# Assignment 8.2

## Project-2: Project Milestone – 3- Final Paper

Sucharitha Puppala

Data Science, Bellevue University

DSC680-T301 Applied Data Science (2233-1)

Prof. Catherine Williams

February 04, 2023.

Project-2

**Topic:**

Prediction of Water Quality using machine learning algorithms.

**Business Problem:**

Let's have a quick look into the topic 'Drinking Water. 'Drinking water is water that is used to drink or for food preparation; potable water is water that is safe to be used as drinking water. The amount of drinking water required to maintain good health varies, and depends on physical activity level, age, health-related issues, and environmental conditions. Water is the most important nutrient for the body. It has many benefits for your health and helps to protect you from illness and disease. Water is also an essential part of a healthy lifestyle.

Potable water is also known as drinking water and comes from surface water and groundwater sources. This water is treated to levels that meet state and federal standards for consumption.

Water potability testing looks for coliform bacteria, improper pH, sodium, chloride nitrates, sulfate, manganese, iron, water hardness, and the total dissolved solids in the water.

Why is Water Quality testing important? Water quality testing is important as many tiny microorganisms and substances naturally get into a water supply that may be detrimental to a person's health. Many of these substances that get added to water can cause digestive issues, illness and (in some severe cases) death.

These effects of water contamination can be efficiently tackled if the data of the water from different water sources is collected and is analyzed that helps in prediction of water quality beforehand. Using machine learning models the data can be analyzed well for the water quality prediction.

Project-2

The project mainly focuses on identifying the machine learning model that best fits in predicting the water quality.

The topic raises the following questions:

1. Which features play an important role in predicting the water quality?
2. Which model best fits for the water quality prediction?

**Dataset:**

The dataset used for this project is collected from Kaggle.com.

https://www.kaggle.com/datasets/adityakadiwal/water-potability

1. pH value:

PH is an important parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status. WHO has recommended maximum permissible limit of pH from 6.5 to 8.5. The current investigation ranges were 6.52–6.83 which are in the range of WHO standards.

2. Hardness:

Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. Hardness was originally defined as the capacity of water to precipitate soap caused by Calcium and Magnesium.

3. Solids (Total dissolved solids - TDS):

Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. The water with high TDS

value indicates that water is highly mineralized. Desirable limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which prescribed for drinking purpose.

4. Chloramines:

Chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are most commonly formed when ammonia is added to chlorine to treat drinking water. Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.

5. Sulfate:

Sulfates are naturally occurring substances that are found in minerals, soil, and rocks. They are present in ambient air, groundwater, plants, and food. Sulfate concentration in seawater is about 2,700 milligrams per liter (mg/L). It ranges from 3 to 30 mg/L in most freshwater supplies, although much higher concentrations (1000 mg/L) are found in some geographic locations.

6. Conductivity:

Pure water is not a good conductor of electric current rather's a good insulator. Increase in ions concentration enhances the electrical conductivity of water. Electrical conductivity (EC) actually measures the ionic process of a solution that enables it to transmit current. According to WHO standards, EC value should not exceeded 400 μS/cm.

7. Organic_carbon:

Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water. According to US EPA < 2 mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is use for treatment.

Project-2

8. Trihalomethanes:

THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppm is considered safe in drinking water.

9. Turbidity:

The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter. The mean turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.

10. Potability:

Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.

**Methods:**

As the problem is a classification problem I would like to use the following models K- Nearest Neighbor Classifier, Random Forest Classifier, and XGB Classifier. All the models will be created, trained, and fit with the test and train datasets respectively. For the model evaluation, evaluation metrics Accuracy, Precision, Recall, and F1 Score will be calculated with the test data. Confusion Matrices that summarizes the performance of the model build are plotted for each model respectively.

