Journal of Ecohumanism
2025
Volume: 4, No: 1, pp. 3057– 3067
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i1.6103

Water Potability Classification System: Regular Water Monitoring Computational Cost Reduction by Utilizing Recursive Feature Elimination with Cross-Validation and SelectKBest for Feature Reduction

Indrabayu[1], Fahruddin[2], Nurhalisa[3], Herlina Abdul Rahim[4], Mohamed Sultan Mohamed Ali[5], Syahidah Nurani Zulkifli[6]

Abstract

Drinking water treatment process requires periodic measurements every hour to meet drinkable water standards. This study aims to classify water potability in Gowa and perform feature selection to identify the most optimal parameters. The research uses the SVM and XGBoost for classification and employs RFECV and SelectKBest for feature selection. The results show that most correlations between parameters and the target are weak, indicating that each parameter operates independently and has unique value in determining drinking water potability. The study achieves high model accuracy, with 95.8% for SVM and 97.8% for XGBoost. After feature selection, the final accuracy for the SVM model is 95.8% using the SelectKBest with 3 selected features: turbidity, free chlorine, and temperature. Using the RFECV, the accuracy is 96% with 5 selected features: turbidity, temperature, free chlorine, alkalinity, and TDS. For XGBoost, the final accuracy after feature selection is 97.8% using the SelectKBest with 5 selected features: turbidity, free chlorine, temperature, pH, and alkalinity. The RFECV feature selection for XGBoost also maintains the same accuracy of 97.8%. Based on the results, XGBoost performs slightly better than SVM, but RFECV improves SVM accuracy while maintaining XGBoost accuracy. The SelectKBest method also maintains the accuracy for both models.

Keywords *: SVM, XGBoost, RFECV, SelectKBest, Water Potability*

Introduction

Water can be sourced from various sources, such as groundwater, river water, lake water, rainwater, and others. The most commonly used sources of drinking water are groundwater and surface water. However, some areas have water quality that is not guaranteed to be safe for consumption, such as in urban areas. Water can become contaminated through human activities, including industrial waste disposal, fuel storage system leaks, or seepage from landfills (Ministry of Health Republic of Indonesia, 2024). This contamination can introduce harmful substances and bacteria into the water, as seen in the Gowa area, where water sourced from the polluted Jeneberang River is distributed. To address this, it is necessary to treat raw water to make it potable.

Based on interviews conducted at the IPA Pandang-Pandang PDAM Gowa, one of the drinking water providers in Gowa Regency, South Sulawesi, Indonesia, their guidelines specify that raw water, once treated into clean water, must be inspected routinely every hour. Additionally, they conduct a complete parameter check twice a year. These measures are taken to ensure that the water distributed to customers is safe for consumption. Manually checking water quality periodically is challenging because it requires measuring each parameter and verifying whether it meets the standards every hour. Therefore, technology is needed to monitor water quality automatically using machine learning.

In water quality assessment, the measurement of various parameters is required. According to the regulation of the Ministry of Health, Republic of Indonesia, No. 2 of 2023, the mandatory parameters for potable water quality consist of 19 types of parameters, including microbiological, physical, and chemical parameters (Ministry of Health Republic of Indonesia, 2023). Additionally, there are special parameters set by the regional government based on the geohydrological conditions of the area. IPA Pandang-Pandang PDAM

[1] Hasanuddin University, Email: indrabayu@unhas.ac.id
[2] Hasanuddin University, Email: fahruddin_science@unhas.ac.id
[3] Hasanuddin University, Email: nurhalisa20d@student.unhas.ac.id
[4] Universiti Teknologi Malaysia, Email: herlina@utm.my
[5] Universiti Teknologi Malaysia, Email: sultanali@utm.my
[6] Universiti Teknologi Malaysia, Email: snurani2@gmail.com

Journal of Ecohumanism
2025
Volume: 4, No: 1, pp. 3057– 3067
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i1.6103

Gowa measures water quality using 27 parameters, although this measurement is only conducted twice a year due to limitations in tools and resources.

