

GUJARAT TECHNOLOGICAL UNIVERSITY

BE - SEMESTER-V (NEW) EXAMINATION – SUMMER 2024

Subject Code:3151608

Date:16-05-2024

Subject Name:Data Science

Time:02:30 PM TO 05:00 PM

Total Marks:70

Instructions:

1. Attempt all questions.
2. Make suitable assumptions wherever necessary.
3. Figures to the right indicate full marks.
4. Simple and non-programmable scientific calculators are allowed.

		Marks
Q.1	(a) Explain the measure of variation.	03
	(b) Explain four 'V' of big data analytics.	04
	(c) List out the major categories of Sampling Method. Explain in detail.	07
Q.2	(a) What is the use of percentile? Calculate quartile (Q1, Q2, Q3) for given data: Data: 1, 3, 3, 4, 5, 6, 6, 7, 8, 8	03
	(b) Explain the categories of Machine Learning Algorithms.	04
	(c) Define random variable. Explain the types of random variable by taking suitable example.	07
	OR	
	(c) Discuss Monty Hall problem using Bayes' Theorem.	07
Q.3	(a) State True or False: "A high R Square value is not necessarily a good indicator of the correctness of the model; it could be a spurious relationship.". Justify.	03
	(b) Differentiate between descriptive analytics and predictive analytics.	04
	(c) Use maximum likelihood estimation (MLE) to estimate the parameter of Binomial Distribution.	07
	OR	
Q.3	(a) What is Outlier? List out the distance measure which are useful in identifying the outlier.	03
	(b) Explain Central Limit Theorem(CLT) in detail.	04
	(c) What is the significance of Probability Distribution Function(PDF) and Cumulative Distribution Function? Discuss PDF and CDF for Normal Distribution and Exponential Distribution.	07
Q.4	(a) Define Entropy, Information Gain and Gini Index.	03
	(b) Differentiate between Logistic Regression and Linear Regression.	04
	(c) Discuss measure of shapes in detail.	07
	OR	
Q.4	(a) Explain Ordinal Scale, Interval Scale and Ratio Scale with example.	03
	(b) Discuss various approaches of variable selection used in Logistic Regression.	04

- (c) Discuss Chebyshev's Theorem. Solve: Amount spent per month by a segment of credit card users of a bank has a mean value of 12000 and standard deviation of 2000. Calculate the proportion of customers who are spending between 8000 and 16000. **07**
- Q.5** (a) Discuss the significance of pruning in Decision Tree. **03**
 (b) Write down steps for generating decision trees. **04**
 (c) Explain Random Forest Algorithm in detail. **07**
- OR**
- Q.5** (a) Explain the significance of using Ensemble Methods for classification. **03**
 (b) Differentiate between Gain Chart and Lift Chart. **04**
 (c) Explain Chi-Square Automatic Interaction Detection (CHAID) in detail. **07**