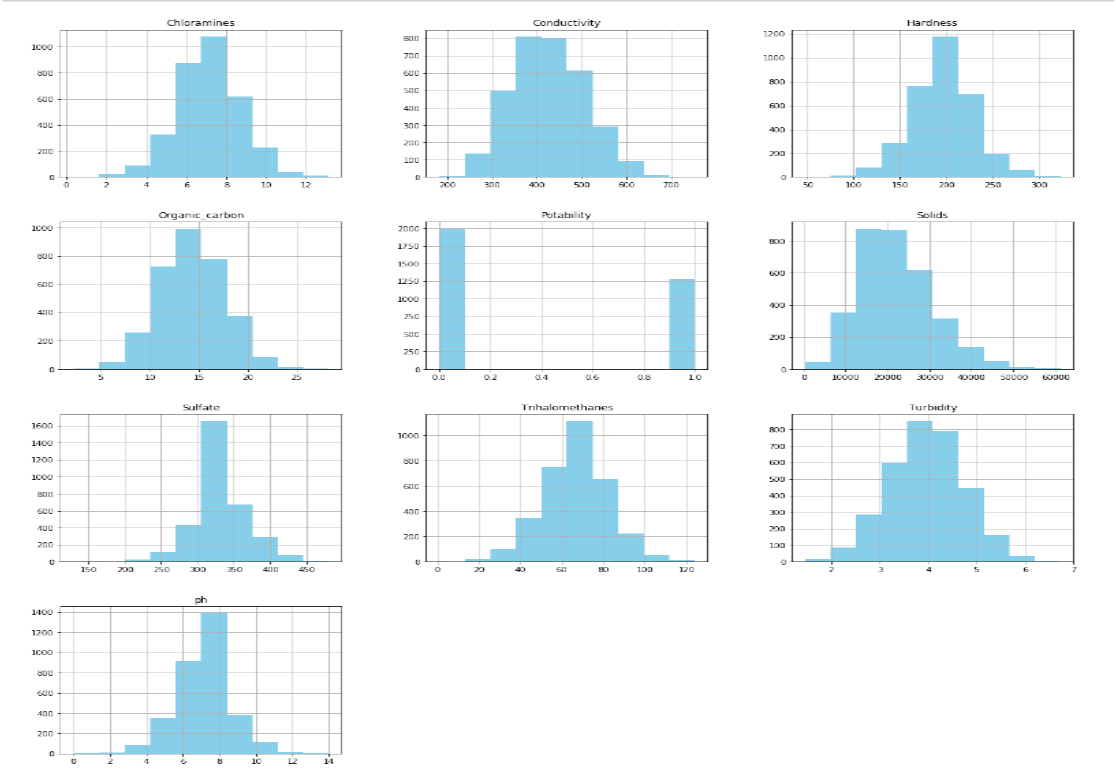
**Analysis:**

The analysis of the project is done using the CRISP DM methodology. The data file water_potability.csv is loaded for analysis. The shape of the dataset is obtained, and the dataset is of size 3276*10(i.e. 3276

rows and 10 columns). Initially the dataset is checked for the null values if any present in the dataset, when observed ph, Sulphates and Trihalomethanes columns are having more null values. All the null values are being replaced with the mean values of the columns respectively for further analysis. Then check for duplicate values are done, when observed there are no duplicate values in the dataset. The unique values present in the dataset are extracted and the dataset data types are obtained, all the data types are of float data types. In the Exploratory Data Analysis (EDA) Histograms, Pie plot, Box plots, Pair plot, and Heat maps are used for the analysis.

The following figure shows the histogram plot of all the numerical data types present in the data set.

*Figure 1: Histograms of numerical features of the dataset.*



From the above histogram plot we see the distribution of the variables in the dataset, we can see that some features are slightly skewed which is negligible.

The following plots show the distribution of the target variable “Potability” in the dataset selected.

*Figure 2: Distribution of the target variable “Potability”*

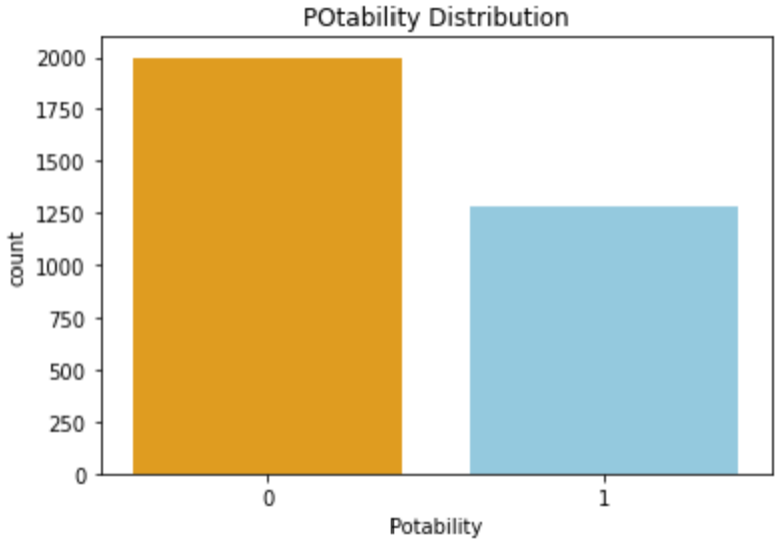
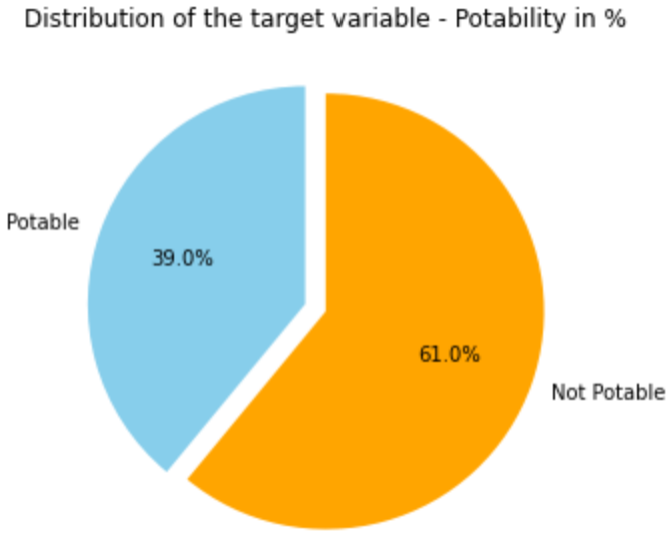



*Figure 3: Distribution of the target variable “Potability” in percentages.*



From the above plots we can say that the data set contains 39% of the water samples that are potable i.e. safe for drinking and the remaining 61% of the water samples are not potable i.e. not safe for drinking.

