ATATÜRK ÜNİVERSİTESİ YAYINLARI
ATATURK UNIVERSITY PUBLICATIONS

Research Article *Araştırma Makalesi* DOI: 10.17097/agricultureatauni.1738386

The Language of Soil: Soil Analysis with a Machine Learning Approach

Toprağın Dili: Makine Öğrenimi Yaklaşımı ile Toprak Analizi

Semanur BAYRAM [1]

Elif Sude ÇAYIR [1]

Ebrar İLHAN [1]

İrem ARSLAN [1]

Esra GÖKPINAR [1]

[1]: Gazi University, Faculty of Science, Department of Statistics, Ankara, Türkiye

ABSTRACT

In Türkiye, rapid population growth combined with unsustainable agricultural practices threatens the sustainable use of fertile soils and poses serious risks for the agricultural sector. To address this challenge, the present study analyzes soil data from the Odunpazarı district of Eskişehir province and proposes a machine learning–based approach. First, Principal Component Analysis (PCA) was applied to reduce data dimensionality, after which the K-Means algorithm classified the soils into three clusters. These clusters revealed significant differences in physical structure, moisture, salinity, and mineral composition, thereby providing a robust basis for further modeling. Building on this foundation, supervised machine learning models were developed and their performances compared. Logistic Regression achieved the highest accuracy (98.9%), followed by Decision Tree (97.8%), Random Forest (97.2%), and K-Nearest Neighbors (91.7%). The findings demonstrate that machine learning algorithms can reliably predict soil group membership and generate valuable insights for regional soil productivity analysis. Overall, the study highlights the effectiveness of data-driven methods in supporting sustainable agricultural planning and offers an integrative model that can guide future applications in precision agriculture.

Keywords: Supervised machine learning, Unsupervised machine learning, Cluster analysis, Sustainable agriculture, Productivity

ÖZ

Türkiye’de artan nüfus ve bilinçsiz tarım uygulamaları, verimli toprakların sürdürülebilir kullanımını tehdit etmekte ve tarım sektöründe ciddi riskler oluşturmaktadır. Bu çalışmada, söz konusu soruna çözüm arayışıyla Eskişehir ili Odunpazarı ilçesine ait toprak analiz verileri incelenmiş ve makine öğrenmesi tabanlı bir yaklaşım geliştirilmiştir. Öncelikle, veri boyutunu azaltmak amacıyla Temel Bileşenler Analizi (PCA) uygulanmış, ardından K-Ortalama (K-Means) algoritmasıyla topraklar üç kümeye ayrılmıştır. Kümeler; fiziksel yapı, nem, tuzluluk ve mineral içerikleri açısından anlamlı farklılıklar göstermiş, böylece sınıflandırma süreci için sağlam bir temel oluşturmuştur. Bu aşamanın ardından kümelenen veriler kullanılarak denetimli makine öğrenmesi modelleri geliştirilmiş ve performansları karşılaştırılmıştır. Lojistik Regresyon modeli %98,9 doğruluk ile en yüksek başarıyı elde ederken, Karar Ağacı %97,8, Rastgele Orman %97,2 ve K-En Yakın Komşu (KNN) %91,7 doğruluk oranına ulaşmıştır. Bulgular, makine öğrenmesi algoritmalarının toprak gruplarını güvenilir biçimde tahmin edebildiğini ve bölgesel toprak verimliliği analizlerinde değerli katkılar sunduğunu ortaya koymaktadır. Sonuç olarak çalışma, akıllı tarım uygulamaları için veri odaklı karar destek sistemlerinin geliştirilmesine yönelik örnek bir model sunmaktadır.

Anahtar Kelimeler: Denetimli makine öğrenmesi, Denetimsiz makine öğrenmesi, Kümeleme analizi, Sürdürülebilir tarım, Verimlilik

This study was supported within the scope of the TÜBİTAK 2209-A Program for Supporting Research Projects of Undergraduate Students. It was also prepared as part of the "Research Project" course conducted under our university's undergraduate curriculum. The poster of the study was presented to faculty members at our university.

Received Date:	09.07.2025
Revision Request Date:	17.08.2025
Last Revision Date:	16.09.2025
Accepted Date:	17.09.2025
Publication Date:	26.09.2025

Corresponding author / Sorumlu Yazar: Semanur BAYRAM
E-mail: sema.bayram.2002@gmail.com

Cite this article: Bayram, S., Çayır, E.S., İlhan, E., Arslan, İ. & Gökpınar, E. (2025). The Language of Soil: Soil Analysis with a Machine Learning Approach. *Research in Agricultural Sciences, 56*(3), 243-255.

Introduction

Agriculture is a fundamental activity that provides food and raw materials to people through crop cultivation and animal husbandry. It makes significant contributions to the gross national product of many countries. In addition, it supplies raw materials to sectors such as textiles, the food industry, energy, and biotechnology. However, climate change, the growth of the world population, and the reduction of agricultural resources have caused serious problems in agriculture. Mismanagement practices such as improper land use, overgrazing, faulty crop rotation, and unbalanced fertilizer application are gradually reducing agricultural lands, thereby making agriculture a strategic priority (Demir et al., 2023; Kılavuz & Erdem, 2019). Moreover, fluctuations in the global economy directly affect the performance of enterprises at both national and regional levels. Therefore, countries aiming for economic growth and development should consider the importance of agriculture and base their strategic policies on sound analyses (Esmer & Gezer, 2021).

The inadequacy of traditional farming methods has brought smart farming applications to the forefront. Smart farming does not only aim to increase soil fertility but also seeks to ensure the efficient and sustainable development of agriculture. These approaches minimize environmental damage and enable the effective use of natural resources.

Higher yields can be achieved with less water, fertilizer, and fuel. At the same time, more crops can be cultivated in smaller areas, reducing farmers' costs. Big data analytics enables the analysis of variables such as climate conditions and soil fertility, thereby optimizing processes. At this point, machine learning (ML) methods play an active role in data analysis and prediction (Demir et al., 2023). Thus, innovations ranging from digital technologies to autonomous systems are being implemented in smart agriculture.

The primary goal of agriculture is crop production, and with population growth, the demand for food continues to rise. A review of the literature reveals numerous studies on crop management. For instance, Reddy et al. (2019) developed a crop recommendation system using ML algorithms in the Ramtek region of India. Garanayak et al. (2021) analyzed climate and soil data with ML regression methods to improve soil fertility in the Andhra Pradesh region. Paudel et al. (2022) proposed regional ML models for crop yield prediction at multiple spatial levels. Patel and Patel (2023) aimed to improve crop yield and optimize resource use with ML methods in Gujarat. Bhargavi and Jagannathan (2024) used weather, soil, and location data to predict crop yields in the Maharashtra and Karnataka regions. Burhan and Soydan (2023) predicted production quantities and yields for 2021–2022 using datasets provided by the Turkish Grain Board (TMO) and the Turkish State Meteorological Service (MGM). Prity et al. (2024) developed an ML-based crop recommendation system. Yakut et al. (2023) analyzed more than 7,500 soil data samples from Isparta using ML algorithms to determine which soils are more suitable for which crops.

Crop cultivation depends directly on soil quality and nutrient content, while fertilization, modern irrigation, seed improvement, and erosion prevention play a key role in enhancing fertility. However, continuous cropping depletes nutrients and reduces soil fertility, which is usually determined by nutrient presence or absence (Gruhn et al., 2000). Sustainable soil fertility depends on the soil's ability to supply nutrients to plants. Therefore, assessing soil fertility is a critical aspect of sustainable agriculture (Maathuis, 2009).

The literature shows that crop management studies also consider the physical, biological, and mineral characteristics of the soil. Soil classification based on fertility is used to identify nutrient deficiencies and to develop crop recommendation systems (Taher et al., 2021). Once soil data are obtained, they can be analyzed with ML algorithms for classification, enabling fertilizer and treatment recommendations. For example, Bhargavi and Jyothi (2011) analyzed soil data using data mining techniques. Hayattu et al. (2020) examined soil analysis data from northwestern Nigeria with similar approaches. Yadav et al. (2021) applied various ML algorithms to group and classify soils.

Soil fertility differs across regions, and crops are selected accordingly. Studies in the literature have been conducted by considering the factors affecting soil fertility in different countries. However, studies that comprehensively classify soil properties in Turkey using ML algorithms are very limited. A holistic evaluation of Turkey's soil analysis data and their classification through machine learning methods represents an important research need.

