World Academy of Science, Engineering and Technology
International Journal of Electrical and Computer Engineering
Vol:17, No:10, 2023

Determination of Water Pollution and Water Quality with Decision Trees

Çiğdem Bakır, Mecit Yüzkat

***Abstract*—**With the increasing emphasis on water quality worldwide, the search for and expanding the market for new and intelligent monitoring systems has increased. The current method is the laboratory process, where samples are taken from bodies of water, and tests are carried out in laboratories. This method is time-consuming, a waste of manpower and uneconomical. To solve this problem, we used machine learning methods to detect water pollution in our study. We created decision trees with the Orange3 software used in the study and tried to determine all the factors that cause water pollution. An automatic prediction model based on water quality was developed by taking many model inputs such as water temperature, pH, transparency, conductivity, dissolved oxygen, and ammonia nitrogen with machine learning methods. The proposed approach consists of three stages: Preprocessing of the data used, feature detection and classification. We tried to determine the success of our study with different accuracy metrics and the results were presented comparatively. In addition, we achieved approximately 98% success with the decision tree.

***Keywords*—**Decision tree, water quality, water pollution, machine learning.

I. Introduction

IN many parts of the world, especially with the development of industry, water pollution is increasing. Industrial deterioration, waste and chemical pollution types that cause water pollution adversely affect human health [1]. For this reason, it is necessary to determine the factors that cause water pollution by constantly monitoring water resources. Analyzing the water, examining the water surface and coastal areas is very important today [2].

Chemical fertilizers, drugs, vegetable wastes, heavy metals, toxic substances, sewage waste, changes in the aquatic ecosystem and fuel consumption cause water pollution [3]. The purpose of our study is to detect water pollution, which adversely affects human health due to such reasons, with Machine Learning (ML) algorithms and to take the necessary precautions, because water pollution affects not only humans but also other living things. For example, plants may not receive the substances they need for photosynthesis. With the decrease in oxygen in the environment, living things in the seas may face the danger of extinction. The lives of fish and other sea creatures are at risk. Living things that feed on dirty water can invite viruses, bacteria and many microbes. This can cause many malignant diseases in both humans and animals.

In addition, access to safe drinking water is essential for health and is one of the basic human rights [4]. Dirty water is an important factor in the transmission of diseases such as cholera, diarrhea, dysentery, hepatitis, typhoid and polio. Inadequate and inappropriate water supply exposes individuals to serious health problems. Improved water supply, sanitation and proper management of water resources will not only increase the growth of countries, but also contribute to a great reduction in poverty [5].

Access to safe and easy water is important for public health [6]. Healthy drinking water should be odorless, colorless and clear. It should not contain disease-causing microorganisms. It should be sufficiently soft. It should not contain elements such as hydrogen sulfide, iron and manganese. There should be no harmful chemicals to health. The aim of our study is to investigate the factors that cause water pollution with ML algorithms in order to get rid of the destructive effects of water pollution, which harm all living things, and to analyze and evaluate the factors that cause water pollution to ensure that the necessary measures are taken immediately. One of the most important steps among the studies to prevent water pollution is to evaluate the water pollution event in a model. With the model we have created, it is aimed to reduce the destructive effects of water pollution on living things.

II. Related Works

Radhakrishnan and Pillai presented a comparison of water quality classification models using Support Vector Machine (SVM), decision tree and naive Bayes ML algorithms [7]. Properties used to determine water quality are pH, dissolved oxygen (DO), biochemical oxygen demand (BOD) and electrical conductivity (EC). Classification models are trained based on the arithmetic water quality index (WAQQI). In the study, two data sets were taken into account for testing the water quality. The first data set belongs to the samples obtained from the Narmada River, consisting of 28 different water quality parameters. The second data set consists of combining the water quality values collected from certain regions of India and belonging to the past years; it contains eight different parameters. Both datasets are taken from the website of the Government of India. Two performance parameters are defined to compare the efficiency of the three algorithms; these are, balanced accuracy score and confusion matrix. Based on these parameters, Table I was obtained for two data sets. Based on the results obtained, the decision tree algorithm was found to be the most appropriate classification model. Naive Bayes, on the

Ç. B. is with the Dumlupinar University, Engineering Department, Computer Engineering, Kutahya, Turkey (corresponding author, phone: 90-536-2072948; e-mail: cigdem.bakir@ dpu.edu.tr).

M.Y. was with Muş Alpaslan University, Engineering Department, Software Engineering, Turkey (e-mail: myuzkat@yildiz.edu.tr).

other hand, was found to be unsuitable.

TABLE I
CONFUSION MATRIX

Predicted Values	Real Values	
	True Positive	False Positive
	False Negative	True Negative

Ragi et al. tried to determine the water quality using Artificial Neural Network (ANN) [8]. The method they propose eliminates the chemical method for assessing water quality parameters. At the same time, this model is cost-effective. This article uses known parameters such as pH, EC, TDS, and includes a method for estimating unknown parameters such as basicity, chloride, sulfate using the Levenberg-Marquardt algorithm, provided by the Pollution Control Board of India. From samples obtained from different regions, 876 values were used for each trait. The Levenberg–Marquardt algorithm proposed in the study is a mixture of the Gauss–Newton algorithm and the gradient decent method. The biggest contribution of this study to the literature is the preprocessing technique, in which many mathematical operations are used to find the best input combination to the neural network. This causes the accuracy value of the results obtained from the ANN to increase.

Saghebian et al. present a quality estimation approach based on the United States Salinity Laboratory diagram from groundwater data from the agricultural districts of Ardebil province in the northwest region of Iran [9]. Performance evaluation is based on the number of correctly classified samples and kappa statistics. A decision tree-based approach has been found to be suitable for classification. Principal Component Analysis (PCA) was also used to identify important parameters for groundwater quality classification. EC and 11 months of accumulated precipitation can be used to assess groundwater quality. With the developed model, other parameters can be ignored by only looking at these two parameters, thus reducing laboratory costs and shortening the time between sampling time and obtaining laboratory results. The Ardebil basin is located in the Ardebil province in the northwestern region of Iran. Since precipitation is one of the main feeding sources for this basin, monthly precipitation data have been determined as one of the main features in the estimation of water quality. For this reason, precipitation data between the years 1995-2010 were obtained from 34 rain stations through the Ministry of Energy of Iran. In addition, the 15-year EC and sodium absorption rate (SAR) of 73 discharge wells in the same region were obtained to be used in the classification model. In other words, in this study [9], a total of two data sets were processed. In this study, a powerful, simple and applicable decision tree method has been developed for the classification of Ardebil's groundwater quality. This model was used with pre-measured hydrochemical data. Using this model, operators can determine the water quality class just by looking at two parameters. These parameters are EC and 11 months total precipitation data.

