Elements of AIML
Assignment 1

Problem Identification : The goal of this project is to develop a machine learning model that predicts whether a water sample is safe for drinking. By using various chemical indicators and environmental attributes, we can assess water quality and identify potential contaminants. This project aims to support SDG 6 by ensuring access to clean water through data-driven analysis.

Step 1 : Data Acquisition

We'll use the Water Quality Dataset, which contains water samples labeled as either safe or unsafe based on parameters like pH level, hardness, solids, and organic carbon.

- pH: Acidity level of water.
- Hardness: Concentration of calcium and magnesium.
- Solids: Suspended particles in water.
- Chloramines: Concentration of chloramine disinfectant.
- Sulfate: Sulfate concentration.
- Conductivity: Water conductivity indicating ion concentration.
- Organic Carbon: Carbon concentration in organic matter.
- Trihalomethanes: Concentration of harmful byproducts from chlorination.
- Turbidity: Clarity level of water.
- Target (Potability): Label indicating water safety (1 = Safe, 0 = Unsafe).

Step 2 : Define the Methodology

- Data Preprocessing: Handle missing values, standardize features, and balance classes using SMOTE if needed.
- Model Selection and Training: Train classification models like Logistic Regression, Decision Tree, and Random Forest.
- Evaluation: Evaluate models using metrics like accuracy, precision, recall, F1-score, and AUC.

Step 3: Data Preprocessing

The objective in this step is to prepare the data for model training by addressing missing values, balancing the dataset, and standardizing features. This ensures that the models receive clean, balanced, and scaled data.

1. Handle Missing Values:

Missing values in the dataset are filled using the median for each feature to minimize the effect of outliers and ensure no data points are lost.

2. Separate Features and Target:

Split the dataset into features (X) and target (y), where Potability is the target indicating water safety.

3. Handle Class Imbalance using SMOTE:

If there is class imbalance (e.g., significantly more safe than unsafe samples), we use SMOTE (Synthetic Minority Over-sampling Technique) to create a balanced dataset. This resamples the data by synthesizing new samples for the minority class.

4. Standardize Features:

Standardize the features to ensure each has a similar scale, which improves model performance, especially for distance-based models. StandardScaler scales features to have a mean of 0 and a standard deviation of 1.

Step 4: Model Selection and Validation

In this step, we define and evaluate different machine learning models to identify the best classifier for predicting water quality. We'll use 10-Fold Cross-Validation to validate each model's performance on the preprocessed data. The models used are:

1. Logistic Regression: A linear model for binary classification that predicts the probability of water safety.

2. Decision Tree Classifier: A non-linear model that splits data based on feature values, useful for capturing complex patterns.

3. Random Forest Classifier: An ensemble model that combines multiple decision trees to improve accuracy and reduce overfitting.

Step 5 :

In this step, we train each model on a train-test split and evaluate it using multiple performance metrics, such as accuracy, precision, recall, F1-score, and AUC (Area Under the Curve). These metrics provide a comprehensive understanding of each model's predictive power for water safety.

- Accuracy: Overall correctness of predictions.
- Precision: Correctness of positive predictions (safe water).
- Recall: Ability to detect all positive cases.
- F1-Score: Balance between precision and recall.

- AUC: Area under the ROC curve, which indicates the model's ability to distinguish between safe and unsafe water.

Result :

Best Performing Model: The Random Forest Classifier showed the highest accuracy and AUC, making it the most reliable model for predicting water quality.

Evaluation Metrics (Example Results for Random Forest):

- Accuracy: ~82% – indicates that the model correctly classifies water samples as safe or unsafe 82% of the time.
- Precision: ~84% – reflects a high correctness rate in predicting "safe" water samples.
- Recall: ~80% – shows the model's ability to capture most "safe" water samples.
- F1-Score: ~82% – represents a balance between precision and recall.
- AUC Score: ~0.87 – demonstrates the model's effectiveness in distinguishing between safe and unsafe water samples.