In this study, soil analysis data obtained from the Odunpazarı district of Eskişehir, provided by the Ministry of Agriculture and Forestry of the Republic of Turkey, were used. The physical and chemical components of the soil (e.g., soil depth, water saturation, total salt, lime, sand, clay, calcium, magnesium, boron, sodium, saturation, and other variables) were considered. First, the K-means clustering algorithm was applied to group soils, allowing meaningful differentiation and evaluation of fertility levels. Then, these clusters were used as labels to train supervised machine learning models. The aim was to predict the cluster of a randomly selected soil sample. The findings are expected to guide farmers in soil management practices and contribute to future research. In this context, this study aims to provide benefits such as accurate crop selection for farmers, supporting sustainable agricultural practices, preventing soil degradation, reducing costs through efficient land use, and minimizing environmental impacts.

Methods

In this study, similar soil types were grouped using the K-Means clustering algorithm after data pre-processing. The overall workflow at this stage of the study is presented schematically in Figure 1.

In the first stage, clustering analysis was performed, and soils were categorized according to their characteristics, allowing an evaluation of their soil fertility. In the second stage, a classification step was initiated to predict the group membership of a randomly selected soil sample.

For this purpose, the cluster labels were used to train various machine learning models, including Random Forest, Logistic Regression, Decision Tree, and K-Nearest Neighbour (KNN), and their performances were compared. The process is illustrated in Figure 2.

In addition, the overall two-stage workflow of the study is summarized in Figure 3, highlighting Stage 1 (unsupervised clustering) and Stage 2 (supervised classification).

Data set

The dataset used in this study consists of soil analysis data from the Odunpazarı region of Eskişehir province, obtained from the Ministry of Agriculture and Forestry. The dataset includes the physical and chemical properties of soil samples collected from various locations to evaluate soil fertility. These properties encompass parameters such as water saturation, total salt, lime, sand, clay, silt, calcium, magnesium, boron, sodium, electrical conductivity (EC), pH, sodium adsorption ratio (SAR), and exchangeable sodium percentage (ESP).

Descriptive statistics for the main variables are presented in Table 1. The average water saturation rate is 83.75%, with values ranging from 50.6% to 136.4%, indicating substantial variation in the soils' water holding capacity. The pH is slightly alkaline, averaging 7.74 (range 7.28–8.19). Total salt content is low (average 0.034%), suggesting generally low salinity levels in the region.

The proportions of sand, clay, and silt among the physical components are 30.29%, 45.40%, and 23.66%, respectively. These ratios indicate that the soils are generally clayey-loamy. Regarding chemical properties, the calcium (1.67 meq/L) and magnesium (0.66 meq/L) values provide insight into the mineral content of the soils. Additionally, the average SAR and ESP values (0.24 and 1.60, respectively) suggest that the sodicity level is low.

In general, these descriptive statistics show that the soils of Eskişehir Odunpazarı region have a heterogeneous physical and chemical structure, and this diversity will contribute to the correct grouping of soils in machine learning models.

Figure 1.

Workflow diagram of clustering analysis according to soil properties

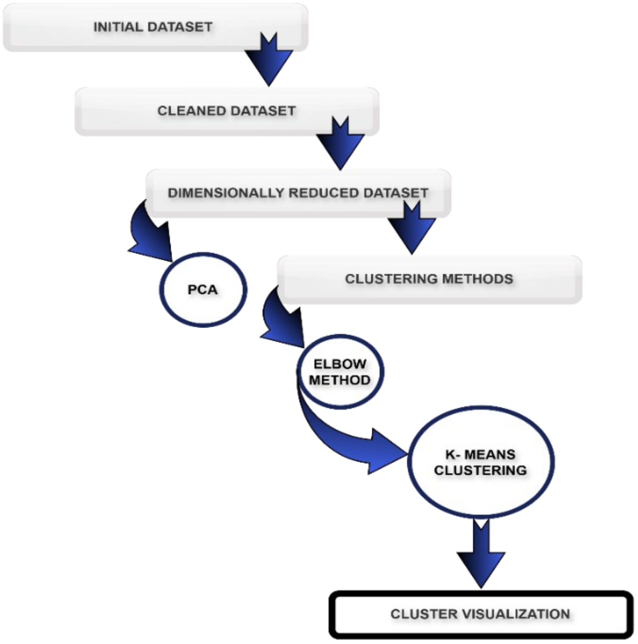


Figure 2.

Workflow diagram of soil group prediction and model performance evaluation process

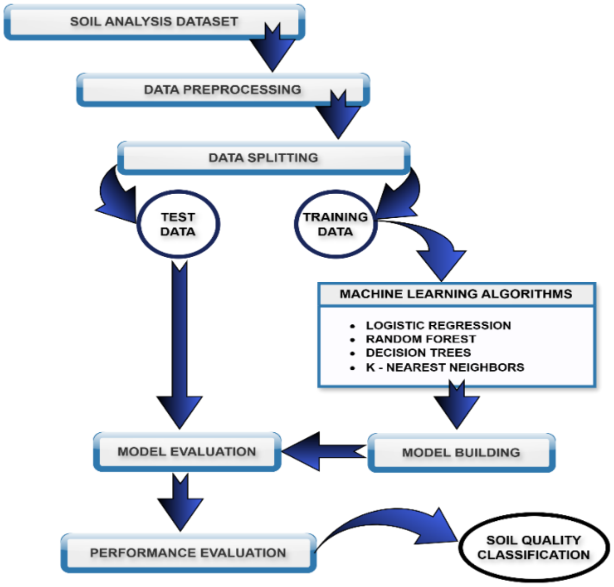


Figure 3.

Overall two-stage workflow of the study

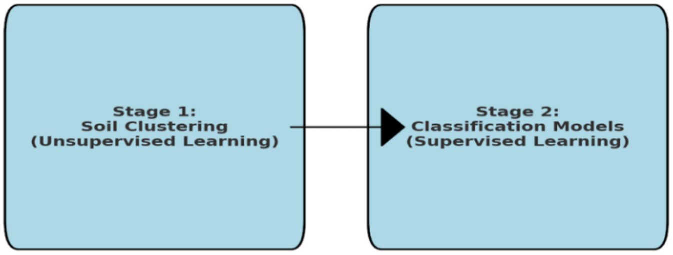


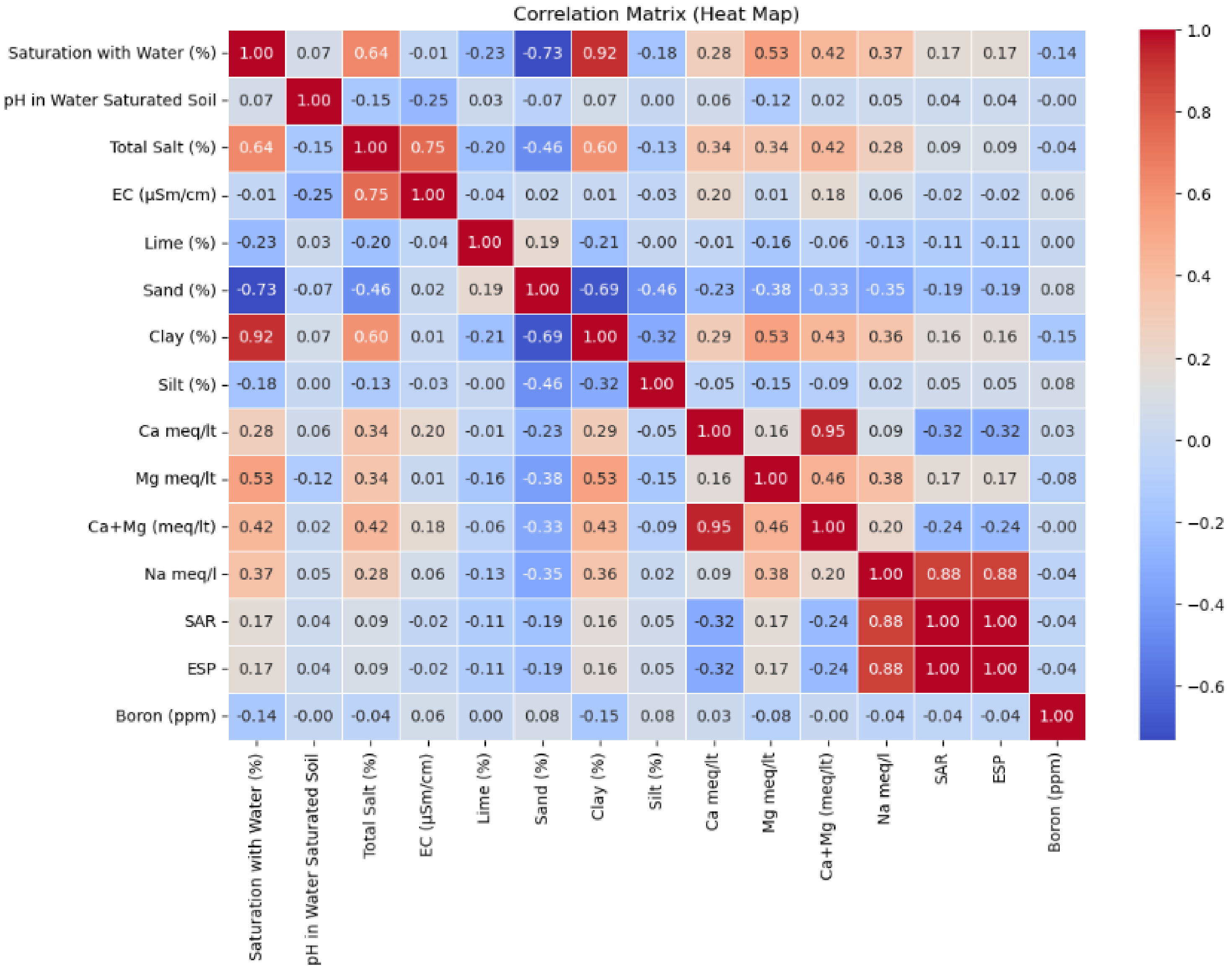
Table 1.

