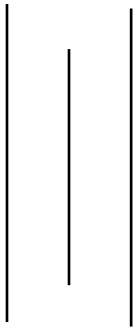
BIRATNAGAR INTERNATIONAL COLLEGE | ing | UNIVERSITY OF WOLVERHAMPTON

Report For Classification

Water Potability Prediction Using Machine Learning



Module: Concepts and Technologies of AI (5CS037)
Assessment: Final Portfolio – Classification Task
Student Name: Bhawanath Sapkota
Student ID: 2513314
Module Leader: Siman Giri
Tutor: Ayush Regmi

BIRATNAGAR INTERNATIONAL COLLEGE | ing | UNIVERSITY OF WOLVERHAMPTON

Table of contents

Abstract

The aim of this project is to find out whether the water is drinkable or not by using machine learning classification models. The Water Potability Data Set, which has 3,200 instances and nine numerical attributes, was used as the basis for the data set used in this study. This data is important in fulfilling the Sustainable Development Goals of the United Nations in achieving Goal 6, which is Clean Water and Sanitation (United Nations, n.d.).

In this study, EDA and data preprocessing were done. The multi-layer perceptron model is also employed to create a neural network model. Two machine learning classification models, Random Forest and Logistic Regression, are created for this project. The metrics used to evaluate the performance of machine learning models are F1 score, accuracy, precision, and recall.

The results of the machine learning models in this project show that the Random Forest model outperforms the other machine learning models, while the Logistic Regression model is more sensitive in identifying drinkable water. From the results, it can be asserted that, as evidenced in this research, machine learning classification methods can be employed to predict drinkable water.

List of Figures

1. Introduction

1.1 Problem Statement

However, the issue of access to drinking water is currently the major challenge facing the world because, through contaminated water, one can develop health problems or waterborne diseases (WHO, 2023). Thus, it is significant to assess the quality of water, ensuring safe health among individuals. However, it is tedious and not necessarily feasible and effective to assess the quality of water within a laboratory environment.

To ensure timely decision-making on the safety of the water for human consumption, the main objective of this study was to use machine learning and assess the quality of water as potable or non-potable based on the assessment of its physicochemical properties.

1.2 Dataset

The dataset used, namely "Water Potability," is obtained from a repository of data. The various numerical characteristics of the data, such as pH, hardness, particles, chloramines, sulphate, conductivity, organic carbon, trihalomethanes, turbidity, etc., have a significant influence on the determination of the potability of the water. (Kaggle, n.d.). The feature of interest of the dataset is the potability of the water, i.e., the consumability of the water. The dataset is used for the attainment of the United Nations' Sustainable Development Goal 6: Clean Water and Sanitation, as it improves the accessibility of drinking water by making predictions on the water testing system. (United Nations, n.d.).

1.3 Objective

The objectives of this project are:

- To explore and understand the given water quality data using EDA
- To develop classification models for predicting potability
- To compare neural network models and classical machine learning models
- To determine the best model using performance metrics

2. Methodology

This section describes the systematic steps followed to prepare the data, explore its characteristics, build classification models, and evaluate their performance. Each stage was carefully designed to ensure reliable and reproducible results.

2.1 Data Preprocessing

Data preparation was also carried out to ensure that the quality of the data was maintained. Ensuring that there were no missing values was also part of the data pretreatment. The characteristics, such as pH, sulfate, and trihalomethanes, had missing values. The missing values were handled by the use of median imputation. Median imputation is a technique used to handle missing values. It helps to reduce extreme values. The scaling of the feature was also handled. The scaling of the feature was carried out by the use of StandardScaler, which normalized all the features . Neural networks and logistic regression are two models that require attention to the value of the feature. Equal weight was given to each feature by scaling the feature. The dataset was split into a set used for testing and a set used for training. The dataset was split by the use of a stratified sampling technique. The splitting of the dataset was carried out to ensure that the original distribution was maintained. Class weighting was used to balance the classes. Class weighting was used to ensure that there was no bias to a particular class. Class weighting was used to ensure that the model was able to identify potable water samples.

