European Journal of Forest Research (2024) 143:1575–1591
https://doi.org/10.1007/s10342-024-01716-7

RESEARCH

Using machine learning algorithms to cluster and classify stone pine (*Pinus pinea* L.) populations based on seed and seedling characteristics

Servet Caliskan[1] · Elif Kartal[2] · Safa Balekoglu[1] · Fatma Çalışkan[3]

Received: 20 February 2024 / Revised: 15 June 2024 / Accepted: 21 June 2024 / Published online: 29 June 2024
© The Author(s) 2024

Abstract

The phenotype of a woody plant represents its unique morphological properties. Population discrimination and individual classification are crucial for breeding populations and conserving genetic diversity. Machine Learning (ML) algorithms are gaining traction as powerful tools for predicting phenotypes. The present study is focused on classifying and clustering the seeds and seedlings in terms of morphological characteristics using ML algorithms. In addition, the k-means algorithm is used to determine the ideal number of clusters. The results obtained from the k-means algorithm were then compared with reality. The best classification performance achieved by the Random Forest algorithm was an accuracy of 0.648 and an F1-Score of 0.658 for the seed traits. Also, the best classification performance for stone pine seedlings was observed for the k-Nearest Neighbors algorithm (k = 18), for which the accuracy and F1-Score were 0.571 and 0.582, respectively. The best clustering performance was achieved with k = 2 for the seed (average Silhouette index = 0.48) and seedling (average Silhouette Index = 0.51) traits. According to the principal component analysis, two dimensions accounted for 97% and 63% of the traits of seeds and seedlings, respectively. The most important features between the seed and seedling traits were cone weight and bud set, respectively. This study will provide a foundation and motivation for future efforts in forest management practices, particularly regarding reforestation, yield optimization, and breeding programs.

Keywords Afforestation · Breeding · Machine learning · Phenotyping · Supervised learning · Unsupervised learning

Introduction

Pine nuts are essential for pine reproduction and afforestation (Huang et al. 2022). Stone pine *(Pinus pinea* L.) is a forest tree of considerable ecological value, and its non-wood products are of significant value in Mediterranean forests (Fig. 1a and b) (Calama et al. 2016). Nuts of stone pines are among the key components of Mediterranean forest ecosystems and have been cultivated by Mediterranean peoples for millennia. Its distinctive umbrella-shaped appearance and frequent cultivation for ornamental reasons contribute to its prominence in the Mediterranean region. Harvested for their edible nuts, the cones of stone pines have been a part of human consumption since the Paleolithic era (Mutke et al. 2005). *Pinus pinea* nuts are known for their elevated market value, rendering them an appealing crop choice due to their profitability. This species exhibits resilience under adverse conditions, thriving in impoverished or eroded soils. Its natural resistance to pests and diseases minimizes the necessity for intensive cultivation practices, while its remarkable drought resistance (Çalışkan and Boydak 2017) qualifies it

Communicated by Marta Pardos.

✉ Safa Balekoglu
safa.balekoglu@iuc.edu.tr

Servet Caliskan
servetc@iuc.edu.tr

Elif Kartal
elifk@istanbul.edu.tr

Fatma Çalışkan
fatmac@istanbul.edu.tr

[1] Faculty of Forestry, Silviculture Department, İstanbul University-Cerrahpaşa, Bahcekoy, Sariyer, İstanbul, Türkiye

[2] Faculty of Economics, Department of Management Information Systems, İstanbul University, Fatih, İstanbul, Türkiye

[3] Faculty of Science, Department of Mathematics, İstanbul University, Fatih, İstanbul, Türkiye

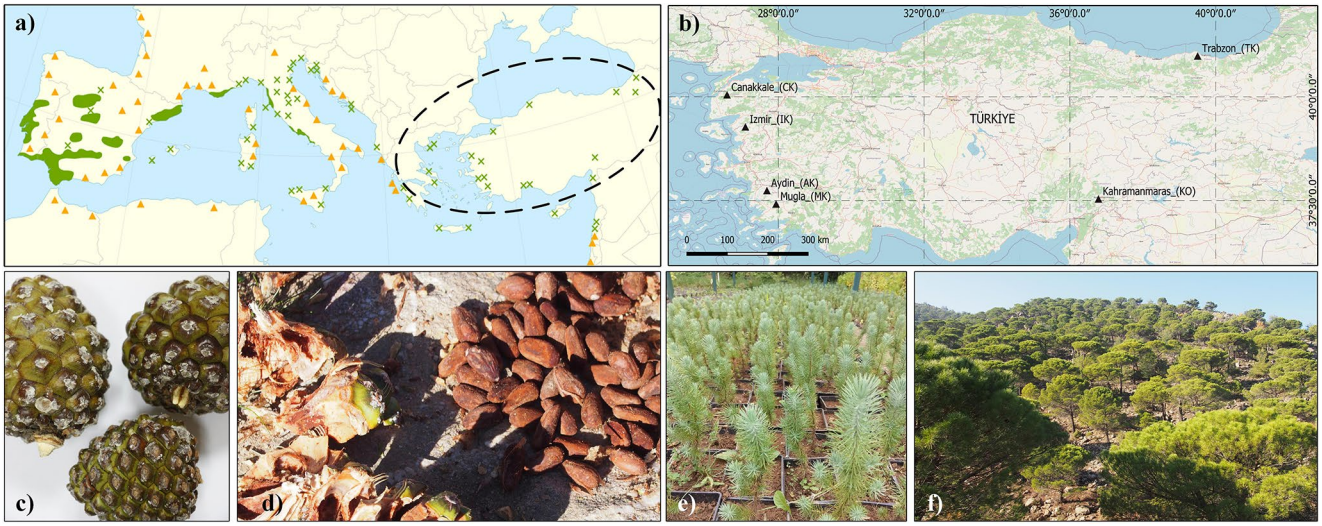


Fig. 1 Distribution map of stone pines (*Pinus pinea*) in the Mediterranean Basin (**a**). (modified from Caudullo et al., 2021). Geographic locations of six natural stands of stone pines (*Pinus pinea*) across Türkiye (**b**). KO Kahramanmaras-Onsen¸ MK Muğla-Katrancı, AK Aydın-Koçarlı, IK İzmir-Kozak, TK Trabzon-Kalenema, CK Çanakkale-Kirazlı. The cones of stone pines generally take three years from pollination to ripening (**c**). Extracted seeds from the cone of stone pines (**d**). One-year-old stone pine seedlings belong to six different populations were grown in the nursery (**e**). The view of natural stands of stone pines in Kahramanmaraş (KO) (**f**)

as a promising candidate for horticultural cultivation. However, there is current evidence of growth and pine nut yield decline due to biotic factors such as *Diplodia* (Caballol et al. 2022; Hlaiem et al. 2023), *Leptoglossus occidentalis* (Bracalini et al. 2013; Lesieur et al. 2014; Farinha et al. 2018) and climatic factor (Parlak 2017; Balekoglu et al. 2020).

The Mediterranean region has approximately 860,000 hectares of forests dominated by stone pines. These forests span from the Atlantic coast of Portugal to the Black Sea and Mount Lebanon (Mutke et al. 2012). The distribution of these areas is primarily distributed approximately 450,000 hectares in Spain, 195,000 hectares in Portugal (ICNF 2013), 176,000 hectares in Türkiye (OGM 2021), and 40,000 hectares in Italy (Pereira et al. 2015).

Stone pines exhibit notable phenotypic plasticity and adaptability coupled with relatively limited genetic variability (Fallour et al. 1997; Vendramin et al. 2008). Genetic diversity is commonly acknowledged as a crucial factor for adjusting to various environmental conditions. Species that are both genetically poor and extensively distributed are uncommon, and the stone pine stands out in this regard because of its significantly low genetic diversity. Nevertheless, the species displays a significant degree of variation in its adaptive characteristics, as highlighted by Vendramin et al. (2008).

There are two basic conditions for obtaining the highest yield in terms of quantity and quality from afforestations to be made in a certain area and for the health of these afforestations (Boydak and Çalışkan 2014, 2015). The first of these is the selection of the most suitable sources in terms of quantity and quality according to our objectives, the breeding of these sources, and the afforestation with the seedlings of the seeds collected from these sources. The second condition is to use the seedlings collected from the seed sources at certain elevations and horizontal distances according to this source, taking into account the soil and climatic conditions. Otherwise, seedling survival, development, form, and resistance to biotic and abiotic pests will decrease significantly.

On the other hand, geographic variation is a concept that describes the differences in morphological traits between populations of a species across their natural distribution. These variations can be observed in both the generative organs and vegetative parts of the trees. Studying geographic variation is essential for understanding the diversity within a species and can provide valuable insights for selecting suitable populations for provenance trials. Provenance trials are experiments that evaluate the performance of seedlings from different origins or from different seed stands when grown in different locations. These trials help determine the optimal horizontal and vertical distances, or "seed transfer zones," for transporting seeds. Provenance trials are considered the most reliable method for establishing seed transfer zones. By assessing the growth and resistance of seedlings from different origins in various locations, provenance trials can provide valuable information for selecting the most suitable seed sources for afforestation. This information helps ensure that the planted seedlings are well adapted to the local environment and have the best chance of survival and growth (Boydak and Çalışkan 2021). Knowing the seed and seedling characteristics of populations is important for determining differences in applications in seed transfer zones and geographic variation studies (Boydak and

Çalışkan 2021). Several studies have investigated the use of machine learning (ML) algorithms in afforestation and breeding programs. Precision morpho-physiological trait measurements and genotyping data are regarded as crucial inputs in contemporary breeding programs (Duc et al. 2023; Niknejad et al. 2023). In the last decade, ML algorithms have been extensively used in agriculture and forestry (Chen et al. 2023; Montagnoli et al. 2016). ML algorithms are also used for variety classification based on the shape characteristics of the seeds of some species (Yang et al. 2019; Osako et al. 2020).

