WATER POTABILITY

NAME: HARSHIT KATARIYA

ROLL No: 210107033

DATE OF SUBMISSION: 30th March 2024

COURSE NAME: CL653

**PROJECT OVERVIEW**

INTRODUCTION:

Potable water, also known as drinking water, is water that's safe to drink and typically comes from surface and groundwater sources. It's treated to remove contaminants like microorganisms, bacteria, harmful chemicals, viruses, and fecal matter that can cause a variety of waterborne diseases. Access to safe drinking water is essential for human health and well-being. Unsafe drinking water is a major global health concern, affecting millions of people worldwide. Contaminated water can lead to illness, even death, and economic hardship. By providing clean water, this project can help to improve public health, reduce healthcare costs, and promote sustainable development.

OBJECTIVE:

1.Data Analysis and Insights: analyzing vast datasets on water quality, including historical data, sensor readings, and environmental factors. This analysis could help identify trends, predict potential contamination events, and pinpoint areas with the greatest need for intervention.

2.Optimizing Treatment Processes: training my model on data about various water treatment methods and their effectiveness in removing specific contaminants. This would allow me to recommend the most efficient and cost-effective treatment options based on the specific water source and its unique challenges.

**DESCRIPTION OF PROJECT**

Factors affecting water potability are:

1. pH: - The pH of water is an important metric in determining its acid–base balance. It can also be used to determine if the water is acidic or alkaline. The pH maximum allowable limit has been set at 6.5 to 8.5 by the WHO.

2. Hardness: - Calcium and magnesium salts are the main causes of hardness. These salts are dissolved in water as it passes through geologic layers. The length of time water spends in contact with hardness-producing material influences the amount of hardness in raw water.

3. Solids: - Potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulphates, and other inorganic and organic minerals or salts can all be dissolved in water. These minerals gave the water an unpleasant taste and a diluted tint. This is an important criterion for water usage. The presence of a high TDS value in water suggests that it is heavily mineralized. The desirable limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which prescribed for drinking purposes.

4. Chloramines: - The main disinfectants used in public water systems are chlorine and chloramine. When ammonia is added to chlorine to purify drinking water, chloramines are generated. Chlorine levels in drinking water up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are deemed safe.

5. Sulfate: - Minerals, soil, and rocks all include sulphates, which are naturally occurring compounds. They can be found in the environment, groundwater, plants, and food. Sulfate's main commercial application is in the chemical industry. Seawater contains roughly 2,700 milligrams per liter (mg/L) of sulphate. Most freshwater supplies have concentrations of 3 to 30 mg/L, while some geographic regions have significantly greater quantities (1000 mg/L).

6. Conductivity: - Pure water is a good insulator rather than a good conductor of electric current. The electrical conductivity of water improves as ions concentration rises. The electrical conductivity of water is usually determined by the number of dissolved particles in it. Electrical conductivity (EC) is a measurement of a solution's ionic process, which allows it to transmit current.

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.

8. Trihalomethanes:- THMs are compounds that can be found in chlorine-treated water. THM levels in drinking water vary depending on the amount of organic matter in the water, the amount of chlorine needed to treat the water, and the temperature of the water being treated. THM levels in drinking water up to 80 ppm are deemed safe.

9. Turbidity:- The amount of solid stuff in the suspended state determines the turbidity of water. It is a test that determines the quality of waste discharge in terms of colloidal matter by measuring the light emitting properties of water.

10. Potability :- Indicates whether or not water is fit for human consumption, with 1 denoting drinkable and 0 denoting unfit for human consumption.

Water sources worldwide face increasing threats from contamination due to Industrial wastes, agricultural runoff and poor infrastructure. These factors coupled with traditional monitoring methods can lead to Ineffective treatment of water and High operational costs.

Access to safe drinking water is a fundamental human right and essential for public health. Contaminated water has severe consequences such as health impacts, economic burden and environmental degradation.

This Project aims to address these challenges by leveraging AI/ML to create a more proactive and data-driven approach to water treatment.

**DATA SOURCE**

Description of Data Source(s)

1.Public Datasets: There are many public datasets available that contain water quality metrics.

2.Water Treatment Plants: Water treatment plants regularly test the water they treat for various chemical characteristics.

3.Environmental Agencies: Environmental agencies often monitor the quality of natural bodies of water and groundwater.

