37	61.23	**54.09**	52.25

MTG = Thousand Grain Mass in grams; HW = Hectoliter Weight; DEM = Days between Emergency and Maturation; PH = percentage of lodging; GY = Grain yield in kilograms per hectare; DEF = Days from Emergence to Flowering; DFM = Days from Flowering to Maturation; PH = Plant Height; LRS = Leaf Rust Severity; SRS = Stem Rust Severity and LS = Leaf Spots; E: environments. E1 and E3: no fungicide; E2 and E4: with fungicide.

The results show that the most important variable using the RBF was MTG, in the E2, E3 and E4 environments, with estimates of 58.97%, 47.98% and 40.97%, respectively. In practice, MTG influences the grain yield in white oats, since the higher MTG, consequently, the higher GY. This justifies the results of this study in white oats in the prediction of GY.

The results obtained corroborate the expectation about the RBF in quantifying and revealing the importance of the characteristics using the strategy of causing disturbances from 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 in white oats. Techniques that show the impact of interruption or disturbance in the information of a given input in the estimation of the coefficient of determination and partition of the connection weights of the ANN were presented. These techniques were effective in estimating the true importance of phenotypic traits. Therefore, there is a certain agreement between the results found by the two computational intelligence methodologies of PMC networks and RBF networks.

3.4. Importance of variables in predicting by approach Machine Learning

Table 4 shows the means of the relative contributions of the explanatory variables for grain yield prediction by estimating the minimum squared error increment percentage (IMSE), which is constructed by swapping the values of each variable in the data set, and comparing with the prediction of the original non-permuted dataset of 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^2indicated greater importance of that variable for the model, in the machine learning approach the importance of the explanatory variable it is related to the estimation of the average decrease in the precision of the model through the IMSE so that the higher this estimate, the greater the importance of the variable.

Table 4
Average estimate of the relative contributions of the explanatory variables for grain yield prediction in white oat using a machine learning approach, in four environments corresponding to without and with fungicide application.

VA	**E1**			**E2**			**E3**			**E4**		
	BA	**RF**	**BO**	**BA**	**RF**	**BO**	**BA**	**RF**	**BO**	**BA**	**RF**	**BO**
MTG	**7.58**	**7.94**	**12.37**	**10.47**	**9.84**	**10.89**	1.04	1.53	**4.49**	3.55	3.21	3.75
HW	**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**
PH	**10.74**	**10.3**	**9.65**	**10.01**	**8.29**	**9.32**	-0.93	-0.24	2.72	0.8	-0.45	1.02
LP	2.83	2.49	5.94	1.36	1.08	2.79	3.04	3.04	2.96	0.36	-0.66	0.89
LRS	**20.87**	**20.1**	**29.59**	**9.27**	**9.29**	**16.05**	**9.91**	**10.91**	**12.29**	**4.19**	**4.58**	**7.02**
SRS	7.32	7.76	5.60	3.09	2.25	3.65	**3.52**	**3.97**	**3.71**	0.8	1.62	2.04
LS	3.11	3.67	4.69	3.62	2.91	3.74	3.22	2.95	3.30	3.99	3.49	3.59

MTG = Thousand Grain Mass in grams; HW = Hectoliter Weight; DEM = Days between Emergency and Maturation; PH = percentage of lodging; GY = Grain yield in kilograms per hectare; DEF = Days from Emergence to Flowering; DFM = Days from Flowering to Maturation; PH = Plant Height; LRS = Leaf Rust Severity; SRS = Stem Rust Severity and LS = Leaf Spots; FA: random forest; BA: Bagging; BO: Boosting; VA: auxiliary variable; E: environments. E1 and E3: no fungicide; E2 and E4: with fungicide.