2.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was carried out in this regard. EDA helped to understand the dataset before modeling. EDA plays an important role in identifying class imbalance, feature distributions, correlations, and outliers, all of which influence preprocessing and modeling decisions.

Figure 1: Class Distribution of Water Potability

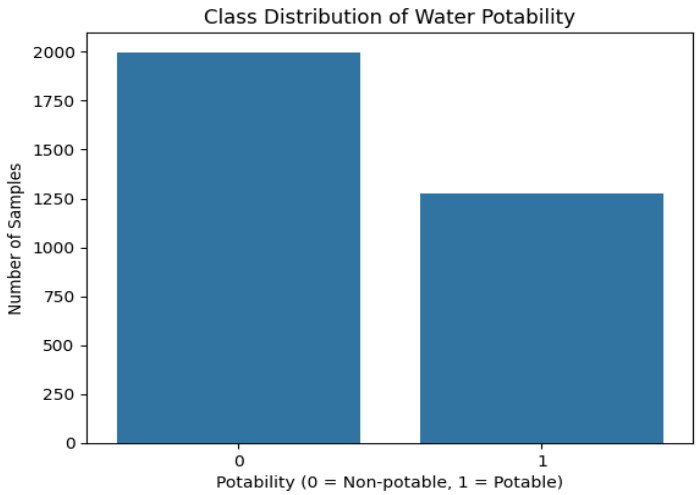


Figure 1 shows the distribution of potable and non-potable water samples. It is also important to note that this distribution of data is imbalanced, i.e., there are more non-potable water samples compared to potable water samples.

Figure 2: Distribution of Water Quality Features

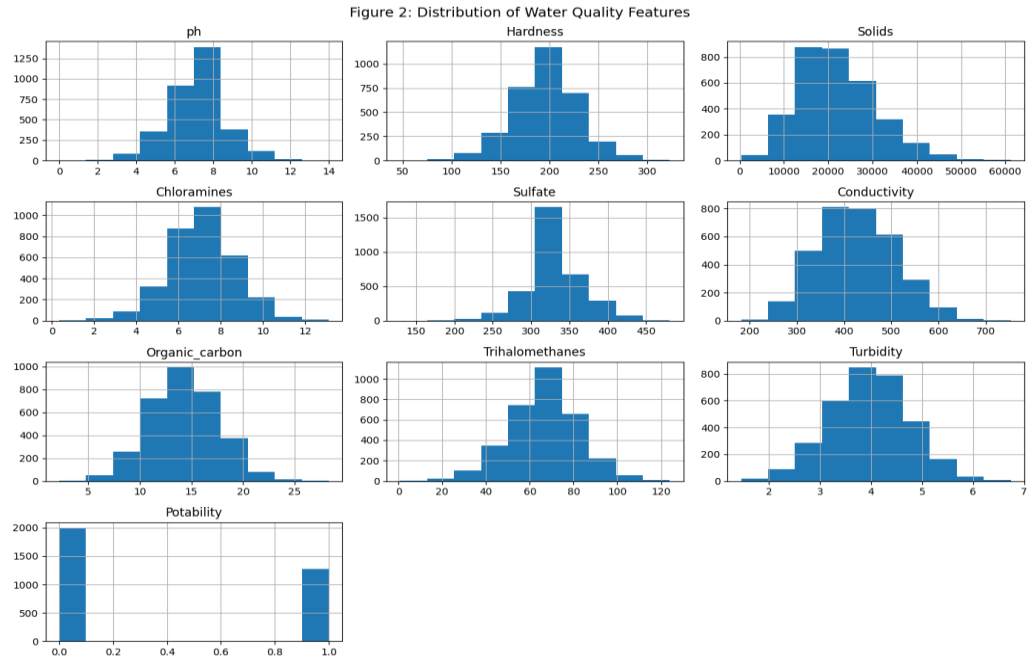


Figure 2 shows histograms of different water quality features. It can be noted that there are several attributes that are skewed and not normally distributed. This justifies the need for median imputation and feature scaling.

