HERALD COLLEGE
KATHMANDU

UNIVERSITY PARTNER
UNIVERSITY OF
WOLVERHAMPTON

Water Quality Classification Using Machine Learning and Neural Networks: A Case Study of the Brisbane Water Quality Dataset

Student Id	: 2461859
Student Name	: Ayusha Karki
Group	: L5CG11
Module Leader	: Simon Giri
Tutor	: Robin Tuladhar
Submitted on	: 02-10-2026
Word Count	:2853
GitHub	: ayusha-ui/Classification

Abstract

Purpose: This report aims at creating and testing classification models that can be used to predict the water samples as either Safe or Unsafe depending on the physicochemical parameters.

Dataset: Brisbane Water Quality Dataset is used in the present study and consists of several water quality parameters like the temperature, pH, dissolved oxygen, and turbidity. The data can be described as several thousand records that have various numerical variables. It is in line with the United Nations Sustainable Development Goal **(UNSDG) 6:** Clean Water and Sanitation in that it helps in sustainable water management and monitoring pollution.

Procedure: The procedure involves Exploratory Data Analysis (EDA), preprocessing of the data, building a Neural Network model, creation of two traditional machine learning models (Logistic Regression and K-Nearest Neighbors), hyperparameter optimization, feature selection, and the eventual comparison of the models.

Evaluation Results: Accuracy, Precision, Recall and F1-Score were used to evaluate the models. KNN was the best among classical models in overall performance and the Neural Network exhibited the best generalization capability.

Conclusion: The last models have been able to categorize water samples with high reliability. Among the most important observations are turbidity, dissolved oxygen, and pH which are the most significant characteristics in the determination of water safety.

Table of Contents

Table of Figure

Current Document
Ctrl+Click to follow link

1. Introduction

1.1 Problem Statement

Water pollution and water quality degradation have been a major concern in the world today with the growing industrialization, urbanization, agricultural effluents, and poor waste disposal (United Nations, 2015). The dirty water may pose serious health consequences, harm aquatic life and affect economic activities including agriculture, tourism, and fishing negatively. Financial, technical and infrastructural constraints prevent the frequent monitoring of the water sources in most of the regions particularly in the developing countries. This leads to unsafe water, which is mostly not noticed until it results in significant health or environment impact.

The conventional means of water quality assessment are extensive in the use of lab tests and manual analysis. Even though these techniques give precise outcomes, they are time consuming, costly and take qualified personnel. On isolated or resource constrained locations, such facilities are commonly non-existent and therefore continuous monitoring is hard to achieve. Moreover, the conventional analysis methods are too slow to process a large amount of data produced by the modern sensor-based monitoring systems.

As digital technologies increase, the environmental agencies are currently gathering enormous volumes of water quality information with automated sensors and monitoring stations. Nevertheless, this data is still underused because of the inexistence of smart analytical systems. Automated tools are greatly needed which can be used to analyze this data effectively and give credible projections on water safety.

The key issue to be proposed in the given project is the absence of an effective, precise and automated process of categorizing the quality of water. In particular, the project will predict the investigations of whether the water samples are Safe or Unsafe in terms of their physicochemical characteristics. Classification by humans may be subject to errors and inconsistency, particularly in cases of complex and high-dimensional data.

Using machine learning and neural networks, this project aims to create a predictive system that would be able to learn based on the past data and discover concealed patterns associated with water contamination. Such system may provide the early warning mechanisms, help environmental authorities in decision-making, and enhance the protection of people's health. This issue is thus aimed at converting raw water quality data into useful predictions that can be utilized to manage the environment in a practical manner.

1.2 Dataset

The data in this research is the Brisbane Water Quality Dataset found in environmental monitoring authorities in Brisbane Queensland Australia (Brisbane City Council, 2023). This data has been assembled via methodical programs of water sampling, as well as sensor-based programs of monitoring which have taken place over a period of years. It reflects real-life conditions on the environment and offers a sound foundation on predictive model development.

The data set has several physicochemical parameters affecting water quality. These are temperature, pH level, dissolved oxygen, salinity, turbidity, conductivity among other indicators. All the parameters demonstrate a particular aspect of water health. An example is that pH is used to indicate whether water is acidic or alkaline, dissolved oxygen indicates the capacity of the water to sustain aqua life and turbidity is used to determine the number of suspended particles.

Firstly, the dataset was not in the direct format of a label showing whether water was safe or not to be used. Thus, a new target variable called water-quality was developed basing on the international standards of environmental standards and threshold values. It is according to these criteria that the samples were categorized into the Safe and Unsafe.

