Reg No.:_____ Name:_____

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

Eighth Semester B.Tech Degree Examination June 2022 (2015 Scheme)

Course Code: CS402

Course Name: DATA MINING AND WAREHOUSING

Max. Marks: 100 Duration: 3 Hours

PART A

Answer all questions, each carries 4 marks.

		Marks
1	What is a data warehouse system? What are its features?	4
2	How missing values are handled during data pre-processing?	4
3	Define the terms *Fact* and *Dimension* with respect to a data cube. Give examples for each.	4
4	What is the significance of pruning in decision tree induction? Give an example	4
5	Suppose you build a classification model and you are going to publish your work in journal. What will be the different metrics that you are using for evaluating the model? Write the definition of each measure.	4
6	What are the termination conditions of back propagation approach?	4
7	What is Link mining? Explain the challenges of link mining	4
8	What are the *four* different aspects to compare different clustering methods? Explain them.	4
9	How do you interpret the following association rule $computer \Rightarrow antivirus\_software\ [support = 2\%, confidence = 60\%]$. What is the use of threshold in support and confidence?	4
10	Describe the steps involved in text mining.	4

PART B

Answer any two full questions, each carries 9 marks.

11	a)	Explain with neat diagram, the knowledge discovery process.	5
	b)	Suppose that a data warehouse for Big-University consists of the following four dimensions: student, course, semester, and instructor, and two measures count and avg_grade. When at the lowest conceptual level (e.g., for a given student,	4

course, semester, and instructor combination), the avg_grade measure stores the actual course grade of the student. At higher conceptual levels, avg_grade stores the average grade for the given combination.

Draw a snowflake schema diagram for the data warehouse.

12 a) What is meant by data discretization? Describe any two methods for data discretization 5

b) Consider the following prices given in dollars. 4

1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30. Draw the histogram

i) with singleton buckets

ii) with equal width of $10

13 a) Write notes on applications and issues in Data Mining Process 5

b) Explain the various data reduction strategies used in preprocessing 4

PART C

Answer any two full questions, each carries 9 marks.

14 a) Given the following table of data 9

T_id	Refund	Marital Status	Taxable Income	Class
1	yes	single	Y	No
2	no	married	Y	No
3	no	single	N	No
4	yes	married	Y	No
5	no	Divorced	N	Yes
6	no	married	N	No
7	yes	Divorced	Y	No
8	no	single	N	Yes
9	No	Married	N	No
10	no	single	N	yes

Find out the probability for the attribute values, X: (Refund=No, Status= single, Taxable income= No) to belong to the class= yes and class = no

15 a) What is meant by attribute selection in decision tree induction? Explain, in detail, any two approaches for attribute selection 5

b) Explain the rule induction by Sequential Covering Algorithm. Discuss its advantage over if-then rule method. 4

16 a) Explain the classification by SVM when data are NOT linearly separable 5

b) How can we use confusion matrix as a tool for analyzing classifiers? 4

PART D

Answer any two full questions, each carries 12 marks.

17 a) Consider the transaction database given below. (12)

Set Support threshold=50%, Confidence= 60%.

Transaction	List if Items
T1	I1,I2,I3
T2	I2,I3,I4
T3	I4,I5
T4	I1,I2,I4
T5	I1,I2,I3,I5
T6	I1,I2,I3

i) Find the frequent item set using Apriori Algorithm.

ii) Generate strong association rules

18 a) How we can define a social network in a data mining perspective? 3

b) Explain the characteristics of social networks 3

c) Explain BIRCH algorithm in detail 6

19 a) With suitable examples describe K-Means clustering algorithm 6

b) Given the following distance matrix, construct dendrogram using single linkage, hierarchical clustering algorithm 6

Item	A	B	C	D	E
A	0	9	3	6	11
B	9	0	7	5	10
C	3	7	0	9	2
D	6	5	9	0	8
E	11	10	2	8	0

****