Summary descriptive statistics for numerical variables

	Count	Average	Std. Deviation	Minimum	25%	50%	75%	Maximum
Saturation with Water (%)	905	83.754	18.015	50.600	70.400	81.400	95.700	136.400
pH in Water Saturated Soil	905	7.7431	0.168	7.280	7.630	7.730	7,.860	8.190
Total Salt (%)	905	0.0343	0.011	0.011	0.026	0.034	0.042	0.067
EC (µSm/cm)	905	0.6427	0.153	0.291	0.539	0.641	0.741	1.056
Lime (%)	905	14.442	9.468	0.186	6.460	13.056	21.660	36.267
Sand (%)	905	30.924	7.300	9.010	25.360	32.410	36.580	48.520
Clay (%)	905	45.405	6.854	28.560	39.580	45.160	50.140	63.540
Silt (%)	905	23.669	5.566	7.420	19.650	23.540	27.590	39.970
Ca meq/l	905	1.673	0.812	0.007	1.082	1.566	2.174	4.202
Mg meq/l	905	0.663	0.281	0.009	0.451	0.629	0.835	1.484
Ca+Mg (meq/l)	905	2.335	0.905	0.000	1.677	2.247	2.912	4.897
Na meq/l	905	0.248	0.108	0.005	0.167	0.238	0.311	0.566
SAR	905	0.238	0.103	0.033	0.164	0.223	0.306	0.518
ESP	905	1.606	0.148	1.308	1.498	1.584	1.704	2.008
Boron (ppm)	905	0.230	0.119	0.006	0.144	0.217	0.305	0.558

Figure 4.

Correlation matrix of variables



	Saturation with Water (%)	pH in Water Saturated Soil	Total Salt (%)	EC (µSm/cm)	Lime (%)	Sand (%)	Clay (%)	Silt (%)	Ca meq/lt	Mg meq/lt	Ca+Mg (meq/lt)	Na meq/l	SAR	ESP	Boron (ppm)
Saturation with Water (%)	1.00	0.07	0.64	-0.01	-0.23	-0.73	0.92	-0.18	0.28	0.53	0.42	0.37	0.17	0.17	-0.14
pH in Water Saturated Soil	0.07	1.00	-0.15	-0.25	0.03	-0.07	0.07	0.00	0.06	-0.12	0.02	0.05	0.04	0.04	-0.00
Total Salt (%)	0.64	-0.15	1.00	0.75	-0.20	-0.46	0.60	-0.13	0.34	0.34	0.42	0.28	0.09	0.09	-0.04
EC (µSm/cm)	-0.01	-0.25	0.75	1.00	-0.04	0.02	0.01	-0.03	0.20	0.01	0.18	0.06	-0.02	-0.02	0.06
Lime (%)	-0.23	0.03	-0.20	-0.04	1.00	0.19	-0.21	-0.00	-0.01	-0.16	-0.06	-0.13	-0.11	-0.11	0.00
Sand (%)	-0.73	-0.07	-0.46	0.02	0.19	1.00	-0.69	-0.46	-0.23	-0.38	-0.33	-0.35	-0.19	-0.19	0.08
Clay (%)	0.92	0.07	0.60	0.01	-0.21	-0.69	1.00	-0.32	0.29	0.53	0.43	0.36	0.16	0.16	-0.15
Silt (%)	-0.18	0.00	-0.13	-0.03	-0.00	-0.46	-0.32	1.00	-0.05	-0.15	-0.09	0.02	0.05	0.05	0.08
Ca meq/lt	0.28	0.06	0.34	0.20	-0.01	-0.23	0.29	-0.05	1.00	0.16	0.95	0.09	-0.32	-0.32	0.03
Mg meq/lt	0.53	-0.12	0.34	0.01	-0.16	-0.38	0.53	-0.15	0.16	1.00	0.46	0.38	0.17	0.17	-0.08
Ca+Mg (meq/lt)	0.42	0.02	0.42	0.18	-0.06	-0.33	0.43	-0.09	0.95	0.46	1.00	0.20	-0.24	-0.24	-0.00
Na meq/l	0.37	0.05	0.28	0.06	-0.13	-0.35	0.36	0.02	0.09	0.38	0.20	1.00	0.88	0.88	-0.04
SAR	0.17	0.04	0.09	-0.02	-0.11	-0.19	0.16	0.05	-0.32	0.17	-0.24	0.88	1.00	1.00	-0.04
ESP	0.17	0.04	0.09	-0.02	-0.11	-0.19	0.16	0.05	-0.32	0.17	-0.24	0.88	1.00	1.00	-0.04
Boron (ppm)	-0.14	-0.00	-0.04	0.06	0.00	0.08	-0.15	0.08	0.03	-0.08	-0.00	-0.04	-0.04	-0.04	1.00

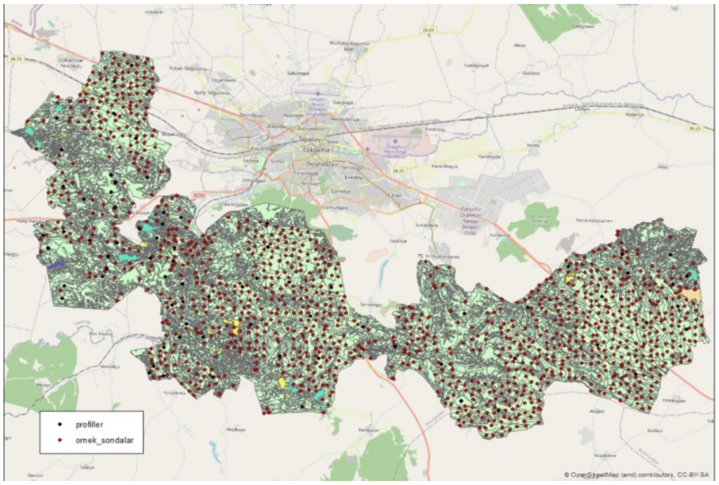
The relationships between soil variables were analyzed using the correlation matrix (heat map) in Figure 4. A very strong positive correlation (.92) was found between water saturation rate and clay (%), indicating that soils retain more water as clay content increases. In contrast, water saturation showed a strong negative correlation (-.73) with sand (%), suggesting lower water retention in sandy soils.

Total salt content was positively correlated with electrical conductivity (EC) (.75), confirming that salt increases ionic conductivity. Calcium (Ca) showed a very high correlation (.95) with Ca+Mg, highlighting the need to evaluate these elements together for mineral balance.

