


Name:	 UPES UNIVERSITY WITH A PURPOSE
Enrolment No:	

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
School of Computer Science

End Semester Examination, December 2020

Course : Data Mining & Prediction by Machines	Semester : III
Program : B. Tech CSE AIML	Time : 03 Hours
Course Code : CSAI 2005	Max. Marks : 100
Instructions :	

SECTION A

		Marks	
Q1	Why data mining is in high demand and what kind of data can be mined?	05	CO1
Q2	What is the difference between classification and regression? Why data classification is known as a two-step process?	05	CO3
Q3	Given two objects represented by the tuples (22, 1, 42,10) and (20, 0, 36, 8): a) Compute the Euclidian distance between the two objects. b) Compute the Manhattan distance between the two objects.	05	CO2
Q4	What are the different methods of measuring central tendency of data set?	05	CO2
Q5	How bagging method can be useful to improve the accuracy of classification machine models?	05	CO3
Q6	Give an application example of where the border between normal objects and outliers is often unclear, so that the degree to which an object is an outlier has to be well estimated.	05	CO4

SECTION B

Q7	Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location into three clusters: A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9): The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. a) Write down k-means algorithm b) Use k-means algorithm for the three cluster centers after the first round execution c) Find the final three clusters	4+4+2 =10	CO4
----	---	----------------------	------------

Q8	What do you mean by Process Standardization? Briefly explain the CRISP-DM phases and tasks.	10	CO1																																			
Q9	<p>Explain KNN algorithm. Why it is also called Lazy Learner? What are the points to be subjected when choosing the value of k? For the below problem predict for the class of Davis using KNN and assume the value of k=3.</p> <table border="1" data-bbox="472 485 1024 877"> <thead> <tr> <th>Customer</th> <th>Age</th> <th>Income (K)</th> <th>No. of cards</th> <th>Response</th> </tr> </thead> <tbody> <tr> <td>John</td> <td>35</td> <td>35</td> <td>3</td> <td>Yes</td> </tr> <tr> <td>Rachel</td> <td>22</td> <td>50</td> <td>2</td> <td>No</td> </tr> <tr> <td>Ruth</td> <td>63</td> <td>200</td> <td>1</td> <td>No</td> </tr> <tr> <td>Tom</td> <td>59</td> <td>170</td> <td>1</td> <td>No</td> </tr> <tr> <td>Neil</td> <td>25</td> <td>40</td> <td>4</td> <td>Yes</td> </tr> <tr> <td>David</td> <td>37</td> <td>50</td> <td>2</td> <td>?</td> </tr> </tbody> </table>	Customer	Age	Income (K)	No. of cards	Response	John	35	35	3	Yes	Rachel	22	50	2	No	Ruth	63	200	1	No	Tom	59	170	1	No	Neil	25	40	4	Yes	David	37	50	2	?	10	CO3
Customer	Age	Income (K)	No. of cards	Response																																		
John	35	35	3	Yes																																		
Rachel	22	50	2	No																																		
Ruth	63	200	1	No																																		
Tom	59	170	1	No																																		
Neil	25	40	4	Yes																																		
David	37	50	2	?																																		
Q10	<p>A data mining model is trained to predict COVID in patients. The test dataset consists of 100 people. Following is the confusion matrix for the same.</p> <table border="1" data-bbox="431 1016 1089 1257"> <tr> <td colspan="2" rowspan="2"></td> <td colspan="2">Predicted</td> </tr> <tr> <td>Negative</td> <td>Positive</td> </tr> <tr> <td rowspan="2">Actual</td> <td>Negative</td> <td>60</td> <td>22</td> </tr> <tr> <td>Positive</td> <td>8</td> <td>10</td> </tr> </table> <p>Calculate the following evaluation parameters :</p> <ol style="list-style-type: none"> 1. True positive 2. True Negative 3. False Positive 4. False Negative 5. Precision 6. Recall 7. Accuracy 8. F1 Score 9. Sensitivity 10. Specificity 			Predicted		Negative	Positive	Actual	Negative	60	22	Positive	8	10	10	CO4																						
				Predicted																																		
		Negative	Positive																																			
Actual	Negative	60	22																																			
	Positive	8	10																																			
Q11	<p>For a given data set compute dissimilarity between two binary attributes.</p> <table border="1" data-bbox="203 1507 1292 1793"> <thead> <tr> <th>Name</th> <th>T1</th> <th>T2</th> <th>T3</th> <th>T4</th> <th>T5</th> <th>T6</th> </tr> </thead> <tbody> <tr> <td>A</td> <td>1</td> <td>0</td> <td>1</td> <td>0</td> <td>0</td> <td>0</td> </tr> <tr> <td>B</td> <td>1</td> <td>0</td> <td>1</td> <td>0</td> <td>1</td> <td>0</td> </tr> <tr> <td>C</td> <td>1</td> <td>1</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> </tr> </tbody> </table> <p style="text-align: center;">OR</p>	Name	T1	T2	T3	T4	T5	T6	A	1	0	1	0	0	0	B	1	0	1	0	1	0	C	1	1	0	0	0	0	10	CO2							
Name	T1	T2	T3	T4	T5	T6																																
A	1	0	1	0	0	0																																
B	1	0	1	0	1	0																																
C	1	1	0	0	0	0																																

What are the various steps to perform data pre-processing? Explain each steps with suitable examples.

Section C

Q12

- a) Consider the data as two-dimensional data points. Given a new data point, $x = (1.4, 1.6)$ as a query, rank the database points based on similarity with the query using Euclidean distance, Manhattan distance, Minkowski distance.
- b) Following dataset is used to learn a decision tree which predicts if a student passed data mining and prediction by machine (Yes or No), based on their previous GPA (High, Medium, or Low) and whether or not they studied. Draw the decision tree for the same. Also, show the calculations regarding entropy and information gain.

GPA	Studied	Passed
Low	False	No
Low	True	Yes
Medium	False	No
Medium	True	Yes
High	False	Yes
High	True	Yes

20

CO3

Or

Explain the decision tree induction algorithm. Create a complete decision tree of the following data set using ID3 algorithm. (based on the parameter **Information Gain**)

Sore throat	Fever	Swollen Glands	Congestion	Headache	Diagnosis
Yes	Yes	Yes	Yes	Yes	Strep throat
No	No	No	Yes	Yes	Allergy
Yes	Yes	No	Yes	No	Cold
Yes	No	Yes	No	No	Strep throat
No	Yes	No	Yes	No	Cold
No	No	No	Yes	No	Allergy
No	No	Yes	No	No	Strep throat
Yes	No	No	Yes	Yes	Allergy
No	Yes	No	Yes	Yes	Cold
Yes	No	No	Yes	Yes	Cold