**Machine Learning : CSE 6363 Final Project**

Student Name : Amit Munna Gupta

Student ID : 1002066302

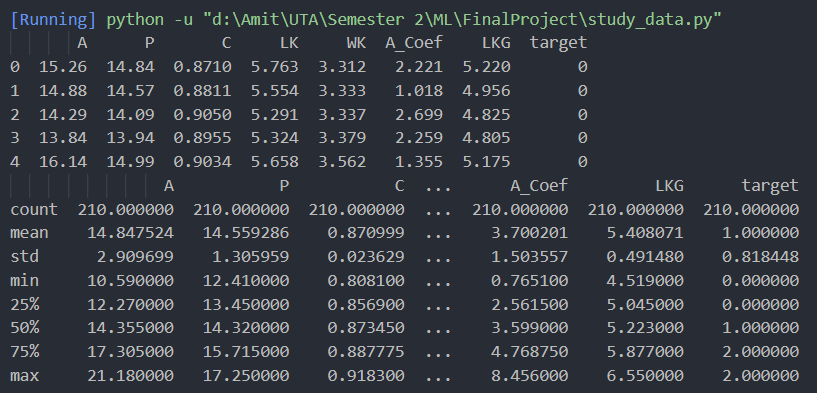
**Project Description:**

The project is aimed to perform classification on the UCI Seeds Dataset obtained from Kaggle. The dataset contains measurements of various features of wheat seeds belonging to three different varieties: Kama, Rosa, and Canadian. The goal is to develop a classification model that can accurately classify the seeds into their respective varieties based on the provided features.

