ISNN:2181-1296



SHAROF RASHIDOV NOMIDAGI
SAMARQAND DAVLAT UNIVERSITETI

САМАРКАНДСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ ИМЕНИ ШАРОФА РАШИДОВА

SAMARKAND STATE UNIVERSITY NAMED
AFTER SHAROF RASHIDOV

ILMIY AXBOROTNOMA

НАУЧНЫЙ ВЕСТНИК
SCIENTIFIC JOURNAL

ANIQ VA TABIIY FANLAR SERIYASI

Matematika, Mexanika, Informatika
Fizika, Kimyo, Biologiya, Geografiya

№ 1/1 (155)
2026

ISSN 2181-1296

ILMIY AXBOROTNOMA

НАУЧНЫЙ ВЕСТНИК SCIENTIFIC JOURNAL

2026-yil, 1-son (155/1) ANIQ VA TABIIY FANLAR SERIYASI

Matematika, Mexanika, Informatika, Fizika, Kimyo, Biologiya, Geografiya

Samarqand viloyat matbuot boshqarmasida ro‘yxatdan o‘tish tartibi 09-25.
Jurnal 1999-yildan chop qilina boshlagan va OAK ro‘yxatiga kiritilgan.

BOSH MUHARRIR **R. I. XALMURADOV, t.f.d., professor**
BOSH MUHARRIR O‘RINBOSARLARI: **H.A. XUSHVAQTOV, f-m.f.d., dotsent**

TAHRIRIYAT KENGASHI:

Sh.A.Alimov	- O‘zFA akademigi	**M.X.Ashurov**	- O‘zFA akademigi
S.N.Lakayev	- O‘zFA akademigi (SamDU)	**B.J.Axmedov**	- O‘zFA akademigi
M.M.Kamilov	- O‘zFA akademigi	**Q.G‘.G‘ulomov**	- O‘zFA akademigi
A.G.Yagola	- Moskva davlat universiteti professori (Rossiya)	**A.A.Zaxidov**	-Texas Nanotexnologiya instituti professori (AQSh)
Alberto Del Bimbo	-Florensiya universiteti professori (Italiya)	**N.B. Ferapontov**	- Moskva davlat universiteti professori (Rossiya)
Sh.Tugizov	-Kaliforniya universiteti professori (AQSh)	**V.L.Popov**	-Berlin texnika universiteti professori (Germaniya)
A.S.Soleev	- f.-m.f.d., professor (SamDU)	**N.F.Bunkin**	- Moskva davlat texnika universiteti professori (Rossiya)
I.A.Ikromov	- f.-m.f.d., professor (SamDU)	**Haji Akber Aisa**	-Xinjiang tibbiyot universiteti professori (Xitoy)
B.X.Xo‘jayarov	- f.-m.f.d., professor (SamDU)	**E. A. Abduraxmonov**	- k.f.d., professor (SamDU)
L.Sabirov	-f.-m.f.d., professor (SamDU)	**N. K. Muxamadiyev**	- k.f.d., professor (SamDU)
A.Jumaboyev	-f.-m.f.d., professor (SamDU)	**L. A. Bulavin**	- Kiev milliy universiteti
X.Bozorov	-k.f.d., professor (SamDU)	**X. Q. Xaydarov**	- b.f.d., professor (SamDU)
O.Q.Quvondiqov	-f.-m.f.d. professor (SamDU)	**Z. I. Izzatullayev**	- b.f.d., professor (SamDU)
A.Z.Xasanov	-f.-m.f.d. professor (SamDU)	**A.B.Hasanov**	-tex.f.d., professor (SamDU)
Muhammad Suzuri bin Hitam	- Terengganu universiteti professori (Malayziya)	**Sh. T. Xoliqulov**	- q.-x.f.d., professor (SamDU)
K.D.Kuliyev	- f.-m.f .n. professor (SamDU)	**S. B. Abbosov**	- g.f.d., professor (SamDU)
A.M.Xolxo‘jayev	-f.-m.f.d., professor (SamDU)	**M.E. Mo‘minov**	-f.-m.f.d., professor (SamDU)
J.Abdullayev	-f.-m.f.d., professor (SamDU)	**I.I.Jumanov**	-tex.f.d., professor (SamDU)
A.R.Axatov	-tex.f.d., professor (SamDU)	**N.S.Mamatov**	-tex.f.d., professor (TIQXMMI)
D.B.Xursanov	- g.f.f.d., dotsent (SamDU)	**A.Sh.Yarmuxamedov**	-f-m.f.n., dotsent (SamDU)
F.Nazarov	-tex.f.d., dotsent (SamDU)	**X.S.Xaydarov**	-f-m.f.n., dotsent (SamDU)

Obuna indeksi – yakka tartibdagi obunachilar uchun **- 5583,**
tashkilot, korxonalar uchun **- 5584**