ML algorithms are rapidly gaining traction as powerful tools for predicting phenotypes. As the volume and complexity of biological data continue to increase, ML algorithms have demonstrated a remarkable ability to extract patterns and make predictions, offering unparalleled insights into the relationships between genes, environments, and observable traits. In addition, several mathematical models and deep learning methods have been applied to analyze the dynamics of pine wilt disease and its effects on forest ecosystems (Rao et al. 2023; Shah et al. 2024). ML algorithms predict the stand volume growth of several national-scale forests (Tian et al. 2022). In recent years, significant breakthroughs have been made in ML algorithms and deep learning for drought modeling, and knowledge-based systems with big data platforms, including machine learning and deep learning, are highly important for future research (Prodhan et al. 2022).

Throughout the life of a plant, seed characteristics play a crucial role in its survival strategies (Caliskan and Makineci 2014; Balekoglu et al. 2021). The development of high nut yielding stone pine populations requires genetic diversity in seed and seedling traits across different stone pine populations to identify superior genotypes. Due to data collection challenges, our understanding of seed and seedling morphological traits is limited. In recent years, the use of ML algorithms to classify and predict the viability of seeds using data obtained from images has increased (Kusumaningrum et al. 2018; Zhang et al. 2018; Fan et al. 2020; Ma et al. 2020; Wang et al. 2021; Qi et al. 2024). On the other hand, research on ML-based classification of morphological seed traits from different origins or populations is limited (Nie et al. 2019).

The present study aims to classify and cluster the seeds and seedlings obtained from different stone pine populations in terms of morphological characteristics. To achieve this aim, it is intended to utilize the recently increasing trend of ML algorithms. The first hypothesis of this study is to test whether the seeds and the seedlings can be classified in terms of morphological characters using ML algorithms. The second hypothesis is to find the ideal cluster number with the k-means algorithm and to compare the results obtained from the k-means algorithm with reality.

Materials and methods

Seed material

The collection of seed material from six natural stands of stone pine (*Pinus pinea* L.) across Türkiye was carried out by Çalışkan et al. (2018) (Fig. 1b). The locations and geoclimatic traits of the six populations are provided in Table 1. To ensure representative sampling, mature cones were harvested from 15 to 25 mature trees within each population, with a minimum distance of 50 to 100 m between each tree. For each tree, ten mature cones were placed in separate, marked cloth bags and taken to the laboratory. The samples were stored at room temperature until further analysis. Subsequently, seeds were manually extracted from 50 randomly selected cones within each of the six populations (Fig. 1c, d and f).

Seed and seedling measurements

In each population, fifty cones were chosen randomly and dissected entirely to assess the seed yields of the six populations. The viable seeds were differentiated by submerging each cone's seeds in water. An objective assessment of the seed measurements was made by sampling ten seeds per cone (Ganatsas et al. 2008). Seed measurements were taken on a total of 286 seeds. The measurements of seedlings were carried out on 216 seedlings (9 seedlings × 4 replications × 6 populations) in the first and second years. Abbreviations and units for the measured cone, seed, and seedling characteristics are given in Table 2. A summary of seed and seedling characteristics across six stone pine populations is given in Table 3.

Plastic containers (18 cm tall and 190 cm^3) were used for sowing the seeds. After one year, the seedlings (Fig. 1e) were transferred to plastic containers measuring 20 cm in height and 12.5 cm in diameter with a capacity of 2.5 l. A 2:1:1 mixture of soil, peat, and river sand was placed in these containers. For the trial, a randomized block design with five replications was used. The plants were grown outdoors at the Istanbul Forest Nursery (the nursery is located at an altitude of 126 m above sea level at 41°10′56″N, 28°59′14″E), where they received daily irrigation and no additional fertilization. The physical and chemical compositions of the container media were as follows: 68.2% sand, 14.5% silt, 17.3% clay, pH 6.9, electrical conductivity (EC) 160.1 µhos/cm, 2.4% carbon content, 0.2% nitrogen content, and no calcium carbonate. According to Thornthwaite's classification,

Table 1 Geographical and climatic characteristics of the six stone pine (*Pinus pinea*) populations across their natural distribution in Türkiye. P is the mean annual rainfall in mm; M is the mean of the maxima of the hottest month; m is the mean of the minima of the coldest month; PE is the sum of rainfall in June, July, and August; ME is the mean of the maxima temperature of June, July, and August for the period 1987–2016; S is the summer drought index value; and Q is the humidity category value. *A* Bioclimate zones, according to Emberger (Daget et al. 1988)

Populations	Latitude Longitude Altitude (m)	*P* mm	M °C	m °C	PE mm	ME °C	S[a]	S Summer Drought Index[a]	Q[a]	Q Humidity category[a]
Kahramanmaraş-Önsen (KO)	37.50758 36.78308 800	742.38	36.62	1.53	7.35	40.30	0.18	2	72.39	Subhumid
Muğla-Katrancı (MK)	37.38645 27.93964 608	632.98	36.36	2.23	31.45	39.71	0.79	2	63.41	Subhumid
Aydın-Koçarlı (AK)	37.70999 27.68841 730	609.34	36.90	4.17	13.80	40.53	0.34	2	63.38	Subhumid
İzmir-Kozak (IK)	39.24608 27.09746 555	611.99	34.01	3.25	26.49	38.96	0.68	2	68.18	Subhumid
Trabzon-Kalenema (TK)	40.95831 39.50467 284	742.31	27.75	4.19	105.70	30.10	3.51	4	108.96	Humid
Çanakkale-Kirazlı (CK)	40.02023 26.59359 276	605.03	31.23	3.35	38.64	34.73	1.11	3	74.71	Subhumid

the climate around the nursery is humid, mesothermal, and maritime, with a moderate water deficit during the summer months. The average annual rainfall was recorded at 1096 mm, with an average temperature of 13.2 °C (Baylan and Ustaoğlu 2020).

Analyses

Descriptive statistical analyses of the seed and seedling datasets were also conducted. The minimum, maximum, mean, median, 1st and 3rd quartile values were calculated for the datasets. Then, the correlations between the attributes were analyzed with the Pearson correlation coefficient, and the significance levels (*p* values) of these coefficients were calculated.

ML algorithms were used to solve classification and clustering problems. In the present study, supervised learning algorithms k-Nearest Neighbors (k-NN), Naive Bayes (NB), Support Vector Machines (SVM), C5.0, Classification and Regression Trees (CART), and Random Forest (RF) were used to predict the populations of seeds and seedlings. A brief overview of the algorithms is summarized below, and more details about the algorithms can be found in (Breiman et al. 1984; Vapnik 1995; Han and Kamber 2006; Harrington 2012; Murphy 2012; Balaban and Kartal 2018; Jafarzadegan et al. 2020; Kartal et al. 2020; Quinlan 2022). The k-NN algorithm classifies instances in a dataset based on a given distance measure. The class of an instance of an unknown class is determined based on the class of its k nearest neighbors. In this process, Euclidean distance is utilized for the k-NN algorithm, and the minimum k, which gives the highest accuracy, is chosen for the k-NN analysis. NB is a statistical classification method that uses Bayes' theorem-based probability value calculations. Its most important assumption is that attributes are conditionally independent given the target value. SVMs are supervised learning techniques often used for classification and regression tasks. The goal is to divide the data points by a hyperplane between two classes. The principle of margin maximization attempts to provide the best separation between the two classes. C5.0 is a decision tree algorithm. Like SVMs, it can be used for classification and regression tasks. The decision tree is built using concepts such as entropy and information gain. The CART algorithm is used to build classification and regression trees. Most often, trees are built with the help of the Gini index, a measure of the impurity of attributes. Only binary branching is performed in the tree. RF is an ensemble learning algorithm. An RF model is built by combining many decision trees. It is used to solve both classification and regression problems. The attributes used in the trees and the observations used in the training process of the trees are randomly selected. The final decision is based on the consensus of the decision trees.

In the present study, the stratified hold-out method was employed as the ML algorithms performance evaluation technique in classification. The dataset is randomly divided into training and test sets. At this stage, the following were considered: (1) The data used in training is not included in the test dataset. This means the data used in the testing phase is completely new for the model. (2) An observation

Table 2 List of morphometric traits of *P. pinea* seeds and seedlings with measurements and units. [A] and [b] indicate ≈ 20% and 10% moisture content, respectively*. No terminal buds were observed in the first year

Morphometric	Abbreviation	Unit	Measurement
Fruit traits			
Cone diameter	CD	mm	Three measurements across the widest portion with a digital caliper
Cone length	CL	mm	Measured using a digital caliper
Cone weight	CW	g	3-year-old cones were weighed with a precision scale (accuracy of 0.01 g)[a]
Seeds cone^{-1}	SN	No.	All seeds per cone were extracted and counted
Empty seeds	EN	No.	All extracted seeds were floated, and floating seeds per cone were counted.
Seed diameter	SD	mm	Three measurements across the widest portion with a digital caliper
Seed length	SL	mm	Measured using a digital caliper
Seed weight	SW	g	Weighed of sound seeds with precision scale (accuracy of 0.01 g)[b]
Seed empty weight	EW	g	Weighed with precision scale (accuracy of 0.01 g)[b]
Pine nut weight	PW	g	Five sound (filled) seeds and empty seeds were weighed from each cone. Calculated differences average five sound seed weights and empty seed weights.[b]
Seedling traits			
First-year-length	L1	cm	Measured by the scale at the end of the first vegetation period in 216 seedlings (9 seedlings x 4 replications x 6 populations)
First-year-diameter	D1	mm	Digital caliper measured root collar diameters at the end of the first vegetation period in 216 seedlings (9 seedlings x 4 replications x 6 populations)
First-year-branch-number	B1	No.	Counted in 216 seedlings (9 seedlings x 4 replications x 6 populations) at the end of the first vegetation period
Second-year-length	L2	cm	Measured by the scale at the end of the second vegetation period in 216 seedlings (9 seedlings x 4 replications x 6 populations)
Second-year-diameter	D2	mm	Digital caliper measured root collar diameters at the end of the second vegetation period in 216 seedlings (9 seedlings x 4 replications x 6 populations)
Second-year-branch-number	B2	No.	Counted in 216 seedlings (9 seedlings x 4 replications x 6 populations) at the end of the second vegetation period
Bud set	BS	1/0	Observed in 216 seedlings (9 seedlings x 4 replications x 6 populations) at the end of the second vegetation period*

selected for the training/test datasets was chosen to appear only once in the respective dataset (without replacement). (3) The proportion of class labels for the target attribute was also approximately maintained in both training and test datasets. The ML algorithms were trained using 80% of the datasets, while the remaining 20% were reserved for testing the performance of the models. The evaluation results belong to the test set. Furthermore, the importance of the attributes in the datasets was evaluated with CART, C5.0, and RF algorithms in this study. The percentage of training set samples that fall into all terminal nodes following the split is considered to assess the importance of predictors with C5.0 (Kuhn and Quinlan 2023). The overall variable importance with CART is calculated as the total of the goodness of split measures for each split in which it was the primary variable, in addition goodness multiplied by (adjusted agreement) for all splits in which it was a surrogate, since a variable might occur in the tree multiple times, either as a primary or surrogate variable (Therneau and Atkinson 2022). Finally, the mean decrease in Gini (index) is used in RF as the impurity measure. It can be defined as the average decrease in node impurity, weighted by the proportion of samples reaching that node in each decision tree in the RF (Gómez-Ramírez et al. 2020; Liaw and Wiener 2002). The

