**Summer 2023: ML 5710**

**(Assignment 2)**

**Anudeep Allamsetty - 700741171**

1. **Pandas** 
   1. Read the provided CSV file ‘data.csv’.

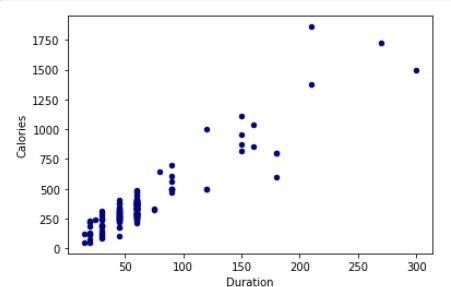
<https://drive.google.com/drive/folders/1h8C3mLsso-R-sIOLsvoYwPLzy2fJ4IOF?usp=sharing>

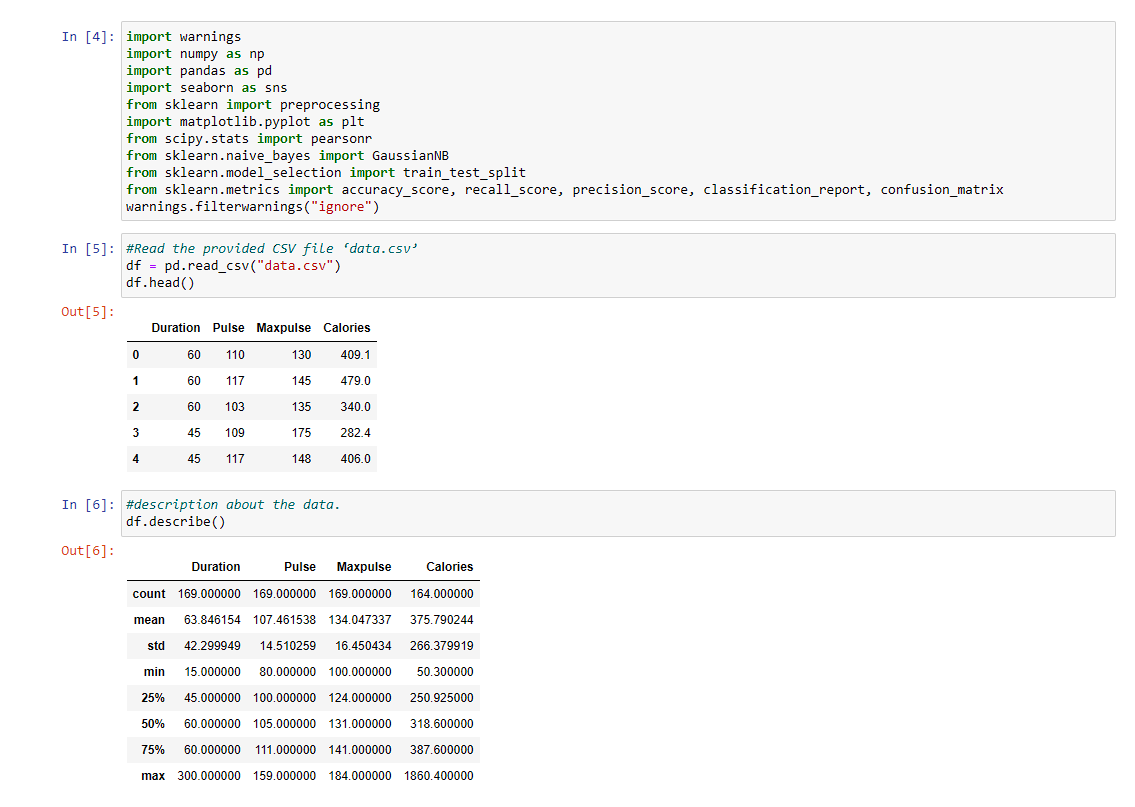
* 1. Show the basic statistical description about the data.
  2. Check if the data has null values.