Figure 3: Correlation Heatmap of Features

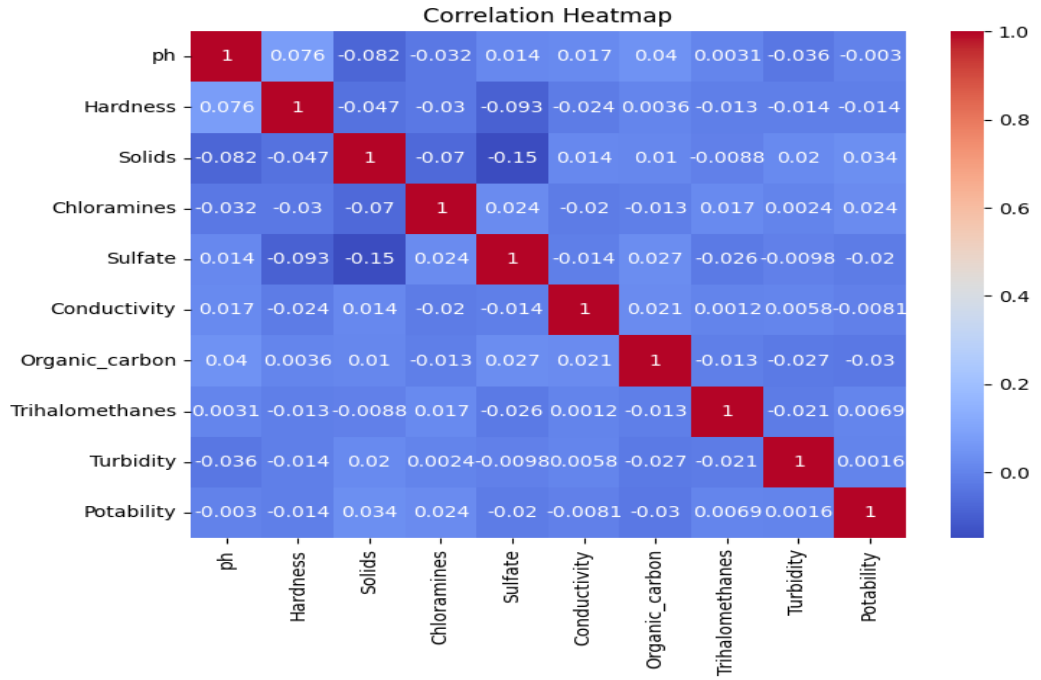


Figure 3 presents the correlation heatmap among water quality features. Most features show weak to moderate correlations, indicating limited multicollinearity. This also supports the usage of an ensemble-based model like Random Forest, which works effectively for weakly correlated features as well as non-linear relationships.

