Reg No.:_______________ Name:__________________________

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

SIXTH SEMESTER B.TECH DEGREE EXAMINATION(R&S), MAY 2019

Course Code: IT304

Course Name: Data Warehousing and Mining

Max. Marks: 100 Duration: 3 Hours

PART A

Answer any two full questions, each carries 15 marks. Marks

1 a) Explain Three tier Data warehouse architecture and its components. (6)

b) Briefly explain the common types of data transformation techniques with suitable examples. (5)

c) Differentiate OLTP and OLAP (4)

2 a) Explain thedifferent schemas used for multi-dimensional databases. (6)

b) Discuss the issues to be considered during data cleaning. Explain how to handle noisy data in data cleaning process. (6)

c) Mention some popular data mining tools. (3)

3 a) Explain the relevance of data preprocessing in datamining. Explain the methods to handle missing values in a data set before mining process? (6)

b) Suppose that a data warehouse consists of **three dimensions time, doctor and patient** and two measures count and charge, where charge is the fee that a doctor charges a patient for a visit. Draw schema diagram for above data warehouse using Snowflake model. (5)

c) Mention challenges and applications of data warehousing (4)

PART B

Answer any two full questions, each carries 15 marks.

4 a) For the given data set, find the best split attribute at root level using ID3 algorithm. (9)

Gender	**Car Ownership**	**Travel Cost**	**Income**	**Transport Mode(Class)**
Male	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	1	Expensive	High	Car
Male	2	Expensive	Medium	Car
Female	2	Expensive	High	Car

Female	1	Cheap	Medium	Train
Male	0	Standard	Medium	Train
Female	1	Standard	Medium	Train

b) Differentiate classification and prediction. Mention the major issues in classification and prediction. (6)

5 a) Consider a training data set consisting of the fauna of the world. Each unit has three features named “Swim”, “Fly” and “Crawl”. Use Naive Bayesian algorithm to classify a particular species if its features are (Slow, Rarely, No). (10)

Sl No	**Swim**	**Fly**	**Crawl**	**Class**
1	Fast	No	No	Fish
2	Fast	No	Yes	Animal
3	Slow	No	No	Animal
4	Fast	No	No	Animal
5	No	Short	No	Bird
6	No	Short	No	Bird
7	No	Rarely	No	Animal
8	Slow	No	Yes	Animal
9	Slow	No	No	Fish
10	Slow	No	Yes	Fish
11	No	Long	No	Bird
12	Fast	No	No	Bird

b) Explain the relevance of attribute selection measures used in Decision Tree. How does Information gain differ from Gain ratio? (5)

6 a) The sales of a company (in million dollars) for each year are shown in the table below. (8)

x (year)	2005	2006	2007	2008	2009
y (sales)	12	19	29	37	45

a) Find the least square regression line y = a x + b.

b) Use the least squares regression line as a model to estimate the sales of the company in 2012.

b) Explain Back Propagation algorithm used in Neural Networks with an example. (7)

PART C

Answer any two full questions, each carries20 marks.

7 a) Suppose that our task is to cluster data pointsinto two clusters.Let the data pointsare :{ 2, 4, 10, 12, 3, 20, 30, 11, 25}. Let 2 and 4 are initial cluster centroids.Apply two rounds of k-means algorithm and find a set of clusters.Use Euclidean distance as the measure. (10)

b) Explain the concepts of any one density-based clustering technique. (6)

c) Differentiate web content mining and web usage mining. (4)

8 a) A database has five transactions. Let min_sup=60% and min_conf=80%. (15)

F1	Category
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}

i) Find all frequent item sets using Apriori algorithm.

ii) List all the strong association rules.

b) Explain the relevance and salient features of WeKa in datamining. (5)

9 a) Explain the importance of web structure mining. Also explain any two techniques used in web structure mining. (10)

b) Explain Weighted Graph Partitioning with an example. (5)

c) Explain the different data types and data handling functions used in R package. (5)

****