Table 3 Information in terms of statistically comparing the mean values of the analyzed traits between the six populations of stone pine. For the traits of abbreviations used, please see Table 2

Traits	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
SD (mm)	6.52	8.22	8.63	8.62	9.00	10.25
SL (mm)	12.77	16.67	17.71	17.58	18.74	21.08
SW (g)	0.36	0.69	0.81	0.80	0.90	1.20
EW (g)	0.16	0.50	0.60	0.59	0.68	0.95
CD (mm)	53.60	68.78	75.20	73.71	78.99	86.61
CL (mm)	58.24	88.32	98.26	97.47	106.54	150.40
CW (g)	87.06	189.78	240.14	245.11	290.51	435.00
SN (#)	1.00	8.25	18.00	27.74	42.75	109.00
EN (#)	1.00	9.00	18.00	23.66	35.75	76.00
PW (g)	0.01	0.15	0.20	0.20	0.25	0.44
L1 (cm)	6.20	14.57	16.50	16.28	18.60	23.60
L2 (cm)	18.60	27.85	31.35	31.81	35.23	52.30
D1 (mm)	1.18	2.53	2.91	2.94	3.33	4.86
D2 (mm)	3.17	5.27	5.81	5.80	6.41	7.90
B1 (#)	0	3	4	4	5	8
B2 (#)	0	4	6	7	9	19
BS (#)	0	0	0	0	0	1

Table 4 The confusion matrix. True positives (TP): positive cases correctly classified as positive; false positives (FP): negative cases incorrectly classified as positive; false negatives (FN): positive cases incorrectly classified as negative; true negatives (TN): negative cases correctly classified as negative

		Actual / Reference Values	
		Positive	Negative
Predictions	Positive	TP	FP
	Negative	FN	TN

mean decrease in Gini is the total decrease in node impurity caused by variable splitting averaged across all trees (Han et al. 2016). The greater the mean decrease in the Gini value, the purer the variable.

In classification tasks, the performance of the ML algorithm is evaluated using the confusion matrix given in Table 4 (Murphy 2012). In binary classification, where there are two class labels - one negative and one positive - four possible outcomes can arise, as indicated in Table 4.

Using the given confusion matrix, the most popular performance evaluation metrics, namely accuracy, error rate, True Positive Rate (TPR)/recall/hit rate/sensitivity, Positive Predictive Value (PPV)/precision, and F1-Score/F-Measure/F-Score, can be calculated as following formulas (1, 2, 3, 4 and 5):

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$\text{Error} = 1 - \text{Accuracy} \tag{2}$$

$$\text{True Positive Rate (TPR)} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{Positive Predictive Value (PPV)} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{F1} - \text{Score} = \frac{2 \times TPR \times PPV}{TPR + PPV} \tag{5}$$

In this study, in addition to the criteria formulated above, False Positive Rate (FPR)/fall-out, True Negative Rate (TNR)/specificity/selectivity, False Negative Rate (FNR)/miss rate, Negative Predictive Value (NPV), False Discovery Rate (FDR), and False Omission Rate (FOR) metrics were also used in the performance evaluation stage for classification. For more information on these, which depends on Table 4 as the previous ones, please see Dua and Chowriappa (2013) and Snodgress (2023). Since there were more than two class labels in both data sets (six populations) in this study, macro-averaging was applied, and the model performance evaluation criteria calculated for each class label were averaged.

Then, to analyze the variations in seed and seedling traits across the populations, the k-means algorithm was used as an unsupervised learning algorithm. We extracted real population labels from the datasets and performed clustering analyses. The k-means algorithm is based on assigning similar instances in the dataset to the same clusters. Given a user-defined number of clusters, the instances in the dataset are grouped. The algorithm initially selects k cluster centers. Then, each instance is assigned to the closest center, and the cluster centers are updated. In this process, Euclidean distance is used for the k-means algorithm in this study. This process is repeated until no sample moves between clusters or reaches the maximum number of iterations (Shmueli et al. 2018). Although the datasets included six populations,

the k-means algorithm's parameter k, which indicates the number of clusters, was tested for numbers between 2 and 20 with the average Silhouette width (Rousseeuw 1987) to determine the ideal number of clusters. Then, we additionally examined how close the clustering results obtained with k-means are to reality from 2 to 6 (since, in real life, there are six classes in the datasets).

In this study, data normalization, the min-max normalization method, was applied to both seed and seedling traits datasets in the data preprocessing stage. Considering the algorithm's performance in clustering seed and seedling traits, datasets are used without and with normalization, respectively (Figs. 2 and 3). Analyses were performed in RStudio with the R programming language (Posit 2023; R Core Team 2023). The following R packages were used: C5.0 (Kuhn and Quinlan 2023), caret (Kuhn 2008), class (Venables and Ripley 2002), cluster (Maechler et al. 2022), clusterSim (Walesiak and Dudek 2020), cpfa (Snodgress 2023), e1071 (Meyer et al. 2023), factoextra (Kassambara and Mundt 2020), ggplot2 (Wickham 2016), randomForest (Liaw and Wiener 2002), and rpart (Therneau and Atkinson 2022). Figures 2 and 3 show flowcharts of classification and clustering.

Principal Component Analysis (PCA) is a statistical method that reduces the dimensionality of a dataset by generating linear combinations of the original variables known as principal components (Greenacre et al. 2022). These components capture the largest variation in the data, resulting in a simplified approximation of the original dataset (Greenacre et al. 2022). It converts data to a new coordinate system and performs an orthogonal linear transformation (Öngen Bilir and Kardeş 2023). This study applied PCA to both datasets following clustering analysis for data visualization. A summary of the ML algorithms used in the study is given in Table 5.

Results

The means and ranges of variation for the seed and seedling traits are shown in Table 3. Table S1 and Table S2 show the correlation analysis of these traits. The results indicate that nearly all the correlations among the seed traits were statistically significant. Moreover, correlation analysis of the seed and seedling traits revealed that the seed traits had greater r values than the seedling traits. The correlations between the seedling traits were not greater than 0.6, while the correlation between the seed traits reached 0.9.

Classification of stone pines seeds and seedlings

The k parameter of the k-NN algorithm was tested from 2 to 20 (Fig. S1). The best performance was achieved with $k=2$ for seed traits (accuracy = 0.52) and $k=18$ for seedling traits (accuracy = 0.57), comparable to the other algorithms' performances.

Tables 6 and 7 show the model performance evaluation for stone pine seeds and seedlings, ranked by accuracy and then F1-Score from highest to lowest. With an accuracy of 0.648 and an F1-Score of 0.658, the pine seed model achieved the best performance in the RF algorithm. Additionally, the best classification performance for stone pine seedlings was observed for the k-NN algorithm ($k=18$), for which the accuracy and F1-Score were 0.571 and 0.582, respectively.

Tables 8 and 9 show that the CART, C5.0, and RF algorithms report the importance of the attributes in stone pine seeds and seedlings, respectively. The importance of an attribute for the model increases with the given importance values, which are described in the Analyses Section. It can be seen that the classification performance of these algorithms is higher for the pine seeds dataset. Therefore, cone diameter (CD), seeds cone (SN), and (empty seeds) EN are ranked as the three most important attributes for the pine seeds dataset. These attributes can be taken into consideration in future studies.

Clustering of stone pine seeds and seedlings

The clustering quality of the k-means algorithm was evaluated with the average Silhouettes for different k values. The average Silhouette Index values should be as close to 1 when assessing clustering quality. An average Silhouette Index value close to -1 indicates that a sample is not in the appropriate cluster (Rousseeuw 1987). The k parameter of the k-means algorithm was tested from 2 to 20 (Fig. S2). The best performance was achieved with $k=2$ for the seed (average Silhouette Index = 0.48) and seedling (average Silhouette Index = 0.51) traits.

