SITEKNIK
Sistem Informasi, Teknik dan Teknologi Terapan
E-ISSN: 3032-3991 P-ISSN: 3090-1626
Vol. 2. No. 4 October 2025. Pages. 317-323

Water Quality Analysis and Consumption Feasibility Using Support Vector Machine and CatBoosting with Hyperparameter Tuning

Christa Putri Rahayu[1]✉, Kusnawi[2]

christa.staput@students.amikom.ac.id, khusnawi@amikom.ac.id
[1,2] Informatics, Universitas Amikom Yogyakarta, Indonesia

Keywords: Water Quality, Support Vector Machine, CatBoosting, SMOTE, Hyperparameter Tuning

Submitted: 18/08/2025
Revised: 25/09/2025
Accepted: 13/10/2025

Abstrak

Water quality analysis plays an important role in determining the suitability of water for human consumption. This study aims to build a machine learning model that is able to classify water quality based on several parameters such as pH, hardness, solids content, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity. The dataset used comes from Kaggle with a total of 3,276 sample data. The two main algorithms applied in this study are Support Vector Machine (SVM) and CatBoost. The research process includes data preprocessing, data balancing using SMOTE, modeling, and model performance evaluation. Hyperparameter tuning is applied to both algorithms to improve performance. The results show that CatBoost has the best performance with an accuracy of 95.8% after hyperparameter tuning, compared to SVM which achieved an accuracy of 77.9%. In addition, CatBoost excels in all evaluation metrics, including precision, recall, and F1-score.

Corresponding Author:
Christa Putri Rahayu, Kusnawi
Informatics Study Program, Faculty of Computer Science, Amikom University YogyakartaJl. Ring Road Utara, Condong Catur, Sleman, Yogyakarta Telp:(0274) 884201 –207
Email: christa.staput@students.amikom.ac.id

INTRODUCTION

Water is a natural resource that is crucial for human life and other living things(Simbolon, 2024). However, the availability of clean water is decreasing due to urban growth that ignores water catchment areas(Riyantoko et al., n.d.). The clean water crisis is a global problem, with only 1% of the earth's total water being suitable for consumption. According to the WHO, approximately 663 million people experience difficulty accessing clean water(PENGAMANAN KUALITAS AIR MINUM, n.d.).

Safe drinking water must meet quality standards based on physical, chemical, and microbiological parameters. In Indonesia, these standards are regulated by the Minister of Health Regulation Number 492/MENKES/PER/IV/2010. If water does not meet these standards, it can endanger human health(Azmi et al., 2022).

Machine learning has great potential in water quality analysis, with the ability to identify patterns in data and classify water based on applicable standards. This study used the water_potability.csv dataset from Kaggle, which contains ten water quality parameters. The methods used were Support Vector Machine (SVM) and CatBoost. SVM is effective in handling high-dimensional data and can be used for nonlinear data using the kernel trick(Bidang Komputer Sains dan Pendidikan Informatika et al., n.d.). Meanwhile, CatBoost is a boosting algorithm that combines gradient boosting and decision trees, improving prediction efficiency and accuracy(*UDARA KOTA PALEMBANG Oleh : NURCHAERANI KADIR*, 2024).

This research aims to develop a machine learning model capable of classifying water quality based on available attributes, evaluate the accuracy of the SVM and CatBoost algorithms, and provide deeper insights into the factors influencing water quality through data analysis.

RESEARCH METHODS

CatBoost is gradient boosting -based machine learning algorithm designed For process feature categorical and numeric in a way efficient(*CatBoost*, n.d.). Algorithm CatBoost developed For increase Machine Learning model performance with focus on speed , accuracy , and ability handling feature categorical (Adi, 2023).

$$ctri = \frac{countinclass + prior}{totalcount + 1} \quad (1)$$

Information :
$ctri$ = i-th data in feature categorical
C$ount$ inclass = shows how many times the label value exceeds i For object with mark feature categorical
Prior = something number constant set by the initial parameters
T$otalCount$ = total number of objects that have mark appropriate features with mark feature .

