Smart Agricultural Technology 4 (2023) 100192

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

Drought tolerance classification of grapevine rootstock by machine learning for the São Francisco Valley

Nina Iris Verslype [a,*], André Câmara Alves do Nascimento [b], Rosimar dos Santos Musser [a], Raphael Miller de Souza Caldas [a], Luiza Suely Semen Martins [c], Patrícia Coelho de Souza Leão [d]

[a] Departament of Agronomy, Federal Rural University of Pernambuco, Recife, Brazil
[b] Department of Computing, Federal Rural University of Pernambuco, Recife, Brazil
[c] Department of Biology, Federal Rural University of Pernambuco, Recife, Brazil
[d] Agricultural Research Centre for Semi-arid Tropics, Petrolina, Brazil

ARTICLE INFO

Edited by Stephen Symons.

Keywords:
Hydric stress
Vitis spp.
Climatic changes
Artificial intelligence
Supervised learning
Algorithm

ABSTRACT

Machine Learning (ML) algorithms are increasingly being used in several areas of agricultural studies, such as plant breeding. ML can assist in the recognition of relevant patterns or groups, or even in the prediction of the outcome under new settings, thus accelerating experiments and interpreting their results. The identification and selection of drought-tolerant grapevine rootstock (*Vitis* spp.) have become more relevant in late years, motivated mostly by global climate change scenarios. However, the grapevine is a perennial species, with polygenic characteristics and a complex traits inheritance by offspring, thus making it very challenging to discover new, drought tolerant cultivars. For this reason, this study's main objective was to compare the performance of six machine learning models on the prediction of drought tolerance levels of grapevine rootstock cultivars. A data set with forty-five distinct cultivars was used to evaluate the methods, and the best performing model (AUC 0.9857) was used to predict the drought tolerance class of three cultivars (IAC 313, IAC 572, and IAC 766) whose drought tolerance level was still unknown. The results predicted a high drought tolerance for IAC 313 and IAC 766 cultivars, and a low tolerance for IAC 572.

1. Introduction

Grapevine is considered a crop of notable socioeconomic importance [1], and one of the most valuable perennial crops in the world, due to the high added value and versatility of the products [2]. Climate plays a crucial role in viticulture and is associated with the geography of wine [3]. Currently, it is singled out as one of the most critical aspects that directly interfere with the ripening and the final quality of the grapes to produce a specific style of wine [4,5]. Climate also directly affects the choice of cultivars, planting location, vegetative potential, phytosanitary behavior, yield, and even the management and cultural practices adopted [3–6].

Climate change makes agricultural production extremely vulnerable, especially in terms of water availability, an essential component of life [5,7]. More specifically, one of the most pronounced abiotic stresses for plants is drought, which causes considerable losses in world agricultural production [8]. Thus, a slight climate variation can directly affect the production and quality of grapes [4,3]. Besides, global warming projections indicate a high variation on rainfall patterns and intensity for the next decade [4,3,9,10]. Therefore, mitigating the effects of climate change has become one of the main demands of the winery sector [4,3,1].

The scarcity of water resources emphasizes the importance of using water more efficiently in the winery sector, as most grape-growing regions around the world experience drought at some point [11–13]. Although several approaches can be employed to mitigate the drought problem in viticulture worldwide, using drought-tolerant rootstocks can be one of the most sustainable solutions to improve adaptability of vineyards [10,14].

Drought tolerance in grapevines is a polygenic characteristic, controlled by many genes [15,16]. This hinders the identification and selection of the most promising genotypes, due to the interaction of the

Abbreviations: DT, Decision Tree; KNN, K-Nearest Neighbors; LDA, Linear Discriminant Analysis; RF, Random Forest; SVM, Support Vector Machines; XGB, XGBoost.

* Corresponding author.
E-mail address: nina.verslype@ufrpe.br (N.I. Verslype).

https://doi.org/10.1016/j.atech.2023.100192
Received 10 October 2022; Received in revised form 23 January 2023; Accepted 29 January 2023
Available online 6 February 2023

Table 1
Genetic origin and drought tolerance of the forty-five grapevine rootstock cultivars (*Vitis* spp.) considered in the training data set.

Cultivars	Pedigree	Drought tolerance
VR 039–16	*Vitis vinifera* L. x *Muscadinia rotundifolia* Michaux[1,;5]	Low[2,;7]
VR 043–43	*Vitis vinifera* L. x *Muscadinia rotundifolia* Michaux[1]	Low[3]
101–14 MGt	*Vitis riparia* Michaux x *Vitis rupestris* Scheele[1,;3,;5,;6]	Low[2,;3,;4,;5,;8,;10]
106–8 MGt	*Vitis riparia* Michaux x Cordifolia rupestris de grasset n°1[1]	High[6]
110 R	*Vitis berlandieri* Planchon x *Vitis rupestris* Scheele[1,;5,;8]	High[2,;4,;5,;6,;7,;8,;13]
1103 P	*Vitis berlandieri* Planchon x *Vitis rupestris* Scheele[1,;5,;6,;8]	High[2,;3,;5,;7,;8]
1202 C	*Vitis vinifera* L. x *Vitis rupestris* Scheele[1]	High[7]
125 AA	*Vitis berlandieri* Planchon x *Vitis riparia* Michaux[1]	Low[4]
140 Ru	*Vitis berlandieri* Planchon x *Vitis rupestris* Scheele[1,;6]	High[2,;4,;5,;7,;8,;12,;13]
157–11 C	*Vitis berlandieri* Planchon x *Vitis riparia* Michaux[1]	Medium[4]
1613 C	Solonis (*Vitis riparia* Michaux x *Vitis longii* Prince) x Othello[1,;5]	Medium[9]
161–49 C	*Vitis riparia* Michaux x *Vitis berlandieri* Planchon[1,;6]	Low[5]
1616 C	Solonis x *Vitis riparia* Michaux[1]	Low[2]
196–17 Cl	1203C x *Vitis riparia* Michaux[1]	High[6,;7]
216–3 Cl	1616C x *Vitis rupestris* Scheele[1]	Medium[5,;6,;7]
26 G	*Vitis vinifera* L. x *Vitis riparia* Michaux[1]	Medium[12]
3306 C	*Vitis riparia* Michaux x *Vitis rupestris* Scheele[1]	Medium[8]
3309 C	*Vitis riparia* Michaux x *Vitis rupestris* Scheele[1,;5,;6]	Low[4,;5,;7,;11,;13]
34 EM	*Vitis berlandieri* Planchon x *Vitis riparia* Michaux[1]	Low[4,;5]
41 B MGt	*Vitis vinifera* L. x *Vitis berlandieri* Planchon[1]	High[5,;10]
420 A MGt	*Vitis berlandieri* Planchon x *Vitis riparia* Michaux[1,;3,;5,;6,;9]	Medium[4,;5,;10,;12]
44–53 M	*Vitis riparia* Michaux x Malegue 144[1]	High[2,;5,;6,;7]
5 BB	*Vitis berlandieri* Planchon x *Vitis riparia* Michaux[1,;5,;6]	Medium[2,;4]
5 C	*Vitis berlandieri* Planchon x *Vitis riparia* Michaux[1,;5]	Low[5,;8]
62–66 C	*Vitis vinifera* L. X *Vitis cordifolia* Michaux[1]	High[11]
8 B	*Vitis berlandieri* Planchon x *Vitis riparia* Michaux[1]	Medium[5]
93–5 C	*Vitis vinifera* L. x *Vitis rupestris* Scheele[1]	Low[6]
99 R	*Vitis berlandieri* Planchon x *Vitis rupestris* Scheele[1,;5,;6,;8]	High[8]
Ganzin 1	*Vitis vinifera* L. x *Vitis rupestris* Scheele[1]	High[7]
Dogridge	*Vitis rupestris* Scheele x *Vitis candicans* Engelmann[1,;5]	High[9]
Fercal	*Vitis berlandieri* Planchon x 31R[1]	Medium[2,;5,;7,;8]
Freedom	Fresno 1613–59 x Dogridge[1]	Medium[2,;8]
Golia	Castel 156–12 x *Vitis rupestris* Scheele[1]	Low[4]
Gravesac	161–49C x 3309C[1]	Medium[6,;7,;8]
Harmony	1613C x Dogridge[1]	Medium[2]
Riparia Gloire	*Vitis riparia* Michaux[1,;5,;6]	Low[2,;5,;7,;10,;12,;13]
Rupestris du lot	*Vitis rupestris* Scheele[1,;5,;6,;8]	Medium[5,;7,;9]
Salt Creek	*Vitis champinii* Planchon[1,;5]	High[9,;8]
Schwarzmann	*Vitis riparia* Michaux x *Vitis rupestris* Scheele[1,;5]	Low[5,;8]
SO4	*Vitis berlandieri* Planchon x *Vitis riparia* Michaux[1,;5,;6]	Low[3,;5,;8,;10]
Sori	Solonis x *Vitis riparia* Michaux[1]	Medium[7,;12]
Vitis champini	*Vitis champinii* Planchon[1]	High[13]
IAC 313	Golia x *Vitis cinerea* Engelmann[1,;3]	Unknown*
IAC 572	*Vitis caribaea* De Candolle x 101–14MGt[1,;3]	Unknown*
IAC 766	106–8MGt x *Vitis caribaea* De Candolle[1,;3]	Unknown*