Based on Table 4, the variables that obtained the highest IMSE estimate in all machine learning methodologies in relation to environments without fungicides were: LRS, HW, PH, and MTG; DEF, SRS, and LRS, E1, and E3, respectively. The variable that showed to be more efficient in these environments was LRS. This justifies that this variable can be used in the indirect selection process when the target variable of prediction is GY. To environments with fungicides, the most important variables were: MTG, DFM, PH, and LRS; DEF, DFM, DEM, and LRS, which are represented by E2 and E4, respectively. For this environment with fungicide, the variables DFM and LRS proved to be efficient in estimating the prediction of grain yield in white oat.

The random forest and bagging methodologies were coincident in quantifying the same explanatory variables. Similar result is reported by [3, 4]. Regarding the boosting procedure, the results show discrepancies. On the other hand, this procedure was more consistent in variable prediction. In this procedure to estimate the importance of a variable using GY as a predictive target, the variables: MTG, HW, PH, and LRS; MTG, DEF and LRS stood out in the environment without fungicides, represented by E1 and E3, respectively. To the fungicide environment, the important variables were: MTG, DFM, PH, and LRS; DEF, DFM, DEM, and LRS, respectively. When using the boosting procedure, the variable that stood out in all environments was LRS. This justifies that this variable can be used to predict GY in white oats.

The bagging technique involves generating several distinct training sets from the original dataset. Final predictions are calculated by averaging all generated predictions. This is useful for decision tree and artificial neural network techniques that are sensitive to small changes in training data [23].

3.5. Importance of variables, in reduced models, in prediction by approach

3.5.1. Machine Learning

This topic considered the biometric technique that led to the best GY prediction results and the information regarding the importance of predictors, which was bagging.

The average estimate of the relative contributions of the explanatory variables for grain yield prediction in white oat using the bagging technique, after eliminating auxiliary variables of smaller relative contributions, in four environments corresponding to without and with fungicide application is shown in Table 5. The choice of this technique (bagging) was based on the estimate of the coefficient of determination (Table 1), which was greater than 90%, and the elimination of auxiliary variables of the smallest relative contributions established by Table 4.

Table 5
Estimate of the coefficient of determination for the training set, in four environments corresponding to the data set of experiments without and with fungicide in two agricultural years, to predict the grain yield in white oat (*Avena sativa* L.) utilizing the bagging technique.

Predictors	**E1**	**E2**	**E3**	**E4**
R^2 (v = 10)	**94.61**	**93.89**	**92.70**	**92.98**
Deleted	DFM	LP	PH	LP
R^2 (v = 9)	94.85	94.34	92.83	93.05
Deleted	DFM, LP	LP, HW	PH, DFM	LP, PH
R^2 (v = 8)	94.26	93.50	92.03	93.11
Deleted	SRS, LS	SRS, LS	SRS, LS	SRS, LS
R^2 (v = 8)	94.95	94.40	91.74	92.84

HW = Hectoliter Weight; LP = percentage of lodging; DEF = Days from Emergence to Flowering; DFM = Days from Flowering to Maturation; PH = Plant Height; SRS = Stem Rust Severity and LS = Leaf Spots; E: environments. E1 and E3: no fungicide; E2 and E4: with fungicide; R^2: coefficient of determination; v: variables.

The importance of predictors through the elimination of auxiliary variables of smaller relative contributions was quantified by using it in several ways. The first, based on the elimination of only one of the predictor variables (DFM, LP, PH, and LP), in E1, E2, E3, and E4, respectively, and then two variables that contributed the least. Finally, we opted for the elimination of the SRS and LS variables, which showed a lower estimate of the percentage of minimum squared error increment in all environments.

After eliminating auxiliary variables with smaller relative contributions, the maximum estimate of the coefficient of determination is similar when we use all auxiliary variables to predict GY (Tables 1 & 5).

4. Discussion

The literature has highlighted machine learning techniques as efficient tools in quantifying the relative importance of variables, in view of their simplicity, the non-use of assumptions about the distribution of explanatory variables, and also due to their robustness in relation to quantity, redundancy and environmental influences [3, 4, 22, 24]. Furthermore, such techniques 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 [25]. The bagging technique shows good predictive performance in practice; it works well for multidimensional problems and can be used with output from multiple classes, categorical predictors, and unbalanced problems [26]. Satisfactory results of variable selection using the bagging and random forest algorithm in the presence of correlated predictors were reported by [26]. Discriminatory power, redundancy, precision, and complexity can influence the indices or statistics used to quantify the importance of auxiliary traits in predicting a main characteristic.