Priyadarshini et al. used ML methods to reduce water pollution for sustainable urban development [10]. In this study, two approaches are proposed. In the first approach, Random Forest (RF), decision trees, SVM, and ANN methods were used, and in the second approach, they tried to find the pollution in sea water with these methods. The F1 approach is used as the accuracy metric. They also wanted to reduce water pollution with different analyses.

Wu et al. used a real-time remote controlled multispectral unmanned aerial vehicle (UAV) to monitor changes in water quality [11]. They performed laboratory tests for measurement and sampling from the water. They separately observed four factors that determine water quality. They calculated the accuracy metrics of four water quality parameters using Extreme Gradient Boosting (XGB), RF and ANN ML algorithms. In their study, they obtained the trophic status from 45 different points and tried to determine the most suitable model to show the changes in water quality according to the space.

Hu et al. used an ensemble ML model to identify the main sources of nitrogen and phosphorus in the lakes [12]. They determined the water quality, environmental input and meteorological conditions in Taihu Lake using six ML techniques based on 13 years of historical data. The XGBoost model they used gave the most successful results for total nitrogen (TN) and total phosphorus (TP) estimation. The results show that lake TN is mainly affected by internal load and inflowing river water quality, while lake TP is mainly affected by internal sources. The U estimate is important for lake eutrophication control for both elements.

III. METHOD

A. Datasets

The information about the dataset we used in our study is given below:

The dataset contains the features described below for a total of 3276 different samples. According to these characteristics, it will be used to classify water as: Drinkable (1) or Non-Drinkable (0).

pH Value: The measure used to understand the acid or alkaline value of a substance is called pH. The pH of all nutrients and nutritional values is formed in the range of 0 to 14. Nutritional values between 0 and 7 are called acidic, and nutritional values between 7 and 14 are called basic. The pH value of pure water is 7; that is, neutral. In the researches [5] and [7], it is mentioned that the pH value of water resources should be between 6.5 and 8.5. If the pH goes too far out of this range, this water may not be safe to drink and may pose some health risks.

Degree of hardness: The hardness is due to the calcium and magnesium salts in the water. Although the degree of hardness of water has no known effect on human health, it is important to determine the hardness and, if necessary, remove it, since it causes the devices that use water to break down in the industry.

Dissolved solids: Water contains various minerals such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, and sulfate. It has the ability to dissolve some organic minerals or salts, as these minerals produce an

undesirable taste and dilute color in the appearance of water. This is an important parameter for water use. Water with a high dissolved solids value indicates that the water is highly mineralized. The desired limit for CCM is 500 mg/l and the maximum limit is 1000 mg/l, which is recommended for drinking purposes.

Chloramines: Chlorine and chloramine are the main disinfectants used in public water systems. Chlorination process is widely used in the treatment of drinking water, industrial water resources, pool water and wastewater. Chlorine levels of up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) in drinking water are considered safe.

Sulfate: Sulfate ion is an ion that is very common in natural waters and its concentration varies from a few mg per liter to several thousand mg. The main sources of sulfate in groundwater are gypsum and anhydrite. Sulphate from natural sources can be found in water to a certain extent. However, excess sulfate in drinking water makes the water taste bitter and can sometimes cause stomach and intestinal problems.

Conductivity: The conductivity of water depends on the dissolved ions in it. An increase in the conductivity of drinking water is a sign that the water is polluted, so the conductivity should not be above the established standards. In general, the amount of dissolved solids in water determines the EC. According to WHO standards, the EC value should not exceed 400 µS/cm [5].

Organic carbon: Total organic carbon is a measurement of the amount of organic compounds present in a water sample. Organic carbon-containing components may be dissolved or undissolved in water as suspended solids or liquids. This organic matter can enter the water naturally and through man-made sources/processes.

Trihalomethanes: Trihalomethanes are compounds formed by chlorine with organic or synthetic substances during disinfection processes. The most well-known is chloroform. THM levels of up to 80 ppm are considered safe in drinking water [5].

Turbidity: The clarity of the water is important for domestic consumption and most production sites. The water we drink should be colorless. If there is coloration in the water, it means that there are some metal ions (such as iron, manganese, chromium, nickel) dissolved in the water. Suspended substances such as sand, organic and inorganic substances, soluble colored organic compounds, plankton and other microscopic organisms cause turbidity in water. According to the World Health Organization, turbidity should be less than 5.00 NTU [5].

B. Method

Orange3 software will be used in the analysis of the data set. Orange3 is a component-based visual programming software package used in the fields of data mining and ML, where we can create clusters, groups, classes with data, perform visual and statistical analysis, and create models. While making these analyses, structures called widgets of the Orange3 software are used.

Orange3 consists of a canvas interface where the user places a widget and creates a data analysis workflow. Widgets can be used for reading data, displaying data table, selecting features, learning predictors, comparing learning algorithms, visualizing data items, etc. It provides basic functions; for example, the user can explore the visualizations interactively or feed the selected subset to other widgets.

The decision tree method will be used in the classification of the data set. Decision trees classification is a classification method that creates a model in the form of a tree structure consisting of decision nodes and leaf nodes according to features and targets [14]. The decision tree algorithm is developed by dividing the data set into small pieces. A decision node may contain one or more branches. The first node is called the root node. A decision tree can consist of both categorical and numerical data [15].