Support Vector Machine (SVM) is machine lots of powerful learning used For linear and nonlinear classification , as well as task detection regression and outliers. How SVM works in non-linear problems is with enter kernel concept in space dimensionless height . In a dimensional space this , later will searching for separator or often called a hyperplane. A hyperplane can maximize distance or margin between data class . Best hyperplane between second class can found with measure the margin and then look for point maximum(Pratiwi, 2020).

Following This is a number of general kernel functions :

Polynomial Kernel with Variables Free q

$$K(\overrightarrow{Xi}, \overrightarrow{Xj}) = (\overrightarrow{Xi}, \overrightarrow{Xj} + 1)^{\varphi} \quad (2)$$

Information :

(Xi) $^{\rightarrow}$ = Data vector in observation i-th
(Xj) $^{\rightarrow}$ = Data vector in observation to -j
q = Free parameter that determines degree of polynomial .
K((Xi) $^{\rightarrow}$,(Xj) $^{\rightarrow}$) = Calculated kernel value For observation i and j.
Gaussian Kernel or RBF

$$K(\overrightarrow{Xi}, \overrightarrow{Xj}) = exp\left(\frac{\left\|\vec{X} - \overrightarrow{Xi}\right\|^2}{2\sigma^2} \right) \quad (3)$$

Information :

$(Xi)^{\rightarrow}$ = Data vector in observation i-th

$(Xj)^{\rightarrow}$ = Data vector in observation to -j

$\|X^{\rightarrow} - (Xi)^{\rightarrow}\|$ = Euclidean distance between two vectors .

exp = Function exponential applied to the result calculation

σ = Gaussian scale (bandwidth) parameter that controls level influence of neighboring data .

$K((Xi)^{\rightarrow},(Xj)^{\rightarrow})$ = Calculated kernel value For observation i and j.

Object study This focus on analysis water quality using the water_potability dataset from Kaggle, which consists of of 10 columns and 3275 data. This dataset covers various parameters, such as pH, hardness, solids, chloramines , sulfate , sonoductivity , organic carbon , trihalomethanes and turbidity, with target potability labels indicating water suitability for consumption . Research this aim for analyze connection between these parameters and create a predictive model use Support Vector Machine (SVM) and Catboosting algorithms for classification.

