KUMARAGURU
college of technology
character is life

Register Number:……..…

M.TECH DEGREE EXAMINATIONS: DEC 2022

(Regulation 2018)

First Semester

DATA SCIENCE

P18ITI1203**:** Data Mining Techniques

COURSE OUTCOMES

CO1: Explain the techniques for data pre processing

CO2: Apply association rules algorithm for correlation analysis

CO3: Apply decision tree algorithm for classification

CO4: Apply Bayesian networks algorithm for classification

CO5: Apply various clustering algorithms for different datasets

CO6: Estimate the classifier accuracy with training, testing and cross validation datasets

Time: Three Hours **Maximum Marks: 100**

Answer all the Questions:-

PART A (10 x 1 = 10 Marks)

1. The following items consists of two statements, one labeled as the "Assertion (A)" and the other as "Reason (R). You are to examine those two statements carefully and select the answers to these items using the codes given below Codes: CO1 [K1]

Assertion (A): A relational database is a collection of tables, each of which is assigned a unique name.
Reason (R): Each table consists of a set of attributes (*columns* or *fields*) and usually stores a large set of tuples (*records* or *rows*).

a) both A and R are individually true and R is the correct explanation of A
b) both A and R are individually true but R is not the correct explanation of A
c) A is true but R is false
d) A is false but R is true.

2. A _____ is a set of mathematical functions that describe the behavior of the objects in a target class in terms of random variables and their associated probability distributions. CO1 [K1]

a) Hierarchical model
b) Descriptive model
b) Statistical model
d) Linear model

3. _____ methods smooth a sorted data value by consulting its "neighborhood," that is, the values around it. CO2 [K1]

a) Regression
b) Outlier analysis
c) Extraction
d) Binning

4. Matching type item with multiple choice code CO4 [K1]

List I	List II
A. Nominal attribute	i. Boolean
B. Binary Attributes	ii. Integer or real values
C. Ordinal Attributes	iii. Symbols or names of things
D. Numeric Attributes	iv. Meaningful order or ranking

	A	B	C	D
a)	ii	iii	i	iv
b)	iii	i	iv	ii
c)	ii	iii	i	iv
d)	iv	i	ii	iii

5. The following items consists of two statements, one labeled as the "Assertion (A)" and the other as "Reason (R). You are to examine those two statements carefully and select the answers to these items using the codes given below Codes: CO4 [K1]

Assertion (A): Normalization, where the attribute data are scaled so as to fall within a smaller range, such as 1.0 to 1.0, or 0.0 to 1.0.

Reason (R): Discretization, where the raw values of a numeric attribute (e.g., *age*) are replaced by interval labels (e.g., 0–10, 11–20, etc.) or conceptual labels (e.g., *youth, adult*, *senior*).

a) both A and R are individually true and R is the correct explanation of A

b) both A and R are individually true but R is not the correct explanation of A

c) A is true but R is false

d) A is false but R is true.

6. _____ techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. CO3 [K1]

a) Data Reduction

b) Data extraction

c) Data Loading

d) Data Transformation

7. Which of the following are extended data types? CO2 [K1]

i) sequential and time-series patterns ii) structural (e.g., tree, lattice, graph) patterns

iii) spatial (e.g., colocation) patterns iv) approximate patterns

a) i, ii, iii

b) i, iii, iv

c) i, ii, iv

d) ii, iii, iii

8.	ID3 uses _____ as its attribute selection measure.		CO3	[K1]
	a) Gain Ratio	b) Gini Index		
	c) Information Gain	d) Repetition		
9.	Rule _____ and _____ are two measures of rule interestingness.		CO6	[K1]
	a) Bagging, Boasting	b) Univariate, Multivariate		
	c) Oversampling, Undersampling	d) Support, Confidence		
10.	_____ is an interactive approach based on multidimensional visualization techniques and allows the user to incorporate background knowledge about the data when building a decision tree.		CO5	[K1]
	a) Bayesian classifiers	b) IF-THEN Rules for Classification		
	c) Perception-based classification (PBC)	d) Bagging		

PART B (10 x 2 = 20 Marks)

11.	Explain in detail about multidimensional data mining.	[CO1]	[K2]
12.	Differentiate between data characterization and discrimination.	[CO1]	[K2]
13.	What are the major tasks in data preprocessing?	[CO2]	[K2]
14.	How Principal Components Analysis (PCA) used for data dimensionality reduction?	[CO2]	[K2]
15.	Given a 5-GB data set with 50 attributes (each containing 100 distinct values) and 512 MB of main memory in your laptop, outline an efficient method that constructs decision trees in such large data sets. Justify your answer by rough calculation of your main memory usage.	[CO3]	[K3]
16.	Explain about confusion matrix with an example.	[CO3]	[K2]
17.	List the techniques to improve classification accuracy.	[CO4]	[K2]
18.	List the various requirements for Cluster Analysis	[CO4]	[K2]
19.	Discuss about the types of outliers.	[CO5]	[K2]
20.	What are challenges of Outlier Detection in High-Dimensional Data is performed?	[CO5]	[K2]

PART C (10 x 5 = 50 Marks)

21.	Explain how data mining as a step in the process of knowledge discovery in large data bases.	CO1	[K2]
22.	Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. (a) What is the mean of the data? What is the median? (b) What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.). (c) What is the midrange of the data?	CO1	[K3]

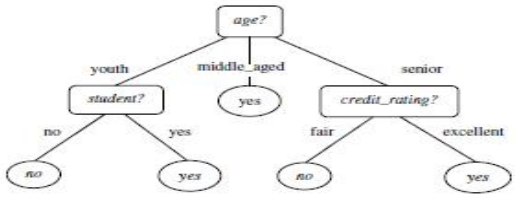
(d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?
(e) Give the five-number summary of the data.
(f) Show a boxplot of the data.
(g) How is a quantile–quantile plot different from a quantile plot?

23. Differentiate between snowflake schema and fact constellation query model with neat diagram. CO2 [K_2]

24. Explain about FP-growth algorithm with an example. CO2 [K_2]

25. Explain in detail about metrics for evaluating classifier performance CO3 [K_2]

26. How to extract rule from a decision tree? Explain for the below decision tree. CO3 [K_3]



27. Describe about k-Means: A Centroid-Based Technique for clustering. CO4 [K_2]

28. Differentiate agglomerative versus divisive hierarchical clustering. CO4 [K_2]

29. How supervised methods are used for outlier detection methods? CO5 [K_2]

30. Explain about proximity-based methods with an example. CO5 [K_2]

Answer any TWO Questions
PART D (2 x 10 = 20 Marks)

31. Briefly describe about attribute selection measures for the following class labeled training tuples from the *AllElectronics* Customer Database. CO2 [K_3]

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

32. Enumerate how support vector machines (SVMs) is used for the classification of both linear and nonlinear data. CO3 [K_3]

33. Explain in detail about Partitioning methods and Hierarchical methods for clustering. CO5 [K_3]

****