Information obtained in:.
[1] Maul et al. [22].
[2] Sunridge Nurseries [23].
[3] Embrapa Grape and Wine (2016).
[4] Vicopad [24].
[5] Villa [25].
[6] Storm & Krasokhina [26].
[7] Audeguin et al. [27].
[8] Wine Australia [28].
[9] Satisha et al. [29].
[10] ATVB [30].
[11] Chevalier [31].
[12] Rebschule Mueller [32].
[13] Carroll [52].
* No information in the literature.

plant with the environment [8,15,10,16]. The objectives of a rootstock breeding program must be decided effectively, given the large number of possible combinations, as well as that, the requirements to assess a wide variety of aspects, for example, drought tolerance, ease of propagation, compatibility with the graft, longevity of the scion, tolerance to environmental factors, resistance to pests and diseases in the region, as well as other plant characteristics and their influence on fruit.

Machine learning is an efficient methodology that has been increasingly used in several areas of study, such as in plant breeding [17]. Since it makes it possible to identify, predict and classify genotypes according to the needs of the plant breeding program [18,17]. This study focused on use of machine learning algorithms to predict the drought tolerance levels of three Brazilian grapevine rootstocks cultivars (IAC 313, IAC 572, and IAC 766), in order to rank the best drought-tolerant rootstocks to be cultivated in the São Francisco Valley sub-middle (Latitude 9° S, Longitude 40° W) characterized by a semi-arid tropical climate, Brazil [19,20]. These tropical rootstocks were selected in our study due to the lack of information about their drought tolerance degree and importance in the region of interest.

2. Materials and methods

2.1. Plant material

Forty-five grapevine rootstock cultivars were evaluated, listed in Table 1, with their drought tolerance degree and genetic origin, in order to build a data set, which was later used to train a machine learning model. Five instances of each cultivar (i.e., plants) were used as samples, totaling 210 instances in the training data set. Three criteria were adopted to choose these cultivars: (i) these are the most common cultivars in the São Francisco Valley region; (ii) their availability among the accessions from the Active Germplasm Bank of Embrapa [21]; and (iii) the availability of information about their drought tolerance in the literature. The trained model was then used to predict the drought tolerance level of the three cultivars IAC 313 (Tropical), IAC 572 (Jales), and IAC 766 (Campinas).

2.2. Data set characteristics

The data set holds fifty-nine variables, including physiological, biochemical, nutritional, and morpho-agronomic characteristics, manually curated by the authors from scientific articles, books, grapevine nurseries websites, thesis, conferences, and scientific research centers. The collected features are listed below:

2.2.1. Physiological characteristics

Stomatal conductance under unstressed conditions (g_s) and drought stress both in μmol CO2.m-2.s-1; Transpiration rate under unstressed conditions (E) and drought stress both in mmol H2O.m-2.s-1; Photosynthesis rate under unstressed conditions (A) and drought stress both in μmol $m-2.s-1$; Intrinsic Water use efficiency under unstressed conditions (WUE) and drought stress both in μmol CO2 mol-1 H2O; Instantaneous water use efficiency under unstressed conditions (iWUE) and drought stress both in μmol CO2 mmol-1 H2O; Osmotic potential under unstressed conditions (Ψos) and drought stress both in Mpa,

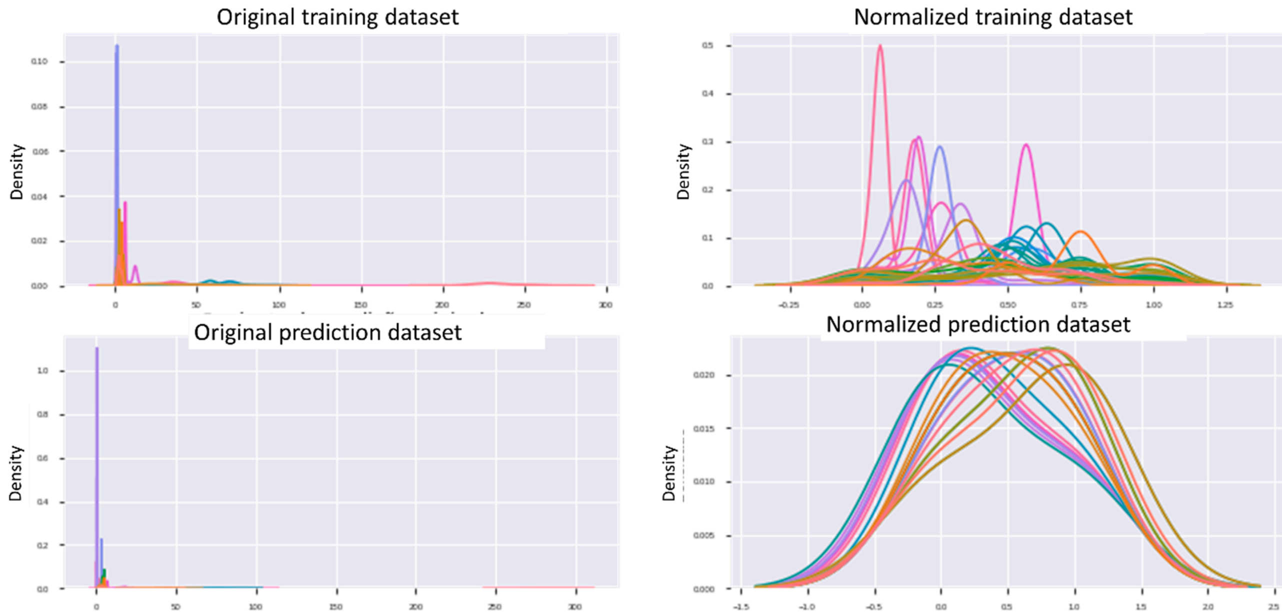

Fig. 1. Data set before and after normalization. Source: Verslype et al. (2021).