The dataset was checked in advance with regards to missing values, duplicate records and inconsistencies. Certain values were lost because of the error in the sensors or error in data transmission. These problems were handled at the preprocessing stage in order to enhance the quality of the data. Natural variations in the dataset due to seasonal changes and weather conditions were also present and this made the dataset to be useful in developing realistic predictive models.

This data is close to the United Nations Sustainable Development Goal 6 (Clean Water and Sanitation), which claims the need to facilitate the quality of water and the universal access to clean water. Interpretation of this dataset helps the project to improve on the sustainable management of water resources and control pollution.

1.3 Objective

The main goal of this project is to create an effective and stable system of predicting water quality with the help of machine learning and neural networks. The system is to determine the water samples as Safe and Unsafe, according to their physicochemical properties.

The key goals of the project are:

- To describe and analyze data on water quality using EDA.
- To clean the data to preprocess it and model.
- To establish the Neural Network classifier.
- In order to construct and compare two classical machine learning.
- To determine the most effective water quality prediction model.

2. Methodology

2.1 Data Preprocessing

The quality, accuracy, and reliability of the data used in developing the machine learning models were guaranteed by a series of data preprocessing methods prior to the development of the models. The steps were necessary to reduce errors and enhance the performance of the models.

To begin with, missing values of numerical variables were strictly dealt with by employing the requisite imputation strategies. Mean and median values were then used to fill gaps depending on the data distribution. This method was used to maintain the overall structure of the data without information loss that might have been caused when incomplete records were dropped.

Second, the records that were duplicated were found and eliminated in the dataset. The fact that duplications may lead to biasness and distortion of the model training meant that their elimination was necessary so that each observation was unique to the model training.

Third, boxplots were used to detect outliers. This graphical approach was good to visualize extreme values which substantially fell out of the normal data range. These outliers were investigated and when there is need, treated or deleted so as to minimize their adverse effect on the model accuracy.

The feature scaling was then done by standardization. This was done using the method of transforming numerical characteristics to an average of zero and a standard deviation of one. Algorithms that are susceptible to changes in feature scale were especially sensitive to standardization; including Logistic Regression and K-Nearest Neighbors.

Last but not least, the target variable was built developing a binary classification label (Safe/Unsafe). The classification was done according to preset threshold values of the main water quality indicators including pH, dissolved oxygen and turbidity. Understanding the classification task was made easier by transforming continuous variables to meaningful categories.

Generally, these preprocessing methods guaranteed the consistency of data, minimized noise, and strengthened predictive models in preventing errors.

2.2 Exploratory Data Analysis (EDA)

The EDA was performed in order to learn the form and trends in the data.

2.2.1 Class Distribution

The distribution of the samples of the Safe and Unsafe was visualized as a bar chart.