Therefore, this research will conduct feature selection to identify the parameters that should be prioritized for regular water monitoring. Previous studies have explored similar approaches, such as classifying drinking water quality by comparing the performance of three machine learning models: J48, Naive Bayes, and MLP. These models were trained using different configurations of features selected through Pearson Correlation Coefficient-based feature selection. The results indicated that MLP achieved the best performance with all features, yielding a precision value of 0.673 and a recall of 0.869. This study aims to classify water potability using feature selection methods that go beyond linear correlations like Pearson, specifically employing Recursive Feature Elimination with Cross-Validation (RFECV) and SelectKBest (Abuzir & Abuzir, 2022). RFECV method is employed because it uses cross-validation to estimate the model's performance at each iteration of feature/parameter removal, helping to avoid overfitting and improving estimation accuracy. Additionally, RFECV has been shown to enhance the performance of Support Vector Machines (SVM). For example, research by Irfan Pratama demonstrated that using RFECV for feature selection in employee demographics and employment track record data led to an optimal selection of six features out of twenty-nine, improving model accuracy by 3% compared to models without feature selection (I. Pratama et al., 2022). In addition, research with the same method was also carried out by Arief Riski Indra Pratama, who optimized rainfall classification using SVM and RFE. The study found that applying RFE to SVM improved accuracy by 2%, from 77% to 79%. The best features identified were three out of ten, namely average temperature (Tavg), duration of sunshine (ss), and minimum temperature (Tn) (Pratama et al., 2022). In addition to RFECV, this research will also use the SelectKBest method for feature selection, allowing for a performance comparison with the RFECV method.

Objectives

RO 1. How is the correlation between features and targets?

RO 2. How to build a water potability classification system using machine learning?

RO 3. How to find the parameters that need to be prioritized in the water potability classification process?

RO 4. How accurate is the classification system for water potability through feature selection and classification without feature selection?

Methodology

Design System

The design flow of the system that will be carried out in the research is shown in figure 1.

Journal of Ecohumanism
2025
Volume: 4, No: 1, pp. 3057– 3067
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i1.6103