Genetic improvement for desired traits in different crops has been a time-consuming, laborious and expensive process. Breeders study generations of plants and identify and modify desired genetic traits as they assess how traits are expressed in offspring [27]. The application of computational intelligence and machine learning to identify ideal sets of observable characteristics (phenotypes) can allow informed decisions and achieve highly relevant results in breeding programs. In addition, these methodologies can help predict auxiliary traits with the best performance under different agricultural management practices.

We compare different approaches to selecting or discarding variables that have been recently proposed to identify relevant predictive variables within a regression problem. Furthermore, we included in our comparison a traditional method that aims to find a small subset of important variables with optimal predictive performance in the white oat crop. It is noteworthy that the characteristics used in this study are difficult to obtain and their evaluation can be costly if there is a greater number of genotypes to be evaluated. In this context, the study of the most important characteristics in the prediction becomes necessary, since it is possible to reduce the physical effort, cost, use of labor, and time in the experimentation [27].

Therefore, our study presents the performance of some methodologies to assess the relative contributions of each variable through computational intelligence and machine learning in white oat culture. It is considered that the approach to estimate 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 is grain yield.

Methodologies based on machine learning and computational intelligence do not depend on stochastic information and tend to be more efficient, while conventional methodologies depend on the normal distribution of phenotypic characteristics. Furthermore, machine learning and computational intelligence methodologies make no assumptions about the model and can capture complex factors in predictive models. In machine learning, a priori knowledge of prediction is not needed if the data produces these effects, and no assumptions are made about the distribution of phenotypic values [10]. Machine learning algorithms have the advantage of modeling data non-linearly and non-parametrically [28]. Unlike many traditional statistical methods, these algorithms are built with the advantage of handling noisy, complex, and heterogeneous data [29]. Researchers now have the ability to identify the individual and interactive contributions of predictor variables to the white oat crop using artificial intelligence and machine learning.

5. Conclusion

Computational intelligence and machine learning methodologies were able to quantify the importance of explanatory variables in predicting white oat grain yield. The net with only one hidden layer was efficient to determine the relative importance of variables in white oat.

The bagging technique showed a high estimate of the coefficient of determination higher than the others. Simpler models, excluding predictors, are as efficient as more complex models, indicating that quantifying the importance of predictors is important to minimize costs, ensuring the same levels of efficiency as the predictive model.

Declarations

ACKNOWLEDGMENT

The authors would like to thank the Research Support Foundation of the State of Minas Gerais, the National Council for Scientific and Technological Development and the Coordination for the Improvement of Higher Education Personnel for the financial support. This study was financed in part by the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Financial Code 001.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interests.

AUTHOR CONTRIBUTIONS

All authors contributed equally to the idea and preparation of the manuscript, and all authors read and approved the manuscript.

DATA AVAILABILITY

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

References

1. Conab: Companhia Nacional de Abastecimento, Conab - Página inicial: access in 6/10/2021.
2. Corazza, T., Carvalho, I.R., Silva, J.A.G., Szareski, V.J., Segatto, T.A., Port, E.D., Loro, M.V., Almeida, H.C.F., Oliveira, A.C., Maia. L., Souza, V.Q., 2021. Genetic parameters and multi-trait selection of white oats for forage. Genetics and Molecular Research. 20(1). http://dx.doi.org/10.4238/gmr18451.
3. Silva Júnior, A.C., Silva, M.J., Cruz, C.D., Sant'Anna, I.C., Silva, G.N., Nascimento, M., Azevedo, C.F., 2021. Prediction of the importance of auxiliary traits using computational intelligence and machine learning: A simulation study. PLoS One, 21, p. EMID:a920715476.
4. Silva Junior, A.C., Sant'Anna, I.C., Silva, G.N., Cruz, C.D., Nascimento, M., Lopes, L.B., Soares, P.C., 2022. Computational intelligence and machine learning to study the importance of characteristics in flood-irrigated

rice. Acta Scientiarum-Agronomy (*in prelo*).