Table 2
Hyperparameters analyzed by grid search for each algorithm.

Algorithm	Hyperparameters
DT	Criterion: [gini, entropy]; Max_features [auto, log2, sqrt, None]
RF	N_estimators: [1, 5,10, 100, 1000]; Max_features: [1, 2, 3]
XGB	N_estimators: [10, 100, 500]; Max_features: [auto, log2, sqrt, None]
SVM	C: [0.001, 0.1, 1,10, 20, 100]; Kernel: [rbf, linear]
LDA	N_components: [10, 15, 20, 25, 30]
KNN	N_neighbors: [3, 7, 10]; Weights: [uniform, distance]

Hydraulic conductance (Kl) in kg/MPa-1;

2.2.2. Biochemical characteristics

abscisic acid under unstressed conditions (ABA), drought stress and rehydration condition of the plant all in ng. g^{-1}; Proline under unstressed conditions and drought stress both in mg. g^{-1};

2.2.3. Nutritional characteristics

percentage of nitrogen, phosphorus, and potassium macronutrients content in the leaf petiole; and also, the uptake ability of the plant to absorb nitrogen, phosphorus, and potassium, on a rating scale assigned between 1 and 5 (high = 5, medium-high = 4, medium = 3, medium-low = 2 and low = 1);

2.2.4. Morpho-agronomic characteristics

stomata density per mm^2; flower sex (male, hermaphrodite or female); grape buds, grape maturity and leaf fall in days; active limestone tolerance in percentage; wood production per plant; root distribution scaled in a range from 1 to 6 (very deep=6, deep=5, moderate-deep=4, moderate=3, shallow=2, very shallow=1); geotropic angle in degree; percentage root distributions to 60 cm and 100 cm; Chlorotic Power Index (CPI) in percentage; Drought tolerance scaled in a range from 1 to 3 (high =3, medium =2 and low = 1); Vegetative cycle (early, intermediate, late); and the variables listed below a scale of scores ranging from 1 to 5 (high=5, medium-high=4,medium=3, low-medium=2 and low=1) were assigned, for Anthracnose, Downy mildew, Fusarium, Phylloxera and Nematode resistance; lime and total limestone tolerance; ease of rooting; ease of branch-grafting; acid, sandy, wet, clay, salinity, calcareous and compactness soil tolerance; vigor; Iron chlorosis tolerance.

2.3. Data analysis methodology

All analysis were performed using the Python language on the Google Colaboratory[1] platform.

2.3.1. Pearson correlation analysis

Pearson's correlation coefficient (r) was calculated among the considered variables, using the corr() function available in the pandas library [33], in order to examine the correlation between the variables and discard highly correlated characteristics. Pearson's correlation coefficients are calculated according to Eq. (1) [34].

$$r = \left(\sum xy - \left(\sum x * \sum y \big/ N\right)\right) \Big/ \sqrt{\left(\sum x^2 - \left(\left(\sum x\right)^2 \big/ N\right)\right) * \left(\sum y^2 - \left(\left(\sum y\right)^2 \big/ N\right)\right)}, \tag{1}$$

where r corresponds to the correlation coefficient that can vary between −1 and +1; N is the number of observations; x and y are the values of both variables.

Subsequently, a heat map chart, available in the seaborn library [35], was generated to improve the visualization of Pearson's correlation coefficients.

2.3.2. Data set pre-processing

Pre-processing occurred in the same way for the training set, with 42

[1] https://colab.research.google.com/

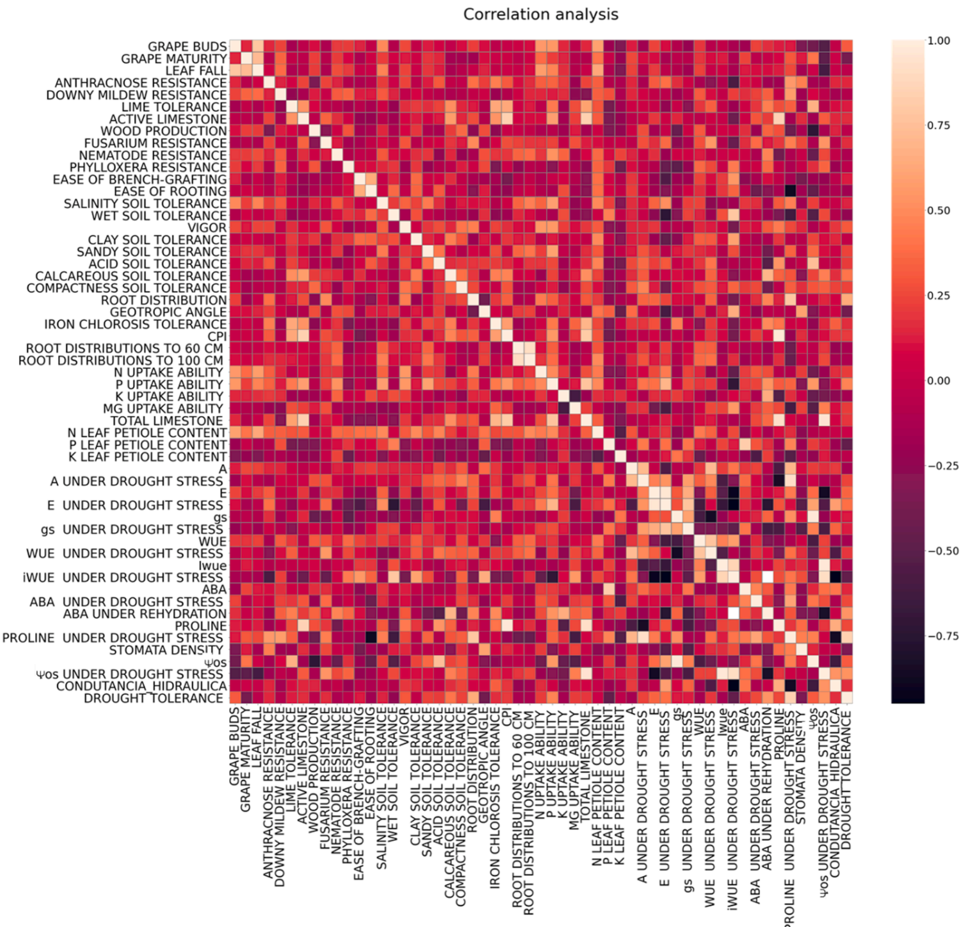

Fig. 2. Correlation coefficient analysis between the sixty-one variables evaluated.