Figure 4: Boxplots of Water Quality Features

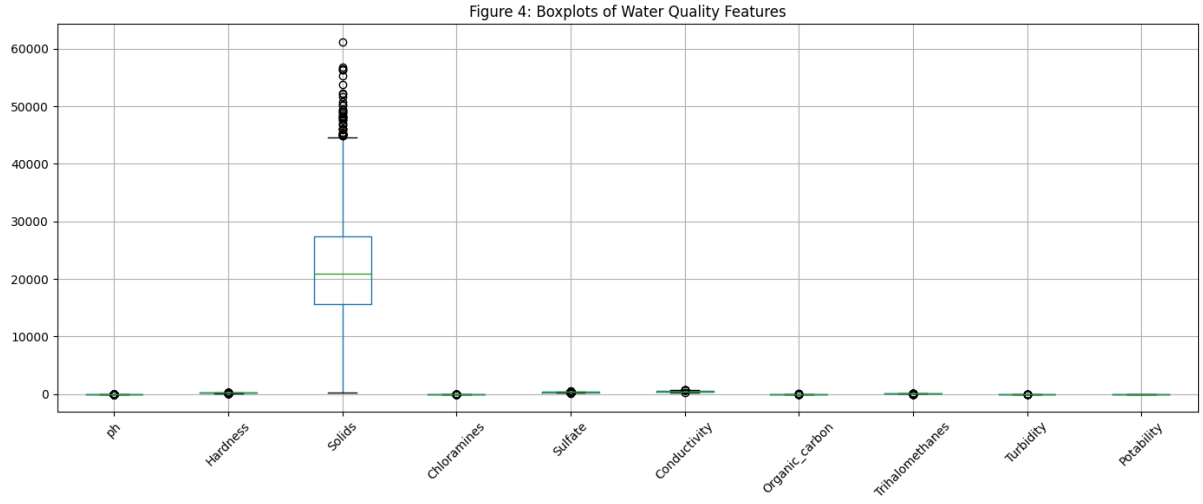


Figure 4 highlights the presence of outliers in several features. These values likely represent natural environmental variations rather than data errors. As a result, robust preprocessing techniques were applied instead of removing the outliers to preserve important information..

2.3 Model Building

After preprocessing and data exploration, multiple classification models were developed to predict water potability.

2.3.1 Neural Network Model (MLP)

Multi-Layer Perceptron (MLP) has been implemented to satisfy the requirement of a neural network model. It has been implemented with two hidden layers consisting of 64 and 32 neurons, utilizing the ReLU activation function. Finally, a sigmoid function has been used as the output layer for a binary classification problem. Binary cross-entropy has been used as the loss function. The Adam optimizer has been used in this regard for efficient optimization of the model weights. (TensorFlow, n.d.)

2.3.2 Classical Machine Learning Models

To facilitate a comparative analysis, two classical machine learning models have been designed and tested.

Model 1: Logistic Regression

The rationale for choosing the Logistic Regression algorithm is based on the fact that it is easier to implement, requires low computational cost, and is easier to interpret the results (Pedregosa et al., 2011). Considering the fact that the data set had an unequal number of potable and non-potable water samples, the idea of class weighing was incorporated in the experiment to properly weigh the minority class.

Model 2: Random Forest

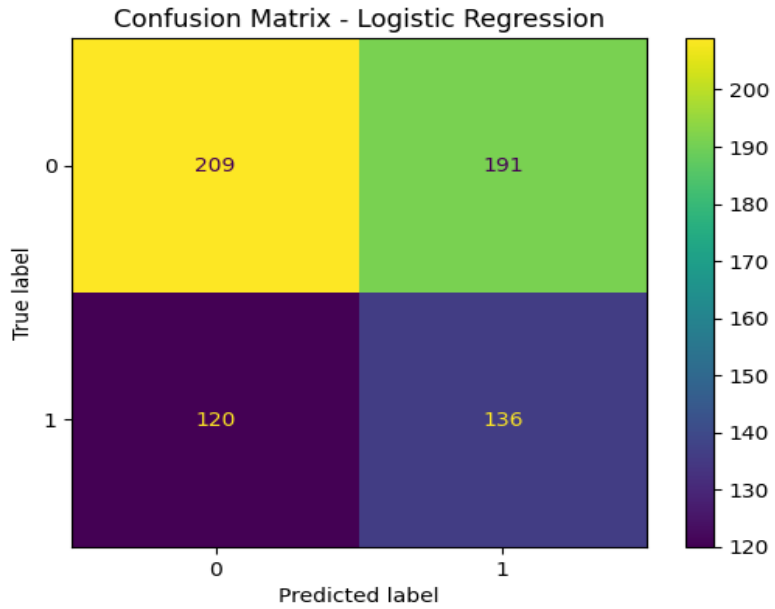
Random Forest is an algorithm that comes under the category of Ensemble Learning, which is a part of Machine Learning, where the results from all the decision trees are combined to predict the target result (Pedregosa et al., 2011). This algorithm is specifically designed to process numerical data, which is available in a structured manner, and can also be used to process non-linear relationships, which are available in the data set. Moreover, the importance of each of the water quality parameters can be measured using this algorithm.

