PHASE 2 – INNOVATION:

TOPIC: WATER QUALITY ANALYSIS

1.Data Collection:

The dataset used in this project comes from Kaggle which is a high-fidelity database with five-star research rating in google scholar. There are nine training variables, such as pH, hardness, solids, chloramines etc… The dataset link is Water Potability Dataset

2. Data Preprocessing:

Once the dataset has been collected, it is important to clean it and prepare it for machine learning. This may involve removing missing values, outliers, and scaling the data.

3. Exploratory Data Analysis (EDA):

Use interactive data visualization tools to create an immersive and user-friendly EDA interface. There are a variety of EDA techniques that we can use, such as histograms, box plots, scatter plots, and heatmaps.

4. Feature Engineering:

This involves creating new features or deriving relevant features from existing ones.

5. Machine Learning:

- Split the dataset into training and testing sets for model evaluation.
- The model of machine learning can be divided into supervised learning, semi-supervised learning, and unsupervised learning. There are many kinds of algorithms for each learning mode, such as decision trees, naive Bayes, random forests, etc. Each algorithm model has its own advantages and disadvantages.
- Among the many machine learning methods, artificial neural networks and support vector machine algorithms became popular in machine learning due to their large processing data and fast calculation speed. Therefore, selecting the above two algorithms to judge the drinking ability of water resources is expected to better achieve the desired purpose.

6. Model Evaluation:

Once the model has been trained, we can evaluate its performance on the testing set. We can use a variety of metrics to evaluate the model's performance, such as accuracy, precision, recall, F1-score, and ROC AUC.

7. Model Tuning:

If the model's performance is not satisfactory, we can try tuning the hyperparameters of the model. Hyperparameters are parameters that control the learning process of the model.

8. Interpretation and Insights:

Once we have trained a satisfactory model, we can interpret the model results to understand which water quality parameters are most influential in determining potability.

9. Continuous Improvement:

We can implement a feedback loop for continuous model improvement. If possible, connect the model to water treatment facilities and use the feedback from water quality improvements to adapt the model in real-time.