ORIGINAL ARTICLE

INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE

ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 03,2026

Journal: https://www.academicpublishers.org/journals/index.php/ijai

DETERMINING THE POTABILITY OF WATER USING MACHINE LEARNING BASED ON DATA PREPROCESSING

I.D. Khamidov

A.A.Akhatov

M.Kh. Salimova

Samarkand State University named after Sharof Rashidov, Samarkand.
xamidovikromjon007@gmail.com,
axatov_abror@samdu.uz,
maxbubacoc@gmail.com

Abstract. In this article, the problem of water composition analysis was studied based on a machine learning decision tree algorithm. Access to clean and safe drinking water is one of the basic human needs, but millions of people around the world still do not have access to potable water. Water quality testing is important in ensuring that it is safe for consumption. The study analyzed a data set based on several physicochemical parameters of water and built a machine learning model based on them. This model allows for a preliminary assessment of the potability of water. The main task of the study is to develop a method that allows for quick, reliable monitoring of water quality, which will be especially useful for regions with limited access to water testing.

Keywords: water quality, potability determination, machine learning, decision tree algorithm, data preprocessing, classification model, water quality assessment.

Annotatsiya. Ushbu maqolada suv tarkibini tahlil qilish muammosi mashinaviy o‘qitishning qaror daraxti algoritmi asosida o‘rganildi. Toza va xavfsiz ichimlik suvi insonning asosiy ehtiyojlaridan biri hisoblanadi, biroq dunyo bo‘ylab millionlab odamlar hali ham ichishga yaroqli suvdan foydalanish imkoniyatiga ega emas. Suv sifatini tekshirish uning iste’mol uchun xavfsizligini ta’minlashda muhim ahamiyatga ega. Tadqiqotda suvning bir qator fizik-kimyoviy parametrlariga asoslangan ma’lumotlar to‘plami tahlil qilindi va ular asosida mashinaviy o‘qitish modeli qurildi. Ushbu model suvning ichishga yaroqliligini dastlabki baholash imkonini beradi. Tadqiqotning asosiy vazifasi suv sifatini tezkor va ishonchli monitoring qilish imkonini beruvchi usulni ishlab chiqishdan iborat bo‘lib, bu ayniqsa suvni sinovdan o‘tkazish imkoniyati cheklangan hududlar uchun juda foydalidir.

Kalit so‘zlar: suv sifati, ichishga yaroqlilikni aniqlash, mashinaviy o‘qitish, qaror daraxti algoritmi, ma’lumotlarga dastlabki ishlov berish, tasniflash modeli, suv sifatini baholash.

Аннотация. В данной статье исследуется проблема анализа состава воды на основе алгоритма дерева решений машинного обучения. Доступ к чистой и безопасной питьевой воде является одной из базовых потребностей человека, однако миллионы людей во всем мире до сих пор не имеют доступа к пригодной для питья воде. Проверка качества воды играет важную роль в обеспечении её безопасности для потребления. В исследовании был проанализирован набор данных, основанный на ряде физико-химических параметров воды, и на их основе построена модель машинного обучения. Данная модель позволяет проводить предварительную оценку пригодности воды для питья. Основной задачей исследования является разработка метода, позволяющего осуществлять быстрый

INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE

ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 03,2026

Journal: https://www.academicpublishers.org/journals/index.php/ijai

и надёжный мониторинг качества воды, что особенно актуально для регионов с ограниченным доступом к тестированию воды.

Ключевые слова: качество воды, определение пригодности для питья, машинное обучение, алгоритм дерева решений, предварительная обработка данных, модель классификации, оценка качества воды.

INTRODUCTION. Water is one of the most important natural resources necessary for human life and health. Nevertheless, today in many regions of the world the problem of providing the population with clean drinking water remains. Industrial waste, various chemicals, fertilizers used in agriculture and other anthropogenic factors lead to pollution of water sources. Therefore, monitoring the composition of water and assessing its suitability for drinking is one of the important issues. Determining water quality is often carried out through laboratory analysis. While such methods give accurate results, they are time-consuming and require certain financial costs. In some regions, the lack of modern laboratory equipment limits the possibility of rapid water testing. Therefore, there is a need to search for simpler and faster methods for assessing water quality [1]. In recent years, artificial intelligence and machine learning technologies have been widely used in various fields. These approaches allow us to identify relationships between indicators based on data and make preliminary assessments of certain processes. Therefore, the use of machine learning methods in determining the potability of water based on its physicochemical parameters is one of the important directions.

In this study, a data set containing several physicochemical parameters of water was studied. Based on this data, a model based on the decision tree algorithm was built, and the possibility of determining the potability of water was also analyzed. The results obtained show that the use of machine learning methods allows simplifying the process of water quality assessment and providing rapid assessment [2-3].