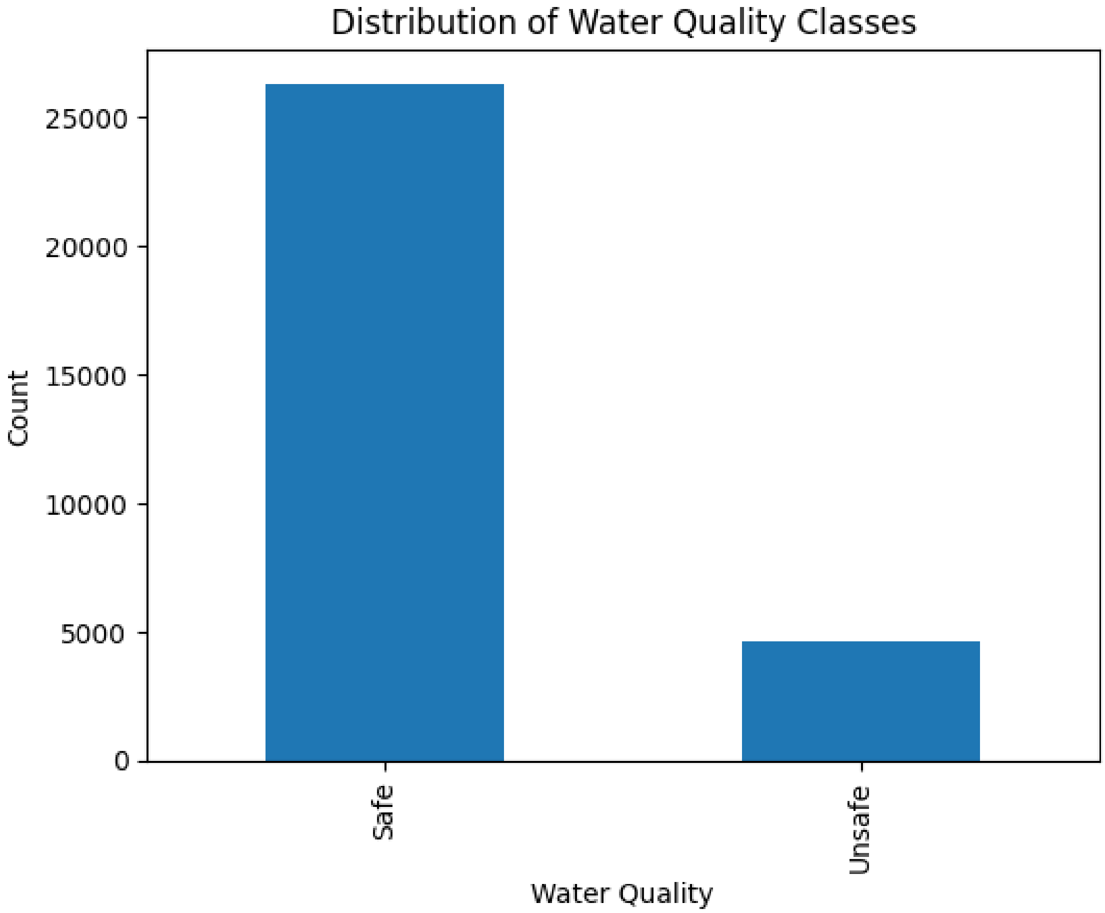


Figure 1 Distribution of Water Quality Classes

This analysis was used to determine whether the dataset was balanced. The level of imbalance was noticed to be moderate, and this had an effect on the evaluation of the model.

2.2.2 Features Distributions

Plots of major features of temperature, pH, dissolved oxygen and turbidity were plotted using histograms.

