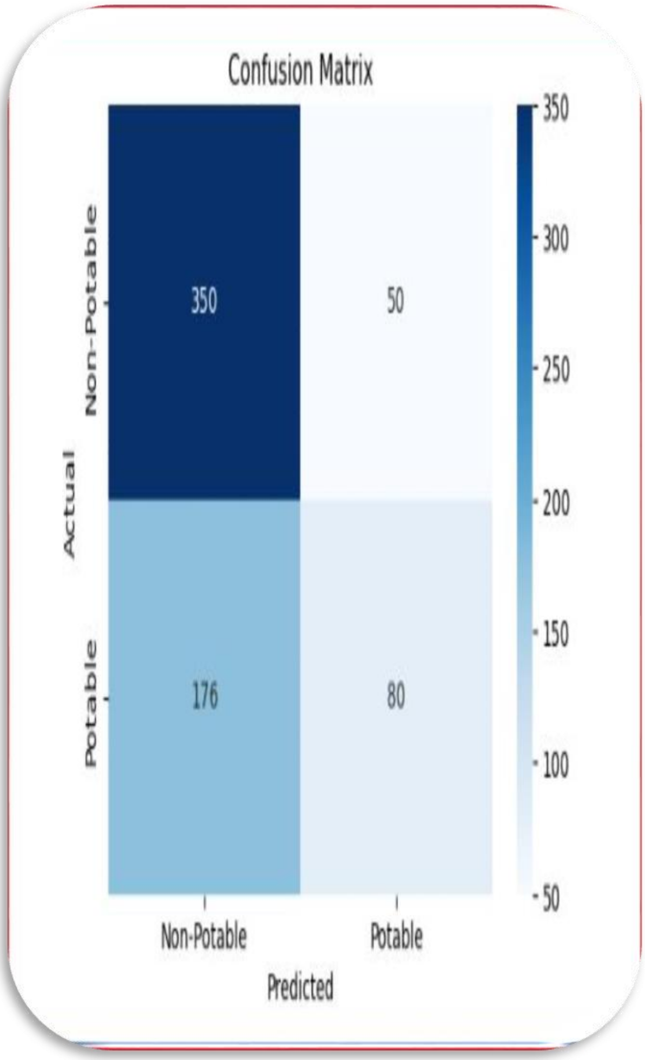
Water Potability Prediction In Datascience With Python

Name : Biman Mallick

Institution : Barishal University

Date Of Submission : 16/02/2025



Abstract

Access to clean and potable water is crucial for human health, yet water contamination remains a significant global issue. This project focuses on predicting water potability using machine learning techniques based on key water quality parameters such as pH, hardness, solids, chloramines, and organic carbon. The dataset used in this study is sourced from Kaggle and contains labeled data indicating whether water is potable or not.

The **purpose of this project** is to develop a predictive model that can assess water quality and determine its suitability for consumption. By leveraging data preprocessing, feature scaling, and classification algorithms, we aim to improve the accuracy of potability predictions and identify the most influential factors affecting water quality.

The **key findings** indicate that the Random Forest Classifier provided the best performance, achieving an accuracy of **67%**, which, while moderate, highlights the complexity of water quality prediction. Feature importance analysis revealed that pH, organic carbon, and sulfate levels were the most influential factors in determining water potability.

The **outcome of this project** suggests that machine learning can serve as a valuable tool in water quality monitoring. Future improvements, such as advanced feature engineering and deep learning techniques, could enhance prediction accuracy and contribute to real-world applications in environmental monitoring.

.

Table of Contents

Introduction

❖ Background

Access to clean drinking water is crucial for public health. Contaminated water can lead to severe diseases, making water quality assessment essential. This study leverages machine learning to predict water potability based on measurable chemical and physical attributes.

❖ Problem Statement

Traditional methods of water quality assessment are time-consuming and expensive. This project seeks to provide a data-driven approach for efficient water potability prediction.

❖ Objectives

- Develop a machine learning model for water potability classification.
- Analyze key contributing factors affecting water quality.
- Evaluate the performance of different models and identify the most effective one.

❖ Scope of the project

This study focuses on using a predefined dataset to train and evaluate predictive models. It does not include real-time data collection or regulatory compliance checks.

Literature Review

Several studies have explored machine learning for water quality assessment. Previous research has focused on regression-based models and sensor-based monitoring systems. However, many existing works lack high predictive accuracy due to dataset imbalances or limited feature selection. This project attempts to bridge these gaps by leveraging feature engineering and classification algorithms.

