5CS037 - Concepts and Technologies of AI.

Classification Task Report

Name : Samip Regmi

Group: L5CG4

University Id: 2511198

Identification number: np02cs4a240105 Submitted To: AyushRegmi

Abstract

Access to clean and safe drinking water is a critical global challenge. This study focuses on predicting water potability using machine learning classification techniques based on physicochemical properties such as pH, hardness, and total dissolved solids. The dataset used consists of 3,276 records with 10 features and aligns with the United Nations Sustainable Development Goal (UNSDG) 6: Clean Water and Sanitation.

data preprocessing, handling of class imbalance, and feature scaling, several classification models including Logistic Regression, Random Forest, and a Multilayer Perceptron (MLP) neural network were developed and evaluated. Model performance was assessed using accuracy, precision, recall, and F1-score.

The results indicate that the neural network achieved the most balanced overall performance among the evaluated models, particularly in handling the imbalanced dataset. Logistic Regression also demonstrated stable and competitive performance, while tree based models showed sensitivity to data imbalance and variance, resulting in reduced generalization performance. Overall, the findings highlight the importance of proper preprocessing, feature selection, and model tuning in improving predictive performance for water quality classification tasks.

Keywords: Water Potability, Machine Learning, Classification, Neural Network, Logistic Regression, Water Quality, Random Forest

Table Of Contents

List of Figures

1. Introduction

1.1 Problem Statement

Access to clean and safe drinking water remains a major global challenge, affecting health and well being worldwide. Predicting water potability based on measurable physicochemical properties can help ensure water safety and support sustainable water management.

1.2 Dataset

The dataset used in analysis is Water Quality Classification which was obtained from Kaggle. It contains data for Water Potability, The dataset aligns with United Nations Sustainable Development Goals (UNSDG) by UNSDG 6 which is clean water and sanitation

1.3 Objective

The objective of this analysis is to build a predictive classification model that estimates the water potability by the features in the dataset.

PH
Hardness
Solids
Chloramines
Sulfate
Conductivity
Organic_carbon
Trihalomethanes
Turbidity
Potability

2. Methodology

2.1 Data Preprocessing

The dataset contained missing values. A threshold-based approach was applied:

- Columns with more than 10% missing values were imputed using the mean.
- Rows with less than 10% missing values were removed.
- Data was then scaled using StandardScaler to ensure uniform feature contribution.

2.2 Exploratory Data Analysis (EDA)

EDA was conducted to understand data distribution and relationships:

- Target distribution showed class imbalance which was solved using Synthetic Minority Over Sampling Technique (SMOTE).
- Correlation analysis indicated weak linear relationships between features.
- Visualization helped justify the use of non-linear models.

2.3 Model Building

Neural Network Model

- Model: Multi Layer Perceptron (MLP classifier)
    - Architecture: Configured with hidden layer using relu optimization and adam solver
    - Performance: Achieved an accuracy of 0.65

Classical Machine Learning Models

- Logistic Regression: A linear model used as a baseline
    - Random Forest: An ensemble method capable of capturing non linear relationships.

2.4 Model Evaluation

Models were evaluated using:

- Accuracy
- Precision
- Recall
- F1-Score

These metrics were selected due to class imbalance in the dataset.

2.5 Hyperparameter Optimization and Feature Selection

Optimization: Techniques like GridSearchCV and RandomizedSearchCV were applied to find optimal parameters for Logistic Regression and RandomForest.

2.6 Feature Selection

Methods like RFE (Recursive Feature Elimination) were used to select most relevant features

Solids
Chloramines
Sulfate
Organic_carbon

3. Data Visualization

- Distribution of the target variable portability

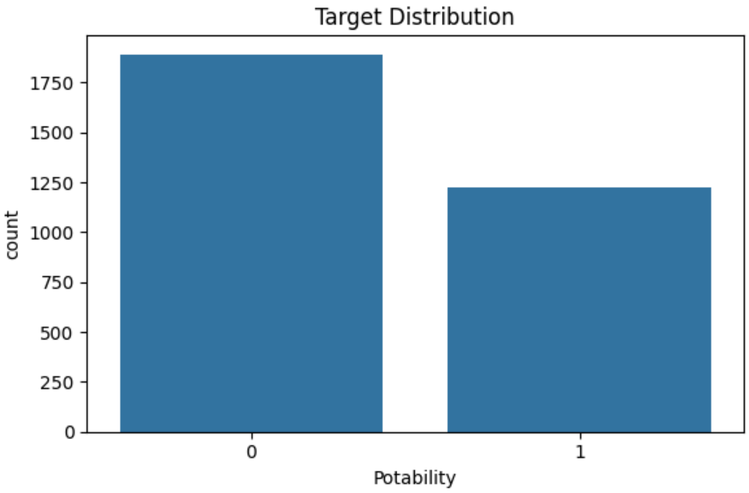

Figure 1: Distribution of the target variable Potability (0 = Non-potable, 1 = Potable). The plot shows class imbalance in the dataset.