M U N D A R I J A
C O N T E N T S

№	Maqola nomi	Article title	Bet/ page
1.	$CsPbBr_3$ kvant nuqtalariga asoslangan lyuminissent quyosh konsentratorlarining optik xossalarini optimallashtirish **Nurimov A.D., Davurov Sh.M., Sherqulov E.S., Juraboyev L.A., Zakhidov A.**	Optimising optical performance of $CsPbBr_3$ quantum dots-based luminescent solar concentrators **Nurimov A.D., Davurov Sh.M., Sherqulov E.S., Juraboyev L.A., Zakhidov A.**	**5-10**
2.	Impuils-davriyli lazer nuri ta'sirida rux oksidi va rux selenid mikrokukunlarida uyg'otilgan fotolyuminesensiya **Rakhmatullaev I.A., Khaydarov Kh.S., Hoshimov Q.H., Abdusattorov B.A.**	Photoluminescence in zinc oxide and zinc selenide micropowders under pulsed-periodic laser excitation **Rakhmatullaev I.A., Khaydarov Kh.S., Hoshimov Q.H., Abdusattorov B.A.**	**11-15**
3.	Perovskit–polimer kompozitiga asoslangan ko'k bir qatlamli yorug'lik chiqaruvchi diodlar **Rahmonova M.O., Muxamadiyev N.Q.**	Blue single-layer light-emitting diodes based on a perovskite–polymer composite **Rahmonova M.O., Muxamadiyev N.Q.**	**16-23**
4.	Chegaraviy oqim bilan berilgan chiziqsiz diffuziya sistemasining sonli tadqiqoti **Yarmetova D.**	Numerical study of a nonlinear diffusion system with boundary fluxes **Yarmetova D.**	**24-28**
5.	Iqlim o'zgarishi sharoitida samarali haroratlar yig'indisi taqsimoti (Samarqand viloyati misolida) **Kholbaev G., Eshmirzaev D., Egamberdiev Kh., Shermukhammedov U., Arzikulov M., Yuldasheva Sh.**	Spectral properties of schrödinger operator for a two-fermion system on a lattice: eigenvalues and resonances **Kholbaev G., Eshmirzaev D., Egamberdiev Kh., Shermukhammedov U., Arzikulov M., Yuldasheva Sh.**	**29-36**
6.	Janus kinaza ingibitorlari sifatida 4,6-dialmashgan xinazolin hosilalarini dizayni va sintezi **Gaybullaev Sh., Takhirov Y., Zohidov K., Chao Niu, Bozorov Kh.**	Design and synthesis of 4,6-disubstituted quinazoline derivatives as janus kinase inhibitors **Gaybullaev Sh., Takhirov Y., Zohidov K., Chao Niu, Bozorov Kh.**	**37-43**
7.	Suzuki-miyaura usulida 3-almashingan-pirido[2,3-*d*]pirimidinonlarning sintezi **Khushnazarov Z., Turgunov D., Bozorov Kh.**	Synthesis of 3-substituted-pyrido[2,3-*d*]pyrimidinones by the suzuki-miyaura method **Khushnazarov Z., Turgunov D., Bozorov Kh.**	**44-48**
8.	2-metil-pirazolo[3,4-*d*]pirimidinonlarning sintezi **Ruzieva F., Turgunov D., Bozorov Kh.**	Synthesis of 2-methyl-pyrazolo[3,4-d]pyrimidinones **Ruzieva F., Turgunov D., Bozorov Kh.**	**49-53**
9.	Organo/noorganik ms-schiff gibrid sorbentning metall kationlariga nisbatan sorbsiya xossalarini baholash **Kuliboyeva N., Uzokov J., Mukhamadiyev A., Ruziyev I., Mukhamadiyev N.**	Evaluation of the sorption properties of organo/inorganic ms-schiff hybrid sorbents towards metal cations **Kuliboyeva N., Uzokov J., Mukhamadiyev A., Ruziyev I., Mukhamadiyev N.**	**54-64**

№	Maqola nomi	Article title	Bet/ page
10.	Qaror daraxtlariga asoslangan anjuman usullari yordamida ma'lumotlarni oldindan qayta ishlash va tasniflash algoritmlari **Akhatov A., Nazarov F.**	Data preprocessing and classification algorithms using ensemble methods based on decision trees **Akhatov A., Nazarov F.**	**65-74**
11.	Fotokatalitik faollikka ega TiO_2 nanoklasterlarini struktur va elektron xossalarini kvant kimyoviy usullarda baholash **Mamaziyayeva S., Uzokov J., Toshpulatova L., Mukhamadiyev N., Mukhamadiyeva Kh.**	Evaluation of the structure and electronic properties of photocatalytic tio2 nanoclusters using quantum chemical methods **Mamaziyayeva S., Uzokov J., Toshpulatova L., Mukhamadiev N., Mukhamadieva Kh.**	**75-81**
12.	Cu(II) ionlarining Piren sorbsiyasini sinergetik ta'sirini Langmuir-Freundlich modellashtirish asosida baholash **Vokhidov A., Kosimov S., Khalilov S., Mukhammadiyev N.**	Evaluation of the synergetic effect of pyrenee sorption of cu(ii) ions based on langmuir-freundlic modeling **Vokhidov A., Kosimov S., Khalilov S., Mukhammadiev N.**	**82-90**
13.	Magnit suyuqliklarining magneto-optik yorug'lik o'tkazuvchanligi **Kuvandiqov O., Kirgizov S., Mamasoliyev B., Kirgizova M.**	Evaluation of the synergetic effect of pyrenee sorption of cu(ii) ions based on langmuir-freundlic modeling **Kuvandikov O., Kirgizov S., Mamasoliev B., Kirgizova M.**	**91-97**
14.	Sun'iy intellekt mexanizmlari asosida suv tarkibini aqlli tahlil qilish usullari **Salimova M., Nazarov F.**	Methods for intelligent analysis of water composition based on artificial intelligence mechanisms **Salimova M., Nazarov F.**	**98-107**
15.	Glukoza asosida uglerodli kvant nuqtalarining fotolyuminestentsiya xossalari **Sobirova F., Ishankulov A., Khalilov K., Galyametdinov Y.**	Photoluminescence properties of glucose-based carbon quantum dots **Sobirova F., Ishankulov A., Khalilov K., Galyametdinov Y.**	**108-112**
16.	Bir o'lchamli kvadrat panjara zinasida diskret Shredinger operatorining xos qiymatlari **Muminov M.I., Sadullaeva M.S.**	On eigenvalues of the discrete schrödinger operator on a ladder of one-dimensional square lattice **Muminov M.I., Sadullaeva M.S.**	**113-120**
17.	O'zbekistonda aholi hududiy harakatining mintaqaviy xususuyatlari **Mukhamedov O.L.**	Regional characteristics of population territorial movement in Uzbekistan **Mukhamedov O.L.**	**121-126**

UDC: 004.8 https://doi.org/10.59251/2181-3973.2025.v1.138.1. 3759

DATA PREPROCESSING AND CLASSIFICATION ALGORITHMS USING ENSEMBLE METHODS BASED ON DECISION TREES

A.Akhatov, F.Nazarov
Samarkand State University named after Sharof Rashidov, Samarkand, Uzbekistan.

A.Akhatov Samarkand State University named after Sharof Rashidov, Samarkand, Uzbekistan.