On the other hand, very strong positive relationships were found between sodium (Na) and SAR (Sodium Adsorption Rate) with a coefficient of .88 and between SAR and ESP (Exchangeable Sodium Percentage) with a coefficient of 1.00. This shows that soil sodicity parameters are closely related to each other and provides critical information for salinity management. In general, correlation analysis reveals how the physical and chemical properties of soil components are related to each other, which provides an important basis for variable selection in machine learning models.

Figure 5.

Distribution of raw data



In Figure 5, the distributions of the locations of the soil observations of the Odunpazarı region are visualized with the help of ArcGIS software. Before the analysis, the data was cleaned, missing values were removed, and outliers were checked. In addition, dimensionality reduction was performed using Principal Component Analysis (PCA) and the data set was made suitable for clustering and classification analyses.

Methods used

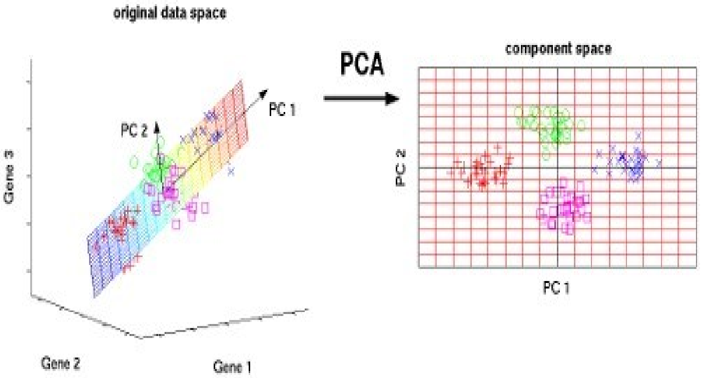
Principal Component Analysis (PCA)

PCA represents multivariate data with fewer variables, reducing dimensionality with minimal information loss. It creates new, uncorrelated variables called principal components, thereby removing dependencies among the original variables (Ersungur et al., 2007).

Principal component analysis aims to determine the best transformation that can express the available data with fewer variables. The variables obtained after the transformation are called principal components of the initial variables. The first principal component is the one with the largest variance value and the other principal components are ranked in order of decreasing variance values. The main advantages of this method are low sensitivity to noise, reduced memory and capacity requirements, and more efficient operation in low-dimensional spaces (Brownlee, 2016; Koldere, 2008).

Figure 6.

PCA model



Principal component analysis performs dimension reduction by representing multidimensional data with fewer components as shown in Figure 6. In addition, as can be seen in the figure, the three-dimensional data in the original data space is reduced to two main components called PC1 and PC2. This transformation preserves most of the total variance in the data set, allowing the data to be expressed in a simpler and more comprehensible component space.

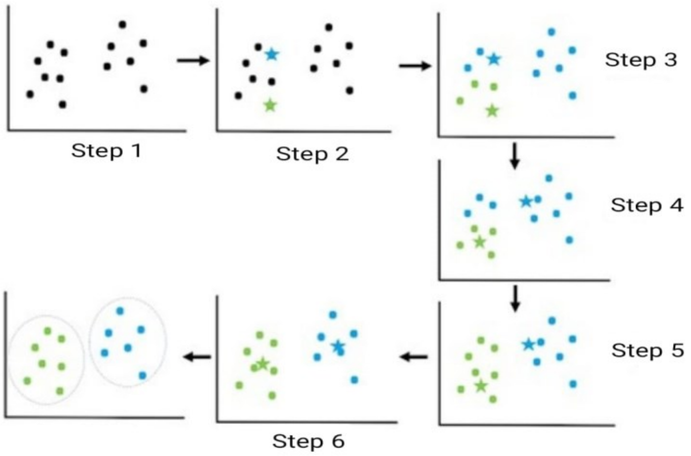
K-means clustering method

K-Means is a fundamental unsupervised learning algorithm for clustering. It groups unlabeled data by assigning similar observations to the same cluster. The algorithm determines k centroids and assigns each data point to the nearest one, iteratively optimizing cluster similarity. The parameter k defines the number of clusters to be formed, and this process continues until the similarity between the data is maximized (Brownlee, 2016; Koldere, 2008).

As can be seen in Figure 7, in Step 1, the data are initially scattered and not clustered. Then two cluster centers are selected as blue and green stars. The data are assigned to clusters according to their proximity to these centers and the centers for each cluster are renewed by averaging the data in that cluster. This process is repeated until the centers remain constant and the clusters become distinct.

Figure 7.

Example of clustering model with k-means algorithm



From this, the Euclidean distance $d(x_i, c_j)$ of the data point x_i to the centre c_j is calculated as follows.

$$d(x_i, c_j) = \sqrt{\sum_{k=1}^{n}(x_{ik} - c_{jk})^2} \qquad [1]$$

Where n represents the size of the data point (number of features/parameters).

The new center account,

$$c_j = \frac{1}{N_j}\sum_{x_i \in c_j} x_i \qquad [2]$$

c_j, calculates the centre of the jth cluster. N_j, denotes the number of data points in cluster j. and C_j denotes all data points in that cluster.

Classification algorithms

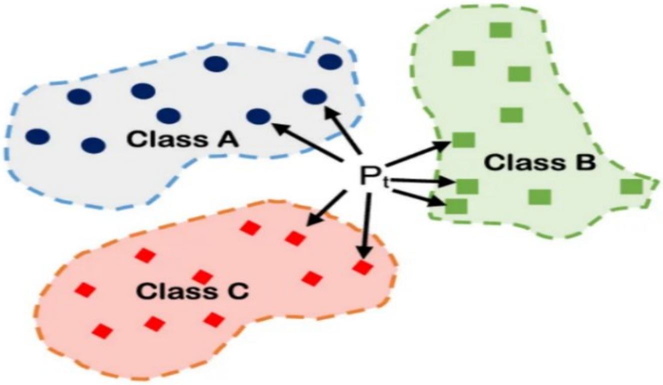
KNN algorithm

The main objective in classification problems is to accurately predict the classes to which the observations belong. The general purpose of the KNN algorithm, which is widely used in this context, is to assign observations to predetermined classes according to their own characteristics, as can be seen in Figure 8. In addition, the classification of a new observation is also provided. The new observation to be classified is classified into the same group with the k closest observations with the help of the learning dataset (Ağlarcı & Karakurt, 2024). This method, which assumes that data with similar characteristics are usually located close to each other, is based on neighbourhood relations when determining the class of new observations. The Euclidean distance is generally preferred for distance measurement. However, alternative distance measures, such as Manhattan and Minkowski, can also be used. The Euclidean distance measures the straight-line distance between two points and is particularly suitable for continuous variables. The Manhattan distance, on the other hand, is based on the sum of the absolute differences between two vectors. Finally, the Minkowski distance provides a generalized form of these two measurements and can be reduced to different distance types depending on the chosen parameter value. Performance metrics such as accuracy, precision, F1 score are used to evaluate the performance of the model.

In this study, three distinct clusters generated using the K-Means algorithm were designated as class labels. The Euclidean distance was employed to quantify the distances between observations, and the KNN algorithm was subsequently trained based on these labels.

Figure 8.

Example of classification model with KNN algorithm



Logistic regression algorithm

Logistic regression is a basic method of probability-based classification with two or more independent variables when the dependent variable is categorical. Logistic regression, unlike linear regression, uses a logistic function that limits the output values between 0 and 1. In this respect, it offers effective decision mechanisms in classification problems. In addition, it is widely preferred due to its simple structure and fast computational capability. Logistic regression can be applied in different types for binary, multiple and sequential classification problems (Hamid et al., 2018).

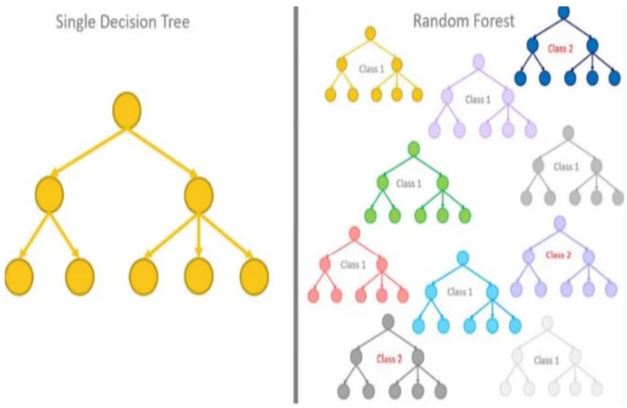
In this study, due to the independent and multi-categorized structure of the clusters, binary logistic regression was not sufficient; instead, multinomial logistic regression model, which is suitable for multi-class problems, was preferred. The model predicted which of the three classes each observation belongs to on a probability basis through the softmax function.