Box plots are plotted to know the distribution of the target variable “Portability” with the features available from the dataset. Appendix –A gives the box plots of all the features with the target variable.

When observed the box plots it is observed that the features are distributed equally with respect to the Potability of the water. To know the correlation of the variables in the dataset the correlation heat map and pair plots are plotted. Appendix- B gives the heat map and the pair plot of the dataset. From the correlation heat map we can say that the feature ‘Hardness is having good correlation but with very low value of correlation coefficient. Boxplots are plotted to identify the outliers. Appendix – C provides the Box plots. When observed the box plot we can see that the feature ‘Solids’ are having some outliers.

The outliers are removed by IQR (Inter Quartile Range Method) and when observed there are 47 outliers present in the Solids column of the dataset, which are not modified as the impact is assumed to be very low on the analysis.

The dataset is split into test and train datasets. Initially the dataset is split into features(X) with the target variable dropped from the dataset and target(y) variable by considering only the 'Potability' column from the dataset. The data set is to be split into training and test datasets using 'Potability' as the target variable. Here the dataset is split into 80% training dataset and 20% test dataset, taking the test size as 0.2 and the random_state as 42. X_train, X_test, y_train, y_ test are created using train_test_split()

The train and test data sets shapes are obtained for better understanding the split of the dataset.

The features train and test datasets are standardized using StandardScaler. The X_train dataset is transformed and fit to the standard scaler and the X_test dataset is just transformed but not fit.

With this the data is ready for model building.

All the models i.e. K- Nearest Neighbor Classifier, Random Forest Classifier, and XGB Classifier are built. All the models are created, trained, and fit with the train datasets. For the selection of the best model that fits our data, evaluation metrics that are most commonly used for classification models Accuracy, Precision, Recall, and F1 Score are calculated with the test data. Confusion Matrices that summarizes the performance of the model build are plotted for each model respectively. After observing the evaluation metrics calculated for each model with the test dataset and train datasets, the best model that fits the dataset, is selected.

**K Nearest Neighbor Classifier:**

In the KNN classifier model the below are the evaluation metrics obtained for the selected dataset.

As mentioned above the most common evaluation metrics in classification problems, Accuracy which is the total number of correct predictions divided by the total number of predictions made for a dataset is calculated and is observed to be about 0.53(53%) which is very low.

Precision this is the ratio of True Positives to all the positives predicted by the model.

Low precision: the more false positives the model predicts, the lower the precision.

When observed the value of the precision it is very low i.e. 0.335 which is not desirable as many false positives indicate wrong prediction of the water potability that are actually not potable, which may be dangerous for human health.

Recall (Sensitivity) this is the ratio of True Positives to all the positives in your Dataset.

Low recall: the more False Negatives the model predicts, the lower the recall.

When observed the recall value for the KNN model it is 0.246 which is very low indicating that there are many samples of water that are not potable but actually they are potable. This makes wrong prediction of potable water as not potable and decreasing the sources of potable water.

A good F1 score means that you have low false positives and low false negatives, indicating you are correctly identifying real threats, and you are not disturbed by false alarms. A F1 score is considered good when it's 1, while the model is bad when it's 0.

When observed the F1 – Score for the KNN model we can see that the value is 0.284 which is very low and the model is a failure for the prediction of the water potability for the dataset selected.

Confusion matrix for the KNN classifier model is plotted and is available for reference in the Appendix-D.

From the confusion matrix plot for the KNN classifier we can see that there are many actual negatives and predicted positives and very less actual and predicted values as positives indicating that there are very less samples of water that are potable which is not desirable.