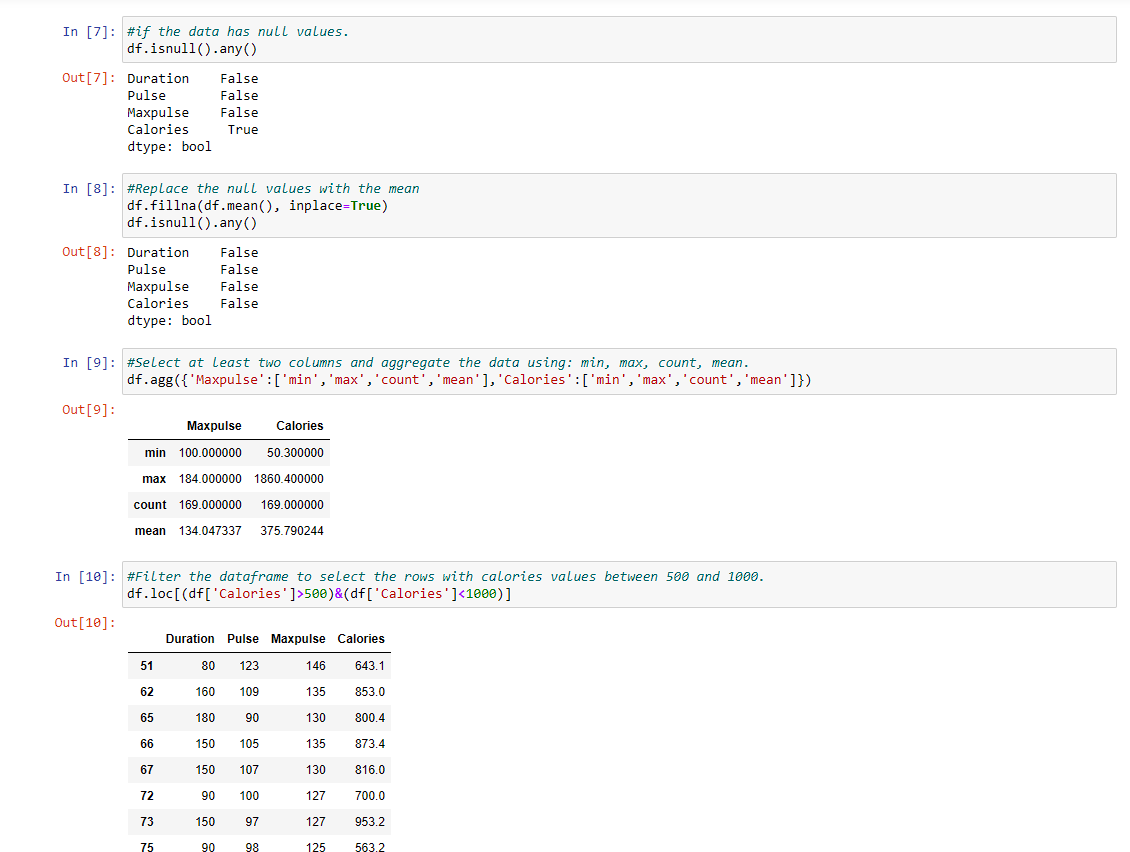
a. Replace the null values with the mean

