UNIVERSITY OF WOLVERHAMPTON

HERALD COLLEGE KATHMANDU

Academic Year	Module	Assessment Number	Assessment Type
2024	5CS037 Concepts and Technologies of AI	Final	Report

Title: Report on Classification Task

Student Id : 2408624
Student Name : Prayas Shrestha
Section : L5CG15
Module Leader : Mr. Siman Giri
Tutor : Mr. Siman Giri

Abstract

This report investigates the use of classification techniques to predict safe drinking water sources.

This analysis uses the "Water Potability" dataset and this dataset includes physicochemical attributes of water samples. The workflow includes data preprocessing, thorough exploratory data analysis (EDA), Logistic Regression and Random Forest model building and hyperparameter tuning and feature selection.

Model evaluation used accuracy and precision and it also used recall and the F1-score. Logistic Regression classifiers were less accurate than Random Forest classifiers.

The study shows that classification models reliably and accurately predict water potability. More advanced feature selection and additional tuning can further improve results.

1. Introduction

1.1 Problem Statement

This project will classify each water sample. The classification will be based on the sample's physicochemical properties, determining if it is potable or not.

1.2 Dataset

The dataset used is the "Water Potability" dataset. It was obtained from an online repository. It includes pH, hardness and solids. These are among other water quality indicators. This dataset supports the UNSDG goal of Clean Water and Sanitation by clarifying several aspects of water quality.

1.3 Objective

The goal is to construct a classification model that precisely predicts water potability based on its several characteristics.

2. Methodology

2.1 Data Preprocessing

Median imputation handled all missing values.
Clipping some outliers reduced skewness.
Normalization standardized the features. The features' disparate scales were affected by this process.

2.2 Exploratory Data Analysis (EDA)

Descriptive statistics were performed in several ways for an understanding of data distribution. Feature distributions were visualized by means of boxplots and histograms. Correlation analysis served the purpose of unmasking the supersets of relationships among them.

2.3 Model Building

The two-classification models that were implemented are:
Logistic regression (baseline model)
Random forest classifier (more complex model)

Data were split into 80% for training and 20% for testing for the purposes of model training and evaluation.

2.4 Model Evaluation

Modeling methods:
Accuracy: Defines the overall correctness of predictions.
Precision: Represents the correct percentage of classifying positives.
Recall: Measures of the capacity to identify true positives.
F1 Score: The harmonic means of precision and recall.

Random forests outperformed logistic regression in all metrics.

2.5 Hyperparameter Tuning

GridSearchCV along with RandomizedSearchCV were used for the hyperparameter optimization of both models. The major improvement in accuracy was attributed to the working of the optimal hyperparameters.

2.6 Feature Selection

Recursive Feature Elimination (RFE) was the technique used for selection of the necessary features.

3. Conclusion

3.1 Summary of the Key Findings

The dataset was preprocessed because of the presence of missing values and outliers.

Random Forest proved to be better than Logistic Regression.

Hyperparameter tuning boosted the accuracy of the model.

3.2 Final Model

The model that performed best was the Random Forest Classifier with optimized hyperparameters.

3.3 Challenges

Dealing with missing values and outliers took partially a long time.

The performance complexity balancing needed tuning.

3.4 Future Work

Consider trying XGBoost and other classification algorithms. Another area of consideration is feature engineering, which may contribute to better accuracy.

4. Discussion

4.1 Model Performance

Random Forests outperformed Logistic Regression and presented higher generalizability.

4.2 Hyperparameter Tuning and Feature Selection Effects

Hyperparameter tuning utilized improved performance, and feature selection helped reduce unnecessary complexity.

4.3 Result Interpretation

The assigned models were able to classify water potability satisfactorily as per

expectations.

4.4 Limitations

High missing values on some features may reduce reliability.
The dataset may not be representative of all source waters.

4.5 Suggestions for Future Studies

Conduct experiments to classify using deep learning mode.
Use domain knowledge to refine the feature selection process.
This report presents a structured approach to the classification of water potability and discusses possible areas of improvement in the future.