MANIPAL ACADEMY OF HIGHER EDUCATION

THIRD SEMESTER B.TECH END SEMESTER MAKEUP EXAMINATIONS, JAN 2024

INTRODUCTION TO DATA ANALYTICS [CSE 2126]

Marks: 50   **Duration: 180 mins.**

A

Answer all the questions.

Instructions to Candidates: Answer ALL questions Missing data may be suitably assumed

1)   How to read and write data from/to the given files, illustrate with appropriate python commands:

       a) csv file b) Text file(only read operation) and c) html file. (5)
   A)

   B)   Develop a python script to create an array of 3X3 with numbers between 1 to 9 and to find a summary of the array which include the max., min. and sum of each of the rows of the array without using the built-in functions, max(), min() and sum(). (3)

   C)   Consider the stationary stock at two stores, which are stored as dictionaries as: Store1: pen(50), pencil(25) and Store2: pencil(60), rubber(30). Convert the stock data into series with index, pen, ink, pencil, and rubber. Include the store series into a data frame. Display the item with maximum stock in each store. (2)

2)

   A)   a)Design a Python Function to output a single string from the two given input strings, separated by a space and swap the first two characters of each string. [sample s1: 'book'; s2: 'pen', O/p: 'peok bon']. (5)

          b)Develop a Python script that accepts a string and calculates the number of digits and letters.

   B)   Develop a Python script to read a sequence of digits into a list. The end of input is specified by 'done'. Handle the non-digits by exception handling. Use the functions process(), which finds the sum of digits and the number which is the result of the concatenation of positive digits and main(). [Sample Input: 1 2 3 -1 d done, Output: Sum is 5 and the number is 123]. (3)

   C)   Given the roll no, name and marks in three different subjects, Maths, Physics and Chemistry of FIVE students, develop a python script to create a data frame of students' data with roll no as the index. Add a new column Avg_Marks to the data frame by calculating the mean of all the marks obtained by each student. (2)

3)   Develop a Pandas script to split the following dataframe given in Table 1 into groups by school code. Compute the mean, min, and max value of age with customized column names(as Mean_age, Min_age, Max_age) for each school. Also, calculate mean,min,max values of age grouped by school code and tabulate the same.

A)

Table 1

	school	class	name	date_of_Birth	age	height	weight	address
S1	s001	V	Alberto Franco	15/05/2002	12	173	35	street1
S2	S002	V	Gino Mcneill	17/05/2002	12	192	32	street2
S3	S003	VI	Ryan Parkes	16/02/1999	13	186	33	street3
S4	S001	VI	Eesha Hinton	25/09/1998	13	167	30	street1
S5	S002	V	Gino Mcneill	11/05/2002	14	151	31	street2
S6	S004	VI	David Parkes	15/09/1997	12	159	32	street4

(5)

B)   Develop a python script, which computes frequency of each digit present in a string and also it replaces blank spaces in a string with the least frequent digit. [Sample input: '12233 55'; output: '12233155'].

(3)

C)   Design a Python script to plot two lines with x-axis, y-axis labels, title and set line markers.

(2)

4)   Calculate the following metrics for the Dog prediction dataset results, given in Table 2:

(5)

Table 2

A)

Index	1	2	3	4	5	6	7	8	9	10
Actual	Dog	Dog	Dog	Not Dog	Dog	Not Dog	Dog	Dog	Not Dog	Not Dog
Predicted	Dog	Not Dog	Dog	Not Dog	Dog	Dog	Dog	Dog	Not Dog	Not Dog

    a. TP, TN, FN,FP values and tabulate the confusion matrix.

b. Accuracy, Precision, Recall and F1-score.

B) Develop a Python script to remove the intersection of a second set with a first set and display its contents. [Sample input: {1, 4, 5}, {4, 5, 6}; output: set1: {1}, set2: {4, 5, 6}.] (3)

C) Illustrate the usage of groupby operation in Pandas with an example. (2)

5) Apply k-means clustering technique (with k=3) for the following two dimensional data set : (2,3),(5,6), (8,9),(12,15),(15,18),(18,21), (25,30),(30,35),(35,40),(40,45). Use Manhattan distance (d = |x1 - x2| + |y1 - y2|) with max. 3 iterations. Consider (8,9),(15,18), (40,45) as initial centroids. (4)

A)

B) Develop a python code,that prompts user to enter a text filename and displays the number of vowels and consonants present in the file. Handle the FileNotFoundError Exception if user specified text file that doesn't exist. Text file may contain lowercase and /or uppercase characters. (3)

C) A dataset of 10 points along with their class labels is given in the Table 3. The Manhattan distance of new observation with each of these data points are also tabulated in Table 3. Compute the class of the new observation for different values of k (1,3 and 5) using knn-classifier. (3)

Table 3

Datapoints	Class	Manhattan Distance of new observation
1	A	1
2	B	5
3	A	3
4	C	8
5	A	10
6	B	7
7	C	6

8	B	2
9	A	12
10	B	4

-----End-----