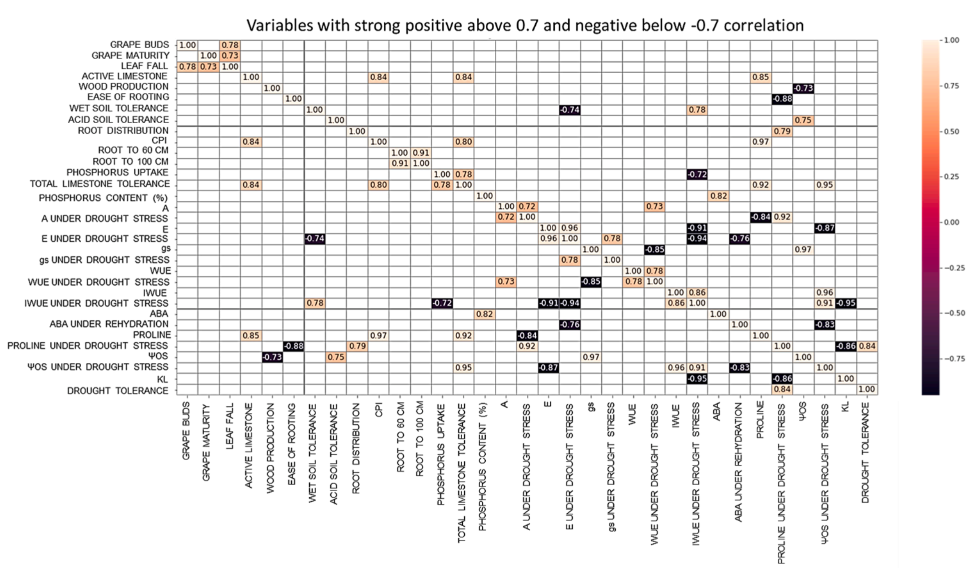

Fig. 3. Variables with strong positive and negative correlation. Source: Verslype et al. (2021).

cultivars, and the prediction set, containing three cultivars (IAC 313, IAC 572, and IAC 766). For this, the missing values were filled in by the mean of each variable column, then duplicated rows were removed and finally, normalization was performed using MinMaxScaler, Fig. 1), whose transformation is given by the Eq. (2) and ((3), available in the scikit-learn library [36].

$$X_{std} = \frac{(x - x.\min)}{(x.\max \ - \ x.\min)} \tag{2}$$

$$X_{scaled} = X_{std} \times (\max - \min) + \min, \tag{3}$$

Table 3
Accuracy, precision (P), recall (R), and f1-score obtained by 10-fold cross-validation evaluation on the six algorithms to predict drought tolerance classes for grapevine rootstock cultivars.

Algorithm	Accuracy (%)	P	R	f1-*score*
DT	82.86	0.86	0.83	0.83
RF	**98.10**	**0.98**	**0.98**	**0.98**
XGB	96.67	0.97	0.97	0.97
SVM	87.14	0.89	0.87	0.87
LDA	64.29	0.67	0.64	0.65
KNN	85.71	0.89	0.86	0.86

where: min, max corresponds to the sample range.

2.3.3. Comparison of supervised learning algorithms

Six algorithms were considered to evaluate their ability to predict drought tolerance classes in grapevine rootstocks, namely: *Decision Tree* (DT) [37], *Random Forest* (RF) [38], *K-Nearest Neighbors* (KNN) [39], *XGBoost* (XGB) [40], *Support Vector Machines* (SVM) [41] and *Linear Discriminant Analysis* (LDA) [42]. These algorithms were select because they are easily accessible by the public, and all algorithms are available in the scikit-learn library [36], except for XGB available in the xgboost library [40]. To determine the best algorithm in our setting, a cross-validation evaluation procedure was performed, and grid search analysis Table 2), available in the scikit-learn library. [36]. The precision rate (P), recall (R), accuracy, and f1-score metrics obtained by all models in the considered data set were evaluated. Are described these evaluation metrics in Eqs. (4), (5, (6), and (7) [36,43,44].

$$P = \frac{TP}{(TP + FP)}, \tag{4}$$

where: *P* represents the precision, *TP* indicates the number of true positives, and *FP* the number of false positives.

$$R = \frac{VP}{(TP + FN)}, \tag{5}$$

where: *R* represents *recall*, *TP* is the number of true positives, and *FN* the number of false negatives.

$$accuracy = (TP + VN)/(TP + TN + FP + FN), \tag{6}$$

where: *TP* is the number of true positives, *TN* the number of true negatives, *FP* the number of false positives, and *FN* the number of false negatives.

$$f1 = 2 \times (precis \sim ao \times recall)/(precis \sim ao + recall), \tag{7}$$

where: *f1* represents that obtained by the harmonic mean between recall and accuracy.

2.3.4. Drought tolerance prediction

A predictive drought tolerance analysis was performed on the three Brazilian grapevine rootstocks cultivars IAC 313, IAC 572, and IAC 766, using the most efficient algorithm and its respective hyperparameter configuration, provided by the grid search (Table 2).

3. Results and discussion

The Pearson's correlation coefficient analysis (Fig. 2) indicated 38 cases with a strong correlation between the 61 variables evaluated in this study, of which 12 had a negative correlation close to −1, and 26 had a positive correlation close to 1.

In this regard, correlation analysis is interesting since it allows one to explain and determine the degree of relationship between the variables in the data set [34]. Thus, coefficients close to 1 and −1 make it possible to obtain a good prediction of one variable from the other [45]. Besides,

Table 4
Evaluation of the best hyperparameters for each model obtained by *Grid search* to predict drought tolerance classes for grapevine rootstock cultivars.

