

# Water Potability Prediction – Detailed Machine Learning Report

## 1. Introduction

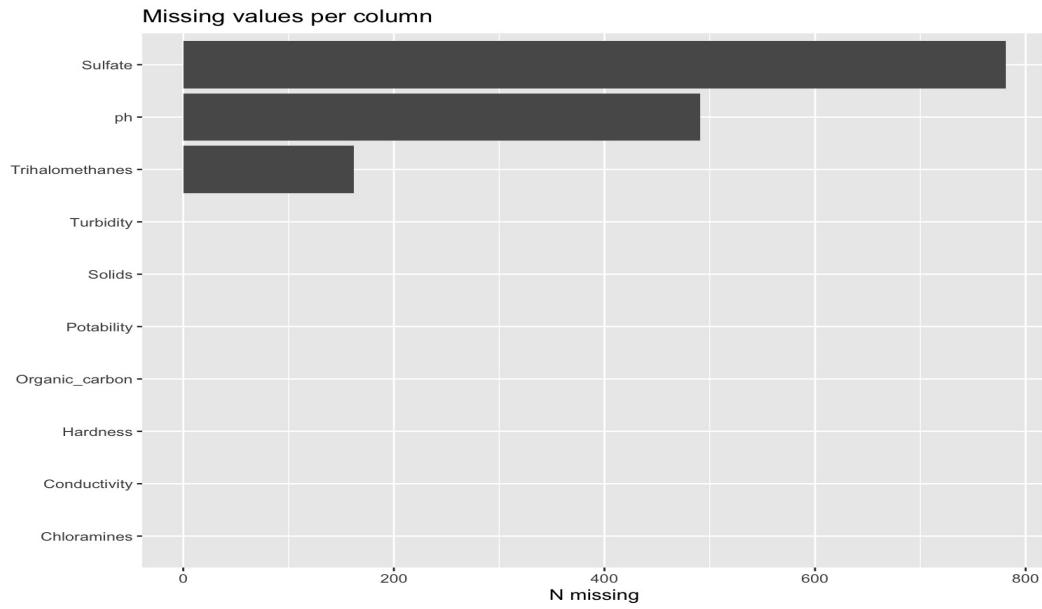
Access to safe drinking water is a critical public health priority. The goal of this project is to use machine learning techniques to predict whether a given water sample is potable or not. The dataset contains 3,276 water samples with nine chemical features including pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity. This report presents a complete machine learning workflow consisting of data preprocessing, exploratory data analysis, model development, tuning, evaluation, and interpretation of results.

## 2. Data Preprocessing

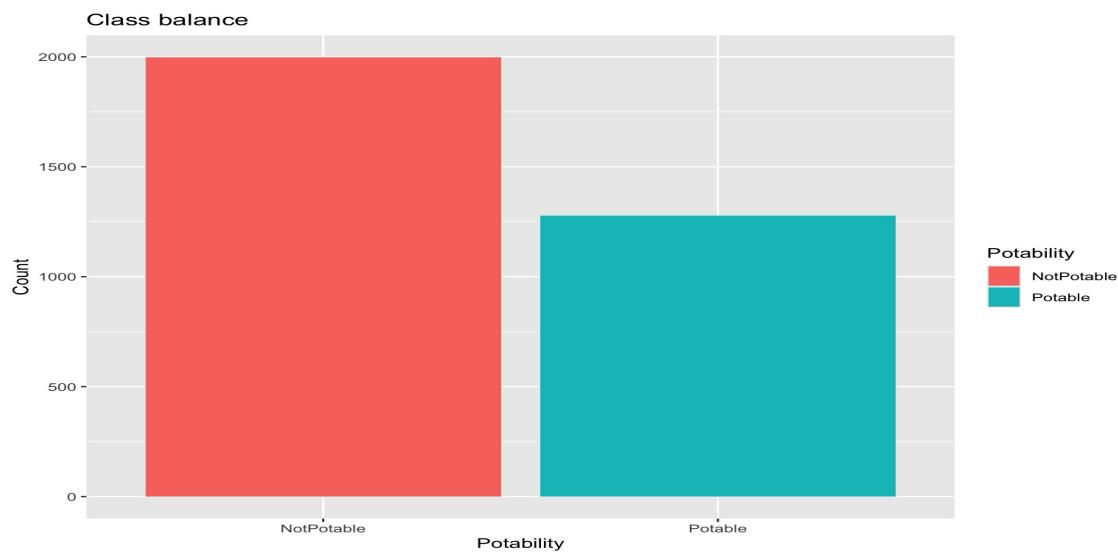
The initial dataset contained missing values in the pH, Sulfate, and Trihalomethanes columns. Since these variables are numerical and continuous, median imputation was applied. Median imputation is robust to outliers and appropriate for skewed distributions. After imputation, all predictor variables were centered and scaled using the caret preprocessing methods. This ensures that models like kNN, logistic regression, and random forest operate effectively without being influenced by differing feature scales. A stratified train-test split of 70% training and 30% testing was used to preserve the underlying class distribution in both sets.