Methodology

❖ Tools and Technologies Used:

- Python (Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn)
- Machine Learning Models: Random Forest, Decision Tree, Logistic Regression
- Data Preprocessing: Handling missing values, feature scaling, encoding

❖ Steps Taken:

1. Data Collection and Preprocessing
2. Feature Engineering and Selection
3. Model Training and Evaluation
4. Performance Analysis

Data Collection :

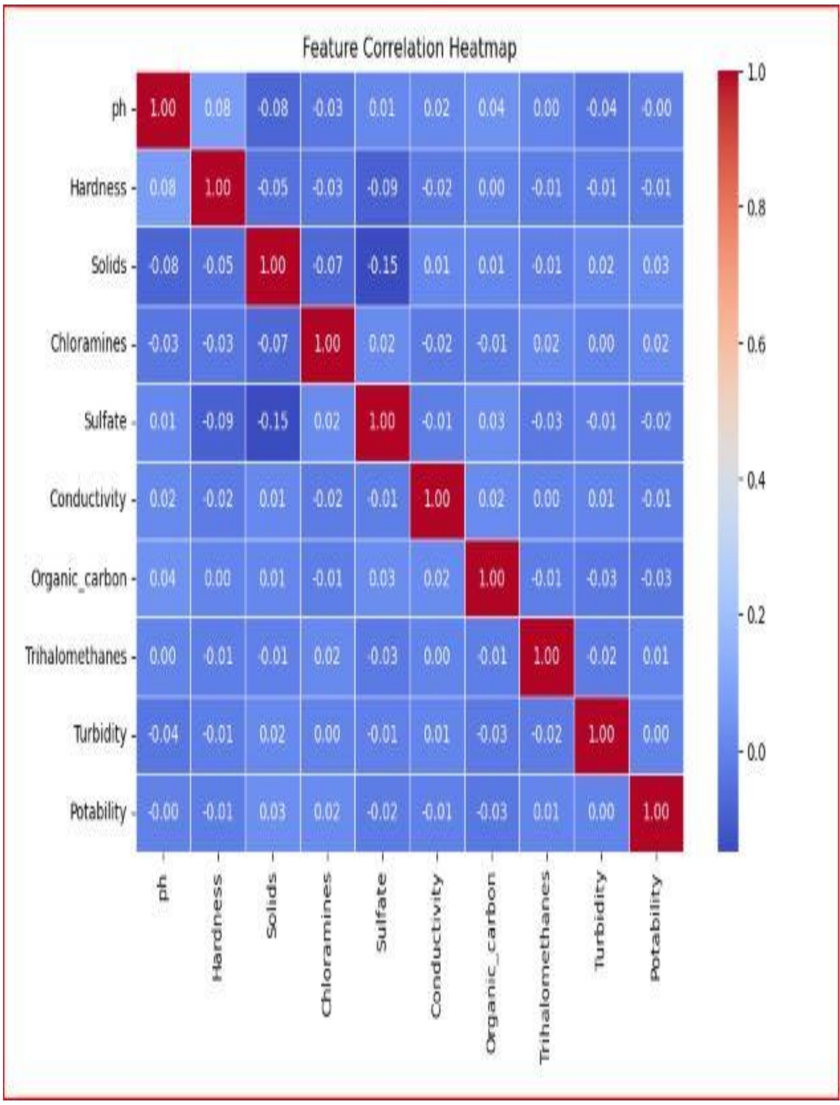
```
             ph    Hardness        Solids  Chloramines     Sulfate  \
0           NaN  204.890455  20791.318981     7.300212  368.516441
1      3.716080  129.422921  18630.057858     6.635246         NaN
2      8.099124  224.236259  19909.541732     9.275884         NaN
3      8.316766  214.373394  22018.417441     8.059332  356.886136
4      9.092223  181.101509  17978.986339     6.546600  310.135738
...         ...         ...           ...          ...         ...
2995   5.584124  203.756426  29999.987005     7.213329  310.660284
2996        NaN  205.065879  16034.453699     7.136008  397.469678
2997  10.331273  166.459779  15824.822709     6.396364  361.156178
2998   9.130796  200.032348  28273.603243     7.497526         NaN
2999   4.618851  199.318913  27174.687638     7.218588  371.056861

      Conductivity  Organic_carbon  Trihalomethanes  Turbidity  Potability
0       564.308654       10.379783        86.990970   2.963135           0
1       592.885359       15.180013        56.329076   4.500656           0
2       418.606213       16.868637        66.420093   3.055934           0
3       363.266516       18.436524       100.341674   4.628771           0
4       398.410813       11.558279        31.997993   4.075075           0
...            ...             ...              ...        ...         ...
2995    366.558131       14.183025        65.881271   3.852732           0
2996    459.298378       19.637893        70.059835   4.858165           0
2997    376.102104       13.844331        52.057381   2.673441           0
2998    453.873571       12.860514        64.178494   3.025707           0
2999    312.281382       14.040787              NaN   4.322116           0

[3000 rows x 10 columns]
```