In Fig. S3, it seems more appropriate to divide the data into the two most distinct clusters (the average Silhouette Index is 0.35 for $k=6$ and 0.48 for $k=2$ for the stone pine seeds); however, since it is known that there are six populations in the dataset, the evaluation results for $k=6$ are interpreted. In Fig. S4, it seems more appropriate to divide the data into the two most distinct clusters (average Silhouette Index = 0.51); however, since it is known that there are six populations in the dataset, the evaluation results were interpreted for $k=6$. According to the graph, for $k=6$, there are 19, 42, 15, 34, 63 and 43 observations in the 1st, 2nd, 3rd, 4th, 5th and 6th clusters, respectively. The

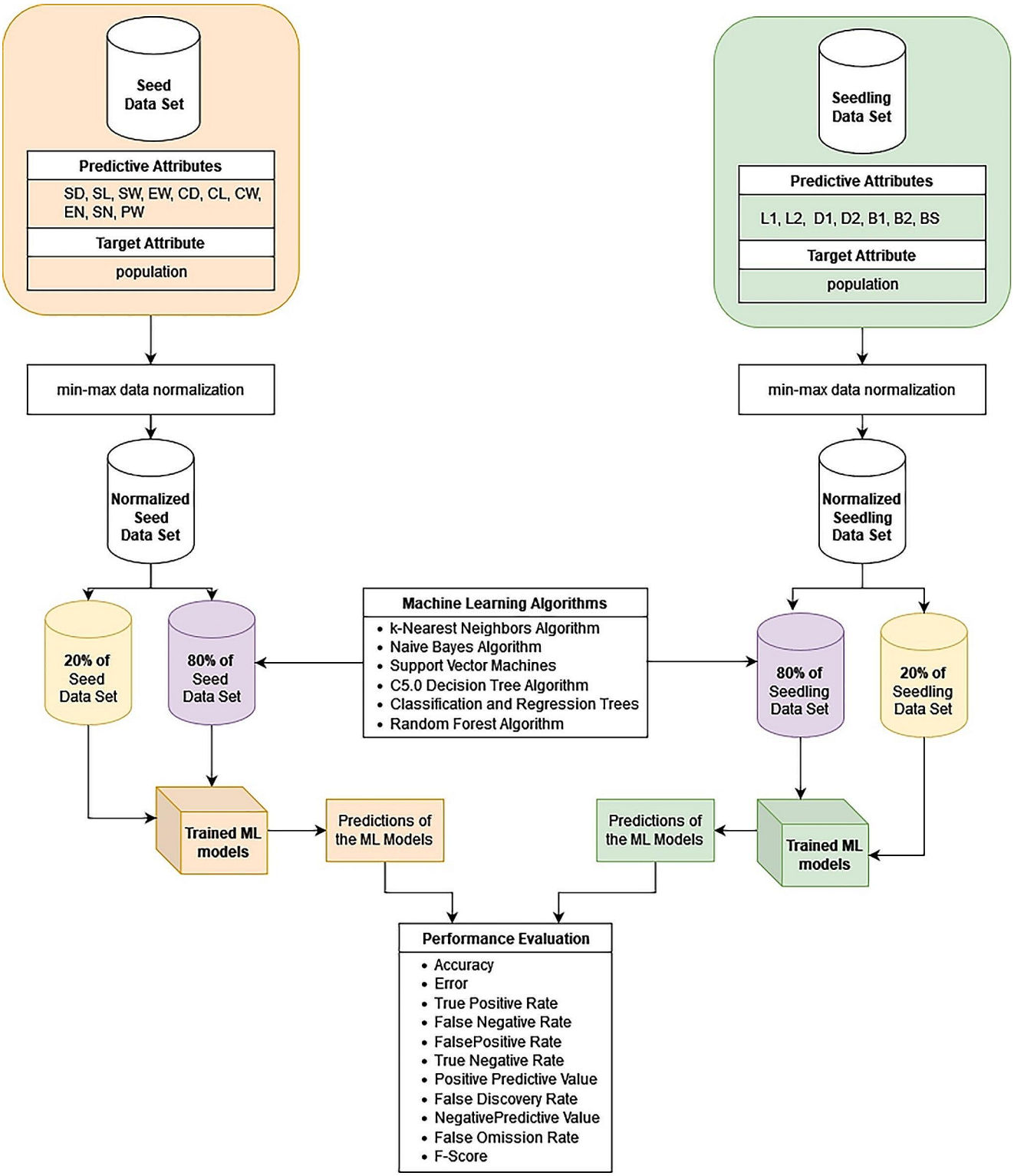


Fig. 2 Flowchart of the classification process. Six ML algorithms were applied to classify the stone pine (*Pinus pinea*) population. For the abbreviations used, please see Table 2

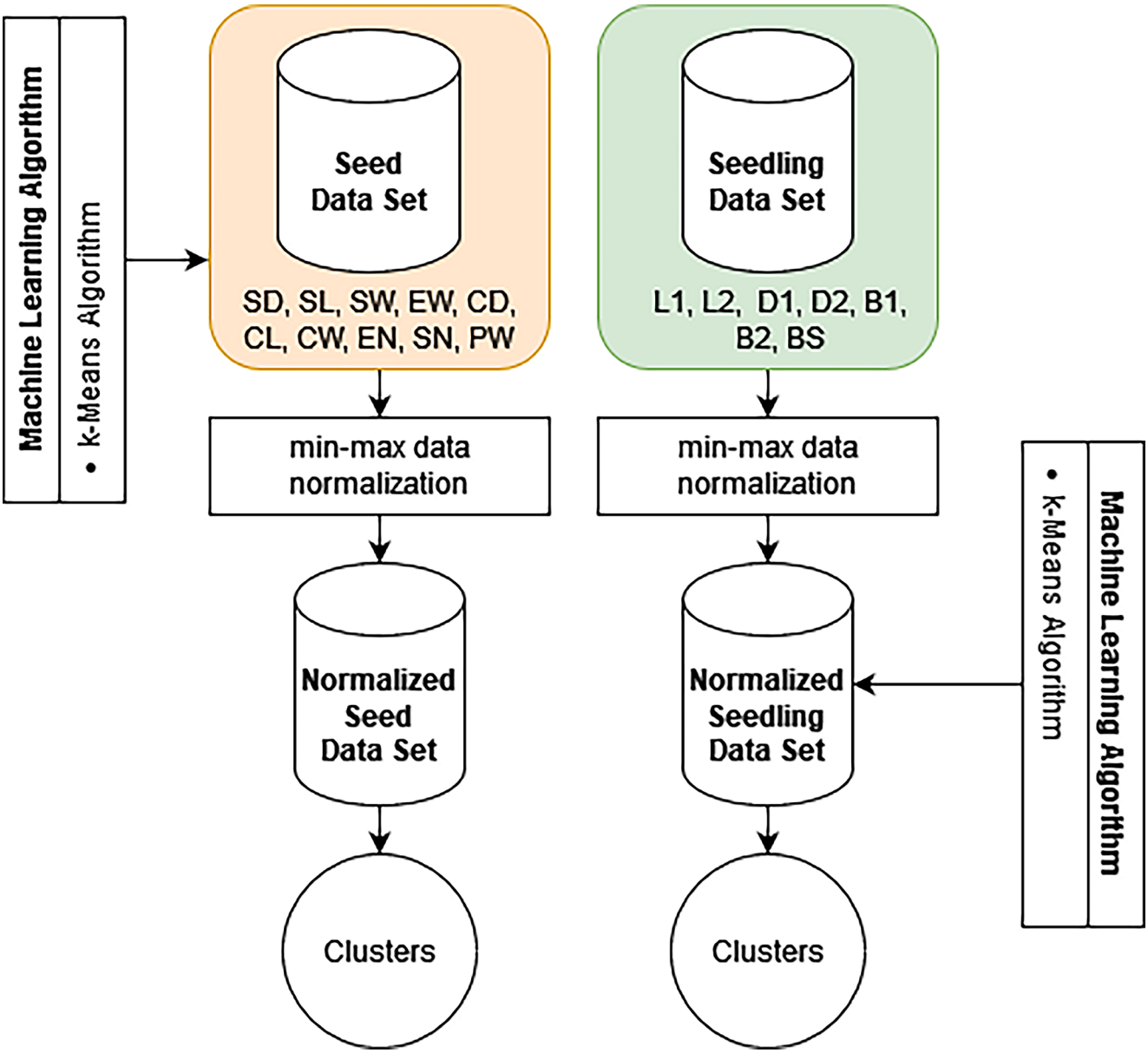


Fig. 3 Flowchart of the clustering of stone pine (*Pinus pinea*) populations. For the abbreviations used, please see Table 2

Table 5 The fundamental characteristics of the ML algorithms used in the study

Features	Machine Learning Algorithms							
	CART	C5.0	k-means	k-NN	NB	PCA	RF	SVM
Bayes Theorem Used	-	-	-	-	+	-	-	-
Classification	+	+	-	+	+	-	+	+
Clustering	-	-	+	-	-	-	-	-
Decision Trees Used	+	+	-	-	-	-	+	+
Dimension Reduction	-	-	-	-	-	+	-	-
Ensemble Learning	-	-	-	-	-	-	+	-
Feature Importance	+	+	-	-	-	-	+	-

Table 6 Evaluation results for the six-model performance of stone pine seeds

	RF	SVM	C5.0	CART	NB	k-NN (k = 2)
Accuracy	0.648	0.648	0.630	0.574	0.537	0.463
Error	0.352	0.352	0.370	0.426	0.463	0.537
TPR	0.651	0.654	0.627	0.576	0.547	0.474
FPR	0.096	0.094	0.104	0.126	0.145	0.175
TNR	0.904	0.906	0.896	0.874	0.855	0.825
FNR	0.349	0.346	0.373	0.424	0.453	0.526
PPV	0.666	0.639	0.660	0.597	0.519	0.480
NPV	0.903	0.910	0.899	0.877	0.863	0.822
FDR	0.334	0.361	0.340	0.403	0.481	0.520
FOR	0.097	0.090	0.101	0.123	0.137	0.178
F1-Score	0.658	0.647	0.643	0.586	0.532	0.477

RF: Random Forest, SVM: Support Vector Machines: C5.0: Decision Tree Algorithm, CART: Classification and Regression Trees, NB: Naive Bayes, k-NN: k-Nearest Neighbors, True Positive Rate (TPR)/recall/hit rate/sensitivity, False Negative Rate (FNR)/miss rate, False Positive Rate (FPR)/fall-out, True Negative Rate (TNR)/specificity/selectivity, Positive Predictive Value (PPV)/precision, FDR: False Discovery Rate, NPV: Negative Predictive Value, FOR: False Omission Rate

Table 7 Results of the six models' classification performance assessment for seedlings of stone pine

	k-NN (k:18)	SVM	C5.0	RF	CART	NB
Accuracy	0.571	0.548	0.476	0.429	0.310	0.286
Error	0.429	0.452	0.524	0.571	0.690	0.714
TPR	0.571	0.548	0.476	0.429	0.310	0.286
FPR	0.125	0.137	0.176	0.210	0.307	0.225
TNR	0.875	0.863	0.824	0.790	0.693	0.775
FNR	0.429	0.452	0.524	0.571	0.690	0.714
PPV	0.593	0.567	0.492	0.406	0.265	-
NPV	0.872	0.861	0.823	0.800	0.703	0.687
FDR	0.407	0.433	0.508	0.594	0.735	-
FOR	0.128	0.139	0.177	0.200	0.297	0.313
F1-Score	0.582	0.557	0.484	0.417	0.286	-

k-NN: k-Nearest Neighbors, SVM: Support Vector Machines, C5.0: Decision Tree Algorithm, RF: Random Forest, CART: Classification and Regression Trees, NB: Naive Bayes, True Positive Rate (TPR)/recall/hit rate/sensitivity, False Negative Rate (FNR)/miss rate, False Positive Rate (FPR)/fall-out, True Negative Rate (TNR)/specificity/selectivity, Positive Predictive Value (PPV)/precision, FDR: False Discovery Rate, NPV: Negative Predictive Value, FOR: False Omission Rate

Table 8 The order of attributes in the stone pine seeds dataset according to the importance values obtained with the classification and regression trees (CART), C5.0, and Random Forest (RF) algorithms

CART	Importance	C5.0	Importance	RF	Mean Decrease Gini
CD	45.867	CD	100.000	EN	31.178
EN	38.324	SN	100.000	SN	26.888
SL	36.992	EN	98.710	CD	25.833
CW	32.906	SL	40.090	SL	22.618
CL	28.463	SD	22.840	CL	20.093
SW	23.148	CL	21.980	SD	16.986
EW	22.191	SW	21.550	EW	14.080
SN	20.618	EW	11.640	SW	13.639
SD	14.010	PW	3.450	CW	11.784
PW	2.222	CW	0.000	PW	9.102

average Silhouette Index values of these clusters were 0.28, 0.14, 0.23, 0.25, 0.15, and 0.11. The percentage of variance explained by each principal component for the seed and seedling datasets is shown in Fig. S5.

