Seat No.: _____ Enrolment No._____

GUJARAT TECHNOLOGICAL UNIVERSITY

BE – SEMESTER- V EXAMINATION-SUMMER 2023

Subject Code: 3151608 Date: 23/06/2023

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)**	What is Data Science? Give some application of Data Science.	**03**
	(b)	Differentiate structured and unstructured data.	**04**
	(c)	Discuss the various types of data measurement scales with example.	**07**
Q.2	**(a)**	Define probability mass function, degree of freedom and joint probability.	**03**
	(b)	Consider the budget (in lakh) of Bollywood movies as below: 150, 60, 90, 180, 50, 69, 76, 88, 94, 63, 113, 126, 200, 91, 75 Draw the Box plot (Whisker plot) on above data.	**04**
	(c)	What do you mean by symmetric, positively skewed and negatively skewed data distributions in data science? Give example of data for each distribution.	**07**
		OR	
	(c)	Consider the marks of 30 students in data science subject as below:	**07**

40	30	45	62	85	95	45	78	76	82	68	84	88	84	79
49	67	85	83	92	73	88	46	88	74	88	96	88	36	88

I. Calculate mean, median and mode.

II. Calculate 95^{th} and 50^{th} percentile of the marks.

III. Calculate the inter quartile range (IQR).

Q.3	**(a)**	Illustrate Geometric Distribution with memoryless property.	**03**
	(b)	Differentiate Probabilistic Sampling and Non-Probability Sampling.	**04**
	(c)	Fashion Trends Online (FTO) is an e-commerce company that sells women apparel. It is observed that about 10% of their customers return the items purchased by them for many reasons (such as size, color, and material mismatch). On a particular day, 20 customers purchased items from FTO. Calculate: **I.** Probability that exactly 5 customers will return the items. **II.** Probability that a maximum of 5 customers will return the items. **III.** Probability that more than 5 customers will return the items purchased by them. **IV.** Average number of customers who are likely to return the items. **V.** The variance and the standard deviation of the number of returns.	**07**

OR

Q.3 **(a)** Discuss Poisson distribution with example. **03**

(b) Differentiate stratified sampling and cluster sampling. **04**

(c) Explain bar chart, pie-chart and scatter plot with proper example in data science. **07**

Q.4 **(a)** Consider the following transactional dataset: **03**

TID	Items Bought
T1	{M,O,N,K,E,Y}
T2	{D,O,N,K,E,Y}
T3	{M,A,K,E,O}
T4	{M,U,C,K,Y}
T5	{C,O,K,I,E,M}

Calculate support, confidence and lift of rule M→O.

(b) What is sampling in data science? Illustrate the steps used in any sampling process. **04**

(c) Consider following data (X Y)=(20, 40), (40,85), (25, 50), (48, 90), (20, 41), (30, 62). Where X represents the marks of a student in midterm exam and Y represents the final exam marks. Predict the final exam marks of a student who received 45 marks in midterm exam using linear regression. **07**

OR

Q.4 **(a)** Explain support, confidence and lift in association rule learning. **03**

(b) Discuss marginal probability and conditional probability with proper example. **04**

(c) Following table shows the data regarding number of visitors in theme park on various days. **07**

Sunny?	High Temperature?	Weekend?	Lots of Visitors?
Yes	Yes	Yes	Yes
Yes	No	Yes	Yes
No	Yes	No	Yes
Yes	Yes	No	Yes
Yes	Yes	No	Yes
Yes	No	No	No
No	No	Yes	No

Using Bayesian theorem, decide the number of visitors (Lots of visitors= yes or no) under following conditions:

I. Sunny=No, High Temperature= yes and Weekend=yes

II. Sunny=No, High Temperature= no and Weekend=no

Q.5 **(a)** Differentiate regression and correlation. **03**

(b) Discuss precision, recall, F-score and specificity with respect to evaluation of classification result. **04**

(c) Draw the decision tree for data given in below table using Gini Index as an attribute selection measure. Also decides the Class of tuple with Owns Home= Yes, Married= No, Gender= Female, Employed= No. **07**

Owns Home?	Married?	Gender	Employed?	Class
Yes	Yes	Male	Yes	B
No	No	Female	Yes	A
Yes	Yes	Female	Yes	C
Yes	No	Male	No	B
No	Yes	Female	Yes	C
No	No	Female	Yes	A
No	No	Male	No	B
Yes	No	Female	Yes	A
No	Yes	Female	Yes	C
Yes	Yes	Female	Yes	C

OR

Q.5 **(a)** Discuss Z-score and Cook's distance as outlier analysis for regression. **03**

(b) Consider the classification result where 80 data is classified as a positive class and out of this classification, 60 data is correctly classified (originally data is from positive class) and 20 data are incorrectly classified (originally data is from negative class). Also, 90 data is classified as a negative class and out of this classification, 80 data is correctly classified (originally data is from negative class) and 10 data is wrongly classified (originally data is from positive class). Calculate precision, recall, F-score and accuracy of this classification result. **04**

(c) Discuss decision tree algorithm for classification of data with example. **07**

****