**ML ASSIGNMENT -5**

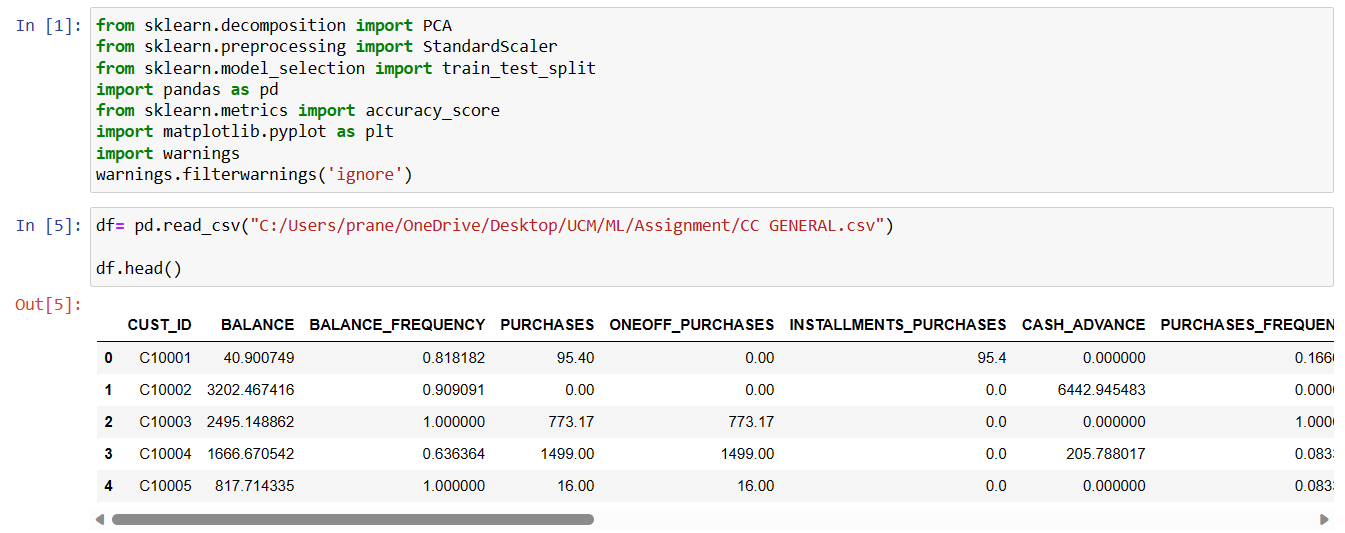
1. **Principal Component Analysis**

**a. Apply PCA on CC dataset.**

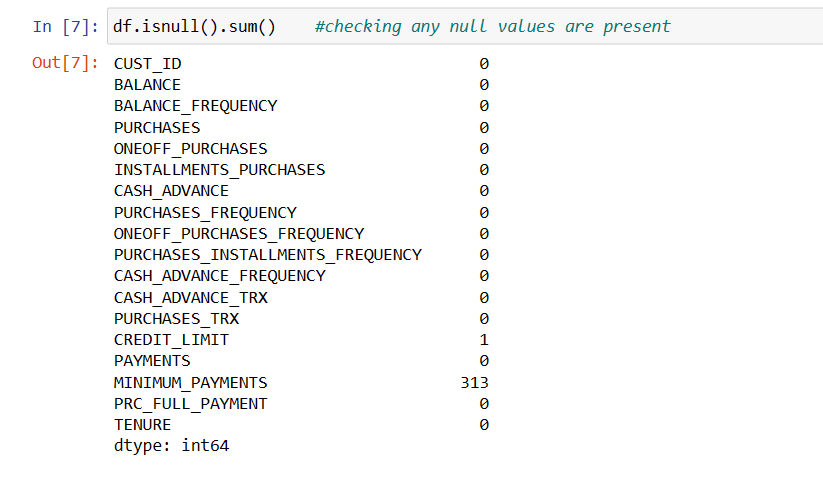
We need to install few python modules to perform data analysis and apply machine learning algorithms to data.