Table 9 The order of attributes in the stone pine seedlings dataset according to the importance values obtained with the classification and regression trees (CART), C5.0, and Random Forest (RF) algorithms

CART	Importance	C5.0	Importance	RF	Mean Decrease Gini
L2	18.35	L2	100.00	L2	32.34
BS	12.08	D2	82.18	D2	24.73
D2	11.45	D1	78.16	D1	23.27
B2	9.77	B2	75.29	B2	21.53
D1	8.65	B1	62.64	L1	19.59
L1	4.39	BS	50.00	B1	12.95
B1	4.16	L1	39.08	BS	9.04

A PCA plot was generated to determine how the variables are related to each other. The PCA results for stone pine seeds and seedlings are presented in Fig. 4. Positively correlated variables are plotted on the same side of the graph, while negatively correlated variables are plotted on the opposite sides of the graph.

By considering the two principal components obtained with PCA in both datasets (Fig. 4), the clusters obtained

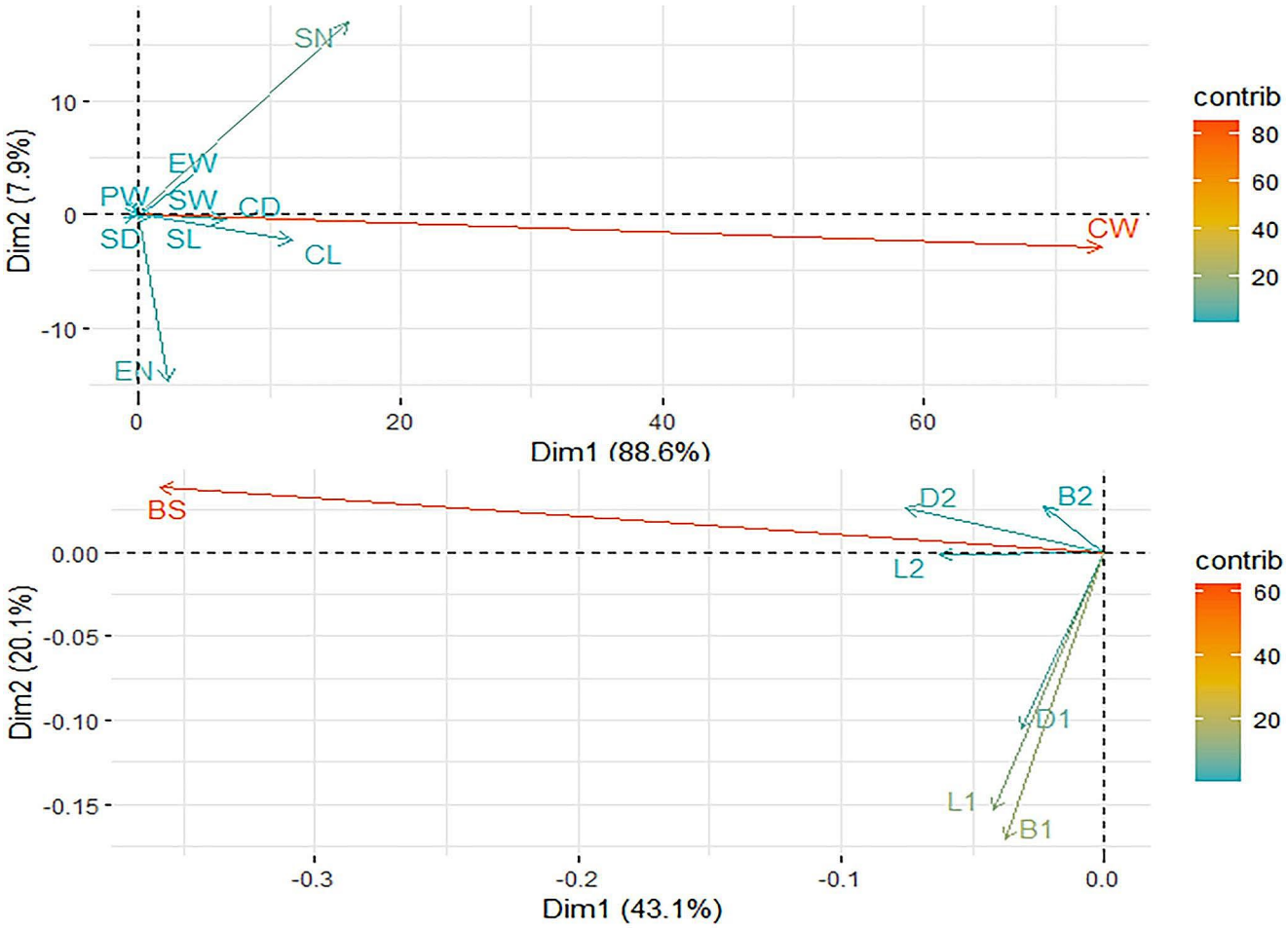


Fig. 4 Principal component analysis (PCA) of stone pine seeds (top) and seedlings (bottom). For the abbreviations used, please see Table 2

from the k-Means algorithm and the actual seedling and seedling traits classes were visualized in Fig. 5. Population clusters are plotted in 2D space using the PCA coordinates of two principal components. The clusters obtained from the k-means algorithm with PCA and the populations in the seed (Fig. 5-Left) and seedling (Fig. 5-Right) datasets were visualized. The colors in the graph represent the clusters revealed by the k-means algorithm, while the labels indicate the actual populations to which the examples belong.

Table 10 shows the percentage of observations in each cluster (k = 2 to k = 6) belonging to a particular class (KO, MK, CK, AK, IK, TK) for seeds/seedlings. For example, in Table 10, the cluster labeled 1 for seed traits by the k-means algorithm contains 61.70% KO, 38.46% MK, 96.23% CK, 20.83% AK, 34.04% IK, and 64.10% TK for k = 2.

Discussion

Geographic variation describes the differences in morphological traits between populations of a species across their natural distribution. These variations can be observed in both the generative organs and vegetative parts of the trees and are essential for understanding the diversity within a species. They can also provide valuable insights for selecting suitable populations for breeding practices.

Chirici et al. (2016) and McRoberts et al. (2016) highlighted the effectiveness of the k-NN for predicting forest attributes. Begum et al. (2015) also underscore the efficiency of the k-NN algorithm in data classification. These studies collectively suggest the value of k-NN in both forestry and plant phenotyping applications. Skowronski et al. (2021) found that ML algorithms demonstrated superior results to traditional discriminant functions in classifying populations with different degrees of similarity. The k-NN, RF, SVM, and NB algorithms exhibited the highest classification accuracy, surpassing traditional statistical techniques and other ML algorithms. Researchers confirmed that ML algorithms