2.4 Model Evaluation

Confusion matrices along with various classification metrics such as accuracy, precision, recall, and F1 scores have been used in this regard in order to

evaluate the models. This is because these metrics provide an efficient way of comparing and contrasting the models.

Figure 5: Confusion Matrix – Logistic Regression



As can be noted from the graph in Figure 5, it can be observed that the Logistic Regression model has performed better in terms of recall, thus implying that the model is better at detecting potable water samples. However, it can be noted that an increase in the sensitivity of the model towards non-potable water samples results in an increase in the number of misclassifications of non-potable water samples.

Figure 6: Confusion Matrix – Random Forest

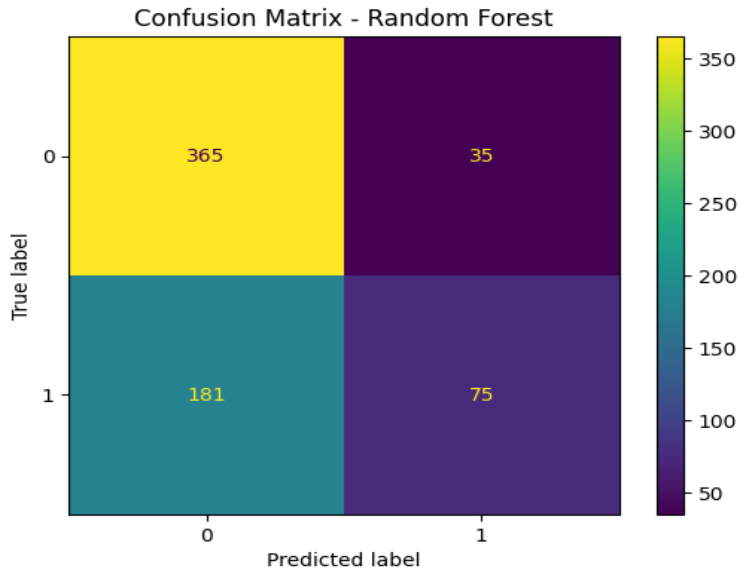


Figure 6 demonstrates that the Random Forest model correctly classifies a greater proportion of samples overall. This results in improved accuracy and F1-score when compared to Logistic Regression, although the recall value is slightly lower.

2.5 Hyperparameter Optimization

In an attempt to improve the performance of the models, it should be noted that GridSearchCV has been employed with 5-fold cross-validation. The complexity of the Logistic Regression model can be controlled by taking into consideration the regularization parameter. The parameters of the Random Forest model have been tuned in an attempt to improve the performance of the model. The F1 score metric has been employed in order to strike a balance between precision and recall of the models, particularly when there is an imbalance in the dataset.

2.6 Feature Selection

The purpose of feature selection is to make the models simpler while retaining high predictive accuracy.