Literature Review on Determining Water Potability Using Machine Learning Techniques. In recent years, machine learning algorithms have been widely used in assessing water quality and determining its potability. Such approaches allow for quick and efficient determination of the potability of water based on its physicochemical parameters.

A study by Jaymin Patel, Karan Shah, Shubham Vaghela, Mehul Aglodiya, and Rohan Bhattad used several machine learning algorithms to determine the potability of water. The study compared the Decision Tree, Support Vector Machine, and Random Forest algorithms. According to the results of the study, the Decision Tree algorithm showed an accuracy of about 63% in determining the potability of water [4].

A study by Emna Ben Rachid used various machine learning algorithms to predict the potability of water. The results of the study emphasized that ensemble algorithms showed high efficiency in classifying water quality data [5].

A study by Umar Zaky analyzed the efficiency of classification algorithms based on water quality indicators. The study showed that various machine learning methods can be used to classify water samples into potable and non-potable classes [6].

In a study conducted by Dwi Ayu Shofiana, M. Caniadi, R. Sholehurrohman, and Aristoteles, ensemble algorithms based on decision trees were studied for water quality determination. The results of the study showed that the Random Forest and XGBoost algorithms provide effective results in water quality classification [7].

INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE

ORIGINAL ARTICLE

ISSN: 2692-515-x, Impact Factor: 9.23

American Academic publishers, volume 6, issue 03,2026

Journal: https://www.academicpublishers.org/journals/index.php/ijai

MATERIALS AND METHODS. This study includes the steps from preprocessing the dataset to training and evaluating the model for the process of analyzing water composition based on machine learning. In the work performed, an approach based on the machine learning decision tree algorithm was used to analyze the composition of water and assess its potability [8]. Several modern methods of preprocessing the dataset were considered and the most suitable methods were selected.

Main parameters characterizing water quality and their measuring devices. Determining its physicochemical parameters is important in characterizing water composition. These parameters are the main indicators of the structural state of water. The determination of the physicochemical parameters of water is carried out using special laboratory methods and various measuring devices. Below is information about the main parameters characterizing the composition of water and the measuring devices used to determine them.

1. pH value. The pH value is an important parameter in assessing the acid-base balance of water. It also serves as a key indicator of whether water is acidic or alkaline. The World Health Organization (WHO) recommends a pH range of 6.5 to 8.5 for drinking water. The pH value is usually measured using a pH meter (Figure 1).

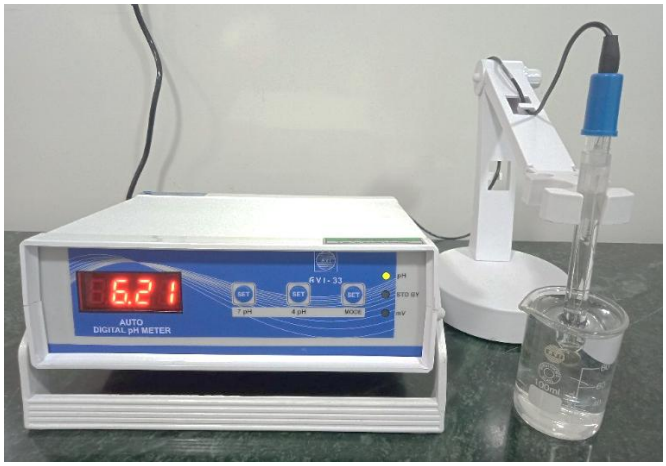

Figure 1. Laboratory pH meter device used to measure pH.

2. Hardness. The hardness of water is mainly due to the presence of calcium and magnesium salts. These salts dissolve from the geological formations through which the water flows and are added to the water. The length of time the water is exposed to hardness-forming substances plays an important role in determining the level of hardness in raw (untreated) water. Initially, the concept of hardness was defined as the ability of water to precipitate soap (not form foam) under the influence of calcium and magnesium ions. Water hardness is usually determined using titration equipment (Figure 2).

INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE

ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 03,2026

Journal: https://www.academicpublishers.org/journals/index.php/ijai

Figure 2. Titration equipment used to determine water hardness.

3. Solids. Water has the ability to dissolve many inorganic and some organic minerals and salts. These include potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates, and other substances. These minerals can cause unpleasant taste and discoloration of water. Therefore, the amount of solids in water is an important indicator in assessing the suitability of water for human consumption [9].

4. Chloramines. Chlorine and chloramines are the main disinfectants used in public water supply systems. Chloramines are often formed by the addition of ammonia to chlorine during the treatment of drinking water.