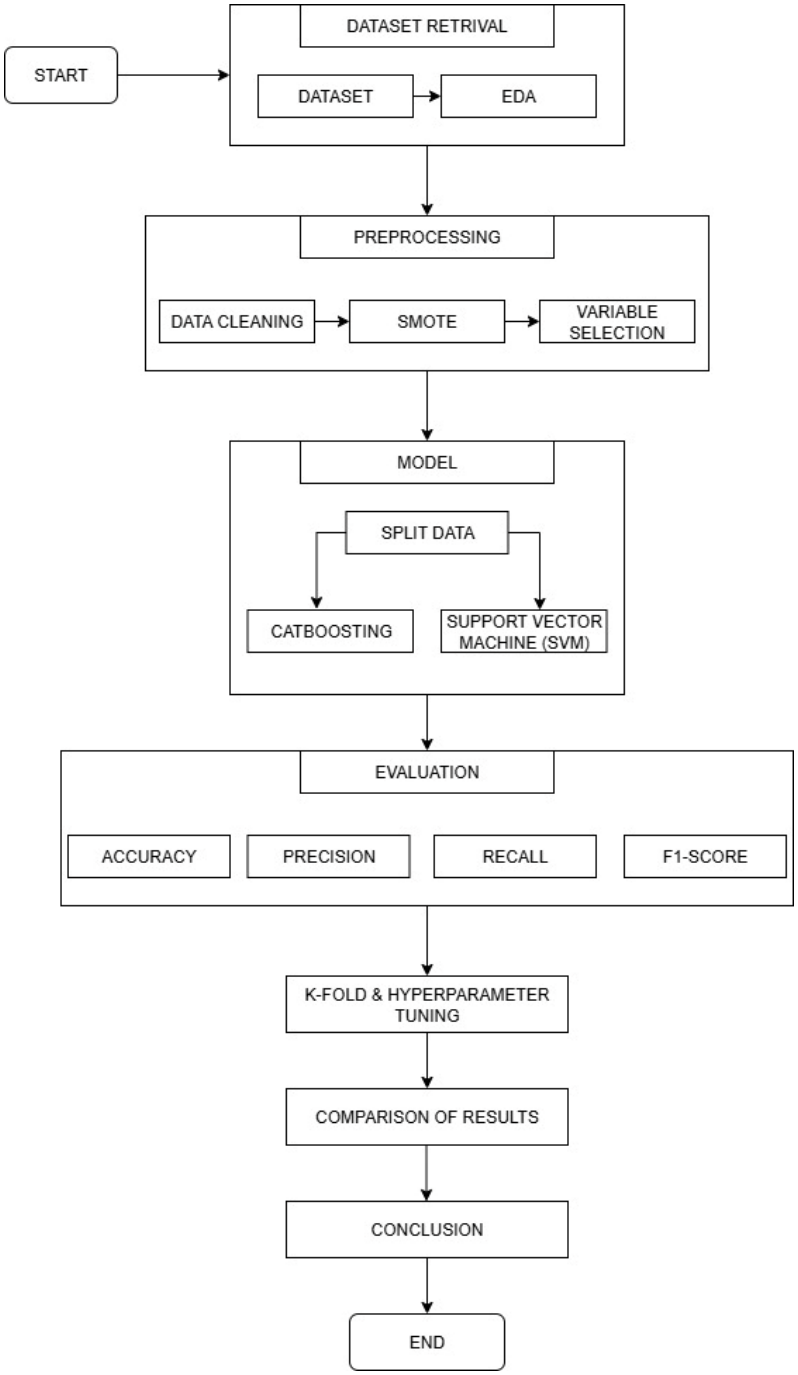

Figure 1. Research Flow

In this research, data cleaning was carried out to handle missing values and ensure the dataset was consistent. The process involved deleting rows that contained missing values, thereby producing a cleaner dataset that could be used effectively for model training and evaluation. This step was essential to avoid biased or inaccurate results caused by incomplete information.

SMOTE (Synthetic Minority Over-sampling Technique) was also applied during preprocessing to address the problem of imbalanced data. By generating synthetic samples for the minority class, SMOTE helped balance the dataset, which in turn improved the classification performance. This step ensured that the models could learn more effectively from both majority and minority classes, resulting in more reliable predictions.

The dataset was then split into three parts: 80% for training, 10% for validation, and 10% for testing. This division aimed to provide sufficient data for training the models while still reserving portions for validation and testing. The validation set was used to fine-tune the models, while the testing set served to evaluate their final performance.

After splitting the data, the modeling process was conducted using two machine learning algorithms, namely Support Vector Machine (SVM) and CatBoost. Both algorithms were chosen for their strong capabilities in classification tasks, allowing the research to compare their performance in predicting water quality based on the given parameters.

To evaluate the models more comprehensively, K-fold cross-validation was applied. This method divided the dataset into several folds, ensuring that all data points were used as both training and testing data alternately. By doing so, it reduced the risk of overfitting and provided a more accurate measure of the model's overall performance.

Hyperparameter tuning was performed to optimize the performance of the machine learning models. By adjusting key parameters, the models were able to work more effectively, resulting in improved accuracy and predictive capabilities. This step played a significant role in enhancing the overall performance of both SVM and CatBoost.