Received: 15.01.2026
Revised: 29.01.2026
Accepted: 15.02.2026
Published: 30.03.2026

Publisher: Samarkand State University named after Sharof Rashidov.

Abstract. This research paper proposes a methodology for processing datasets based on modern methods. Iterative imputation based on decision trees was used to systematically replace missing values in the dataset. The remaining irregular part of the missing values was filled in using the SVD imputation method based on Matrix Factorization. The data's extreme values and outliers were stabilized using a logarithmic transformation. The prepared data was trained on decision trees and ensemble models based on them. The effectiveness of the proposed ensemble approaches in classifying crop types based on soil data was evaluated on training, validation, and test sets, and the results were analyzed.

Key words. Machine learning, ensemble learning, logarithmic transformation, iterative imputation, decision trees, dataset.

Qaror daraxtlariga asoslangan anjuman usullari yordamida ma'lumotlarni oldindan qayta ishlash va tasniflash algoritmlari

Annotatsiya. Ushbu tadqiqot ishida ma'lumotlar to'plamiga ishlov berishning zamonaviy usullarga asoslangan metodologiyasi taklif etildi. Ma'lumotlar to'plamida yo'qotilgan qiymatlarni tartibli tiklash uchun qaror daraxtlariga asoslangan iterative to'ldirish qo'llanildi. Matritsali faktorizatsiya qilishga asoslangan SVD imputatsiya usuli yordamida yo'qotilgan qiymatlarni qolgan tartibsiz qismi to'ldirildi. Ma'lumotlardagi chetlanishlar va ekstremal qiymatlarni barqarorlashtirish uchun logarifmik transformatsiya qo'llanildi. Tayyorlangan ma'lumotlar qaror daraxtlari va ularga asoslangan ansabml modellarida o'qitildi. Taklif etilgan ansambl yondashuvlarning tuproq ma'lumotlari asosida ekin turlarini klassifikatsiya qilishdagi samaradorligi o'quv, validatsiya va test to'plamlarida baholanib olingan natijalar tahlil qilindi.

Kalit so'zlar. Mashinaviy o'qitish, ansambl o'qitish, logarifmik transformatsiyalash, iterativ imputatsiya, qaror daraxtlari, ma'lumotlar to'plami.

Алгоритмы предварительной обработки и классификации данных с использованием ансамблевых методов, основанных на деревьях решений

Аннотация. В данной работе предложена методология, основанная на современных методах обработки наборов данных. Для систематической замены пропущенных значений в наборе данных использовался итеративный метод импутации на основе деревьев решений. Оставшаяся нерегулярная часть пропущенных значений была заполнена с помощью метода импутации SVD на основе матричной факторизации. Для стабилизации выбросов и экстремальных значений в данных использовалось логарифмическое

преобразование. Подготовленные данные были использованы для обучения деревьев решений и ансамблевых моделей на их основе. Эффективность предложенных ансамблевых подходов в классификации типов сельскохозяйственных культур на основе данных о почве оценивалась на обучающей, валидационной и тестовой выборках, а результаты анализировались.

Ключевые слова. Машинное обучение, ансамбл обучение, логарифмическое преобразование, итеративная импутация, дерево решений, набор данных.

Introduction. In today's advanced digital era, ready-made data has become the main resource for decision-making systems in almost all areas. Practical experiences in working with different types of data show that the high capabilities of a model depend on the quality of the data. In this case, missing values in the dataset, imprecise or erroneous measurements, and extreme observations can violate statistical assumptions, reduce the generalizability of the model, and lead to misinterpretation of the results. Therefore, data processing is examining processes ranging from simple fills to matrix-based fills [1]. One of the most pressing problems today is to eliminate problems in data preprocessing without losing their correlations. Data processing stabilises model parameters and makes correct and reliable decisions throughout training, regardless of model selection [2].

Decision tree-based algorithms are well suited to these practical needs, as they easily accommodate mixed-type features and require less standardization of measurements. In addition, data analysis through ensemble approaches based on decision trees is considered an effective solution for informed decision making [3].

The main goal of this research is to provide modern, reliable, and efficient methods for processing these datasets, which involves developing models for intelligent data analysis using algorithms and methods based on decision trees. The research tasks are as follows:

- imputation of missing values in the dataset;
- detection and stabilization of extreme values;
- training data in decision tree-based classifiers;
- model parameter tuning and evaluation of results.

Literature review. A lot of scientific research is being conducted in the field of data processing and analysis for decision support systems based on artificial intelligence and machine learning. They study in detail the effectiveness of decision trees and ensemble methods, their capabilities in classification work, and new algorithmic optimization approaches.

A decision tree classification algorithm based on granular matrices was proposed by L. Meng et al. This approach has shown high classification accuracy compared to classical tree methods such as ID3 and C4.5. According to the results, the proposed algorithm is more efficient than traditional tree methods, better models complex structures, and provides classification with higher accuracy [4]. Strong SVD techniques were demonstrated to be able to fill in missing values with excellent accuracy and stability even in data with outliers in a work by Marisol García-Peña and Sergio Arciniegas-Alarcón [5].

Long Zhao and colleagues proposed a decision tree model integrated with the boosting algorithm and tested it on a credit rating task. The CART tree-based boosting model in their investigation demonstrated a 90.95% classification accuracy [6].

Scientists such as Paul D. Yoo, Seongil Han, and Haemin Jung used two-stage rSVD and GAIN approaches to recover missing values in data. As a result, their developed methodology recorded significantly lower RMSE compared to traditional imputation methods [7].

Scientists such as Kanimozhi Gunasekaran, Ayhan Arisoy, and Walid Abdullah have been conducting several research studies on developing ensemble algorithms based on decision trees for analyzing soil fertility in agriculture [8-9].

Materials and methods. The methodology proposed in this research work includes scientifically based steps such as preprocessing the dataset and training the final models using decision tree-based methods. The proposed methodology is presented in Figure 1 below, and each step in the study was carried out in this given sequence.

