Name: **Enrolment No:**	UPES

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

Course: Data Mining & Business Intelligence **Semester: I**
Programme: M.Tech. (Computer Science & Engineering) **Code: CSDA 7001**
Time: 03 hrs. **Max. Marks: 100**
Instructions:

SECTION A

S. No.		Marks	CO
Q1.	a) Bronze, Silver, and Gold medals as awarded at the Olympics will be termed as …….. (***Ordinal / nominal***) attribute. b) Suppose a group of 15 cricket innings scores of a batsman has been recorded in sorted form as ***15, 29, 35, 42, 53, 56, 71, 82, 82, 95, 105, 128, 159, 193, 258.*** We wish to partition them into three bins. The elements in bins using **equi-depth** partitioning will be (…….), (…….) and (…….) AND using **equi-width** partitioning will be (…….), (…….) and (…….). c) ***(Mean/Median/Mode)*** ……is the <u>most sensitive</u> to the outliers while (***Mean/Median/Mode***) ……..is <u>least sensitive</u>. d) Clustering is an example of……..learning while classification is an example of ……..learning.(***Supervised/Unsupervised***)	**4**	**CO1/ CO2/ CO3**
Q2.	Suppose that you are employed as a data-mining consultant for an ***Internet search engine company***. Describe how data mining can help the company by giving **<u>one specific example</u>** of how techniques, such as clustering, classification, association rule mining, and anomaly detection can be applied.	**4**	**CO1/ CO2/ CO3**
Q3.	Analyze the role of data mining and business intelligence applications in banking.	**4**	**CO4**
Q4.	What do you understand by term "Big-Data"? Justify your answer by taking suitable examples.	**4**	**CO5**
Q5.	Critically comment on the statement ***"Data Pre-processing takes 60% of the total efforts in data-mining"*** The answer must contain the examples to validate the justification.	**4**	**CO1**
	SECTION B		
Q6.	Explain linear and non-linear regression using an example. Also, explain classification by back-propagation method of neural networks.	**10**	**CO3**
Q7.	Describe OLTP and OLAP. How Meta-data models are used in data mining, illustrate.	**10**	**CO1**
Q8.	Elaborate how data mining and business intelligence applications play an importance role in production with respect to various industries.	**10**	**CO4**
Q9.	What is a frequent item in a transactional data set? Find all frequent item-sets in the following data set by using Apriori algorithm. Given min_support Count = 2	**10**	**CO2**

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

OR

What is a frequent item in a transactional data set? Find all frequent item-sets in the following data set by using FP-tree algorithm. Given min_support Count = 3

TID	***Items bought***	***(ordered) frequent items***
100	***{f, a, c, d, g, i, m, p}***	***{f, c, a, m, p}***
200	***{a, b, c, f, l, m, o}***	***{f, c, a, b, m}***
300	***{b, f, h, j, o, w}***	***{f, b}***
400	***{b, c, k, s, p}***	***{c, b, p}***
500	***{a, f, c, e, l, p, m, n}***	***{f, c, a, m, p}***

SECTION-C

Q10.

Consider the training examples shown in below table for a binary classification problem:

Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

a) What is the entropy of this collection of training examples with respect to positive class?
b) What are the information gains of **a_1** and **a_2** relative to these training examples?
c) What is the best split (among **a_1 , a_2 and a_3)** according to the information gain?

8+8+4 | **CO3**

Write a comprehensive note on Hadoop architecture, which also includes its storage system.

20 | **CO5**

	OR	**OR**	
Q11.	Write the lucid notes on the following: i) Spatial Mining ii) Web Mining iii) Text Mining iv) Distributed File System	**5 *4**	