**ELEMENTS OF AIML**

**ASSIGNMENT 1:**

**NAME: Ritika Mathur**

**Sap ID: 500120608**

**Roll No.: R2142230375**

**The aim of this assignment is to create a Machine Learning solution for a problem related to an SDG, beginning by identifying a relevant issue and then acquiring data. The dataset is then cleaned and preprocessed along with handling issues like class imbalance. Multiple ML models have also been applied and validated through K Fold Cross Validation, and their performance is then compared to addresses the problem effectively.**

**AIM:**



**This project aligns with SDG 6**  Sustainable Development Goal 6 is a United Nations goal to ensure access to clean water and sanitation for all.



**Chosen dataset:** “**[Drinking\_Water\_Potability](https://www.kaggle.com/datasets/artimule/drinking-water-probability)**”

**🧷:** [**https://www.kaggle.com/datasets/artimule/drinkingwaterprobability**](https://www.kaggle.com/datasets/artimule/drinking-water-portability)

**GitHub Link:** <https://github.com/Mathuritika/Assignment-one/blob/main/Ritika_proj.ipynb>

**Introduction:**

Water potability is a major issue in the world due to the widespread contamination from industrial waste, inadequate sanitation, etc. which impacts both rural and urban areas. Unsafe drinking water poses serious risks to health like waterborne diseases cholera, dysentery and typhoid. These diseases lead to numerous preventable deaths each year. It is essential to address water potability issue for public health, sustainable development, and ensuring a safe water supply for the world and coming generations.

By developing a ML model using the “*drinking\_water\_potability”* dataset from Kaggle, I aim to create a tool that can predict the potability of water based on its chemical and physical parameters – pH, hardness, solids, Chloramines, Sulphates, Conductivity etc. This model helps identify unsafe water sources which will allow authorities to target areas that require urgent attention and hence allocate resources accordingly, ultimately providing safe drinking water for all.

**Methodology:**

* **Data Gathering and Preprocessing**: First, the dataset is searched from Kaggle and then preprocessed, which includes handling missing values and scaling features. Then, the features are analyzed for their importance and contribution to the prediction model.
* **Model Selection and Training**: ML models like KNN, Logistic Regression, and SVM are then evaluated. Eventually, the model which performs the best is chosen based on cross validation results.
* **Evaluation and Validation**: After model selection, the model performance is evaluated using metrics like accuracy, precision, recall, confusion matrix and AUC etc. for classification tasks. For regression, performance is assessed using MSE (mean squared error), (MAE) mean absolute error, etc. For clustering models, metrics like WCSS (within cluster sum of squares) and silhouette score are used.
* **Deployment and Use**: The model is then deployed to help identify the status of water potability in regions with varying water quality for a sustainable water management aligned with SDG 6.



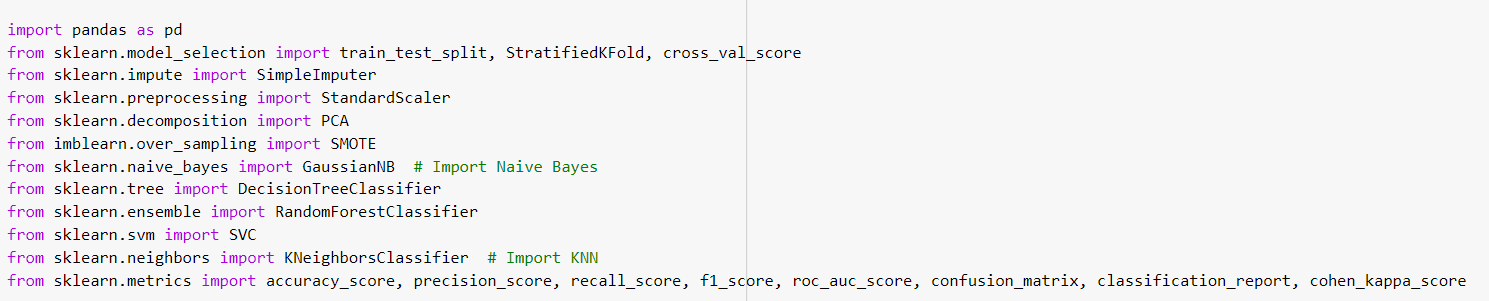
**Code Snippets and explanations:**

**1. Dataset downloading and preprocessing:**

First downloading the dataset using `kagglehub` and unzipping. The dataset contains features related to water quality and the target variable is "Potability" meaning whether the water is drinkable or not.

The data is then loaded to a Pandas dataframe, where I am at handling the missing values using `SimpleImputer`. This essentially replaces missing data with the mean of the columns respective of them to ensure no rows are left out due to missing values.

