05/07/2022
FN

Code No: R2022422 **R20** **SET - 1**

II B. Tech II Semester Regular Examinations, June/July - 2022
DATA WAREHOUSING AND MINING
(Common to CSE(AIML),CSE(AI),CSE(DS),CSE(AIDS),AIDS & AIML)

Time: 3 hours **Max. Marks: 70**

Answer any **FIVE** Questions each Question from each unit
All Questions carry **Equal** Marks

~~~~~~~~~~~~~~~~~~~~~~~~~

UNIT-I

1 a) Explain in brief about Data mining task primitives. [7M]
b) Describe any five advanced data base systems and applications. [7M]

Or

2 a) Briefly discuss about the Metadata Repository data warehouse implementation methods. [7M]
b) Explain the implementation process of Data Warehousing. [7M]

UNIT-II

3 a) Explain the types of data and attributes with respect to quality and quantity. [7M]
b) Explain the various data smoothing techniques that are used to handle noise data [7M]

Or

4 a) What is the need of dimensionality reduction? Explain any two techniques for dimensionality reduction. [7M]
b) "Data mining is a confluence of multiple disciplines". Explain this with an example. [7M]

UNIT-III

5 a) Given a decision tree, there exists an option of (i) converting the decision tree to rules and then pruning the resulting rules, or (ii) pruning the decision tree and then converting the pruned tree to rules? What advantage does (i) have over (ii)? [7M]
b) Describe the criteria used to evaluate classification and prediction methods. [7M]

Or

6 a) Consider the characteristics of a 'Smart Mobile' and implement decision tree classifier to assign the price ranges. [7M]
b) Explain the ID3 algorithm for the induction of decision trees. [7M]

UNIT-IV

7 a) Discuss the Apriori algorithm with a suitable example and explain how its efficiency can be improved. [7M]
b) How to construct Frequent pattern tree to find frequent item sets? Explain the role of support in it? [7M]

Or
~~~~~~~~~~~~~~~~~~~~~~~~~

8 a) Write a algorithm to discover frequent item sets without candidate generation and explain it with a example.

b) Generate association rules from two step algorithm.

UNIT-V

9 a) Define partitioning methods. Describe anyone partition based clustering algorithm.

b) What is cluster analysis? Describe the dissimilarity measures for interval-scaled variables and binary variables.

Or

10 a) With a suitable example, explain the k-means clustering algorithm.

b) Write the time and space complexity for DBSCAN algorithm.

Code No: R2022422 **R20** **SET - 2**

II B. Tech II Semester Regular Examinations, June/July - 2022

DATA WAREHOUSING AND MINING

(Common to CSE(AIML),CSE(AI),CSE(DS),CSE(AIDS),AIDS & AIML)

Time: 3 hours **Max. Marks: 70**

Answer any **FIVE** Questions each Question from each unit
All Questions carry **Equal** Marks

UNIT-I

1	a)	With neat sketch explain the architecture and functions of data warehouse.	[7M]
	b)	Define measure. What are the different categories of measures?	[7M]
		Or	
2	a)	What are the different OLAP operations on multidimensional data?	[7M]
	b)	Explain how the evolution of database technology led to data mining.	[7M]
		UNIT-II	
3	a)	Write a short note on: i)Data Preprocessing ii)Data Discretization iii)concept hierarchy.	[7M]
	b)	Define dimensionality reduction. What are the different methods used in it?	[7M]
		Or	
4	a)	Draw and explain the architecture of a typical data mining system.	[7M]
	b)	Define Attribute subset selection and data preprocessing? Explain why data preprocessing is necessary.	[7M]
		UNIT-III	
5	a)	Specify the reasons for model overfitting and explain the methods to solve this problem.	[7M]
	b)	What is misclassification rate of a classifier? Describe sensitivity and specificity measures of Naïve Bayes Classifier.	[7M]
		Or	
6	a)	Explain the different measures for selecting best split in decision tree induction.	[7M]
	b)	Discuss the methods that are commonly used to evaluate the performance of a classifier.	[7M]
		UNIT-IV	
7	a)	What is market based analysis? Give an example and explain the techniques to perform this.	[7M]
	b)	Explain the rule generation and frequent item set generation steps with suitable example?	[7M]
		Or	
8	a)	Develop the Apriori algorithm for generating frequent-item set.	[7M]
	b)	Explain how to find frequent item sets using FP-Growth algorithm.	[7M]