5. Parmley, K.A., Higgins, R.H., 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.
6. Garson, G.D., 1991. Interpreting neural network connection weights. Artificial Intelligence Expert. 6:46-51.
7. Goh, A.T.C., 1995. Back-propagation neural networks for modeling complex systems. Artificial Intelligence in Engineering. 9:143-51. http://dx.doi.org/10.1016/0954-1810(94)00011-S.
8. 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. 9:4846. http://dx.doi.org/10.3390/app9224846.
9. González-Camacho, J.M., Campos, G., Pérez, P., Gianola, D., Cairns, J.E., Mahuku, G., Babu, R. and Crossa, J., 2012. Genome-enabled prediction of genetic values using radial basis function neural networks. Theoretical and Applied Genetics. 125(4): 759-771.
10. Matlab (R2016b). 2016. Natick, Massachusetts: The MathWorks Inc.
11. Cruz, C.D. 2016. Genes Software – extended and integrated with the R, Matlab and Selegen. Acta Scientiarum. 38(4):547-552. https://doi.org/10.4025/actasciagron.v38i4.32629.
12. Silva, G.N., Tomaz, R.S., Sant'anna, I.C., Nascimento, M., Bhering, L.L., and 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.
13. Sant'Anna, I.C., Tomaz, R.S., Silva, G.N., Nascimento, M., Bhering, L.L., Cruz, C.D., 2016. Superiority of artificial neural networks for a genetic classification procedure. Genet. Mol. Res. 14, 9898–9906.
14. 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 Caixeita, E.T., 2020. Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. Scientia Agricola. 78: 1–8. https://doi.org/10.1590/1678-992x-2020-0021.
15. Sant'Anna, I.C., Silva, G.N., Nascimento, M., Cruz, C.D., 2020. Subset selection of markers for the genome-enabled prediction of genetic values using radial basis function neural networks. Acta Scientiarum-Agronomy. 43: e46307. https://doi.org/10.4025/actasciagron.v43i1.46307.
16. Rosado, R. D.S., Cruz, C. D., Barili, L. D., Carneiro, J. E. S., Carneiro, V. Q., Silva, J. T., Nascimento, M.2020.Artificial Neural Networks in the Prediction of Genetic Merit to Flowering Traits in Bean Cultivars. Agriculture. 10, 638; doi:10.3390/agriculture10120638.
17. 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. http://dx.doi.org/10.1371/journal.pone.0210531.
18. Santos, R.P., Dean, D.L., Weaver, J.M., and 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. 39:234-245. http://dx.doi.org/ 10.1080/02286203.2018.1558736.
19. Costa, W.G.D., Barbosa, I.P., de Souza, J.E., Cruz, C,D., Nascimento, M., de Oliveira, A.C.B., 2021. Machine learning and statistics to qualify environments through multi-traits in Coffea arabica. PLoS One. 12;16(1):e0245298. http://dx.doi.org/10.1371/journal.pone.0245298.

20. Chagas, C. S., Fernandes Filho, E. I., Vieira, C. A. O., Schaefer, C. E. G. R., Carvalho Júnior, W. 2010. Topographic attributes and Landsat7 data in digital soil mapping using neural networks. Pesquisa Agropecuária Brasileira. 45: 497-507. http://dx.doi.org/10.1590/S0100-204X2010000500009
21. 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.
22. 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. http://dx.doi.org/10.1016/j.geoderma.2017.11.004.
23. Song, H., Liu, A., Li, G., Liu, X., 2021. Bayesian bootstrap aggregation for tourism demand forecasting. IntJ Tourism Res. 1–14. https://doi.org/10.1002/jtr.245314.
24. Tan K., Li, E., Du, Q., Du, P. 2014. An efficient semi-supervised classification approach for hyperspectral imagery. ISPRS Journal of Photogrammetry and Remote Sensing. 97:36–45. http://dx.doi.org/10.1016/j.isprsjprs.2014.08.003.
25. González-Recio, O., Forni, S., 2011. Prediction across the genome of discrete traits using Bayesian regressions and machine learning. Genet Sel Evol. 43:7. https://doi.org/10.1186/1297-9686-43-7.
26. 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.
27. 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. 74(4). http://dx.doi.org/10.1590/1678-992X-2015-0451.
28. 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 et al., 2019. Modeling hyperspectral response of water-stress induced lettuce plants using artificial neural networks. Remote Sens. 11:2797.
29. 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.

