Water Quality Prediction using KNN

Student Name: Rohit Kumar Dayalani

Roll No: 220107072

Final Project submission

Course Name: Applications of AI and ML for Chemical Engineering

Course Code: CL653

Submission Date: April 25, 2025

Contents

1 Executive Summary

This project addresses the critical issue of water potability prediction by developing a machine learning model using the K-Nearest Neighbors (KNN) algorithm. By leveraging a publicly available dataset containing physicochemical properties of water samples, the model determines whether a sample is safe for drinking. The approach emphasizes the importance of data preprocessing, including handling missing values and normalization. The outcomes demonstrate the potential of AI tools in environmental monitoring, specifically in enhancing decision-making in water treatment and ensuring public health.

2 Introduction

2.1.1 Background

Clean water accessibility is a fundamental need, particularly in chemical engineering processes where water purity directly impacts product quality and environmental safety. Industrialization has worsened water contamination issues, necessitating automated tools to detect unsafe drinking water in real-time.

2.1.2 Problem Statement

To develop a reliable machine learning model that predicts whether a water sample is potable based on its chemical properties, thus aiding in early identification of unsafe drinking water.

2.1.3 Objectives

- To predict water potability using KNN based on chemical features.
- To preprocess and normalize the dataset effectively.
- To evaluate model performance using standard metrics.
- To provide a potential real-time decision-making tool for water management systems.

3 Methodology

3.1.1 Data Source

The dataset was sourced from Kaggle:

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

It includes 10 parameters like pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity, and a binary target variable (Potability).

3.1.2 Data Preprocessing

- Missing values were handled using mean imputation.
- Normalization was performed using MinMaxScaler for KNN compatibility.
- Feature distribution and correlation were visualized using Seaborn.

3.1.3 Model Architecture

- **Algorithm**: K-Nearest Neighbors (KNN)
- **Why KNN?**: It's non-parametric, simple, and effective on small datasets with non-linear boundaries.
- **Parameters**: Distance metric (Euclidean), number of neighbors (k), normalized features.

4 Implementation Plan

4.1.1 Development Phases

- Phase 1: Data Collection and Cleaning
- Phase 2: EDA and Normalization
- Phase 3: Model Development using KNN
- Phase 4: Evaluation and Visualization

4.1.2 Model Training

- Data split: 80% train / 20% test
- Hyperparameter tuning (planned): different k-values and distance metrics
- Training on normalized dataset for KNN performance

4.1.3 Model Evaluation

- Accuracy
- Precision
- Recall
- F1-Score
- Confusion Matrix

5 Testing and Deployment

5.1.1 Testing Strategy

Model was tested using holdout test set and classification metrics to validate predictions on unseen data.

5.1.2 Ethical Considerations

- Ensuring correct prediction to avoid health hazards
- Transparency of AI decisions
- Addressing class imbalance to avoid bias

6 Results and Discussion

6.1.1 Findings

- The model was moderately accurate, with performance impacted by class imbalance.
- Certain features like sulfate, chloramines, and solids showed higher influence.
- Visualization revealed separability in feature space after normalization.

6.1.2 Comparative Analysis

While KNN performed reasonably, future phases may include testing models like Random Forest or SVM for improved robustness.

6.1.3 Challenges and Limitations

- Dataset imbalance led to reduced recall on minority class
- KNN is computationally intensive for larger datasets
- Potential overfitting without careful k-value selection

7 Conclusion and Future Work

The project demonstrates how machine learning, specifically KNN, can assist in water potability classification. It offers a foundation for real-time monitoring tools and underscores the role of AI in sustainable water management. Future work involves tuning the model further, addressing class imbalance using SMOTE, and exploring advanced classifiers.

8 References

• Kaggle Water Potability Dataset: (https://www.kaggle.com/datasets/devanshibavaria/water-potability-dataset-with-10-parameteres)

• Scikit-learn Documentation:(https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html)

• Medium article on water potability ML :(https://medium.com/analytics-vidhya/predicting-water-potability-using-machine-learning-9f7d8c3d3857)

9 Appendices

A. Dataset
Kaggle – 9 features + target (Potability: 0/1)

B. Key Code

- fillna(mean)
- MinMaxScaler()
- KNN(n=5)
- classification_report()

C. Visuals

- Heatmap
- pH Plot
- Confusion Matrix

D. Metrics
Class 0: F1 = 0.68
Class 1: F1 = 0.53

10 Auxiliaries

Web link: https://depelctive-project.onrender.com/

Data Source: https://www.kaggle.com/datasets/devanshibavaria/water-potability-dataset-with-10-parameteres

Python file: https://github.com/dayalanirohit/depelctive-project.git