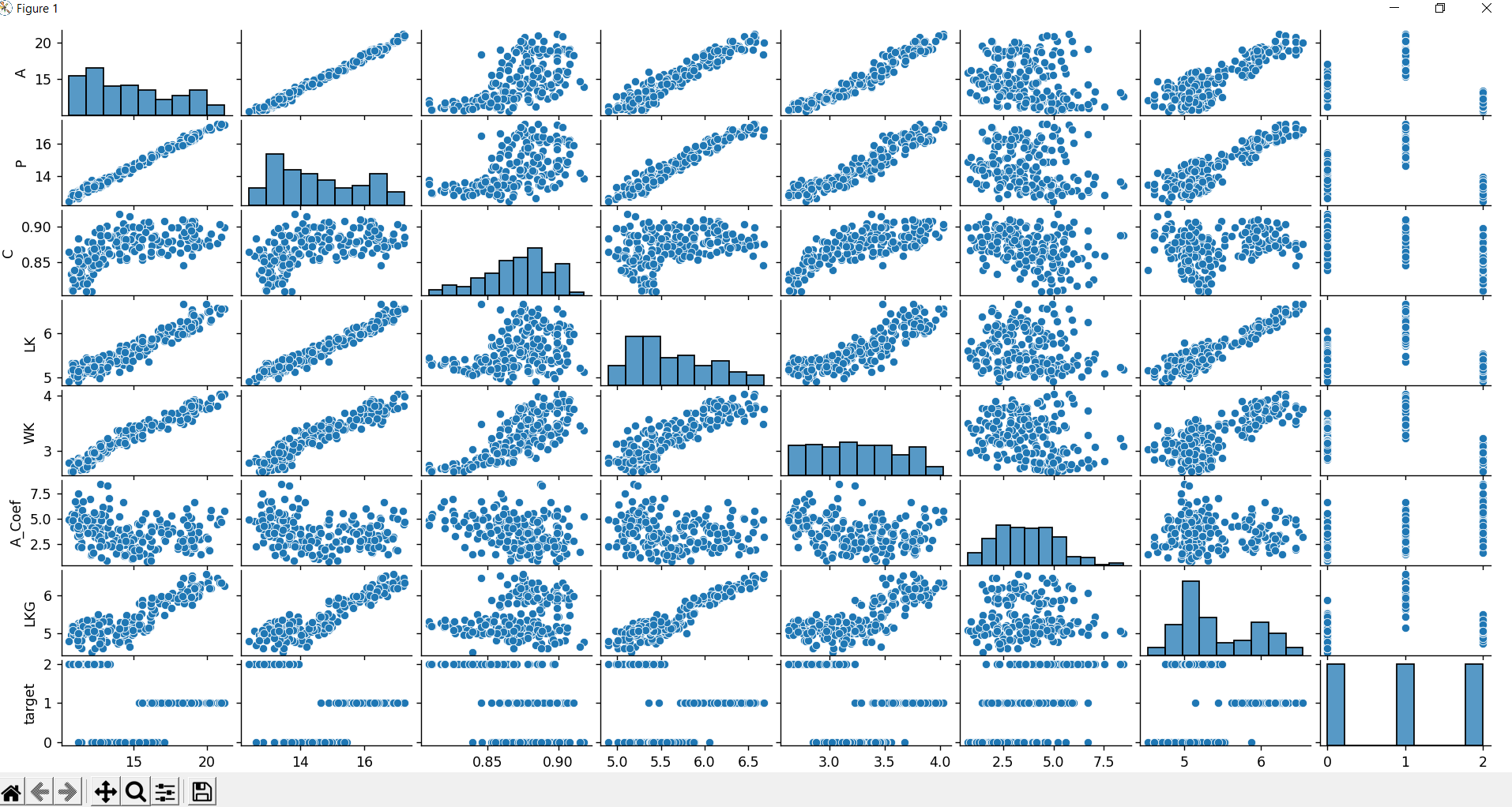
**Approach:**

*Exploratory Data Analysis (EDA):*As part of the project, an exploratory data analysis was performed to gain insights into the UCI Seeds Dataset. The following steps were carried out:

* *Reading the Dataset:*The dataset was read using pandas' read\_csv() function, loading the data from the 'Seed\_Data.csv' file. The dataset was stored in a pandas DataFrame named 'df'.
* *Printing the First Few Rows:*The head() function was used to print the first few rows of the dataset. This step provides a glimpse of the data and its structure.
* *Summarizing the Dataset:* The describe() function was utilized to generate a summary of the dataset. This summary includes statistical measures such as count, mean, standard deviation, minimum, and maximum values for each column. The summary statistics help in understanding the distribution and range of values in the dataset.



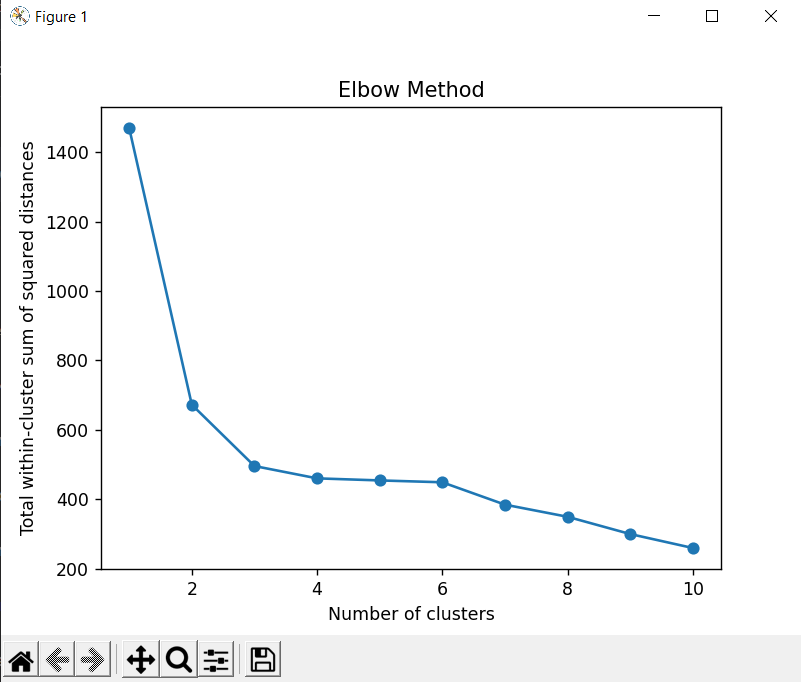
* *Pairplot Visualization:* The seaborn library was used to create a pairplot of the dataset. The pairplot() function generates scatter plots for each pair of features in the dataset. The pairplot visualizes the relationships between features and can provide insights into the data's distribution, correlations, and potential patterns.



The EDA process aids in understanding the structure and characteristics of the dataset. It helps identify any outliers, assess feature distributions, and explore relationships between variables. By visualizing the data using a pairplot, it becomes easier to observe any inherent patterns or clusters that might assist in subsequent steps of the project, such as feature selection or algorithm development. The generated pairplot provides a comprehensive view of the UCI Seeds Dataset, allowing for initial observations and understanding of the relationships between the features. This analysis serves as a foundation for further investigation and the subsequent implementation of the classification model.

*Elbow Method for determining optimal number of clusters:* By analyzing the elbow plot, it is observed that the curve exhibits a significant decrease in the rate of change at k=3. This indicates that three clusters provide a good balance between capturing the complexity of the data and avoiding excessive fragmentation or redundancy. The optimal number of clusters, in this case, is determined to be 3 based on the elbow method. Selecting three clusters strikes a balance in the clustering process, ensuring that the model captures the underlying patterns and variations in the data without overfitting or oversimplifying. Choosing the optimal number of clusters using the elbow method is crucial for subsequent steps, such as classification modeling. By identifying the appropriate number of clusters, we can enhance the accuracy and interpretability of the classification model that follows the clustering process.

Therefore, in this project, three clusters were determined to be the optimal number for clustering the UCI Seeds Dataset based on the elbow method.