## 3. Exploratory Data Analysis

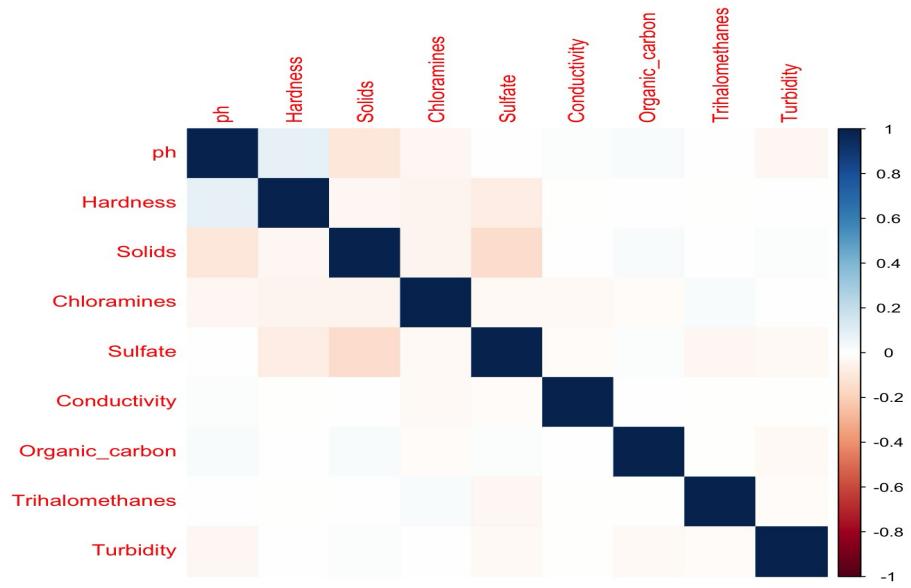
Exploratory data analysis (EDA) helps uncover the structure, distribution, and relationships in the data. Several visualizations were produced to understand missing values, class balance, feature correlations, and distributions.



