**Lambton College in Mississauga**

AML 2203 – Advanced Python

Term Test 1

Winter 2022

March 10, 2022 duration of test: in-class

Student name: Aadarsha chapagain Student no.:C0825975

* **This is an individual and online test.**
* **Late or emailed submissions are NOT marked.**
* **For the programming questions, create python programs and execute them using your IDE.**
* **Take screenshots (MUST be clear) of all outputs and save them in this test Word document.**
* **Copy/paste your code (MUST be clear)from the Visual Studio to this test Word document and save.**
* **Submit the test Word document it to Moodle/week 9/** **Midterm Test W22.**

**TIP: STOP WRITING 10 MINUTES BEFORE THE DEADLINE TO MAKE SURE YOUR submission IS ON-TIME.**

* **Copying from others and/or internet violates the College’s academic integrity and will be subject to the College plagiarism policy.**

**Q.1: Highlight or type in the correct answer next to each question for the following questions: (7 marks)**

1. Which of the following is not a file format?

a)CSV

b)TSV

c)KSV  
d)XML

Answer:

2) To which of the file format the below, belongs to

K=**{“Name”:”Benjamen”,**

**“Class”:”Eleventh”,**

**“Course”:”CSE”**

**}**

a)CSV

b)Json

c)TSV

d)HTML

3) what is the extension of the lossless compression format?

a).zip

b).xml

c).xslv

d).html

4) Fill in the blanks with right option:

\_\_\_\_\_\_\_\_\_\_\_\_ is not predefined and \_\_\_\_\_\_\_\_\_ is self descriptive.

a)XML,HTML

b)HTML,XML

c)XML,XML

d)HTML,XML

5)What are the methods to generate datasets for classification and regression in Sklearn?

a)make\_regression

b)make\_classification

c)make\_circles

d)All of the above

6) What is the output of the following?

**import numpy as np**

**k=np.array([[1,2,3],[9,9,9]])**

**print(k.shape)**

a)(3,2)

b)(2,3)

c)(3,3)

d(2,2)

7) What is the output of the following?

**import numpy as np**

**k=np.array([(1,2,3),(9,9,9)])**

**print(k[:,:2])**

a) [[2 1]

[9 9]]

b) [[1 2]

[9 9]]

c) [[2 3]

[9 9]]

d) [[3 1]

[9 9]]

8) What is the output of the following?

**import numpy as np**

**k=np.array([(9,16,25)])**

**print(np.sqrt(k))**

a) [[3. 4. 5.]]

b) [[4. 4. 5.]]

c) [81. 256. 625.]]

d) [[3. 4. 25.]]

9) What is the output of the following?

**import numpy as np**

**l=np.matrix('1 2 3; 4 5 6; 7 8 9')**

**print(np.sum(l.diagonal()))**

a) 12

b) 13

c) 15

d) 14

10) What is the output of the following?

**import numpy as np**

**k=np.array([(1,2,3,4,5,6)])**

**print(k.reshape(3,2))**

a) [[1 2]

[3 4]

[5 6]]

b)

[[1 2 3]

[4 5 6]]

c) [[3 4]

[3 4]

[5 6]]

d) [[4 5 6]

[1 2 3]]

11)

Graphical user interface, text, application

Description automatically generated

What is the file format of the **res** variable?

a)xml

b)yaml

c)json

d)html

12)

**import pandas as pd**

**df=pd.read\_csv(“adam.csv”)**

**fis=df[[“name”]]**

a) The data type of fis is Series

b) The data type of fis is Dataframe

c) The data type of fis is Array

d) The data type of fis is list

13)

**Import pandas as pd**

**df=pd.read\_csv(‘silk.csv’)**

**df.describe()**

a) df.describe() returns describes the structure of dataframe

b) df.describe() returns summary statistics of dataframe

c) df.describe() returns count of number of rows of dataframe

d) df.describe() returns count of number of columns of dataframe

14) Which of the following are data discretization methods?

a)kmeans transformation

b)uniform transformation

c)quantile transformation

d)All of the above

15)Which of the following are used to determine the outliers?

a)Box-Plot

b)Y-score

c)Inter quartile range

d)Both a and c

**Q.2: This part requires creating python programs for the following tasks. Paste code and screenshots of outputs below question. (8 marks)**

1. Create an array using numPy and perform the following operations on it
2. Create a one dimensional array (vector) iterate through all the elements and find the mean of the vector.

Code:

sum =0

# One dimensional array with numpy

one\_vector = np.array([1, 2, 3,4,4,10])

# iteration over each element

for item in one\_vector:

  sum +=item

# Mean is sum of element divided by its length

mean =sum/len(one\_vector)

print("Mean:",mean)

output:

Text

Description automatically generated

1. Create a matrix and perform the below operations on it
2. Get the diagonal elements and calculate the sum of all the diagonal elements.