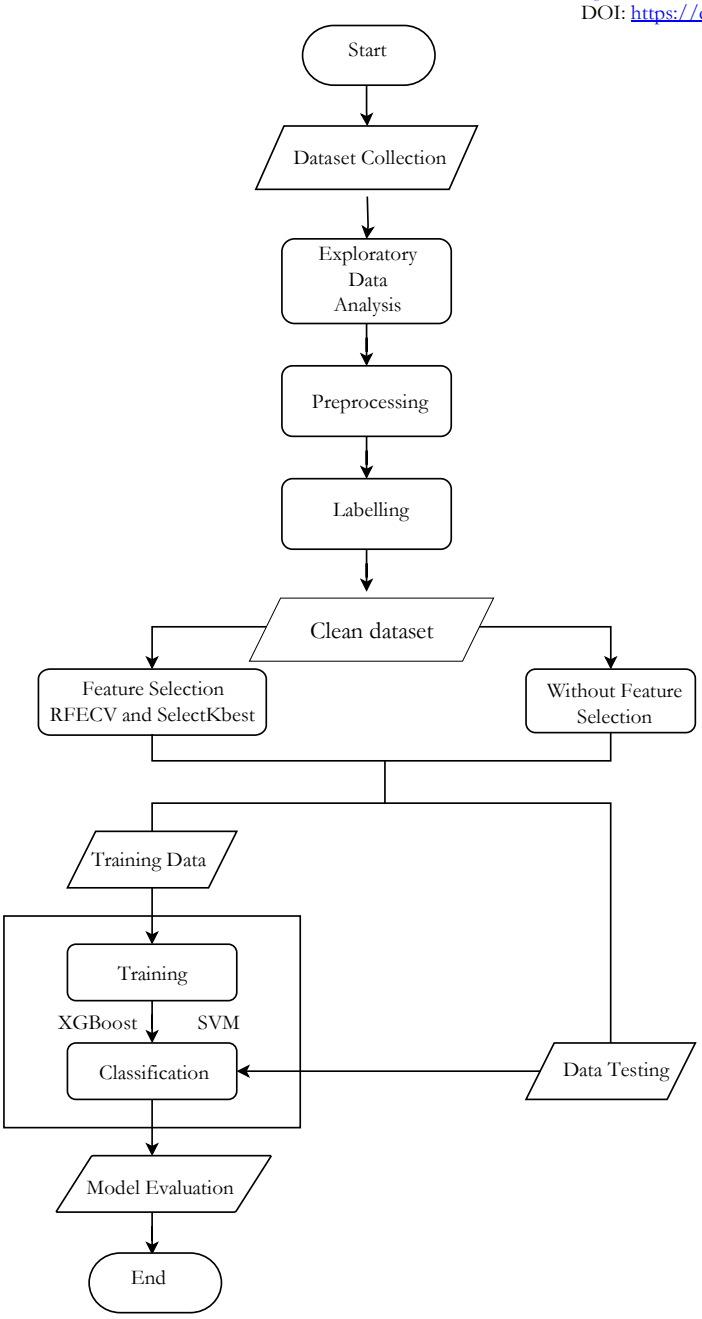

Figure 1. Design System Flow

Data Collection

The primary data used is the daily clean water monitoring data taken from PDAM and samples of water that has been distributed to customers consisting of 6 water parameters, namely turbidity, pH, temperature, free chlorine and tds. Data collection was carried out at the Water management installations of Pandang-Pandang, Gowa, South Sulawesi, Indonesia. The data consists of 2488 data collected from 2020 to July 2024. Meanwhile, secondary data is daily average air temperature data in the areas where water inspections are conducted accessed through the NASA Power Single Point Data Access service, which is a service provided by NASA (National Aeronautics and Space Administration, USA) with a latitude of -5.215 and a longitude of 119.4567 (*NASA POWER*, 2024). Air temperature data is used in the process of labeling water temperature data. Defenitions of features are described below.

Journal of Ecohumanism
2025
Volume: 4, No: 1, pp. 3057– 3067
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i1.6103

Table 1. Defenitions of Features

Features	Defenitions
Turbidity	Measurements that use the effect of light as the basis for the state of a water sample on the NTU (Nephelometrix Turbidity Unit) scale
pH	A standard used to state the level of acidity or alkalinity possessed by a sample of water in the form of a value.
Temperature	Measurement of the hot or cold intensity of a water sample.
Air Temperatures	Average air temperature in the water measurement area
Free Chlorine	Chlorine in water that acts as hypochlorous acid that functions as a disinfectant
Alkalinitas	Measures of water capacity to neutralize acids
Total Dissolve Solid (TDS)	The amount of dissolved solids in the form of organic ions, compounds, and colloids in the water sample.

Exploratory Data Analysis

After data collection, the data is analyzed to gain insights into its key characteristics. This approach typically uses statistical graphs and data visualization techniques to summarize and present the data in an easy-to-understand manner. At this stage, several functions are carried out such as the identification of missing values and blank lines, the identification of outliers using the Interquartile Range (IQR), feature correlation with the pearson correlation method and data distribution.

Data Preprocessing

After the exploration of data analysis, the next stage is the data preprocessing. This stage is a data processing stage to clean data from problems found in the previous stage into data that is ready for further processing. The preprocessing stage is carried out such as data cleaning that includes deleting blank rows and handling missing values, then the data that has been cleaned will go through a normalization process using the standard scaling method.

Data Labelling

After the data is cleaned, then data labeling is carried out based on the Minister of Health of the Republic of Indonesia No. 2 of 2023 concerning Drinking Water Quality Standards. After that, all water parameters that meet the standards will be labeled as "potable" with a value of 1 while if any of the parameters are not eligible then they will be labeled as "non-potable" with a value of 0 (Patel et al., 2023). Drinking water quality standards are shown in table 2.

Table 2. Drinking Water Quality Standards

Parameters	Standards	Unit
Turbidity	< 3	NTU
pH	6.5 – 8.5	-
Temperature	Air temperature ± 3	°C
Free Chlorine	0.3 – 0.5 with a contact time of 30 minutes	mg/L
Total Dissolve Solid (TDS)	<300	mg/L

Source: Regulation of the Ministry of Health Republic Indonesia Number 2 of 2023

Model Development

The development model in the training process uses the SVM and XGBoost methods. SVM is a powerful method for classification development. It aims to create a decision boundary between two classes that allows for the prediction of labels or more vector features (Huang et al., 2018). The RBF kernel is used to