Data set "CC" was imported using the read\_csv

The top rows of a data set are returned by the pandas library's head () method.

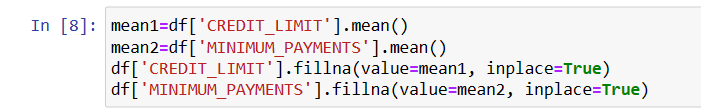


isnull() method looks for any values in the data collection.



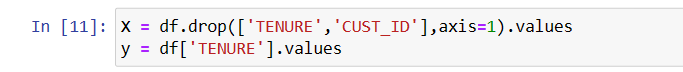
There are a few null entries in the minimum payment and credit limit columns of this data collection.

In order to replace these null values, the fillna() method uses the column mean value.



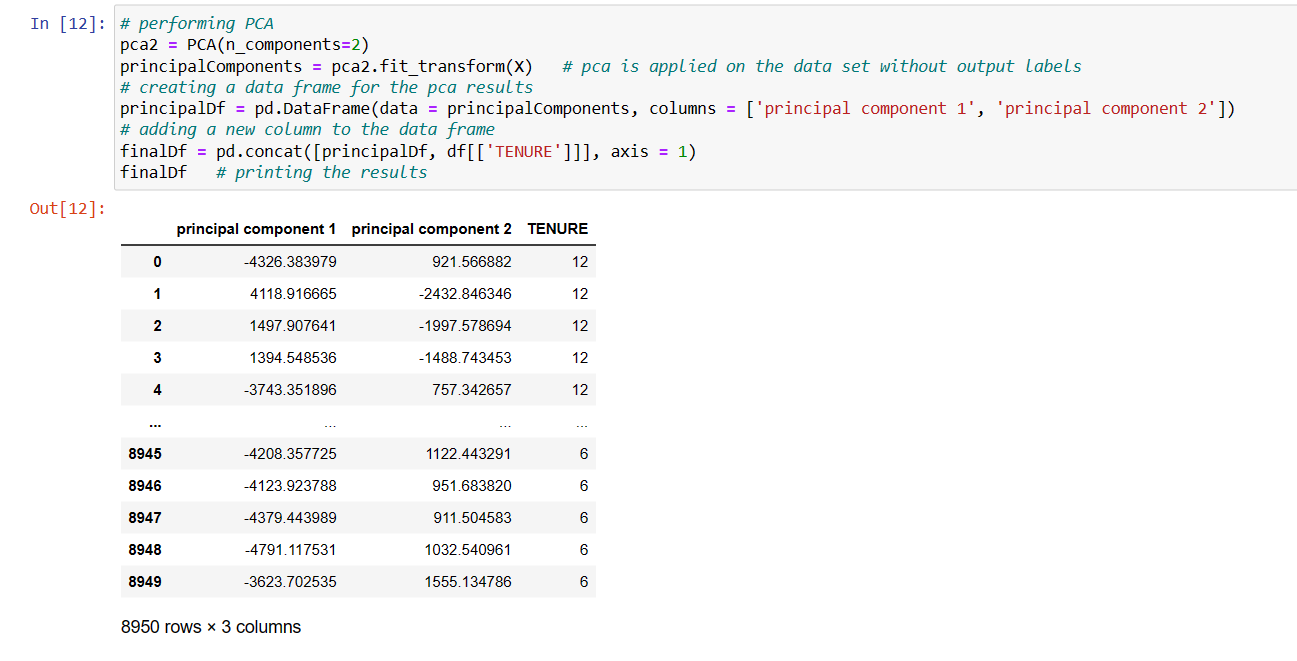
We don't need the output labels to run PCA on this data set

We eliminated a few unneeded columns by using the drop method.