4.Research Studies: There are many research studies that involve collecting and analyzing water quality data.

Data Characteristics

1.Volume: The dataset is relatively small, with 3276 entries and 10 columns. It is using approximately 256.1 KB of memory.

2.Variety: The dataset contains a mix of numerical data (both float and integer types). The columns represent various attributes related to water quality, such as pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, turbidity, and potability. Some columns have missing values, indicating that the data collection process was not entirely complete or that some measurements were not available.

3.Velocity: The velocity of the data refers to the speed at which the data is being generated and processed. In this case, as we are dealing with a static dataset, the velocity is not applicable. If this were a real-time water quality monitoring system, the velocity could be high, with data being generated and processed in real-time.

**DESCRIPTION OF DATA**

Nature of Data:

The data provided is a static or steady-state dataset. This means that it is a snapshot of the water quality parameters at a certain point in time and does not change or update over time.

In the context of this project, a steady-state dataset means that we can perform exploratory data analysis, data cleaning, and modeling without worrying about the data changing during our analysis. However, it also means that our analysis and any predictive models we build will only reflect the state of the water quality at the time the data was collected.

If the water quality parameters change significantly over time or if the factors influencing water potability change, our analysis and models may not remain accurate or relevant. For example, if a new source of pollution emerges after the data was collected, our models may not account for its impact on water potability.

Data preprocessing:

1.Handling Missing Values: The 'ph', 'Sulfate', and 'Trihalomethanes' columns have missing values. We need to decide how to handle these.

2.Data Normalization: The columns in the dataset have different ranges. For example, 'Solids' has a much larger range than 'ph'. Some machine learning algorithms perform better when the features have the same scale, so we might need to normalize or standardize the data.

3.Outlier Detection: We might need to check for outliers in the data that could skew our analysis.

4.Feature Engineering: Depending on the type of analysis we're doing, we might need to create new features from the existing ones

**Strategies for AI/ML Model Development**

Given that the goal is to predict the potability of water, which is a binary classification problem (potable or not potable), we can consider the following machine learning models:

1.Logistic Regression: This is a simple and fast model for binary classification problems. It's a good starting point for this problem as it can provide a baseline for model performance.

2.Decision Trees: Decision trees can handle both numerical and categorical data, are easy to understand and interpret, and can model non-linear relationships. They can also provide feature importance, which can give us insights into which water quality parameters are most important in predicting potability.

3.Random Forest: This is an ensemble method that combines multiple decision trees to reduce the risk of overfitting and improve prediction performance. It can also handle both numerical and categorical data and provide feature importance.

General approach for this project:

1. Split the Data

2. Preprocess the Data

3. Choose a Model: RANDOM FOREST

4. Train the Model: We'll train the model using the training data.

5. Tune Hyperparameters

6. Evaluate the Model

**7.** Iterate: Based on the model's performance, we might go back and try a different model, preprocess the data in a different way, or engineer new features. This is an iterative process

EVALUATION METRICS

1.Accuracy: This is the simplest and most intuitive metric - it's the proportion of correct predictions out of all predictions. However, it can be misleading if the classes are imbalanced.

2.Precision: Precision is the proportion of true positive predictions (water is potable and we predicted it as potable) out of all positive predictions. This metric is important if the cost of a false positive (predicting water is potable when it's not) is high.

3.Recall (Sensitivity): Recall is the proportion of true positive predictions out of all actual positives. This metric is important if the cost of a false negative (predicting water is not potable when it is) is high.

4.F1 Score: The F1 score is the harmonic mean of precision and recall. It's a good metric to use if we need to balance precision and recall and if there is an uneven class distribution.

5.Area Under the ROC Curve (AUC-ROC): The ROC curve plots the true positive rate against the false positive rate at different threshold settings. The AUC-ROC measures the entire two-dimensional area underneath the entire ROC curve. AUC-ROC is a good metric when we need to balance sensitivity (true positive rate) and specificity (true negative rate), and it's not sensitive to imbalanced classes.

Validation Strategy:

Cross-Validation: We use cross-validation. The most common form of cross-validation is k-fold cross-validation, where the dataset is divided into 'k' subsets. The model is trained 'k' times, each time using a different subset as the test set and the remaining data as the training set. The final performance is the average performance across the 'k' iterations. This gives us a more robust estimate of the model's performance.