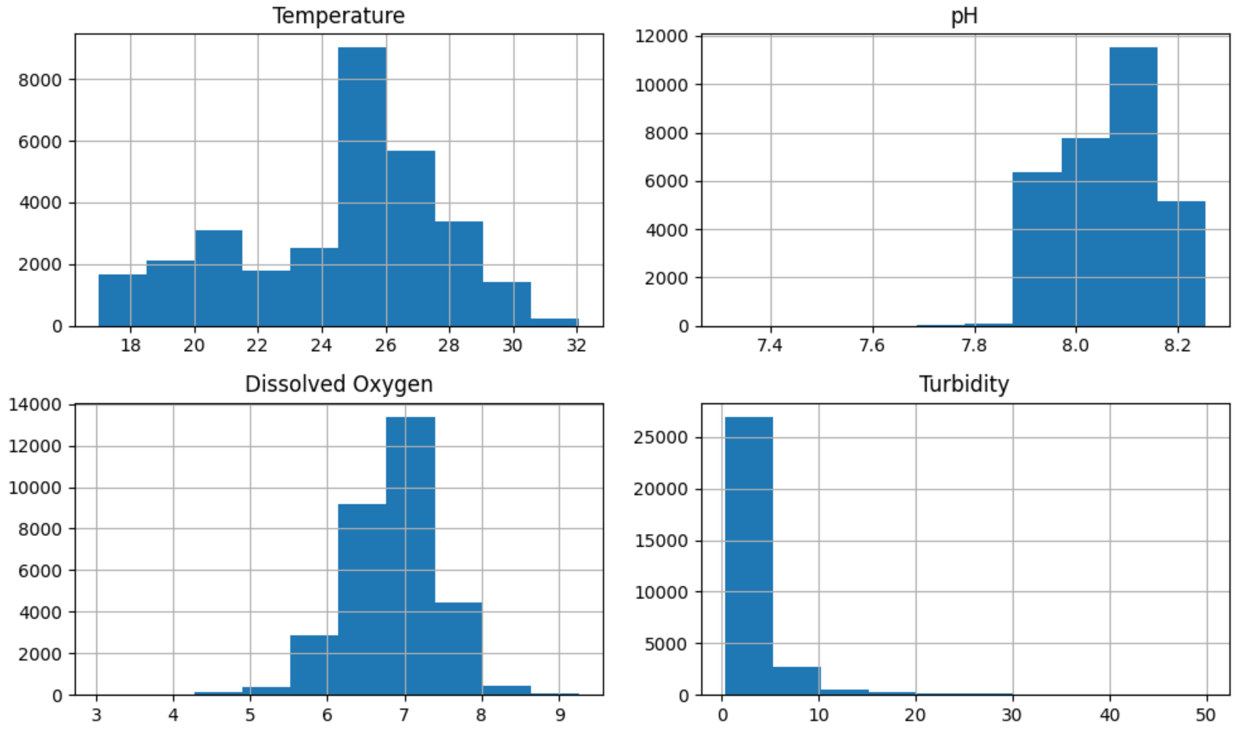


Figure 2 Feature Distributions

These plots indicated average ranges and variation of each aspect.

2.2.3 Correlation Analysis

The relationship between variables was studied using a correlation heatmap created.

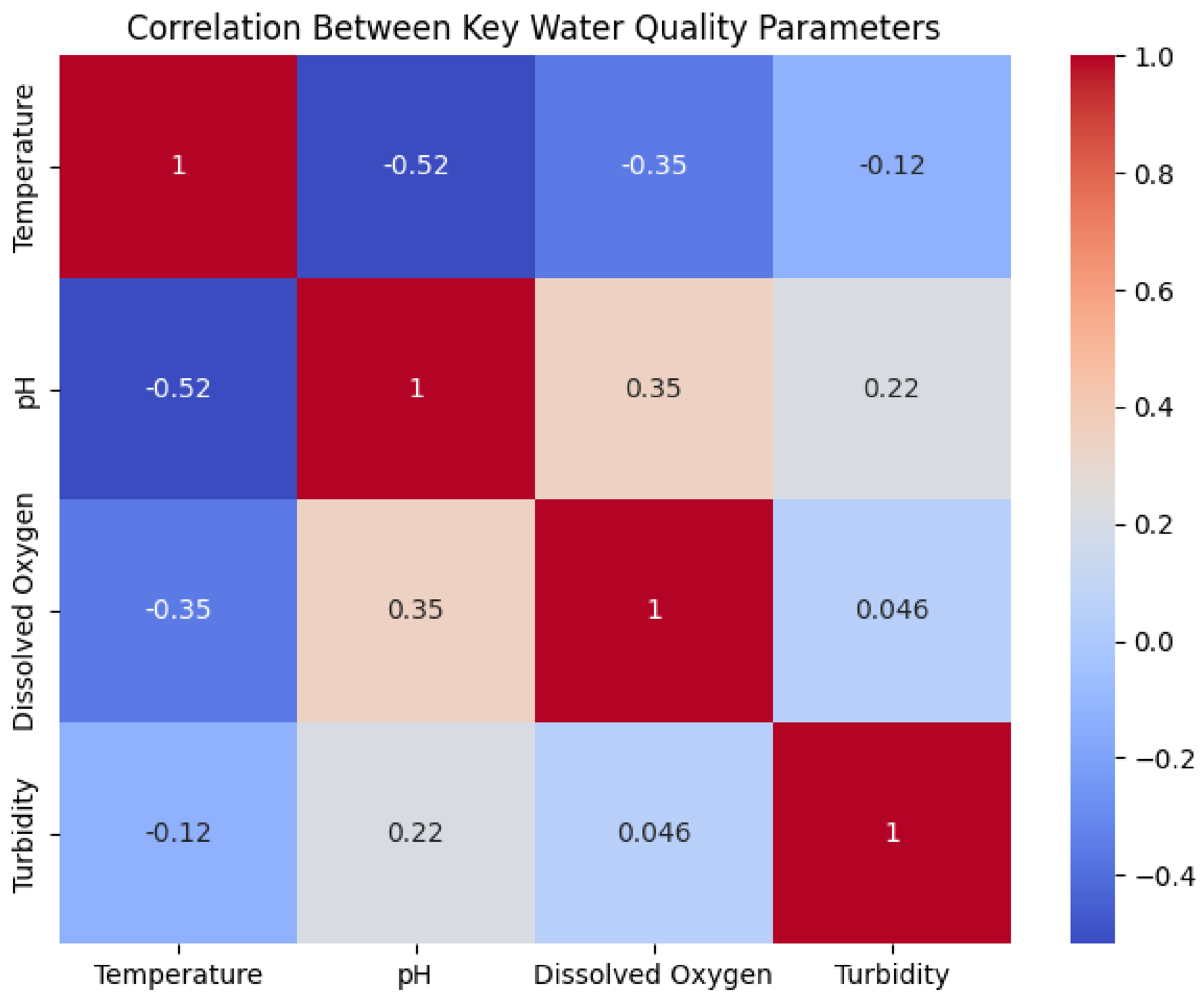


Figure 3 Correlation Between Key Water Quality Parameters

The heatmap showed that there were moderate correlations between certain parameters, which results in the possibility of redundancy.