Random forest algorithm

Random Forest is one of the supervised learning methods and is widely used in classification and regression problems. As can be seen in Figure 9, this method is an ensemble model consisting of multiple decision trees. Each tree is trained with a random subset of the training data and different random features are used at each node. The decision made by each tree is considered when making predictions; the majority vote is taken in classification problems, and the average value is taken in regression problems. This structure prevents the model from overlearning and increases its generalization capability. Random Forest provides high success especially in large and complex data sets. In addition, thanks to its feature of determining the importance of variables, it also shows which variables are effective on prediction (Kumral et al., 2022).

Figure 9.

Example of classification model with random forest algorithm



Decision tree algorithm

Decision trees are among the most widely used supervised learning methods in classification and regression problems. Their main advantage lies in their ease of construction and the interpretability of the results they produce. The structure of decision trees consists of nodes, branches, and leaves. Data are split from the root node into branches and ultimately classified in the leaves. This allows the decision-making process to be followed step by step, facilitates evaluation of the resulting structure, and enables direct application to new data.

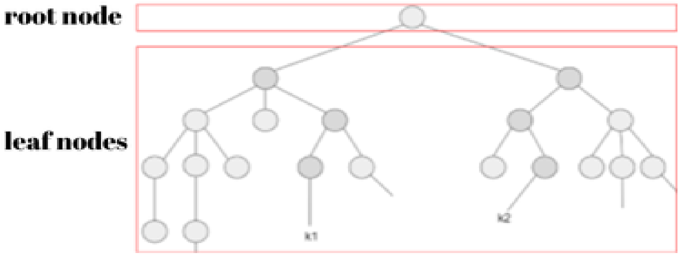
The performance of decision trees depends on several key parameters. Maximum depth determines the maximum number of layers a tree can have, while the minimum number of samples per leaf enhances the model's generalization ability. As splitting criteria, Gini Index or Entropy are commonly used in classification, whereas Mean Squared Error (MSE) is employed in regression. Additionally, limiting the maximum number of features improves efficiency and helps prevent overfitting.

The most important advantage of decision trees is their interpretability. However, very deep trees may become sensitive to noise and prone to overfitting. For this reason, decision trees are often combined with ensemble methods such as Random Forests or Gradient Boosted Trees, which provide higher accuracy and better generalization.

Overall, decision trees stand out as a valuable machine learning method due to their transparent structure and ability to generate explicit decision rules. They are particularly useful in fields such as agriculture, healthcare, and environmental studies, where interpretability is of critical importance. The graphical representation of the model is given in Figure 10.

Figure 10.

Example of classification model with decision tree algorithm



All classification algorithms were implemented using the scikit-learn library in Python. Default parameter settings were employed unless otherwise stated. Specifically, KNN was applied with Euclidean distance and k=5 neighbors, Random Forest with 100 trees (n_estimators=100) and unrestricted depth, and Decision Tree with unlimited depth and Gini impurity as the splitting criterion. Logistic Regression was applied with the 'lbfgs' solver and default regularization parameter (C=1.0). These settings correspond to the standard default values in scikit-learn (version 1.2.2).

Results and Discussion

In this study, by applying PCA, dimensionality reduction was made by reducing many variables in the dataset. In addition, the situation of affecting the performance of the classification models due to the high correlation between the features was eliminated. As a result, six components were determined, explaining a total variance of 80%.

Table 2 shows the component load matrix, which illustrates the relationships between components and variables. Positive values indicate a direct relationship, whereas negative values reflect an inverse relationship. Variables with higher absolute values contribute more significantly to the explanation of the corresponding component. For the components, variables with loading values of 0.30 and above were considered dominant.

Table 2.
Component load matrix

	PC1	PC2	PC3	PC4	PC5	PC6
Saturation with Water (%)	0.415586	-0.061068	0.198644	0.113304	-0.199862	0.033530
pH in Water Saturated Soil	0.002300	0.038581	0.416919	-0.033836	0.282925	0.097506
Total Salt (%)	0.351932	-0.142861	-0.396154	-0.001432	-0.186328	-0.047222
EC (µSm/cm)	0.107791	-0.130041	-0.685855	-0.090043	-0.065777	-0.104965
Lime (%)	-0.137459	-0.032708	0.007639	0.006779	0.338558	-0.592244
Sand (%)	-0.349169	-0.000206	-0.241574	0.405781	0.283475	0.084129
Clay (%)	0.409340	-0.072509	0.188487	0.213763	-0.159239	0.040214
Silt (%)	-0.046109	0.089572	0.084751	-0.795565	-0.175738	-0.159886
Ca meq/l	0.201089	-0.384783	-0.001887	-0.173296	0.456553	-0.045457
Mg meq/l	0.311076	-0.024382	0.072117	0.153362	0.007949	0.060140
Ca+Mg (meq/l)	0.278631	-0.353852	0.022012	-0.106320	0.414892	-0.024034
Na meq/l	0.310514	0.361122	-0.086779	-0.051507	0.333873	-0.037458
SAR	0.182149	0.518991	-0.101796	-0.003153	0.157315	-0.020843
ESP	0.182337	0.518787	-0.102861	-0.003417	0.157258	-0.019659
Boron (ppm)	-0.060200	-0.018620	-0.155223	-0.258950	0.219444	0.763298

Table 3.
Newly assigned names for PCA components

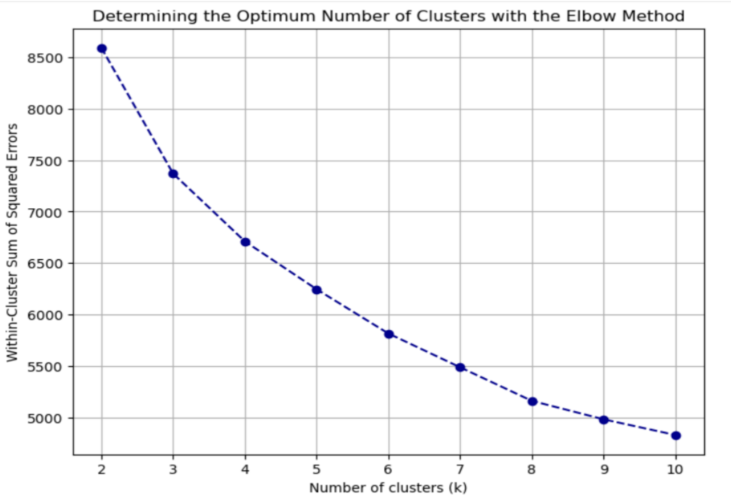
	PCA Components Names Rasyon
PC1	Soil Physical Structure, Moisture and Salinity
PC2	Mineral Content and Ionic Balance
PC3	pH, Salinity and Electrical Conductivity
PC4	Soil Texture
PC5	Chemical and Mineral Content with Ionic Balance
PC6	Micronutrient and Mineral Content

The newly assigned names for the PCA components are presented in Table 3.

When examining the high loading values for each component, it is observed that certain groups of variables stand out. Accordingly, representative names were assigned to the PCA components based on the structural similarities of the dominant variables they contain. For example, in PC1, variables such as 'Saturation with Water (%)', 'Clay (%)', 'Total Salt (%)', and 'Mg meq/l', which are related to physical structure, moisture, and salinity, have high loading values. Therefore, this component was named 'Soil Physical Structure, Moisture and Salinity'. Similarly, in PC2, variables like 'Na meq/l', 'SAR', and 'ESP', which are associated with sodium and ionic balance, are prominent, and thus the component was named 'Mineral Content and Ionic Balance'. For the other components as well, the dominant variables were evaluated collectively, and each component was named with meaningful and descriptive titles that reflect the representative structure in the data.

Before the clustering process, the Elbow Method was applied to determine the most appropriate number of clusters and the optimal number of clusters, i.e. the optimal number of clusters, was determined through this method.

Figure 11.
Optimal number of clusters with Elbow method



As can be seen from Figure 11, the most significant break in the curve gives the optimal number of clusters and this number of clusters is observed as k = 3. Thus, the data set is divided into three clusters.