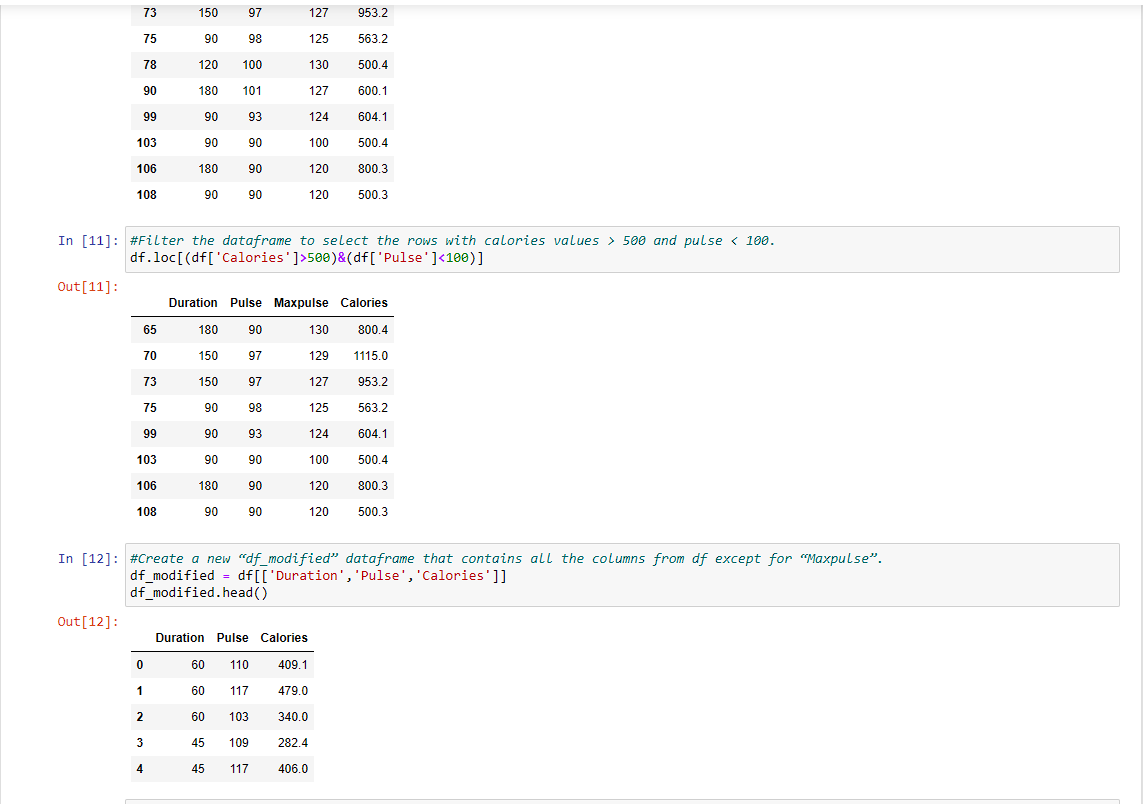
* 1. Select at least two columns and aggregate the data using: min, max, count, mean.
  2. Filter the dataframe to select the rows with calories values between 500 and 1000.
  3. Filter the dataframe to select the rows with calories values > 500 and pulse < 100.
  4. Create a new “df\_modified” dataframe that contains all the columns from df except for “Maxpulse”.
  5. Delete the “Maxpulse” column from the main df dataframe
  6. Convert the datatype of Calories column to int datatype.
  7. Using pandas create a scatter plot for the two columns (Duration and Calories).

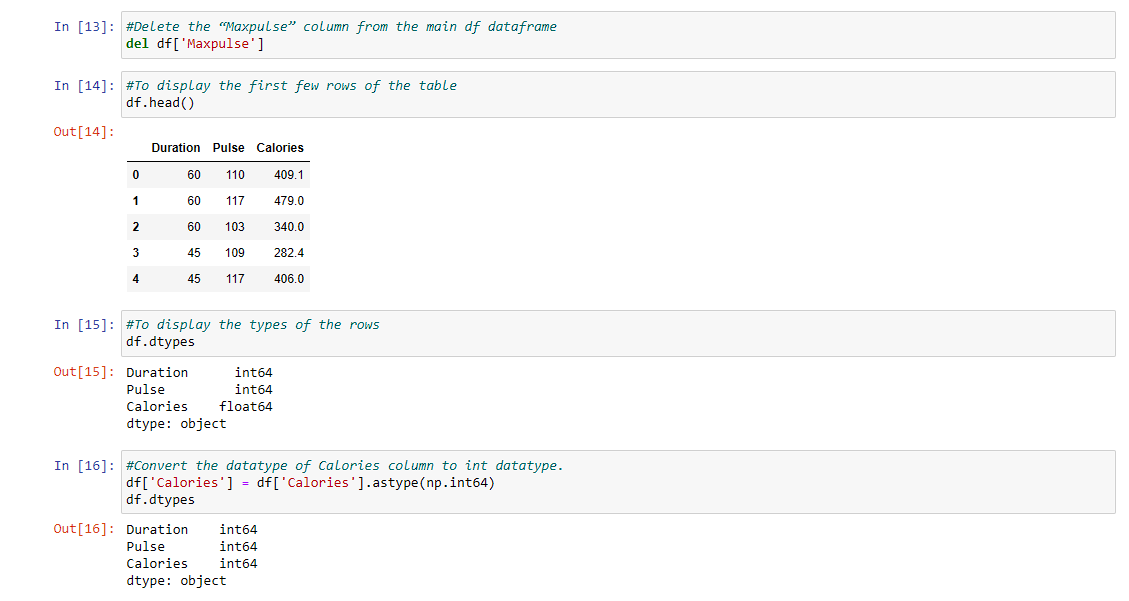
a. Example:

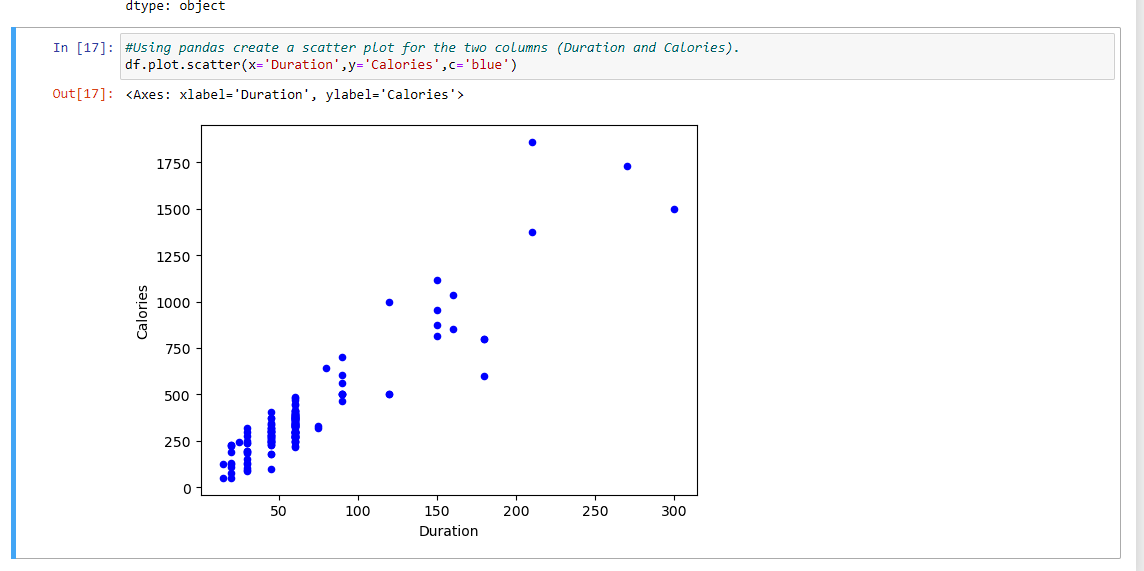












The program imports the necessary libraries for machine learning, data processing, visualization, and handling errors.A dataset is loaded from a CSV file and kept in a DataFrame with the name df.To provide a brief overview of the data, the code shows the top five rows of the DataFrame.For numerical columns in the DataFrame, it computes descriptive statistics like count, mean, standard deviation, minimum, quartiles, and maximum.The program scans the DataFrame for any missing values and returns a boolean value (True/False) for each column to indicate whether or not it has any.The code substitutes the mean value of each column for any missing data. After handling them, it checks once more to see if any missing values are still present.The code combines information such as minimum, maximum, count, and mean for the DataFrame's "Maxpulse" and "Calories" columns.The DataFrame is filtered according to predetermined criteria, such as choosing rows where the 'Calories' column is greater than 500 and less than 1000, or where 'Calories' is larger than 500 and 'Pulse' is less than 100.The duration, pulse, and calories columns from the original DataFrame are the sole columns in the newly formed DataFrame, df\_modified. This altered DataFrame's initial few rows are shown.The DataFrame's 'Maxpulse' column gets removed.The code shows data and changes the 'Calories' column's data type to a 64-bit integer type (int64).