**Random Forest Classifier Model:**

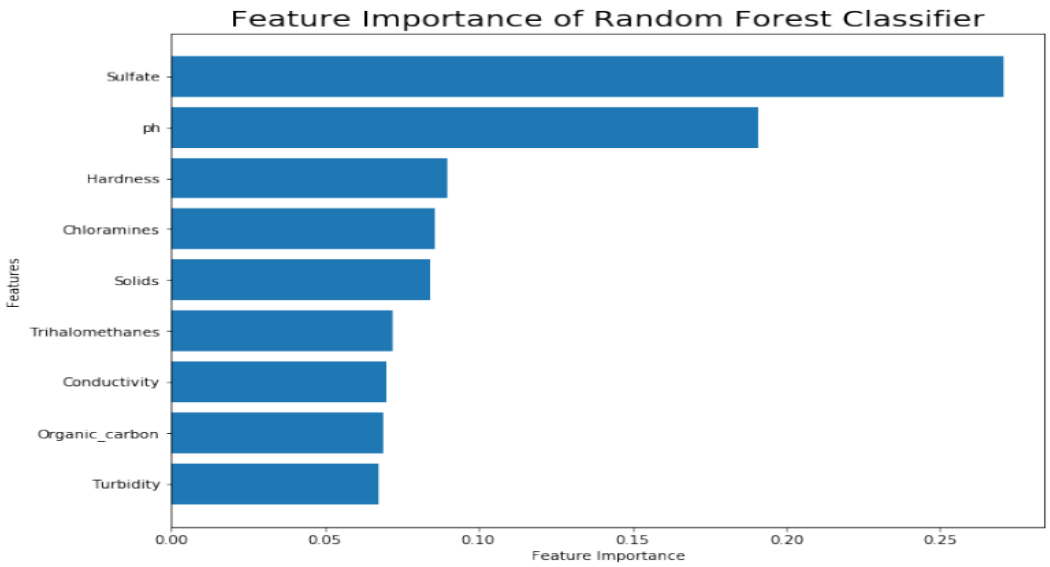
When observed the accuracy of the Random Forest Classifier model it is 0.807(nearly 81%) which is very good accuracy score. The Precision, Recall, and F1 Scores are also good indicating that the model performed well with the selected dataset.

Confusion Matrix for the Random Forest Classifier is plotted which is available for reference in the Appendix-D. When observed the Confusion Matrix we can see that there are more true negatives indicating that many of the water samples are not potable. The numbers of true positives from the

water samples that are potable are more in count when compared to the KNN classifier model which is a desirable one.

Feature importance plot of the Random Forest Classifier model is plotted. Below is the feature importance plot where we can see that the features Sulfate and ph are playing an important role in the prediction of the water potability using Random Classifier model.

*Figure 4: Feature importance plot of Random Forest Classifier*

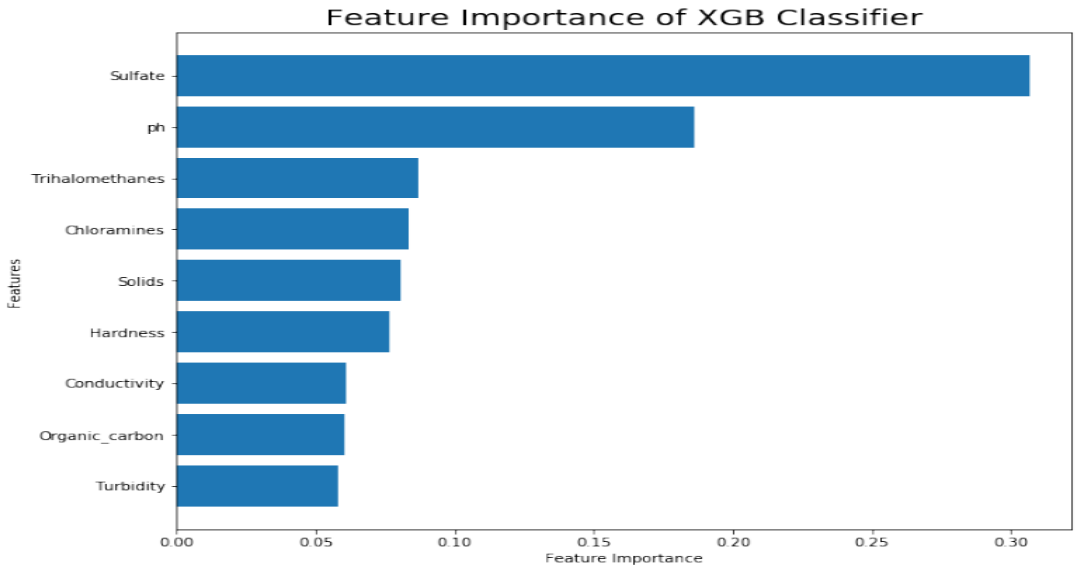



**XGB Classifier:**

When observed the accuracy of the XGB Classifier model the accuracy is about 0.80 (80%) which is a good accuracy score. The Precision score, Recall, and F1 Score are calculated and when observed they had good scores.

Confusion matrix for the XGB Classifier model is plotted and when observed the Confusion Matrix plot we find that the number of true positives i.e. the number of samples of water that are potable are good in count. The plot of Confusion Matrix is available in Appendix-D for reference.

*Figure 5: Feature importance plot of XGB Classifier*




Feature importance plot for the XGB Classifier model is plotted and from the above we can see that the features Sulfate and Ph from the selected dataset play an important role in the prediction of Water Potability using XGB Classifier model.

The results table for all the three models is available in the Appendix-E for reference.