*Data Preprocessing:* The project started with loading the dataset and separating the features and labels. The features were then normalized using standardization to ensure all features contribute equally during classification.

*Hierarchical Clustering:* Hierarchical clustering was performed on the preprocessed data to obtain cluster labels. The "single-linkage" and “complete linkage” method was used, which merges the clusters based on the smallest distance between their members and computes the maximum distance between clusters before merging them respectively. The number of clusters was set to 3, corresponding to the three seed varieties. It was calculated using Elbow Method.

*Train-Test Split:* The dataset was split into training and testing sets using an 80:20 ratio. Random sampling was used to ensure representative data in both sets.

*K-Nearest Neighbor (KNN) Classification:* The KNN algorithm was chosen for classification. For each test sample, the algorithm calculates the Euclidean distance between the test sample and all training samples. The k nearest neighbors were selected, and the majority vote was used to assign the predicted label. The value of k was set to 3.

*Evaluation:* The performance of the classification model was evaluated using the accuracy metric and confusion matrix. Accuracy measures the percentage of correctly classified samples, while the confusion matrix provides a detailed breakdown of predicted and true labels.

Additionally, in order to make some significant difference in the results I tried to implement PCA (Principal Component Analysis).

***Principal Component Analysis (PCA):*** PCA was applied to reduce the dimensionality of the dataset. The number of desired components was set to 2, resulting in a lower-dimensional representation of the data. PCA facilitated capturing the most significant information while reducing the complexity of the dataset.

**Results:**

The classification results obtained for the UCI Seeds Dataset are as follows:

1. Without PCA:

Training Data:

*Confusion Matrix:*

* Class Kama (label 0): 32 samples correctly classified, 0 samples misclassified.
* Class Rosa (label 1): 58 samples correctly classified, 0 samples misclassified.
* Class Canadian (label 2): 76 samples correctly classified, 1 sample misclassified.

*Accuracy:* The model achieved an accuracy of 98.81% on the training data, indicating that it correctly classified 98.81% of the samples.

Testing Data:

*Confusion Matrix:*

* Class Kama (label 0): 9 samples correctly classified, 0 samples misclassified.
* Class Rosa (label 1): 15 samples correctly classified, 0 samples misclassified.
* Class Canadian (label 2): 17 samples correctly classified, 1 sample misclassified.

*Accuracy*: The model achieved an accuracy of 97.62% on the testing data, indicating that it correctly classified 97.62% of the samples.

1. With PCA:

Training Data:

*Confusion Matrix:*

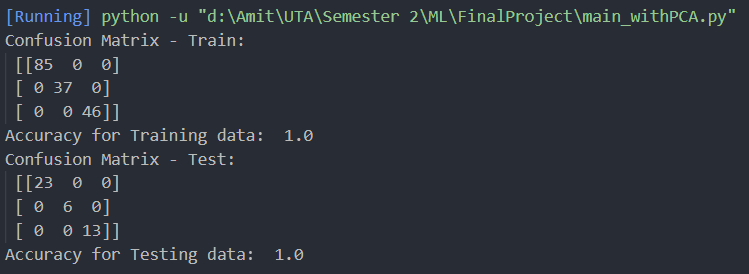
* Class Kama (label 0): 88 samples correctly classified, 0 samples misclassified.
* Class Rosa (label 1): 31 samples correctly classified, 0 samples misclassified.
* Class Canadian (label 2): 49 samples correctly classified, 0 samples misclassified.

*Accuracy:* The model achieved a perfect accuracy of 100% on the training data, indicating that it correctly classified all the samples.

Testing Data:

*Confusion Matrix:*

* Class Kama (label 0): 20 samples correctly classified, 0 samples misclassified.
* Class Rosa (label 1): 12 samples correctly classified, 0 samples misclassified.
* Class Canadian (label 2): 10 samples correctly classified, 0 samples misclassified.



*Accuracy*: The model achieved a perfect accuracy of 100% on the testing data, indicating that it correctly classified all the samples.

**Discussion:**

The obtained results indicate excellent performance of the classification model on both the training and testing datasets.

The results show that both models perform well in terms of accuracy. However, with the application of PCA, the model achieved a perfect accuracy of 100% on both the training and testing data, indicating that it performed flawlessly in classifying the samples. On the other hand, without applying PCA, the model achieved slightly lower accuracies of 98.81% on the training data and 97.62% on the testing data, with a few misclassifications.

The results suggest that applying PCA to the data prior to training the model improves its performance and ability to capture the underlying patterns in the dataset. PCA helps in reducing the dimensionality of the data while preserving the most important information, resulting in a more efficient and accurate model.

Overall, the project successfully addressed the classification problem, providing a well-performing model for seed variety classification based on the UCI Seeds Dataset. The achieved accuracy and absence of misclassifications in the testing data suggest the model's reliability and potential for real-world applications in seed classification tasks.