**Water Quality Prediction Project Documentation**

1. **Problem Statement**

Issue of poor water quality is an age-old issue that most developing countries have battled with. According to Hirani and Dimble (2019) humans have battled with water quality for thousands of years. India as a developing country is not left out of the issues of water pollution which has become become a major national issue in the country. According to the Borgen magazine, “around 80% of India’s water is severely polluted because people dump raw sewage, silt and garbage into the country’s rivers and lakes.” Due to this situation, water has become undrinkable with many having to depend on illegal and expensive sources of water. Some images of rivers as shown below show how populated some waterbodies have become in India.



Figure 1: The Yamuna River, a tributary of the Ganges, is widely recognized as India’s most polluted river. Photo © J. Carl Ganter / Circle of Blue



Figure2:<https://blog.ipleaders.in/how-and-where-a-common-citizen-can-file-a-complaint-against-water-pollution-in-india/>

The problem of water pollution or decreased water quality has become very huge that the existence of biodiversity and multiple human communities have been threatened. This is such bad that almost 1.5 million children have been reported to have died due to catching diarrhea in India. A World Bank report also shows that population reduces the growth of an economy. Thus, a high increase in water pollution the higher loss in economic growth of the economy.

For these reasons it is important for the ministry in charge of water and sanitation as well as the major stakeholders in providing water to know the quality of the water bodies and the quality of water they produce for the consumption of the people of India.

Improved water supply and sanitation, and better management of water resources, can greatly reduce illnesses such as cholera, diarrhea, dysentery, hepatitis A, typhoid, and polio. Consequently, the country's economic growth will be boosted leading to poverty reduction.

So, being motivated by the above reasons stated, we use this Water Quality dataset collected for India to understand what constitutes to safe, Potable water and apply machine learning models to classify the water according to the water quality index. This model can be used as an aid to help experts quickly determine water quality and make decisions as to whether it is good for human consumption.

1. **Dataset**

The Water Quality dataset used was downloaded from Kaggle. It contained 534 records and 11 columns. Figure 3 shows a view of the first 23 records of the dataset

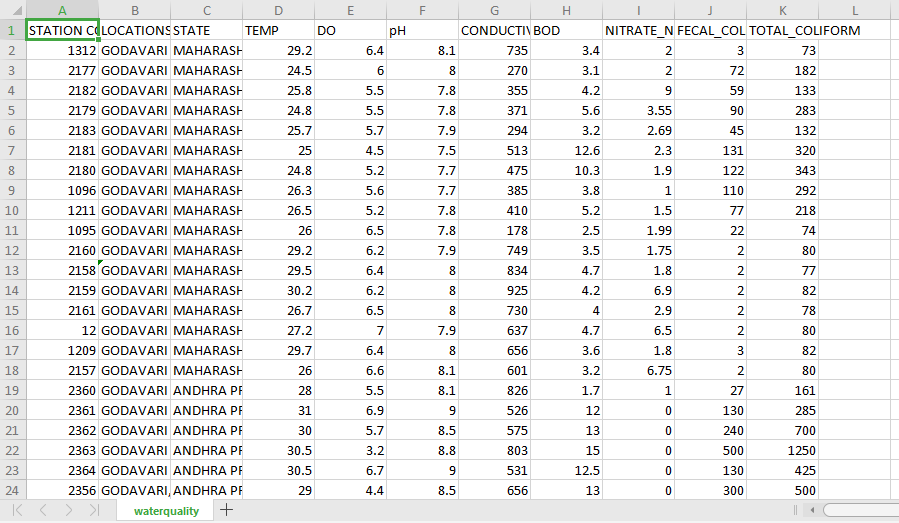


Figure 3: Dataset

The following are the description of the various features or variables in the dataset

1. STATION CODE: Code for the station where data was recorded
2. LOCATION: Location of the recorded data
3. STATE: State of the location where the data was recorded
4. TEMP: Temperature of the water in degree celcius.
5. DO: Dissolved Oxygen in water.
6. ph: pH of 1. water (0 to 14).
7. Conductivity: Electrical conductivity of water.
8. BOD: Biological oxygen demand is the amount of oxygen consumed by bacteria in water
9. NITRATES\_N\_NITRITE\_N: Nutrient nitrate in water.
10. FECAL\_COLIFORM: they may indicate the presence of other pathogenic bacteria in water and are considered a more accurate indication of animal or human waste than total coliforms
11. TOTAL\_COLIFORM: include all bacteria found in water that has been influenced by human and animal waste
12. **Solution**