Journal of Ecohumanism
2025
Volume: 4, No: 1, pp. 3057– 3067
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i1.6103

solve the problem of data that is not linearly separated (Ma'ruf et al., 2019). The RBF kernel uses two parameters, namely Gamma and Cost. SVM training is carried out using tuning parameters, namely using the RBF (Radial Basis Function) kernel with a value of C= 1000, gamma = 'scale' and max_iter= 1000. Here is the equation of the RBF kernel:

$$\text{Radial Basis Function (RBF): } K(= \exp\vec{x}_i, \vec{x}_j)\left(-\gamma \times \left\|\vec{x}_i - \vec{x}_j\right\|^2\right) \quad (1)$$

where

$K(\vec{x}_i, \vec{x}_j)$ = RBF kernel function

$\vec{x}$ = Input vector

Exp = Exponential function

Γ = Gamma parameter, determining the influence of individual training samples on the function of the decision

While the method XGBoost is one of the implementations of gradient boosting which is known as one of the best performing algorithms used to supervised learning. This algorithm can be used for prediction and classification problems, and it also has high execution speeds outside of core computing (Ibrahem Ahmed Osman et al., 2021). XGBoost is an improved algorithm based on gradient boosting decision tree and can build boosted trees efficiently and operate in parallel. Inside the regression tree, Nodes The inside represents the values for the attribute test and leaf nodes with a score that represents the decision (Karo, 2020). XGBoost training process uses parameters with values n_estimators=100, learning_rate = 0.1 and max_depth = 5. The way XGBoost makes predictions can be seen in equation (2).

$$\hat{y}_i = \sum_{k=1}^{n} f_k(x_i), f_k \in F \quad (2)$$

where

y_i = XGBoost model prediction value

x_i = Input vector

k = Index on each function fk

f_k = function k^{th}, decision tree function k^{th}

$f_k(x_i)$ = The function k^{th} decision tree when given an input x_i

Journal of Ecohumanism
2025
Volume: 4, No: 1, pp. 3057– 3067
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i1.6103

F = The set of all decision tree functions

The objective function of XGBoost can be seen in equation (3)

where

$$Obj(\theta) = L(\theta) + \Omega(\theta) \quad (3)$$

Obj(θ) = Objective function

θ = Parameter model

L(θ) = Loss/loss function

Ω(θ) = Regularization functions that control the complexity of the model

$L(\theta) = \sum_{i=1}^{n} l(yi, y\hat{}i)$ is a loss function, y^i is the prediction and yi is the target and Ω(θ) is a regularization that adds up every complexity of each decision tree. Then the model is trained in an additive way. Then let $= \sum_{k=1}^{K} \Omega(fk)$ $\hat{y}i^{(t)}$ be the prediction of instance i in iteration t, and can be expressed in equation (4).

$$\hat{y}i^{(t)} = \hat{y}i^{(t-1)} + ft(xi) \quad (4)$$

where

$\hat{y}i^{(t)}$ = Prediction for the i^{th} instance/ data on the t^{th} iteration

$\hat{y}_i^{(t-1)}$ = Prediction for the i^{th} instance/ data in the previous iteration(t-1)

ft = New decision tree function added in t^{th} iteration

xi = input for instance/i^{th} data

ft(xi) = Prediction of the new tree for the i^{th} instance

Figure 2 shows the general architecture of the XGBoost model can be seen.

Journal of Ecohumanism
2025
Volume: 4, No: 1, pp. 3057– 3067
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i1.6103

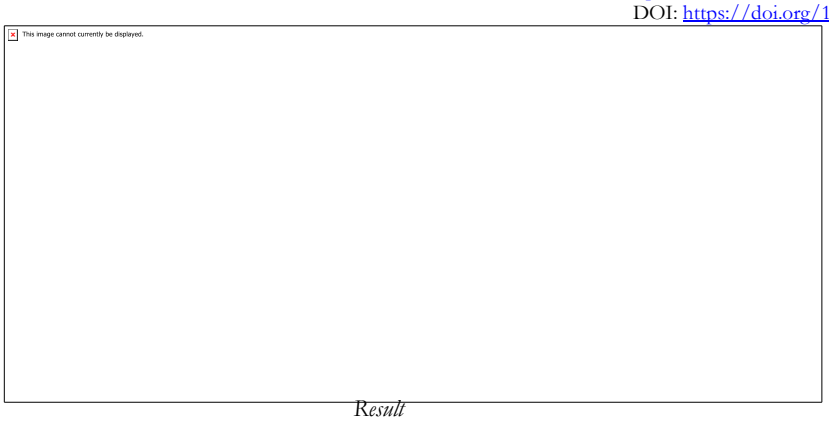

Figure 2. General Architecture of the XGBoost model (Wang et al., 2019)