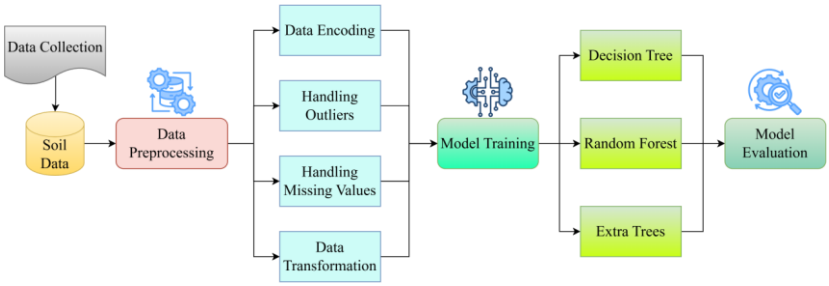


Figure 1. General stages of the methodology for data preprocessing and training of Decision Tree–based algorithms

In this methodology, each stage of data pre-processing is aimed at increasing model stability, generalizability, and reliability of results. The stages of this methodology were carried out in the given sequence, which are described in detail below.

Dataset and Data Preprocessing Techniques. The dataset used in the study included soil properties data from various crop fields. Crop types were selected as the target variable in the dataset. The dataset used contains 19 soil properties of seven crop species, namely Wheat, Maize, Carrot, Bean, Pepper, Potato and Cotton, with a total of 18346 rows. General information about the dataset is given in Table 1 below.

Table 1. Description and properties of the dataset.

Variable type	Features	Unit	Count	Data type
Input features	Sample	-	16307	object
	Layer	cm	18092	object
	Mechanical composition	-	17982	object
	CF	%	17893	float64
	Soil Density	g/cm3	17814	float64
	pH	-	17967	float64
	EC	mS/cm	17854	float64
	N	mg/kg	17966	float64
	P	mg/kg	17945	float64
	K	mg/kg	17995	float64
	Humus	%	17837	float64
	Mg	mg/kg	17882	float64
	S	mg/kg	17848	float64
	Zn	mg/kg	17641	float64
	Mn	mg/kg	17722	float64
	B	mg/kg	17780	float64
	Fe	mg/kg	17679	float64
	Cu	mg/kg	17779	float64
	Microorganisms	CFU/g	17246	float64
Target variable	Crop	-	18237	object

As can be seen from Table 1, the dataset contains missing values and categorical data. In order for these types of values not to affect the accuracy of machine learning models, they must first be identified and then processed using appropriate methods [10].

Data Encoding. The categorical values in the dataset were converted to numeric values using the Ordinal Encoding method. Ordinal Encoding is a method of encoding categorical variables into numeric values based on the order or level of the categories. This method was used because the categorical values in the dataset used in the study corresponded to this category [11]. In this case, the Ordinal Encodig function for the categorical character c and its set of possible values $K_c = \{k_1, k_2, .., k_n\}$ was determined based on equation 1 below:

$$f_c : K_c \to \{1, 2, .., n\}, \quad f_c(k_i) = i \tag{1}$$

where K_c is a set of categorical symbols, k_i is the k categorical value in the i row.

Using the given formula 1, the categorical values $f_c(X_{ij})$ were converted into numerical values X_{ij} while maintaining the ordinal relationships.

Handling Extreme and Missing Values. Extreme values are values that deviate sharply from the distribution of observations, that are very large or very small compared to other values. To effectively identify such values and find solutions to them, the IQR method, which is a statistical method, was used. This method was implemented in the following three steps:

1. The interquartile range was determined as follows, with the first quartile of the data being Q_1 and the third quartile being Q_3:

$$IQR = Q_3 - Q_1 \tag{2}$$

2. When determining the limit values, the lower limit of L was set as $L = Q_1 - 1.5 \times IQR$ and the upper limit of U was set as $U = Q_3 + 1.5 \times IQR$, respectively.

3. When determining extreme values, any observation X_{ij} was considered an extreme value if it satisfied the conditions given in formula 3 below:

$$X_{ij} < Q_1 - 1.5 \times IQR \text{ or } X_{ij} > Q_3 + 1.5 \times IQR \tag{3}$$

The classical Tukey method uses 25% and 75% quantiles to identify extreme values. In this study, 20% and 80% quantiles were chosen, the main reason being that the soil data have a complex distribution [12-13]. To identify missing values in a dataset, the presence identifier of the observations matrix M is $M_{ij} = 1$ if X_{ij} is present, otherwise $M_{ij} = 0$, indicating a missing value. The number of missing values for each column in the dataset was determined by the following expression 4:

$$M_j = \sum_{i=1}^{n} I(X_{ij}) \tag{4}$$

here M_j is the number of missing values in column j, I is an indicator function that takes the value 1 if the condition is met, and 0 otherwise.

According to formulas (3) and (4) given above, extreme and missing values in the dataset were determined. The distribution of identified extreme and missing values across the columns of the dataset is shown in Figure 2 below.

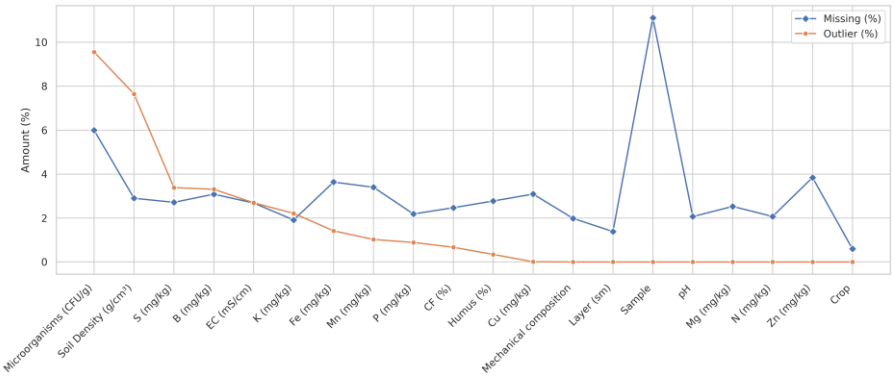


Figure 2. The total percentage of extreme and missing values in the dataset.