**Conclusion:**

There are several factors that can cause water contamination in different water bodies. From the above analysis we can conclude that the Random Forest Classifier Model and XGB Classifier model performed well for the water quality dataset selected and from the feature importance plots of both the models we can say that Sulfate and ph plays an important role in prediction of water quality. The KNN Classifier performance is very low with very low accuracy score.

**Assumptions:**

Here in this analysis few values of the features are having null values and cannot be considered as null. Hence the mean of the respective column is calculated and replaced the null values with the mean value for better analysis. The dataset selected for this project didn't have the details regarding the source or the process of collecting the samples of the water. The method of measurement of the values present in the dataset is assumed to be valid and performed the analysis.

**Limitations:**

The study's limitations are observed in the heterogeneity between the models that is difficult to compare them. Land-use change and climate change affect hydrological components, and consequently river discharge and pollutant transport. Therefore, it is essential to take into account land-use and climate changes, which may improve the accuracy of the ML models.

**Challenges/ Issues:**

The amount of uncertainty in any model is difficult to know exactly, but there are specific types of uncertainties which can affect the model's performance.

**Future Uses/Additional Applications**

As part of additional applications, I would like to use the functions and Pipelines for transforming the data and to improve the code readability. There are few other features that can be added to the dataset, and more samples can be added to the dataset which helps in better prediction of water quality.

**Recommendations**

For future recommendations of this project, I would like to perform the Hyper parameter tuning on the models and check the model performance, and for evaluation purpose Cross Validation can be included for better model evaluation. With hyper parameter tuning the performance of the models used in this analysis can be improved.

**Implementation Plan**

Though the Random Forest Classifier and XGB Classifier models performed well with the features available in the dataset, the models are not recommended for deployment, as there are a lot of other features that are to be added to the dataset in prediction of water quality. However these models can be used for understanding the features that are more responsible for the predicting the water quality of different sources of water.

**Ethical Considerations:**

Ethical considerations while handling the water data are transparency, justice and fairness, responsibility and accountability, non – maleficence. I have taken care to handle the data with transparency without any misrepresentation of the data.

**References**

*Kaggle.com, Diabetes dataset, https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset*

*Google, Wastewater Digest, https://www.wwdmag.com/editorial-topical/what-is-articles/article/10940236/what-is-potable-water*

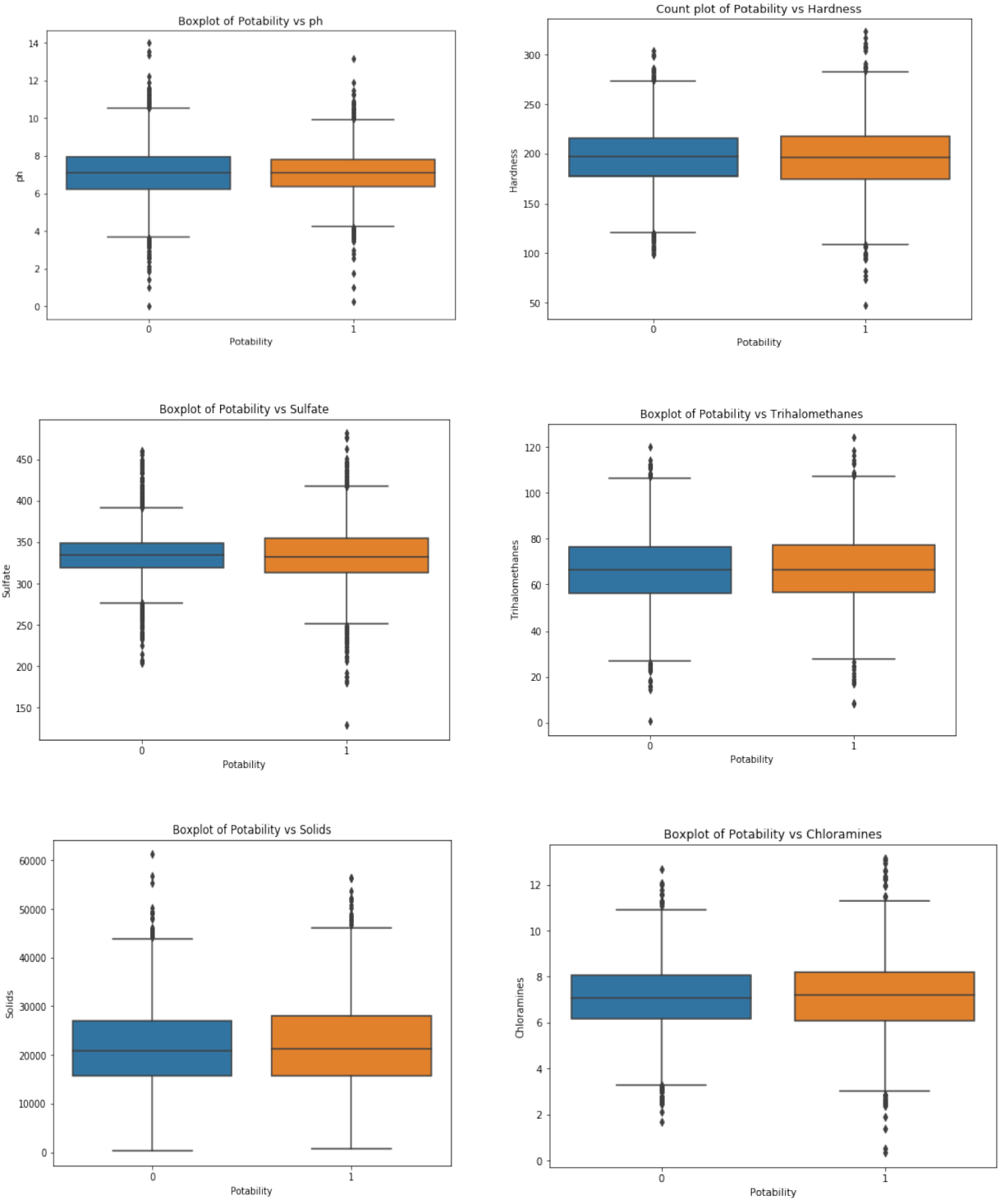
*Google, c and j water softeners, The importance of water quality testing, https://candjwater.com/2022/02/01/the-importance-of-water-quality-testing/*