Finally, an evaluation and comparison of the results were conducted to measure the accuracy, precision, recall, and F1-score of both algorithms. This analysis provided insights into the effectiveness of the models before and after hyperparameter tuning. By comparing these metrics, the research was able to highlight the improvements gained and determine which algorithm performed best in classifying water quality.

RESULTS AND DISCUSSION

Data Cleaning

At this data cleaning stage done For overcome missing value problem. The number of missing values found in the dataset used covers pH as much as 491, sulfate as many as 781, and trihalomethanes as many as 162. Following from the missing value data that can be obtained seen under this.

```
Missing values di setiap kolom sebelum penghapusan:
ph                491
Hardness            0
Solids              0
Chloramines         0
Sulfate           781
Conductivity        0
Organic_carbon      0
Trihalomethanes   162
Turbidity           0
Potability          0
dtype: int64
```

Figure 2. Missing Value

For handle problem above, then from That researchers choose For deletion each row that contains missing values.

SMOTE

SMOTE is A method data processing for can handle absence balance class in the dataset, data imbalance occurs moment amount sample data in class majority and minority Far more A little with class the majority of those who make the model are less capable analyze class minority(Agung et al., 2024). Smote this done For overcome problem absence balance of data in the dataset. Smote works too For increase performance classification on the previous dataset No balanced. Here difference After and before smote can be seen in Figure 3.

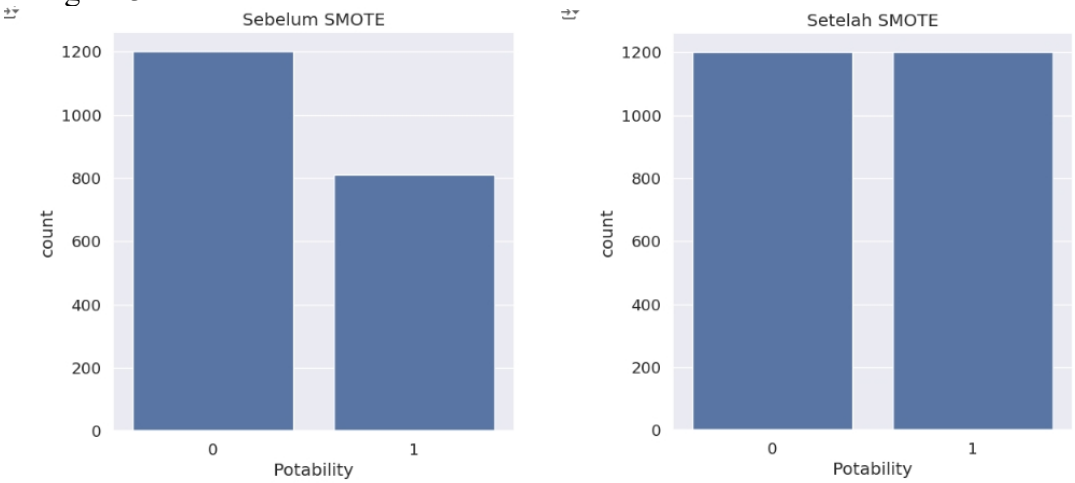

Figure 3. SMOTE

From the picture above can seen that (1) in the picture first (before SMOTE) bar plot shows difference for potency 0 it is at count 1200 while potency 1 is at count 800, (2) in the picture second (after SMOTE) can seen that potability 0 and 1 are at count 1200 or balanced.

Split D ata

At the split data stage it can be said, right? For divide the data into 3, namely, the training set of 80% is used For train the model, Validation set of 10% is used For evaluate model performance during training, Testing Set of 10% is used For test performance end of model on unprocessed data Once seen previously.

SVM

This Support Vector Machine (SVM). can Work with good on multidimensional data sets high. Initially, the Support Vector Machine method only used For linear data classification. However Now, SVM is developed for nonlinear data with applying kernel tricks. How it works method This is find hyperplane and margin for maximize between class classification.

