Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division
Second Semester 2019-2020
M.Tech (Data Science and Engineering)
Mid-Semester Exam (EC-2 Make-up)

Course No.	: DSECLZC415
Course Title	: Data Mining
Nature of Exam	: Open Book
Weightage	: 30%
Duration	: 90 minutes
Date of Exam	: 4/07/2020 (AN), 2:00 pm to 3:30 pm

No. of Pages = 3
No. of Questions = 4

Note to Students:
1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. **All parts of a question should be answered consecutively. Each answer should start from a fresh page.**
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Q.1. Answer the following:

a) Consider the following ordered list of observations of a variable: **[2+1+2+2=7 marks]**

25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 41, 42, 42, 99

1) What is the five-number summary for the given data?
2) Draw boxplot.
3) Considering univariate normal distribution, estimate the range of values that helps in estimating whether the given value is an outlier or not.
4) Explain, how do the outliers affect the measures of central tendency (Mean, and Median) of data? Comment using the given data set.

b) As many data mining algorithms cannot handle missing values, analyst sometimes remove all observations (rows) that contain missing values before the analysis. Give two potential disadvantages of this procedure. **[2 marks]**

c) Consider the following marks details of some students

Name	Marks
Abishek	*20*
Ramesh	30
Vinod	55
Rahul	75
Anu	95
Kavita	40
Latha	45
Pravin	57
Sonu	32
Sneha	65

Use min-max normalization to transform the above marks onto the range [50, 70]. **[3 marks]**

Q.2. The following table shows the house area(in sq meters) and house rent(in thousand rupees) obtained for a metropolitan city. **[4+1+1=6 marks]**

Area	172	150	181	174	194
Rent	42	35	46	40	50

a) Assuming linear relationship, Use the method of least squares to get an equation for the prediction of rent based on the area.
b) Predict the rent of a house with 180 sq. meter area.
c) Do you notice any limitations with the solution?

Q.3. Answer the following: **[1.5*4=6 marks]**

Suppose you are a Data scientist. You are building a Classifier that can predict whether a person is likely to default or Not based on certain parameters/attribute values. Assume, the class variable is "Default" and has two outcomes, {"yes", "no"}

- Own_House = Yes, No
- Marital Status = Single, Married, Divorced
- Annual Income = Low, Medium, High
- Currently Employed = Yes, No

Suppose a rule-based classifier produces the following rules:

1. Own_House = Yes → Default = Yes
2. Marital Status = Single → Default = Yes
3. Annual Income = Low → Default = Yes
4. Annual Income = High, Currently Employed = No → Default = Yes
5. Annual Income = Medium, Currently Employed = Yes → Default = No
6. Own_House = No, Marital Status = Married → Default = No
7. Own_House = No, Marital Status = Single → Default = Yes

Answer the following questions. **Make sure to provide a brief explanation or examples to illustrate the answer.**

a) Are the rules mutually exclusive?
b) Is the rule set exhaustive?
c) Is ordering needed for this set of rules?
d) Do you need a default class for the rule set?

Q.4. Based on the information given in the table below, find the most similar and most dissimilar persons among them. Apply min-max normalization on income to obtain [0,1] range. Consider Profession and City as nominal, and Age as an ordinal variable with ranking order of [Youth, Middle-Aged, Senior]. Give equal weight to each attribute. **[6 marks]**

Name	Profession	Age	City	Income
Sam	Doctor	Middle Aged	Mumbai	80000
John	Engineer	Youth	Delhi	50000
Mary	Politician	Senior	Bangalore	70000
Sonam	Doctor	Middle Aged	Delhi	89000
Sujoy	Carpenter	Middle-aged	Gurgaon	20000