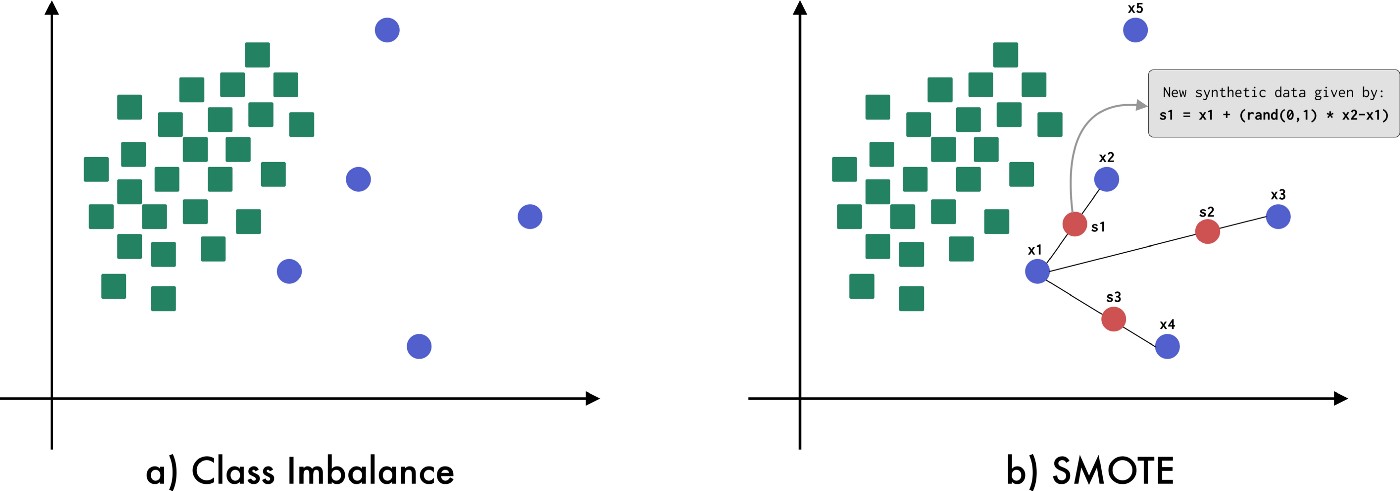


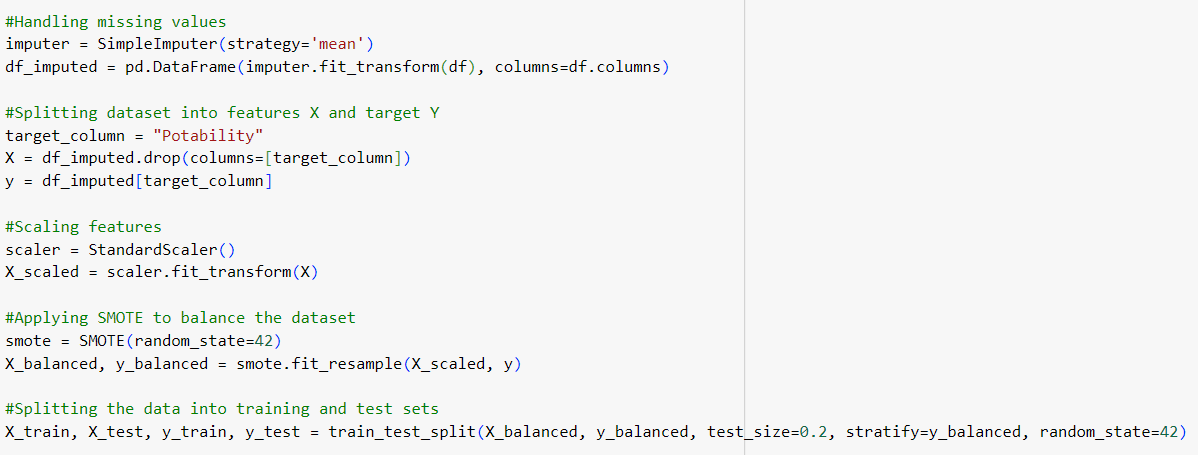


**2. Feature selection and scaling:**

Then splitting into features X and target Y. The feature matrix, X contains all columns excluding target - Potability and the target variable is stored in Y. Then scaling the features using `standard sclaler` to standardize them, essential for algorithms like SVM , KNN

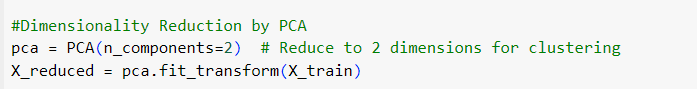
For class imbalance which is common for real world datasets, SMOTE which stands for Synthetic Minority Oversampling Technique is used. This creates artificial data points for the minority class, balancing the dataset by plotting new points in between two existing data points as shown in the figure below :





