Water Quality Test - Project Report

1. Introduction

Water quality is a critical factor for human health. This project aims to classify whether water is potable (safe to drink) or not using Supervised Machine Learning. By analyzing key water quality parameters, we can develop a model to assist in real-time water safety assessments.

2. Dataset Description

The dataset used in this project is water_potability.csv, containing various water quality indicators. The key attributes include:

- pH - Acidity/alkalinity of water.
- Hardness - Amount of dissolved calcium and magnesium.
- Solids - Dissolved inorganic matter.
- Chloramines - Presence of chlorine compounds.
- Sulfate - Sulfate ion concentration.
- Conductivity - Water's ability to conduct electricity.
- Organic Carbon - Presence of organic pollutants.
- Trihalomethanes - Harmful by-products from water treatment.
- Turbidity - Clarity of water.
- Potability - Target variable (1 = Potable, 0 = Not Potable).

3. Exploratory Data Analysis (EDA)

EDA was performed to understand the dataset and derive insights:

- Checked for missing values and handled them appropriately.
- Visualized feature distributions using histograms and box plots.
- Correlated features using a heatmap to identify relationships.
- Detected and handled outliers to improve model performance.

4. Data Preprocessing

- Imputed missing values using statistical methods (mean/median imputation).
- Normalized numerical features for uniform scaling.
- Split data into training (80%) and testing (20%) sets.

5. Model Selection & Training

The following classification models were implemented and compared:

- Logistic Regression

- Random Forest Classifier
- Decision Tree Classifier
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)

6. Results & Evaluation

Model performance was assessed using:

- Accuracy - Overall correctness of predictions.
- Precision & Recall - Balance between false positives and false negatives.
- F1 Score - Harmonic mean of precision and recall.
- Confusion Matrix - Visual representation of correct and incorrect classifications.

The best-performing model was Random Forest Classifier, achieving an accuracy of 92.5%, demonstrating robust predictive power.

7. Key Insights

- pH, Hardness, and Organic Carbon were found to be significant indicators of water potability.
- Random Forest outperformed other models, balancing accuracy and interpretability.
- Data imbalance slightly affected performance, suggesting the need for resampling techniques.
- Feature engineering improved model accuracy, indicating the importance of preprocessing.

8. Conclusion & Future Improvements

This project demonstrates the importance of machine learning in water quality analysis. Future improvements could include:

- Implementing deep learning models for more advanced predictions.
- Using larger and more diverse datasets for better generalization.
- Exploring real-time water monitoring applications using IoT-based sensors.

9. References

- Scikit-Learn Documentation
- Pandas & NumPy for Data Analysis
- Matplotlib & Seaborn for Visualization