Reg No.:_______________ Name:__________________________

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

Sixth semester B.Tech examinations (S), September 2020

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) What are the major challenges of mining a huge amount of data in comparison with mining a small amount of data? (5)

b) How is data warehouse different from a database? How are they similar? (5)

c) What are the different types of applications where data mining can be directly applied? (5)

2 a) Distinguish between OLTP & OLAP. (8)

b) Use the two methods below to normalize the following group of data: 200, 300, 400, 600, 1000 (7)

i) min-max normalization by setting min=0 and max=1

ii) z-score normalization

3 a) What is multidimensional schema? (2)

b) Write short notes on Star, Snowflake and Data constellation schema. (9)

c) Suppose that a data warehouse consists of the four dimensions date, spectator, location and game, and the two measures count and charge, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults or seniors, with each category having its own charge rate. Draw a star schema diagram for the data warehouse. (4)

PART B

Answer any two full questions, each carries 15 marks.

4 a) Use Naive Bayes algorithm to determine whether a red domestic SUV car is stolen or not using the following data: (9)

Example No.	Colour	Type	Origin	Whether stolen
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

b) Explain in detail about Support vector machine. (6)

5 a) Use ID3 algorithm to construct a decision tree for the data in the following table. (9)

Age	**Competition**	**Type**	**Class(Profit)**
Old	Yes	Software	Down
Old	No	Software	Down
Old	No	Hardware	Down
Mid	Yes	Software	Down
Mid	Yes	Hardware	Down
Mid	No	Hardware	Up
Mid	No	Software	Up
New	Yes	Software	Up
New	No	Hardware	Up
New	No	Software	Up

b) Write the steps used for constructing a decision tree using ID3 algorithm. (6)

6 a) How to select the best splitting criterion in Decision tree? (3)

b) What is overfitting in neural network training? Which are the two approaches to avoid overfitting? (4)

c) Obtain a linear regression for the data in the table below assuming that y is the independent variable. (5)

x	1.0	2.0	3.0	4.0	5.0
y	1.00	2.00	1.30	3.75	2.25

d) Mention the issues regarding classification and prediction. (3)

PART C

Answer any two full questions, each carries 20 marks.

7 a) Write note on CRM data mining models. (7)

b) Draw the classification framework for data mining techniques in CRM and explain in detail. (6)

c) Explain the different stages of Customer life cycle with a neat diagram? (7)

8 a) Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8) (4)

i) Compute the Eucleidian distance between the two objects.

ii) Compute the Manhattan distance between the two objects.

b) Use K-means clustering algorithm to divide the following data into *two* clusters and also compute the representative data points for the clusters assuming the initial cluster centre as (2,1) and (2,3). (8)

X1	1	2	2	3	4	5
X2	1	1	3	2	3	5

c) Mention any *four* features of R programming. (4)

d) Differentiate web content mining and web structure mining. (4)

9 a) Write an algorithm for k-nearest neighbour classification given k and n, the number of attributes describing each tuple. (8)

b) How density based clustering varies from other methods? (8)

c) List the advantages and disadvantages of K-means clustering. (4)

****