Total No. of Questions: 6 *Total No. of Printed Pages:3*

Enrollment No.....

Faculty of Engineering
End Sem Examination May-2024
IT3ED03 Data Analytics

Programme: B.Tech. Branch/Specialisation: IT

Duration: 3 Hrs. **Maximum Marks: 60**

Note: All questions are compulsory. Internal choices, if any, are indicated. Answers of Q.1 (MCQs) should be written in full instead of only a, b, c or d. Assume suitable data if necessary. Notations and symbols have their usual meaning.

Q.1 i. What are the primary tasks of data mining? **1**
(a) Classification, regression, clustering
(b) Summarization, visualization, prediction
(c) Reporting, querying, modeling
(d) ETL (Extract, Transform, Loa(d), data warehousing, OLAP (Online Analytical Processing)

ii. What are some challenges commonly encountered in data mining projects? **1**
(a) Lack of computational resources (b) Data quality issues
(c) Overfitting (d) All of these

iii. Variable binning is used for: **1**
(a) Handling missing values
(b) Removing outliers
(c) Converting numerical variables into categorical variables
(d) Scaling numerical variables

iv. Feature creation refers to: **1**
(a) Removing unnecessary features from the dataset
(b) Creating new features from existing ones
(c) Standardizing features for analysis
(d) Normalizing features to a specific range

v. In PCA, how many principal components are typically retained? **1**
(a) One
(b) Two
(c) A subset based on explained variance threshold
(d) All of these

P.T.O.

vi. Which of the following is a common factor rotation method? **1**
(a) Standardization (b) Normalization
(c) Varimax (d) Interpolation

vii. The critical value in hypothesis testing is determined by: **1**
(a) The level of significance and the sample size
(b) The sample mean and the population mean
(c) The standard deviation of the sample
(d) The confidence interval and the margin of error

viii. When the p-value is less than the significance level, what action should be taken? **1**
(a) Fail to reject the null hypothesis
(b) Reject the null hypothesis
(c) Accept the null hypothesis
(d) Retest the hypothesis with a larger sample size

ix. In a chi-square test, what is the null hypothesis? **1**
(a) There is no difference between observed and expected frequencies
(b) There is a difference between observed and expected frequencies
(c) There is no association between two categorical variables
(d) There is an association between two categorical variables

x. When is a one-tailed test appropriate? **1**
(a) When the researcher is unsure about the direction of the effect
(b) When the researcher is specifically interested in one direction of the effect
(c) When the sample size is small
(d) When the data is normally distributed

Q.2 i. Define business intelligence in the context of predictive analytics. **2**
ii. How are statistics and data mining related? Explain in detail. **3**
iii. Discuss the significance of predictive analytics in modern business contexts, providing examples. **5**
OR iv. Provide examples of short-term prediction applications across different industries, emphasizing the importance of accuracy and timeliness. **5**

Q.3 i. What is the significance of rank-ordered statistics in data preprocessing? **2**
ii. Explain how skewness and kurtosis measures contribute to the understanding of data distribution shapes, and how they affect data preprocessing decisions. **8**

OR iii. Explain the purpose and techniques of data transformation, including normalization, standardization, variable scaling, and variable binning, with examples. **8**

Q.4 i. Define communalities in factor analysis. **2**
ii. What are the criteria for selecting principal components in PCA? **3**
iii. Evaluate the practical implications of dimension reduction methods such as PCA and factor analysis in real-world data analysis scenarios, considering their benefits and limitations. **5**
OR iv. Discuss the criteria used for selecting principal components in PCA and their significance in dimensionality reduction. **5**

Q.5 i. Define the margin of error in the context of estimation. **4**
ii. Define Type I and Type II errors in hypothesis testing and discuss their practical implications in decision-making. **6**
OR iii. Describe the process of constructing a confidence interval for estimating the mean, including the formula and interpretation of results. **6**

Q.6 i. When would you use a paired t-test instead of a two-sample t-test? **2**
ii. Discuss the assumptions underlying the t-test and Z-test., How violations of these assumptions can impact the validity of the test results? **8**
OR iii. Consider a dataset containing the scores of two groups of students (Group A and Group B) in a math exam. Perform a two-sample t-test to determine if there is a significant difference in the mean scores between the two groups. Use the following information: **8**
(a) Group A: Mean score = 85, Standard deviation = 10, Sample size = 30.
(b) Group B: Mean score = 78, Standard deviation = 8, Sample size = 25.
(c) Assume equal variances for both groups. Use a significance level of 0.05. Calculate the t-statistic and determine whether the difference in means is statistically significant.

