Summary of Data Analysis and Modeling Process

1. ***Data Exploration and Preprocessing***:

- Loading the dataset.
- Checking basic statistics and distributions of the data including mean, standard deviation, and other descriptive statistics.
- Handling null values and duplicates.
- Generating descriptive statistics and visualizing data distributions using histograms and correlation matrices.
- Data normalization and standardization where necessary.

2. ***Feature Engineering:***

- Creating new features like Total_Nutrients, Temperature_Humidity, and Log_Rainfall based on existing data.
- Analyzing and selecting features based on their importance and relevance to the prediction goals.

3. ***Model Training:***

- Splitting the data into training and test sets.
- Training multiple machine learning models including Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Gradient Boosting classifiers.
- Utilizing techniques like cross-validation and grid search to fine-tune the hyperparameters of the models.

4. ***Model Evaluation***

- Evaluating models based on accuracy, precision, recall, F1-score, and other relevant metrics.
- Using confusion matrices to visualize the performance of each model.
- Comparing different models to select the best performer based on the test data.

5. ***Feature Importance and Optimization:***

- Analyzing the importance of each feature in the RandomForest model to understand their impact on the prediction.
- Optionally, removing less important features and retraining the model to see if performance improves.

6. ***Cluster Analysis***:

- We tried using the clustering approach as well, since all the predictors are numerical we tried using the KMeans approach. By plotting the WCSS plot we could observe the elbow point at 4, so we decided to go ahead with 4 clusters. Then we plotted the silhoutte plot using the silhoutte scores of each cluster, But the average silhoutte score is 0.498 which is a less amount. Furthermore we could observe that the number of observations in each cluster is also vary significantly as well. Also proportions for each class for each cluster is significantly differ as well but initially we had same number of observations in each class(200). Therefore we decided not to go ahead with the clustering approach.

7. ***PCA***

- When we observe the Heatmap of the variables we can see that there is considerable amount of correlation between several variables.(Total_Nutrients with P and K,Temperature_Humidity with Temperature and Humidity, P with K, log_rainfall with Rainfall)
- therefore we used pca to overcome the multicollinearity of the predictors, but the test accuracy does not incease by following this method.