

Summer 2023: ML 5710

(Assignment 2)

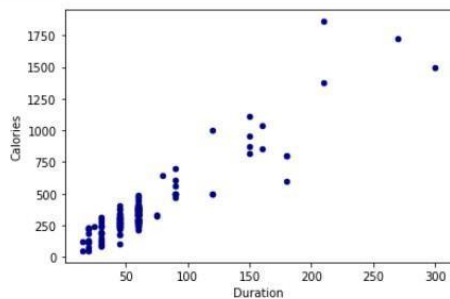
Programming elements:

Classification

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 and 1000.
6. Filter the dataframe to select the rows with calories values > 500 and pulse < 100.
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. Example:



2. Scikit-learn

1. Implement Naïve Bayes method using scikit-learn 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

```
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

```
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?

** Follow the IPC rubric guidelines.

Submission Guidelines:

1. Use the same set of steps performed in project submission, along with the video presentation.