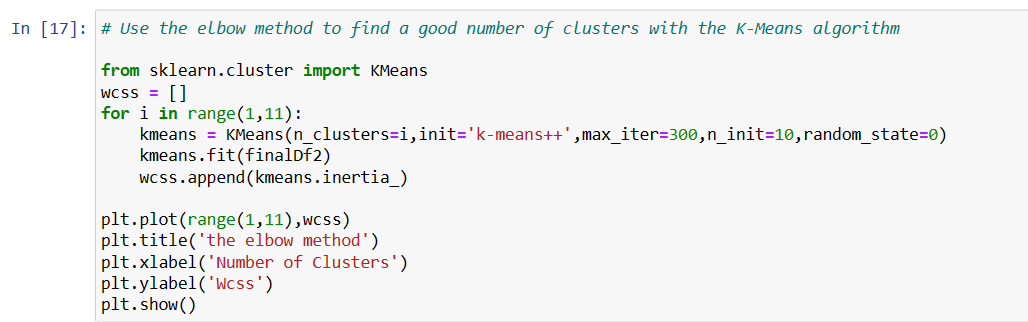
To run PCA on the data set, we imported the PCA method from the Sklearn Python package. PCA results in a data frame with features having highest variance with other features by ignoring the duplicate features.

Here, we kept the k value at 2, which reduced the data's dimension to two components.

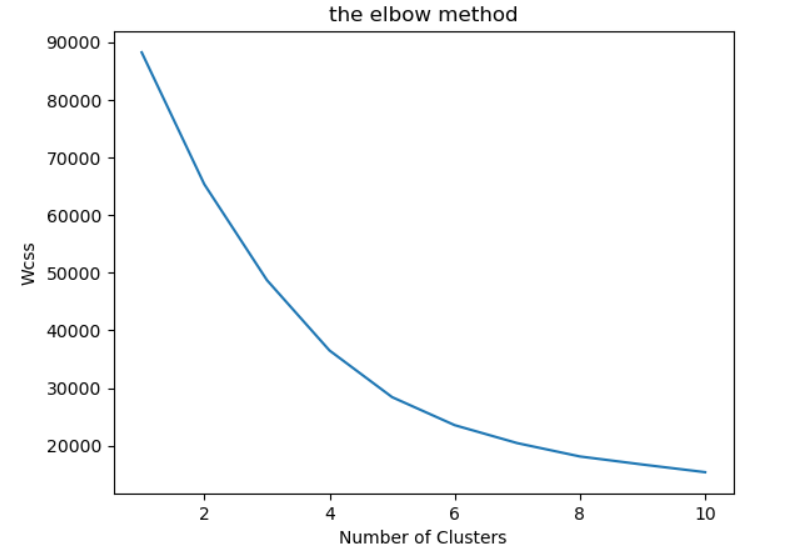


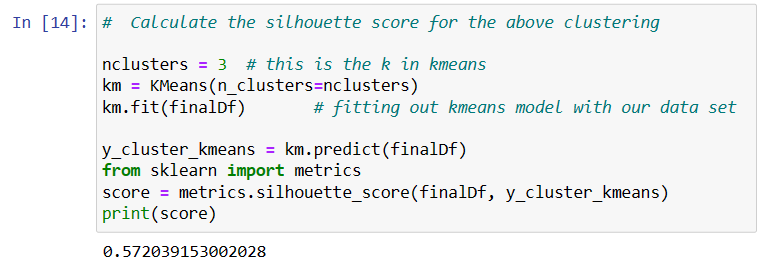
1. Before running the k-means algorithm on a data set, we must use the elbow approach to determine how many clusters are necessary to put our data into.

This elbow approach generates a graph from which we must determine the k value, or the number of clusters.

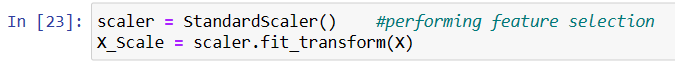


According to the graph below, the wcss value begins to decrease linearly once there are two clusters. Consequently, 3 clusters, or a k value of 3, are needed to suit our data. The number of clusters in the k-means algorithm is k.

