Project Documentation: Predicting Potability of Water

Table of Contents

1. Introduction

2. Data Overview

3. Data Preprocessing

3.1 Data Cleaning

3.2 Handling Missing Values

3.3 Outlier Detection and Treatment

3.4 Feature Engineering

4. Model Selection and Tuning

4.1 Model Selection

4.2 Hyperparameter Tuning

5. Model Evaluation

5.1 Performance Metrics

5.2 Results

6. Application Development

7. Conclusion and Future Work

1. Introduction

Water is utilized for drinking, household usage, food production, or recreational activities, it is crucial for the public's health to have access to safe, readily available water. Drinking water, also referred to as potable water, is obtained from surface and underground sources and treated to satisfy state and federal criteria for consumption. Natural water is purified to remove viruses, germs, hazardous compounds, feces, and microbes. Consuming raw, untreated water might result in stomach issues including diarrhea, vomiting, or fever. This is an effort to use machine learning to evaluate whether the water that is accessible is drinkable or not.

2. Data Overview

Source: public dataset from Kaggle.

Size: Total Number of columns :- 10,

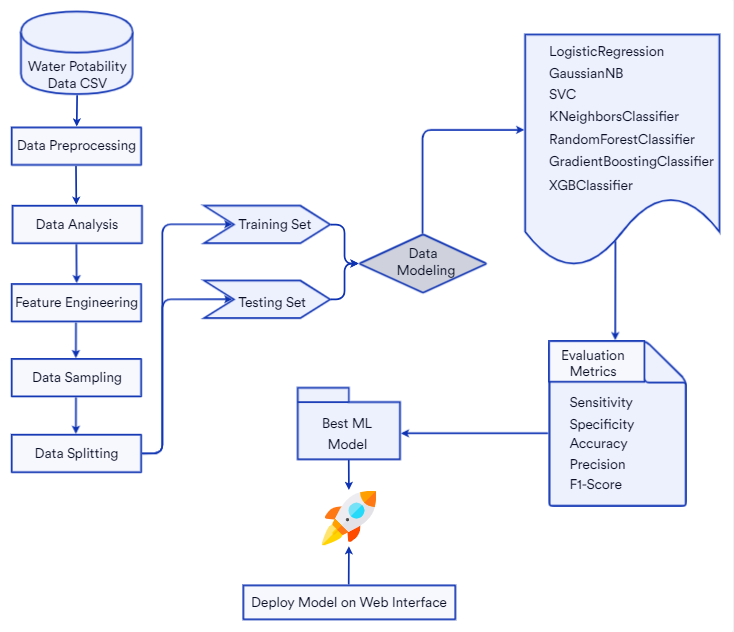
Total number of records :- 3276.

Features:

1. ph
2. Hardness
3. Solids
4. Chloramines
5. Sulfate
6. Conductivity
7. Organic\_carbon
8. Trihalomethanes
9. Turbidity

Target Variable: Potability (0 for non-potable, 1 for potable)

3. Architecture



4. Data Preprocessing

4.1 Data Cleaning

Data cleaning is the process of identifying and correcting errors, inconsistencies, and inaccuracies in a dataset. It ensures that the data is of high quality and reliable for analysis and modeling.

4.2 Handling Missing Values

**Identify Missing Values:**

Use functions like isna() or isnull() to detect missing values in your dataset.

**Imputation:**

Mean/Median Imputation: Replace missing values with the mean or median of the feature. This is suitable for numerical data. For the water potability data use case, we will apply median imputation to handle missing values.

4.3 Outlier Detection and Treatment

Utilizing box plot and z-score, outlier detection is carried out. To preserve the data in a small dataset, all outliers were imputed to the median value.

4.4 Feature Engineering

A correlation matrix is used to discover relationships between the various dataset parameters.

5. Model Selection and Tuning

5.1 Model Selection

Evaluated various classification models like Logistic Regression, Decision Trees, Support Vector Classifier, Random Forest Classifier, KNN Classifier, XGBoost. Random Forest Classifier demonstrated superior performance in terms of accuracy and robustness.

5.2 Hyperparameter Tuning

GridSerchCV and manual tuning with a lot of trials resulted in the final model. After tuning the Classifier we came to a conclusion that the following parameters provide best outcome.

RandomForestClassifier(n\_estimators=170,criterion='log\_loss').

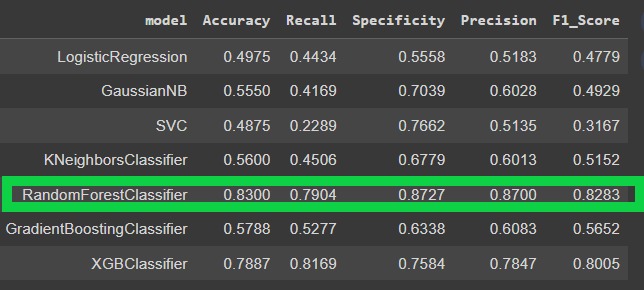
6. Model Evaluation

6.1 Performance Metrics

Assessed model performance using metrics such as Accuracy, Recall, Specificity, Precision, F1\_Score.

Defined the confusion matrix to visualize model performance.

6.2 Results



7. Application Development

A simple web app built using Streamlit to take the nine parameters as input, pass the inputs to the model which then gives the prediction. The model gives a binary classification i.e. 0 or 1. A text output is shown to the user along with the probability of the prediction.

8. Future Work

Future work may involve incorporating additional features or exploring more advanced modeling techniques.

Further, a IOT device providing the input parameters can be built using the required censors which can then be paired with this model to provide real time results. If the device is made to be pocket friendly it can be carried and used on the go.

9. Conclusion

successfully developed a robust Random Forest Classifier model for predicting water potability. The model demonstrates high accuracy and performance, providing a reliable tool for water quality assessment. The implementation of median imputation further enhanced its predictive capabilities