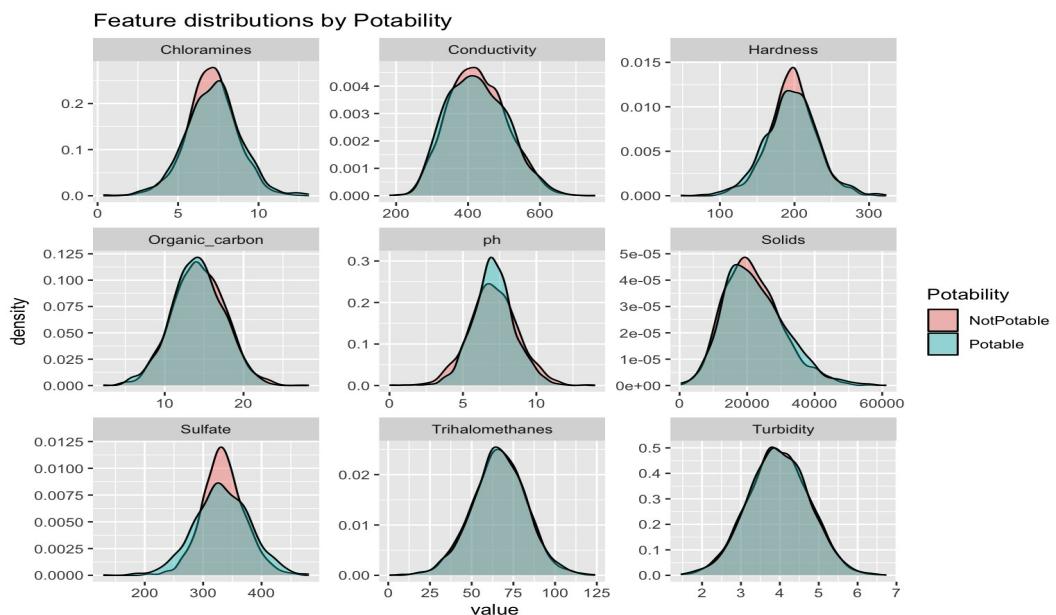
**Figure 1. Missing Values per Column**



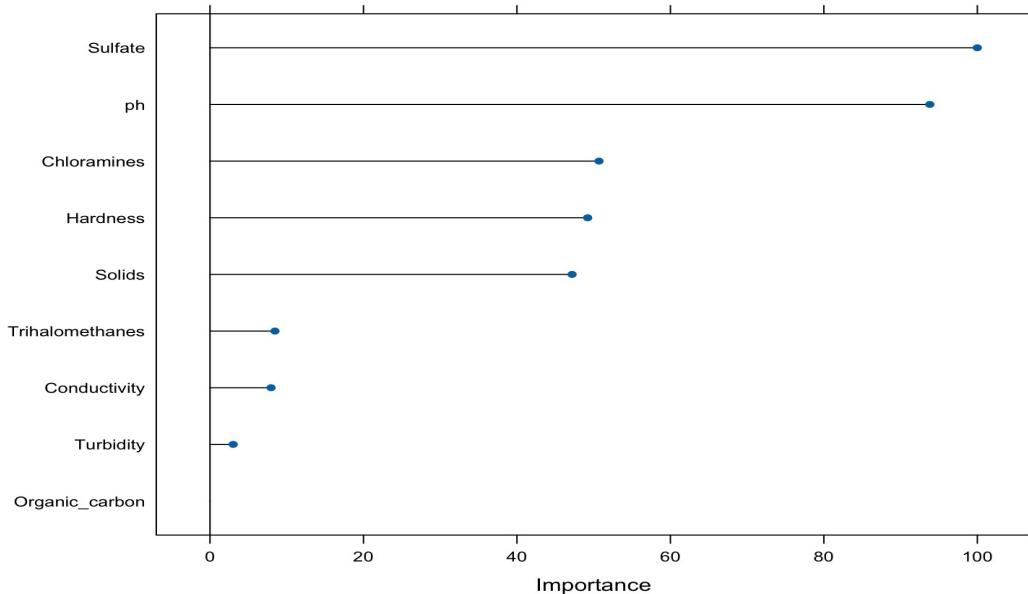
**Figure 2. Class Balance**



**Figure 3. Correlation Heatmap**



**Figure 4. Feature Distributions by Potability**



**Figure 5. Random Forest Feature Importance**

## 4. Machine Learning Models

Three models were trained on the processed dataset: Logistic Regression, k-Nearest Neighbors (kNN), and Random Forest. All models were evaluated using repeated 5-fold cross-validation. Logistic Regression was chosen as a baseline linear model, kNN as a distance-based nonlinear model, and Random Forest as an ensemble model capable of capturing interactions among features.

### 4.1 Logistic Regression

Logistic Regression is a simple, interpretable baseline model. However, it performed poorly on this dataset with an AUC of 0.491, suggesting linear separability assumptions do not hold for water quality data.

### 4.2 k-Nearest Neighbors (kNN)

kNN was trained with automated hyperparameter tuning using `tuneLength = 10`, which searches for optimal k values. It achieved an AUC of 0.639—better than logistic regression but still not ideal due to overlapping feature distributions.

### **4.3 Random Forest**

Random Forest outperformed all models with an AUC of 0.678. It captures nonlinear relationships and handles feature interactions effectively. A grid search over mtry values from 2 to 9 was performed, with mtry = 7 selected as the optimal value.

## **5. Final Model Evaluation**

The tuned Random Forest model was evaluated on the test set. Two evaluations were conducted: one using the default 0.5 probability threshold, and one using an optimized threshold computed via Youden's J statistic.

### ***Default 0.5 Threshold Results***

- Accuracy: 65.99% • AUC: 0.6776 • Sensitivity: 0.376 • Specificity: 0.841

### ***Optimized Threshold Results***

Using Youden's J, the threshold increased sensitivity significantly:

- Accuracy: 66.6%
- AUC: 0.675
- Sensitivity: 0.462
- Balanced Accuracy: 0.629

## **6. Conclusion**

This project demonstrates a complete machine learning pipeline for predicting water potability. Random Forest was identified as the best-performing model due to its ability to learn complex nonlinear patterns in the data. Although overall predictive accuracy is moderate, the model provides meaningful insights into which chemical components—such as sulfate, pH, and chloramines—have the strongest influence on potability. This pipeline can serve as a foundation for more advanced predictive systems in environmental and public health applications.

### #Team Contribution

Simran conducted EDA and visualizations. Ashwin developed and evaluated ML models. Chandhana analyzed feature importance and interpretation. Sai performed Random Forest tuning and ROC threshold analysis.