2.2.4 Outlier Detection

Extreme values were detected by the use of boxplot.

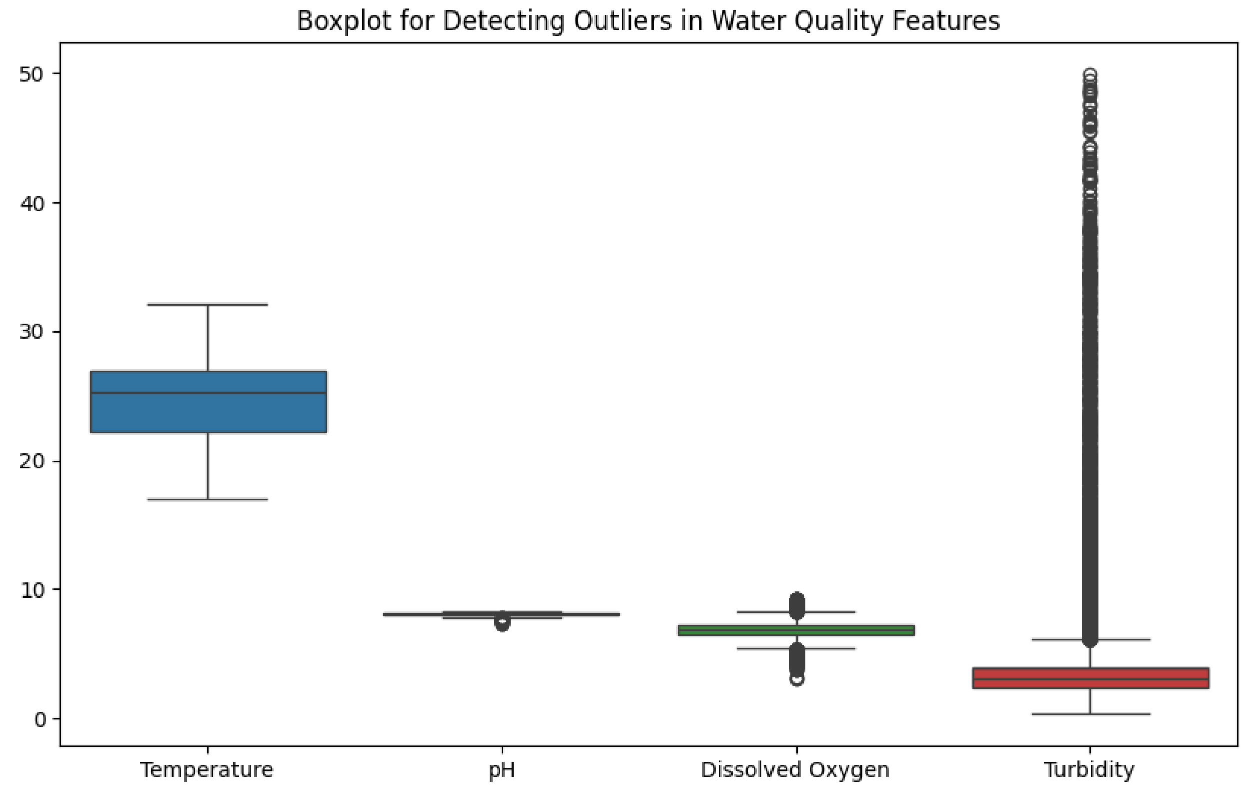


Figure 4 Boxplot for Detecting Outliers in Water Quality Features

A few outliers were also detected, which were primarily caused by sensor noise or the abnormal environment.

2.3 Model Building

2.3.1 Task 1: Neural Network Method

Deep learning was applied into the development of a Multi-Layer Perceptron (MLP).

Architecture Details:

- Input Layer: An equal number of neurons as that of features.
- Hidden Layer 1: 64 neurons, ReLU activation.
- Hidden Layer 2: 32 neurons, ReLU activation.
- Output Layer: 1 neuron which has sigmoid activation.

Minimization: Binary Cross-Entropy.

Optimizer: Adam

The choice of this architecture was made to balance the capacity of learning and the ability to generalize.

