

Name:  
Enrolment No:



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

Course: Data Mining & Predictive Analytics  
Program: B.Tech. (CSE with ECARA)  
Course Code: CSIT 401

Semester : VII  
Time : 03 hrs.  
Max. Marks: 100

Instructions:

SECTION A

		Marks	CO
Q 1	Define term <i>data mining</i> . What are different types of data in context of data mining? Differentiate between structured and unstructured data.	1+2+1	CO1
Q 2	Define <i>Sampling</i> . Discuss and distinguish among various sampling techniques.	1+3	CO2
Q 3	What are <i>frequent items</i> in a transaction? How association rules are generated from frequent items.	1+3	CO3
Q 4	What do understand by <i>model overfitting</i> ? Enlist couple of reasons for it.	1+3	CO4
Q 5	Discuss various challenges posed to data mining.	04	CO1

SECTION B

Q 6	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 THREE specific examples of how techniques, such as clustering, classification, association rule mining, and anomaly detection can be applied.	10	CO2
Q 7	Discuss different types of data on which mining can be performed and thus define various applications of data mining.	10	CO1
Q 8	What are major challenges in cluster analysis? Illustrate k-Means Clustering algorithm in brief. For your support, you may take an example. <b>OR</b> Illustrate Naïve Bayes Classifier. Discuss its working philosophy by taking suitable example.	10	CO3
Q 9	How accuracy of a classifier model can be increased? Discuss TWO popular ensemble methods.	2+8	CO4

SECTION-C

Q10	Create various possible association rules including 3 items from the market basket transactions shown in the below table. Use <i>Apriori</i> algorithm for generation of frequent item sets. Show all steps explicitly. (Assume $min\_sup = 2$ , $confidence = 100\%$ )	20	CO3
-----	--	----	-----

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

**OR**

Calculate all frequent item sets (1/2/3...) using *FP-Tree algorithm* in the following transactional data set, assuming minimum support count as **3**.

<u>TID</u>	<u>Items bought</u>
100	{f, a, c, d, g, i, m, p}
200	{a, b, c, f, l, m, o}
300	{b, f, h, j, o}
400	{b, c, k, s, p}
500	{a, f, c, e, l, p, m, n}

Q11	<p>Write lucid notes on</p> <ul style="list-style-type: none"> <li>a) Handling missing values</li> <li>b) Nearest-neighbor classifier</li> <li>c) Rule Induction using CHID</li> <li>d) ROC curves</li> </ul>	20	CO1, CO2, CO4
-----	---	----	---------------------