Several advantages of the decision tree as a classification tool have been noted in the literature [16]:

1. Decision trees are self-explanatory and easy to follow when compressed. In other words, if the decision tree has a reasonable number of permissions, it can be grasped by non-professional users. Also, decision trees can be transformed into a set of rules. Therefore, this representation is considered to be understandable.
2. Decision trees can handle both nominal and numeric input attributes.
3. The decision tree representation is rich enough to represent any discrete value classifier.
4. Decision trees can handle potentially erroneous datasets.
5. Decision trees are capable of processing datasets that may have missing values.
6. Decision trees are considered a non-parametric method. This means that decision trees have no assumptions about the field distribution and classifier structure.
7. When the cost of classification is high, decision trees can be attractive in that they only demand the values of features in a single root-to-leaf path.

On the other hand, decision trees have disadvantages such as [17]:

a) Most algorithms (like ID3 and C4.5) only require the target attribute to have discrete values.
b) Because decision trees use the "divide and conquer" method, they tend to perform well if there are several highly relevant features, but less if there are many complex interactions. One of the reasons is that other classifiers can compactly describe a classifier that would be very difficult to represent using a decision tree.

IV. Experimental Study

The flow diagram of the application created in the Orange3 tool is given in Fig. 1 [13]. The tools we used in our study are given in detail below:

File: The file structure is given in Fig. 2. This allows the dataset to be loaded from a directory. This dataset can be in xlsx, csv, txt or URL format. With the widget, the features, types and roles in the data set can be displayed.

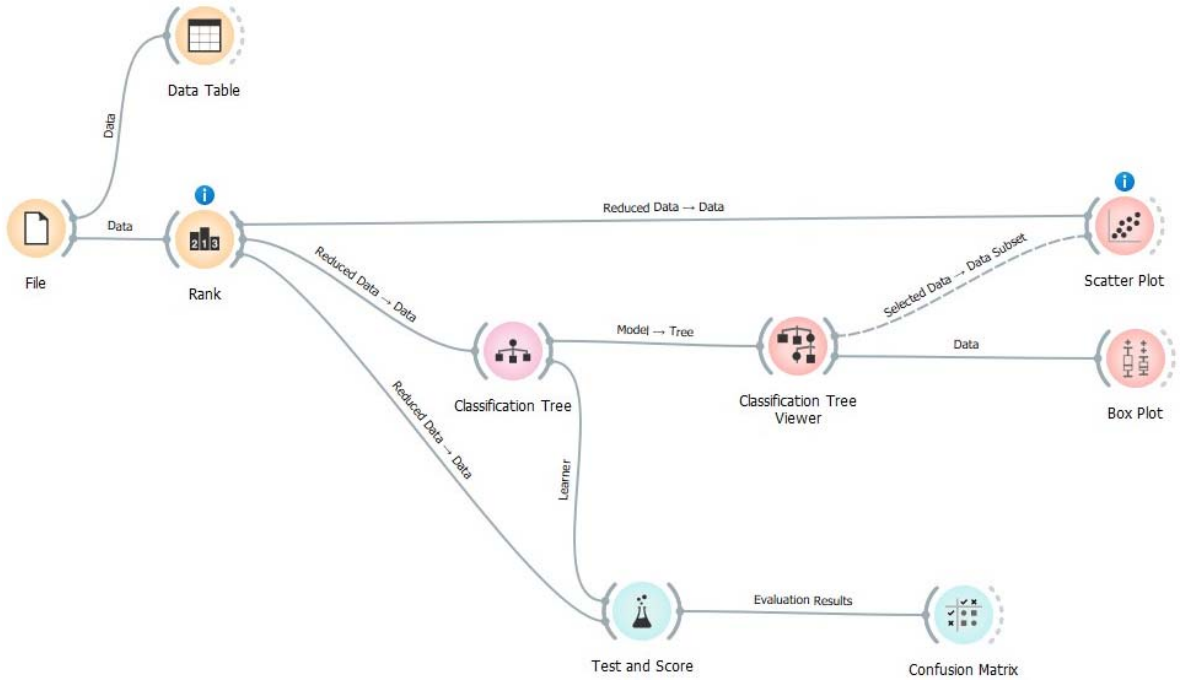


Fig. 1 Orange3 flow diagram

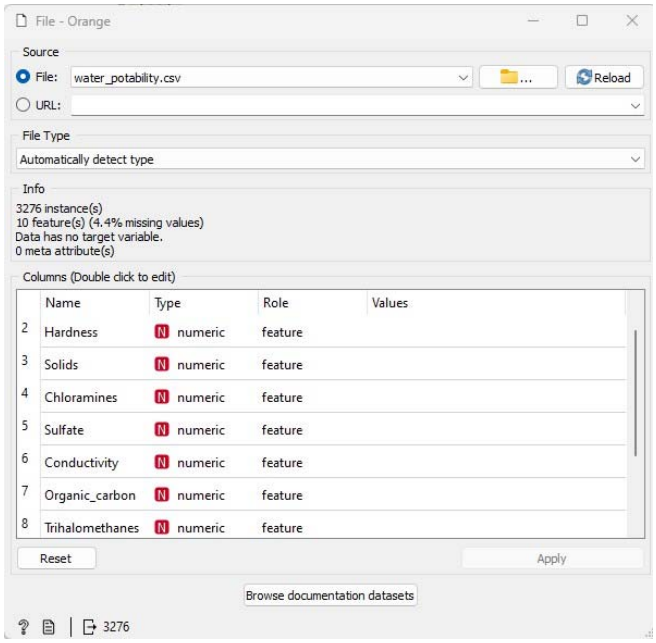


Fig. 2 File structure

There are 10 features and 3276 instances in our dataset. One of the properties is set to target. According to the potability value, it is determined whether the instance is potable or not.