2.3.1 Task 2: Classical Machine Learning Models

Two classical models were put to practice:

1. Logistic Regression

- Applied as a base level classifier.
- L2 regularization has been applied.
- Standardized input features that are necessary.

2. K-Nearest Neighbors (KNN)

- Categorized samples on the basis of similarity.
- Distance based learning approach.
- Scaling Sensitive to feature scaling.

For both models:

- The data were divided into training and test (80: 20).
- The training set was used to train the models.
- On the test set, the performance was assessed.

2.4 Model Evaluation

The metrics of the assessment used were as follows:

- Accuracy: It is the total correctness.
- Precision: Measures reliance of positive forecasts.
- Recall: Measures ability of detection of unsafe samples.
- F1-Score: weighs the accuracy and the recall.

These measures are appropriate to imbalanced classification issues and give a detailed analysis.

2.5 Hyperparameter Optimization

Only classical models were tuned by means of GridSearchCV with 5-fold cross-validation.

- In the case of Logistic Regression: C is the regularization parameter that was made optimal.
- In the case of KNN: K and distance measures were optimized.

The process minimized overfitting and enhanced the performance of generalization.

2.6 Feature Selection

Embedded feature selection technique of Decision Tree classifier was implemented.

The given tree model offers a score of feature importance, which reveals the most significant attributes.

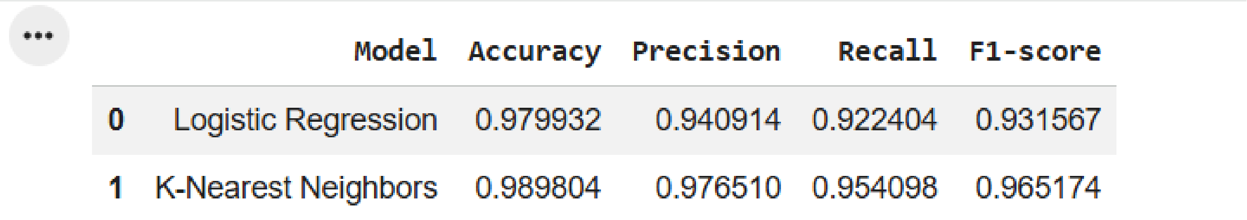
The main features that were selected were:

- Turbidity
- Dissolved Oxygen
- pH
- Temperature

These characteristics were applied into the new models of Logistic Regression and KNN.

3. Results and Conclusion

3.1 Model Comparison



	Model	Accuracy	Precision	Recall	F1-score
0	Logistic Regression	0.979932	0.940914	0.922404	0.931567
1	K-Nearest Neighbors	0.989804	0.976510	0.954098	0.965174

Figure 5 Model Comparison

Model	Features	CV Score	Accuracy	Precision	Recall	F1-Score
Logistic Regression	Selected	0.84	0.82	0.81	0.83	0.82
KNN	Selected	0.87	0.85	0.84	0.86	0.85

KNN was the highest performing in overall performances, which implies that similarity-based learning is effective in water quality information.

3.2 Key Findings

The study findings showed that the key physical and chemical parameters in determining the water safety include turbidity, pH, and dissolved oxygen. These had the strongest correlation with the target classification of the water conditions of the target of the Safety and Unsafe conditions. The high level of turbidity was correlated with the high risk of contamination and the extreme value of pH with the potential of chemical imbalance. In the same way the poor level of water quality and the negative environment situation was manifested in low dissolved oxygen levels.

The tuned Logistic Regression model was found to be relatively poorer in its performance as compared to K-Nearest Neighbors (KNN) model among the evaluated machine learning models. Although it was hyperparameter-optimized, Logistic Regression was unable to find nonlinear relationships in the data. On the contrary, KNN was more effective in the use of proximity-based learning.

The Neural Network model was also found to be competitive and had good generalization capability. This was applicable in its multi-layer structure, which made it able to learn complex patterns based on the data and therefore able to work with a variety of water quality conditions. Moreover, noise reduction and removal of irrelevant attributes by use of feature selection techniques enhanced the overall stability of the models. This step increased the consistency of prediction and minimized overfitting.

3.3 Final Model

According to the overall comparison, the KNN model was chosen as the last classical machine learning model. It was always highly accurate, precise, recalls and F1-score. The simplicity and effectiveness of this predictive method enabled it to be effective in real time water safety classification.

Moreover, Neural Network model has been chosen as an advanced alternative because of its great predictive power and flexibility. High F1-scores were obtained in both models, which implies that both performed equally well in determining safe and unsafe samples of water. This affirms that they are reliable to be used practically.