5. Sulfate. Sulfates are naturally occurring substances that are present in minerals, soil, and rocks. Sulfates are also found in ambient air, groundwater, plants, and food.

6. Organic carbon. Total Organic Carbon (TOC) in source waters is formed mainly due to decaying natural organic matter (NOM) and artificial (synthetic) sources. TOC represents the total amount of carbon present in organic compounds in water.

7. Electrical conductivity. Pure water does not conduct electricity well, but on the contrary, it is a good insulator. An increase in the concentration of ions in water increases its electrical conductivity. Usually, the amount of dissolved solids in water determines the level of electrical conductivity [10]. Electrical conductivity is measured using a conductivity meter (Figure 3).

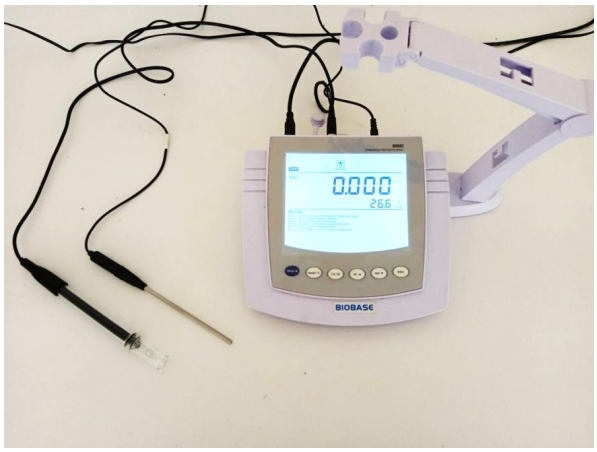

Figure 3. Conductivity meter device used to determine the electrical conductivity of water.

8. Trihalomethanes (THMs). Trihalomethanes are chemicals that can form in water treated with chlorine. The concentration of THMs in drinking water depends on several factors.

ORIGINAL ARTICLE
INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE
ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 03,2026
Journal: https://www.academicpublishers.org/journals/index.php/ijai

9. Turbidity. The turbidity of water depends on the amount of suspended solids in it. Turbidity is an indicator of the light transmission properties of water, and this test is used to assess the quality of wastewater discharge in relation to colloidal substances [11, 12]. This indicator should be below the standard of 5.00 NTU recommended by the World Health Organization (WHO). The level of turbidity of water is measured using a turbidity meter (Figure 4).

Figure 4. Turbidity meter (nephelometer) device used to determine the level of turbidity of water.

Dataset and its initial analysis. The water quality dataset hosted on the Kaggle platform, which is publicly available, was used for this study [13]. This dataset contains information on the physicochemical parameters and their potability status for various water samples. Each water sample consists of 9 physicochemical parameters, which include pH, hardness, solids, chloramines, sulfates, electrical conductivity, organic carbon, trihalomethanes and turbidity, and the dataset consists of a total of 3276 rows. Each water sample is marked as potable (1) or unpotable (0). General information about the dataset is given in Table 1 below.

Table 1. General information about the dataset

Feature Category	Features	Units	Count	WHO standarts	Data Type
Chemical properties	pH	-	2785	6.5–8.5	float64
	Hardness	mg/L	3276	200	float64
	Chloramines	ppm	3276	4	float64
	Sulfates	mg/L	2495	1000	float64
	Organic Carbon	ppm	3276	10	float64
	Trihalomethanes	ppm	3114	80	float64
Physical properties	Solids	ppm	3276	1000	float64
	Conductivity	µS/cm	3276	400	float64
	Turbidity	NTU	3276	5	float64
Target Variable	Potability	-	3276	-	int64

Table 1 above lists the characteristics of water quality, as well as the threshold values recommended by the WHO [14]. As can be seen from this table, the dataset contains missing values for the columns pH, Sulfates, and Trihalomethanes. These missing values need to be processed before modeling. Commonly used methods include imputation, such as using the mean, median, or most frequent value, or more advanced methods such as KNN imputation.

ORIGINAL ARTICLE
INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE
ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 03,2026
Journal: https://www.academicpublishers.org/journals/index.php/ijai

In this study, the target variable was chosen as the indicator of "potability". This variable is expressed as a binary classification, with a value of 0 indicating that the water is not potable and a value of 1 indicating that it is potable. The distribution of potability in the dataset can be seen in Figure 5 below.