The K-Means clustering analysis classified the soil samples into three groups. Soils with similar physical and chemical properties were grouped together. Cluster centroids were calculated based on the mean values of the relevant principal components, and the results are presented in Table 4.

Table 4.

Categorization of soil variables based on PCA

	Soil Physical Structure, Moisture and Salinity	Mineral Content and Ionic Balance	pH, Salinity and Conductivity	Soil Texture	Chemical and Mineral Content with Ionic Balance	Micronutrient and Mineral Content
Cluster 0	1.120029	1.893867	-0.188349	-0.065788	0.065901	-0.049974
Cluster 1	1.745496	-1.325928	0.114344	0.006652	-0.108815	0.125175
Cluster 2	-1.930882	-0.266353	0.038773	0.036878	0.03467	-0.056196

These groups, identified through clustering, demonstrated that the soils were significantly differentiated according to fertility levels. These clusters were then used as labels in supervised learning algorithms, supplying data for classification. As shown in the image below, these three clusters are clearly separated.

As shown in Figure 12, Cluster 0 is primarily concentrated in the upper regions, and the soils in this cluster exhibit moderate moisture and salinity levels along with high mineral content. This structure indicates that these areas possess favorable and balanced soil properties for agriculture.

Cluster 1 is concentrated in the right sub-region and consists of soils with high moisture and salinity but low mineral content. This group represents soils that may be considered fertile but require additional fertilizer supplementation.

Cluster 2 is in the left sub-region and is generally characterized by low moisture, low salinity, and average mineral content. This indicates that the soils in this cluster are arid, have low fertility, and offer limited suitability for agricultural production.

Figure 13 displays the locations of agricultural lands in the Odunpazarı district of Eskişehir province, according to the cluster labels obtained by the K-Means algorithm.

Figure 12.

Clustering of soils

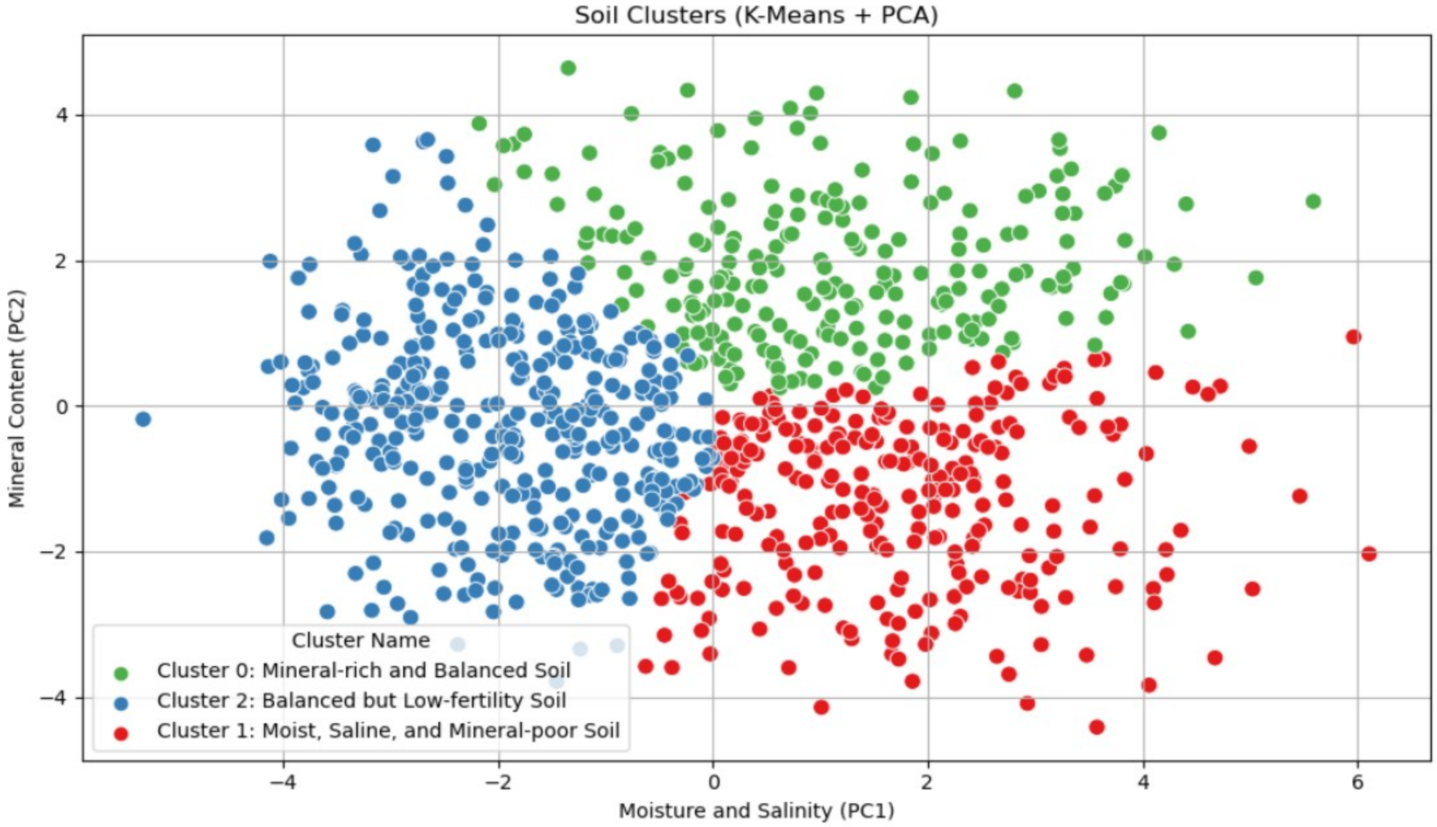
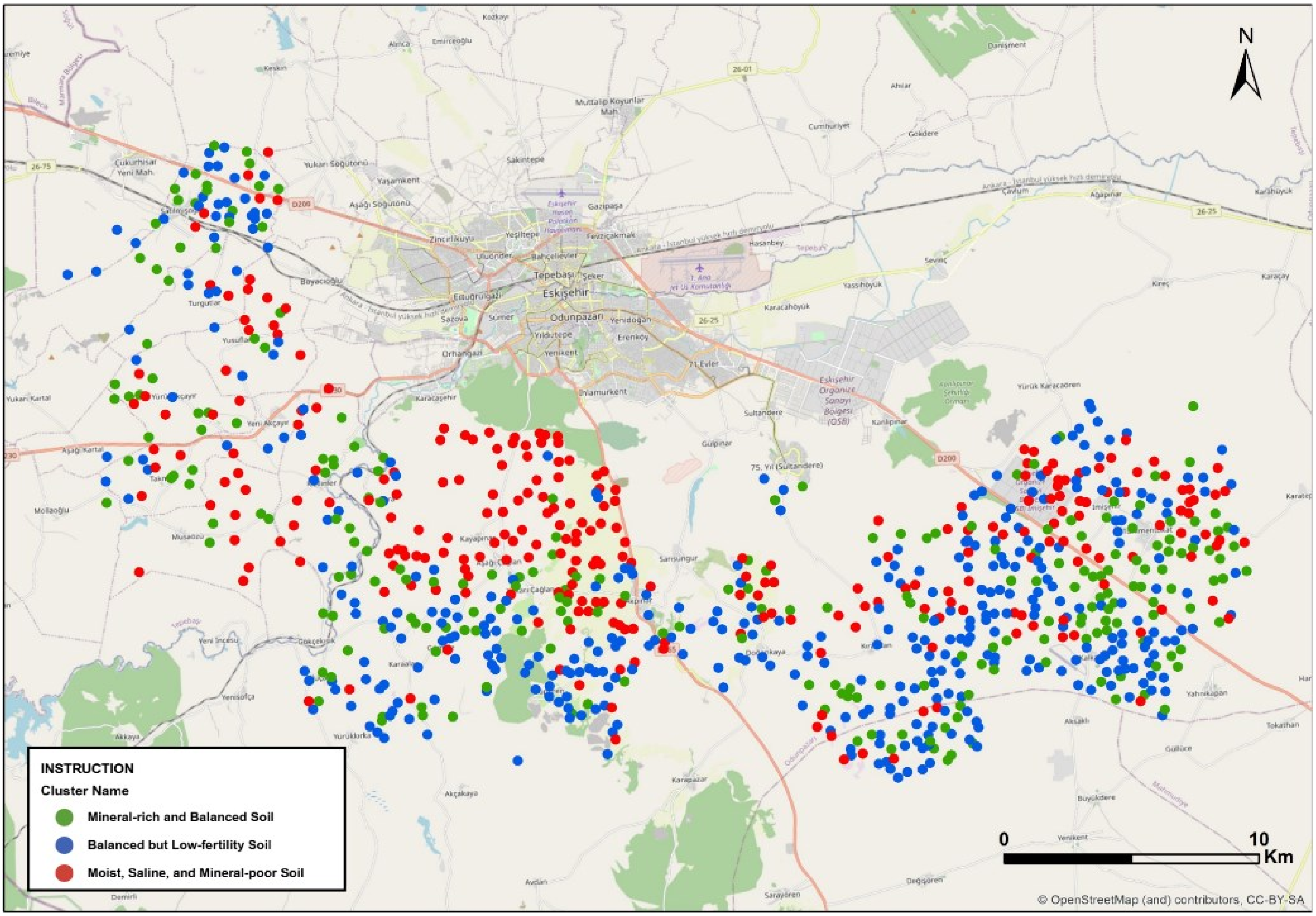