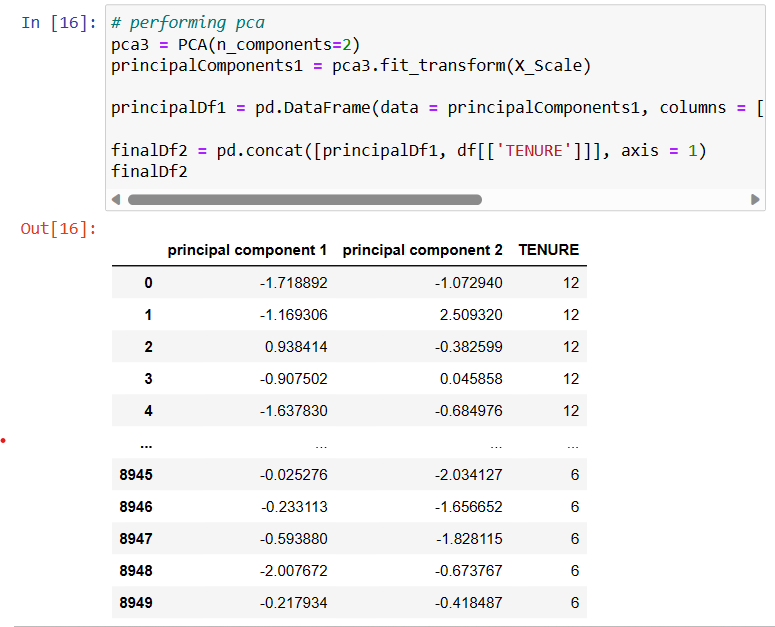




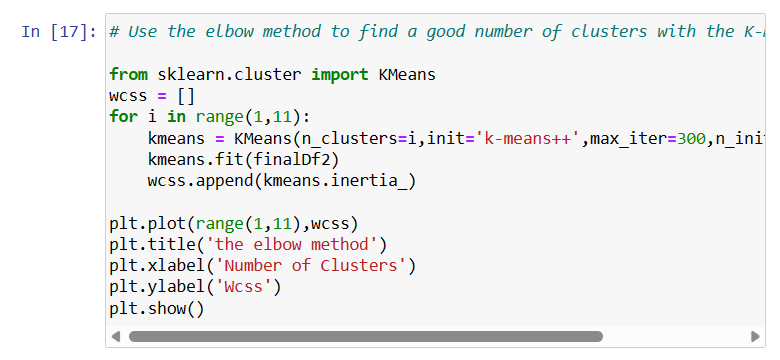
c.



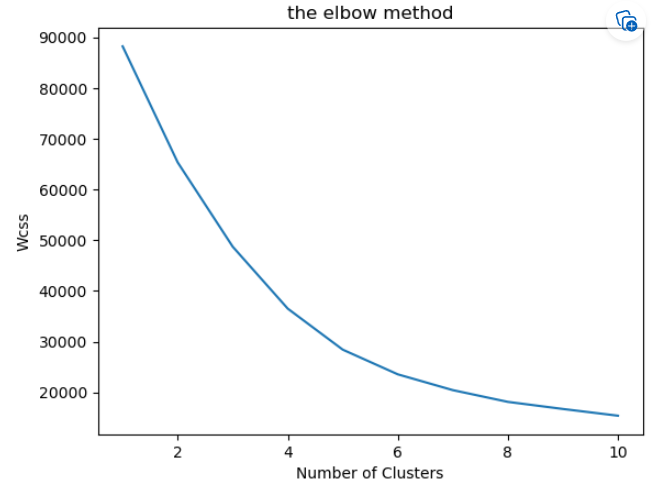
We are performing PCA on the feature scaled data set using the PCA method.



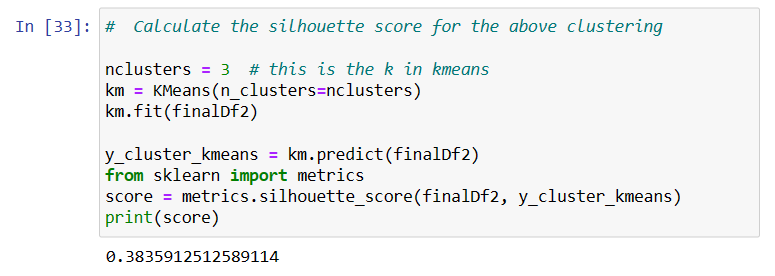
Before running the k-means algorithm on a data set, we must use the elbow approach to determine how many clusters are necessary to put our data into.