1. **Scikit-learn**

1. Implement Naïve Bayes method using scikit-learn library.

* + - * 1. Use the glass dataset available in [Link a](https://umkc.box.com/s/ea6wn1cidukan67t02j60nmp1ljln3kd)lso provided in your assignment.
        2. Use **train\_test\_split** to create training and testing part. 2. Evaluate the model on testing part using score and

classification\_report(y\_true, y\_pred)

1. Implement linear SVM method using scikit library

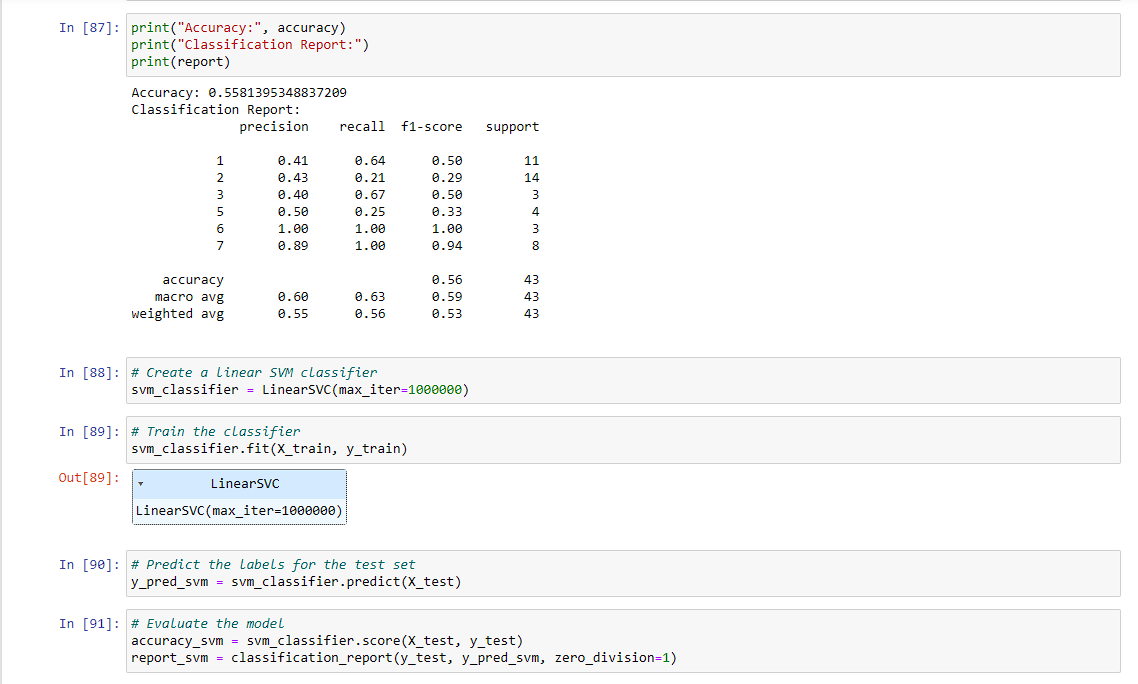
* + - 1. Use the glass dataset available in [Link a](https://umkc.box.com/s/ea6wn1cidukan67t02j60nmp1ljln3kd)lso provided in your assignment.
      2. Use **train\_test\_split** to create training and testing part. 2. Evaluate the model on testing part using score and

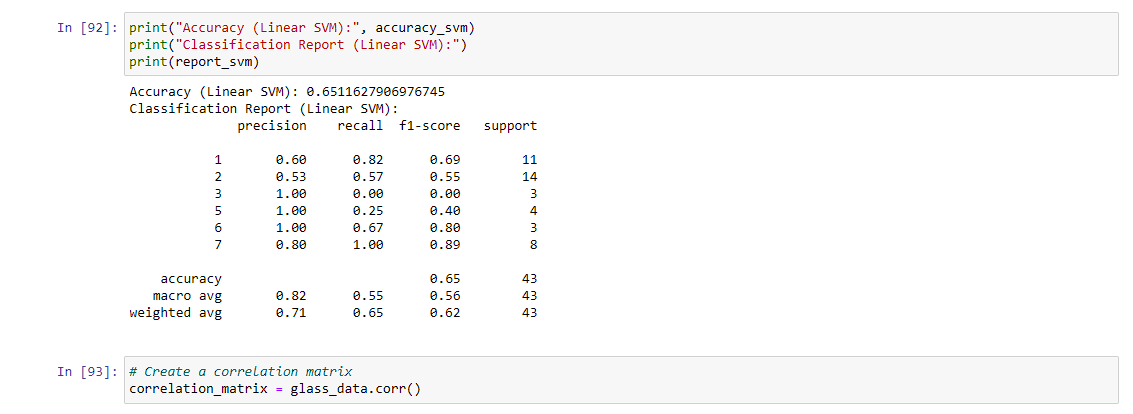
classification\_report(y\_true, y\_pred)