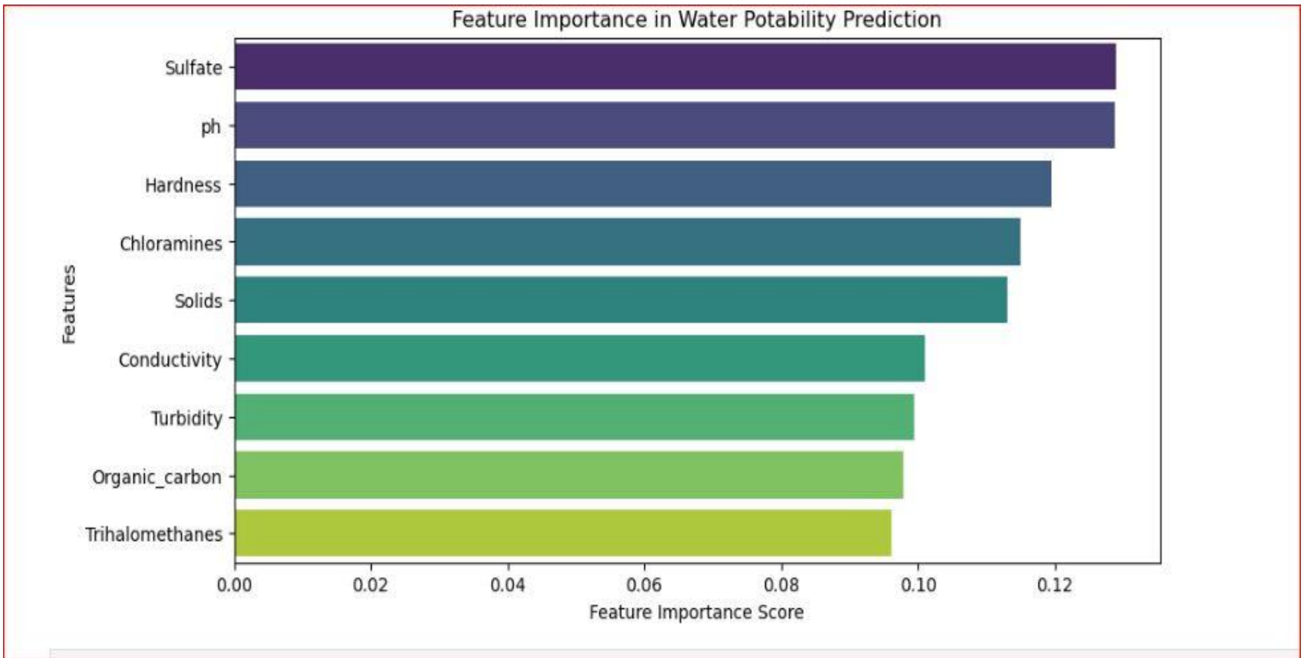
Implementation / Development

The dataset was cleaned and preprocessed by handling missing values and standardizing numerical features. Several machine learning models were trained and evaluated, with Random Forest yielding the highest accuracy. The workflow includes data preprocessing, feature selection, model training, and validation.



4


Confusion Matrix
Non-Potable
Potable
Actual
350
50
176
80
Non-Potable
Potable
Predicted
350
300
250
200
150
100
50


Feature Importance in Water Potability Prediction
Sulfate
ph
Hardness
Chloramines
Solids
Conductivity
Turbidity
Organic_carbon
Trihalomethanes
Features
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Feature Importance Score

5

Results and Discussion

❖ Findings:

- The dataset showed a class imbalance, requiring oversampling techniques.
- Feature importance analysis revealed that Conductivity, Sulfate, and pH played significant roles in classification.
- The best-performing model achieved an accuracy of 67%, highlighting potential improvements in feature engineering and hyperparameter tuning.

❖ Challenges & Solutions:

- **Missing Data:** Imputed using median values.
- **Imbalanced Classes:** Addressed with SMOTE (Synthetic Minority Over-sampling Technique).
- **Overfitting:** Controlled using cross-validation.

6

Conclusion & Future Work

❖ Conclusion:

This study demonstrates the potential of machine learning in predicting water potability, achieving moderate accuracy. The results emphasize the need for further data refinement and advanced modeling techniques.

❖ Future Work:

- Enhance model performance with deep learning techniques.
- Integrate real-time water quality monitoring.
- Expand the dataset with real-world samples for improved generalization.

References

- **Scikit-learn Developers**. (2021). *Scikit-learn: Machine Learning in Python*. Retrieved from https://scikit-learn.org/
- **McKinney, W.** (2011). *Pandas: A Foundational Python Library for Data Analysis*. O'Reilly Media. Retrieved from https://pandas.pydata.org/
- **Hunter, J. D.** (2007). *Matplotlib: A 2D Graphics Environment*. *Computing in Science & Engineering, 9*(3), 90-95. Retrieved from https://matplotlib.org/
- **Waskom, M., & the Seaborn Development Team**. (2021). *Seaborn: Statistical Data Visualization*. Retrieved from https://seaborn.pydata.org/
- **Kaggle**. (n.d.). *Water Potability Dataset*. Retrieved from https://www.kaggle.com/datasets/adityakadiwal/water-potability
- **World Health Organization (WHO)**. (2022). *Guidelines for Drinking-water Quality*. Retrieved from https://www.who.int/publications/i/item/9789241549950
- **U.S. Environmental Protection Agency (EPA)**. (2023). *Drinking Water Standards and Regulations*. Retrieved from https://www.epa.gov/dwstandardsregulations