Name: Enrolment No:	UPES

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
End Semester Examination, December 2018

Course: Data Mining & Prediction Modeling **Semester: V**
Programme: B. Tech CSE+BAO **Course Code:CSIB 338**
Time: 03 hrs. **Max. Marks: 100**
Instructions:

SECTION A

S. No.	All the questions are compulsory and carries equal marks.	Marks	CO
Q 1	Write down in brief the importance of statistics in data mining.	4	CO2
Q 2	Define KDD. How it is different from data mining?	4	CO1
Q 3	Explain Gaussian Distribution. How it is useful for prediction?	4	CO2
Q 4	Differentiate between Chi-Square and Lift with the help of simple example.	4	CO2
Q 5	Is it possible that Support will be equal to Confidence? If yes write down the condition.	4	CO3

SECTION B

	All questions are compulsory and carry 10 marks		
Q 6	Assume 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) Use smoothing by bin means to smooth the data, using a bin depth of 3.Illustrate your steps. Comment on the effect of this technique for the given data. (b) How can you determine outliers in the data? (c) What other methods are there for data smoothing?	5+3+2	CO2
Q 7	Explain the basis of Model Evaluation and selection. Assume 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 Among M1 and M2 which one is more preferable model?	5+5	CO4
Q 8	Define a pattern in data mining. Assume that a person is driving a car, while listening to music. The car's driver assistance system detects a pedestrian ahead on the road and estimates that the driver needs to apply his brakes within 10-15 ms to avoid an accident. Which of the following alarms will be most effective in this case? 1. Audio alarm, e.g. beeps from the speakers. 2. Visual alarm, e.g. flashing light on the panel. 3. Vibro-tactile alarm, e.g. a vibration on the seat.	3+7	CO1

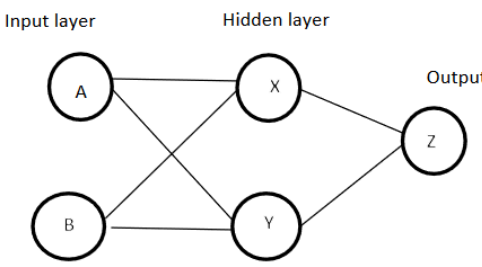
	Justify your answer. Assume that the car has all the three provisions.		
Q 9	Discuss the Process Standardization? Briefly explain the CRISP-DM phases and tasks with diagram. (2+8) **OR** Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location into 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 (4) b) Use k-means algorithm for the three cluster centers after the first round execution (4) c) Find the final three clusters (2)	**10**	**CO1** **CO3**
SECTION-C			
	All questions are compulsory and carry 20 marks		
Q 10	Classification predicts categorical labels (classes), prediction models continuous-valued functions. Justify the statement. Explain the Naïve Bayesian Classification Algorithm. With the help of given table, for the day <sunny, cool, high, strong>, what's the play prediction?	**(2+4+8 +6)**	**CO2, CO3**

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

For the same data justify another model and compare the result of both.

Q 11

20 **CO3**



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

Learning rate=0.35

Biases are $\theta_x = \theta_y = \theta_z = 0.9$. 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 $W_{AX} = 0.4$, $W_{AY}=0.1$. Weights of bottom input node are $W_{BX}=0.8$ and $W_{BY}=0.6$. Weights of top hidden node is $W_{XZ}=0.3$ and that of bottom hidden node is $W_{YZ}=0.9$.

Solve this problem and prove that backpropagation is helping to reduce the loss function for the data set.

	OR "The support vector machine is highly accurate classification method", justify the statement. SVM classifier suffers from slow processing when training with a large data set, why? How we can solve this problem and make the SVM scalable. Categorize the types of hyperplane, if any. Explain with the concept of projection (orthonormal).		

Name: **Enrolment No:**	UPES

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
End Semester Examination, December 2018

Course: Data Mining & Prediction Modeling **Semester: V**
Programme: B. Tech CSE+BAO **Course Code:CSIB 338**
Time: 03 hrs. **Max. Marks: 100**
Instructions:

SECTION A

S. No.	**All the questions shall be compulsory and carries equal marks.**	**Marks**	**CO**
Q 1	What do you mean by "Data Migration Tool"? Discuss any one tool of such type	**4**	**CO2**
Q 2	Differentiate between KDD and Data Mining.	**4**	**CO1**
Q 3	Explain Normal Distribution. How it is useful for prediction?	**4**	**CO1**
Q 4	What is data mining? Briefly explain major issues of data mining	**4**	**CO2**

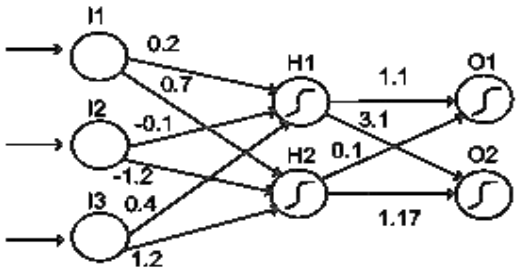
<table>
<tr><td>Q 5</td><td>Briefly explain purpose of CHAID.</td><td>4</td><td>CO4</td></tr>
<tr><td colspan="4">SECTION B</td></tr>
<tr><td></td><td>All questions are compulsory and carry 10 marks</td><td></td><td></td></tr>
<tr><td>Q 6</td><td>Differentiate between Classification and Clustering. Briefly describe the major steps of Feed forward Back Propagation algorithm for classification</td><td>3+7</td><td>CO3</td></tr>
<tr><td>Q 7</td><td>Explain the basis of Model Evaluation and selection. Suppose there are two models M1 and M2.

