1st Stage Project

Project Title: Water Quality Prediction using KNN

Name: Rohit Kumar Dayalani
Roll No: 220107072

1 Introduction

Access to clean and safe drinking water is a fundamental human necessity. However, many regions around the world struggle with water contamination due to industrial waste, agricultural runoff, and insufficient treatment facilities. Ensuring water potability is crucial for health, economic development, and sustainability. With the emergence of data-driven techniques, machine learning provides a promising approach to automate the detection of water quality based on physicochemical properties.

2 Objectives

- To predict whether a water sample is potable or not based on its chemical attributes.
- To evaluate the performance of K-Nearest Neighbors (KNN) in classifying water potability.
- To assess the model using appropriate metrics like accuracy, precision, recall, and F1-score.
- To understand the importance of data preprocessing in predictive performance.

.

3 Problem Statement

To develop a reliable machine learning model that predicts the potability of water samples using measurable physicochemical features, ensuring early detection of unsafe drinking water.

4 Significance of the Study

- Helps in real-time identification of unsafe water for public use.
- Reduces reliance on complex laboratory setups for initial screening.
- Enhances decision-making in water treatment plants and environmental monitoring.
- Supports sustainable development by improving water management.

5 Data Description and Preprocessing

D**ata Source**: Kaggle(https://www.kaggle.com/datasets/devanshibavaria/water-potability-dataset-with-10-parameteres)

Key Features:

- pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic_carbon, Trihalomethanes, Turbidity
- Target: Potability (0 = not safe, 1 = safe)

Steps Taken:

- Loaded dataset using Pandas.
- Identified and handled missing values via mean imputation.
- Normalized data using MinMaxScaler for KNN effectiveness.
- Visualized distributions and feature correlations using Seaborn.
- Checked class imbalance.

6 Methodology Flowchart

Raw Data → Data Cleaning & Imputation → Normalization →Train-Test Split → Model Training with KNN → Evaluation

7 Model Selection & Rationale

Model Used: K-Nearest Neighbors (KNN)

Why KNN?

- Simple and intuitive.
- Non-parametric, suitable for real-world datasets with non-linear decision boundaries.
- Performs well with smaller datasets and normalized features.

8 Model Training & Validation

- Split: 80% training, 20% testing.
- Hyperparameters to tune in next phase: K value, distance metric (Euclidean, Manhattan).
- Evaluation on test set using accuracy, classification report, and confusion matrix.

9 Evaluation Metrics

- **Precision/Recall/F1**: Varied due to class imbalance.
- Further tuning and handling of imbalance planned (e.g., SMOTE or resampling)

10 Deployment Strategy(Planned)

- Simple UI for users to input water sample attributes and get potability prediction.
- Deploy model using Flask or Streamlit for demonstration purposes.

11 Tools and Libraries Used

- **Python**
- **NumPy, Pandas** – Data manipulation
- **Matplotlib, Seaborn** – Visualization
- **Scikit-learn** – Modeling and evaluation

12 Scalability and Optimization

- Normalize features for KNN effectiveness.
- Explore other classifiers like Random Forest or SVM in the next phase.
- Optimize K-value and try feature selection to reduce dimensionality.

13 Use Case in Chemical Engineering

- Monitoring and control of water purification in chemical processing.
- Detection of pollutant levels in effluents.
- Assisting engineers in water treatment design using predictive analytics.

14 Expected Impact

- Improved water safety assessment.
- Scalable solution for remote areas with limited lab facilities.
- Contributes to environmental sustainability and health safety

15 Conclusion

This project demonstrates the feasibility of applying machine learning techniques like KNN for predicting water potability. Early results are promising, and further improvements will focus on model tuning, addressing class imbalance, and enhancing interpretability.

16 References

- Kaggle Dataset – Water Potability Dataset
🔗 https://www.kaggle.com/datasets/adityakadiwal/water-potability

- Scikit-learn Documentation – KNeighborsClassifier
🔗 https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

- Scikit-learn User Guide – Nearest Neighbors
🔗 https://scikit-learn.org/stable/modules/neighbors.html

- Article – Predicting Water Potability using Machine Learning (Medium)
🔗 https://medium.com/analytics-vidhya/predicting-water-potability-using-machine-learning-9f7d8c3d3857