We used a machine learning approach to train a model to make the predictions on the quality of water. Several models were trained and tested. Steps followed in reaching a conclusion are outlined as follows:

**Step 1: Exploratory Data Analysis**

We checked for missing values and tried to understand the data. Please check the dataset section for a description of the data that was used. There were missing values in the Temperature, Dissolved Oxygen, Biological Oxygen Demand, Nitrates, Fecal Coliform, total coliform and electrical conductivity features. Exploring the data, we discovered that these values were missing per the state in which the water sample was collected.

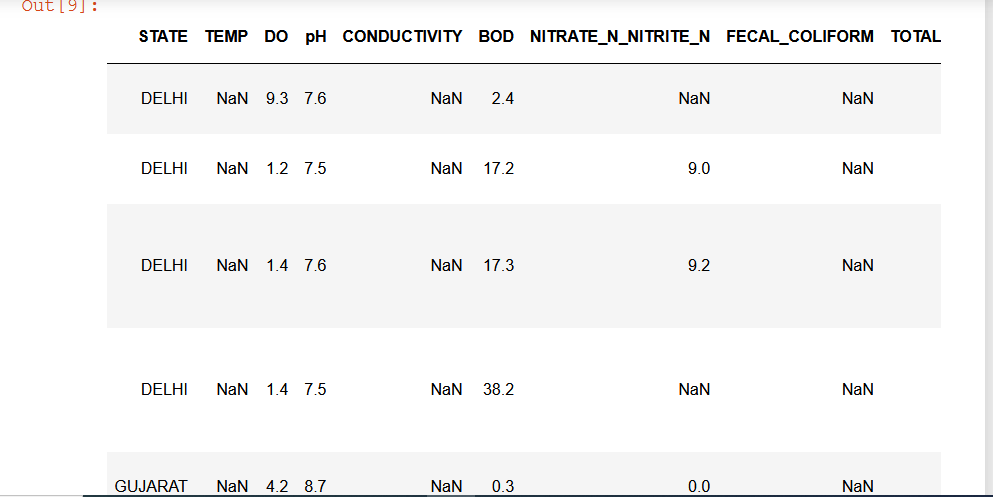


Figure 4: Some missing values in the Dataset

From Figure 4, Delhi and Gujarat states has missing values for some features. The problem of missing values were solved by filling them with the median of a parameter(feature) for a state. In other words, to fill the missing values for Temp feature, we calculated the median of Temp values for the Delhi state and used the median value obtained to fill the missing values. This was how all missing values filled. In cases where median was not possible or could not be calculated because the state did not have any recordings, the rows were dropped. Next we dropped all categorical variables like Location and State were dropped because it will not be needed in the prediction of water quality. Likewise, the Station code feature. Total Coliform variable was dropped due to the presence of fecal\_coliform variable. In the absence of feacal\_coliform, one may use the total coliform feature. Using both may cause the data to be biased.

**Step 2: Feature Engineering**

After cleaning the data, we realized that to predict water quality, we needed to calculate water quality index. Therefore, we feature engineered the water quality index column by using proved methods (brown et al WQI method) and approved standards according to World health Organisation (WHO), Bureau of Indian Standard (BIS) and ICMR standards. Using the water quality index, we were able to calculate the water quality for each record. Another feature was engineered by encoding the water quality which was string to numeric. The last two columns: WQI and Quality\_le in xxxx were the engineered columns.