Based on Figure 2, the final prediction of XGBoost is generated from the sum of the results of all decision trees used where each tree (Tree 1,2,...n) is trained to correct the prediction error of the previous tree.

Feature Selection Method

The feature selection methods used are RFECV and SelectKbest. RFE is a method of feature selection wrapper which iteratively removes the least important features based on model performance to identify and rank the most significant predictors (Harif & Kassimi, 2024). The RFECV method works by iteratively removing non-important features based on model performance. It then automatically generates the optimal feature using cross-validation. Additionally, RFECV can also eliminate dependencies and collinearities that exist in the model (Shi dkk., 2024). In the SVM model, this method uses LinearSVC as an estimator because it has the ability to give the important weight of the feature. Then the value of step = 1, the number of k-folds = 5 and the scoring "accuracy" is set. Meanwhile, the estimator used in the XGBoost model is XGBClassifier.

The SelectKbest method is one of the feature selection methods that selects the k feature with the highest top score which is calculated based on univariate statistical analysis which is an analysis of variables one by one (Desyani et al., 2020). This method works by selecting the best features based on the ANOVA test, then eliminating features that are not included in the best features based on the specified number of features. SelectKbest selects the top k features with the greatest relevance to the target variable (Fitri et al., 2023). In this feature selection method, the final model performance is tested one by one by determining the value of k starting from 1 to the entire number of features. This is done to find out the k value with the highest performance. The model will be trained using all parameters and parameters of the feature selection. So in this research, there are six test scenarios carried out, namely SVM without feature selection, XGBoost without feature selection, SVM with RFECV feature selection, SVM with SelectKbest feature selection, XGBoost with RFECV feature selection and XGBoost with SelectKbest feature selection.

Model Evaluation

This stage will be evaluated using the Confusion Matrix with the calculation of Accuracy, Precision, Recall, and F1-Score. This metric is used because it provides detailed information about the correct and incorrect predictions for each class. This is very important because it can have a bad impact on the health of people who consume water that is detected incorrectly. Classification performance evaluation is used on classification results using the selection feature and classification results without the selection feature (final model). The two will be compared to see which model works optimally. The results of the comparison will be interpreted with a classification report and heatmap visualization.

Journal of Ecohumanism
2025
Volume: 4, No: 1, pp. 3057– 3067
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i1.6103