Algorithm	Best result (%)	Best parameter	Hyperparameters best configuration
SVM	83.33	{'C': 100, 'kernel': 'rbf'}	SVC (C = 100, break_ties=False, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape='ovr', degree=3, gamma='auto', kernel='rbf', max_iter=−1, probability=False, random_state=None, shrinking=True, tol=0.001, verbose=False)
RF	98.57	{'max_features': 2, 'n_estimators': 1000}	RandomForestClassifier (bootstrap=True, ccp_alpha=0.0, class_weight=None, criterion='gini', max_depth=None, max_features=2, max_leaf_nodes=None, max_samples=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=1000, n_jobs=None, oob_score=False, random_state=None, verbose=0, warm_start=False)
KNN	90.48	{'n_neighbors': 3, 'weights': 'distance'}	KNeighborsClassifier (algorithm='auto', leaf_size=30, metric='minkowski', metric_params=None, n_jobs=None, n_neighbors=3, p = 2, weights='distance')
XGB	96.19	{'max_features': 'auto', 'n_estimators': 500}	XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1, gamma=0, learning_rate=0.1, max_delta_step=0, max_depth=3, max_features='auto', min_child_weight=1, missing=None, n_estimators=500, n_jobs=1, nthread=None, objective='multi:softprob', random_state=0, reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None, silent=None, subsample=1, verbosity=1)
LDA	67.14	{'n_components': 10}	LinearDiscriminantAnalysis (n_components=10, priors=None, shrinkage=None, solver='svd', store_covariance=False, tol=0.0001)
DT	86.19	{'criterion': 'entropy', 'max_features': None}	DecisionTreeClassifier (ccp_alpha=0.0, class_weight=None, criterion='entropy', max_depth=None, max_features=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, presort='deprecated', random_state=None, splitter='best')

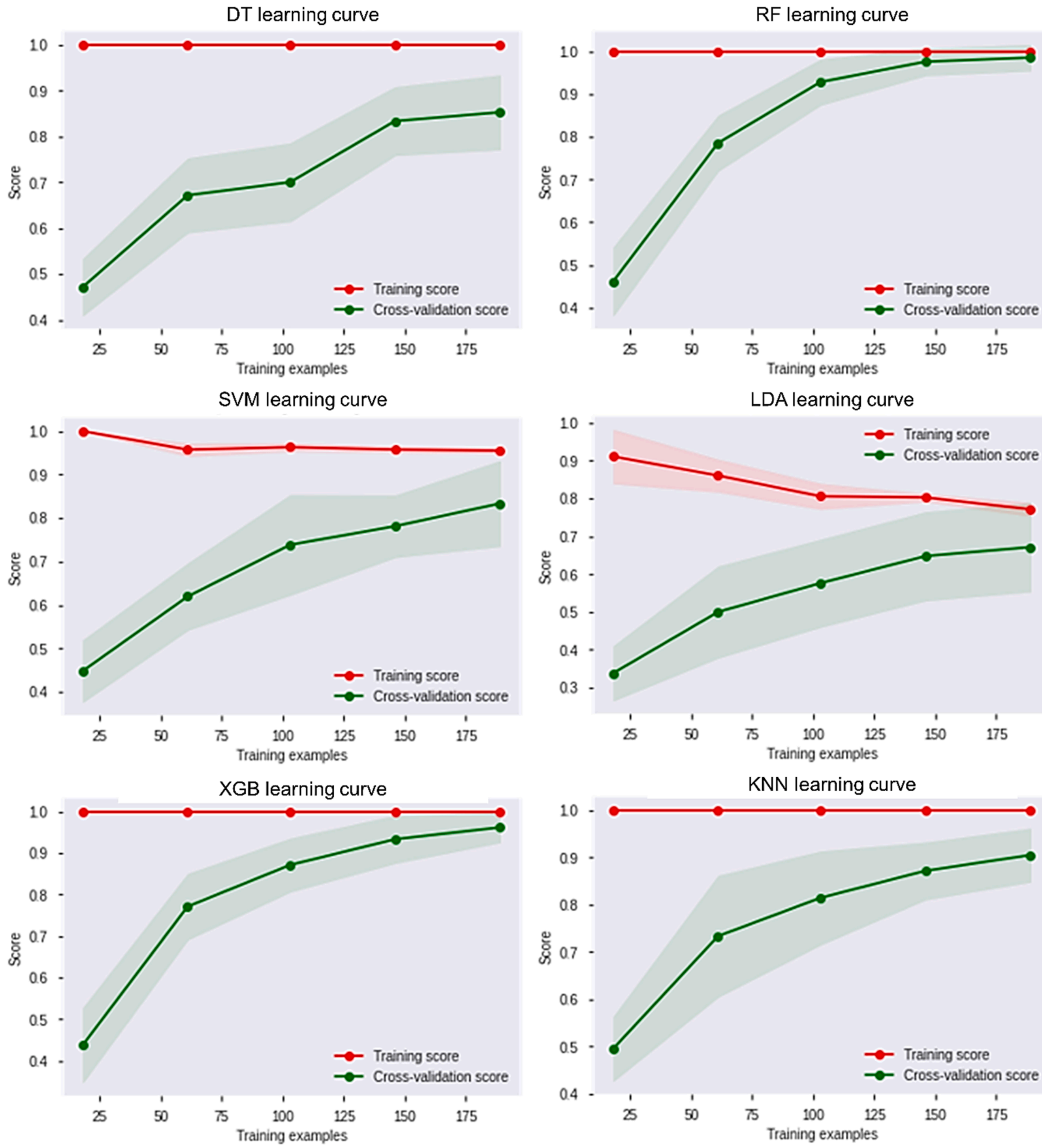

Fig. 4. Learning curve of the six algorithms on the data set, through 10-fold cross-validation evaluation, and score metric (R^2) to predict drought tolerance classes for grapevine rootstock cultivars. Where: DT - *Decision Tree*, RF – *Random Forest*, XGB - *XGBoost*, SVM – *Support Vector Machines*, LDA - *Linear Discriminant Analysis* e KNN - *K-Nearest Neighbors*. Source: Verslype et al. (2021).

Table 5
RF algorithm classification report for 10-fold cross-validation.

Class	Precision	Recall	f1-*score*	Support
High	1.00	0.99	0.99	70
Low	0.96	1.00	0.98	70
Medium	1.00	0.97	0.99	70
Accuracy			0.99	210
Macro average	0.99	0.99	0.99	210
Weighted average	0.99	0.99	0.99	210

redundant features can be detected, for example, the root distributions to 100 cm and 60 cm can be highlighted among the strong correlations, due to their positive correlation ($r = 0.91$). This finding indicates that there no need to evaluate both features, resulting in cost savings and a significant reduction in analytical time. Another pair of features that this analysis pointed as highly correlated were between the grape buds and grape maturity variables, as they presented, respectively ($r = 0.78$) and ($r = 0.73$) of positive correlation with the variable onset of leaf fall, demonstrating that only the evaluation of the variable early leaf fall would be sufficient.

The proline under drought stress conditions has a high correlation with drought tolerance ($r = 0.84$). This interaction is important since proline could be used as an indirect selection criterion for the search for drought-tolerant varieties. As drought tolerance is a polygenic trait which is difficult to identify and select for superior materials in plant breeding programs ([10,16]; Cantu & Walker, 2019). This high correlation could be explained by the fact that proline is considered an osmotically active substance, which is accumulated in high levels in the cytoplasm when the grapevine is in water restriction periods, thus enabling it to maintain balance the water potential within the plant cell and the turgor pressure ([16]; Keller, 2015; [46]).

