WATER QUALITY ANALYSIS

Phase 1: Problem Definition and Design Thinking

Project Definition:

The objective of this project is to conduct a comprehensive analysis of water quality Data to assess itssuitability for specific purposes, with a primary focus on Determining its potability for drinking. The project encompasses multiple key Components, including defining clear analysis objectives, collectingrelevant water Quality data, designing effective visualizations, and building a predictive model. Below,we outline the project's objectives and the approach to achieving them:

Analysis Objectives

1. Assessing Water Potability: The primary goal is to determine whether the water is Potable ornon-potable based on defined standards. We need to categorize water Samples as safe for drinking or not.

2. Identifying Deviations from Standards: Identify and flag water samples that Deviate from
regulatory standards. This includes recognizing parameters that fall Outside acceptable ranges.

3. Understanding Parameter Relationships: Explore and understand the relationships Between various water quality parameters, such as pH, hardness, and solids content. This insight can aidin identifying the factors contributing to water potability.

Data Collection

To achieve our objectives, we will collect water quality data from the provided Dataset, which includes various parameters, such as pH, hardness, solids, and more. This dataset is crucial for our analysis and modeling efforts. It is essential to ensure That the data is complete, reliable, and well-documented.

Data set link: https://www.kaggle.com/datasets/adityakadiwal/water-potability

Visualization Strategy:

Visualizations play a vital role in conveying the results of our analysis effectively.

WeWill employ suitable tools and techniques to visualize different aspects of

the data:

1. Parameter Distributions: Histograms, density plots, and box plots will be used to Visualize the distributions of key parameters. This will help us understand the central Tendencies and spreadof the data.

2. Correlations: Heatmaps and scatter plots will be utilized to visualize the correlations Betweenvarious water quality parameters. This will assist in identifying dependencies And potential multicollinearity.

3. Potability Assessment: We will use bar charts or pie charts to represent the Proportion of potable and non-potable water samples. This visualization will provide A clear understanding ofwater safety.

Predictive Modeling:

To predict water potability accurately, we will implement machine learning Algorithms. Here's ourapproach to predictive modeling:

1. Feature Selection: We will select relevant features (input variables) from the dataset Based on their significance in predicting water potability. Feature engineering may be Employed to createnew informative features.

2. Data Splitting: The dataset will be divided into training and testing sets. This division Will allow us to train our model on one subset and evaluate its performance on Another, ensuring that ourmodel generalizes well.

3. Model Selection: We will experiment with different machine learning algorithms Suitable for binary classification tasks. Random Forest, Logistic Regression, and Support Vector Machines(SVM) are among the potential options.

4. Model Training and Evaluation: The selected model(s) will be trained on the Training dataset andevaluated using appropriate performance metrics such as Accuracy, precision, recall, F1-score, and ROC-AUC.

5. Hyperparameter Tuning: Fine-tuning of model hyperparameters will be performed To optimizepredictive performance.

6. Model Interpretability: We will seek to understand the factors contributing to Predictions byanalyzing feature importance scores and decision boundaries Generated by the model.

With a well-defined project scope and design thinking, we are poised to embark on a Structured anddata-driven analysis of water quality, ultimately providing valuable Insights into water potability andsafety. The subsequent phases of the project will Build upon this foundation, including data preprocessing, exploratory data analysis, Model development, and documentation.