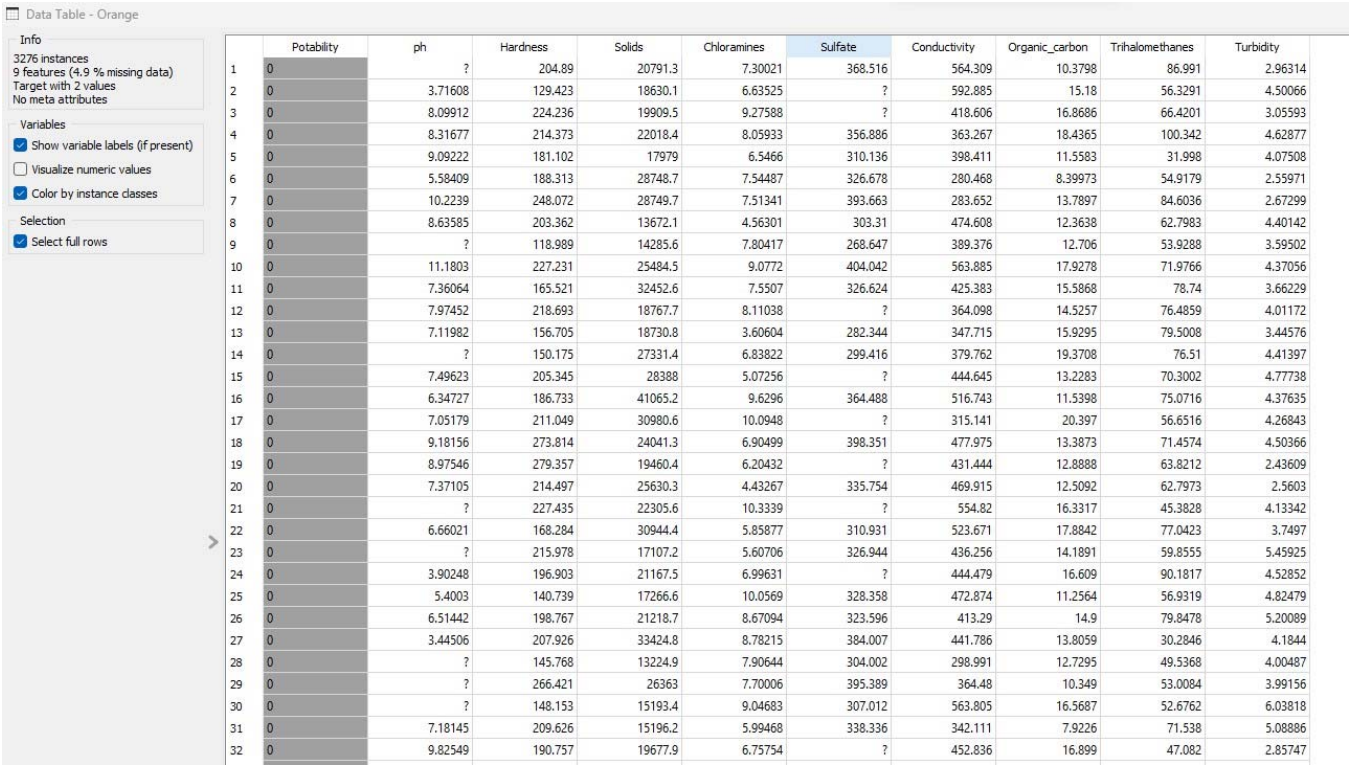
Data Table: Data table structure is given in Fig. 3. In a table, it shows the attributes and their values. When you click on it, the category information of the samples and the property values of each sample are displayed in the window that opens.

Feature Statistics: The feature statistics structure is given in Fig. 4. It shows basic statistical information for data characteristics, which are: dispersion, mean, median, dispersion, min, max and missing values. By looking at this information, the features that we will remove from the data set or keep in the data set can be determined, and a new data set can be created with only these features.

Rank: Rank structure is given in Fig. 5. It classifies the features in the data set according to the target class, using methods and scoring them from the most important to the least important. It disables features unimportant for the classification operation.

Classification Tree: The classification tree structure is given in Fig. 6. It represents the decision tree algorithm. It allows us to divide our dataset into classes.

Classification Tree Viewer: The classification tree viewer structure is given in Fig. 7. It is used to visualize classification and regression problems.



Data Table - Orange

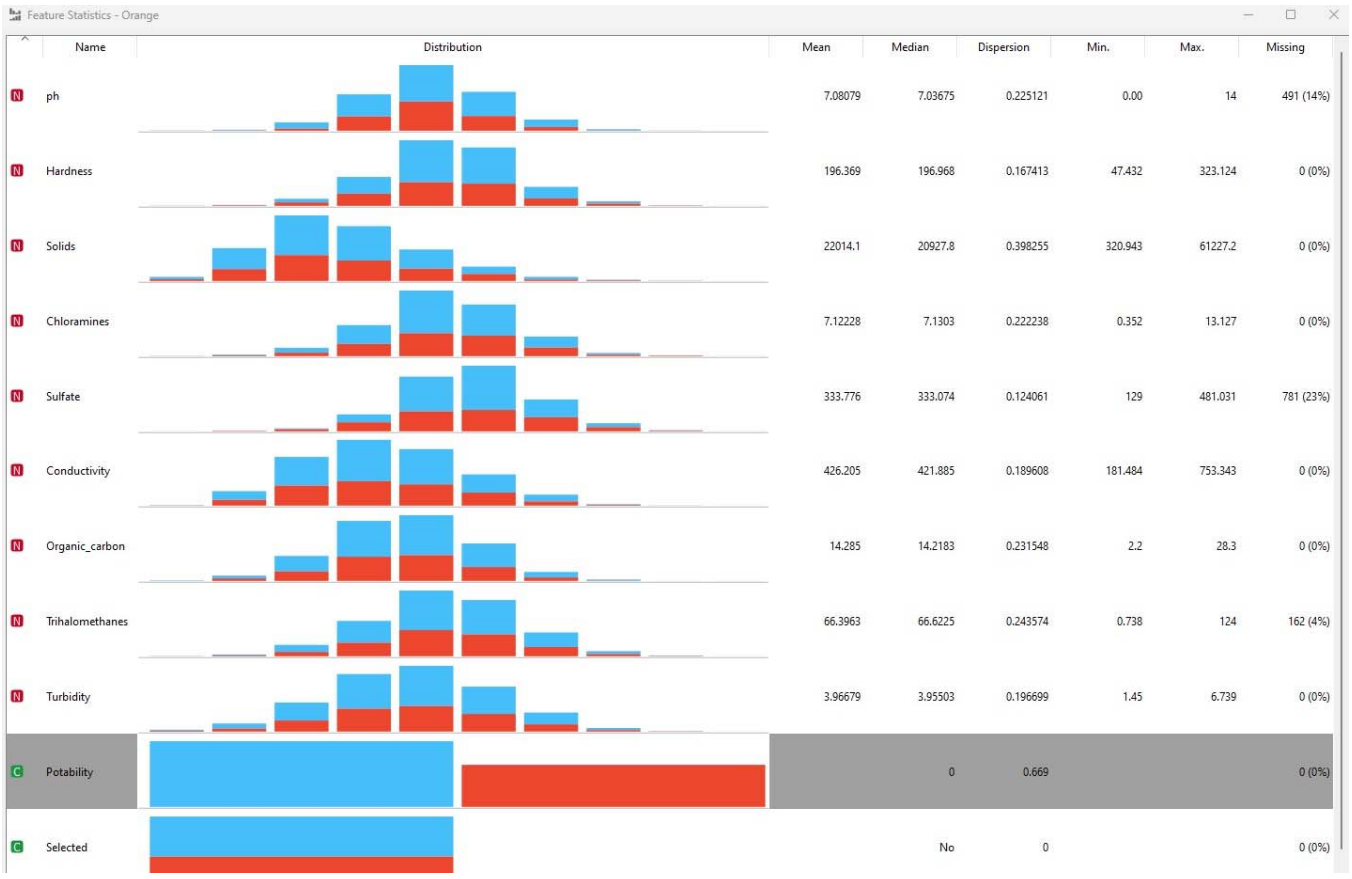
Info
3276 instances
9 features (4.9 % missing data)
Target with 2 values
No meta attributes

