Data Mining lab

Assignment 4

Name: Ayush Kumar

Regno: 20214284

Group: CSE-6B1

Q1.

1.

import numpy as np import pandas as pd

my\_column\_names = ['Reg\_no', 'Name', 'Subject1', 'Subject2', 'Subject3', 'Subject4' ] my\_data = np.array( [ [2022001,'Abhijeet',65,65,69,81],

[2022002,'Ajeet',75,75,90,81],

[2022003,'Amit',75,5,69,87],

[2022004,'Ranjeet',55,65,79,91],

[2022005,'Santosh',85,85,60,61],

[2022006,'Satyam',73,75,68,51],

[2022007,'Shivam',85,85,50,40],

[2022009,'Shyam',75,65,69,81],

[2022010,'Yash',85,75,89,61] ] )

my\_dataframe = pd.DataFrame(data=my\_data, columns=my\_column\_names)

my\_dataframe = my\_dataframe.astype({'Subject1':int, 'Subject2':int, 'Subject3':int, 'Subject4':int}) print(my\_dataframe)

2.

my\_dataframe["Total"] = 0

my\_dataframe["Total"] = my\_dataframe["Total"].astype(int) print(my\_dataframe)

3.

my\_dataframe["Total"] = (my\_dataframe["Subject1"] + my\_dataframe["Subject2"] + my\_dataframe["Subject3"]

+ my\_dataframe["Subject4"])/4

print(my\_dataframe)

4.

|  |  |  |
| --- | --- | --- |
| my\_dataframe["Grade"] = np.select(  [ | | |
| my\_dataframe["Total"] | >= | 90, |
| my\_dataframe["Total"] | >= | 80, |
| my\_dataframe["Total"] | >= | 70, |
| my\_dataframe["Total"] | >= | 50, |

my\_dataframe["Total"] < 50

],

["A", "B", "C", "D", "E"]

)

print(my\_dataframe)

5.

print("Subset:")

print(my\_dataframe[['Reg\_no', 'Name', 'Grade']])

6.

# Sort my\_dataframe based on the 'Grade' column in ascending order

sorted\_df = my\_dataframe.sort\_values(by='Grade', ascending=True)

print(sorted\_df)

Q2.

1. ​

import pandas as pd

df= pd.read\_csv("weather-numeric.csv") df

1. and 3.

Y = df.iloc[:, -1]

X = df.iloc[:, :-1]

4.

from sklearn.model\_selection import

train\_test\_split

5.

X\_train1, X\_test1, Y\_train1, Y\_test1 = train\_test\_split(X, Y, test\_size=0.2, random\_state=1) X\_train1.to\_csv('train1.csv', index = False) X\_test1.to\_csv('test1.csv', index = False)

X\_train2, X\_test2, Y\_train2, Y\_test2 = train\_test\_split(X, Y, test\_size=0.2, random\_state=2) X\_train2.to\_csv('train2.csv', index = False) X\_test2.to\_csv('test2.csv', index = False)

X\_train3, X\_test3, Y\_train3, Y\_test3 = train\_test\_split(X, Y, test\_size=0.2, random\_state=3) X\_train3.to\_csv('train3.csv', index = False) X\_test3.to\_csv('test3.csv', index = False)

X\_train4, X\_test4, Y\_train4, Y\_test4 = train\_test\_split(X, Y, test\_size=0.2, random\_state=4) X\_train4.to\_csv('train4.csv', index = False) X\_test4.to\_csv('test4.csv', index = False)

X\_train5, X\_test5, Y\_train5, Y\_test5 = train\_test\_split(X, Y, test\_size=0.2, random\_state=5) X\_train5.to\_csv('train5.csv', index = False) X\_test5.to\_csv('test5.csv', index = False)

Q3.

import pandas as pd import numpy as np

cols = ['Dataset', 'ENN', 'AllKNN', 'SMOTE'] dt = np.array([

['Pima', 0.9552, 0.9452, 0.9352],

['Glass', 0.9773, 0.9773, 0.9673],

['Wisconsin', 0.7864, 0.7864, 0.7864]

]

)

df = pd.DataFrame(data=dt, columns=cols)

df = df.astype({'ENN':float, 'AllKNN':float, 'SMOTE':float})

print(df)

1.

ranked\_df = df.iloc[:, 1:].rank(axis=1, ascending=False, method='min') ranked\_df.insert(0, 'Dataset', df['Dataset'])

print(ranked\_df)

2.

average\_ranks = ranked\_df.iloc[:, 1:].mean(axis=0,).round(2) print(average\_ranks)

3.

highest\_performer = average\_ranks.idxmin() lowest\_performer = average\_ranks.idxmax()

print("\nHighest Performer:", highest\_performer) print("Lowest Performer:", lowest\_performer)