Thereby analyzing the variables with high correlation (Fig. 3), 14 of them were discarded, namely: E under drought stress, A under drought

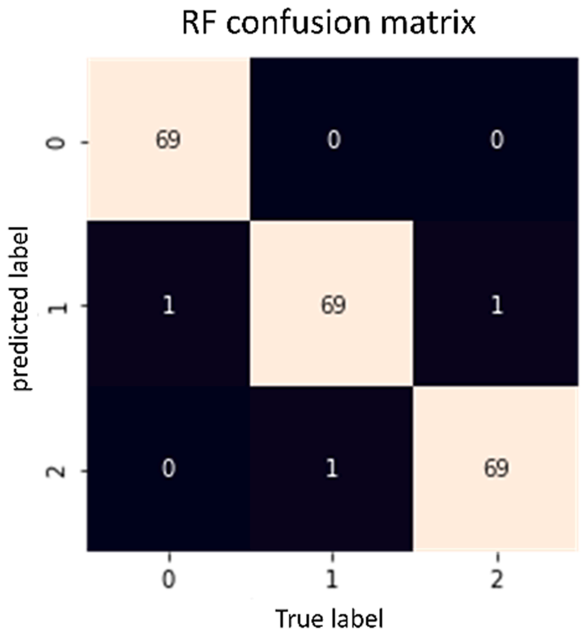

Fig. 5. RF algorithm confusion matrix. Where 0 – represents the high drought tolerance class, 1 – low, and 2 – medium. Source: Verslype et al. (2021).

stress, EUA under drought stress, EiUA under water stress, proline under unstressed and drought stress conditions, ABA content under unstressed conditions, osmotic potential under unstressed and drought stress conditions, IPC, total limestone tolerance, root distribution to 100 cm, grape buds and grape maturity.

Performance differences in the six algorithms on the training data set were identified (Table 3). The accuracy metric range was between 98.10% and 64.29%, precision between 98% and 67%, and recall between 98% and 64%. For Skansi [44], these evaluation metrics are an important cue to obtain a meaningful performance comparation of a set of machine learning classifiers. Since the accuracy allows identifying how good the classifier is, in average, in the task to label unseen instances, the precision determines the model's ability to avoid erroneous results like not labeling a positive sample as negative, and the recall indicates the success of the model to find all positive samples [36,44].

In this sense, the cross-validation experiments have shown that the RF algorithm as is the best classifier. Due to the highest mean values achieved for accuracy (98.10%), precision (98%), and recall (98%). Nevertheless, the LDA classifier had the lowest accuracy (64.29%), precision (67%), and recall (64%) values among all the models evaluated, indicating a low rate of correctness in the indication of the three drought tolerance classes.

Grid search analysis also was applied to the six algorithms (Table 2) to indicate the best performance algorithm to predict drought tolerance classes. The results obtained confirm that the LDA had the worst performance, reaching only 67.14% accuracy as the best possible result for all the configurations of hyperparameters tested on the data set and consequently the lowest learning curve. Meanwhile, the RF and XGB algorithms performed the best results with 98.57% and 96.19% correctness to predict the three drought tolerance classes (Table 4). The best parameter and its respective configuration of hyperparameters and the learning curve plotted in Fig. 4 indicate a higher learning rate over time.

Table 5 presents the performance of the best model, stratified by class. The RF algorithm achieved a f1-score of 0.99 for the high tolerance class, 0.98 for low, and 0.99 for medium drought tolerance. According to Pedregosa *et al.* [36], The f1-score values near to one indicate a better performance, while those near zero indicate the worst score in the model. In this sense, these results indicate a high success rate to classify each of the three drought tolerance classes by the RF algorithm.

A confusion matrix consists of a visualization of the number of erroneously labeled and correctly labeled results, represented in the matrix diagonal [44]. The confusion matrix for the best model can be seen in (Fig. 5), demonstrating a low rate of erroneously classified results by the RF algorithm, which shows itself as a good classifier for our problem.

The RF algorithm assigned a rating scale of importance to all the features evaluated in the data set. Among these, we observed that the variables cutting production, the onset of leaf fall, anthracnose resistance, geotropic angle, calcareous soil tolerance, and root distribution up to 60 cm were the first variables with the notable importance in their decision-making, as we can see it in (Fig. 6).

The RF decision-making high importance attribution to wood production per plant can be explained by the fact plants with more efficient use of water produce more dry matter per gram of transpired water. Given the need for an average intake of 1 liter for each 2 gs of grapevine dry mass-produced [47,16]. Consequently, these more vigorous plants can tolerate conditions with lower water availability in general [47,25]. While, the feature of leaf fall is related to drought tolerance, as at this stage the grapevines direct the photoassimilates to the roots. This enables the absorption area = to increase and recover from water stress [47].

The greater importance attributed to the variables roots distribution up to 60 cm and geotropic angle in RF decision-making can be explained

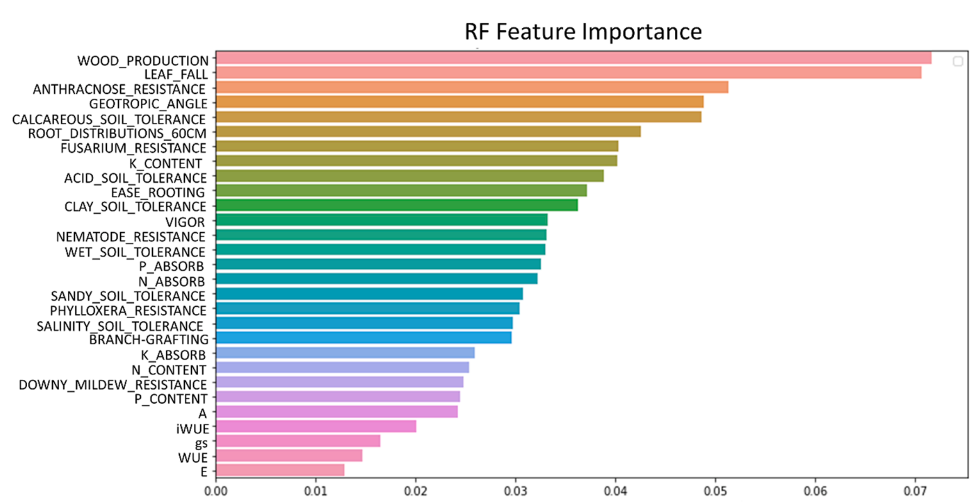

Fig. 6. Feature importance to the RF algorithm decision making to predict drought tolerance classes for grapevine rootstock cultivars. Source: Verslype et al. (2021).

by the geotropic angle that determines the depth capacity of the grapevine root system [25]. In this sense, the literature has evidenced the closed geotropic angles allow the deepening of the roots and consequently make it possible to reach deeper layers of soil remain moist, as well as to avoid the absorption of harmful elements to plants in saline soils through selective absorption [29,16,48,25]. These features enable rootstocks to better adapt to water stress conditions, and consequently, a greater drought tolerance [29,25].

The RF model also assigned a high importance to the variable anthracnose resistance, considered a disease with a wide impact on grapevines, as it mainly affects young and tender tissues in the aerial part of the plant [49]. The drought stress associated with anthracnose predisposition occurrence was studied by Erbaugh et al. [50] for the *Cornus florida* L. species, which observed a significant increase in disease severity in plants shaded and under drought conditions. In addition, Hsiang [51] mentions how the *Poa annua* L. species becomes unable to respond quickly to anthracnose infection and overcome the disease when subjected to severe stresses such as cyclical water stress. In this sense, it may be a signal of some degree of interference between the two variables in grapevines.

