Spring 2023: Machine Learning 5710

(Assignment 4)

Murali Yaswanth Motamarri (700741647)

**Video link:** [**https://drive.google.com/file/d/1vShaLoDAa2zjF5bmiLuB\_Lv7OMYkLwdM/view?usp=share\_link**](https://drive.google.com/file/d/1vShaLoDAa2zjF5bmiLuB_Lv7OMYkLwdM/view?usp=share_link)

**Github link:**

# Pandas

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

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

Text

Description automatically generated

* 1. Show the basic statistical description about the data.

Graphical user interface

Description automatically generated with medium confidence

* 1. Check if the data has null values.
     1. Replace the null values with the mean

Graphical user interface, text, application, email

Description automatically generated

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

Text

Description automatically generated

* 1. Filter the dataframe to select the rows with calories values between 500 and 1000.

Graphical user interface, text, application

Description automatically generated

* 1. Filter the dataframe to select the rows with calories values > 500 and pulse < 100

Graphical user interface, text, application, email

Description automatically generated

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

Graphical user interface, text, application, email

Description automatically generated

* 1. Delete the “Maxpulse” column from the main df dataframe

Graphical user interface, text, application

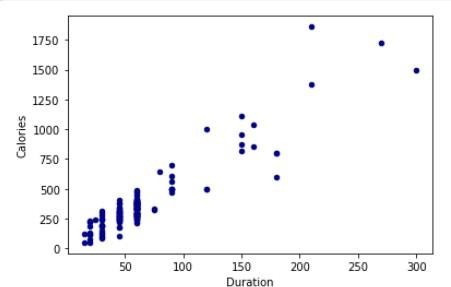
Description automatically generated

* 1. Convert the datatype of Calories column to int datatype.

Graphical user interface, text, application, chat or text message

Description automatically generated

* 1. Using pandas create a scatter plot for the two columns (Duration and Calories).
     1. Example:



Chart, scatter chart

Description automatically generated

1. **(**Titanic Dataset**)**
   1. **Find the correlation between ‘survived’ (target column) and ‘sex’ column for the Titanic use case inclass.**

Graphical user interface, application, Word

Description automatically generated

* + 1. **Do you think we should keep this feature?**

Ans: we should not keep this feature because the percentage is 54.

* 1. Do at least two visualizations to describe or show correlations.

Graphical user interface

Description automatically generated with low confidence

Graphical user interface, application

Description automatically generated

Chart, bar chart

Description automatically generated

* 1. Implement Naïve Bayes method using scikit-learn library and report the accuracy.

Graphical user interface, text, application

Description automatically generated

1. **(**Glass Dataset**)**
   1. Implement Naïve Bayes method using scikit-learn library.
      1. Use the glass dataset available in [Link](https://umkc.box.com/s/ea6wn1cidukan67t02j60nmp1ljln3kd) also 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)

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated

1. Implement linear SVM method using scikit library
   1. Use the glass dataset available in [Link](https://umkc.box.com/s/ea6wn1cidukan67t02j60nmp1ljln3kd) also 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)

Text

Description automatically generated

Table

Description automatically generated

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

Calendar

Description automatically generated with medium confidence

Graphical user interface, application

Description automatically generated

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

From the above accuracy scores we can say that Naïve bayes algorithm got better accuracy than Linear SVM, because Linear SVM requires large amount of data when compared with naïve bayes algorithm. SVM is more expensive than Naïve bayes algorithm. So Naïve bayes is better than Linear SVM algorithm.