Results and Discussion

Correlation Feature Analysis

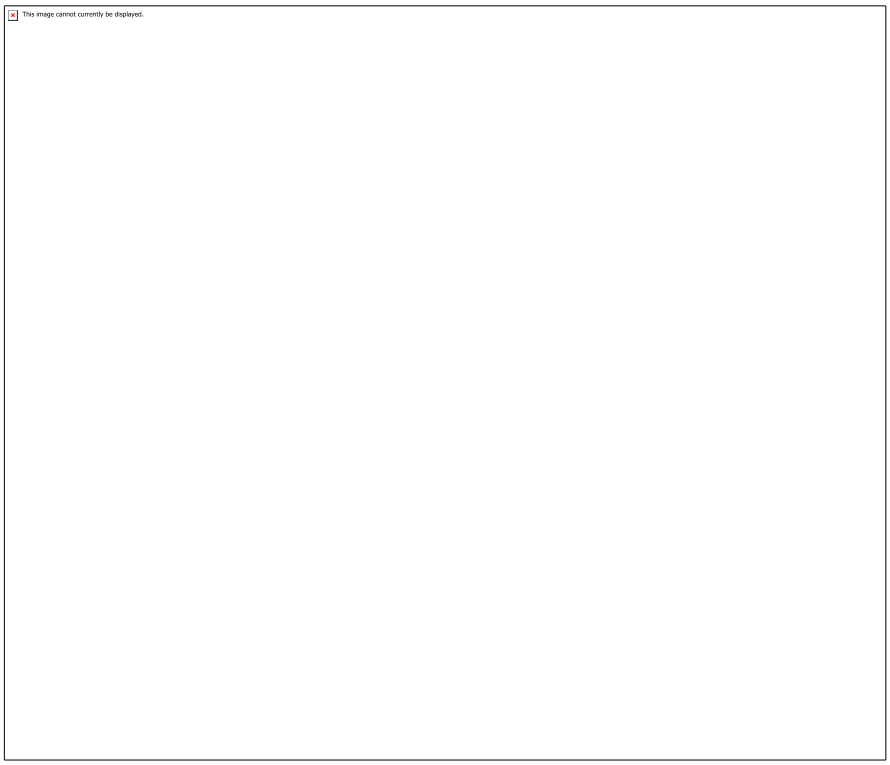

Figure 3. Correlation Feature

Based on Figure 3, it can be seen that there is no strong correlation between each features (>0.75) of water parameters. Based on the resulting feature correlation heatmap, it can be seen that water temperature and air temperature have a fairly strong positive correlation with a value of 0.45. This is because water temperature tends to be influenced by the temperature of the surrounding air. In addition, TDS was also moderately positively correlated with water temperature (0.39) and air temperature (0.37). Then turbidity had the strongest negative correlation with water temperature (-0.24). This suggests that murkier water tends to be slightly colder. While other parameters have a weak correlation so that these parameters hardly affect each other. Based on the results of the analysis of the correlation of features shown on the heatmap, it shows that most of the parameters tend to be independent of each other. The correlation of features to the target also shows a fairly low correlation. Based on the resulting feature correlation heatmap, it can be seen that there are two features that have the highest negative correlation with the feature, namely turbidity with a value of -0.29 and free chlorine with a value of -0.24. While the others had a positive correlation with a very low value and approached 0, namely pH with a value of 0.07, then alkalinity with a value of 0.05 and tds with a value of 0.04. This means that the determination of the water potability cannot be determined based on one of the water parameters alone.

Feature Selection

The following is the cross-validation score in the RFECV feature selection process using two estimator models, namely LinearSVC and XGBClassifier, shown in the following figure.

Journal of Ecohumanism
2025
Volume: 4, No: 1, pp. 3057– 3067
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i1.6103

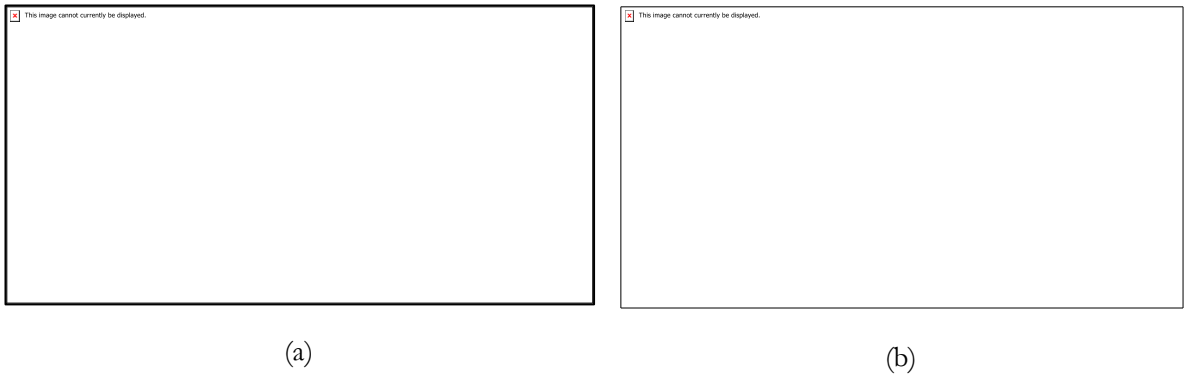

Figure 4. Cross-validation Scores (a) RFECV with SVM & (b) RFECV with XGBoost

In figure 4(a) is the cross-validation score in the selection of RFECV features with the LinearSVC estimator, the results show that the optimal number of features is 5 features with a score of 0.74. The selected features are 'Turbidity (NTU)', Temperature (°C), 'Free chlorine (mg/L)', 'Alkalinity (ppm)' and 'TDS (ppm)'. This means that there is only one feature that is selected, namely 'ph'. While figure 4(b) is the cross-validation score in the selection of RFECV features with the XGBClassifier estimator, the result is the same as SVM, namely there are 5 optimal features with a much higher score of 0.98. In addition, the scores with features 5 and 6 have the same value, which means that the model is able to maintain the model's performance even if more features are included.

