Name:

Enrolment No:

UPES
UNIVERSITY WITH A PURPOSE

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
End Semester Examination, December 2019

Course: B.Tech CSE+BAO	**Semester: III**
Program: Data Mining and Prediction Modeling	**Time : 03 hrs.**
Course Code: CSBA3001	**Max. Marks: 100**

Instructions:

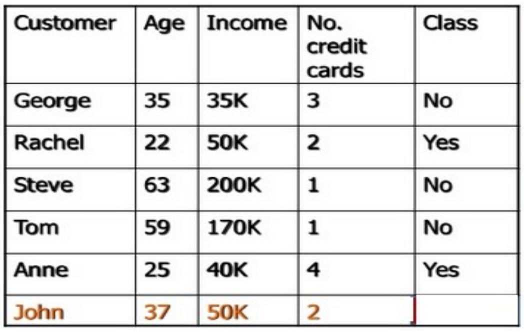
SECTION A

<table>
<tr><th>S. No.</th><th></th><th>Marks</th><th>CO</th></tr>
<tr><td>Q 1</td><td>Define Data Mining. Write down five application of it.</td><td>4</td><td>CO1</td></tr>
<tr><td>Q 2</td><td>Write down major issues of data mining.</td><td>4</td><td>CO1</td></tr>
<tr><td>Q 3</td><td>Write down the techniques to Improve Classification Accuracy.</td><td>4</td><td>CO4</td></tr>
<tr><td>Q 4</td><td><table>
<tr><td>Time Point</td><td>Reliance Industries</td><td>ONGC</td></tr>
<tr><td>Jan 2019</td><td>6</td><td>20</td></tr>
<tr><td>Feb 2019</td><td>5</td><td>10</td></tr>
<tr><td>March 2019</td><td>4</td><td>14</td></tr>
<tr><td>April 2019</td><td>3</td><td>5</td></tr>
<tr><td>May 2019</td><td>2</td><td>5</td></tr>
</table>
It is given the average stock price of Reliance and ONGC for five consecutive months. Find it either the stock price are independent to each other or not.</td><td>4</td><td>CO2</td></tr>
<tr><td>Q 5</td><td>Differentiate between regression and association with formula.</td><td>4</td><td>CO2</td></tr>
</table>

SECTION B

Q 6	What do you mean by Process Standardization? Briefly explain the CRISP-DM phases and tasks.	10	CO1
Q 7	Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8): (a) Compute the Euclidean distance between the two objects. (b) Compute the Manhattan distance between the two objects. (c) Compute the Minkowski distance between the two objects, using q D 3. (d) Compute the supremum distance between the two objects	2.5x4= 10	CO2
Q 8	Explain the basis of Model Evaluation and selection. Suppose there are two models M1 and M2. For M1: TP=6954, FN=46, FP=412 and TN=2588 For M2: TP=6800, FN=134, FP=566 and TN=2500 Calculate Accuracy, Recall, Specificity, Sensitivity and Z-Score. Among M1 and M2 which one is more preferable model?	10	CO4
Q 9	Explain KNN algorithm. Why it is also called Lazy Learner? What are the points to	10	CO3

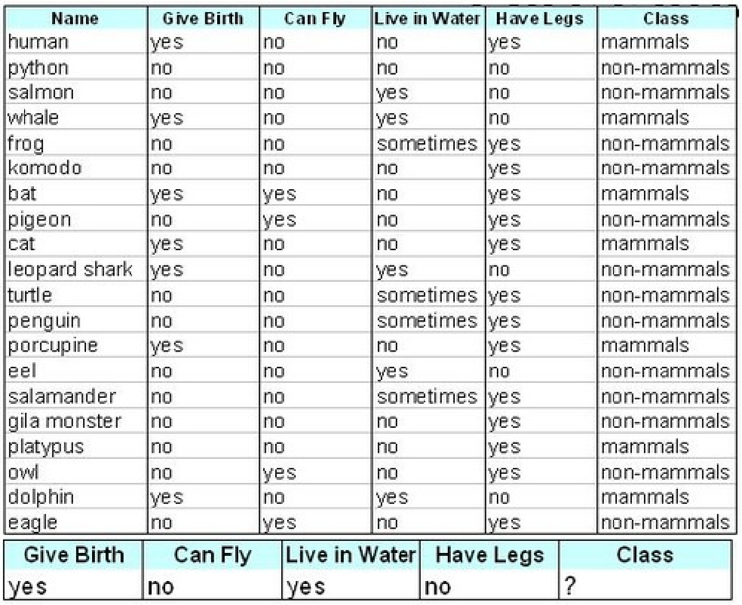
be subjected when choosing the value of k?



Customer	Age	Income	No. credit cards	Class
George	35	35K	3	No
Rachel	22	50K	2	Yes
Steve	63	200K	1	No
Tom	59	170K	1	No
Anne	25	40K	4	Yes
John	37	50K	2	

OR

Discuss Bayesian Classification Algorithm. Apply this algorithm for given data set:

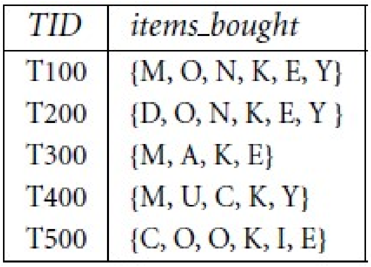


Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

SECTION-C

Q 10 A database has five transactions. Let min_sup 60% and min_conf 80%.



TID	*items_bought*
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I, E}

10+10 =20

CO3

Find frequent itemsets of set 3, using Apriori and FP-Growth, respectively. Compare the efficiency of the two mining processes.

Q 11

a) Explain and discuss the SVM Classification algorithm with advantages and limitations.

b) When an anthropologist finds skeletal remains, they need to figure out the height of the person. The height of a person (in cm) and the length of their metacarpal bone 1 (in cm) were collected and are in table. Find a regression equation between the height of a person and the length of their metacarpal. Then use the regression equation to find the height of a person for a metacarpal length of 44 cm and for a metacarpal length of 55 cm. Which height that you calculated do you think is closer to the true height of the person? Why?

Data of Metacarpal versus Height

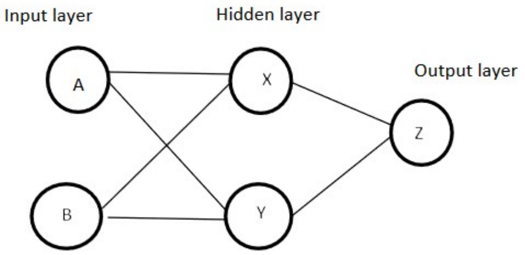
Length of Metacarpal (cm)	Height of Person (cm)
45	171
51	178
39	157
41	163
48	172
49	183
46	173
43	175
47	173

20

CO3

OR

.



Input		Output
A	B	Z
0	0	0
0	1	1
1	0	1
1	1	1

Learning rate=0.35

Biases are θx = θy = θz = 0. Neural Network of above diagram has two nodes (A,B) in the input layer, two nodes in the hidden layer (X,Y)and one node in the output layer (Z). The values given to weights are taken randomly and will be changed during back propagation iterations. Initial weights of the top input nodes taken at random are 0.4, 0.1 .Weights of bottom input node are 0.8 and 0.6. Weights of top hidden node is 0.3 and that of bottom hidden node is 0.9.