According to the graph below, the wcss value begins to decrease linearly once there are two clusters. Consequently, 3 clusters, or a k value of 3, are needed to suit our data.

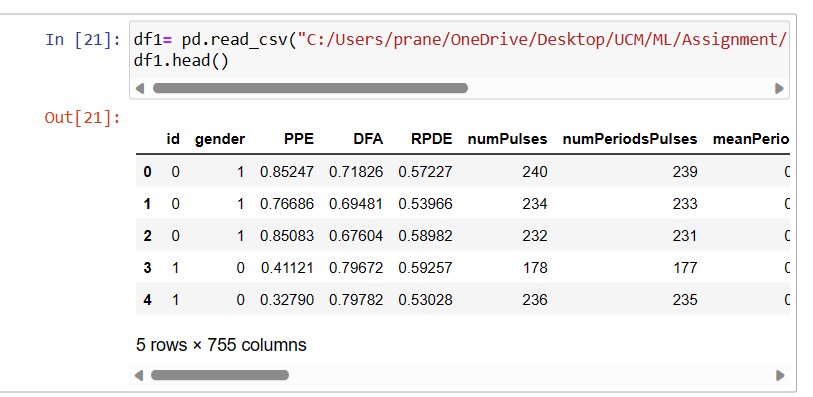


Using KMeans method of sklearn library, I applied K- Means algorithm by taking k value as 3 on data set, we got after performing feature scaling and PCA. After performing k-means on this data we got a silhouette score of 38%.

****

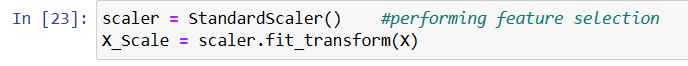
1. **Use pd\_speech\_features.csv.**

A csv file was imported using the read\_csv function. The top rows of a data collection are returned by the pandas library's head() function.



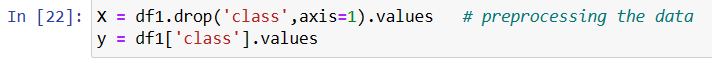
**a. Perform Scaling**

We scaled the data set's features using the StandardScalar technique. All features' ranges are normalized using feature scaling.

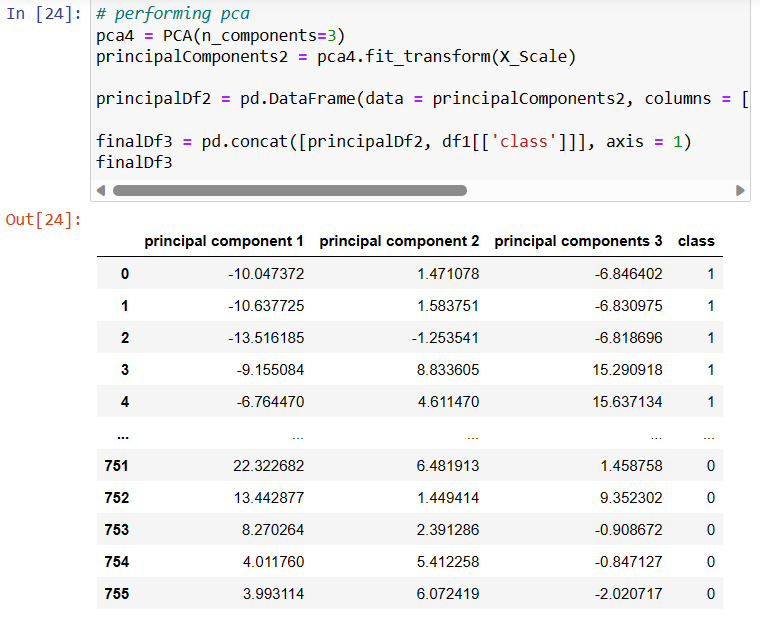


**b. Apply PCA (k=3)**

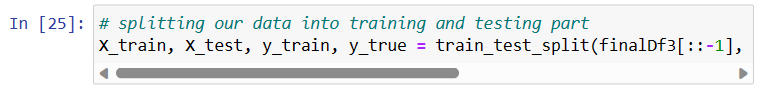
We don't need the output labels to run PCA on this data set since PCA doesn't rely on the output labels. We eliminated an unneeded class column by using the drop technique.