As mentioned above, after evaluating accuracy, precision, recall, and f1-score metrics, the RF was identified as the best model produced in the scenarios and data sets test to predict drought tolerance classes in grapevine rootstocks. The prediction of drought tolerance classes indicated by the RF algorithm was high tolerance for cultivars IAC 313 and IAC 766, and a low tolerance for cultivar IAC 572. Despite being an initial study, this approach with supervised algorithms proved to be a helpful and accessible strategy for the breeder, that can help develop these cultivars by predicting, identifying and selecting promising genotypes. We would like to acknowledge some limitations of the present work, which still needs validation of the predicted classes for the IAC 313, 766 and 572 through studies in field conditions under drought conditions. Besides, the training data set is rather a limited sample of grapevine rootstocks. In future works, we intend to increase the size of the training data set or explore data augmentation strategies.

4. Conclusion

This paper proposed a pipeline to predict rootstock drought tolerance for the São Francisco Valley in Brazil, through the use of machine learning methods. A manually curated data set was produced, which was used to evaluate several machine learning algorithms. The results indicated that the best performing classifiers were RF, followed by XGB with 98.57% and 96.19% correctness to predict drought tolerance classes of grapevine rootstocks.

The best RF hyperparameter configuration was max_*features=2* and the number of estimators equal to 1000 for the evaluated scenarios and data sets. While, the LDA classifier had the lowest efficiency, with a hit rate of 67.14%, to predict the classes of tolerance to water deficit.

The trained model was used to predict drought tolerance of 03 the grapevine rootstock cultivars that have never been experimentally evaluated. The model classified grapevine rootstock cultivars IAC 313 and IAC 766 with high drought tolerance, while IAC 572 with a low tolerance. In this regard, cultivars IAC 313 and IAC 766 could be the best option for vintners in the São Francisco Valley. It is important to note that this is a preliminary study, and these predictions, are an indicative of the real drought tolerance of the three cultivars, requiring more research to validate such results.

Machine learning algorithms demonstrated to be a helpful tool in plant breeding studies that can contribute to identifying and selecting drought-tolerant grapevine rootstock genotypes in breeding programs. As a suggestion to extend this study, we envision the validation of the predictions obtained by the ML model in field tests under drought stress, as well as an increase of the training data set.

Declaration of Competing Interest

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

Data availability

The supplementary material contains the reference list used to assemble the data set.

Acknowledgments

We would like to thank the National Council for Scientific and Technological Development (CNPq), for granting a master's scholarship to the first author.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.atech.2023.100192.

References

[1] L. Zhang, E. Marguerit, L. Rossdeutsch, N. Ollat, G.A. Gambetta, The influence of grapevine rootstocks on scion growth and drought resistance, Theoret. Experiment. Plant Physiol. 28 (2016) 143–157.

[2] M.A. Walker, C. Heinitz, S. Riaz, J. Uretsky, Grape Taxonomy and Germplasm, in: D Cantu, MA Walker (Eds.), The Grape Genome, Springer, Switzerland, 2019, pp. 25–38.

[3] G.V. Jones, R. Reild, A. Vilks, Climate, grapes, and wine: structure and suitability in a variable and changing climate, in: Percy H Dougherty (Ed.), The Geography of Wine: regions, Terroir and Technique, Springer, Netherlands, 2012, pp. 109–136.

[4] H. Fraga, A.C. Malheiro, J. Moutinho-Pereira, J.A. Santos, An overview of climate change impacts on European viticulture, Food Energy Security 1 (2013) 94–110.

[5] G.V. Jones, Grapevines in a changing environment: a global perspective, in: H Gerós, MM Chaves, HM Gil, S Delrot (Eds.), Grapevine in a Changing environment: a Molecular and Ecophysiological Perspective, John Wiley & Sons, UK, 2016, pp. 23–34.

[6] M.S.B. Moura, A.H.C. Teixeira, J.M. Soares, Exigências climáticas, in: J.M. Soares, P.C.S. Leão (Eds.), A Vitivinicultura No Semiárido Brasileiro, Embrapa Informação Tecnológica, Petrolina, Embrapa Semi-Árido, Brasília, DF, 2009, pp. 36–69.

[7] R. Kathpalia, S.C. Bhatla, Plant Water Relations, in: SC Bhatla, MA Lal (Eds.), Plant physiology, Development and Metabolism, Springer, Singapore, 2018, pp. 3–36, p.

[8] M. Ashraf, Inducing drought tolerance in plants: recent advances, Biotechonol. Adv. 28 (2010) 169–183.

[9] G.V. Jones, L.B. Webb, Climate change, viticulture, and wine: challenges and opportunities, J. Wine Res. 21 (2010) 103–106.

[10] E. Marguerit, O. Brendel, E. Lebon, &.C. Van Leeuwen, N. Ollat, Rootstock control of scion transpiration and its acclimation to water deficit are controlled by different genes, New Phytol. 194 (2012) 416–429.

[11] M.M. Chaves, O. Zarrouk, J.M. Costa, T. Santos, A.P. Regalado, M.L. Rodrigues, C. M. Lopes, Grapevine under deficit irrigation: hints from physiological and molecular data, Annals Appl. Biol. 1 (2010) 661–676.

[12] K. Fort, H. Fraga, D. Grossi, M.A. Walker, Early measures of drought tolerance in four grape rootstocks, J. Amer. Soc. Hort. Sci. 1 (2017) 36–46.

[13] O. Zarrouk, J.M. Costa, R. Francisco, C. Lopes, M.M. Chaves, Drought and water management in Mediterranean vineyards, in: H Gerós, MM Chaves, HM Gil, S Delrot (Eds.), Grapevine in a Changing environment: a Molecular and Ecophysiological Perspective, John Wiley & Sons, UK, 2016, pp. 67–88.

[14] N. Ollat, A. Peccoux, D. Papura, D. Esmenjaud, E. Marguerit, J.P. Tandonnet, L. Bordenave, S.J. Cookson, F. Barrieu, L. Rossdeutsch, J. Lecourt, V. Lauvergeat, P. Vivin, P.F. Bert, S. Delrot, Rootstocks as a component of adaptation to environment, in: H. Gerós, M.M. Chaves, H.M. Gil, S. Delrot (Eds.), Grapevine in a Changing environment: a Molecular and Ecophysiological Perspective, John Wiley & Sons, UK, 2016, pp. 102–152.

[15] S. Dayer, I. Reingwirtz, A.J. McElrone, G.A. Gambetta, Response and recovery of grapevine to water deficit: from genes to physiology, in: D. Cantu, M. Walker (Eds.), The Grape Genome. Compendium of Plant Genomes, Springer, Switzerland, 2019, pp. 223–245.

[16] I. Serra, A. Strever, P.A. Myburgh, A. Deloire, The interaction between rootstocks and cultivars (*Vitis vinifera* L.) to enhance drought tolerance in grapevine, Aust. J. Grape Wine Res. 20 (2014) 1–14.