**3. Dimensionality reduction:**

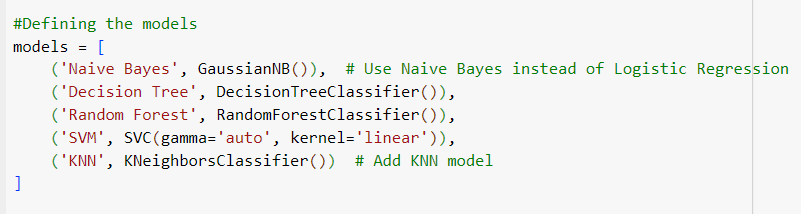
Then, applying PCA - Principal Component Analysis to reduce the dimensionality of the data to two dimensions for clustering done at the end. Hence, with reduced dimensions, we will be able to observe patterns in the dataset.



**4. Model selection:**

Classification models such as - Naive Bayes, Decision Tree, Random Forest, Support Vector Machine - SVM and K Nearest Neighbors – KNN are then defined.

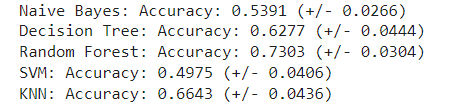
Which are then evaluated using Stratified K Fold crossvalidation (20 folds). This ensures the class distribution is preserved in each fold. Eventually, the accuracy is used as the performance metric during evaluation process.



**5. Model evaluation on test data:**

With an accuracy of about 73%, Random forest model performs the best based on cross validation. THence, choosing this model and then training on the full training set to evaluate on the test set. The evaluation metrics - accuracy, precision, recall, F1 score, AUC score, and Cohen's Kappa give an understanding to us of the model's performance.

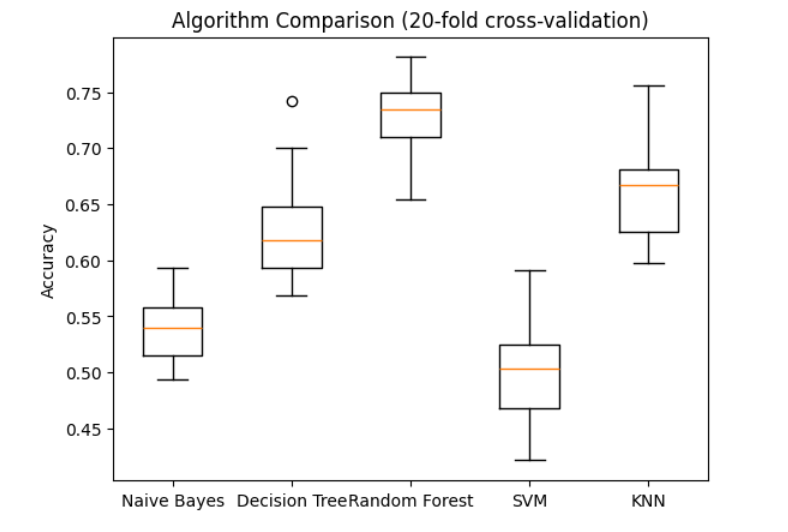
* Accuracy tells us the overall correctness of the model.
* Precision and Recall focus on the model’s ability to identify true positives and avoid false positives/negatives which becomes important in imbalanced datasets.
* F1Score provides a balanced measure that combines precision and recall.
* AUC evaluates the model’s ability to discriminate between the classes.
* Cohen’s Kappa accounts for agreement between predicted and true values

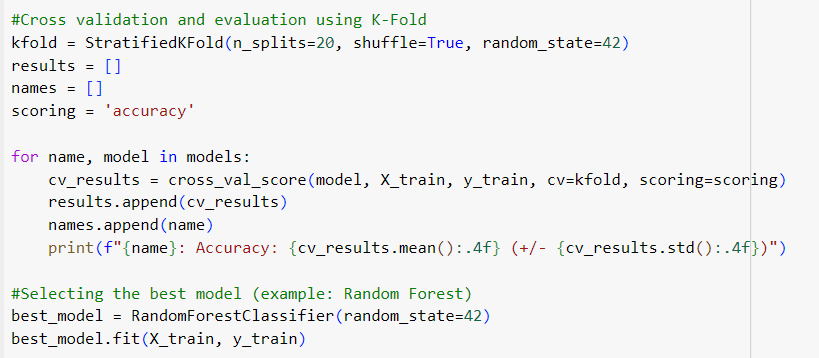


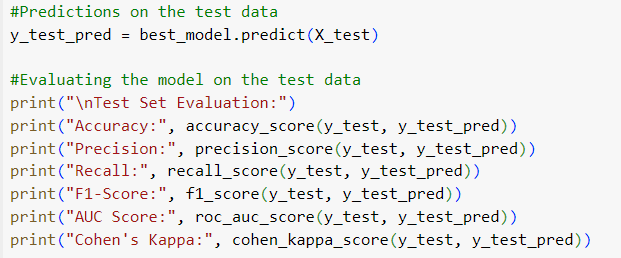
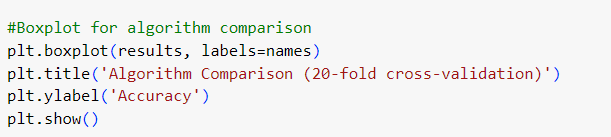
**6. Cross validation and performance evaluation:**

Results of every model are then plotted using a boxplot. This compares their performance and selects model that consistently performs well across the 20 folds, in a visual manner.