Code No: R2022422 **R20** **SET - 2**

II B. Tech II Semester Regular Examinations, June/July - 2022

DATA WAREHOUSING AND MINING

(Common to CSE(AIML),CSE(AI),CSE(DS),CSE(AIDS),AIDS & AIML)

Time: 3 hours **Max. Marks: 70**

Answer any **FIVE** Questions each Question from each unit

All Questions carry **Equal** Marks

~~~~~~~~~~~~~~~~~~~~~~~~~~~~

UNIT-I

1 a) With neat sketch explain the architecture and functions of data warehouse. [7M]

b) Define measure. What are the different categories of measures? [7M]

Or

2 a) What are the different OLAP operations on multidimensional data? [7M]

b) Explain how the evolution of database technology led to data mining. [7M]

UNIT-II

3 a) Write a short note on: i)Data Preprocessing ii)Data Discretization iii)concept hierarchy. [7M]

b) Define dimensionality reduction. What are the different methods used in it? [7M]

Or

4 a) Draw and explain the architecture of a typical data mining system. [7M]

b) Define Attribute subset selection and data preprocessing? Explain why data preprocessing is necessary. [7M]

UNIT-III

5 a) Specify the reasons for model overfitting and explain the methods to solve this problem. [7M]

b) What is misclassification rate of a classifier? Describe sensitivity and specificity measures of Naïve Bayes Classifier. [7M]

Or

6 a) Explain the different measures for selecting best split in decision tree induction. [7M]

b) Discuss the methods that are commonly used to evaluate the performance of a classifier. [7M]

UNIT-IV

7 a) What is market based analysis? Give an example and explain the techniques to perform this. [7M]

b) Explain the rule generation and frequent item set generation steps with suitable example? [7M]

Or

8 a) Develop the Apriori algorithm for generating frequent-item set. [7M]

b) Explain how to find frequent item sets using FP-Growth algorithm. [7M]
~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Code No: R2022422

R20

SE

UNIT-V

9 a) Differentiate the k-means clustering and kernel k-means clustering.

b) Describe any one Hierarchical clustering algorithm.

Or

10 a) Explain different types of clustering techniques with merits and demerits.

b) Consider three points [X1 , X2 , X3] with the following coordinates as a two dimensional sample for clustering : X1 = (0.5,2.5) X2 = (0,0) X3 = (1.5,1) Illustrate the K-means partitioning algorithm using the above data set.

Code No: R2022422 R20 SET - 3

II B. Tech II Semester Regular Examinations, June/July - 2022

DATA WAREHOUSING AND MINING

(Common to CSE(AIML),CSE(AI),CSE(DS),CSE(AIDS),AIDS & AIML)

Time: 3 hours Max. Marks: 70

Answer any **FIVE** Questions each Question from each unit

All Questions carry **Equal** Marks

UNIT-I

1	a)	Give a contrast for OLAP and OLTP.	[7M]
	b)	Explain the implementation process of Data Warehousing.	[7M]

Or

2	a)	Explain the three-tier data warehouse architecture.	[7M]
	b)	What is concept hierarchy? Describe the OLAP operations in the multidimensional data model.	[7M]

UNIT-II

3	a)	Define Data Cleaning. Discuss in brief about approaches to fill missing values.	[7M]
	b)	Explain about Data discretization and data summarization.	[7M]

Or

4	a)	Define data integration. Discuss the issues to be considered for data integration.	[5M]
	b)	Discuss in detail about the available techniques for concept hierarchy generation for categorical data.	[9M]

UNIT-III

5	a)	Given two classifiers with variation in performance. Then discuss various methods used to evaluate their performance.	[9M]
	b)	Write the Decision tree induction algorithm and discuss its issues.	[5M]

Or

6	a)	Explain the different attribute types that are used in attribute test condition in the decision tree.	[7M]
	b)	Define the terms Model overfitting and Model underfitting. Compare them.	[7M]

UNIT-IV

7	a)	Differentiate the mining frequent item sets with and without candidate generations.	[7M]
	b)	Explain how to generate rules from frequent item sets.	[7M]

Or

8 a) The price of each item in a store is non-negative. For each of the following cases, identify the kind of constraint they represent and briefly discuss how to mine such association values efficiently: [7M]

i) containing atleast one Ninetendo game

ii) Containing items the sum of whose price is less than $150.

b) Explain how the frequent item set is generated using FP-Growth algorithm. [7M]

UNIT-V

9 a) Consider five points [X1 , X2 , X3 , X4 , X5] with the following coordinates as a two dimensional sample for clustering : X1 = (0.5,2.5) X2 = (0,0) X3 = (1.5,1) X4 = (5,1) X5 = (6,2) .. Illustrate the K-means partitioning algorithm using the above data set. [12M]

b) Define core, border and noise points with respect to clustering. [2M]

Or

10 a) Write about Minimum, Maximum and Average links used in clusterings. [7M]

b) Explain DBSCAN algorithm with its strength and weaknesses. [7M]

LIBRARY 28-06-22

Code No: R2022051 **R20** **SET - 1**

II B. Tech II Semester Regular Examinations, June/July - 2022

PROBABILITY AND STATISTICS