*Google, Using Machine Learning Models for Predicting the Water Quality Index in the La Buong River, Vietnam, https://www.mdpi.com/2073-4441/14/10/1552*

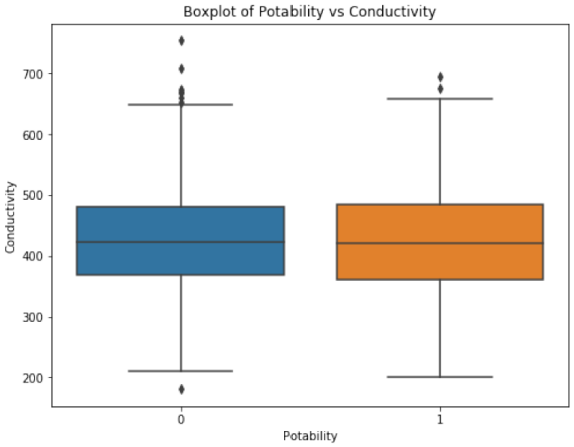
*Google, KDNuggets, More Performance Evaluation Metrics for Classification Problems You Should Know, Clare Liu, Fintech Industry on September 20, 2022 in Machine Learning, https://www.kdnuggets.com/2020/04/performance-evaluation-metrics-classification.html#:~:text=The%20key%20classification%20metrics%3A%20Accuracy,Receiver%20Operating%20Characteristic%20(ROC)%20curve*

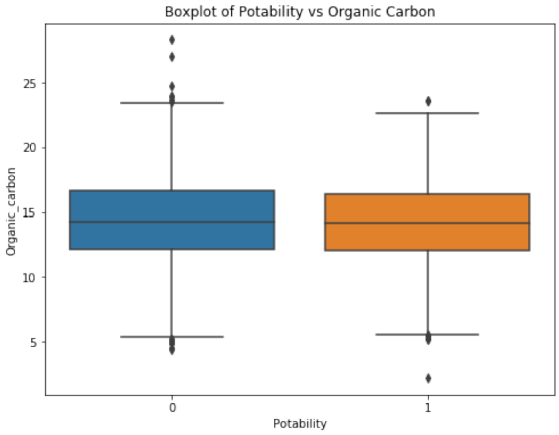
**Appendix –A :**

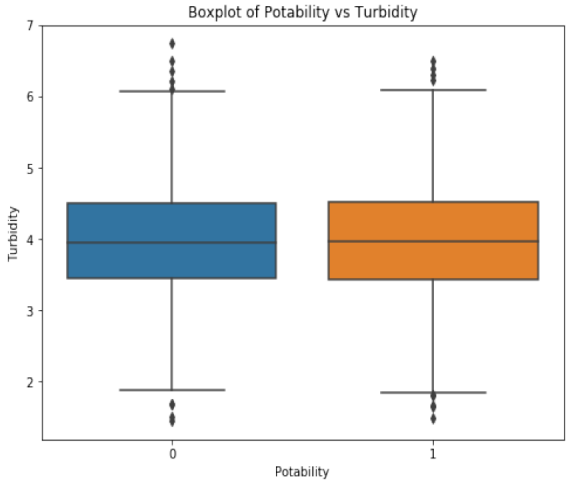
*Figure 6: Box Plots of the distribution of features with the target variable*



Project-2


Boxplot of Potability vs Conductivity
Conductivity
Potability
0
1
200
300
400
500
600
700


Boxplot of Potability vs Organic Carbon
Organic_carbon
Potability
0
1
5
10
15
20
25