- Correlation matrix showing pairwise relationship among features in dataset, we can see there is no any strong linear relationship among features indicating that there is non linear relationship among the features

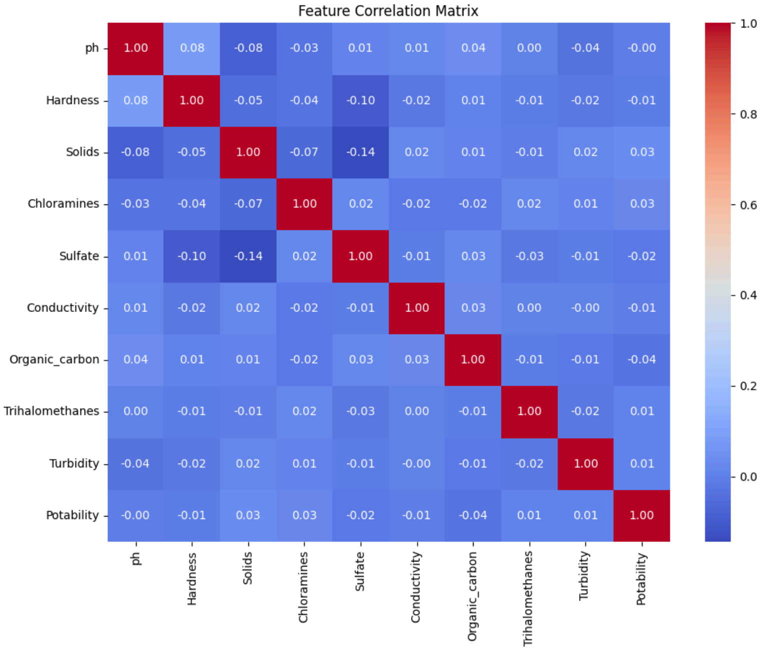

Figure 2: Correlation matrix showing the pairwise relationships among features in the dataset. Positive correlations are shown in warm colors, while negative correlations are shown in cool colors.

- Distribution of key chemical features

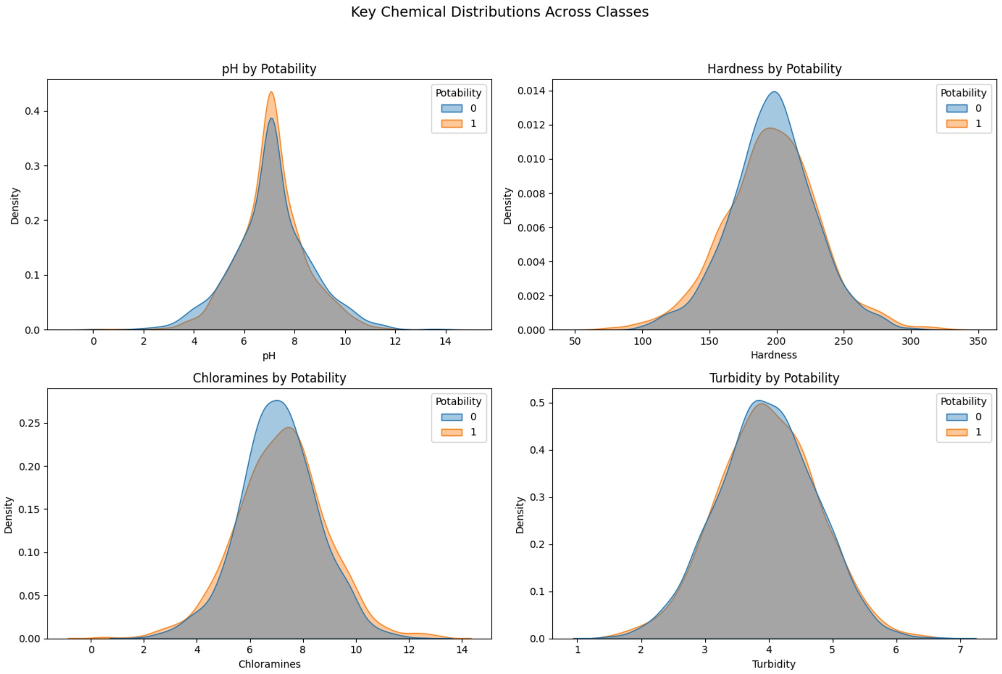

Figure 3: plots illustrating the distribution of key chemical features (pH, Hardness, Chloramines, and Turbidity) across potable and non-potable water samples.