Figures

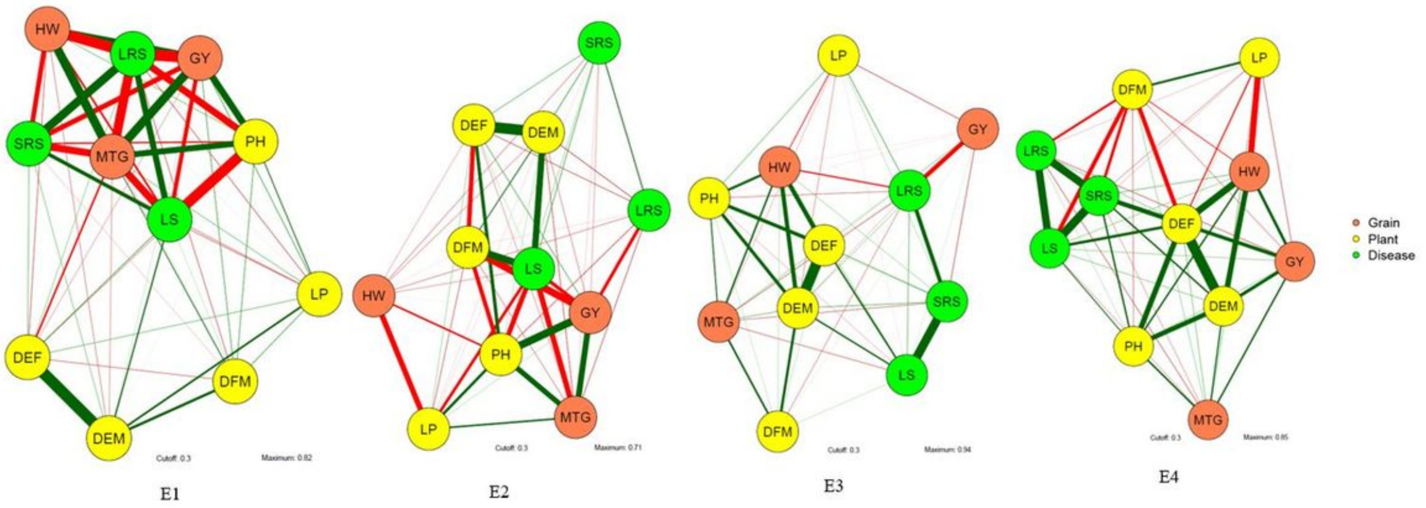

Figure 1

Phenotypic correlation network for the three distinct groups in four environments corresponding to without and with fungicide in two agricultural years, to predict grain yield in white oat (*Avena sativa* L.). The line width is proportional to the strength of the correlation. E1 and E3; E2 and E4 represent the environments without and with fungicide, respectively. The orange color represents the grain characteristics; The yellow color represents the plant characteristics and the green disease severity. MTG = Thousand Grain Mass in grams; HW = Hectoliter Weight; DEM = Days between Emergency and Maturation; PH= percentage of lodging; GY = Grain yield in kilograms per hectare; DEF = Days from Emergence to Flowering; DFM= Days from Flowering to Maturation; PH= Plant Height; LRS= Leaf Rust Severity; SRS=Stem Rust Severity and LS= Leaf Spots.