Boxplot of Potability vs Turbidity
Turbidity
Potability
0
1
2
3
4
5
6
7

**Appendix – B:**

*Figure 7: Pair Plot of the water quality dataset.*

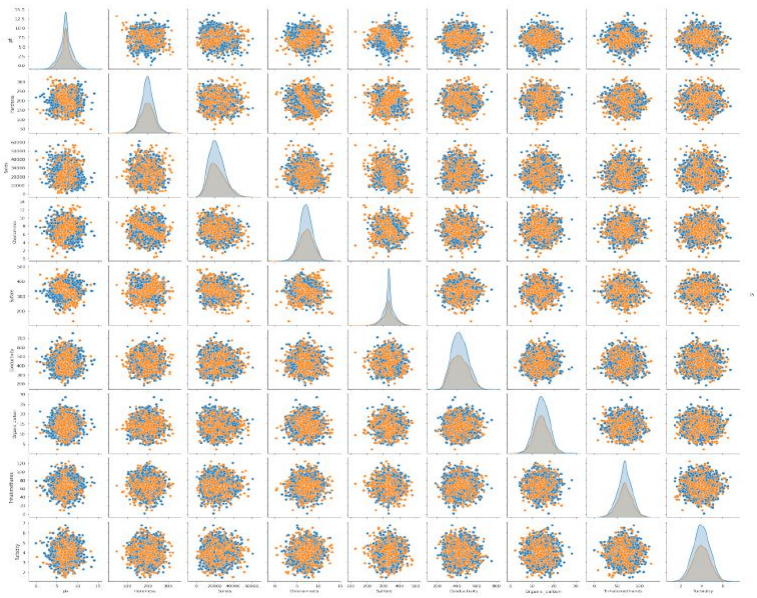
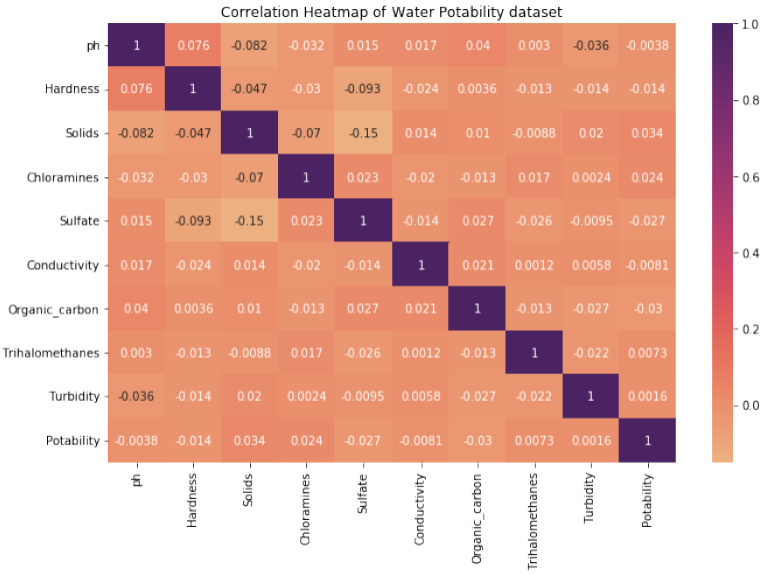
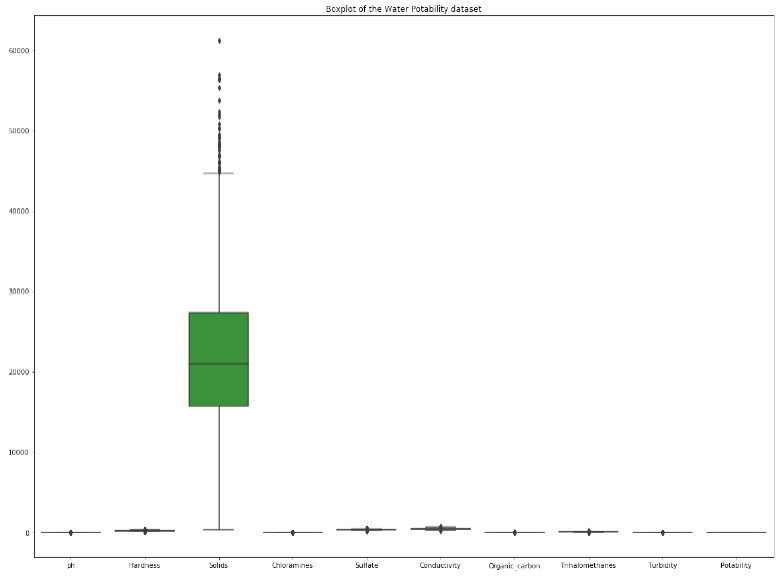