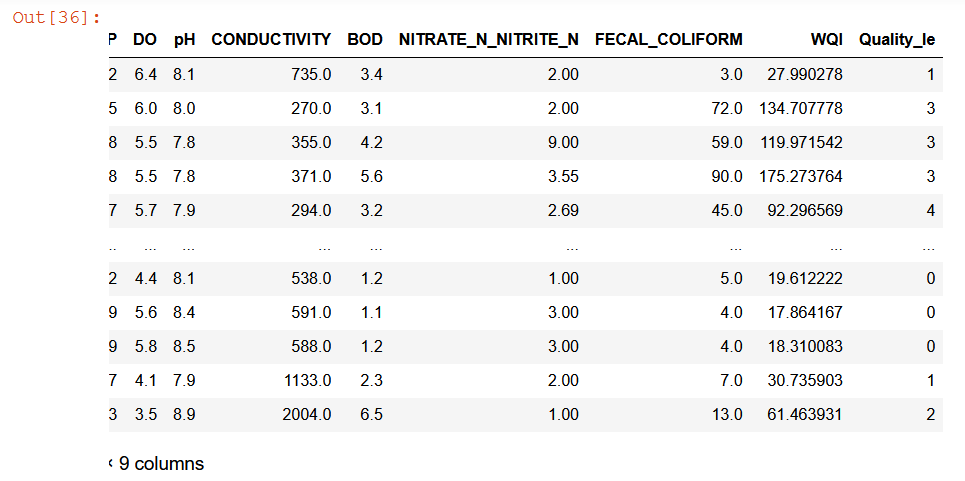


Figure 5: Feature Engineered columns WQI and Quality\_le

**Step 3: Model Training and Testing**

A linear regression model was trained to predict the water quality index. The output was passed to a classification model to predict the actual quality of the water. 80 percent of the data was used as training data and the remaining 20 as testing data.

After splitting the data, each was normalized using the standard scaler, the first model which is the linear regression model was used to predict the water quality index. The water quality index calculated or predicted categorized data into 5 classes namely Excellent, Good, Poor, Very Poor and Unsuitable.

For a classification model to finally classify the data into one of these classes, we trained and tested five models: A logistic Regression Model, A Support Vector Machine Model, A K-Nearest Neighbor Model, A Random Forest Classifier and a Gradient Boosting Classifier. These models were selected beacuse they support multiclass classification. The best Model was the Gradient Boosting Classifier with a precision score of 0.99. xxxx. The two models were saved and was used to test another dataset. And the models performed well.

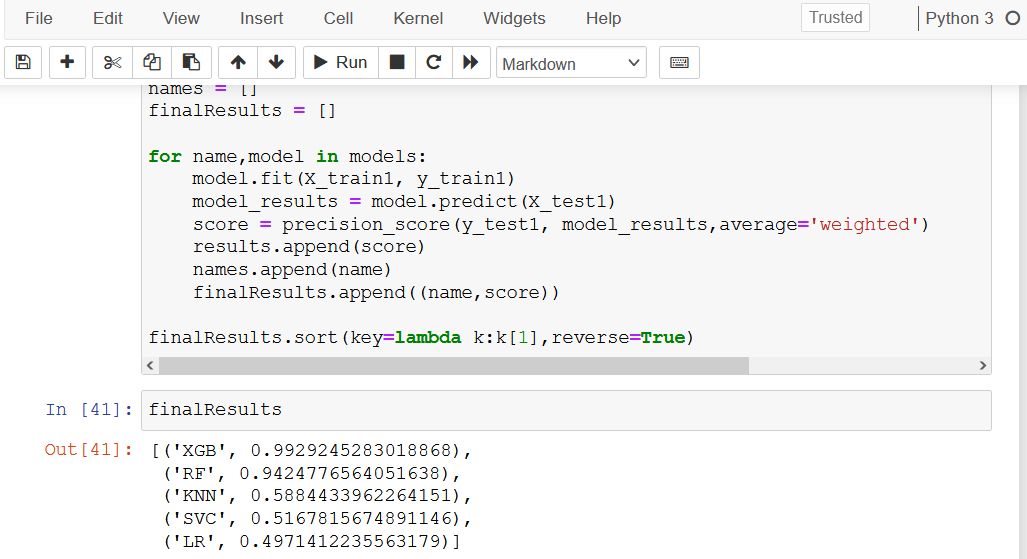


Figure 6: Results of testing the classification models selected

**Assumptions**

No assumptions were made

**Conclusion**

Machine learning is a great technique that can be used for predictions and help solve real life problems Several machine learning models were used to help predict water quality in India. Ultimately, the Gradient Boosting Classifier was identified as the best model. This model can be used to predict water quality in Ghana as well if data becomes available.