CONCEPT NOTE

Water Quality Data Analytics (SDG 6)

Concept of the Project

Access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection. The project aims to develop a system to monitor and predict the water quality using data analytics and machine learning. By leveraging data analysis tools and methodologies, the project seeks to propose actionable solutions that align with UN Sustainable Development Goal 6 (SDG 6) – water and sanitation. SDG 6 seeks to ensure safe drinking water and sanitation for all, focusing on the sustainable management of water resources, wastewater and ecosystems, and acknowledging the importance of an enabling environment.

Problem Statement

Water quality is a critical factor for public health, environmental sustainability, and economic activities. However, monitoring and maintaining water quality is a complex and resource-intensive process. Traditional methods of water quality monitoring are often reactive, involving manual sampling and laboratory analysis, which can be time-consuming, costly, and not always representative of real-time conditions. Additionally, the increasing pollution from industrial, agricultural, and urban sources poses significant challenges in ensuring safe and clean water for various uses. The project seeks to develop an advanced system that leverages data analytics and machine learning to monitor and predict water quality in real time.

Project Objective

The primary objective of the project is to develop a system to monitor and predict water quality using data analytics and machine learning. The specific objectives are:

- To collect and analyze water quality data from reliable sources.
- To develop predictive models for future water quality levels based on current data.
- To propose actionable solutions and policy recommendations to mitigate urban pollution.
- To assess the potential impact of these solutions on achieving SDG 6.

Data Sources

1. **Kaggle:** Water Quality dataset depicting drinking water probability on Kaggle.
2. **World Health Organization (WHO):** Water quality guidelines, parameters and reports.
3. **Government Websites:** Dataset from governmental organization like Central Pollution Control Board in India.

Content

The water_potability.csv file 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. 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. The length of time water is in contact with hardness producing material helps determine how much hardness there is in raw water. 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. These minerals produced un-wanted taste and diluted color in appearance of water. This is the important parameter for the use of water. 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. The principal commercial use of sulfate is in the chemical industry. 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. Generally, the amount of dissolved solids in water determines the electrical conductivity. 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.

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.

Features Description

1. ph: pH of water (0 to 14).
2. Hardness: Capacity of water to precipitate soap in mg/L.
3. Solids: Total dissolved solids in ppm.
4. Chloramines: Amount of Chloramines in ppm.
5. Sulfate: Amount of Sulfates dissolved in mg/L.
6. Conductivity: Electrical conductivity of water in μS/cm.
7. Organic_carbon: Amount of organic carbon in ppm.
8. Trihalomethanes: Amount of Trihalomethanes in μg/L.
9. Turbidity: Measure of light emitting property of water in NTU.
10. Potability: Indicates if water is safe for human consumption. Potable-1 and Not potable-0

*ppm: parts per million
μg/L: microgram per litre
mg/L: milligram per litre

Tools for Analysis

The following tools and technologies will be used for data analysis:

1. **Python:** For data cleaning, analysis, and visualization, using libraries such as Pandas, NumPy, Matplotlib, and Seaborn.
2. **Jupyter Notebooks:** For documenting the analysis process and visualizations.
3. **Scikit-learn:** For developing predictive models and machine learning algorithms.
4. **PowerBI:** For creating interactive dashboards and visualizations to present the findings.

Methodology

The project will be conducted in the following sections:

Section 1: Project Introduction

- Topic: Data Science Project on Water Quality

- Objective: To analyze and improve water quality using data science techniques.

Section 2: Data Collection and Preprocessing

- Data Sources: Water samples, sensor data, historical records
- Preprocessing Steps:
    - Cleaning: Removing duplicates, handling missing values
    - Normalization: Standardizing data for consistency

Section 3: Feature Engineering

- Feature Selection: Identifying relevant features such as pH levels, contaminants, temperature, etc.
- New Features: Creating new features like contamination index, seasonal variations

Section 4: Model Selection

- Models Used:
    - Linear Regression for continuous predictions
    - Classification models Decision Trees for categorical predictions
- Evaluation Metrics:
    - Accuracy for classification models

Section 5: Model Training and Evaluation

- Training Process: Split data into training and testing sets
- Hyperparameter Tuning: Using techniques like Grid Search or Random Search
- Model Evaluation: Assess model performance using evaluation metrics

Section 6: Results and Interpretation

- Model Performance: Compare different models based on evaluation metrics
- Insights: Key findings and patterns in the data

Section 7: Conclusion and Future Work

- Conclusion: Summarize the findings and their implications
- Future Work: Potential improvements and further research directions

Concept Interpretation

In the transcript, there are several technical terms and data science concepts mentioned amidst the conversation. Here's a detailed explanation and interpretation of some parts of the dialogue for better understanding:

1. Normalization and Data Distribution:

- **Normalization** is a process to transform data into a standard format, making it

easier to analyze and model. It often involves scaling numerical data to a range, such as 0 to 1.

- **Data Distribution** refers to how data points are spread across a range of values. The conversation discusses whether the data is normally distributed, which means data points form a symmetric bell-shaped curve around the mean.

2. Scatter Plot and Box Plot:

- **Scatter Plot** is a type of plot that shows individual data points plotted on a two-dimensional graph, often used to observe relationships between variables.
- **Box Plot** (or box-and-whisker plot) is used to display the distribution of data based on a five-number summary: minimum, first quartile, median, third quartile, and maximum. It helps identify outliers and the spread of data.

3. Training and Testing Data:

- **Training Data** is the subset of data used to train a model. The model learns the patterns and relationships within this data.
- **Testing Data** is a separate subset of data used to evaluate the model's performance. It helps ensure the model generalizes well to new, unseen data.

4. Confusion Matrix:

- A **Confusion Matrix** is a table used to evaluate the performance of a classification algorithm. It summarizes the counts of true positives, true negatives, false positives, and false negatives. It provides insights into the accuracy, precision, recall, and overall performance of the model.

5. Probability and Distribution:

- The conversation mentions various probabilities and distributions, implying a discussion about statistical measures and how data fits into different probabilistic models. This includes checking if data follows a normal distribution or any other statistical distribution.

6. Model Optimization and Feature Selection:

- **Model Optimization** refers to the process of tuning model parameters to improve its performance. This can involve hyperparameter tuning, cross-validation, and other techniques to enhance model accuracy.
- **Feature Selection** is the process of selecting the most important variables (features) that contribute to the model's predictive power. It helps reduce the complexity of the model and improve its performance.

Probable Outcome

The expected outcomes of the project are:

- **Comprehensive Analysis:** A detailed analysis of water quality data identifying key sources and trends of urban pollution.
- **Predictive Models:** Reliable models for predicting future water quality levels and assessing the impact of potential interventions.
- **Actionable Solutions:** Data-driven solutions and policy recommendations to reduce water pollution.
- **Impact Assessment:** Evaluation of the potential impact of proposed solutions on achieving SDG 6.
- **Awareness and Engagement:** Increased awareness among policymakers and the public about the sources and impacts of water pollution, and the benefits of proposed interventions.

By addressing water pollution through data analysis and evidence-based solutions, this project will contribute to creating sustainable and healthier environments, aligning with the objectives of SDG 6: Water and Sanitation.