Water Quality Prediction

Project group members:

- ✓ Ashmita Gupta (A20512498)
- ✓ Sumanth Donthula (A20519856)

1. Project Proposal:

Safe drinking water is a fundamental human right, a necessity for good health, and a component of sensible health protection policies. At the national, regional, and municipal levels, this is crucial as a matter of health and development. Since the decreases in unfavorable health consequences and medical expenses outweigh the costs of implementing the interventions, it has been demonstrated that expenditures in water supply and sanitation can produce a net economic gain in some areas.

In addition, Potability refers to the suitability of water for drinking, and it is a key concern for public health and safety. This model can be used to improve the monitoring, analysis, and management of water quality data, which is essential for ensuring that water is safe for human consumption. For this project, we take Water Quality dataset www.cpcbenvis.nic.in/waterpollution site and would like to analyze the below:

- **Early detection of water contamination:** The model can be trained to detect contaminants in river water at an early stage, which can help prevent widespread contamination and protect public health
- **Water quality monitoring:** This model can be used to monitor water quality in real-time, allowing water treatment plants to take corrective action quickly if the water quality falls below acceptable levels
- **Environmental monitoring:** This model can be used to monitor the impact of human activities on the environment, such as the discharge of pollutants into water bodies, and provide early warning of potential environmental problems

2. What are the key questions we are trying to address with the machine learning model?

- Can we predict whether a water sample is potable or non-potable based on its physicochemical properties? This could involve developing a classification model that uses the measured water quality parameters as input features and predicts whether the water is potable or non-potable.
- Predicting which class does the water sample will fall into using input features and standards set by Government Agency. Knowing classes, we can make inferences on water quality.
- What are the most important factors that influence water quality? This could involve using feature selection or other analysis techniques to identify which water quality parameters have the greatest impact on potability

3. A proposed methodology/approach to the analysis that will be performed:

1. **Data Collection:** We collected water quality data that includes the values of environmental parameters (temperature, dissolved oxygen, pH, conductivity, BOD, Nitrate N + Nitrite N, Fecal Coliform, Total Coliform) from http://www.cpcbenvis.nic.in/water_quality_data.html

2. We had collected the corresponding water quality indicators (pollutant concentrations or water quality class labels). The below data has been obtained from Indian government agency (Central Pollution Control Board of India):

Designated-Best-Use	Class of water	Criteria
Drinking Water Source without conventional treatment but after disinfection	A	Total Coliforms Organism MPN/100ml shall be 50 or less
		pH between 6.5 and 8.5
		Dissolved Oxygen 6mg/l or more
		Biochemical Oxygen Demand 5 days 20C 2mg/l or less
Outdoor bathing (Organized)	B	Total Coliforms Organism MPN/100ml shall be 500 or less pH between 6.5 and 8.5 Dissolved Oxygen 5mg/l or more
		Biochemical Oxygen Demand 5 days 20C 3mg/l or less
Drinking water source after conventional treatment and disinfection	C	Total Coliforms Organism MPN/100ml shall be 5000 or less pH between 6 to 9 Dissolved Oxygen 4mg/l or more
		Biochemical Oxygen Demand 5 days 20C 3mg/l or less
Propagation of Wildlife and Fisheries	D	pH between 6.5 to 8.5 Dissolved Oxygen 4mg/l or more
		Free Ammonia (as N) 1.2 mg/l or less
Irrigation, Industrial Cooling, Controlled Waste disposal	E	pH between 6.0 to 8.5
		Electrical Conductivity at 25C micro mhos/cm Max.2250
		Sodium absorption Ratio Max. 26
		Boron Max. 2mg/l
	Below-E	Not Meeting A, B, C, D & E Criteria

3. **Data Preparation**: We will clean and preprocess the data to ensure that it is accurate, complete, and consistent. This may involve tasks such as removing missing values, handling outliers, normalizing the data, and splitting the data into training and testing sets.
4. **Feature Selection:** We will select the most relevant features (environmental parameters) for predicting the water quality indicators or class labels. This we will do using methods such as correlation analysis, feature importance ranking, or principal component analysis.

5. **Model Selection:** We will choose an appropriate machine learning algorithm for the classification, like logistic regression, naïve bias or decision trees and predicting a suitable model with optimal accuracy and metrics
6. **Model Training:** We will train the selected machine learning model on the training data, using techniques such as cross-validation, hyperparameter tuning, and regularization to optimize the model performance
7. **Model Evaluation:** We will evaluate the performance of the trained model on the testing data, using metrics such as accuracy, confusion matrix, ROC curve, and precision-recall curve. Compare the performance of different models and feature sets to identify the best approach.

4. A metric or set of metrics which will measure analysis results.

We will use four classification metrics, including Accuracy, Recall, Precision, and F1-score to test the classification model performance.

These metrics are expressed as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Recall = \frac{TP}{TP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

5. Software packages, applications, libraries, and associated tools, etc.

Libraries/Packages: ggplot2, islr, tidyhydat, rLakeAnalyzer, dplyr, mlr, plyr, caret, CRAN and some additional packages will be used for the project.
Software: RStudio, R
Project Management and Source Control: Central Pollution Control Board of India, Kaggle Website, and additional analysis pdfs

6. Data sources and reference data with descriptions:

Datasets Type(Water Quality Dataset (Year 2021) and Standard Class): Excel

Dataset Size(N) : 311 rows

Number of Dimensions(p) : 7

Feature Description:

1. **Dissolved Oxygen (mg/L):** Dissolved oxygen is the amount of oxygen that is dissolved in water. It is an important indicator of water quality because it affects the survival of aquatic organisms.

