Name: **Enrolment No:**	UPES UNIVERSITY OF TOMORROW

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
End Semester Examination, December 2022

Course: Data Mining and Prediction Modeling
Semester: **V**
Program: B.Tech.(CSE + BAO)/ B.Tech.(CSE + BAO) (Hons.)
Time **: 03 hrs.**
Course Code: **CSBA 3001** **Max. Marks: 100**

Instructions:

SECTION A
(5Qx4M=20Marks)

S. No.		**Marks**	**CO**
Q1	Data preprocessing is a most important part of KDD process. Enlist FIVE characteristics of data which are essentially required to measure a data quality.	**4**	**CO1**
Q2	What are data mining tasks and two types of these? Discuss and give examples of 3 applications of each category of these tasks.	**4**	**CO2**
Q3	Name **FIVE** measures which are generally used to evaluate the performance of a classifier or classification method	**4**	**CO3**
Q4	What does it mean to deploy a machine learning model?	**4**	**CO4**
Q5	a) Data Mining is ----------in nature. (Hypothetical-based/Exploratory-based) b) Let, in a given data set, the mean **μ** is 40 and standard deviation **σ** is 16, what will be the z-score for the value 85? c) What is the supremum distance between two data points (3,9,1) and (2,6,2)? 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.	**4**	**CO2**

SECTION B
(4Qx10M= 40 Marks)

Q6	For a given dataset *(Youth, Low, No, Excellent, ,?)*, using **Naïve Bayes Classifier** classify whether customer will buy computer or not?	**10**	**CO3**

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Q7.	Write an algorithm for k-nearest neighbor classification given ***k***, the nearest number of neighbors, and ***n***, the number of attributes describing each tuple. ***OR*** Illustrate Neural Network Classifier. Discuss Back Propagation Algorithm and its working philosophy by taking suitable example.	**10**	**CO3**
Q8	How can the efficiency of a classifier be increased? Discuss various methods available to do so. ***OR*** Explain the terms: a) Model evaluation b) Model Validation c) Model Deployment d) Model Performance	**10**	**CO4**
Q9.	What do you understand by Sampling? Discuss various types of Sampling methods.	**10**	**CO2**
	SECTION-C **(2Qx20M=40 Marks)**		
Q 10	Create a complete decision tree of the following data set using C 4.5 algorithm (based on the parameter ***Gain Ratio***) OR Create a complete decision tree of the following data set using ID3 algorithm. (based on the parameter ***Information Gain***)	**20**	**CO3**

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

Q11	Write short note on the following: a) Support Vector Machine b) Artificial Neural Network c) Sampling d) Confusion Matrix	**20**	**CO1, CO4**