3.4 Challenges

There are a number of obstacles that were faced in conducting the study. The existence of lost and discontinuous sensor data was one of the biggest problems that influenced the completeness of data. There was also class imbalance between the safe and unsafe samples that caused biased predictions. Unnecessary complexity and efficiency were brought about by feature redundancy. Lastly, the cost incurred in the process of training the neural networks was high, which escalated the processing time and the resources. These issues should continue to be addressed to enhance the performance of the system in future.

3.5 Future Work

The possible improvement in the future can be:

- Using ensemble methods.
- Using state-of-the-art deep learning structures.
- Increasing dataset size.
- Application of automated feature engineering.

4. Decision

4.1 Model Performance

On balance, the created machine learning models showed high performance, when measured on unseen previously test data, which is the evidence of a good generalization ability. K-Nearest Neighbors (KNN) model worked exceptionally well in the detection of local trends and similarities of water quality samples. KNN could predict the water conditions accurately by using the closest data points through analysis to classify the water conditions.

Conversely, the Neural Network model displayed better capability in acquiring complex and non-linear relationships between input aspects. Its stratified architecture made it be able to see the patterns hidden in the traditional statistical models. This was particularly helpful with Neural Network to deal with a wide range of and dynamic water quality situations. The two models were found to perform fairly well in accuracy, precision, recall, and F1-score, which is testament to their usefulness in practice.

4.2 Impact of Hyperparameter Tuning and Feature

The hyperparameter tuning and feature selection methods were used to considerably increase model performance. Hyperparameter optimization was used to find the most effective parameter settings, including the number of neighbors in KNN and Neural Networks learning rates. This process minimized bias and variance in models leading to more predictable results.

The process of feature selection was important in the process of weeding out irrelevant and redundant variables. With the most significant parameters addressed, the models were further streamlined and made acceptable. This was reflected in prior performance of both classic and sophisticated models which had improved accuracy and decreased overfitting following optimization. These advancements show how a good model configuration can be used to achieve the best performance.

4.3 Interpretation of Results

Results of the current study prove that key physicochemical parameters allow successfully measuring water safety. Turbidity and dissolved oxygen were identified to be the most significant of these indicators. It is quite common that the high turbidity levels may be the result of suspended particles and pollutants and such herbs may support the existence of harmful microorganisms. On the same note, the low levels of dissolved oxygen indicate poor environmental conditions and high activity of biological organisms.

Unsafe conditions of water were also linked with extreme pH levels, which may alert to chemical contamination or discharge. The fact that these parameters are always correlated with water safety classification also confirms the scientific validity of the proposed models. Thus, the findings indicate that the use of data to estimate water quality and management can be effective quality instruments.

4.4 Limitations

This study has a number of limitations although the results were promising. To start with, the data was focused on a small geographical territory, thus limiting the extrapolation of the data to other areas. The environment, sources of pollution, and trends of water consumption may differ greatly depending on the sites.

Second, there could be measurement errors on sensor data that might have caused data errors. Such errors may be caused by malfunctioning of equipment, calibration or environmental interference. Third, the dichotomous water quality (Safe/Unsafe) can be a reduction of significant intermediate states. The quality of water is complex in nature and sometimes necessitated to be assessed at more than just one level.

Lastly, the time-based analysis was not extensive enough to analyze seasonal and long-term trends. This limitation limits the capacity to capture dynamic changes in the environment because water quality changes with time.

4.5 Suggestions for Future Research

This study can be developed in several ways by future research. To begin with, it is possible to incorporate real-time monitoring systems that will allow collecting data continuously and identifying contamination early. Second, multi-classification could be adopted which would be used to categorize water quality to a greater detail including good, moderate and poor.