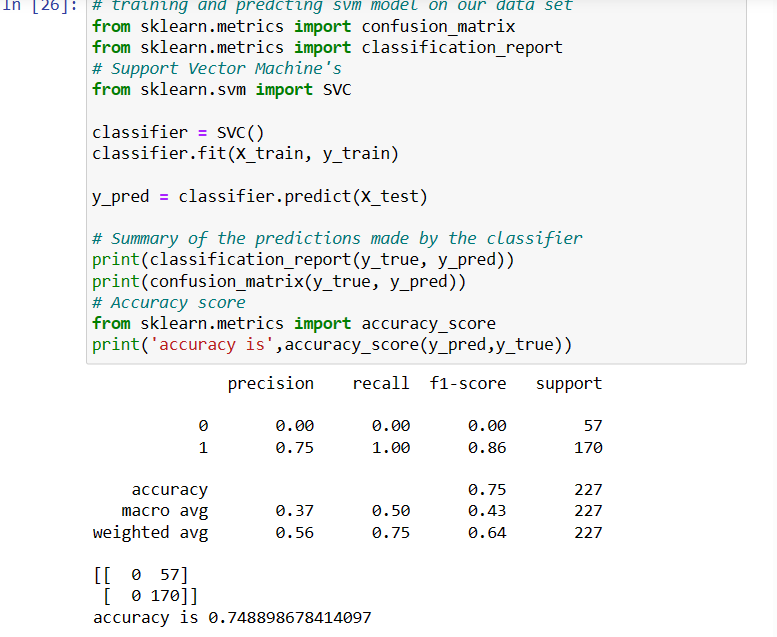
We imported the PCA function from the Sklearn Python library in order to perform PCA on the data set. PCA creates a data frame with features that have the most variance with other attributes by eliminating the duplicate features. Here, we retained the k value at 3, which resulted in a three-component data dimension.



**c.Use SVM to report performance.**



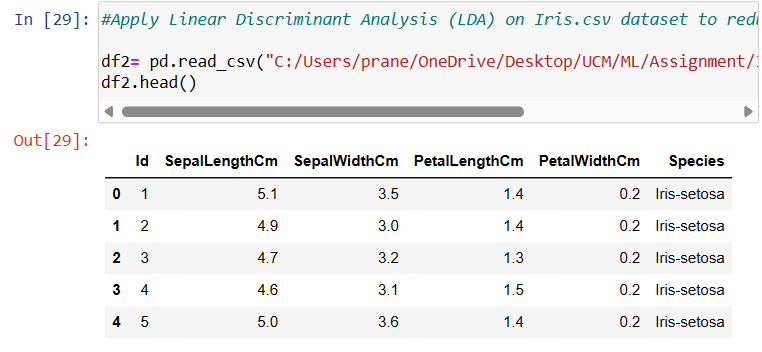
After conducting PCA with the help of the Sklearn module, the support vector machine technique is used on the data set we obtained. When we trained SVM on our data set, we obtained an accuracy of 74.8%.



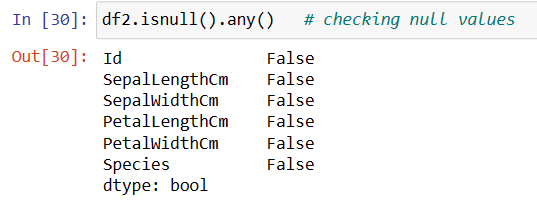
The F1-score, accuracy, recall, and precision metrics are some of the many metrics used to compare the machine learning algorithms. We will choose the method that is the most accurate and comprehensive classifier for the dataset based on those values.