In addition to the RFECV feature selection model, the following is the feature importance score from the results of the SelectKbest feature selection.

Figure 5. Feature Importance Scores SelectKBest

In the image above, it can be seen that there are only two features that have a fairly high importance score with a score of >25, namely 'turbidity' and 'Free Chlorine' while the other features have a very low importance score and are almost the same as the score of <5, namely 'ph', 'temperature', 'alkalinity' and 'tds'.

Journal of Ecohumanism
2025
Volume: 4, No: 1, pp. 3057– 3067
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i1.6103

Model Classification Analysis

Table 3. Model Classification Performance

Method	Accuracy	Precision	Recall	F1-Score	Number of Features	Selected Features
SVM without feature selection	0.9575	0.9575	0.9575	0.9575	6	All features
SVM-RFECV	0.9597	0.9597	0.9597	0.9597	5	Turbidity, temperature, free chlorine, alkalinity and tds.
SVM-SelectKbest	0.9575	0.9575	0.9575	0.9575	3	Turbidity, temperature, free chlorine.
XGBoost without feature selection	0. 9776	0. 9776	0. 9776	0. 9776	6	All features
XGBoost-RFECV	0. 9776	0. 9776	0. 9776	0. 9776	5	Turbidity, temperature, free chlorine, alkalinity and tds.
XGBoost-SelectKbest	0. 9776	0. 9776	0. 9776	0. 9776	5	Turbidity, temperature, free chlorine, alkalinity and tds.

In table 3, it can be seen that all models used scored almost the same evaluation metrics. All of these models have very strong performance where SVM without feature selection and SVM with feature selection using SelectKbest with 3 features have the same metric value of 0.9575. Meanwhile, SVM with feature selection using RFECV has a slightly higher metric of 0.9579 and uses 5 features. This means that feature selection using the RFECV method has succeeded in predicting well and can increase accuracy. On the other hand, XGBoost has the same performance on models without feature selection and with feature selection, which is 0.9776. XGBoost is able to maintain its performance even by selecting one feature. Feature selection with RFECV is able to select one feature, namely ph with the same performance, while feature selection with XGBoost selects one feature, namely tds with the same performance as well. This means that XGBoost is able to maintain the same performance after a selection of features which shows that this model is more stable and less dependent on removed features.

Conclusion

Feature correlation using pearson correlation shows that most of the correlation between parameters and the parameter's correlation to the target is relatively weak, so this shows that each parameter works independently or is not too influenced by other parameters and each parameter has its own unique value in determining the water potability. The SVM and XGBoost models also succeeded in classifying water potability very effectively based on existing parameters with the same evaluation metric results, namely 95.6% using SVM and 97.8% using XGBoost. In addition, feature selection is also successful in selecting features without significantly reducing performance. In SVM, RFECV feature selection succeeded in selecting 1 feature, namely ph with a slightly higher evaluation metric than SVM before feature selection, which was 96%. Meanwhile, SelectKbest managed to select 3 features with the lowest scores, namely ph, alkalinity, and tds with the same evaluation metrics as before the feature selection, which was 95.7%. As

Journal of Ecohumanism
2025
Volume: 4, No: 1, pp. 3057– 3067
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i1.6103

for XGBoost, RFECV and SelectKbest feature selection both selected 1 feature, namely the ph feature in RFECV and tds in SelectKbest with the same evaluation metric of 97.8%. Based on the results of the research conducted, it can be concluded that XGBoost with RFECV is more effective and stable in maintaining its performance in classifying the potability of drinking water even though there is one reduced feature. However, the removed feature does not mean that the feature can be eliminated absolutely in the water quality measurement process because based on the results of the feature correlation, the existing features work independently. So that the results of the feature selection only provide an understanding of what parameters need to be prioritized in periodic water quality measurements.