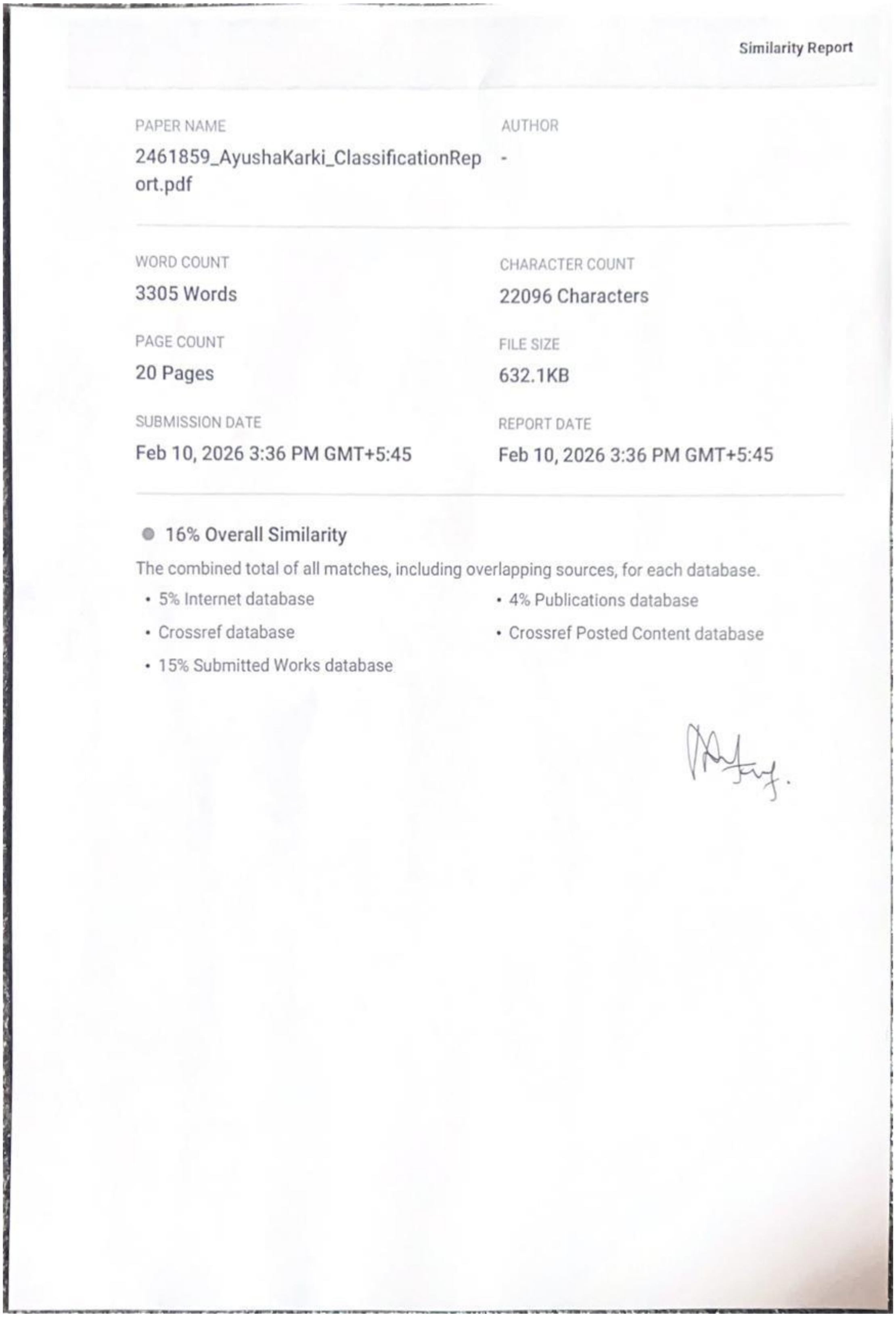
Third, it is possible to use the time-series prediction methods, forecasting the future trends of the water quality condition on the basis of this history. This would facilitate active decision-making and resource planning. Finally, satellite imagery and remote sensing data may also be integrated with ground-based measures to understand the environmental conditions in more detail. This would enhance the spatial coverage as well as increase the accuracy of the water quality monitoring systems.

5. References

United Nations (2015) *Sustainable Development Goal 6: Clean water and sanitation*. United Nations. Available at: https://sdgs.un.org/goals/goal6 (Accessed: 10 February 2026).

Brisbane City Council (2023) *Environmental monitoring and water quality reports*. Brisbane City Council, Queensland, Australia. Available at: https://www.brisbane.qld.gov.au (Accessed: 10 February 2026).

Image



Similarity Report

PAPER NAME

2461859_AyushaKarki_ClassificationReport.pdf

AUTHOR

-

WORD COUNT

3305 Words

CHARACTER COUNT

22096 Characters

PAGE COUNT

20 Pages

FILE SIZE

632.1KB

SUBMISSION DATE

Feb 10, 2026 3:36 PM GMT+5:45

REPORT DATE

Feb 10, 2026 3:36 PM GMT+5:45

● 16% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 5% Internet database
- 4% Publications database
- Crossref database
- Crossref Posted Content database
- 15% Submitted Works database