Name: **Enrolment No:**	UPES UNIVERSITY WITH A PURPOSE

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
End Semester Examination, December 2020

Course: Data Mining & Predictive Analytics **Semester: VII**
Program: B.Tech CSE ECRA **Time : 03 hrs.**
Course Code: CSBA 4005 **Max. Marks: 100**

Instructions:

1. In Section A, you have to write one word/one sentence answers, no explanation, no calculation is required to be furnished.
2. Section B and C are the sections in which you will be writing the answers on A4 sheets and after clicking the picture, upload as per the direction.

SECTION A

S. No.		Marks	CO
Q 1	Discuss whether or not each of the following activities is a data mining task: a) Establishing a cricket player into All-rounder category by analyzing his batting and bowling statistics. b) Monitoring the oxygen level of a patient for abnormalities. c) Predicting the outcome of a hockey match based on past performances of two teams. d) Predicting the outcome of tossing pair of dice. e) Monitoring climatic conditions for tsunami.	**5**	**CO1**
Q 2	Name **THREE** data mining task which are ***descriptive*** in nature.	**5**	**CO2**
Q 3	State True/False a) ROC curve is drawn between TP and TN values. b) Accuracy of a classifier = 1- Error rate c) Accuracy of a Rule-based classifier is written as accuracy(R) = $\mathbf{n_{correct}}$ / $\mathbf{n_{covers}}$ d) DIANA and AGNES are two classification algorithms. e) Full form of ID in ID3 is …….	**5**	**CO3**
Q 4	In k-fold validation method, what does ***k*** denote? What is the value of ***k*** which is taken as standard?	**5**	**CO4**
Q 5	Name **FIVE** important characteristics of structured data.	**5**	**CO1**
Q 6	a) If there are N data objects, each having P attributes, what will be the dimension of Proximity Matrix?. b) In a data set, a variable is having values in the range 1000 to 9000. We wish to normalize these to a new range 0-5. What will be the equivalent of 5000 while we use min-max normalization.? c) What will be the Euclidean distance between two data points (5,7,10) and (6,8,2)? d) If we do the partitioning of dataset, and pick up the proportional volume from each partition, which type of sampling this is called? e) Name THREE data visualization techniques.	**5**	**CO2**

SECTION B

	For a given dataset – Classify whether the following student will buy or not, using Naïve Bayes Classifier *(Senior, high, No, excellent,?)*	**10**	**CO3**

<table>
<tr>
<td>Q 7</td>
<td>
<table>
<tr><th>RID</th><th>age</th><th>income</th><th>student</th><th>credit_rating</th><th>Class: buys_computer</th></tr>
<tr><td>1</td><td>youth</td><td>high</td><td>no</td><td>fair</td><td>no</td></tr>
<tr><td>2</td><td>youth</td><td>high</td><td>no</td><td>excellent</td><td>no</td></tr>
<tr><td>3</td><td>middle_aged</td><td>high</td><td>no</td><td>fair</td><td>yes</td></tr>
<tr><td>4</td><td>senior</td><td>medium</td><td>no</td><td>fair</td><td>yes</td></tr>
<tr><td>5</td><td>senior</td><td>low</td><td>yes</td><td>fair</td><td>yes</td></tr>
<tr><td>6</td><td>senior</td><td>low</td><td>yes</td><td>excellent</td><td>no</td></tr>
<tr><td>7</td><td>middle_aged</td><td>low</td><td>yes</td><td>excellent</td><td>yes</td></tr>
<tr><td>8</td><td>youth</td><td>medium</td><td>no</td><td>fair</td><td>no</td></tr>
<tr><td>9</td><td>youth</td><td>low</td><td>yes</td><td>fair</td><td>yes</td></tr>
<tr><td>10</td><td>senior</td><td>medium</td><td>yes</td><td>fair</td><td>yes</td></tr>
<tr><td>11</td><td>youth</td><td>medium</td><td>yes</td><td>excellent</td><td>yes</td></tr>
<tr><td>12</td><td>middle_aged</td><td>medium</td><td>no</td><td>excellent</td><td>yes</td></tr>
<tr><td>13</td><td>middle_aged</td><td>high</td><td>yes</td><td>fair</td><td>yes</td></tr>
<tr><td>14</td><td>senior</td><td>medium</td><td>no</td><td>excellent</td><td>no</td></tr>
</table>
</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q 8</td>
<td>Discuss various important and desirous characteristics of data which are really required for a reasonable good outcome of any data mining task. Here, in between you can discuss the various types of data for which data mining is applied.</td>
<td>10</td>
<td>CO2</td>
</tr>
<tr>
<td>Q 9</td>
<td>For what purpose k-Means algorithm is used. Write various standard steps of this algorithm. How the values of k are decided?
OR
Perform KNN- Classification algorithm on following dataset and predict the class for X (Height=1633 and Weight=57). Given K=3.
<table>
<tr><th>Height</th><th>Weight</th><th>Class</th></tr>
<tr><td>157</td><td>45</td><td>Underweight</td></tr>
<tr><td>162</td><td>76</td><td>Normal</td></tr>
<tr><td>153</td><td>48</td><td>Underweight</td></tr>
<tr><td>161</td><td>56</td><td>Underweight</td></tr>
<tr><td>155</td><td>51</td><td>Normal</td></tr>
<tr><td>172</td><td>78</td><td>Normal</td></tr>
</table>
</td>
<td>10</td>
<td>CO3</td>
</tr>
<tr>
<td>Q 10</td>
<td>For the given confusion matrix of a binary classification problem, find out Accuracy, Precision, Recall and F-Score for the classifier.
<table>
<tr><th>Actual Class ↓\Predicted Class→</th><th>Covid 19 = Yes</th><th>Covid 19 = No</th></tr>
<tr><td>Covid 19 = Yes</td><td>110</td><td>190</td></tr>
<tr><td>Covid 19 = No</td><td>240</td><td>9460</td></tr>
</table>
OR
Write a short note on the following. Short notes should essentially include the explanations with examples
a) Bagging c) Boosting
b) Precision d) ROC
</td>
<td>10</td>
<td>CO4</td>
</tr>
<tr>
<td>Q11</td>
<td>Explain various steps in KDD process. Also comment on the statement “Data preprocessing takes 60% of the efforts in entire process”.</td>
<td>10</td>
<td>CO1</td>
</tr>
</table>

SECTION-C

Q 12

Create a complete decision tree of the following data set using C 4.5 algorithm (based on the parameter ***Gain Ratio***)

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

20 **CO3**

OR

Q 12

a) Find out all frequent item sets in the follwing data set using ***FP Algorithm*** with min_sup = 25%.

b) Find all association rules with more than 80% confidence.

Transaction ID	Items Bought
1	Chips, Cookies, Regular Soda, Ham
2	Chips, Ham, Boneless Chicken, Diet Soda
3	Ham, Bacon, Whole Chicken, Regular Soda
4	Chips, Ham, Boneless Chicken, Diet Soda
5	Chips, Bacon, Boneless Chicken
6	Chips, Ham, Bacon, Whole Chicken, Regular Soda
7	Chips, Cookies, Boneless Chicken, Diet Soda

20 **CO3**