In Figure 2, the blue line represents the missing values in the dataset, and the red line represents the extreme values as a percentage of the column cross-section. It can be seen from this graph that almost all columns have missing values. Extreme values were mainly found in the Microorganisms (CFU/g) and Soil Density (g/cm³) columns.

Logarithmic Transformation of Data. In many cases, the identified extreme values are simply removed from the dataset. Given that the dataset used in the study consists of soil properties, removing such values can break the correlation between the data. Therefore, such values were approximated to a normal distribution using the Logarithmic transformation (Log-transformation) method. Log-transformation is a statistical method aimed at symmetricizing the distribution of numerical variables by applying a logarithmic function to them [13]. Log-transformation for a numerical variable X_{ij} in a dataset can be expressed as in formula 5:

$$T_{ij} = \ln(1 + X_{ij}), \; X_{ij} \geq 0 \tag{5}$$

here X_{ij} is the original value, T_{ij} is the log-transformed value.

Iterative imputation method based on decision trees for ordinal values. After logarithmic transformation of the data, the missing data in the ordinal columns encoded above using the Ordinal Encoding method were restored using the Iterative Imputation method. In this approach, each column is treated as a target variable and predicted based on the remaining columns [14]. In this case, for a dataset consisting of variables $X=\{X_1, X_2,..., X_n\}$, the missing values in column X_j are determined by equation 6:

$$X_j^t = \mathrm{P}_j^t(X_1^{(t-1)},..., X_{j-1}^{(t-1)}, X_{j+1}^{(t-1)},..., X_n^{(t-1)}) \quad (6)$$

here t is the iteration step, P_j^t is is the prediction model built for column j, and X_j^t is the populated values at iteration t.

In this imputation approach, Extra Trees (Extremely Randomized Trees) was chosen as the P_j prediction model. This ensemble algorithm is based on a set of random decision trees and is expressed as:

$$\mathrm{P}_j(X) = F\{T_1(X), T_2(X),..., T_N(X)\} \quad (7)$$

here $T_N(X)$ is the prediction of N decision trees, N is the number of trees, and F is the operator for selecting the most frequent class.

The Extra Trees algorithm is effective for data with complex relationships by randomly selecting split points and detecting highly nonlinear relationships [15]. In this Extra Trees-based iterative imputation, missing values are filled in with the most common valuet $X_j^0 = F(X_j)$ before the iterative process begins. Then the imputation process was performed for a maximum of $t=10$ iterations. In each iteration, the columns were updated sequentially, and at the end of the iterations, the imputed values $\hat{X}_j = (X_j^t)$ were generated.

Handling Missing Values Using Singular Value Decomposition (SVD) Imputation. After filling in the existing missing values in the ordered columns, the missing values in the remaining columns were restored using the SVD method. This approach uses a low rank hidden structure to approximate the data matrix, with $X=[X_{ij}] \in R^{n\times m}$ representing the missing value data matrix. The goal is to estimate the missing values by approximating the matrix X with a low rank matrix, given a soil dataset Ω consisting of n samples and m features. To do this, the optimization problem in formula 8 is solved:

$$\min_{U,\Sigma,V} \sum_{(i,j)\in\Omega} (X_{ij} - (U\Sigma V^{\mathrm{T}})_{ij})^2 \,,\; U\Sigma V^{\mathrm{T}} = k \quad (8)$$

where $U \in R^{n\times k}$, $\Sigma \in R^{k\times k}$, $V \in R^{m\times k}$, and k is latent factor.

The missing values are first filled in with statistical estimates to create an initial matrix. To extract the underlying hidden structure in the data matrix and simplify the data based on this structure, SVD performs a decomposition process. At eachr t iteration, the matrix is decomposed as follows:

$$X^{(t-1)} = U_k^{(t)} \Sigma_k^{(t)} V_k^{(t)\mathrm{T}} \quad (9)$$

After SVD decomposition, a low rank reconstruction is performed based on the components retained, as shown in expression 10:

$$\hat{X}^{(t)} = U_k^{(t)} \Sigma_k^{(t)} V_k^{(t)\mathrm{T}} \quad (10)$$

The resulting matrix $\hat{X}^{(t)}$ is a smoothed approximation that reflects the hidden relationships between the data, and it offers logically appropriate values for incompletely observed data. In the next step, only the missing elements are extracted from the reconstructed matrix. This process is given by the following expression 11:

$$X_{ij}^{(t)} = \begin{cases} X_{ij} \\ \hat{X}_{ij}^{(t)} \end{cases} \quad (11)$$

In the given expression 11, the observed real values of $X_{ij} \in \Omega$ and $\hat{X}_{ij}^{(t)} \notin \Omega$ are not changed, that is, the algorithm only fills in the empty places based on the hidden structure, while preserving the existing information. After each update, the convergence of the iterative process is

evaluated. For this, the difference between successive iterations is measured using the Frobenius norm, as shown in formula 12:

$$\left\| X^{(t)} - X^{(t-1)} \right\| \tag{12}$$

If this difference is smaller than a predefined threshold, the matrix is considered to be stabilized and the iteration is stopped. Otherwise, the process is continued for a maximum of T iterations [16]. The resulting $\hat{X} = X^{(T)}$ matrix at the end of the iterations represents the final imputed estimates for all missing values and is reinserted into the original dataset.

Iterative imputation based on decision trees was considered a suitable method for ordinal values, and the remaining columns were filled in using the SVD imputation method. The distribution of values in the column section of the dataset prepared based on all the above steps is presented in Figure 3.