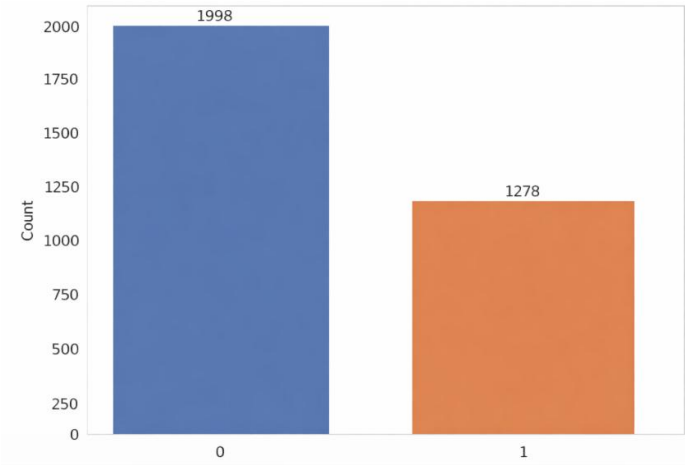

Figure 5. Potability distribution (0-unpotable, 1-potable)

From Figure 5, we can see that the data set is unbalanced, with more samples marked as 0 (unpotable) than 1 (potable). Specifically, approximately 61% of the samples were found to be potable, while the remaining 39% were potable.

Preprocessing the dataset. Initially, the missing values in the data set were filled in using the median method. The median method is the central value of the data set, which represents the element in the middle when the values are arranged in an ordered manner. If the number of values is even, it takes the arithmetic mean of the two middle values. The main reason for choosing the median method is that this method gives better results than the data set with outliers [15].

Outlier detection and processing. An outlier is a value that is too large or too small compared to other values in the dataset. Outlier processing is one of the important steps in the data set processing process. Extreme values can reduce the efficiency of the model or lead to incorrect interpretation of the results. Outlier detection is important in cases where the data distribution is uneven. One of the most popular methods for detecting outliers is the Interquartile Range (IQR) method [16]. Box plots and percentiles are important statistical indicators in the process of detecting outliers using the IQR method. A box plot visually represents the distribution of data by showing:

- Q_1 (25th percentile): 25% of the data are below this point
- Median (50th percentile): the middle value of the data.
- Q_3 (75th percentile): 75% of the data are below this point.
- The interquartile range covers the middle 50% of the data:

$$IQR = Q_3\text{-}Q_1 \quad (1)$$

This method ignores the extreme values of the distribution that are very large or very small and focuses on the central part.

In the IQR method, outliers are determined using the IQR proximity rule. It defines the following 2 limits:

ORIGINAL ARTICLE
INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE
ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 03,2026
Journal: https://www.academicpublishers.org/journals/index.php/ijai

$$L = Q_1 - 1.5 * IQR \quad (2)$$

$$U = Q_3 + 1.5 * IQR \quad (3)$$

here L is the lower limit and U is the upper limit.

If a value is outside these ranges, it is considered an outlier. The coefficient of 1.5 used in the formula is a widely used standard in practice. After we found outliers, the thresholding method was used to process them. The essence of this method is that the outliers are not deleted, but rather are equated to the calculated threshold values.

Working with Imbalanced Classes . In classification problems, it is often the case that the number of classes in a dataset is not equal. This situation is called class imbalance, that is, the number of samples belonging to one class is much larger than that of another. To reduce this problem, the study used the SMOTE (Synthetic Minority Oversampling Technique) method [17].

SMOTE increases the number of samples in a rare class by creating new synthetic samples for this class. This method, instead of simply multiplying the existing data, creates new data points based on the characteristics of nearby neighboring samples.

First, the algorithm analyzes the target variable (y values) to determine which class is the rare class. Then, for each sample belonging to this class, k nearest neighbors are found (usually k = 5). In the next step, a new synthetic point is created between the current sample and one of its neighbors. This new point is located along the line between the two samples and is calculated using the following formula:

$$x = x_i + \lambda^*(x_m - x_i) \quad (4)$$

where x_i is the current (main) sample vector, x_m - x_i is the nearest neighbor sample to $\lambda \in [0,1]$ is a random number between 0 and 1.

The additional samples generated in this way increase the number of rare classes and, as a result, improve the balance of classes in the dataset. This helps to obtain more stable and reliable results for classification models. Table 2 shows the number of samples before and after the application of the SMOTE method to the target variable, the indicator of "potability".

Table 2. Samples in the dataset before and after the SMOTE method

	Before SMOTE	After SMOTE
Not Potable	1998	1998
Potable	1278	1998

Standardization of features. The StandardScaler method was used to scale features with different units of measurement to the same scale:

$$z = \frac{x-\mu}{\sigma} \quad (5)$$

where "μ" is the mean, "σ" is the standard deviation

Correlation analysis. Correlation analysis is used to determine the relationship between features in a data set. For this, a correlation matrix was constructed and the results were visualized using a heat map. Figure 6 shows the correlation between all features. The results showed that there was no strong linear relationship between the features, so there was no need to remove any features from the data set.

INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE

ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 03,2026

Journal: https://www.academicpublishers.org/journals/index.php/ijai

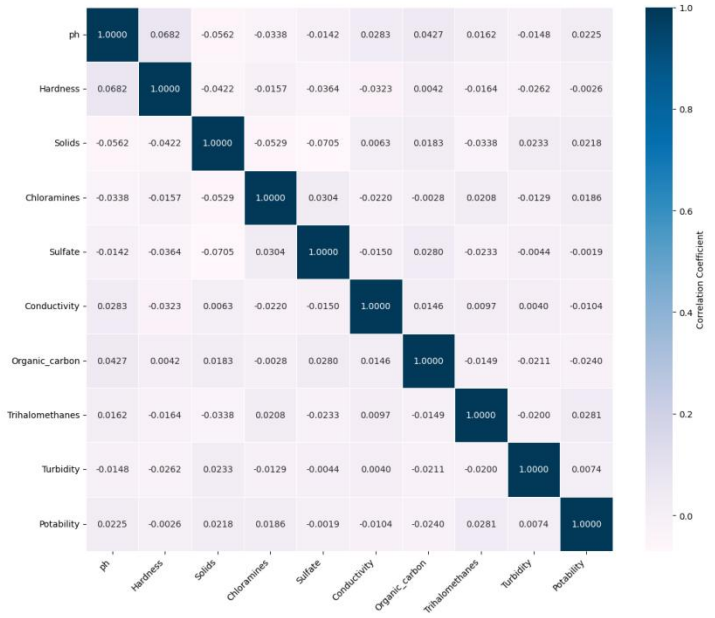

Figure 6. Heatmap of correlation between features

The distribution of values in the column cross-section of the dataset prepared based on all the above steps is presented in Figure 7.

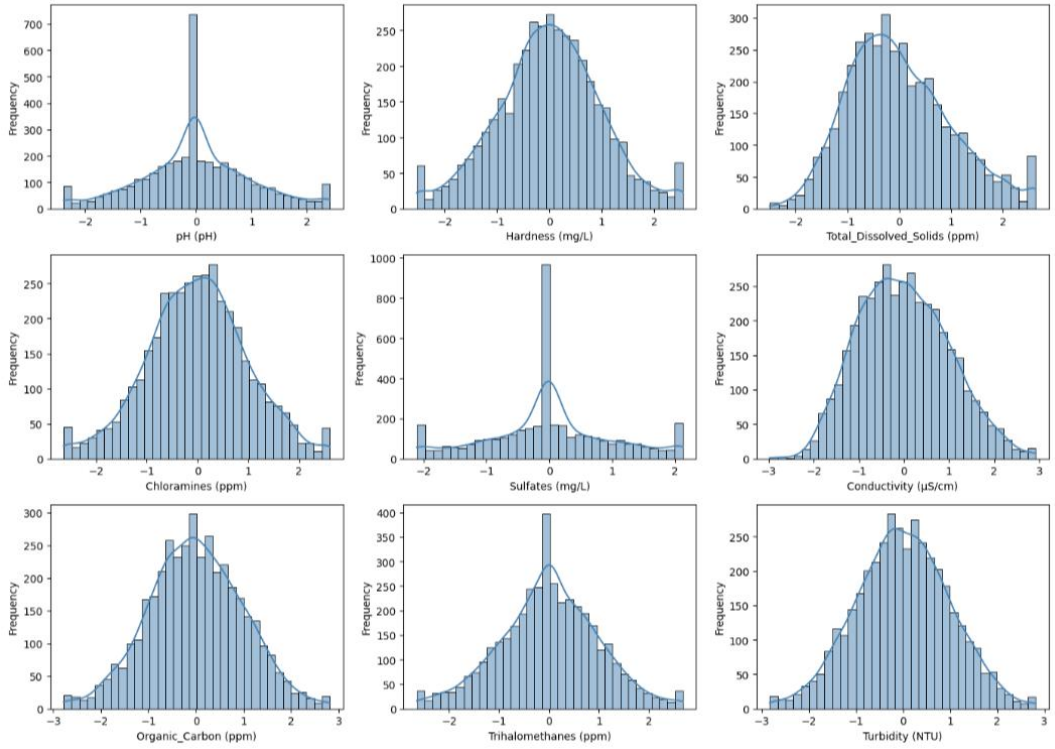

Figure 7. Histogram distribution of water data features in the column section after data processing stages.

