MACHINE LEARNING ASSIGNMENT 4

Name : Supriya Sama
700744510

Video Link:
https://drive.google.com/file/d/1m6eZf-urF7rcTgUlCsS_FG-9v4eo5xco/view?usp=share_link

Pandas

1. Read the provided CSV file 'data.csv'.
Using the import keyword, I imported the pandas module.
2. With describe() function from pandas module we get the statistical description of data which is present in data frame.
3. To check any null values present in data frame we need to use isnull() function.
4. Select at least two columns and aggregate the data using: min, max, count, mean. Using agg() method we can apply certain operation on data.
5. Filter the dataframe to select the rows with calories values between 500 and 1000. Using '& 'operator we can filter the data based on the conditions given.
6. Filter the dataframe to select the rows with calories values > 500 and pulse < 100. Using '& 'operator we can filter the data based on the conditions given.
7. Create a new "df_modified" dataframe that contains all the columns from df except for "Maxpulse". Using copy() method we can copy the data from the original data frame to another data frame.
8. Delete the "Maxpulse" column from the main df dataframe. drop() method can be used to remove a particular column from the data frame.
9. Convert the datatype of Calories column to int datatype. astype() method to convert one data type to other.
10.Using pandas create a scatter plot for the two columns (Duration and Calories). pandas module contains functions to represent the data in visual format. Plot.scatter() method.

[Titanic Dataset]

1. Find the correlation between ‘survived’ (target column) and ‘sex’ column for the Titanic use case in class.
As sex column contains string object so we cannot find the correlation between sex column and survived column. So, first we need to convert into type of objects with which we are comparing and find the correlation with survived column.

```
In [6]: #Import the dataset
        test_df = pd.read_csv('C:\\Users\\supri\\Desktop\\test.csv')
        train_df = pd.read_csv('C:\\Users\\supri\\Desktop\\train.csv')
        combine = [train_df,test_df]

In [7]: train_df['Sex'].str.get_dummies().corrwith(train_df['Survived']/train_df['Survived'].max())

Out[7]: female    0.543351
        male     -0.543351
        dtype: float64
```

a. Do you think we should keep this feature?
As correlation results shows that males were strongly negatively correlated, and females were Strongly positively correlated with their survival. Males are inversely proportional, and females are directly proportional to their survival. So, we need this feature to analysis.

[Glass Dataset]

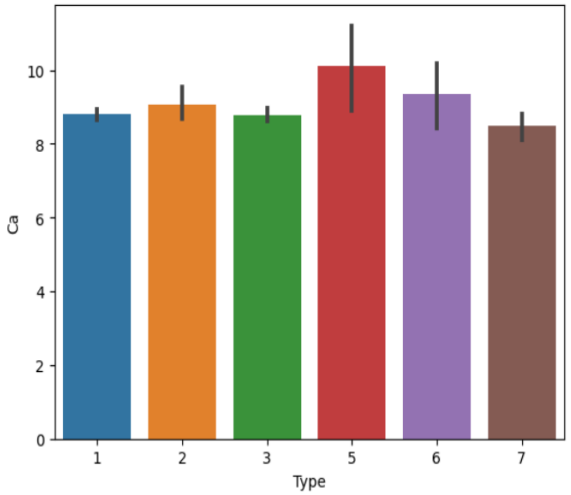
Implement Naive Bayes method using scikit-learn library and implement linear SVM method using scikit library.

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

Seaborn library is used to visually show the correlations between the columns data. Here, I am representing correlation between Type and Ca column using bar plot where Type on x-axis and Ca on y-axis.
Similarly, Regression plot for Type and Fe columns and categorized plot for Type and K columns.

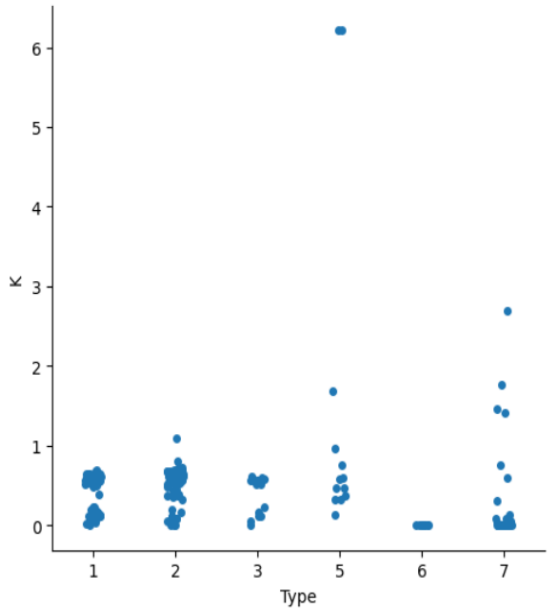
In [12]:
```
#For Visualisation import seaborn library
import seaborn as sns
import matplotlib.pyplot as plt
sns.barplot(x = glass['Type'], y = glass['Ca'])
```

Out[12]: <Axes: xlabel='Type', ylabel='Ca'>

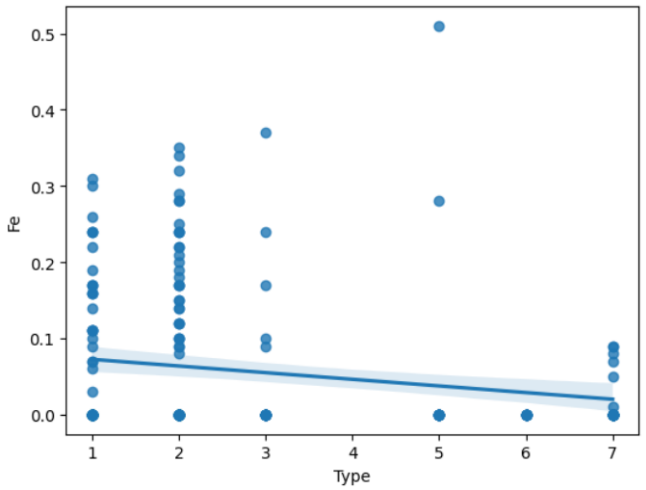


In [13]:
```
sns.catplot(data=glass, x="Type", y="K")
```

Out[13]: <seaborn.axisgrid.FacetGrid at 0x24bcc9bbed0>



In [14]:
```
sns.regplot(x="Type", y="Fe", data=glass);
```



In [ ]:
```

```

Which algorithm got better accuracy?

Among Naive Bayes and Support vector machine algorithms, Support vector algorithm get more accuracy than Naive Bayes.