- Confusion Matrix Of Logistic Regression

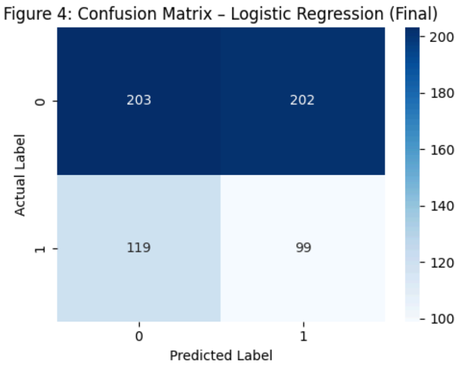

Figure 4: Confusion matrix illustrating the classification performance of the Logistic Regression model on the test dataset. The matrix shows the number of true positives, true negatives, false positives, and false negatives.

- Confusion Matrix Of Random Forest

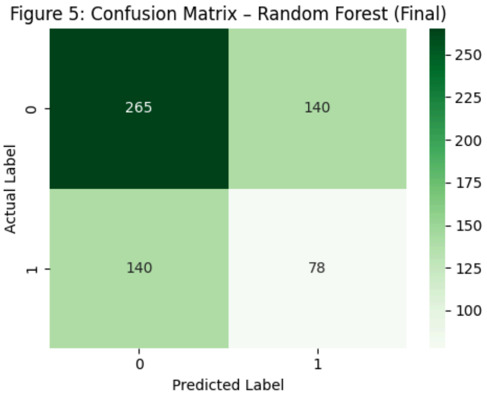

Figure 5: Confusion matrix illustrating the classification performance of the Random Forest model on the test dataset. Compared to Logistic Regression, the model demonstrates improved classification of potable water samples.

4. Results and Discussion

4.1 Model Comparison

The Following table summarizes the performance of the models. Notably the baseline models without any feature selection and hyperparameter optimization outperformed the models after aggressive feature selection and tuning.

Model	Status	Accuracy	Precision (0/1)	Recall (0/1)	F1-Score (Avg)
Neural Network	**Final**	**0.65**	**0.50**	**0.47**	**0.48**
Random Forest	Baseline	0.63	0.71 / 0.47	0.74 / 0.43	0.63 (weighted)
Logistic Regression	Baseline	0.49	0.63 / 0.33	0.50 / 0.46	0.50 (weighted)
Random Forest	Optimized	0.57	0.39	0.41	0.40
Logistic Regression	Optimized	0.48	0.33	0.45	0.38

4.2 Key Findings

- Best Model: The Neural Network achieved the highest accuracy followed closely by the baseline random forest.
- The models built after Hyperparameter Tuning and Feature Selection actually performed **worse**. This suggests that feature selection likely

removed variables that, while individually weak, provided necessary information when combined.

- Precision ensures public safety by minimizing the risk of unsafe water being classified as potable, while recall ensures detection of all safe water sources. Given the safety critical nature of water quality assessment, precision is prioritized over recall in this study.

4.3 Final Model Selection

Based on evaluation metrics, the Neural Network is selected as the final model as it achieved the highest accuracy on the test set.

5. Model Performance

The performance of all models was moderate. The low correlation coefficients observed during EDA suggested that linear models like logistic regression would struggle, which was confirmed by its low accuracy and non linear models like random forest and neural network performed significantly better.

5.2 Impact of Methods

- Feature Selection: This technique appears to have negatively impacted the results. Since all features had low correlation with the target, removing any of them resulted in a loss of information leading to lower scores in optimized models compared to baselines
- Hyper parameter Tuning: While intended to improved results, the search space might have been too congested or the cross validation splits may have differed significantly from the test set distribution

5.3 Interpretation of Results

The results indicate that water potability is a complex, non linear function of chemical properties. No single feature guarantees safety, it is a combination of factors.

5.4 Limitations and Future work

Limitations: The dataset is relatively small (3276 rows) and imbalanced. The moderate accuracy suggests data quality issues or that critical features (e.g., biological contaminants) might be missing.

Future works

- Apply SMOTE (Synthetic Minority Over sampling Technique)

- Experiment with XGBOOST or Gradient Boosting, which often outperforms Random Forest
- Avoiding discarding features based on linear correlation

6. References

- Concepts and Technologies of AI module materials
- Kaggle water potability dataset
- Scikit Learn documentation

7. Github

https://github.com/samTime101/AI-classification