Code:

# Create matrix using numpy

mat\_diag = np.array([[1, 2, 3, 31],

                  [4, 5, 6, 45],

                  [7, 8, 9, 56],

                  [10, 45, 10, 67]])

print("The diagonal element of the matrix is:",mat\_diag.diagonal())

print("The Sum of the diagonal element is:",mat\_diag.diagonal().sum())

Output:

A screenshot of a computer

Description automatically generated with medium confidence

1. Create a python program to read a Jason file (see below) as a data frame then print the last 3 records.  
     
   data = {  
    **"Duration"**:{  
    **"0"**:60,  
    **"1"**:60,  
    **"2"**:60,  
    **"3"**:45,  
    **"4"**:45,  
    **"5"**:60  
    },  
    **"Pulse"**:{  
    **"0"**:110,  
    **"1"**:117,  
    **"2"**:103,  
    **"3"**:109,  
    **"4"**:117,  
    **"5"**:102  
    },  
    **"Maxpulse"**:{  
    **"0"**:130,  
    **"1"**:145,  
    **"2"**:135,  
    **"3"**:175,  
    **"4"**:148,  
    **"5"**:127  
    },  
    **"Calories"**:{  
    **"0"**:409,  
    **"1"**:479,  
    **"2"**:340,  
    **"3"**:282,  
    **"4"**:406,  
    **"5"**:300  
    }  
   }

Code:

import pandas as pd

data = {

  "Duration":{

    "0":60,

    "1":60,

    "2":60,

    "3":45,

    "4":45,

    "5":60

  },

  "Pulse":{

    "0":110,

    "1":117,

    "2":103,

    "3":109,

    "4":117,

    "5":102

  },

  "Maxpulse":{

    "0":130,

    "1":145,

    "2":135,

    "3":175,

    "4":148,

    "5":127

  },

  "Calories":{

    "0":409,

    "1":479,

    "2":340,

    "3":282,

    "4":406,

    "5":300

  }

}

df = pd.DataFrame.from\_dict(data,orient='index')

#Print last three records.

df.tail(3)

Output:

A screenshot of a computer

Description automatically generated with medium confidence

1. The given data set shown below has missing values. Create a python program to fill in the missing data using one of the methods you know from the lecture note.

A screenshot of a computer

Description automatically generated with medium confidence

Code:

df\_missing = pd.DataFrame({'column\_a':[1,2,4,4,np.nan,np.nan,6],

                    'column\_b':[1.2,1.4,np.nan,6.2,None,1.1,4.3],

                    'column\_c':['a',np.nan,'c','d',np.nan,np.nan,'d'],

                    'column\_d':[True, True,np.nan,None,False,True,False]

                    })

# Here we have used forward fill method to fill the missing value

df\_filled =df\_missing.fillna(axis =0,method='ffill')

print('df\_missing:',df\_missing)

df\_filled

Output:

Graphical user interface

Description automatically generated with medium confidence

This code will fill the missing values with the mean from dataset

**Q.3: Discuss and analyze the following: (5 marks)**

In a movie prediction problem, our goal is to predict future movies for new customers based on some history data, we collected from existing customers who live in the same or surrounding neighbourhoods. Discuss and analyze the problem with *outliers* and *missing* *data* if found in the history data. What does it mean – how does it look? How does it impact the quality of the prediction results? and how to solve it? Write one paragraph (150-200 words) that is specific to the movie problem.

**Answer:**

Here in this problem, we are trying to predict the movies for our new customers. In simple words, we are trying to **recommend** the movies to our customer based on some history data. The history data we have may contain the information about our customer and movies they have watched. According to these provided data we are going to recommend the movies to our customer that they might like.

Before building the models, we have to **clean the data and explore** it. During this process we should be able to detect the **outliers and missing data** if there are any in the given dataset. In our case of movie dataset there might be a genre of movie which is **completely different** from the rest of the genres in this case these genres can be considered as outliers. Other outliers might be the age of the customer if we are considering our customer to be either kids or adult then if there are any **infants** in the customer list then we should discard those customers as **outliers**.

Outliers can be detected using various methods such as **visualization** of the data using the graphs, using **mathematical formula** such **as standardization and statistical approach**. In our case for **movie recommendation,** we can visualize out data for genre and consider the genre which is liked by a single person as outlier.

Regarding missing value, there might me some missing values in our dataset we can use multiple methods to remedy those data points. If there are only few missing values in large datasets of movie, we can drop those values. If dropping the datasets is not an option than we can either use central tendency such **as mean, median or previous value, next value or interpolation** to fill out those missing values

The outlier or missing value directly affects our prediction result. It might skew our results and we might not achieve the accurate model if the outliers and missing data are not handled properly. The model will not be able to represent out datasets if outliers and missing data are not taken care of. In our scenario for **movie recommendation**, if we **drop all the rows with missing** value then we might be left with **small datasets** to train the model and the prediction of such model might not be accurate.