**DEPLOYMENT STRATEGY**

1.Integration with Existing Systems: The model needs to be integrated with the existing systems where the data comes from and where the predictions are used.

2.User Interface: Depending on who the end users are, we might need to develop a user interface for the model.

3. Model Maintenance and Updates: Over time, the model might become less accurate as the data and the underlying reality it represents change. To maintain the performance of the model, we need to regularly retrain it on new data. This could be done on a fixed schedule (e.g., every month), or it could be done based on monitoring the model's performance and retraining when the performance drops below a certain threshold.

4.Monitoring and Alerting: After the model is deployed, it's important to continuously monitor its performance to ensure it's working as expected. This could involve tracking key metrics and setting up alerts if the performance drops.

5.Documentation and Training: Finally, it's important to document the model, how it's used, and how it's maintained. This documentation should be kept up to date as the model and the systems around it evolve.

**Scalability and Performance Optimization**

SCALABILITY:

1.Use More Powerful Hardware: The simplest way to scale up is to use more powerful hardware. This could involve using a machine with more memory and a faster CPU, or it could involve using a GPU, which can dramatically speed up certain types of computations.

2.Use More Efficient tree-based algorithms like random forests and gradient boosting machines can handle large datasets more efficiently than algorithms like k-nearest neighbors or support vector machines.

3.Use Data Reduction Techniques: If the dataset is very large, we might be able to reduce its size without losing too much information. This could involve sampling a subset of the data, or it could involve using dimensionality reduction techniques like principal component analysis (PCA) to reduce the number of features.

Performance Optimization:

**1.Using algorithmic optimization like Feature engineering, Feature Selection,** Hyperparameter Tuning, Ensemble Methods etc.

2. Hardware Choices: Using more powerful hardware can allow for more complex models or larger datasets. This could involve using a machine with more memory, a faster CPU, or a GPU.

3.Software Solutions: Using more efficient software can also improve performance. This could involve using a more efficient implementation of an algorithm, using a library that takes advantage of specific hardware features, or using a platform that automates some of the optimization process.

**Use of open-source tools**

1.Scikit-learn: It provides a wide range of algorithms for classification, regression, clustering, and dimensionality reduction, as well as utilities for preprocessing data, tuning hyperparameters, and evaluating models.

2.Pandas: This is a library for data manipulation and analysis in Python. It provided data structures for efficiently storing large datasets and tools for manipulating these datasets.

3.NumPy: This is a library for numerical computing in Python. It provided data structures for storing large multi-dimensional arrays and matrices, and it provides functions for performing mathematical operations on these data structures.

4.Matplotlib and Seaborn: These libraries are used in creating visualizations in Python. Matplotlib provides a wide range of functionality for creating static, animated, and interactive visualizations in Python.

**Purpose and Use Case**

Application:

The model can be applied in a real-world scenario in the water treatment industry. Water treatment plants can use this model to predict the potability of water and take necessary actions if the water is predicted to be non-potable. This can help in ensuring the safety and health of the community by providing clean and safe drinking water.

Impact:

The potential impact of this project is significant. Access to clean and safe drinking water is a fundamental human right, and yet many communities around the world struggle with water quality issues. By accurately predicting the potability of water, this model can contribute to efforts to ensure safe drinking water for all.

Environmental agencies can use this model to monitor and protect the quality of natural water sources. This can help in preserving the health of ecosystems and can also contribute to long-term water security by identifying and addressing pollution issues.

**Conclusion**

The key points of the project are:

1.Data: The model will be trained on a dataset that includes various chemical characteristics of water, such as pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, and Turbidity.

2.Model: The specific machine learning algorithm to be used will be determined based on exploratory data analysis and model performance. Techniques such as feature engineering, feature selection, hyperparameter tuning, and ensemble methods will be used to enhance the model's performance.

3.Application: The model can be used in real-world scenarios by water treatment plants and environmental agencies to predict the potability of water and take necessary actions if the water is predicted to be non-potable.

4.Impact: The project has the potential to make a significant contribution to public health and environmental protection by providing a tool to accurately predict the potability of water based on its chemical characteristics.

The innovation of this project lies in the application of machine learning techniques to the important problem of water quality assessment. By accurately predicting the potability of water, the model can contribute to efforts to ensure safe drinking water for all, which is a fundamental human right and a key public health issue. The project is significant because it addresses a real-world challenge and has the potential to benefit both human health and the environment.