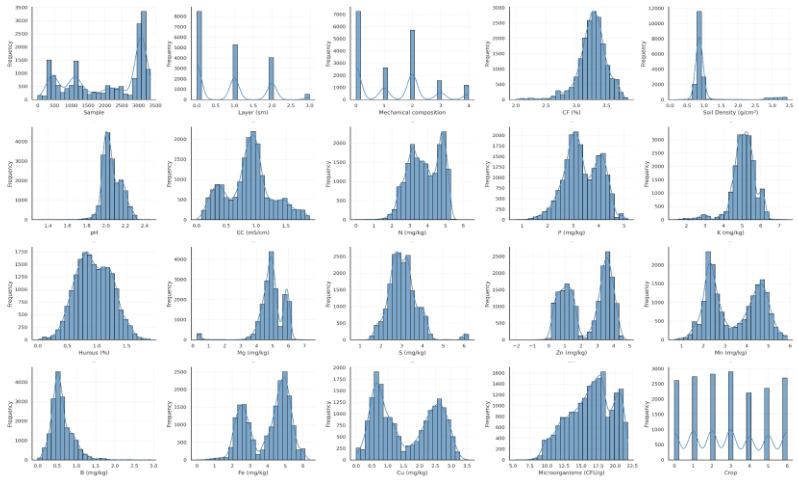


Figure 3. Histograms of soil data features after preprocessing

The results of the graphical analysis showed a significant reduction in the asymmetry of the distribution for most characteristics, which in turn confirms the effectiveness of the Log-transformation of the data and the imputation operations.

Machine Learning Algorithms Based on Decision Trees. A decision tree is a hierarchical algorithm that divides a dataset into small groups based on sequential logical decisions [14-17]. In classification problems, the primary objective of a decision tree is to determine the class to which a new instance belongs. To achieve this, the algorithm performs the following logical steps:

- Initially, all samples are grouped in a single root node;
- at each step, the feature that best separates the data is selected;
- the data are then split into smaller subsets based on the selected feature;
- this process continues until all nodes become pure.

As a result, the leaf nodes at the bottom of the tree represent the final decision [18-19]. The dataset prepared in the study was divided into training, test, and validation sets for training algorithms based on decision trees. In this case, the training dataset is defined as $D = \{X_i, y_i\}_{i=1}^{N}$. Here, X_i is the number of samples, $y_i \in \{1,2,...,C\}$ is the number of classes, and N is the number of samples. The Gini index is used to estimate the degree of mixing of classes at each node in the decision tree, as can be seen in the following expression 13:

$$g(D) = 1 - \sum_{c=1}^{C} p_c^2 \tag{13}$$

here $p_c^2 = \frac{N_c}{N}$ is the proportion of samples belonging to class c, and N_c is is the number of samples of class c.

In a decision tree, if node D is split by feature A, the quality of the split is evaluated using equation 14:

$$\hat{g}(D,A) = g(D) - \sum_{v \in A} \frac{|D_v|}{|D|} g(D_v) \quad (14)$$

here $D_v - A = v$ are samples satisfying the condition.

According to Expression 14 above, the feature that gives the largest $\hat{g}$ value is considered optimal. The main goal of building a decision tree is to create an optimal tree structure by gradually reducing the class uncertainty in the dataset, and this process is given in the following equation 15:

$$\hat{T} = \arg\min_{T} \sum_{t \in T} g(t) \quad (15)$$

here T is the set of all decision trees, t is a particular node in the tree, and $\hat{T}$ is the optimal decision tree.

Once the decision tree is built, it acts as a deterministic classifier, meaning that the model makes a precise decision every time for a new incoming sample. The final prediction for the new incoming sample X is determined by expression 16:

$$\hat{y} = \arg\max_{c} P(y = c \mid t_l) \quad (16)$$

here $\hat{y}$ is the predicted class, $P(y = c \mid t_l) = \frac{N_c^l}{N_l}$ is the probability of class c at this leaf node, and t_l is the leaf node from which the sample was taken.

The model selects the class with the highest probability as the final decision. This makes the classification rule of the decision tree clear and interpretable. To prevent this algorithm from overfitting to the training data, a complexity penalty function is introduced, given by the following expression:

$$R_\alpha(T) = R(T) + \alpha |T| \quad (17)$$

In formula 17 above, $R(T)$ is the learning error of the tree, $|T|$ is the number of leaf nodes in the tree, and α is the control coefficient. This function serves to reduce the error and maintain the simplicity of the model. As a result, the optimal tree satisfies the condition $\hat{T} = \arg\min_{T} R_\alpha(T)$.

A significant drawback of decision tree classification problems is the low stability. Therefore, ensemble approaches have been proposed to improve the stability of decision trees while maintaining their strengths. The proposed approaches are to reduce the error of the overall model by building several independent decision trees and combining their decisions [20].

Random Forest Ensemble Learning Model. This approach is ensemble learning, which consists of many decision trees. In this case, each tree is constructed independently based on random elements. This randomness is mainly due to the data and features. As a result, the trees in the ensemble are not strongly dependent on each other, making their overall decision stable and reliable [21]. If this ensemble consists of B decision trees, that is $\{T_1, T_2, ..., T_B\}$, then each tree T_B is constructed through the processes given in expression 18:

$$D_B \sim D,\ A_B \subseteq A,\ |A_B| = \sqrt{m} \quad (18)$$

here D is the original training dataseti, A is the random feature set, and m is the total number of features.

In the Random Forest ensemble approach, each node is optimally divided by formula 12 above and makes the final decision by voting. This process is given in equation 19 below:

$$\hat{y} = \arg\max_{c} \sum_{B=1}^{B} \Phi(T_B(X) = c) \quad (19)$$

here $T_B(X)$ is the prediction of B - tree, Φ is the indicator function.

In the above expression, if the number of trees is $B = 1$, the model represents the decision tree itself. The Random Forest ensemble reduces variance and the risk of overfitting. However, since each tree searches for the optimal split point, the correlation between individual trees remains high and sometimes sensitive to noise in the data. Therefore, the study also used Extra Trees, a maximally randomized ensemble of decision trees.

Extra Trees (Extremely Randomized Trees) Ensemble Learning Model. Extra Trees is an ensemble learning model based on decision trees and is a more robust randomized method. The main feature of this approach is that the split points at each node are not searched based on optimal criteria, but are selected randomly. In addition, in this method, all trees are trained based on the entire dataset, i.e., randomly selected samples are not used [22]. The split point for the selected j - feature in the Extra Trees node is given by formula 20:

$$H_j \sim U(\min(X_j), \max(X_j)) \quad (20)$$

here H_j is is a randomly chosen split point on the feature j, and U is a uniform distribution.