Variables
Show variable labels (if present)
Visualize numeric values
Color by instance classes

Selection
Select full rows

	Potability	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity
1	0	?	204.89	20791.3	7.30021	368.516	564.309	10.3798	86.991	2.96314
2	0	3.71608	129.423	18630.1	6.63525	?	592.885	15.18	56.3291	4.50066
3	0	8.09912	224.236	19909.5	9.27588	?	418.606	16.8686	66.4201	3.05593
4	0	8.31677	214.373	22018.4	8.05933	356.886	363.267	18.4365	100.342	4.62877
5	0	9.09222	181.102	17979	6.5466	310.136	398.411	11.5583	31.998	4.07508
6	0	5.58409	188.313	28748.7	7.54487	326.678	280.468	8.39973	54.9179	2.55971
7	0	10.2239	248.072	28749.7	7.51341	393.663	283.652	13.7897	84.6036	2.67299
8	0	8.63585	203.362	13672.1	4.56301	303.31	474.608	12.3638	62.7983	4.40142
9	0	?	118.989	14285.6	7.80417	268.647	389.376	12.706	53.9288	3.59502
10	0	11.1803	227.231	25484.5	9.0772	404.042	563.885	17.9278	71.9766	4.37056
11	0	7.36064	165.521	32452.6	7.5507	326.624	425.383	15.5868	78.74	3.66229
12	0	7.97452	218.693	18767.7	8.11038	?	364.098	14.5257	76.4859	4.01172
13	0	7.11982	156.705	18730.8	3.60604	282.344	347.715	15.9295	79.5008	3.44576
14	0	?	150.175	27331.4	6.83822	299.416	379.762	19.3708	76.51	4.41397
15	0	7.49623	205.345	28388	5.07256	?	444.645	13.2283	70.3002	4.77738
16	0	6.34727	186.733	41065.2	9.6296	364.488	516.743	11.5398	75.0716	4.37635
17	0	7.05179	211.049	30980.6	10.0948	?	315.141	20.397	56.6516	4.26843
18	0	9.18156	273.814	24041.3	6.90499	398.351	477.975	13.3873	71.4574	4.50366
19	0	8.97546	279.357	19460.4	6.20432	?	431.444	12.8888	63.8212	2.43609
20	0	7.37105	214.497	25630.3	4.43267	335.754	469.915	12.5092	62.7973	2.5603
21	0	?	227.435	22305.6	10.3339	?	554.82	16.3317	45.3828	4.13342
22	0	6.66021	168.284	30944.4	5.85877	310.931	523.671	17.8842	77.0423	3.7497
23	0	?	215.978	17107.2	5.60706	326.944	436.256	14.1891	59.8555	5.45925
24	0	3.90248	196.903	21167.5	6.99631	?	444.479	16.609	90.1817	4.52852
25	0	5.4003	140.739	17266.6	10.0569	328.358	472.874	11.2564	56.9319	4.82479
26	0	6.51442	198.767	21218.7	8.67094	323.596	413.29	14.9	79.8478	5.20089
27	0	3.44506	207.926	33424.8	8.78215	384.007	441.786	13.8059	30.2846	4.1844
28	0	?	145.768	13224.9	7.90644	304.002	298.991	12.7295	49.5368	4.00487
29	0	?	266.421	26363	7.70006	395.389	364.48	10.349	53.0084	3.99156
30	0	?	148.153	15193.4	9.04683	307.012	563.805	16.5687	52.6762	6.03818
31	0	7.18145	209.626	15196.2	5.99468	338.336	342.111	7.9226	71.538	5.08886
32	0	9.82549	190.757	19677.9	6.75754	?	452.836	16.899	47.082	2.85747

Fig. 3 Data table structure



Feature Statistics - Orange

Name	Distribution	Mean	Median	Dispersion	Min.	Max.	Missing
ph		7.08079	7.03675	0.225121	0.00	14	491 (14%)
Hardness		196.369	196.968	0.167413	47.432	323.124	0 (0%)
Solids		22014.1	20927.8	0.398255	320.943	61227.2	0 (0%)
Chloramines		7.12228	7.1303	0.222238	0.352	13.127	0 (0%)
Sulfate		333.776	333.074	0.124061	129	481.031	781 (23%)
Conductivity		426.205	421.885	0.189608	181.484	753.343	0 (0%)
Organic_carbon		14.285	14.2183	0.231548	2.2	28.3	0 (0%)
Trihalomethanes		66.3963	66.6225	0.243574	0.738	124	162 (4%)
Turbidity		3.96679	3.95503	0.196699	1.45	6.739	0 (0%)
Potability			0	0.669			0 (0%)
Selected			No	0			0 (0%)