For M1: TP=6900, FN=100, FP=412 and TN=2588
For M2: TP=6800, FN=134, FP=500 and TN=2566
Among M1 and M2 which one is more preferable model?</td><td>5+5</td><td>CO4</td></tr>
<tr><td>Q 8</td><td>Analyze and explain Decision Tree Induction Classification algorithm.</td><td>10</td><td>CO3</td></tr>
<tr><td>Q 9</td><td>Discuss the Process Standardization? Briefly explain the CRISP-DM phases and tasks with diagram.

OR

Analyze and explain the SVM Classification algorithm with advantages and limitations.</td><td>10</td><td>CO1

CO3</td></tr>
<tr><td colspan="4">SECTION-C</td></tr>
<tr><td></td><td>All questions are compulsory and carry 20 marks</td><td></td><td></td></tr>
<tr><td>Q 10</td><td>Classification predicts categorical labels (classes), prediction models continuous-valued functions. Justify the statement. Explain the Naïve Bayesian Classification Algorithm.
<table>
<tr><th>Name</th><th>Blood Type</th><th>Give Birth</th><th>Can Fly</th><th>Live in Water</th><th>Class</th></tr>
<tr><td>human</td><td>warm</td><td>yes</td><td>no</td><td>no</td><td>mammals</td></tr>
<tr><td>python</td><td>cold</td><td>no</td><td>no</td><td>no</td><td>reptiles</td></tr>
<tr><td>salmon</td><td>cold</td><td>no</td><td>no</td><td>yes</td><td>fishes</td></tr>
<tr><td>whale</td><td>warm</td><td>yes</td><td>no</td><td>yes</td><td>mammals</td></tr>
<tr><td>frog</td><td>cold</td><td>no</td><td>no</td><td>sometimes</td><td>amphibians</td></tr>
<tr><td>komodo</td><td>cold</td><td>no</td><td>no</td><td>no</td><td>reptiles</td></tr>
<tr><td>bat</td><td>warm</td><td>yes</td><td>yes</td><td>no</td><td>mammals</td></tr>
<tr><td>pigeon</td><td>warm</td><td>no</td><td>yes</td><td>no</td><td>birds</td></tr>
<tr><td>cat</td><td>warm</td><td>yes</td><td>no</td><td>no</td><td>mammals</td></tr>
<tr><td>leopard shark</td><td>cold</td><td>yes</td><td>no</td><td>yes</td><td>fishes</td></tr>
<tr><td>turtle</td><td>cold</td><td>no</td><td>no</td><td>sometimes</td><td>reptiles</td></tr>
<tr><td>penguin</td><td>warm</td><td>no</td><td>no</td><td>sometimes</td><td>birds</td></tr>
<tr><td>porcupine</td><td>warm</td><td>yes</td><td>no</td><td>no</td><td>mammals</td></tr>
<tr><td>eel</td><td>cold</td><td>no</td><td>no</td><td>yes</td><td>fishes</td></tr>
<tr><td>salamander</td><td>cold</td><td>no</td><td>no</td><td>sometimes</td><td>amphibians</td></tr>
<tr><td>gila monster</td><td>cold</td><td>no</td><td>no</td><td>no</td><td>reptiles</td></tr>
<tr><td>platypus</td><td>warm</td><td>no</td><td>no</td><td>no</td><td>mammals</td></tr>
<tr><td>owl</td><td>warm</td><td>no</td><td>yes</td><td>no</td><td>birds</td></tr>
<tr><td>dolphin</td><td>warm</td><td>yes</td><td>no</td><td>yes</td><td>mammals</td></tr>
<tr><td>eagle</td><td>warm</td><td>no</td><td>yes</td><td>no</td><td>birds</td></tr>
</table>
Find the class label of given species:
<table>
<tr><th>Name</th><th>Blood Type</th><th>Give Birth</th><th>Can Fly</th><th>Live in Water</th><th>Class</th></tr>
<tr><td>hawk</td><td>warm</td><td>no</td><td>yes</td><td>no</td><td>?</td></tr>
<tr><td>grizzly bear</td><td>warm</td><td>yes</td><td>no</td><td>no</td><td>?</td></tr>
</table>
For the same data justify another model and compare the result of both.</td><td>(2+6+6 +6)</td><td>CO1, CO3</td></tr>
<tr><td>Q 11</td><td></td><td></td><td>CO1, CO2, CO3</td></tr>
</table>



It is given that**:**

I1=10, I2=30, I3=10, learning rate (l) = 0.1, t1(E)=target value for O1=1, t2(E)=target value for O2=0, ⍉ = 0 for all nodes. Implement feed forward NN to minimize the error.

20

OR

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 and Manhattan distance, implement for both and find the difference in results. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively.

(8+6+6)

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.