In this approach, split points are selected based on probability rather than searching for optimal separation from the data. As a result, the error of individual trees may increase slightly, and the correlation between trees may decrease significantly. In Extra Trees too, the final prediction was determined by the 19th expression through ensemble voting.

The dataset prepared using the above methods was trained on three artificial intelligence models based on decision trees, divided into training, testing, and validation sets, and the corresponding results were obtained. **Results.** The proposed decision tree-based models were trained on a soil dataset, divided into training, validation, and testing sections. In order to ensure the efficiency of the models and prevent overfitting, their main hyperparameters were appropriately selected and adjusted based on the values presented in Table 2.

Table 2. Hyperparameters of models based on decision trees

Hyperparameters	**Decision Tree**	**Random Forest**	**Extra Trees**
B	1	200	500
$D_{\max}$	20	20	∞
N_s	20	20	8
N_l	30	20	6
ω_c	*balanced*	*balanced*	*balanced*
α	0.0005	0.0003	-
S_j	*optimal*	*optimal*	-

The accuracy of the models in classifying crop types based on the given soil data was evaluated based on the Accuracy metric for all datasets. The classification accuracy indicators of models based on decision trees are given in Table 3 below.

Table 3. Accuracy performance of the models on the training, validation, and test sets

Models	**Accuracy**		
	Training	**Validation**	**Test**
Decision Tree	0.8694	0.8456	0.8427
Random Forest	0.9167	0.8942	0.8845
Extra Trees	0.9577	0.9081	0.8970

In addition, Confusion Matrix graphs were constructed for each model to analyze the prediction quality across classes. These graphs can be seen in Figure 4 below.

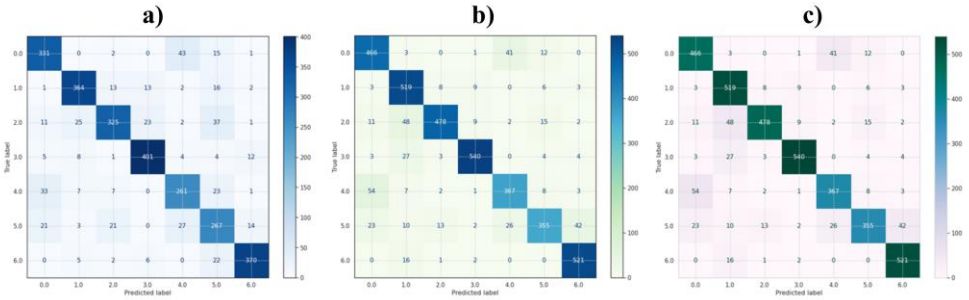


Figure 4. Prediction indicators in the class distribution of a) Decision Tree, b) Random Forest ensemble, and c) Extra Trees ensemble.

In order to further analyze the stability of the models and their learning process, learning curve graphs were constructed, as can be seen in Figure 5 below.

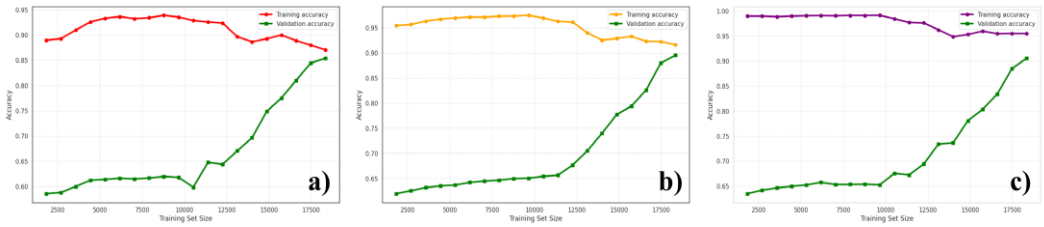


Figure 5. Training and validation learning curves on the reference dataset of a) Decision Tree,
b) Random Forest ensemble, and c) Extra Trees ensemble.

Together with Table 3 above, Figures 4 and 5 demonstrate that decision tree-based machine learning algorithms have produced satisfactory results in the analysis of soil data. Better accuracies were achieved in ensemble approaches than in simple decision trees, with the Random Forest ensemble predicting crops with 88% accuracy based on soil data, and the Extra Trees ensemble predicting crops with almost 90% accuracy. This study demonstrated that data preparation using decision tree-based methods and SVD imputation can provide robust predictions for artificial intelligence models.

Conclusion. The research work developed a comprehensive methodology aimed at solving the problem of missing and extreme values that arise when working with datasets. As a result of these methodological approaches, the quality of the primary data has improved significantly. The results of the decision tree, Random Forest and Extra Trees models built on the prepared data showed the superiority of ensemble approaches over simple decision trees. The Random Forest ensemble model achieved approximately 88% accuracy in the test set and 89% in the validation set. The Extra Trees ensemble achieved the best results with 90% accuracy in the test set and 91% in the validation set. Also, the analysis of learning curves and confusion matrices showed that the models were less prone to overfitting and provided stable predictions across classes. In general, the approach proposed in this study is practically effective in predicting crop types based on soil data analysis. The developed methodology and models can be applied in the future to process and analyze datasets in other areas.