According to the results of graphical analysis, it was observed that the distribution of most features has a shape close to the normal distribution. This indicates that the data cleaning and filling in of missing values processes were effective.

ORIGINAL ARTICLE
INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE
ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 03,2026
Journal: https://www.academicpublishers.org/journals/index.php/ijai

Data partitioning. To train the model, it is necessary to partition the data set. Before applying the model to the machine learning algorithm, a test is conducted with a part of the data in order to evaluate its performance and the evaluation indicators are calculated [18]. Therefore, training and test data were formed from the data set. The training data accounted for 80% of the total data set, and the test data accounted for 20%.

Decision tree algorithm. Decision trees are one of the supervised learning methods [19]. Although this algorithm can be used to solve various problems, it is mainly widely used in classification problems. In a decision tree classifier, nodes represent features of the data set, and paths represent previous decisions or conditions, each node serves to form a final conclusion. A decision tree is a hierarchical algorithm that divides a data set into small groups based on sequential logical decisions. This algorithm consists of the following steps:

- initially, all samples are placed in one common node;
- at each step, the feature that best separates the data is determined;
- based on the selected feature, the data is divided into smaller groups;
- this process continues until the leaf nodes of the tree are formed.

It should be noted that the data set prepared in the study was divided into training and test sets. In this case, the training data set is represented as $D=\{(X_i,y_i)\}_{i=1}^{N}$. Here, X is the number of observations, and N is the number of samples. The Gini index was used to evaluate the degree of mixing of classes at each node in the decision tree [20]. It is determined by the following formula 6:

$$g(D)=1-\sum_{c=1}^{C} p_c^2 \qquad (6)$$

where $p_c=\frac{N_c}{N}$ is the fraction of samples belonging to class c and N_c is the number of samples of class c.

If node D in a decision tree is split by feature A, the quality of the split is evaluated by the following expression:

$$\hat{g}(D,A)=g(D)-\sum_{v\in A}\frac{|D_v|}{|D|}g(D_v) \qquad (7)$$

here $D_v-A=v$ is the set of samples satisfying the condition.

The feature that gives the largest $\hat{g}$ value according to the formula is chosen as optimal. The main goal of building a decision tree is to create an optimal tree structure by gradually reducing the uncertainty of the classes in the data set. This process is expressed as follows:

$$\hat{T}=\arg\min_{T}\sum_{t\in T} g(t) \qquad (8)$$

where 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 constructed, it acts as a deterministic classifier, meaning that the model will make the exact decision for each new input sample. The final prediction for a new input sample X is given by the following expression (9):

$$\hat{y}=\arg\max_{c} P(y=c|t_i) \qquad (9)$$

where $\hat{y}$ is the predicted class, $P(y=c|t_i)=\frac{N_c}{N_t}$ is the probability of class c at this leaf node, and t_l is the leaf node from which the sample fell.

The model makes the final decision based on the class with the highest probability. An important feature of this algorithm is its ability to clearly and clearly explain the decision-

ORIGINAL ARTICLE
INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE
ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 03,2026
Journal: https://www.academicpublishers.org/journals/index.php/ijai

making process. However, in order to prevent the model from overfitting to the training data, a penalty function is used in expression 10 to control the complexity:

$$R_\alpha(T)=R(T)+\alpha|T| \quad (10)$$

In the given formula 10, 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, which serves to reduce the error and maintain the simplicity of the model. As a result, the optimal tree $\hat{T}= \arg\min_T R_\alpha(T)$.

Decision trees are one of the most effective algorithms for solving classification problems. In this study, the decision tree algorithm was used to determine the suitability of water for drinking. However, in some cases, the results of the decision tree may be sensitive to data changes. Nevertheless, its simple structure and understandable decision-making process allow this algorithm to be widely used in practical problems.

Evaluation indicators. The following indicators were used to evaluate the performance of the model.

Precision. Precision is the proportion of cases correctly classified by the classifier among all cases classified as positive. When calculating the Precision value, TP indicates cases correctly classified as positive, and FP indicates cases classified as false positive. The Precision value is calculated using formula (11).

$$\text{Precision} = \frac{TP}{TP+FP} \quad (11)$$

Accuracy. Accuracy is the ratio of all cases correctly classified by the model to the total number of observations. Formula (12) is used to calculate the accuracy value. Here, TP is the number of correctly classified positive cases, TN is the number of correctly classified negative cases, FP is the number of false positive cases, and FN is the number of false negative cases.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad (12)$$

Recall. Recall is the proportion of cases correctly identified by the model out of the cases belonging to the positive class. Formula (13) is used to determine the value of Recall. Here, TP is the number of cases correctly classified as positive, and FN is the number of cases incorrectly classified as negative.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (13)$$

