MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
(A constituent unit of MAHE, Manipal)

VII SEMESTER B.TECH (ELECTRICAL & ELECTRONICS ENGINEERING)
PROCTORED ONLINE MAKEUP EXAMINATIONS, FEBRAUARY 2022

DATA ANALYTICS [ELE 4077]
REVISED CREDIT SYSTEM

Time: 75 Minutes + 10 Minutes **Date: 19 February 2022** **Max. Marks: 20**

Instructions to Candidates:

- Answer **ALL** the questions.
- Missing data may be suitably assumed.
- Time: 75 minutes for writing + 10 minutes for uploading.

1A. The Python code shown in Fig. 1A, is broken down in 4 sections. Analyze each code sections and explain them.

```
#Section-1
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

#Section- 2
advertising = pd.read_csv("advertising.csv")
print(advertising.head())

print('\n')

#Section-3
print(advertising.shape)
print('\n')
print(advertising.info())
print('\n')
print(advertising.describe())
print('\n')

#Section-4
advertising.corr()
sns.heatmap(advertising.corr(),
            cmap="YlGnBu", annot = True)
plt.show()
```

Fig. 1A ***(03)***

1B. The histogram of Fig. 1C, is a plot of the number of students who were absent from Data Analytics course verse number of days they were absent.

i) Construct a box plot to represent the data.

ii) Describe the distribution of the data based on box plot interpretation. ***(03)***

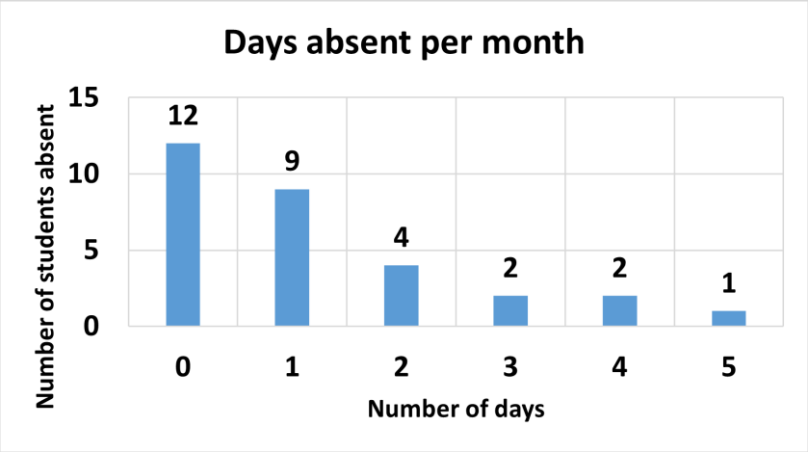

Fig. 1C

1C. The average heights of a random sample of 400 people from a city is 1.75 m. It is known that the heights of the population are random variables that follow a normal distribution with a variance of 0.16.

i) Determine the interval of 95% confidence for the average heights of the population.

ii) With a confidence level of 90%, what would the minimum sample size needed, to be in order for the true mean of the heights to be less than 2 cm from the sample mean? ***(04)***

2A. Assume that you have built a model to predict whether a patient is diabetic. In order to evaluate the performance of the model, you have constructed the confusion matrix as shown in Table 5B.

	Not Diabetic (Predicted)	Diabetic (Predicted)
Not Diabetic (Actual)	3269	366
Diabetic (Actual)	595	692

Table 2A

i) Analyze the accuracy of the model and comment on the fallacy/limitation of relying on accuracy only.

ii) Which other metric (or metrics) will you use to evaluate the model and justify your answer? Also, calculate the values of the metric (or metrics) used. ***(04)***

2B. Marketing head of a pharmaceutical company decided to analyze the relationship between Sales (Y) of medicines and the marketing expenditure (X) for the past several months. You, the data analyst of the company came up with a simple model.

A sample of 5 dataset is shown in Table 4A below. Y-pred is the predicted sales from your model. Use appropriate metric to evaluate the strength of the model and comment on it. ***(03)***

Marketing Budget (X) (In lakhs)	Actual Sales(Y) (In crores)	Predicted Sales (Y-pred)
127.40	10.50	10.08
364.40	21.40	22.59
150.00	10.00	11.27
128.70	9.60	10.15
285.90	17.40	18.45

Table 2B

2C. A dataset has 2 features and 6 observations as shown in the Table 5C. Take the initial centroids as (1, 4) & (6, 2).

i) Using appropriate algorithm, cluster them into 2 categories.

ii) Also calculate the sum of squared errors (SSE) for each interation.

Observation number	X1	X2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0

Table 2C ***(03)***