Acknowlegments

The authors wish to thank the Hasanuddin University and Universiti Teknologi Malaysia (UTM) for financing this research with grant vote no. 04M88 and 4B914.

References

Abuzir, S. Y., & Abuzir, Y. S. (2022). Machine learning for water quality classification. Water Quality Research Journal, 57(3), 152–164. https://doi.org/10.2166/wqrj.2022.004

Desyani, T., Saifudin, A., & Yulianti, Y. (2020). Feature Selection Based on Naive Bayes for Caesarean Section Prediction. IOP Conference Series: Materials Science and Engineering, 879(1), 012091. https://doi.org/10.1088/1757-899X/879/1/012091

Fitri, E. N., Winarno, S., Budiman, F., Rohmani, A., Zeniarja, J., & Sugiarto, E. (2023). Decision Tree Simplification Through Feature Selection Approach In Selecting Fish Feed Sellers. Jurnal Teknik Informatika (Jutif), 4(2), 301–309. https://doi.org/10.52436/1.jutif.2023.4.2.747

Harif, A., & Kassimi, M. A. (2024). Predictive Modeling of Student Performance Using RFECV-RF for Feature Selection and Machine Learning Techniques. International Journal of Advanced Computer Science and Applications, 15(7), 231–240.

Huang, S., CAI, N., & PACHECO, P. P. (2018). Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics & Proteomics, 15(1). https://doi.org/10.21873/cgp.20063

Ibrahem Ahmed Osman, A., Najah Ahmed, A., Chow, M. F., Feng Huang, Y., & El-Shafie, A. (2021). Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Engineering Journal, 12(2), 1545–1556. https://doi.org/10.1016/j.asej.2020.11.011

Karo, I. M. K. (2020). Implementation of XGBoost Method and Feature Importance for Classification in Forest and Land Fires. Journal of Software Engineering, 1(1).

Ma'ruf, F. A., Adiwijaya, & Wisesty, U. N. (2019). Analysis of the influence of Minimum Redundancy Maximum Relevance as dimensionality reduction method on cancer classification based on microarray data using Support Vector Machine classifier. Journal of Physics: Conference Series, 1192, 012011. https://doi.org/10.1088/1742-6596/1192/1/012011

NASA POWER. (2024). Data from NASA Power Data Access Viewer (2020-2024). NASA POWER. https://power.larc.nasa.gov/data-access-viewer/

Patel, S., Shah, K., Vaghela, S., Aglodiya, M., & Bhattad, R. (2023). Water Potability Prediction Using Machine Learning. https://doi.org/10.21203/rs.3.rs-2965961/v1

Pratama, A. R. I., Latipah, S. A., & Sari, B. N. (2022). Optimasi Klasifikasi Curah Hujan menggunakan Support Vector Machine (SVM) dan Recursive Feature Elimination (RFE). JIPI (Scientific Journal of Informatics Research and Learning), 7(2), 314–324. https://doi.org/10.29100/jipi.v7i2.2675

Pratama, I., Chandra, A. Y., & Presetyaningrum, P. T. (2022). Seleksi Fitur dan Penanganan Imbalanced Data menggunakan RFECV dan ADASYN. Journal of Explora Informatics, 11(1), 38–49. https://doi.org/10.30864/eksplora.v11i1.578

Regulation of the Minister of Health Republic Indonesia No. 2 of 2023 concerning Environmental Health Quality Standards. Jakarta: Ministry of Health Republic Indonesia, 2023.

Shi, K., Shi, R., Fu, T., Lu, Z., & Zhang, J. (2024). A Novel Identification Approach Using RFECV–Optuna–XGBoost for Assessing Surrounding Rock Grade of Tunnel Boring Machine Based on Tunneling Parameters. Applied Sciences, 14(6), 2347. https://doi.org/10.3390/app14062347

Wang, Y., Pan, Z., Zheng, J., Qian, L., & Li, M. (2019). A hybrid ensemble method for pulsar candidate classification. Astrophysics and Space Science, 364(8), 139. https://doi.org/10.1007/s10509-019-3602-4