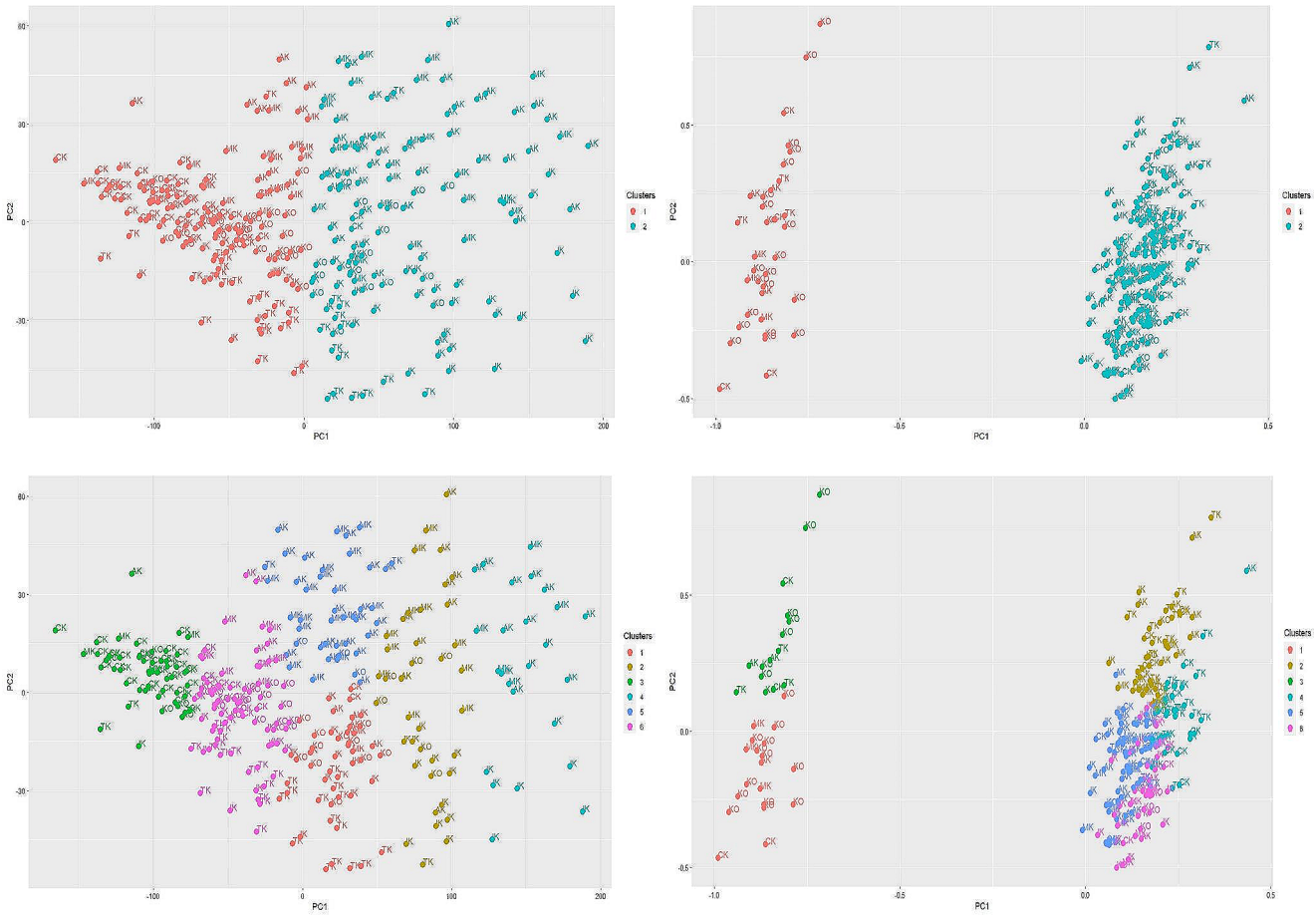


Fig. 5 **Left-**Population clusters plotted in 2D space using principal component analysis (PCA) coordinates of the first two principal components of stone pine seeds (k-parameters of the k-means algorithm are 2 and 6). **Right-** Population clusters plotted in 2D space using PCA coordinates of two principal components of stone pines seedlings (k-parameter of the k-means algorithm is 6-upper; k-parameter of the k-means algorithm is 2-below). The clusters obtained from the k-means algorithm with PCA and the populations in the seedling dataset are visualized in 2 dimensions. The colors in the graph represent the clusters revealed by the k-means algorithm, while the labels indicate the actual populations to which the observations belong

are more accurate at discriminating and classifying populations than statistical techniques without limitations. Rao et al. (2023) reported that SVM and k-NN outperformed the other two ML algorithms. In a study by Sotomayor et al. (2023), five supervised ML algorithms were compared for prediction using multiple abiotic factors, such as topographic, edaphic, and climatic factors. The results showed that the RF model was the most accurate prediction model. As the first hypothesis of our study is to test whether the seeds and the seedlings can be classified in terms of morphological characters with ML algorithms, we obtained the best classification performance with RF since it is an ensemble learning algorithm that uses multiple models in parallel and finally outputs the majority class label for the unlabeled observation. Additionally, the best classification performance for stone pine seedlings was observed for the k-NN algorithm.

Huang et al. (2022) utilized seven distinct pine nut varieties, namely *P. bungeana, P. yunnanensis, P. thunbergii, P. armandii, P. massoniana, P. elliottii, and P. taiwanensis*. Five algorithms, decision tree, RF, multilayer perceptron, SVM, and NB, were employed to classify the pine nut samples. That study demonstrated the effectiveness of ML algorithms for classifying pine nuts (Huang et al. 2022). When evaluating the classification performance of ML algorithms, the ultimate goal is to achieve accuracy as close to 1 as possible. Performance metrics such as sensitivity, precision, and F1-Score provide additional support in assessing the effectiveness of the models. According to the classification results in our study, the accuracy of the selected ML algorithms ranges between $0.571-0.286$ for stone pine seedlings and $0.648-0.463$ for stone pine seeds.

Clustering analysis and PCA were used to evaluate the variability in seed quality among *Pinus patula* clonal seed orchards based on three physical cone characteristics as

Table 10 Actual distribution of the populations in each cluster for k = 2, 3, 4, 5, 6 regarding stone pine seed/seedling traits. Bold values represent the most dominant population in each cluster

		Populations					
	Clusters	KO	MK	CK	AK	IK	TK
k = 2	1	61.70/**58.33**	38.46/8.33	**96.23**/11.11	20.83/8.33	34.04/0	64.10/8.33
	2	38.30/41.67	61.54/91.67	3.77/88.89	**79.17**/91.67	65.96/**100**	35.90/91.67
k = 3	1	57.45/30.56	57.69/61.11	7.55/44.44	54.17/52.78	38.30/**66.67**	**64.10**/22.22
	2	8.51/11.11	26.92/30.56	0.00/44.44	41.67/38.89	**44.68**/33.33	2.56/**69.44**
	3	34.04/**58.33**	15.38/8.33	**92.45**/11.11	4.17/8.33	17.02/0	33.33/8.33
k = 4	1	46.81/25	26.92/**55.56**	22.64/30.56	16.67/50	27.66/41.67	**53.85**/19.44
	2	34.04/5.56	44.23/25	3.77/25	**47.92**/30.56	31.91/5.56	35.90/**33.33**
	3	14.89/**58.33**	9.62/8.33	**73.58**/11.11	2.08/8.33	6.38/0	10.26/8.33
	4	4.26/11.11	19.23/11.11	0.00/33.33	33.33/11.11	**34.04/52.78**	0.00/38.89
k = 5	1	36.17/8.33	3.85/41.67	3.77/22.22	6.25/**44.44**	38.30/25	**58.97**/8.33
	2	14.89/5.56	23.08/25	0.00/22.22	**31.25**/30.56	25.53/5.56	5.13/**33.33**
	3	34.04/**58.33**	15.38/8.33	**92.45**/11.11	2.08/8.33	17.02/0	33.33/8.33
	4	0.00/5.56	11.54/11.11	0.00/30.56	**20.83**/11.11	19.15/30.56	0.00/**38.89**
	5	14.89/22.22	**46.15**/13.89	3.77/13.89	39.58/5.56	0.00/**38.89**	2.56/11.11
k = 6	1	**25.53/36.11**	3.85/8.33	3.77/5.56	8.33/2.78	**25.53/0**	41.03/0
	2	12.77/5.56	21.15/22.22	0.00/22.22	20.83/25	**25.53**/5.56	2.56/**36.11**
	3	10.64/**22.22**	7.69/0	**69.81**/5.56	2.08/5.56	6.38/0	7.69/8.33
	4	0.00/5.56	11.54/11.11	0.00/25	20.83/8.33	**19.15**/8.33	0.00/**36.11**
	5	6.38/11.11	36.54/**47.22**	0.00/27.78	**39.58**/50	0.00/22.22	5.13/16.67
	6	**44.68**/19.44	19.23/11.11	26.42/13.89	8.33/8.33	23.40/**63.89**	43.59/2.78

length, diameter, and weight. Five natural groupings were identified through cluster analysis out of 14 possible clusters. Cone length was found to be most important for group formation, with width and weight having equal effects (Owino et al. 2020). In our study, the axes of PCA for seed traits display a percentage of variance of 88.6%, which could be explained by the first principal component (Dim1) and 7.9% by the second principal component (Dim2). Of the variance, two dimensions account for 97%. On the left side of the plot, the SW, EW, PW, SL, SD, and CD variables are all close to one another. All of the other variables are far from EN. On the other hand, the axes of PCA for seedling traits display a percentage of variance of 43.1%, which was explained by the first principal component (Dim1), and 20.1% was explained by the second principal component (Dim2). Of the variance, two dimensions account for 63%. The variance explained by two dimensions increased to approximately 80% when adding one additional dimension (Fig. S5). On the left side of the plot, the variables BS, L2, D2, and B2 are all close to one another. All the other variables are far from the variables D1, L1, and B1. BS was the most significant parameter influencing the seeding traits (Fig. 4, Fig. S5). The most important features between seed and seedling traits were cone weight (CW) and bud set (BS), respectively. Using the attribute importance ratings given in our study, we can prioritize the attributes in Tables 8 and 9 in their future classification studies on seed and seedling traits. In the clustering analysis, k-means algorithm's performance was evaluated considering our study's second hypothesis. Different clustering analyses were performed on the data sets from 2 to 20 (Fig. S2); however, since there are six different seed and seedling traits, algorithm's performance was especially considered in the case of k = 2, 3, 4, 5, 6. Tables 7 and 8 show the actual distribution of the populations in each cluster obtained by the k-means algorithm for stone pine seed traits and seedling traits, respectively. The most surprising result is that the best clustering performance was achieved with k = 2 for seed and seedling traits. This can be seen in Fig. 5, plotted over the two principal components of the data sets obtained by PCA analyses. For stone pine seed traits, KO, CK, and TK are mostly in Cluster 1, while MK, AK, and IK are mostly in Cluster 2 for k = 2 (Table 10). Table 10 shows that KO is quite distinct from the others (Cluster 1) for seedling traits, while MK, CK, AK, IK and TK are mostly in Cluster 2 for k = 2. In the stone pine seed traits dataset, for k values of 2, 3, 4, 5, and 6, it can be observed that more than 70% of the CK remains isolated in a single cluster. However, a similar situation was not observed in the stone pine seedling traits.

Regarding climatic factors, the main drivers of significant effects on the morphometric traits of cones, seeds, and pine nuts were found to be Q (humidity category) and the mean of the maximum temperature in June, July, and August (ME) (Balekoglu et al. 2020). Within populations, the cone, seed characteristics, and germination behavior of *P. pinea* vary and are correlated with environmental or parent plant variables, such as stand age, canopy cover, and site conditions (Ganatsas et al. 2008). The weight of the pine nuts

was significantly affected by the amount of rainfall in the third year before harvest time (r: 0.889), despite the negative effect of summer rainfall. Additionally, the percentage of sound seeds (filled) per cone was negatively affected by rainfall in June of the third year before harvest (Balekoglu et al. 2020). The distinction of the Çanakkale-Kirazlı (CK) population from other populations in morphological, physiological, and biochemical studies conducted by Balekoglu et al. (2020, 2023a, b) on different populations of stone pines is also supported by the current study with CK is clustered in one group (96.23) (Table 10). According to the k-means algorithm, this study showed that the CK population differed according to seed and cone characteristics.