F1-Score. F1-score is a measure that takes into account both precision and recall indicators when evaluating the effectiveness of a model. Because in some cases, only precision or only recall may not fully reflect the result of the model. Therefore, the F1 indicator is defined as the harmonic mean of these two indicators. The F1 value is calculated using formula (14). This indicator ranges from 0 to 1, and the larger the value, the higher the accuracy of the model.

$$\text{F 1-Score} = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (14)$$

RESULTS. A Decision Tree classification model was used to determine the potability of water. In the process of evaluating this model, the cases before and after processing the dataset were analyzed separately. Initially, the model's result in the case before processing the dataset showed 57.6%. Then, pre-processing steps such as data cleaning and filling in missing values were performed. After these processes, the model was re-trained and evaluated. As a result, processing the dataset had a positive effect on the efficiency of the model, and the accuracy of

ORIGINAL ARTICLE
INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE
ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 03,2026
Journal: https://www.academicpublishers.org/journals/index.php/ijai

the model reached 62.3%. The results of the Decision Tree model before and after processing the dataset are presented in Table 3.

Table 3. Comparison of model accuracy before and after data set processing

Model	Before Resampling (%)	After Resampling (%)
Decision Tree	57.6	62.3

To further analyze the classification results of the model, an error matrix was constructed. This matrix shows the correct and incorrect predictions made by the model. This can be seen in Figure 8.

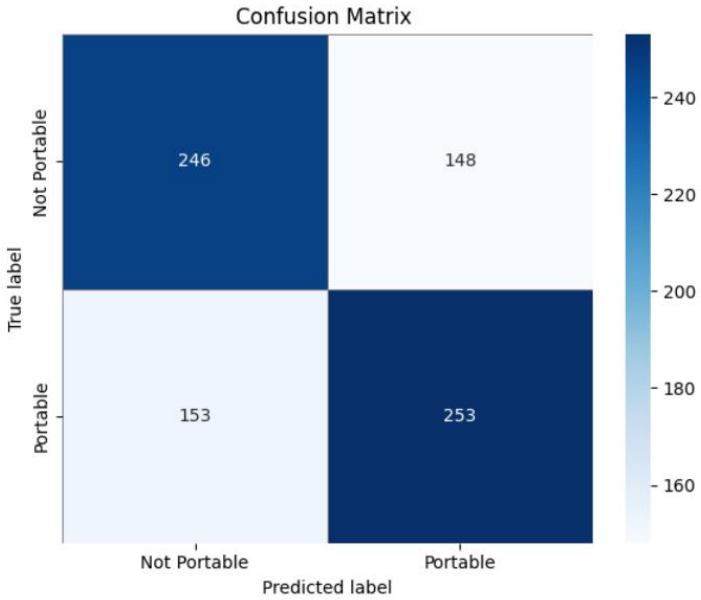

Figure 8. Error matrix for the decision tree model

In order to evaluate the classification performance of the model, the accuracy, precision, recall and F1-score indicators were calculated. The values of these indicators are presented in Table 4.

Table 4. Classification indicators for the decision tree model

Model	Class	Precision	Recall	F1-score	Accuracy
Decision Tree	Not Potable	0.617	0.624	0.620	0.623
	Potable	0.631	0.623	0.627	

According to the overall results, the decision tree model showed moderate efficiency in determining the suitability of water for drinking. Pre-processing of the data set helped to increase the accuracy of the model. Also, the precision, recall and F1-score indicators indicate that the model showed stable results in classifying both classes.

CONCLUSION. In this study, the problem of determining the suitability of water for drinking was analyzed using a machine learning algorithm and the model efficiency was evaluated. During the research, a data set containing several physicochemical properties of water was analyzed and pre-processing stages were performed. This process had a positive effect not only on the quality of the data, but also on the accuracy of the model. In the study, the Decision Tree classification model was used and showed 57.6% accuracy on the unprocessed

ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 03,2026
Journal: https://www.academicpublishers.org/journals/index.php/ijai

data set, while after pre-processing of the data, the accuracy of the model reached 62.3%. Therefore, preprocessing of data is important to improve the model performance.

The model results were also analyzed using the matrix of model results and evaluation indicators such as precision, recall, and F1-score. According to the analysis, the model showed relatively stable results in the classification process for both classes. Overall, this study demonstrated that machine learning algorithms are an effective tool for automating the process of assessing water quality and determining its suitability for drinking. In future studies, there are opportunities to further improve the prediction accuracy using ensemble algorithms or deep learning models.