Figure 7: Feature Importance Using Random Forest

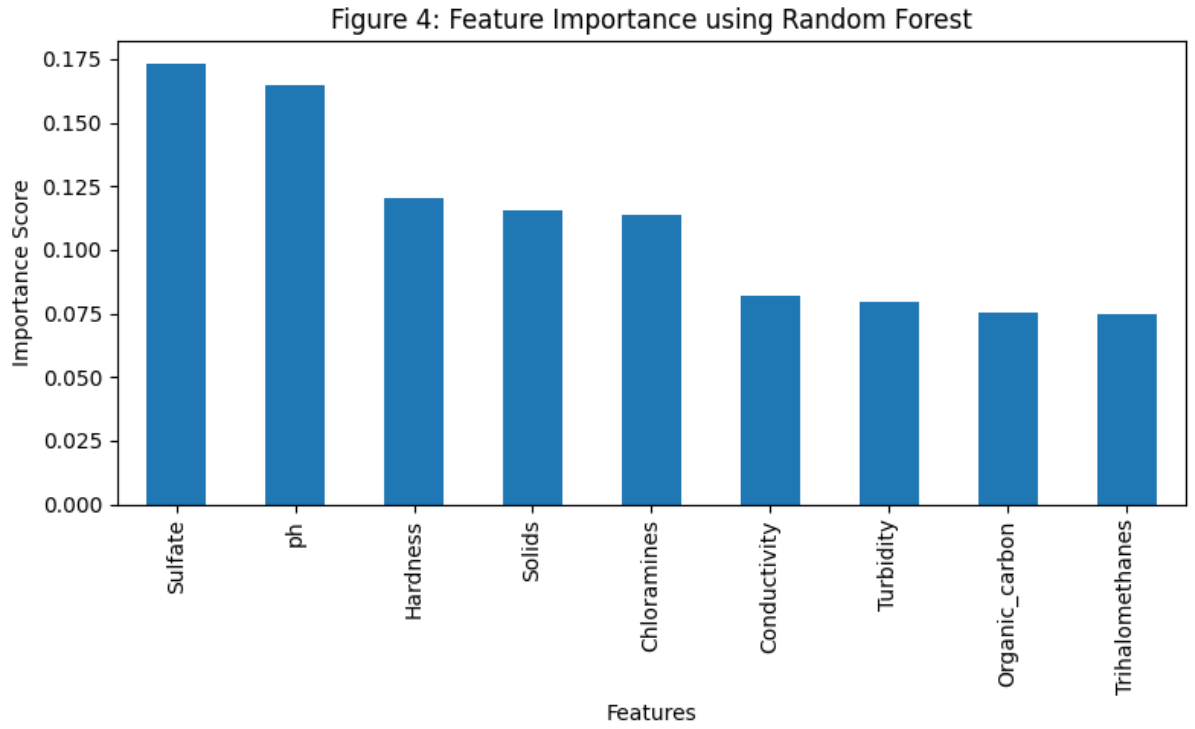


Figure 7 below illustrates the feature importance scores obtained using the Random Forest model. From the figure below, it can be seen that the features with the highest scores are the features that have the greatest impact on the potability of the water.

3. Results and Conclusion

This section reports the final performance outcomes of the classification models and outlines the main findings obtained from the experimental analysis. The models are given the test data and a consistent set of classification metrics to ensure a fair comparison.

3.1 Final Model Comparison

Both models were trained using optimized hyperparameters along with a refined set of selected features. Their performance on the test dataset is summarized in Table 4.

Table 4: Comparison of Final Classification Models

	Model	Algorithm	Accuracy	Precision	Recall	F1-Score
0	Model 1	Logistic Regression	0.541159	0.431193	0.550781	0.483705
1	Model 2	Random Forest	0.669207	0.601036	0.453125	0.516704

The comparison indicates that Random Forest achieved higher accuracy and F1-score, whereas Logistic Regression recorded a higher recall value. However, the Logistic Regression model performed better than the Random Forest model in terms of recall.

3.2 Key Findings

The results obtained from the experiment show the trade-off between the sensitivity of the two models and their accuracy. The Logistic Regression model performed better than the Random Forest model in terms of recall; therefore, it was able to classify potable water samples well. However, this was at the expense of the accuracy of the model.

In contrast, Random Forest achieved higher accuracy and F1-score, demonstrating stronger overall classification performance. Its ensemble-based structure allowed it to capture complex relationships within the data, leading to more balanced predictions across classes.

3.3 Final Model

On the basis of evaluation criteria and cross-validation scores, Random Forest was chosen as the final model. The higher accuracy and F1-score of Random Forest signify improved generalization capabilities and robustness when deployed on structured water quality data. Though Logistic Regression had high recall values, Random Forest offered a better trade-off between precision and recall, thus being more appropriate for this task.