Considering that the shortening of vegetation from southern latitudes to northern latitudes and from low elevations to high elevations will affect the bud set, the geographical distribution of the populations may have caused this distinction (Balekoglu et al. 2020). This study revealed that the morphological characteristics of seedlings in the first year did not correlate with the second-year data. One of the reasons for this difference may be thought to be maternal effects in the first year. In such clustering and classification studies, performing at least two years of seedling morphological measurements is recommended. The results of the present study indicate that the BS can be an important parameter for population discrimination. As a final word, one of the limitations of the present study is that we did not have detailed site characteristics, such as soil and climate properties. The inclusion of soil and climate characteristics in the models can be included in future studies. On the other hand, the analysis in this study is limited by the parameters used. Therefore, researchers who want to conduct similar studies are advised to optimize the model parameters they use in order to avoid the overfitting problem, whether the analysis is descriptive or predictive.

Conclusions

In breeding programs, population selection represents the initial stage. It is of significant importance to ascertain the seed and seedling characteristics of populations, as this enables the differentiation of applications in seed transfer zones and geographic variation studies. Using ML algorithms to classify and cluster seeds and seedling populations in terms of morphological characteristics is beneficial for advancing forestry practices. We employed various ML techniques to analyze seed and seedling trait variations across the populations. The present research employs one unsupervised ML algorithm for seed and seedling clustering and six supervised ML algorithms for seed and seedling classification. The performances of six supervised ML algorithms were compared on the seed and seedling datasets. The RF algorithm achieved the best classification performance in terms of seed traits. Additionally, the best classification performance for stone pine seedlings was observed for the k-NN algorithm. The best clustering performance was achieved with $k=2$ for the seed and seedling traits. According to the PCA, two dimensions accounted for 97% and 63% of the traits of seeds and seedlings, respectively. The most important features between seed and seedling traits were cone weight (CW) and bud set (BS), respectively. The phenotype of a woody plant represents its unique morphological properties. In order to facilitate genetic improvement and conserve genetic diversity, it is essential to employ population discrimination and individual classification. This study will provide an infrastructure and inspiration for the realization of future tasks such as image recognition, image segmentation, breeding of populations, and conserving genetic diversity in forest management practices, especially regarding reforestation, yield optimization, and breeding programs.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10342-024-01716-7.

Acknowledgements We thank Istanbul Forest Nursery Manager Nejdet BALCI and the nursery staff for their great support. This work was partially supported by the Scientific Research Projects Coordination Unit of Istanbul University-Cerrahpaşa [Project No. FBA-2016-21357]. We also thank anonymous reviewers and editors for their valuable comments, significantly improving the original manuscript.

Author contributions All the authors contributed to the study's conception and design. Servet Caliskan: conceptualization, methodology, writing, original draft. Elif Kartal: conceptualization, methodology, analysis, writing, original draft. Safa Balekoglu: Laboratory and nursery activities, data collection, writing, review & editing. Fatma Çalışkan: conceptualization, methodology, writing, review & editing. All the authors have read and approved the final manuscript.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK).

Data availability No datasets were generated or analysed during the current study.

Declarations

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not

included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Balaban ME, Kartal E (2018) Veri Madenciliği Ve Makine Öğrenmesi Temel Algoritmaları Ve R Dili Ile Uygulamaları, 2nd edn. Çağlayan Kitabevi

Balekoglu S, Caliskan S, Dirik H (2020) Effects of geoclimatic factors on the variability in *Pinus pinea* cone, seed, and seedling traits in Turkey native habitats. Ecol Process 9(1):1–13. https://doi.org/10.1186/s13717-020-00264-3

Balekoglu S, Caliskan S, Makineci E, Dirik H (2021) Influence of seed nitrogen and carbon on germination in different populations of stone pine. Erwerbs Obstbau 63:369–374. https://doi.org/10.1007/s10341-021-00593-3

Balekoglu S, Caliskan S, Dirik H, Rosner S (2023a) Response to drought stress differs among *Pinus pinea* provenances. Ecol Manage 531:120779. https://doi.org/10.1016/j.foreco.2023.120779

Balekoglu S, Caliskan S, Makineci E, Dirik H (2023b) An experimental assessment of carbon and nitrogen allocation in *Pinus pinea* populations under drought stress and rewatering treatment. Environ Exp Bot 210:105334. https://doi.org/10.1016/j.envexpbot.2023.105334

Baylan KA, Ustaoğlu B (2020) Emberger biyoiklim sınıflandırmasına göre Türkiye'de Akdeniz Biyoiklim katlarının ve alt tiplerinin dağılışı. Ulusal Çevre Bilimleri Araştırma Dergisi 3(3):158–174

Begum S, Chakraborty D, Sarkar R (2015) Data classification using feature selection and kNN machine learning approach. In *2015 International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 811–814). IEEE

Boydak M, Çalışkan S (2014) Afforestation (in Turkish), 1st ed. Ankara, ISBN: 978-975-93943-8-7

Boydak M, Çalışkan S (2015) Afforestation in Arid and Semi-Arid Regions, first ed. Ankara

Boydak M, Çalışkan S (2021) Afforestation (in Turkish), 2st ed. Ankara

Bracalini M, Benedettelli S, Croci F, Terreni P, Tiberi R, Panzavolta T (2013) Cone and seed pests of *Pinus pinea*: assessment and characterization of damage. J Econ Entomol 106:229–234. https://doi.org/10.1603/EC12293

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. Taylor & Francis

Caballol M, Ridley M, Colangelo M, Valeriano C, Camarero JJ, Oliva J (2022) Tree mortality caused by Diplodia shoot blight on *Pinus sylvestris* and other mediterranean pines. Ecol Manage 505:119935. https://doi.org/10.1016/j.foreco.2021.119935

Calama R, Gordo J, Madrigal G, Mutke S, Conde M, Montero G, Pardos M (2016) Enhanced tools for predicting annual stone pine (*Pinus pinea* L.) cone production at tree and forest scale in Inner Spain. Syst 25:e079. https://doi.org/10.5424/fs/2016253-0967

Caliskan S, Makineci E (2014) Variations in carbon and nitrogen ratios and their effects on seed germination in *Cupressus sempervirens* populations. Scand J Res 29(2):162–169. https://doi.org/10.1080/02827581.2014.881544

Çalışkan S, Boydak M (2017) Afforestation of arid and semiarid ecosystems in Turkey. Turk J Agric for 41:317–330. https://doi.org/10.3906/tar-1702-39

Çalışkan S, Balekoglu S, Dirik H (2018) Seed and cone diversity and germination potential of stone pine provenances in different bioclimatic zones (in Turkish). BAP Project. FBA-2016-21357

Chen S, Dai D, Zheng J, Kang H, Wang D, Zheng X, Gu X, Mo J, Luo Z (2023) Intelligent grading method for walnut kernels based on deep learning and physiological indicators. Front Nutr 9:1075781. https://doi.org/10.3389/fnut.2022.1075781

Chirici G, Mura M, McInerney D, Py N, Tomppo EO, Waser LT, Travaglini D, McRoberts RE (2016) A meta-analysis and review of the literature on the k-Nearest neighbors technique for forestry applications that use remotely sensed data. Remote Sens Environ 176:282–294. https://doi.org/10.1016/j.rse.2016.02.001

R Core Team (2023) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org/

Daget P, Ahdali L, David P (1988) Mediterranean bioclimate and its variation in the Palaearctic region. In: Specht RL (ed) Mediterranean-type ecosystems, a data source book. Kluwer Academic, Dordrecht, pp 139–148

Dua S, Chowriappa P (2013) Data mining for bioinformatics. CRC

Duc NT, Ramlal A, Rajendran A, Raju D, Lal SK, Kumar S, Sahoo RN, Chinnusamy V (2023) Image-based phenotyping of seed architectural traits and prediction of seed weight using machine learning models in soybean. Front Plant Sci 14. https://doi.org/10.3389/fpls.2023.1206357

Fallour D, Fady B, Lefevre F (1997) Study on isozyme variation in *Pinus pinea* L.: evidence for low polymorphism. Silvae Genet 46(4):201–207

Fan Y, Ma S, Wu T (2020) Individual wheat kernels vigor assessment based on NIR spectroscopy coupled with machine learning methodologies. Infrared Phys Technol 105:103213. https://doi.org/10.1016/j.infrared.2020.103213

Farinha AO, Branco M, Pereira MF, Auger-Rozenberg MA, Maurício A, Yart A, Guerreiro V, Sousa EM, Roques A (2018) Micro X-ray computed tomography suggests cooperative feeding among adult invasive bugs *Leptoglossus occidentalis* on mature seeds of stone pine *Pinus pinea*. Agric Entomol 20:18–27. https://doi.org/10.1111/afe.12225

Ganatsas P, Tsakaldimi M, Thanos C (2008) Seed and cone diversity and seed germination of *Pinus pinea* in Strofylia site of the Natura 2000 Network. Biodivers Conserv 17:2427–2439. https://doi.org/10.1007/s10531-008-9390-8

Gómez-Ramírez J, Ávila-Villanueva M, Fernández-Blázquez MÁ (2020) Selecting the most important self-assessed features for predicting conversion to mild cognitive impairment with random forest and permutation-based methods. Sci Rep 10(1):20630. https://doi.org/10.1038/s41598-020-77296-4

Greenacre M, Groenen PJ, Hastie T, d'Enza AI, Markos A, Tuzhilina E (2022) Principal component analysis. Nat Rev Methods Primers 2(1):100. https://doi.org/10.1038/s43586-022-00184-w

Han J, Kamber M (2006) Data Mining: concepts and techniques, 2nd edn. Morgan Kaufmann

Han H, Guo X, Yu H (2016) Variable selection using Mean decrease Accuracy and Mean decrease Gini based on Random Forest. 2016 7th IEEE Int Conf Softw Eng Service Sci (ICSESS) 219–224. https://doi.org/10.1109/ICSESS.2016.7883053

Harrington P (2012) Machine learning in action, 1st edn. Manning Publications Co

Hlaiem S, Yangui I, Della Rocca G, Barberini S, Danti R, Ben Jamaa ML (2023) Diplodia species causing dieback on *Pinus pinea*: relationship between disease incidence, dendrometric and ecological parameters. J Sustainable for 42(1):59–76. https://doi.org/10.1080/10549811.2021.1944879