Catboosting

CatBoost is one of the Boosting algorithm, which is known with reliability and capability highly efficient prediction algorithm. CatBoost is development more carry on from Gradient Boosting and Decision Tree methods. CatBoost perfect draft efficient ensemble learning with use technique lifting sorting and symmetric Decision Tree as classifier weak.

K-Fold Cross Validation

Cross validation is one of the data resampling method used For estimate error actual model predictions as well as For setting model parameters. Applying k-fold cross validation, we can obtained knowledge about how much good model that we use in generalize new data and make more decisions appropriate about its suitability For evaluate performance. Applying k-fold cross validation, can obtained knowledge about how much good model that we use in generalize new data and make more decisions appropriate about its suitability For evaluate performance , as well as For determine amount maximum iteration For used in validate cross Kfold- nya (Wijiyanto et al., 2024).

Hyperparameter Tuning

Hyperparameter tuning works with find the most optimal hyperparameter set. The method is to test the hyperparameters by trial and error until find mark optimal. Study This use method Gridsearch provided by Sckit -learn for find the best hyperparameters. GridsearchCV is hyperparameter method possible tuning for do scanning on a number of hyperparameters selected. In GridsearchCV, a number of hyperparameter combination will applied to models and performance from every combination will evaluated use cross-validation. Combination of hyperparameters with performance best will chosen as the best hyperparameter For model (Muhamad Malik Matin, 2023).

COMPARISON OF RESULTS

Table 1. Comparison Before and After Hyperparameters of SVM and Catboost models

Model	Precision %		Recall%		F1-Score%		Accuracy %
	Class 0	Class 1	Class 0	Class 1	Class 0	Class 1	
SVM (Before)	65%	74%	75%	63%	70%	68%	69%
Catboost (Before)	66%	75%	75%	65%	70%	69%	70%
SVM (After)	74%	82%	82%	75%	78%	78%	78%
Catboost (After)	96%	96%	96%	96%	96%	96%	96%

SVM algorithm before hyperparameter tuning is performed, the results the best is own accuracy of 69.1%. Then the precision shows that class 0 gets 65% mark and got 1st class value 74%, next is a recall that shows that class 0 gets 75% score and class 1 gets value 63%. For the last one is the f1-score which shows that class 0 gets score 70% and get class 1 value 68%. Algorithm Catboost before hyperparameter tuning is performed, this results the best is own accuracy by 70%. Then for precision it shows that class 0 gets score 66% and get 1st class value 75%, then is a recall that shows that class 0 gets 75% score and class 1 gets value 65%. For the last one is the f1-score which shows that class 0 gets score 70% and get class 1 value 69%. SVM algorithm after hyperparameter tuning is performed, the results the best is own accuracy of 77.9%. Then the precision shows that class 0 gets score 74% and got class 1 value 82%, next is a recall that shows that class 0 gets 82% score and class 1 gets value 75%. For the last one is the f1-score which shows that class 0 gets score 78% and get 1st class value 78%. Algorithm Catboost after hyperparameter tuning is performed, the results the best is own accuracy of 95.8%. Then the precision shows that class 0 gets a score of 96% and grade 1 got value 96%, then is a recall that shows that class 0 gets 96% score and class 1 gets value 96%. For the last one is the f1-score which shows that class 0 gets a score of 96 % and 1st class value 96%.

CONCLUSIONS AND SUGGESTIONS

Conclusion

Based on the results and discussion of the research, it can be concluded that the SVM and CatBoost algorithms are capable of classifying water quality based on parameters such as pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity. Both algorithms demonstrate reliable performance in predicting water quality classification accuracy, with CatBoost achieving an initial accuracy of 70%, slightly higher than SVM with 69.1%. The application of hyperparameter tuning further enhanced their performance, showing a significant improvement in accuracy for both models. After tuning, CatBoost's accuracy drastically increased to 95.8%, while SVM's accuracy improved to 77.9%.