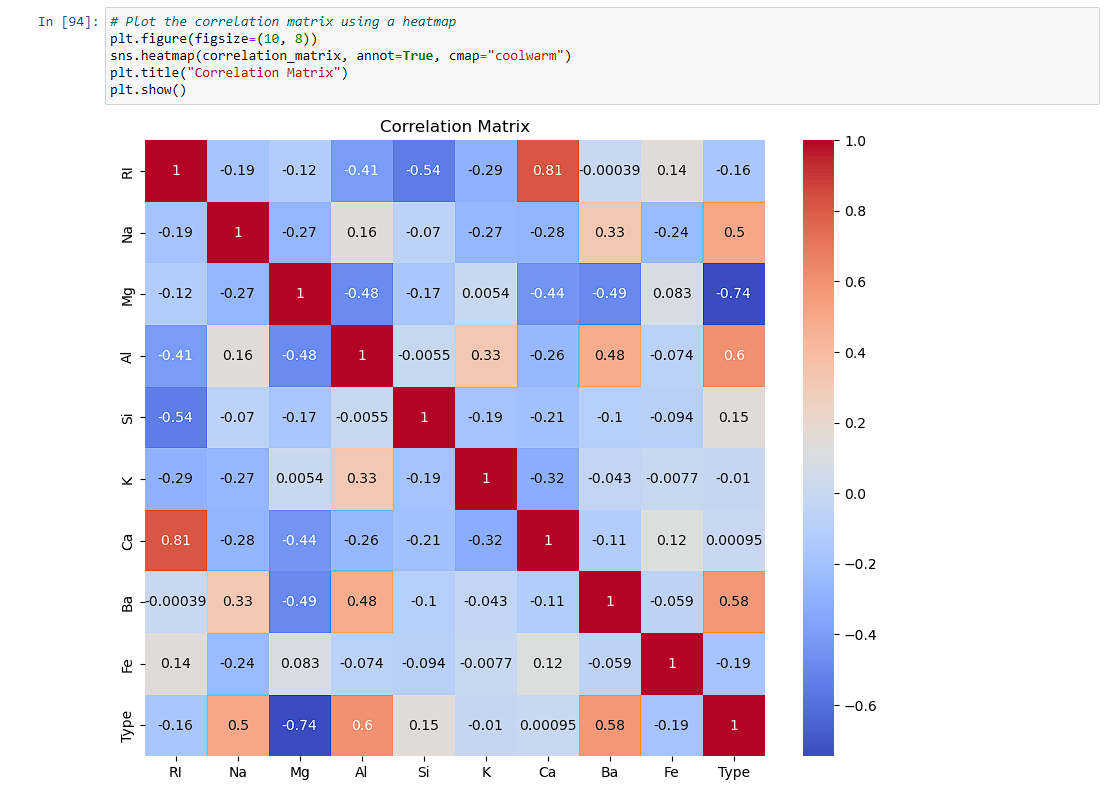
Do at least two visualizations to describe or show correlations in the Glass Dataset.