Water with low levels of dissolved oxygen can harm fish and other aquatic life, while high levels of dissolved oxygen can indicate a healthy aquatic environment.

2. **pH**: pH is a measure of the acidity or alkalinity of water. The pH scale ranges from 0 to 14, with 7 being neutral. Water with a pH below 7 is considered acidic, while water with a pH above 7 is considered alkaline. The pH of water is an important indicator of its suitability for drinking, irrigation, and aquatic life.
3. **Conductivity** (µmho/cm): Conductivity is a measure of the ability of water to conduct electricity. It is influenced by the concentration of ions in the water, such as salts and minerals. Conductivity is an important indicator of water quality because it can affect the taste, suitability for irrigation, and aquatic life.
4. **BOD (mg/L):** Biological Oxygen Demand (BOD) is a measure of the amount of oxygen that is consumed by bacteria and other microorganisms as they decompose organic matter in water. High BOD levels can indicate a high level of pollution in the water, which can harm aquatic life and make the water unsuitable for drinking.
5. **Nitrate N + Nitrite N(mg/L):** Nitrate and Nitrite are forms of nitrogen that are commonly found in fertilizers and other agricultural sources. High levels of nitrate and nitrite in water can indicate contamination from agricultural runoff and can pose a health risk to humans and animals.
6. **Fecal Coliform (MPN/100ml): Fecal** coliform is a group of bacteria that are found in the intestines of warm-blooded animals. High levels of fecal coliform in water can indicate contamination from sewage or animal waste, which can pose a health risk to humans and animals.
7. **Total Coliform (MPN/100ml):** Total coliform is a group of bacteria that are commonly found in soil, water, and on plants. While most types of coliform bacteria are not harmful, their presence in water can indicate that the water has been contaminated and may contain harmful pathogens.

Missing Values:

- Temperature ^{o}C: No Missing Values
- Dissolved Oxygen (mg/L): No Missing Values
- pH: No Missing Values
- Conductivity (µmho/cm): No Missing Values
- BOD (mg/L): No Missing Values
- Nitrate N + Nitrite N(mg/L) : Some Nan Values are present(35 count)
- Fecal Coliform (MPN/100ml) : Some Nan Values are present(10 count)
- Total Coliform (MPN/100ml) : Some Nan Values are present(9 count)

Other notes:

1)We got the data in CSV format we are converting into Excel and loading the data set in R

2)We might add more data as past years data was also present on site

7. Data processing and pipeline - cleaning, imputing, transformation, outlier detection, etc.

Step 1: Data Cleaning: We will clean the data to remove any duplicates, irrelevant or unnecessary columns, and rows with missing values

Step 2: Impute Missing Values: We will impute any missing values in the dataset and we will use median imputation.

Step 3: Feature Transformation: The next step we will take is to transform the features to make them more suitable for a classification model. This would involve techniques like scaling and encoding categorical variables.

Step 4: Outlier Detection and Handling: The next step we will perform is to detect and handle outliers with techniques such as boxplots, scatterplots, and z-scores.

Step 5: Feature Selection: We will find most relevant features for the classification model using correlation analysis

8. **Data stylized facts** – We will perform distributional analysis to examine the distribution of a dataset. For water quality prediction based on features such as temperature, dissolved oxygen, pH, conductivity, BOD, nitrate N + nitrite N, fecal coliform, and total coliform, distributional analysis will be used to understand the distribution of each feature in the dataset.
The distribution of a feature can be visualized using techniques such as histograms, kernel density plots, and boxplots. These visualizations can reveal the shape of the distribution, the presence of outliers or skewness, and the range of values for the feature.

9. **Model selection**– We will use classification model to predict the quality of water samples based on features such as temperature, dissolved oxygen, pH, conductivity, BOD, nitrate N + nitrite N, fecal coliform, and total coliform. The goal of the classification model is to predict the water quality based on the values of these features and categorize them under one class of water quality as per set standards

10. Literature review and related work - existing projects, references, papers, and relevant articles, etc:

Academic Papers:

- https://pdxscholar.library.pdx.edu/cgi/viewcontent.cgi?article=1002&context=reu_reports
- https://www.researchgate.net/publication/351077205_Efficient_Water_Quality_Prediction_for_Indian_Rivers_Using_Machine_Learning
- https://www.mdpi.com/2306-5338/9/5/92

Articles:

- https://www.datascience2000.in/2021/10/water-quality-prediction-using-machine.html
- https://www.researchgate.net/publication/361118196_The_Quality_of_Drinkable_Water_using_Machine_Learning_Techniques

- https://www.sciencedirect.com/science/article/abs/pii/S0022169419308194

Existing projects:

- https://www.kaggle.com/code/maujmishra/water-quality-index-prediction
- https://www.kaggle.com/code/imakash3011/water-quality-prediction-7-model/notebook

Reference Books:

- "Machine Learning Techniques for Water Quality Monitoring" by Yuanyuan Liu, Yongping Li, and Junfeng Ji
- "Machine Learning for Environmental Monitoring" by Michael G. Schrlau and Kalyanmoy Deb.
- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013. Print.