Fig. 4 Feature statistics structure

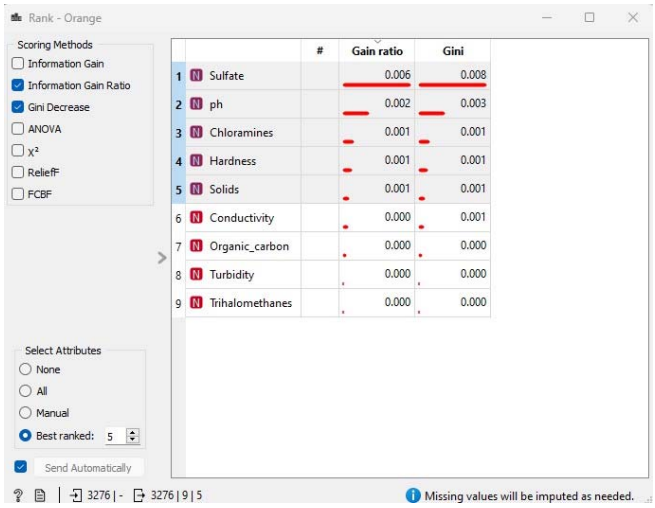


Fig. 5 Rank structure

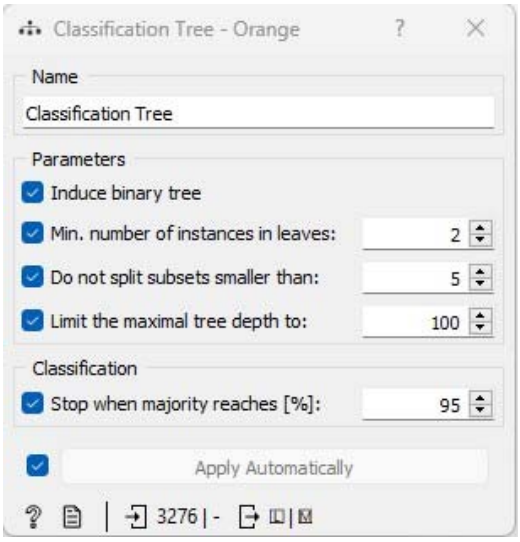


Fig. 6 Classification Tree structure

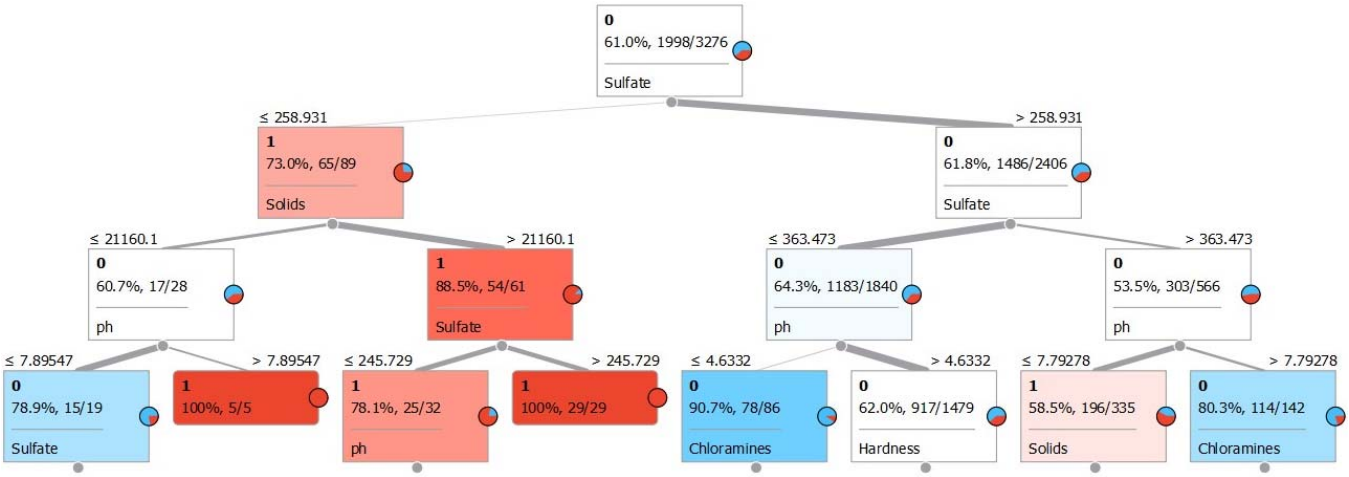


Fig. 7 Classification tree viewer structure

V. Conclusion

Test and score: It tests algorithms with data. As an introduction, it takes the dataset, tests data (if any) and algorithms (Learner) and outputs the test results of the classification algorithms. Along with separate test data, different sampling methods can be used. The widget does two things. First, it provides a table of classification performance measures such as classification accuracy, area under the curve. Second, ROC analysis gives evaluation results from which the performance of classifiers can be analyzed by other widgets such as confusion matrix. Classification results are given in Fig. 8.

The widget supports various sampling methods:

- Cross validation: Splits the data into a specified number of times (usually 5 or 10). The algorithm is tested by excluding samples from 1 layer at a time. The model draws conclusions from samples on other floors and the outside floor is classified. This process is repeated for all floors.
- Attribute cross-validation: Performs cross-validation but multiples are defined by the categorical attribute selected from the meta-properties.
- Random sampling: Randomly divides the data into the training and test set in the given ratio (e.g., 70:30); the whole procedure is repeated a certain number of times.
- Exclude one: Excludes 1 sample at a time. The model draws a conclusion from the other samples and classifies the left-out sample. This method is powerful and reliable, but slow.
- Test on train data: Uses the entire dataset for training and then testing. This method practically always gives false results.
- Test on test data: The above methods use data from data signal only. If it is desired to select the test data from a different file or a different widget, the Separate Test Data signal is selected in the communication channel and the test on test data are used.

