


Name: Enrolment No:	
--------------------------------------	--

UPES
End Semester Examination, December 2023

Course: Data Mining and Prediction Modeling
Semester: V
Program: B. Tech CSE- BAO
Course Code: CSBA 3001

Time: 03 hrs.
Max. Marks: 100

Instructions: All Questions are compulsory. Internal Choice is mentioned in the paper

SECTION A
(5Qx4M=20Marks)

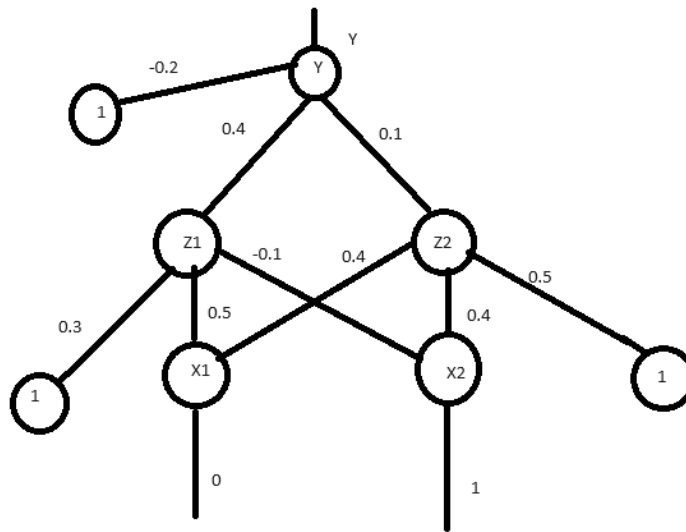
S. No.		Marks	CO
Q 1	Outliers are often discarded as noise. However, one person’s garbage could be another’s treasure. For example, exceptions in credit card transactions can help us detect the fraudulent use of credit cards. Taking fraudulence detection as an example, propose two methods that can be used to detect outliers and discuss which one is more reliable.	4 marks	CO1
Q 2	<p>(1) Movie Recommendation system is an example of:</p> <p style="padding-left: 40px;">Classification Clustering Reinforcement Learning Regression</p> <p>(a) 2 only (b) 1 and 2 (c) 1 and 3 (d) 2 and 3 (e) 1, 2 and 3 (f) 1, 2, 3, and 4</p> <p>(2) Sentiment Analysis is an example of:</p> <p style="padding-left: 40px;">Classification Reinforcement Learning Clustering Regression</p> <p>(a) 1 only (b) 1 and 2 only (c) 1, 2 and 3 (d) 1, 2 and 4 (e) 1,2, 3 and 4</p> <p>(3) Can Decision Trees be used for performing Clustering?</p>	4 marks	CO3

	(a) True (b) False (4) What is the minimum number of variables/features required to perform clustering? (a) 0 (b) 1 (c) 2 (d) 3																
Q 3	Discuss the methods of handling missing data.	4 marks	CO2														
Q 4	Discuss antecedent and consequent with real world examples.	4 marks	CO3														
Q 5	Explain variables and its types.	4 marks	CO1														
SECTION B (4Qx10M= 40 Marks)																	
Q 6	Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. (a) What is the mean of the data? What is the median? (b) What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.). (c) What is the midrange of the data? (d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data? (e) Give the five-number summary of the data. (f) Show a boxplot of the data. (g) How is a quantile-quantile plot different from a quantile plot?	10 marks	CO4														
Q 7	Suppose that the values for a given set of data are grouped into intervals. The intervals and corresponding frequencies are as follows : <table style="margin-left: 20px;"> <thead> <tr> <th>age</th> <th>frequency</th> </tr> </thead> <tbody> <tr> <td>1-5</td> <td>200</td> </tr> <tr> <td>5-15</td> <td>450</td> </tr> <tr> <td>15-20</td> <td>300</td> </tr> <tr> <td>20-50</td> <td>1500</td> </tr> <tr> <td>50-80</td> <td>700</td> </tr> <tr> <td>80-110</td> <td>44</td> </tr> </tbody> </table> Compute an approximate median value for the data.	age	frequency	1-5	200	5-15	450	15-20	300	20-50	1500	50-80	700	80-110	44	10 marks	CO2
age	frequency																
1-5	200																
5-15	450																
15-20	300																
20-50	1500																
50-80	700																
80-110	44																
Q 8	Explain Market basket analysis with example. Define support, confidence, and tilt with mathematical operations.	10 marks	CO5														

Q 9	Compare and Contrast OLAP and OLTP.	10 marks	CO4
-----	-------------------------------------	----------	-----

SECTION-C
(2Qx20M=40 Marks)

Q 10 Using backpropagation_ network, find the new weights the shown in Figure. It is presented with the input pattern [0, 1] and the target output is 1. Use a learning rate $\alpha = 0.25$ and binary sigmoidal activation function.



20 marks

CO5

Q 11
(1)

age	income	student	credit_rating	Class: Buy computer
25	high	no	fair	no
27	high	no	excellent	no
32	high	no	fair	yes
?	medium	?	fair	yes
45	low	yes	fair	yes
43	low	yes	?	no
?	low	yes	excellent	yes
28	medium	no	fair	no
29	low	yes	fair	yes
47	medium	yes	fair	yes
21	medium	yes	excellent	yes
31	medium	no	excellent	yes
36	high	yes	fair	yes
42	high	no	excellent	no

- (a) For the following questions, the reference data set is given at the top in table form. Before proceeding further, work out for the missing values denoted by '?'.
- (b) Use Bayes' model to predict the final decision for the new instance (20, low, yes, fair)
- (c) Propose a method to discretize the numerical attribute 'age' first and then construct a decision tree based on **information gain** for the given data set.
- (d) Reconstruct a decision tree based on **gain ratio** for the given data set. Compare this new decision tree with the one you obtained in Q3.

OR

(2)

Write Short Notes on:

- (a) Applications of Data Mining
(b) Classification and its algorithms

CO5

20 marks

	(c) Clustering and its algorithms (d) Steps involved in Pre-Processing		
--	---	--	--