Water Potability Prediction

Problem Statement

Ensuring access to clean drinking water is vital for health, constituting a fundamental human right and a key element of sound health protection policies. This issue holds significance at the national, regional, and local levels in terms of both health and development. The objective is to predict the potability of water based on given features.
The water_potability.csv file contains water quality metrics for 3276 different water bodies.

Dataset: waterpotability dataset can be downloaded from drive

https://drive.google.com/file/d/1YdQ8Uvc8KYy_u8pNQGoNHaaFH7wEBCCu/view?usp=sharing

1. **Import Libraries/Dataset**

   a. Download the dataset
   b. Import the required libraries

2. **Data Visualization and Exploration**
   a. Print 2 rows for sanity check to identify all the features present in the dataset and if the target matches with them.
   b. Comment on class imbalance with appropriate visualization method.
   c. Provide appropriate data visualizations to get an insight about the dataset.
   d. Do the correlational analysis on the dataset. Provide a visualization for the same. Will this correlational analysis have effect on feature selection that you will perform in the next step? Justify your answer. Answer without justification will not be awarded marks

3. **Data Pre-processing and cleaning**
   a. Do the appropriate pre-processing of the data like identifying NULL or Missing Values, if any, handling of outliers if present in the dataset, skewed data etc.Mention the pre-processing steps performed in the markdown cell. Explore a few latest data balancing tasks and their effect on model evaluation parameters.
   b. Apply appropriate feature engineering techniques for them. Apply the feature transformation techniques like Standardization, Normalization, etc. You are free to apply the appropriate transformations depending on your dataset's structure and complexity. Provide proper justification. Techniques used without justification will not be awarded marks. Explore a few techniques for identifying feature importance for your feature engineering task.

4. **Model Building**

   Split the dataset into training and test sets. Justify your choice of split. You may experiment with different splits to get the final split.

a. Build Model Development using Logistic Regression and Decision Tree

5. Performance Evaluation

b. Do the prediction for the test data and display the results for the inference. Calculate all the evaluation metrics. Comment on the performance of these models.

c. Comment on under fitting/overfitting/just right model. Justify your comment.