**3. Apply Linear Discriminant Analysis (LDA) on Iris.csv dataset to reduce dimensionality of data to k=2.**

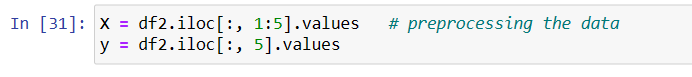
The read\_csv function was utilized to import a csv file. The head() function of the pandas library returns the top rows of a data collection.



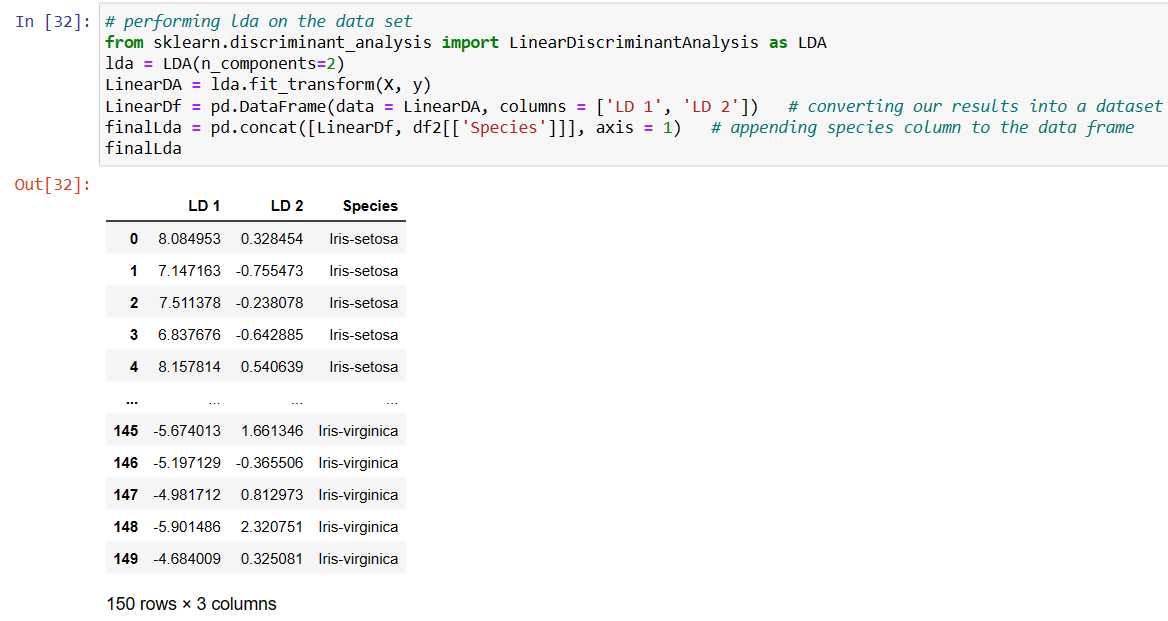
The pandas library's isnull() function looks for any values in the data collection. The iris dataset doesn't contain any null values.



We need the output labels in order to execute LDA on this data set since LDA relies on them to reduce the dimensionality of data based on output classes.



LDA may be carried out in Python by using the LinearDiscriminantAnalysis class of the sklearn.discriminant\_analysis module. We may retrieve the results in two linear discriminates by changing the n\_components parameter to 2. To get our results, we apply the fit and transform procedures.



1. **Briefly identify the difference between PCA and LDA**

In machine learning, the term "dimensionality reduction" refers to the process of gathering a group of key factors in order to minimize the number of random variables taken into account. The two primary techniques for dimensionality reduction are Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA).

LDA is a supervised dimensionality reduction approach, whereas PCA is unsupervised.

Without relying on the output labels, PCA obtains the desired results. By disregarding the duplicates of other characteristics, PCA produces a data set with the greatest variation between the features. PCA does not take the output labels into account since the variation between the features is independent of the result.

The output labels are important for LDA. LDA finds a decision boundary and decreases the feature set dimensions based on the output labels information. In order to make the clusters stand out from one another and the individual components of a cluster as close to the cluster centroid as feasible, the data points are then projected to new dimensions.