CAPSTONE PROJECT 2
Water Quality Prediction Project Proposal
By Bhavani Atyam- PGA36

1. Executive Summary:

The aim is to analyze water quality data access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection. This is important as a health and development issue at a national, regional and local level.

2. Problem Statement:

Content

The water_potability prediction contains water quality metrics for 3276 different water bodies.

1. PH value:

PH is an important parameter in evaluating the acid–base balance of water. WHO has recommended maximum permissible limit of pH from 6.5 to 8.5.

2. Hardness:

Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels.

3. Solids (Total dissolved solids - TDS):

Total dissolved solids in the water. 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, sulphates etc.

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.

5. Sulphate:

Sulphates are naturally occurring substances that are found in minerals, soil, and rocks.

6. Conductivity:

Pure water is not a good conductor of electric current rather's a good insulator.

7. Organic carbon:

Total Organic carbon content in the water.

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.

9. Turbidity:

Clarity of the water, measured by turbidity.

10. Potability:

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

- (0) Water is not safe to drink and
- (1) Water is safe to drink.

3. Data Sources:

Primary Data: This is data collected directly from first-hand sources for the specific purpose of the study. In the context of water Potability, primary data might include measurements like pH, hardness, and chlorine levels taken directly from water samples.

Secondary Data: This is data collected by someone else and repurposed for another analysis or study. Secondary data for water Potability might include public health records, previously published research, or environmental reports.

4. Methodology:

- **Data Preprocessing:** Handle missing values, outliers, and errors in the dataset.
- **Model Implementation:** Develop the chosen model and train it on the preprocessed data.
- **Model Training:** Train selected models using the training dataset, optimizing for early and accurate detection of water Potability.
- **Model Evaluation:** Use metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to assess the model's performance.

5. Expected Outcomes:

- Identifying key chemical properties that have a significant impact on water Potability.
- Developing a predictive model (e.g., using machine learning) to classify water samples as potable or non-potable.
- Insights into how different regions or sources influence water quality.

6. Tools and Technologies:

- Jupiter notebook: For data Preprocessing exploratory data analysis.
- Machine Learning Frameworks: Tensor Flow.
- Programming Language: Python.
- Data visualization: Matplotlib,Seaborn for model performance analysis

7. Risks and Challenges:

- **Missing Data:** Several features have missing values (e.g., Ph, Sulphate, Trihalomethanes), which need to be handled.
- **Data Imbalance:** If the Potability variable is imbalanced, it might affect the performance of predictive models.
- **Feature Correlation:** Some features might be highly correlated, potentially leading to multicollinearity.

8. Conclusion:

This project aims to develop a machine learning model for early detection of water Potability, the dataset provides an opportunity to build a predictive model to assess water quality, but handling missing data and ensuring model robustness against potential data imbalance will be crucial.