(Common to CSE, CST, CSE(AIML), CSE(AI), CSE(DS), CSE(AIDS), CSE(CS), CSE(IOTCSIBCT), CSE(CSBS), CSE(IOT), AIDS, & AIML)

Time: 3 hours **Max. Marks: 70**

Answer any **FIVE** Questions each Question from each unit
All Questions carry **Equal** Marks

1 a) What are the main methods of collecting primary data? State briefly the advantages and disadvantages of each. [7M]

b) What do you understand by skewness. What are the various methods of measuring skewness? [7M]

Or

2 a) What do you understand by a measure of dispersion ? What purpose does a measure of dispersion serve ? [6M]

b) Calculate the mean and standard deviation for the following table giving the age distribution of 542 members. [8M]

Age (in years)	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70	70 - 80	80 - 90
No. of members	3	61	132	153	140	51	2

3 a) Calculate correlation coefficient to the following data ; [7M]

x	10	15	12	17	13	16	24	14	22	20
y	30	42	45	46	33	34	40	35	39	38

b) Fit an exponential curve of the form $y = ab^x$ to the following data: [7M]

x	1	2	3	4	5	6	7	8
y	1.0	1.2	1.8	2.5	3.6	4.7	6.6	9.1

Or

4 a) Obtain the equations of two lines of regression for the following data. Also obtain the estimate of X for Y = 70. [10M]

X	65	66	67	67	68	69	70	72
Y	67	68	65	68	72	72	69	71

b) State the principle of least squares and write the normal equations for the fitting of the Straight line. [4M]

5 a) The probability that a turbine will have a defective coil is 0.10, the probability that it will have defective blades is 0.15, and the probability that it will have both defects is 0.04. [7M]
(i) What is the probability that a turbine will have one of these defects?
(ii) What is the probability that a turbine will have either of these defects?

b) A shipment of 20 similar laptop computers to a retail outlet contains 3 that are defective. If a school makes a random purchase of 2 of these computers, find the probability distribution for the number of defectives. [7M]

Or

SET - 1

R20

Code No:R2022051

6 a) The diameter of an electric cable, say X, is assumed to be a continuous random variable with p.d.f. $f(x) = 6x(1-x), 0 \le x \le 1$. [8M]

(i) Check that $f(x)$ is p.d.f., and

(ii) Determine a number b such that $P(X < b) = P(X > b)$.

b) Given a standard normal distribution, find the area under the curve that lies [6M]

(i) to the right of $z = 1.84$ and

(ii) between $z = -1.97$ and $z = 0.86$.

7 a) Define Population and sample with examples. [6M]

b) For a chi-squared distribution, find [8M]

(i) $\chi^2_{0.025}$ when $v = 15$;

(ii) $\chi^2_{0.01}$ when $v = 7$.

Or

8 The pulse rate of 50 yoga practitioners decreased on the average by 20.2 beats/minute with s.d. of 3.5. (a) If $\bar{x} = 20.2$ is used as a point estimate of the true average decrease in the pulse rate, what can we assert with 95% confidence about the maximum error E. (b)Construct 99% confidence intervals for the true average decrease in pulse rate. [14M]

9 a) A random sample of 100 recorded deaths in the United States during the past year showed an average life span of 71.8 years. Assuming a population standard deviation of 8.9 years, does this seem to indicate that the mean life span today is greater than 70 years? Use a 0.05 level of significance. [7M]

b) An urban community would like to show that the incidence of breast cancer is higher in their area than in a nearby rural area. If it is found that 20 of 200 adult women in the urban community have breast cancer and 10 of 150 adult women in the rural community have breast cancer, can we conclude at the 0.05 level of significance that breast cancer is more prevalent in the urban community? [7M]

Or

10 a) Past experience indicates that the time required for high school seniors to complete a standardized test is a normal random variable with a mean of 35 minutes. If a random sample of 20 high school seniors took an average of 33.1 minutes to complete this test with a standard deviation of 4.3 minutes, test the hypothesis, at the 0.05 level of significance, that $\mu = 35$ minutes against the alternative that $\mu < 35$ minutes. [7M]

b) A manufacturer claims that the average tensile strength of thread A exceeds the average tensile strength of thread B by at least 12 kilograms. To test this claim, 50 pieces of each type of thread were tested under similar conditions. Type A thread had an average tensile strength of 86.7 kilograms with a standard deviation of 6.28 kilograms, while type B thread had an average tensile strength of 77.8 kilograms with a standard deviation of 5.61 kilograms. Test the manufacturer's claim using a 0.05 level of significance. [7M]

Code No:R2022051 **R20** **SET - 2**

II B. Tech II Semester Regular Examinations, June/July - 2022

PROBABILITY AND STATISTICS

(Common to CSE, CST, CSE(AIML), CSE(AI), CSE(DS), CSE(AIDS), CSE(CS), CSE(IOTCSIBCT), CSE(CSBS), CSE(IOT), AIDS, CS, & AIML)

Time: 3 hours **Max. Marks: 70**

Answer any **FIVE** Questions each Question from each unit
All Questions carry **Equal** Marks

