Reg No.:_____ Name:_____

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

B.Tech Degree S8 (S, FE) / S6 (PT) (S, FE) Examination May 2024 (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		How is data warehouse different from a database? How are they similar?	(4)
2		Give any two methods to handle noisy data.	(4)
3		State Entity Identification problem. Take a real world example and show how a data analyst solves it.	(4)
4		What is the significance of tree pruning in decision tree algorithms?	(4)
5		Why linear SVM is known as maximal margin classifier? Explain with suitable figure.	(4)
6		Compare and contrast Eager Classification and Lazy Classification.	(4)
7		Given two objects represented by the tuples (10,20,15,10,5) and (12,24,18,8,7). (a) Compute the Euclidean distance between the two objects. (b) Compute the Manhattan distance between the two objects.	(4)
8		What are the two measures used for rule interestingness?	(4)
9		How can we compute the dissimilarity between two binary objects?	(4)
10		Differentiate web content mining and web structure mining.	(4)

PART B

Answer any two full questions, each carries 9 marks.

11	a)	With the help of suitable diagrams explain the various OLAP operations.	(4)
	b)	Suppose that a data warehouse consists of the three dimensions Course, Teacher, and Student, and the two measures count and fee. Draw a star schema diagram for the data warehouse.	(5)
12	a)	Correlation does not imply causality. Justify the statement with a suitable example.	(4)

b) Consider the contingency table given below. Find out the correlation between the attributes through a Chi-square test. (5)

	game	$\overline{game}$	Σ_{row}
video	4,000	3,500	7,500
$\overline{video}$	2,000	500	2,500
Σ_{col}	6,000	4,000	10,000

13 a) How is data warehouse different from a database? How are they similar? (4)

b) Suppose a group of 12 sales price records has been sorted as follows: (5)

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215.

Partition them into three bins by each of the following methods.

(a) equal-frequency partitioning

(b) equal-width partitioning

PART C

Answer any two full questions, each carries 9 marks.

14 a) What is meant by attribute selection in decision tree induction? Explain, in detail, any two approaches for attribute selection. (4)

b) Given the following table of data. 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. (5)

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

15 a) Write short note on Linear and Non linear regression. (4)

b) A machine learning model is trained to predict tumors in patients. The test dataset consists of 100 people out of which 20 are tumor cases. The model (5)

predicted 15 cases as tumor cases, of which 10 are actually tumor cases. Draw the confusion matrix for the above problem and find the value of precision and recall.

16 a) Consider the following small data table for two classes of woods. Using information gain, construct a decision tree to classify the data set. Which attribute would information gain choose as the root of the tree? What class does the tree infer for the example {Density=Light, Grain=Small, Hardness=Soft}? (5)

Density	Grain	Hardness	Class
Heavy	Small	Hard	Oak
Heavy	Large	Hard	Oak
Heavy	Small	Hard	Oak
Light	Large	Soft	Oak
Light	Large	Hard	Pine
Heavy	Small	Soft	Pine
Heavy	Large	Soft	Pine
Heavy	Small	Soft	Pine

b) Distinguish between hold out method and cross validation method. (4)

PART D

Answer any two full questions, each carries 12 marks.

17 a) Differentiate between Support and Confidence. (5)

b) Using Apriori algorithm identify frequent itemsets for the following transaction table, given min_support = 3. (7)

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

18 a) Explain BIRCH clustering method. (8)

b) What are the advantages of BIRCH compared to other clustering method? (4)

19 a) Explain Apriori based frequent subgraph mining. (6)

b) Write short note on k-means clustering. (6)

****