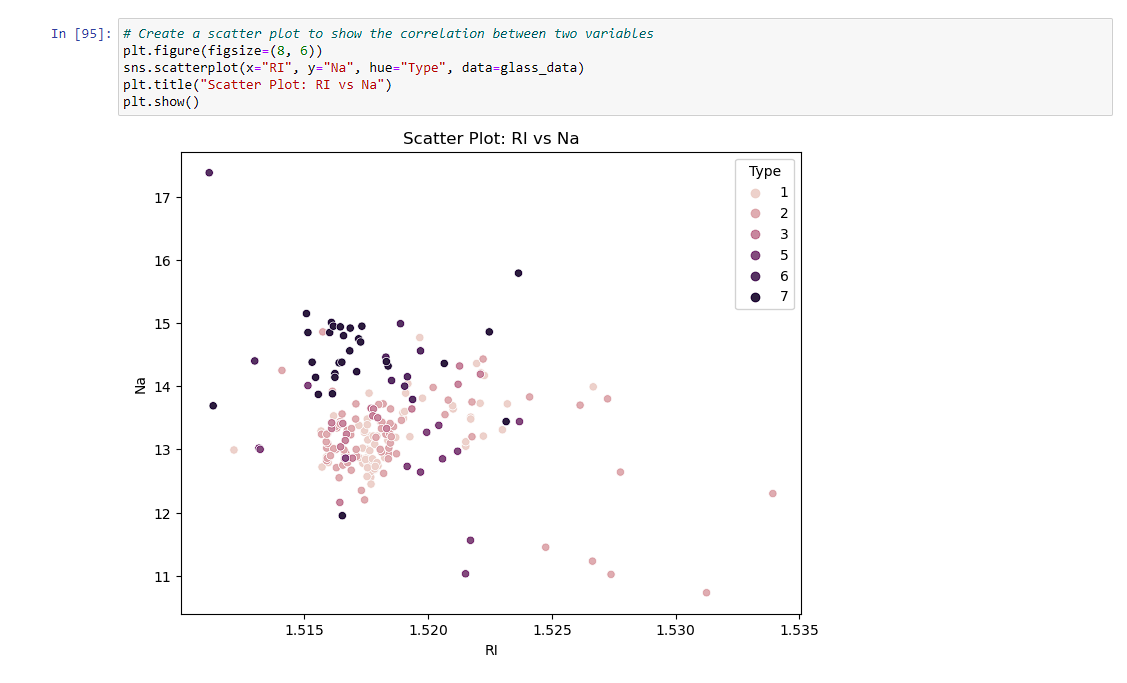
Which algorithm you got better accuracy? Can you justify why?











The code imports the required libraries for data manipulation, machine learning, and visualization, including pandas, scikit-learn, and seaborn.A DataFrame called glass\_data is used to hold the glass dataset after it has been loaded from a CSV file.The dataset is divided into characteristics (X) and the desired outcome (y), where X contains all columns other than the "Type" column and y only the "Type" column.Using the train\_test\_split function of scikit-learn, the dataset is further divided into training and testing sets. A random seed is supplied for repeatability, and the testing set size is set to 20% of the data.The scikit-learn GaussianNB class is used to construct the Naive Bayes classifier (GaussianNB).The fit approach is used to train the Nave Bayes classifier on the training set. The trained Nave Bayes classifier makes predictions on the testing set.The score method, which compares the predicted labels with the actual labels, is used to determine the accuracy of the Naive Bayes classifier.The classification\_report function is used to create the classification report, which contains metrics like precision, recall, and F1-score.The Nave Bayes classifier's accuracy and classification report are printed.The scikit-learn LinearSVC class is used to build a linear SVM classifier (LinearSVC).The fit approach is used to train the linear SVM classifier on the training set. On the testing set, predictions are made using the trained linear SVM classifier.Using the score approach, the linear SVM classifier's accuracy is determined.The classification\_report function is used to create the classification report for the linear SVM classifier.The linear SVM classifier prints its accuracy and classification report.The data visualization process is then carried out by the code utilizing Seaborn and Matplotlib.The corr method, which determines the correlation between all pairs of variables, is used to construct a correlation matrix from the glass dataset.The heatmap function from seaborn is used to visualize the correlation matrix as a heatmap. The correlation between the RI (refractive index) and Na (sodium) variables is displayed in a scatter plot, with the color of each data point denoting the type of glass.Using matplotlib, the final scatter plot and correlation matrix heatmap are shown.