ML Assignment 2

Hima Bindu Desalanka

700739704

1. Pandas

1. Read the provided CSV file ‘data.csv’. https://drive.google.com/drive/folders/1h8C3mLsso-R-sIOLsvoYwPLzy2fJ4IOF?usp=sharing

2. Show the basic statistical description about the data.

3. Check if the data has null values

. a. Replace the null values with the mean

4. Select at least two columns and aggregate the data using: min, max, count, mean.

5. Filter the dataframe to select the rows with calories values between 500 and1000.

6. Filter the dataframe to select the rows with calories values > 500 and pulse

7. Create a new “df\_modified” dataframe that contains all the columns from df except for “Maxpulse”.

8. Delete the “Maxpulse” column from the main df dataframe

9. Convert the datatype of Calories column to int datatype.

10. Using pandas create a scatter plot for the two columns (Duration and Calories).

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

A screen shot of a graph

Description automatically generated with medium confidence

The program imports the required libraries for data processing, error handling, data visualization, and machine learning.A dataset is kept in a DataFrame with the name df after being loaded from a CSV file.The code displays the first five rows of the DataFrame to give a quick overview of the data.In the numerical columns,

It computes descriptive statistics like count, mean, standard deviation, minimum, quartiles, and maximum using the DataFrame.A boolean value (True/False) is returned for each column to indicate whether or not it has any missing values after the program searches the DataFrame for any.If any data is missing, the code fills in the gaps with the mean value of each column. It checks again to see if any missing values are still present after handling them. The code combines data for the "Maxpulse" and "Calories" columns of the DataFrame, including minimum, maximum, count, and mean.The DataFrame is filtered based on preset criteria, such as selecting 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 only columns in the brand-new DataFrame, df\_modified, are the duration, pulse, and calories columns from the original DataFrame. The first few rows of this modified DataFrame are displayed.The 'Maxpulse' column in the DataFrame is deleted.The code displays data and changes the 'Calories' column's data type to a 64-bit integer type (int64).

2. Scikit-learn

1. Implement Naïve Bayes method using scikit-learnlibrary.

a. Use the glass dataset available in Link also provided in your assignment.

b. 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

a. Use the glass dataset available in Link also provided in your assignment.

b. Use train\_test\_split to create training and testing part.

2. Evaluate the model on testing part using score and Do at least two visualizations to describe or show correlations in the Glass Dataset.

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

A screenshot of a computer program

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer program

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, font, screenshot, line

Description automatically generated

A picture containing text, screenshot, square

Description automatically generated

A screen shot of a computer screen

Description automatically generated with low confidence

The code imports the necessary libraries, such as pandas, scikit-learn, and seaborn, for data manipulation, machine learning, and visualization.The glass dataset is loaded from a CSV file and stored in a DataFrame called glass\_data.The dataset is separated into characteristics (X) and the desired result (y), where X includes all columns except for the "Type" column and y only includes the "Type" column.The dataset is further split into training and testing sets using scikit-learn's train\_test\_split function. For consistency, a random seed is provided, and 20% of the data is designated for testing.The Naive Bayes classifier (GaussianNB) is built using the scikit-learn GaussianNB class. .The Naive Bayes classifier is trained on the training set using the fit approach. On the testing set, the trained Nave Bayes classifier makes predictions.The Naive Bayes classifier's accuracy is assessed using the score method, which contrasts the predicted labels with the actual labels.The classification report, which includes metrics like precision, recall, and F1- score, is produced using the classification\_report function.The classification report and accuracy of the Nave Bayes classifier are printed.A linear SVM classifier (LinearSVC) is created using the scikit-learn LinearSVC class.The linear SVM classifier is trained on the training set using the fit approach. The trained linear SVM classifier makes predictions on the testing set. The accuracy and classification report from the linear SVM classifier are printed.The code then uses Seaborn and Matplotlib to carry out the data visualization process.The glass dataset is used to create a correlation matrix using the corr method, which establishes the correlation between each pair of variables.The correlation matrix is shown as a heatmap using the seaborn heatmap function. A scatter plot is used to show the relationship between the RI (refractive index) and Na (sodium) variables, with the color of each data point designating the kind of glass.The final scatter plot and correlation matrix heatmap are displayed using matplotlib.The output showed that the Linear SVM for my model has higher accuracy than the Naive Bayes.

https://drive.google.com/drive/folders/1Fjle7i7Nagu-vEJO8SJ3tuB3On6pob7p?usp=sharing ---> Video link

https://github.com/hbdesalanka/assignment2 --> github link