[17] A. Etminan, A. Pour-Aboughadareh, R. Mohammadi, L. Shooshtari, M. Yousefiazarkhanian, H. Moradkhani, Determining the best drought tolerance indices using artificial neural network (ANN): insight into application of intelligent

agriculture in agronomy and plant breeding, Cereal Res. Commun. 47 (2019) 170–181.
[18] A.H. Beiki, Molecular predicting drought tolerance in maize inbred lines by machine learning approaches, Biorxiv 10 (2019) 1–33.
[19] U.A. Camargo, F.D. Amorim, Análise dos atuais sistemas de produção de uvas para vinho no Vale do São Francisco, in: Workshop Internacional De Pesquisa 1, 2004, pp. 97–101.
[20] P.C.S. Leão, D.J. Silva, Cultivo da videira no Semiárido brasileiro, in: R Pio (Ed.), Cultivo De Fruteiras De Clima Temperado Em Regiões Subtropicais e Tropicais, UFLA, Lavras, 2014, pp. 578–618.
[21] Camargo, U.A., Ritschel, P.S., Maia, J.D.G., Quecini, V., Machado, C.A.E., Bosco, D. D., Sinski, I., Comachio, &.V., Zilio, R. (2017). Banco ativo de germoplasma de uva. Embrapa Uva e Vinho, Bento Gonçalves. Retrieved from https://www.embrapa.br/en/uva-e-vinho/banco-ativo-de-germoplasma-de-uva/pesquisa. Accessed September 16, 2020.
[22] E. Maul, R. Töpfer, F. Röckel, U. Brühl, M. Hundemer, A. Mahler-Ries, M. Walk, S. Kecke, A. Wolck, A. Ganesch, *Vitis* International Variety Catalogue. Federal Research Centre for Cultivated Plants, Institute for Grapevine Breeding, 2020. Retrieved from, https://www.vivc.de/. >. Accessed July 30, 2020.
[23] Sunridge Nurseries (2020). Rootstock chart. Bakersfield, California. Retrieved from https://www.sunridgenurseries.com/. Accessed August 25, 2020.
[24] VICOPAD - Vivai Cooperativi di Padergnone (2020). Scelta del portinnesto. Padergnone, Italy. Retrieved from http://www.vicopad.it/. Accessed August 19, 2020.
[25] P. Villa, Cultivar La Vid, Editorial de Vecchi, USA, 2018, p. 160.
[26] Storm, B., & Krasokhina, C. (2020). Rootstock grape varieties, gardening varieties, wild grapes, breeding varieties. Retrieved from https://vinograd.info/. Accessed August 5, 2020.
[27] L. Audeguin, V. Bécart, E. Besnard, J.M. Boursiquot, G. Ciccolini, S. Dedet, G. Delorme, J. Follet, A. Froelhy, L. Garcin, N. Gonin, T. Lacombe, F. Laffargue, C. Marchal, M. Pesato, C. Sereno, G. Uriel, M. Wentzel, O. Yobregat, Pl@ntGrape: Le Catalogue Des Vignes Cultivées En France, IFV – INRA – Montpellier SupAgro, Montpellier, France, 2020. Retrieved from, http://plantgrape.plantnet-project.org/. Accessed September 15, 2020.
[28] Wine Australia (2016). Grapevine rootstock selector tool. Australia. Retrieved from https://grapevinerootstock.com/. Accessed August 25, 2020.
[29] J. Satisha, G.S. Prakash, G.S.R. Murti, K.K. Upreti, Response of grape rootstocks to soil moisture stress, J. Horticult. Sci. 1 (2006) 19–23p.
[30] ATVB – Association Technique Viticole de Bourgogne (2013) Caractéristiques des porte-greffes et greffons utilisés en Côte-d'Or. Retrieved from http://www.atvb-bourgogne.com/. Accessed August 27, 2020.
[31] A. Chevalier, L'amélioration de la vigne en France et les Travaux de G. Couderc sur l'Hybridation et le Greffage, Revue de Botanique Appliquée et d'agriculture Coloniale 52 (1925) 926–945.
[32] Rebschule Mueller (2020). Unterlagen für den weinbau. Retrieved from http://www.rebschulemueller.de/. Accessed August 15, 2020.
[33] W. McKinney, Data structures for statistical computing in python, in: Stéfan van der Walt, Jarrod Millman (Eds.), Proceedings of the 9th Python in Science Conference, 2010, pp. 56–61, pages.
[34] P.V. Ferreira, Estatística Experimental, Ed.UFAL, Maceió, 2011, p. 559.
[35] M. Waskom, Seaborn: statistical data visualization, Open J. 60 (2021) 3021, https://doi.org/10.21105/joss.03021.
[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. BlondeL, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine Learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[37] L. Breiman, J. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, CRC press, Wadsworth, Belmont, CA, 1984.
[38] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[39] E. Fix, J.L. Hodges, Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.
[40] T. Chen, C. Guestrin, XGBoost: a Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, ACM, 2016, pp. 785–794.
[41] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[42] R.C. Rao, The Utilization of Multiple Measurements in Problems of Biological Classification, J. Roy. Statist. Soc. Ser. C 10 (1948) 159–203.
[43] S. Shalev-Shwartz, S. Ben-David, Understanding Machine learning: from Theory to Algorithms, Cambridge university press, New York, 2014, p. 409.
[44] S. Skansi, Introduction to Deep Learning: from Logical Calculus to Artificial Intelligence, Springer, 2018, p. 196.
[45] D. Forsyth, Probability and Statistics For Computer Science, Springer International Publishing, 2018, pp. 36–42.
[46] L. Taiz, E. Zeiger, I.M. Møller, A. Murphy, Fisiologia e Desenvolvimento Vegetal, Artmed Editora, Porto Alegre, 2017, p. 888.
[47] H. Ojeda, A. Deloire, Z. Wang, A. Carbonneau, Detereminación y control del estado hídrico de la vid: efectos morfológicosy fisiológicos de la restricción hídrica en vides, Viticultura/Enología Profesional 90 (2004) 27–43.
[48] D.R. Smart, E. Schwass, A. Lakso, L. Morano, Grapevine rooting patterns: a comprehensive analysis and a review, Am. J. Enol. Vitic. 57 (2006) 89–104.
[49] M.R. Lima, D.B. Lopes, S.C.C.H. Tavares, D.L. Tessmann, N.F. Melo, Doenças e alternativas de controle, in: J.M. Soares, P.C.S. Leão (Eds.), A Vitivinicultura No Semiárido Brasileiro, Embrapa Informação Tecnológica, Petrolina, Embrapa Semi-Árido, Brasília, DF, 2009, pp. 543–596.
[50] D.K. Erbaugh, M.T. Windham, A.J. Stodola, R.M. Augé, Light intensity and drought stress as predisposition factors for dogwood anthracnose, J. Environ. Hortic. 13 (4) (1995) 186–189.
[51] T. Hsiang, Anthracnose: terrorizing turf grass, GCSAA (2006) 2, p. Retrieved from, https://www.gcsaa.org/. Accessed May 14, 2020.
[52] B. Carroll, Characteristics of *Vitis* grape Rootstocks and Adaptability, Oklahoma State University, Oklahoma, 2016, p. 4.