Huang B, Liu J, Jiao J, Lu J, Lv D, Mao J, Zhao Y, Zhang Y (2022) Applications of machine learning in pine nuts classification. Sci Rep 12(1):8799. https://doi.org/10.1038/s41598-022-12754-9

ICNF (2013) IFN6—Áreas Dos Usos do solo e das espécies florestais de Portugal continental. Resultados preliminares. Instituto da Conservação da Natureza e das Florestas, Lisboa

Jafarzadegan K, Merwade V, Moradkhani H (2020) Combining clustering and classification for the regionalization of environmental model parameters: application to floodplain mapping in data-scarce regions. Environ Modell Softw 125:104613. https://doi.org/10.1016/j.envsoft.2019.104613

Kartal E, Özyaprak M, Özen Z, Şimşek İ, Köse Biber S, Biber M, Can T (2020) Asking the right questions to nominate a student as gifted and talented: a Machine Learning Approach. Int J Inf Techn 13(4):385–400. https://doi.org/10.17671/gazibtd.591158

Kassambara A, Mundt F (2020) Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. https://CRAN.R-project.org/package=factoextra

Kuhn M (2008) Building Predictive models in R using the Caret Package. J Stat Softw 28(5):1–26. https://doi.org/10.18637/jss.v028.i05

Kuhn M, Quinlan R (2023) C50: C5.0 Decision Trees and Rule-Based Models. https://CRAN.R-project.org/package=C50

Kusumaningrum D, Lee H, Lohumi S, Mo C, Kim MS, Cho BK (2018) Non-destructive technique for determining the viability of soybean (Glycine max) seeds using FT-NIR spectroscopy. J Sci Food Agric 98:1734–1742. https://doi.org/10.1002/jsfa.8646

Lesieur V, Yart A, Guilbon S, Lorme P, Auger-Rozenberg MA, Roques A (2014) The invasive Leptoglossus seed bug, a threat for commercial seed crops, but for conifer diversity? Biol Invasions 16:1833–1849. https://doi.org/10.1007/s10530-013-0630-9

Liaw A, Wiener M (2002) Classification and regression by random forest. R News 2(3):18–22

Ma T, Tsuchikawa S, Inagaki T (2020) Rapid and non-destructive seed viability prediction using near-infrared hyperspectral imaging coupled with a deep learning approach. Comput Electron Agric 177:105683. https://doi.org/10.1016/j.compag.2020.105683

Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K (2022) cluster: Cluster Analysis Basics and Extensions. https://CRAN.R-project.org/package=cluster

McRoberts RE, Domke GM, Chen Q, Naesset E, Gobakken T (2016) Using genetic algorithms to optimize k-Nearest neighbors configurations for use with airborne laser scanning data. Remote Sens Environ. https://doi.org/10.1016/j.rse.2016.07.007

Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2023) E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. https://CRAN.R-project.org/package=e1071

Montagnoli A, Terzaghi M, Fulgaro N, Stoew B, Wipenmyr J, Ilver D, Rusu C, Scippa GS, Chiatante D (2016) Non-destructive phenotypic analysis of early-stage tree seedling growth using an automated stereovision imaging method. Front Plant Sci 7:1644. https://doi.org/10.3389/fpls.2016.01644

Murphy KP (2012) Machine learning: a probabilistic perspective. The MIT Press

Mutke S, Gordo J, Gil L (2005) Variability of Mediterranean stone pine cone production: yield loss as response to climate change. Agric Meteorol 132:263–272. https://doi.org/10.1016/j.agrformet.2005.08.002

Mutke S, Gordo J, Bono D, Gil L (2012) Mediterranean Stone pine: botany and horticulture. Hortic Rev 39:153–201. https://doi.org/10.1002/9781118100592.ch4

Nie P, Zhang J, Feng X, Yu C, He Y (2019) Classification of hybrid seeds using near-infrared hyperspectral imaging technology combined with deep learning. Sens Actuators B Chem 296:126630. https://doi.org/10.1016/j.snb.2019.126630

Niknejad N, Bidese-Puhl R, Bao Y, Payn KG, Zheng J (2023) Phenotyping of architecture traits of loblolly pine trees using stereo machine vision and deep learning: stem diameter, branch angle, and branch diameter. Comput Electron Agric 211:107999. https://doi.org/10.1016/j.compag.2023.107999

OGM (2021) Türkiye Orman Varlığı (in Turkish). Orman Genel Müdürlüğü, Ankara. ISBN 978-605-7599-68-1

Öngen Bilir B, Kardeş S (2023) Temel Bileşenler Analizi. In: Özen Z, Kartal E (eds) Denetimsiz Makine Öğrenmesi Algoritmaları: R ve Python Uygulamaları, 1st edn. Nobel Akademik Yayıncılık, pp 1–19

Osako Y, Yamane H, Lin SY, Chen PA, Tao R (2020) Cultivar discrimination of litchi fruit images using deep learning. Sci Hortic 269:109360. https://doi.org/10.1016/j.scienta.2020.109360

Owino JO, Angaine PM, Onyango AA, Ojunga SO, Otuoma J (2020) Evaluating variation in seed quality attributes in *Pinus patula* clonal orchards using cone cluster analysis. J Forests 7(1):1–8

Parlak S (2017) An invasive species: *Leptoglossus occidentalis* (Heidemann) how does it affect forestry activities? Kast Univ J Fac 17:531–542

Pereira S, Prieto A, Calama R, Diaz-Balteiro L (2015) Optimal management in *Pinus pinea* L. stands combining silvicultural schedules for timber and cone production. Silva Fenn 49:1226. https://doi.org/10.14214/sf.1226

Posit (2023) RStudio IDE. Posit. https://www.posit.co/

Prodhan FA, Zhang J, Hasan SS, Sharma TPP, Mohana HP (2022) A review of machine learning methods for drought hazard monitoring and forecasting: current research trends, challenges, and future research directions. Environ Modell Softw 149:105327. https://doi.org/10.1016/j.envsoft.2022.105327

Qi H, Huang Z, Jin B, Tang Q, Jia L, Zhao G, Zhang C (2024) SAM-GAN: an improved DCGAN for rice seed viability determination using near-infrared hyperspectral imaging. Comput Electron Agric 216:108473. https://doi.org/10.1016/j.compag.2023.108473

Quinlan JR (2022) Data Mining Tools See5 and C5.0. https://www.rulequest.com/

Rao D, Zhang D, Lu H, Yang Y, Qiu Y, Ding M, Yu X (2023) Deep learning combined with Balance Mixup for the detection of pine wilt disease using multispectral imagery. Comput Electron Agric 208:107778. https://doi.org/10.1016/j.compag.2023.107778

Rousseeuw P, Rousseeuw PJ (1987) Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Comput. Appl. Math. 20, 53–65. J. Comput Appl Math 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Shah K, Wenqi L, Raezah AA, Khan N, Khan SU, Ozair M, Ahmad Z (2024) Unraveling pine wilt disease: comparative study of stochastic and deterministic model using spectral method. Expert Syst Appl 240:122407. https://doi.org/10.1016/j.eswa.2023.122407

Shmueli G, Bruce PC, Yahav I, Patel NR, Lictendahl KC (2018) Data Mining for Business Analytics, 1st edn. Wiley

Skowronski L, de Moraes PM, de Moraes MLT, Goncalves WN, Constantino M, Costa CS, Costa RB (2021) Supervised learning algorithms in the classification of plant populations with different degrees of kinship. Brazilian J Bot 44(2):371–379. https://doi.org/10.1007/s40415-021-00703-1

Snodgress MA (2023) cpfa: Classification with Parallel Factor Analysis. https://CRAN.R-project.org/package=cpfa

Sotomayor LN, Cracknell MJ, Musk R (2023) Supervised machine learning for predicting and interpreting dynamic drivers of plantation forest productivity in northern Tasmania, Australia. Comput Electron Agric 209:107804. https://doi.org/10.1016/j.compag.2023.107804

Therneau T, Atkinson B (2022) rpart: Recursive Partitioning and Regression Trees. https://CRAN.R-project.org/package=rpart using supervised learning. PloS one 6(5). e14802

Tian H, Zhu J, He X, Chen X, Jian Z, Li C, Xiao W (2022) Using machine learning algorithms to estimate stand volume growth of *Larix* and *Quercus* forests based on national-scale forest inventory data in China. Ecosyst 9:100037. https://doi.org/10.1016/j.fecs.2022.100037

Vapnik V (1995) The nature of statistical learning theory. Springer

Venables WN, Ripley BD (2002) Modern Applied Statistics with S (Fourth). Springer. https://www.stats.ox.ac.uk/pub/MASS4/

Vendramin GG, Fady B, González-Martínez SC, Hu FS, Scotti I, Sebastiani F, Petit RJ (2008) Genetically depauperate but widespread: the case of an emblematic Mediterranean pine. Evolution 62:680–688. https://doi.org/10.1111/j.1558-5646.2007.00294.x

Walesiak M, Dudek A (2020) The Choice of Variable Normalization Method in Cluster Analysis. In: Soliman KS (ed) Education Excellence and Innovation Management: a 2025 vision to Sustain Economic Development during Global challenges. International Business Information Management Association (IBIMA), pp 325–340

Wang C, Liu B, Liu L, Zhu Y, Hou J, Liu P, Li X (2021) A review of deep learning used in the hyperspectral image analysis for agriculture. Artif Intell Rev 54(7):5205–5253. https://doi.org/10.1007/s10462-021-10018-y

Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer-, New York. https://ggplot2.tidyverse.org

Yang X, Zhang R, Zhai Z, Pang Y, Jin Z (2019) Machine learning for cultivar classification of apricots (*Prunus armeniaca* L.) based on shape features. Sci Hortic 256:108524. https://doi.org/10.1016/j.scienta.2019.05.051

Zhang T, Wei W, Zhao B, Wang R, Li M, Yang L, Wang J, Sun Q (2018) A reliable methodology for determining seed viability by using hyperspectral data from two sides of wheat seeds. Sens (Switzerland) 18. https://doi.org/10.3390/s18030813

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.