**Observation**: The best performing model in terms of accuracy and stability across cross validation folds is Random Forest model.





**7. Clustering evaluation:**

Then, KMeans clustering is applied to the reduced 2D dataset (from PCA) to evaluate the clustering performance. The Silhouette Score is used to assess the quality of the clustering. The silhouette score gives us an idea of how well separated the clusters are and how cohesive the points within each cluster are.

**Observations of Silhouette Score (Clustering)**: The Silhouette score means measure of how well separated and well formed the clusters are for k=3 is **0.3343**. A score closer to 1 means clusters are well separated and distinct.

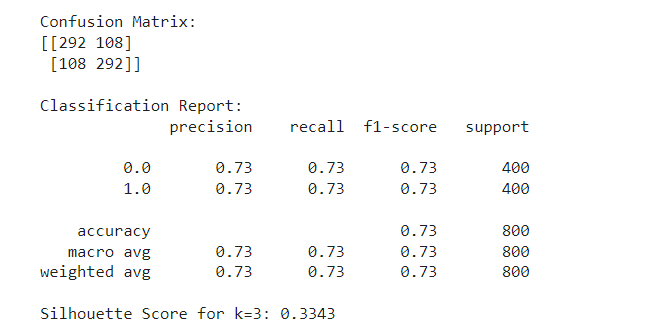
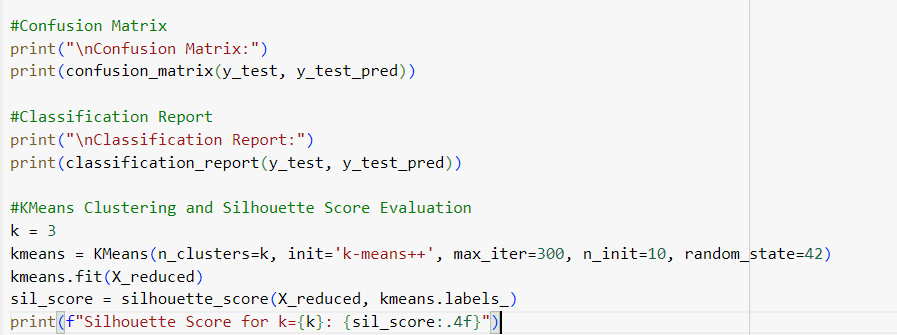
**Observations from the Confusion matrix and Classification reportClassification Report**:

1. **Confusion Matrix**:

* 292 instances of class 0 (water not potable) are correctly classified as 0 and 108 instances of class 0 are misclassified as class 1. Then, 108 instances of class 1 (water potable) are incorrectly classified as class 0. Also, 292 instances of class 1 are correctly classified as 1.

1. **Classification Report**:

* **Precision**: The model's precision is 0.73 for both classes, meaning that when the model predicts water as potable (class 1) or not potable (class 0), 73% of the time the prediction is correct.
* **Recall**: The recall is also 0.73 for both classes. This indicates that the model correctly identifies 73% of the actual potable or not potable instances.
* **F1 score**: The F1 score is 0.73 for both classes. This tells, there is a good balance between precision and recall.
* **Accuracy**: The model has an overall accuracy of 73%, which means it correctly predicts the water potability for 73% of the instances in the test dataset.
* A**UC**: Evaluates the model’s ability to discriminate between the classes. Hence, model hasmoderate ability to distinguish between potable and non-potable water.
* **Cohen’s Kappa:** Accounts for agreement between predicted and true values, accounting for chance. Its obtained value indicates a strong agreement between the predicted and actual values

**CONCLUSION:**

This model includes data preprocessing, model selection, evaluation, and visualization. By using different models and metrics, it tells about the effectiveness of classifiers in predicting water potability. The results suggest that ensemble methods like Random Forest provide the best performance in terms of accuracy and stability, while other models like Decision Trees or Naive Bayes perform poorly when compared. The clustering analysis with KMeans, using PCA for dimensionality reduction also gives a perspective on the dataset and helps in understanding the data distribution.

This is crucial in addressing **SDG 6 - Clean Water and Sanitation** as it contributes in ensuring safe water access by predicting water potability which is the main factor in improving health outcomes and sustainable water management.