DS505
Assignment2
Deadline: 23:59 Hrs, 11.10.2022

Instructions:

1. Dataset links has been provided in blue color
2. A single jupyter notebook can be submitted containing solutions for each question separated by proper markdown cells

Numpy:

Q1.

Dataset Description:

Given file “**terrorismData.csv**”. It is an open source database including information on terrorist attacks around the world from 1970 through 2017. This dataset includes systematic data on domestic as well as international terrorist incidents that have occurred during this period.

Note: Expected is to not use pandas read_csv() function to read given csv file. You can use inbuilt csv reader available in python or numpy.genfromtxt function

A. Find the number of attacks held between day 10 and day 20? (ignoring the year and month) (including both day)

Print count of NumberOFAttack as integer value.

Note: np.count_nonzero() can be used here

B. Find the number of attacks held between 1 Jan 2010 and 31 Jan 2010? (including both dates)

Print count of NumberOFAttack as integer value.

Note: Ignore the case where day is 0. np.datetime64() can be used here

C. Find the casualty in the Red Corridor States? Mainly Red Corridor States include Jharkhand, Odisha, Andhra Pradesh, and Chhattisgarh.

Print count of Casualty as integer value.

Note: Casualty = Killed + Wounded. np.sum() can be used here

D. Find the most frequent day of attack in a terrorismDataset?

Print count of frequent day and number of attacks as integer value.

Note: np.unique can be used here

Pandas:

Q2.
Same dataset as of Q1
Note: Dataset can be read using Pandas read_csv() function. Expected is to solve given problems using only pandas inbuilt functions.

A. The most dangerous city in Jammu and Kashmir and terrorist group which is most active in that city?
Print count of number of attacks in that city as integer value.
Note: Ignoring the unknown terrorist group. Here dangerous related with the number of terrorist attacks.

B. Find out the Country with highest number of terror attacks and in which year the most number of terrorist attacks happened in that country?
Print count of terrorist attacks as integer value.

C. The deadliest attack in the history of HumanKind?
Print count of killed people as integer value.
Note: Here the deadliest attack means, in which the most number of people killed.

D. There was formation of a new government in India on 26 May 2014. So the current government's span is from 26th May 2014 to current. Find out two things from this period -

a. Total number of attacks done in this period in India. Find this count as integer
b. Which terrorist group was most active in this period in India. Most active means, group which has done the maximum number of attacks.

Note: Ignore the unknown group

E. Find the frequency of the Casualty in Red Corridor states and in Jammu and Kashmir? Here frequency is (Total Casualty/Total number of years)
Print frequency as integer value.
Note: Red Corridor States include Jharkhand, Odisha, Andhra Pradesh, and Chhattisgarh. Here Casualty = Killed + Wounded. Don't fill the NaN value present in the Killed and Wounded feature.

Scikit-Learn:

Q3.

A. Linear Regression - **Diabetes Dataset**

Diabetes dataset is one of the datasets available in sklearn. The diabetes dataset consists of 10 physiological variables (age, sex, weight, blood pressure) measure on 442 patients, and an indication of disease progression after one year.

You are given a Training dataset csv file with X train and Y train data. As studied in lecture, your task is to come up with Linear Regression training algorithm and thus predictions for the test dataset given.

Read Instructions carefully -

Use Linear Regression(in scikit learn) as a training algorithm and submit results predicted by that. Print feature importance and different error metrics used.

Submit a csv file with only predictions for X test data. File should not have any headers and should only have one column i.e. predictions. Also prediction values in file should be upto **5** decimal places.

B. Classification - Convert above Linear Regression problem to Binary Classification Problem by considering patients to be diabetic if Y>=138, else non-diabetic.

Apply Cross-validation and report avg performance metrics (Accuracy/Precision/Recall/F1) upon applying both SVM and Decision Tree classifiers. Compare the performances between the two. Print features importance learned by both classifiers.

Note: While calculating different performance metrics, consider diabetic to be a positive class.

NetworkX:

Q4.
Given “**test_graph.txt**” dataset with 516 Nodes and 1188 Edges:

a. Create and Visualize Undirected Network using NetworkX tool.
b. Visualize the Degree distribution of the Network
c. Print Most Influential Node - Degree Centrality
d. Print Most Influential Edge - Edge Betweenness Centrality
e. Print Node with highest Betweenness Centrality
f. Print all Shortest Paths Between Nodes 300 and 400