REFERENCES

[1] World Health Organization. Guidelines for Drinking-Water Quality, 4th ed.; WHO]Press: Geneva, Switzerland, 2017. https://www.who.int/publications/i/item/9789241549950

[2] Sahoo, P. K.; Kim, K.; Powell, M. A. Machine Learning for Water Quality Prediction and Monitoring: A Review. *Water*, 2021, 13(18), 2562. https://doi.org/10.3390/w13182562

[3] Patel, J.; Shah, K.; Vaghela, S.; Aglodiya, M.; Bhattad, R. A Machine Learning-Based Water Potability Prediction Model. *Computers, Materials & Continua*, 2022, 72(3), 5347–5360. https://doi.org/10.32604/cmc.2022.028945

[4] Patel, J.; Shah, K.; Vaghela, S.; Aglodiya, M.; Bhattad, R. A Machine Learning-Based Water Potability Prediction Model. *Computers, Materials & Continua* 2022, 72(3), 5347–5360. https://doi.org/10.32604/cmc.2022.028945

[5] Rachid, E. B. Predicting Water Potability Using Machine Learning Techniques. *Cleaner Engineering and Technology* 2025, 21, 100902. https://doi.org/10.1016/j.clet.2024.100902

[6] Zaky, U. Performance Analysis of the Decision Tree Classification Algorithm on the Water Quality and Potability Dataset. *International Journal of Data and Network Science* 2023, 7(2), 613–620. https://doi.org/10.5267/j.ijdns.2022.12.012

[7] Shofiana, D. A.; Caniadi, M.; Sholehurrohman, R.; Aristoteles. Decision Tree Algorithms in Water Quality Classification: A Comparative Study of Random Forest, XGBoost and C5.0. *Journal of Information Systems Engineering and Management* 2025, 10(1), e10123. https://doi.org/10.29333/jisem/10123

[8] Mienye, I.; Jere, N. A Survey of Decision Trees: Concepts, Algorithms, and Applications. *IEEE Access* 2024, 12, 1–20. https://doi.org/10.1109/ACCESS.2024.3416838

[9] World Health Organization. **Guidelines for Drinking-Water Quality**, 4th ed.; WHO Press: Geneva, Switzerland, 2017. https://www.who.int/publications/i/item/9789241549950

[10] 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

ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 03,2026
Journal: https://www.academicpublishers.org/journals/index.php/ijai

[11] U.S. Environmental Protection Agency (EPA). **Turbidity: Water Quality Parameter Factsheet.** https://www.epa.gov/system/files/documents/2021-07/parameter-factsheet_turbidity.pdf

[12] A. Akhatov, N. Fayzulla and M. Eshmurodov, "Ensemble Machine Learning Algorithms for Crop Types Classification Based on Intelligent Detection of Soil Properties," *2025 International Russian Automation Conference (RusAutoCon)*, Sochi, Russian Federation, 2025, pp. 691-696, doi: 10.1109/RusAutoCon65989.2025.11177338.

[13] Kaggle. Water Potability Dataset. Available online: https://www.kaggle.com/datasets/uom190346a/water-quality-and-potability.

[14] World Health Organization. Guidelines for Drinking-Water Quality, 4th ed.; WHO Press: [15] Geneva, Switzerland, 2017. https://www.who.int/publications/i/item/9789241549950

Sayidqulov, Asliddin, and Fayzullo Nazarov. "Mashinani o'rganish modellari asosida iqtisodiy ko'rsatkichlarni baholashning aqlli algoritmlari." *Потомки Аль-Фаргани* 2 (2025): 49-54.

[16] 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. Detecting Potential Outliers in Longitudinal Data with Time-Dependent Covariates. *European Journal of Clinical Nutrition* 2024, 78(4), 344–350. https://doi.org/10.1038/s41430-023-01393-6

[17] F. Nazarov, A. Sayidkulov and S. Yarmatov, "Intelligent Data Analysis Algorithms Based on Neural Network Models of Machine Learning," *2025 International Russian Automation Conference (RusAutoCon)*, Sochi, Russian Federation, 2025, pp. 393-398, doi: 10.1109/RusAutoCon65989.2025.11177293.

[18] Raschka, S.; Mirjalili, V. Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow, 3rd ed.; Packt Publishing: Birmingham, UK, 2019.

[19] Mienye, I.; Jere, N. A Survey of Decision Trees: Concepts, Algorithms, and Applications. *IEEE Access* 2024, 12, 1–20. https://doi.org/10.1109/ACCESS.2024.3416838

[20] Zhao, L.; Lee, S.; Jeong, S. P. Decision Tree Application to Classification Problems with Boosting Algorithm. *Electronics* 2021, 10(16), 1903. https://doi.org/10.3390/electronics10161903