~~~~~~~~~~~~~~~~~~~~~~~~~

1 a) Distinguish between primary and secondary data and discuss the various methods of collecting primary data. [7M]

b) What do you understand by skewness and kurtosis? Point out their role in analyzing a frequency distribution. [7M]

Or

2 a) What do you understand by dispersion? Explain briefly the various methods used for measuring dispersion. [7M]

b) Calculate the coefficient of skewness based on mean and median from the following distribution [7M]

Class interval	0 - 10	10 - 20	20 – 30	30 - 40	40 - 50	50 - 60	60 - 70	70 - 80
Frequency	6	12	22	48	56	32	18	6

3 a) Obtain the correlation coefficient for the following data : [7M]

x	48	60	72	62	56	40	39	52	30
y	62	78	65	70	38	54	60	32	31

b) Fit a parabola $y = a + bx + cx^2$ to the following data: [7M]

x	2	4	6	8	10
y	3.07	12.85	31.47	57.38	91.29

Or

4 a) In a partially destroyed laboratory, record of an analysis of correlation data, the following only are legible: Variance of $X = 9$, Regression equations: $8X - 10Y + 66 = 0$, $40X - 18Y = 214$. What are: (i) the mean values X and Y, (ii) the correlation coefficient between X and Y, and (iii) the standard deviation of Y? [8M]

b) The ranks of same 16 students in Mathematics and Physics are as follows. Two numbers within brackets denote the ranks of the students in Mathematics and Physics: (1,1) (2,10) (3,3) (4,4) (5,5) (6,7) (7,2) (8,6) (9,8) (10,11) (11,15) (12,9) (13,14) (14,12) (15,16) (16,13) .Calculate the rank correlation coefficient for proficiencies of this group in Mathematics and Physics. [6M]

5 a) The probability that a construction company will get the tender for constructing a flyover is 0.33, the probability that it will get the tender for constructing an underpass is 0.28, and the probability that it will get both tenders is 0.13.
(i) What is the probability that it will get at least one tender?
(ii) What is the probability that it will get neither tender? [7M]

b) Find the mean and the variance of the uniform probability distribution given by [7M]

$$f(x) = \frac{1}{n} \quad \text{for } x = 1, 2, 3, \ldots, n.$$

Or
~~~~~~~~~~~~~~~~~~~~~~~~~