******

Marking Scheme

DATA ANALYTICS (DA) IT3ED03

Q.1	i)	What are the primary tasks of data mining? **a) Classification, regression, clustering**	**1**
	ii)	What are some challenges commonly encountered in data mining projects? **d) All of the above**	**1**
	iii)	Variable binning is used for: **c) Converting numerical variables into categorical variables**	**1**
	iv)	Feature creation refers to: **b) Creating new features from existing ones**	**1**
	v)	In PCA, how many principal components are typically retained? **d) A subset based on explained variance threshold**	**1**
	vi)	Which of the following is a common factor rotation method? **c) Varimax**	**1**
	vii)	The critical value in hypothesis testing is determined by: **a) The level of significance and the sample size**	**1**
	viii)	When the p-value is less than the significance level, what action should be taken? **b) Reject the null hypothesis**	**1**
	ix)	In a chi-square test, what is the null hypothesis? **a) There is no difference between observed and expected frequencies**	**1**
	x)	When is a one-tailed test appropriate? **b) When the researcher is specifically interested in one direction of the effect**	**1**
Q.2	i.	Define business intelligence in the context of predictive analytics.	**2**
	ii.	How are statistics and data mining related? Explain in detail. Statistics 1 Mark Data mining 1 Mark Working Together 1 Mark	**3**
	iii.	Discuss the significance of predictive analytics in modern business contexts, providing examples. Significance of predictive analytics 4 Mark Example 1 Mark	**5**
OR	iv.	Provide examples of short-term prediction applications across different industries, emphasizing the importance of accuracy and timeliness. Short-term prediction applications Give 5 Example (accuracy and Timeliness) 5 Mark each	**5**
Q.3	i.	What is the significance of rank-ordered statistics in data preprocessing?	**2**
	ii.	Explain how skewness and kurtosis measures contribute to the understanding of data distribution shapes, and how they affect data preprocessing decisions. Skewness & Impact on Data Preprocessing 4 Mark Kurtosis & Impact on Data Preprocessing 4 Mark	**8**
OR	iii.	Explain the purpose and techniques of data transformation, including normalization, standardization, variable scaling, and variable binning, with examples. Purpose of Data Transformation 2 Mark Techniques of Data Transformation 6 Mark	**8**
Q.4	i.	Define communalities in factor analysis.	**2**
	ii.	What are the criteria for selecting principal components in PCA? Criteria for selecting principal components in PCA 3 Mark	**3**
	iii.	Evaluate the practical implications of dimension reduction methods such as PCA and factor analysis in real-world data analysis scenarios, considering their benefits and limitations. Benefit 2.5 Mark Limitation 2.5 Mark	**5**
OR	iv.	Discuss the criteria used for selecting principal components in PCA and their significance in dimensionality reduction. Criteria Used for Selecting Principal Component 2.5 Mark Significance in Dimensionality Reduction 2.5 Mark	**5**
Q.5	i.	Define the margin of error in the context of estimation. Define 1 Mark margin of error in estimation with example 3 Mark	**4**
	ii.	Define Type I and Type II errors in hypothesis testing and discuss their practical implications in decision-making. Type I Error 1 Mark Type II Error 1 Mark Practical implications of Type I and Type II errors in decision-making 4 Mark	**6**
OR	iii.	Describe the process of constructing a confidence interval for estimating the mean, including the formula and interpretation of results. Breakdown of the process along with the formula and interpretation of results 6 Mark	**6**
Q.6	i.	When would you use a paired t-test instead of a two-sample t-test?	**2**
	ii.	Discuss the assumptions underlying the t-test and Z-test, and how violations of these assumptions can impact the validity of the test	**8**

results.

Assumptions of the t-test & Impacts of Violations 4 Mark

Assumptions of the Z-test & Impacts of Violations 4 Mark

iii. Consider a dataset containing the scores of two groups of students (Group A and Group B) in a math exam. Perform a two-sample t-test to determine if there is a significant difference in the mean scores between the two groups. Use the following information: **8**

- Group A: Mean score = 85, Standard deviation = 10, Sample size = 30
- Group B: Mean score = 78, Standard deviation = 8, Sample size = 25
- Assume equal variances for both groups. Use a significance level of 0.05. Calculate the t-statistic and determine whether the difference in means is statistically significant.

Given Information **1 Mark**

Write Formula **2 Mark**

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

calculate the t-statistic 5 Mark

t≈2.884

Now, we compare this t-statistic with the critical t-value at a significance level of 0.05 and degrees of freedom (df)) equal to n1 +n2−2. The critical t-value for a two-tailed test with df=53 and a significance level of 0.05 is approximately 2.004.

Since 2.884>2.004, the calculated t-statistic falls in the critical region, which means we reject the null hypothesis. Therefore, we conclude that there is a statistically significant difference in the mean scores between Group A and Group B.

******