The widget also calculates various performance statistics:

- Area under ROC
- Classification accuracy
- F-1

- Precision
- Recall
- Specificity
- Log-loss
- Train-time
- Test-time

Confusion Matrix: Confusion matrices are given in Figs. 9-11, which show the ratio of predicted and actual classes of data. It takes the results of the tested classification algorithms as input (usually from the Test & Score Widget) and outputs a subset of data selected from the confusion matrix.

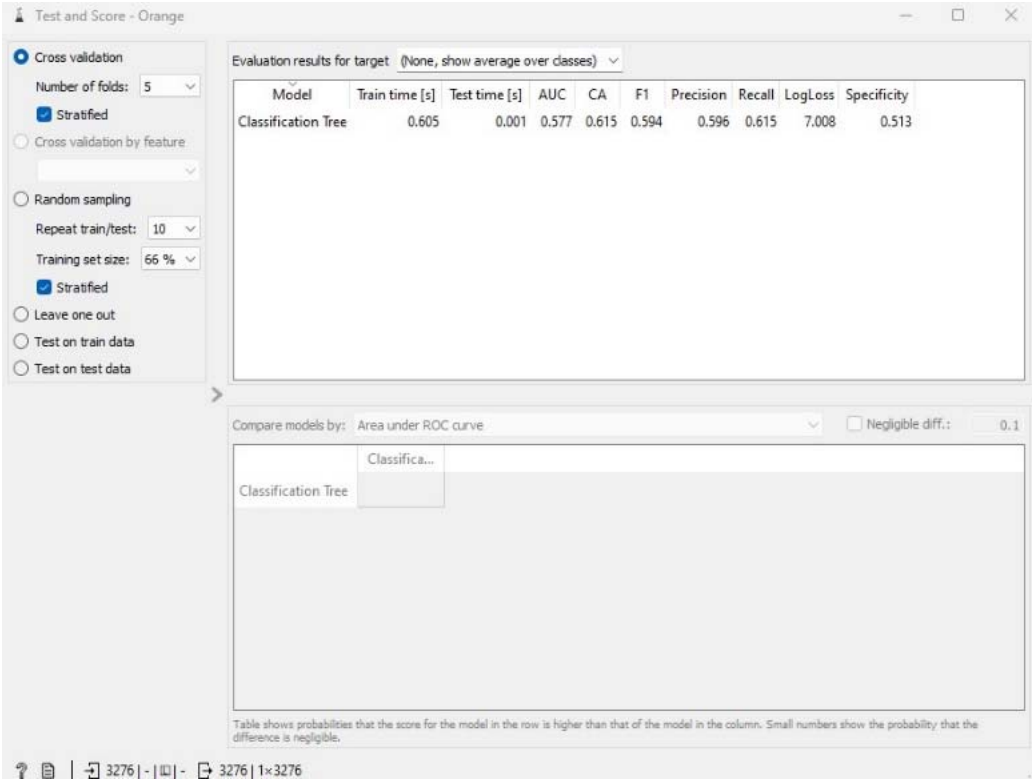


Fig. 8 Classification results

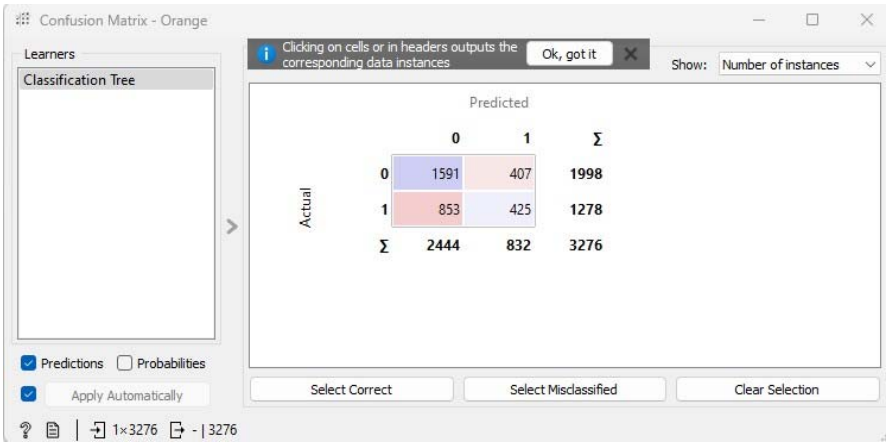


Fig. 9 Confusion matrix 1

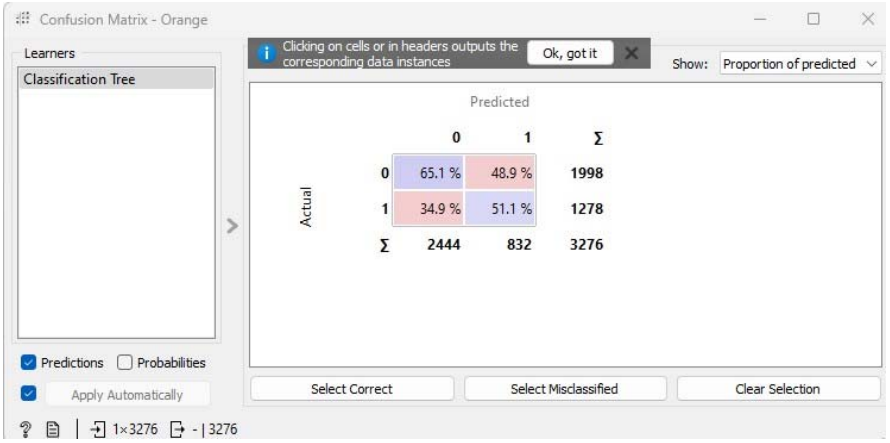


Fig. 10 Confusion matrix 2

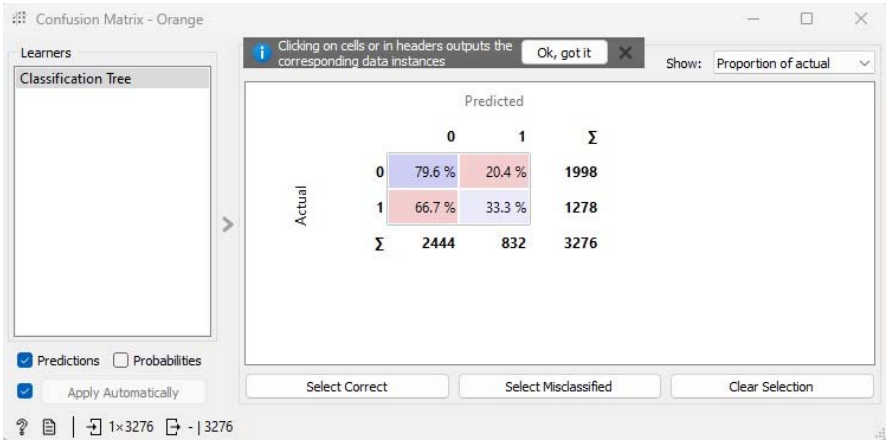


Fig. 11 Confusion matrix 3

REFERENCES

[1] Budka, M., Gabrys, B., & Ravagnan, E. (2010). Robust predictive modelling of water pollution using biomarker data. Water research, 44(10), 3294-3308.

[2] Zharikova, E. P., Grigoriev, J. Y., & Grigorieva, A. L. (2022, February). Artificial Intelligence Methods for Detecting Water Pollution. In IOP Conference Series: Earth and Environmental Science (Vol. 988, No. 2, p. 022082). IOP Publishing.

[3] Priyadarshini, I., Alkhayyat, A., Obaid, A. J., & Sharma, R. (2022). Water pollution reduction for sustainable urban development using machine learning techniques. Cities, 130, 103970.

[4] Muhammad, S. Y., Makhtar, M., Rozaimee, A., Aziz, A. A., & Jamal, A. A. (2015). Classification model for water quality using machine learning techniques. International Journal of software engineering and its applications, 9(6), 45-52.

[5] Shafi, U., Mumtaz, R., Anwar, H., Qamar, A. M., & Khurshid, H. (2018, October). Surface water pollution detection using internet of things. In 2018 15th international conference on smart cities: improving quality of life using ICT & IoT (HONET-ICT) (pp. 92-96). IEEE.

[6] Chen, H., Chen, A., Xu, L., Xie, H., Qiao, H., Lin, Q., & Cai, K. (2020). A deep learning CNN architecture applied in smart near-infrared analysis of water pollution for agricultural irrigation resources. Agricultural Water Management, 240, 106303.

[7] Radhakrishnan, N., & Pillai, A. S. (2020, June). Comparison of water quality classification models using machine learning. In 2020 5th International Conference on Communication and Electronics Systems (ICCES) (pp. 1183-1188). IEEE. DOI: 10.1109/icces48766.2020.9137903

[8] Ragi, N. M., Holla, R., & Manju, G. (2019, May). Predicting water quality parameters using machine learning. In 2019 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT) (pp. 1109-1112). IEEE. DOI: 10.1109/rteict46194.2019.9016825

[9] Saghebian, S. M., Sattari, M. T., Mirabbasi, R., & Pal, M. (2014). Ground water quality classification by decision tree method in Ardebil region, Iran. Arabian journal of geosciences, 7, 4767-4777.