To achieve better and more comprehensive results, it is recommended to use datasets with broader regional coverage and a larger volume of data, enabling the models to be tested on a wider scale. Additionally, future research could consider evaluating other

machine learning algorithms to compare performance and identify the most effective approach in classifying water quality.

Suggestion

Based on the research that has been conducted, several suggestions can be considered for future work. First, it is recommended to use a broader dataset with a larger volume of data in order to test the model's capabilities on a wider scale and obtain more reliable results. This approach would enhance the robustness of the classification process and provide better generalization in real-world applications.

Second, aside from utilizing the SVM and CatBoost algorithms, future research should also explore the use of other machine learning algorithms. By comparing different models, researchers can identify the most effective approach for water quality classification, thereby improving accuracy, reliability, and the overall performance of predictive models..

REFERENCE

Adi, R. P. (2023). *Mengenal CatBoost: Algoritma Boosting yang Membuat Machine Learning Lebih Efektif*. Medium. https://medium.com/@rezapurnama1997/mengenal-catboost-algoritma-boosting-yang-membuat-machine-learning-lebih-efektif-5d679bab4966

Agung, G., Sri, D., Ningsih, A., & Pramartha, C. (2024). *Klasifikasi Kualitas Air Layak Minum menggunakan Algoritma Random Forest Classifier dan GridsearchCV*. *12*(1), 217–226.

Azmi, B. N., Hermawan, A., & Avianto, D. (2022). *Jurnal Sistem dan Teknologi Informasi Analisis Pengaruh PCA Pada Klasifikasi Kualitas Air Menggunakan Algoritma K-Nearest Neighbor dan Logistic Regression*. *7*(2). http://jurnal.unmuhjember.ac.id/index.php/JUSTINDO

Bidang Komputer Sains dan Pendidikan Informatika, P., Akademi Perekam dan Informasi Kesehatan Iris Padang Jl Gajah Mada No, D., & Barat, S. (n.d.). *Jurnal Edik Informatika Data Mining : Klasifikasi Menggunakan Algoritma C4.5 Yuli Mardi*.

CatBoost. (n.d.). https://catboost.ai/docs/en/concepts/algorithm-main-stages_cat-to-numberic

Muhamad Malik Matin, I. (2023). Hyperparameter Tuning Menggunakan GridsearchCV pada Random Forest untuk Deteksi Malware. *Multinetics*, *9*(1), 43–50. https://doi.org/10.32722/multinetics.v9i1.5578

PENGAMANAN KUALITAS AIR MINUM. (n.d.).

Pratiwi, K. S. (2020). *Support Vector Machine Classification with Python*. Medium. https://medium.com/@kurniasp/support-vector-machine-classification-with-python-64521fbd5b57

Riyantoko, P. A., Fahrudin, T. M., Hindrayani, K. M., Data, S., & Timur, J. (n.d.). Analisis Sederhana Pada Kualitas Air Minum Berdasarkan Akurasi Model Klasifikasi Dengan Menggunakan Lucifer Machine Learning. *Seminar Nasional Sains Data*, *2021*.

Simbolon, I. N. (2024). PREDIKSI KUALITAS AIR SUNGAI DI JAKARTA MENGGUNAKAN KNN YANG DIOPTIMALISASI DENGAN PSO. *Jurnal Informatika Dan Teknik Elektro Terapan*, *12*(2). https://doi.org/10.23960/jitet.v12i2.4191

UDARA KOTA PALEMBANG Oleh : NURCHAERANI KADIR. (2024).

Wijiyanto, W., Pradana, A. I., Sopingi, S., & Atina, V. (2024). Teknik K-Fold Cross Validation untuk Mengevaluasi Kinerja Mahasiswa. *Jurnal Algoritma*, *21*(1). https://doi.org/10.33364/algoritma/v.21-1.1618