References

1. Gabr, M.I.; Helmy, Y.M.; Elzanfaly, D.S. Effect of Missing Data Types and Imputation Methods on Supervised Classifiers: An Evaluation Study. Big Data Cogn. Comput. 2023, 7, 55. https://doi.org/10.3390/bdcc7010055.
2. Nazarov, F., Nurmamatov, M., Sariyev, S. (2024). Ma'lumotlarni intellektual tahlil qilish uchun genetik algoritmlar va ularni qo'llanilishi. digital transformation and artificial intelligence, 2(6), 162–168. Retrieved from https://dtai.tsue.uz/index.php/dtai/article/view/v2i630Mienye, Ibomoiye & Jere, Nobert. (2024). A Survey of Decision Trees: Concepts, Algorithms, and Applications. IEEE Access. PP. 1-1. 10.1109/ACCESS.2024.3416838.
3. Meng, L.; Bai, B.; Zhang, W.; Liu, L.; Zhang, C. Research on a Decision Tree Classification Algorithm Based on Granular Matrices. Electronics 2023, 12, 4470. https://doi.org/10.3390/electronics12214470.
4. García-Peña, M., Arciniegas-Alarcón, S., Krzanowski, W. J., & Duarte, D. (2021). Missing-value imputation using the robust singular-value decomposition: Proposals and numerical evaluation. Crop Science, 61(5), 3288-3300. https://doi.org/10.1002/csc2.20508.
5. Zhao, L., Lee, S., & Jeong, S. P. (2021). Decision tree application to classification problems with boosting algorithm. Electronics (Switzerland), 10(16), Article 1903. https://doi.org/10.3390/electronics10161903.
6. Han, Seongil & Jung, Haemin & Yoo, Paul. (2025). SMART: Structured Missingness Analysis and Reconstruction Technique for credit scoring. Scientific Reports. 15. 10.1038/s41598-025-99997-4.
7. Gunasekaran, K., A., K., & Sreevardhan, P. (2025). Real-time soil fertility analysis, crop prediction, and insights using machine learning and deep learning algorithms. Frontiers in Soil Science.

8. W. Abdullah, “A Comparative Study of Machine Learning Models for Soil Fertility Prediction Based on Soil Properties,” Jan. 2025, doi: 10.61356/j.oia.2025.2477.
9. Nazarov Fayzulla , Akhatov Abror. Methods for preparing a dataset for determining soil fertility based on artificial intelligence models. (2025). digital transformation and artificial intelligence, *3*(2), 91-99. https://dtai.tsue.uz/index.php/dtai/article/view/v3i215.
10. Esuli, A., Sebastiani, F. (2009). Encoding Ordinal Features into Binary Features for Text Classification. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds) Advances in Information Retrieval. ECIR 2009. Lecture Notes in Computer Science, vol 5478. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-00958-7_83.
11. Mramba, L. K., Liu, X., Lynch, K. F., Yang, J., Aronsson, C. A., Hummel, S., Norris, J. M., Virtanen, S. M., Hakola, L., Uusitalo, U. M., & Krischer, J. P. (2024). Detecting potential outliers in longitudinal data with time-dependent covariates. European journal of clinical nutrition, 78(4), 344–350. https://doi.org/10.1038/s41430-023-01393-6.
12. West RM. Best practice in statistics: The use of log transformation. Annals of Clinical Biochemistry: International Journal of Laboratory Medicine. 2021;59(3):162-165. doi:10.1177/00045632211050531.
13. Xu, Chao. (2025). Outlier Detection in Asymmetric Distributions - Using a Quantile-Adjusted Boxplot Approach. American Journal Of Big Data. 6. 1-16. 10.71465/ajbd3386.
14. J. Zhang, H. Yuan, J. Si, N. Jiang and S. Wang, "Think Twice Before Imputation: Optimizing Data Imputation Order for Machine Learning," 2025 IEEE 41st International Conference on Data Engineering (ICDE), Hong Kong, Hong Kong, 2025, pp. 1362-1374, doi: 10.1109/ICDE65448.2025.00106.
15. Nikfalazar, S., Yeh, CH., Bedingfield, S. *et al*. Missing data imputation using decision trees and fuzzy clustering with iterative learning. Knowl Inf Syst **62**, 2419–2437 (2020). https://doi.org/10.1007/s10115-019-01427-1.
16. Arciniegas Alarcón, Sergio & García-Peña, Marisol & Krzanowski, Wojtek. (2020). Imputation using the singular value decomposition: variants of existing methods, proposed and assessed. International journal of innovative computing, information & control: IJICIC. 16. 1681-1696. 10.24507/ijicic.16.05.1681.
17. N. Fayzullo, S. Sariyev and Y. Sherzodjon, "Analyzing the Effectiveness of Ensemble Methods in Solving Multi-Class Classification Problems," *2025* International Russian Smart Industry Conference (SmartIndustryCon), Sochi, Russian Federation, 2025, pp. 788-793, doi: 10.1109/SmartIndustryCon65166.2025.10986248.
18. Zhao, L., Lee, S., & Jeong, S.-P. (2021). Decision Tree Application to Classification Problems with Boosting Algorithm. Electronics, 10(16), 1903. https://doi.org/10.3390/electronics10161903.
19. Nazarov F., Yarmatov Sh. (2022). Artificial intelligence technologies in increasing data reliability. Современные инновационные исследования актуальные проблемы и развитие тенденции: решения и перспективы, 1(1), 199–202. извлечено от https://inlibrary.uz/index.php/zitdmrt/article/view/5149.
20. Jijo, Bahzad & Abdulazeez, Adnan. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends. 2. 20-28.
21. Natras, R., Soja, B., & Schmidt, M. (2022). Ensemble Machine Learning of Random Forest, AdaBoost and XGBoost for Vertical Total Electron Content Forecasting. Remote Sensing, 14(15), 3547. https://doi.org/10.3390/rs14153547.
22. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach Learn* **63**, 3–42 (2006). https://doi.org/10.1007/s10994-006-6226-1.

SAMARQAND DAVLAT UNIVERSITETI ILMIY AXBOROTNOMASI

НАУЧНЫЙ ВЕСТНИК SCIENTIFIC REPORTS

Texnik muharrir **Sh.Xoliqulov**

Muharrirlar:

I.Sulaymonov - fil.f.d., dotsent
E.Arziqulov - f.-m.f.d., dotsent
O.Yusupova - fil.f.n., dotsent

Mas'ul muharrir:

A. Sh. Yarmuxamedov - f.-m.f.n.

Muassis: Samarqand davlat universiteti Manzil: 140104,
Samarqand shahri, Universitet xiyoboni, 15.
Telefon: (0 366) 240-38-47, Faks: (0 366) 239-13-87 e-mail:
axborotnoma.uz
SamDU «Ilmiy axborotnoma» jurnali tahririyati kompyuterida terildi. Bosishga 30.03.2026-yilda ruxsat etildi. Qog'oz o'lchami A-4. Nashriyot hisob tabog'i 10,00
Buyurtma raqami 833. Adadi 30 nusxa.

ISSN 2181-1296