[10] Priyadarshini, I., Alkhayyat, A., Obaid, A. J., & Sharma, R. (2022). Water pollution reduction for sustainable urban development using machine learning techniques. Cities, 130, 103970.

[11] Wu, D., Jiang, J., Wang, F., Luo, Y., Lei, X., Lai, C., ... & Xu, M. (2023). Retrieving Eutrophic Water in Highly Urbanized Area Coupling UAV Multispectral Data and Machine Learning Algorithms. Water, 15(2), 354.

[12] Hu, Y., Du, W., Yang, C., Wang, Y., Huang, T., Xu, X., & Li, W. (2023). Source identification and prediction of nitrogen and phosphorus pollution of Lake Taihu by an ensemble machine learning technique. Frontiers of Environmental Science & Engineering, 17(5), 55.

[13] Tunca, S., Wilk, V., & Sezen, B. (2023). Defining Virtual Consumerism Through Content and Sentiment Analyses. Cyberpsychology, Behavior, and Social Networking.

[14] Tangirala, S. (2020). Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. International Journal of Advanced Computer Science and Applications, 11(2), 612-619.

[15] Khraisat, A., Gondal, I., Vamplew, P., Kamruzzaman, J., & Alazab, A. (2020). Hybrid intrusion detection system based on the stacking ensemble of c5 decision tree classifier and one class support vector machine. Electronics, 9(1), 173.

[16] Kherif, O., Benmahamed, Y., Teguar, M., Boubakeur, A., & Ghoneim, S. S. (2021). Accuracy improvement of power transformer faults diagnostic using KNN classifier with decision tree principle. IEEE Access, 9, 81693-81701.

[17] Yadav, D. C., & Pal, S. (2020). Prediction of thyroid disease using decision tree ensemble method. Human-Intelligent Systems Integration,

2, 89-95.

Çiğdem Bakır is an Assistant Professor of Software Engineering at the Engineering Faculty – Kutahya University, Turkey. She received the B.S. degree in computer engineering from the University of Sakarya, in 2010, and the M.S. and Ph.D. degrees in computer engineering from Yildiz Technical University, Istanbul. She is currently pursuing the doctorate degree in computer science with the University of Yildiz Technical, Istanbul. She was a Research Assistant at Yildiz Technical University and Igdir University. She was an Instructor at Erzincan Binali Yildirim, from 2020 to 2021. She has been an Assistant Professor with the Software Engineering Department, Dumlupinar University, since 2021. Her research interests include information security, distributed database, big data, blockchain technology, cloud computing, and computer networks.

Mecit Yüzkat is a Research Assistant of Software Engineering at the Engineering Faculty – Muş Alpaslan University, Turkey. He received the B.S. degree in computer engineering from the University of Trakya, in 2010, and the M.S. and Ph.D. degrees in computer engineering from Yildiz Technical University, Istanbul. He is currently pursuing the doctorate degree in computer science with the University of Yildiz Technical, Istanbul. His research interests include patern recognition, data mining and deep learning.