*Figure 8: Correlation heat map of the diabetes dataset.*



**Appendix – C:**

*Figure 9: Boxplot for identifying the outliers*



**Appendix – D:**

*Figure 10: Confusion Matrix for K – Nearest Neighbor Classifier*

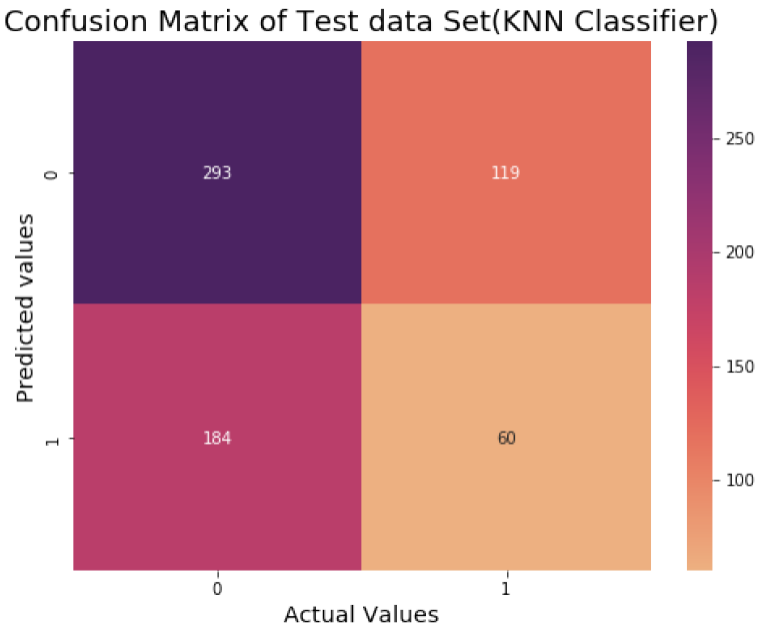



*Figure 11: Confusion matrix for Random Forest Classifier*

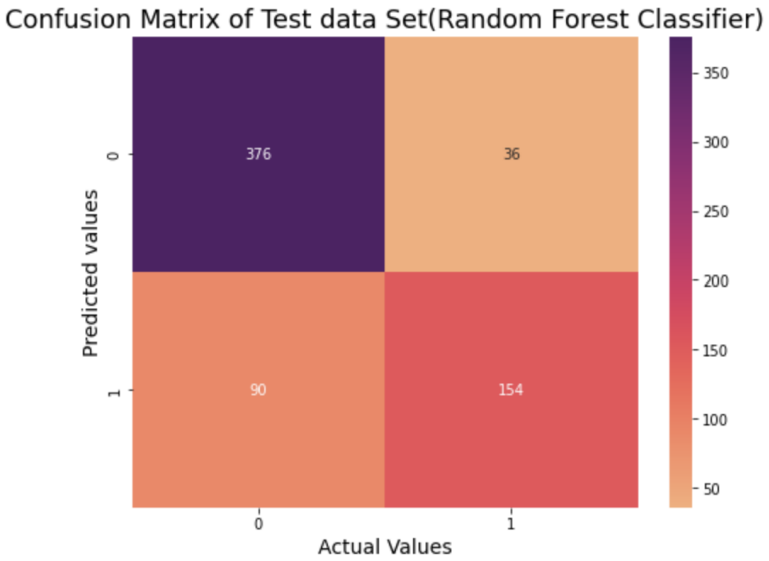
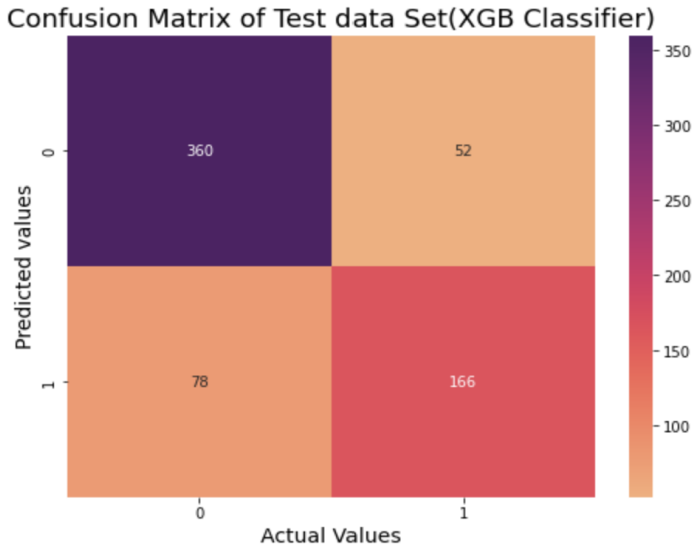


*Figure 12: Feature importance plot of XGB Classifier*



**Appendix – E:**

*Table – 1: Summary of the evaluation metrics of the models.*

| | KNN Classifier | RandomForest Classifier | XGB Classifier |
|---|---|---|---|
| **Model** | KNN Classifier | RandomForestClassifier | XGBClassifier |
| **Accuracy (test)** | 0.53811 | 0.807927 | 0.801829 |
| **Accuracy (train)** | 0.714504 | 1.0 | 1.0 |
| **Precision score** | 0.335 | 0.811 | 0.761 |
| **Recall score** | 0.246 | 0.631 | 0.68 |
| **F1 Score** | 0.284 | 0.71 | 0.719 |