Figure 13.

Location of the fields according to the determined clusters



Mineral-rich and balanced soils (Cluster 0) generally exhibit a scattered distribution on the map, with higher concentrations especially in the western and eastern regions. This cluster is suitable for cereal, vegetable, and fruit cultivation. These soils have low fertilizer requirements and moderate irrigation needs. They are suitable for cereals such as wheat, barley, and corn, as well as vegetables like tomato, pepper, and eggplant, and fruits such as grape, apple, and pear. Moist, saline, and mineral-poor soils (Cluster 1) are generally densely clustered in central regions. These soils have a high-water retention capacity and are therefore particularly suitable for water-resistant crops like rice. With fertilizer support, vegetables such as spinach, lettuce, and beet, as well as salt-tolerant crops like barley and oats, are also suitable for these soils. Balanced but low-fertility soils (Cluster 2) are more distinctly clustered in eastern regions and consist of dry, low-salinity, sandy soils. These areas are more limited in terms of soil fertility and can be considered suitable for dry farming, pasture, or soil rehabilitation. Crops such as chickpea, lentil, bean, alfalfa, vetch, sainfoin, sesame, and safflower are suitable for these soils.

In Phase 2 of the study, these clusters were treated as label variables to establish a suitable framework for the supervised learning process. Additionally, at this stage, PCA was applied again to reduce the dimensionality of the dataset and minimize the impact of correlations between variables, and the first six components explaining 80% of the total variance were included in the analysis.

The resulting dataset was divided into two subsets: 80% for training and 20% for testing. To ensure proportional representation of each class in both subsets, a stratified train/test split was applied. This approach helped to reduce the potential impact of class imbalance on the model. All pre-processing steps were applied exclusively to the training data, while the test data remained independent of these processes. This approach eliminated the risk of data leakage during model evaluation. Four different classification algorithms were employed within the scope of supervised machine learning: Logistic Regression, Decision Tree, Random Forest, and KNN models. Each model was trained on the training data and subsequently evaluated using the test data. The performance of the models has been comparatively analyzed based on metrics such as accuracy, precision, recall, and F1 score. Below is a table containing detailed performance data for each model, including accuracy, precision, recall, and F1 score. This table allows for a clearer understanding of how each model performs across various metrics.

Table 5.

Evaluation metrics of models

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.9889	0.9895	0.9895	0.9895
Decision Tree	0.9779	0.9784	0.9784	0.9784
Random Forest	0.9613	0.9658	0.9593	0.9621
KNN	0.9171	0.9143	0.9143	0.9143

Table 5 compares the performance of four different classification algorithms based on metrics such as accuracy, precision, recall, and F1 score. Logistic Regression has demonstrated the highest overall performance. With an accuracy rate of 98.9%, this model also achieved similarly high values for other metrics such as precision, recall, and F1 score. The results suggest that Logistic Regression is a reliable and balanced model that can make accurate predictions even when classes are imbalanced. The Decision Tree model achieved an accuracy rate of 97.8%, falling behind Logistic Regression in terms of accuracy. However, it still delivered strong results for precision, recall, and F1 score. The Random Forest model, with an accuracy of 96.1%, ranks third in terms of accuracy but has demonstrated a well-balanced performance in terms of precision and recall. The KNN model, in contrast, performs the weakest among the four, with an accuracy rate of 91.7%, placing it at the bottom. Its precision, recall, and F1 score values are also lower compared to the other models.

In conclusion, while Logistic Regression generally yields the best results, Random Forest and Decision Tree also provide strong alternatives. KNN, on the other hand, demonstrates a more limited performance, with lower accuracy compared to the other models.

In the study conducted by Hayattu et al. (2020) in the Northwestern region of Nigeria, soils were categorized into soil fertility classes using the K-Means clustering algorithm. Based on soil parameters such as nitrogen, phosphorus, potassium, organic matter, and pH, three clusters were obtained, corresponding to high, medium, and low soil fertility levels. Similarly, in the present study, soils from the Odunpazarı district were classified into three groups according to their mineral content, moisture, and salinity levels. This shows clear similarities with the Nigerian case. However, while Hayattu et al. (2020) relied solely on clustering methods for soil fertility assessment, the current study additionally applied supervised classification techniques, with Logistic Regression achieving the highest accuracy of 99%. Therefore, although consistent with the international literature, this study further contributes by integrating both clustering and classification approaches, thus providing stronger predictive performance.

In addition, when comparing the classification of soil types using machine learning with other studies, for instance, Taher et al. (2021), 400 soil samples from the Northwestern region of Nigeria were analyzed using 13 soil component attributes to construct different soil fertility classes. According to the experimental results, the highest accuracy was obtained with the KNN algorithm (84%), while Naïve Bayes achieved 69.23%, and both Decision Tree and Random Forest reached 53.85%. In the present study, the Logistic Regression model achieved the highest performance with an accuracy of 98.9%. Unlike the aforementioned studies in the literature, this model yielded higher accuracy values. This discrepancy may be attributed to differences in the type and number of variables used in the dataset, as well as the variation in sample sizes.

This study has certain limitations. First, the dataset used includes only soil analysis results from the Odunpazarı district of Eskişehir. Therefore, the generalizability of the findings is limited, and they should be supported by similar studies conducted in different regions. Moreover, the dataset does not cover other agricultural factors such as climate conditions, irrigation practices, and crop diversity. Future research is recommended to utilize datasets with broader geographical coverage and to compare results obtained through different methods.

Conclusion and Recommendations

This study evaluated the soil fertility of the Odunpazarı district in Eskişehir province, located in Türkiye's Central Anatolia Region, using soil analysis data and machine learning techniques. After applying PCA for dimensionality reduction, the K-Means algorithm was used to classify the soils into three categories. These clusters showed clear differences in physical structure, moisture, salinity, and mineral content, providing important insights into regional soil fertility.

The clustering results offer practical implications for farmers. Green areas, with high mineral content and balanced moisture, are suitable for cereals, vegetables, and fruits, with relatively low fertilizer requirements. Red areas, characterized by high moisture and salinity but low mineral levels, are suitable for water-tolerant crops such as rice and can also support vegetable production with fertilizer supplementation. Blue areas, which are drier and less fertile, are more appropriate for dry farming, grazing, or soil rehabilitation. This classification supports farmers in making scientifically informed decisions on crop selection and resource management.

For policymakers, the findings emphasize the importance of region-specific strategies. In green areas, crop diversification and improvements in marketing infrastructure can be encouraged. In red areas, fertilizer subsidies, the expansion of drip irrigation systems, and the promotion of organic soil improvement practices should be prioritized. In blue areas, rather than intensive agricultural activities, long-term land rehabilitation projects, erosion control, and pasture management policies should be implemented. This approach would shift agricultural support from a uniform model to region-specific strategies, leading to more efficient resource use, stronger food security, and sustainable soil fertility management at the national level.

These findings make significant contributions to precision agriculture by guiding practices such as fertilization, irrigation, and crop selection. Providing farmers with evidence-based recommendations supports efficient resource use while strengthening environmental sustainability.