3.4 Challenges

There were a number of challenges faced during the process of this project. For instance, there was a challenge of class imbalance, which necessitated the choice of evaluation metrics as well as class weighting. Moreover, the small size of the dataset limited the capacity of the model to learn more complex patterns. Another challenge that arose during the project involved the trade-off that exists between precision and recall, where an improvement in one of them caused a corresponding decrease in the other.

3.5 Future Work

Future work could focus on improving model performance by applying advanced ensemble techniques such as gradient boosting algorithms. Additional feature engineering and the use of larger and more diverse datasets could further enhance predictive accuracy. Exploring real-time data collection and deployment of the model in practical water monitoring systems would also be valuable extensions of this work.

4. Discussion

This section will discuss the evaluation of the models that were developed, as well as the limitations of the research. It will be a discussion of the experimental results in relation to the dataset as well as the models used.

4.1 Model Performance

The results obtained from the experiment show that the generalization capability of the model is good for the Random Forest model compared to the Logistic Regression model. This can be identified from the accuracy and F1-score values obtained from the test data set. This is because the Random Forest model can effectively handle complex and non-linear relationships between the water quality feature variables. This can be effectively applied to structured numerical data.

The Logistic Regression model is a simple model. It can be used as a baseline model. The results show that the model can effectively identify the potable water samples. This can be identified from the good recall values obtained from the model.

4.2 Impact of Hyperparameter Tuning and Feature Selection

In particular, the role of hyperparameter tuning was critical in the improvement of the stability and performance of the models. This was done by choosing the most appropriate parameters through cross-validation. As a result, the models became more consistent and reduced the chances of overfitting.

The feature selection process aided in the improvement of the interpretability of the models. This was achieved by selecting the most impactful water quality parameters. It was noted that the reduction of the features did not impact the models significantly. This proved that a smaller set of features was enough for the representation of the data.

4.3 Interpretation of Results

The results obtained are in line with the expected performance for structured numerical data. Ensemble-based models like the Random Forest are expected to perform better than linear models when the relationship between the target variable and the features is not linear. The results obtained also show the significance of employing multiple evaluation metrics for assessing the performance of a model.

4.4 Limitations

Although the results obtained were satisfactory, there are still some limitations. The dataset obtained is relatively small. This could hinder the performance of the model in learning more complex relationships. Class imbalance posed an additional challenge and required careful metric selection and weighting strategies. Furthermore, the analysis relies solely on numerical features, and external environmental or temporal factors were not considered.

4.5 Suggestions for Future Research

Future research could explore more advanced ensemble techniques such as gradient boosting models to further improve performance. Incorporating additional data sources, performing advanced feature engineering, or combining ensemble methods with neural networks could enhance predictive accuracy. Additionally, deploying the model in real-time water quality monitoring systems would be a valuable extension of this work.

References

Kaggle (n.d.) *Water Potability Dataset*. Available at: https://www.kaggle.com/datasets/adityakadiwal/water-potability (Accessed: 30 January 2026).

World Health Organization (WHO) (2023) *Drinking-water*. Available at: https://www.who.int/news-room/fact-sheets/detail/drinking-water (Accessed: 30 January 2026).

United Nations (n.d.) *Sustainable Development Goal 6: Clean Water and Sanitation*. Available at: https://sdgs.un.org/goals/goal6 (Accessed: 30 January 2026).

Pedregosa, F. et al. (2011) 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, 12, pp. 2825–2830. Available at: https://jmlr.org/papers/v12/pedregosa11a.html (Accessed: 30 January 2026).

scikit-learn (n.d.) *scikit-learn documentation*. Available at: https://scikit-learn.org/stable/ (Accessed: 30 January 2026).

McKinney, W. (2010) 'Data Structures for Statistical Computing in Python', *Proceedings of the 9th Python in Science Conference*, pp. 56–61. Available at: https://pandas.pydata.org/docs/ (Accessed: 30 January 2026).

TensorFlow (n.d.) *TensorFlow documentation*. Available at: https://www.tensorflow.org/learn (Accessed: 30 January 2026).