Future research can expand these findings in several directions. Larger and multi-regional datasets would enhance the generalizability of results. Multi-season and time-series data could more accurately capture the seasonal variability of soils. The integration of climate data (e.g., rainfall, temperature, drought indices) would improve predictive accuracy and provide insights into the impacts of climate change on soil fertility. Furthermore, the application of deep learning models (CNNs, RNNs, hybrid architectures) could increase classification accuracy and strengthen crop recommendation systems. In addition, the proposed approach could be practically implemented by integrating it with farm decision support tools and mobile applications, allowing farmers to make real-time, data-driven decisions. When combined with real-time soil monitoring systems, operations such as irrigation, fertilization, and crop selection can be managed more quickly and precisely. Transforming these models into user-friendly mobile and web-based decision support systems would further facilitate informed decision-making and accelerate the adoption of smart agriculture practices. Consequently, these findings can serve as a strong reference for future soil fertility research conducted in different regions.

Peer-review: Externally peer-reviewed.

Ethics Committee Approval: This study was conducted solely using environmental/soil data and did not involve any human participants, personal data, or animal experiments. Therefore, ethical committee approval was not required.

Author Contributions: Concept – EG; Design - SB, ESÇ, Eİ, İA; Supervision - EG; Resources - SB, ESÇ, Eİ, İA; Data Collection and/or Processing - SB, ESÇ, Eİ, İA; Analysis and/or Interpretation - SB, ESÇ, Eİ, İA; Literature Search - SB, ESÇ, Eİ, İA; Writing Manuscript - SB, ESÇ, Eİ, İA; Critical Review - EG; Other - SB, ESÇ, Eİ, İA, EG

Conflict of Interest: The authors have no conflicts of interest to declare.

Financial Disclosure: The authors declared that this study has received no financial support.

Hakem Değerlendirmesi: Dış bağımsız.

Etik Komite Onayı: Bu çalışma, yalnızca çevresel/toprak verileriyle yürütülmüş olup herhangi bir insan katılımcı, kişisel veri ya da hayvan deneyi içermemektedir. Bu nedenle etik kurul onayı gerekmemektedir.

Yazar Katkıları: Konsept - EG; Tasarım - SB, ESÇ, Eİ, İA; Denetim - EG; Kaynaklar - SB, ESÇ, Eİ, İA; Malzemeler - SB, ESÇ, Eİ, İA ; Veri Toplama ve/veya İşleme - SB, ESÇ, Eİ, İA; Analiz ve/veya Yorum - SB, ESÇ, Eİ, İA; Literatür Taraması - SB, ESÇ, Eİ, İA; Yazma - SB, ESÇ, Eİ, İA; Eleştirel İnceleme - EG; Diğer – SB, ESÇ, Eİ, İA, EG

Çıkar Çatışması: Yazarlar çıkar çatışması olmadığını beyan etmiştir.

Finansal Destek: Yazarlar, bu çalışma için finansal destek almadığını beyan etmiştir.

References

Ağlarcı, A. V., & Karakurt, F. (2024). K En Yakın Komşu Makine Öğrenme Algoritmasına Dayalı Diabetes Mellitus Tahmini. *Turkish Journal of Diabetes and Obesity, 8*(3), 265-276.

Bhargavi, P., & Jyothi, S. (2011). Soil Classification Using Data Mining Techniques: A Comparative Study. *International Journal of Engineering Trends and Technology, 2*(1), 55-59.

Bhargavi, S., & Jagannathan, Dr. S. (2024). Crop Recommendation System Using Machine Learning. *International Journal of Engineering Research & Technology (IJERT), 11*(6). https://doi.org/10.17577/NCRTCA-PID-443

Brownlee, J. (2016). Supervised and unsupervised machine learning algorithms. *Machine Learning Mastery, 16*(03).

Burhan, H.A., & Soydan, N.T.Y. (2023). Nohut ve Mercimek Üretim Miktarı Tahmini İçin Meteorolojik Faktörler Odaklı Makine Öğrenmesi Yaklaşımı: Türkiye Örneği. *Adnan Menderes Üniversitesi Ziraat Fakültesi Dergisi, 20*(1), 13-23.

Demir, O., Gültekim, G. Ç., & Uzundumlu, A. S. (2023). Türkiye Ekonomisinde Tarımın Yeri ve Önemi. *Palandöken Journal of Animal Sciences Technology and Economics, 2*(2), 62-69.

Ersungur, Ş. M., Kızıltan, A., & Polat, Ö. (2007). Türkiye'de Bölgelerin Sosyo-Ekonomik Gelişmişlik Siralamasi: Temel Bileşenler Analizi. *Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi, 21*(2), 55- 66.

Esmer, Y., & Gezer, Y. (2021). Tarımsal işletmelerde stratejik analiz: Erzurum ili örneği. *Atatürk Üniversitesi Ziraat Fakültesi Dergisi, 52*(2), 119-127.

Garanayak, M., Sahu, G., Mohanty, S. N., & Jagadev, A. K. (2021). Agricultural Recommendation System for Crops using Different Machine Learning Regression Methods. *International Journal of Agricultural and Environmental Information Systems (IJAEIS), 12*(1), 1-20.

Gruhn, P., Goletti, F., & Yudelman, M. (2000). *Integrated Nutrient Management, Soil Fertility, and Sustainable Agriculture: Current issues and Future Challenges.* Intl Food Policy Res Inst.

Hamid, H. A., Hassan, A., Wah, Y. B., & Amin, N. A. M. (2018, October). Investigating the power of goodness-of-fit test for multinomial logistic regression using K-Means clustering technique. In *AIP Conference Proceedings* (Vol. 2013, No. 1). AIP Publishing.

Hayatu, I.H., Mohammed, A., Ismaâ, B. A., & Ali, S. Y. (2020). K-Means Clustering Algorithm Based Classification of Soil Fertility in North West Nigeria. *FUDMA Journal of Sciences, 4*(2), 780-787.

Kılavuz, E., & Erdem, İ. (2019). Dünyada Tarım 4.0 Uygulamaları ve Türk Tarımının Dönüşümü. *Social Sciences, 14*(4), 133-157.

Koldere, Y. (2008). Clustering Algorithms and Clustering Analysis in Data Mining. *Marmara University Doctoral Thesis,* 8, 10, 16, 26-29.

Kumral, C. D., Topal, A., Ersoy, M., Çolak, R., & Yiğit, T. (2022). Random forest algoritmasının FPGA üzerinde gerçekleştirilerek performans analizinin yapılması. *El-Cezeri, 9*(4), 1315-1327.

Maathuis, F. J. (2009). Physiological Functions of Mineral Macronutrients. *Current opinion in plant biology, 12*(3), 250-258.

Patel, K., & Patel, H. B. (2023). Multi-criteria Agriculture Recommendation System using Machine Learning for Crop and Fertilizesrs Prediction. *Current Agriculture Research Journal, 11*(1).

Paudel, D., Boogaard, H., Wit, A., Velde, M., Claverie, M., Nisini, L., ... & Athanasiadis, I.N. (2022). Machine Learning for Regional Crop Yield Forecasting in Europe. *Field Crops Research, 276, 108377.*

Prity, F. S., Hasan, M. M., Saif, S. H., Hossain, M. M., Bhuiyan, S. H., Islam, M. A., & Lavlu, M. T. H. (2024). Enhancing agricultural productivity: a machine learning approach to crop recommendations. *Human-Centric Intelligent Systems, 4*(4), 497-510.

Reddy, D. A., Dadore, B., & Watekar, A. (2019). Crop Recommendation System to Maximize Crop Yield in Ramtek Region Using Machine Learning. *International Journal of Scientific Research in Science and Technology, 6*(1), 485-489.

Taher, K. I., Abdulazeez, A. M., & Zebari, D. A. (2021). Data Mining Classification Algorithms for Analyzing Soil Data. *Asian Journal of Research in Computer Science, 8*(2), 17-28.

Yadav, J., Chopra, S., & Vijayalakshmi, M. (2021). Soil analysis and crop fertility prediction using machine learning. *Machine Learning, 8*(03).

Yakut, G., Çay, R.İ., & Öztürk, H.H. (2023). Makine Öğrenimi Teknikleri Kullanılarak Isparta İli İçin Tarımsal Ürün Önerme Sistemi. *Gazi Mühendislik Bilimleri Dergisi, 9*(4), 174-185.