Seat No.: _____ Enrolment No._____

GUJARAT TECHNOLOGICAL UNIVERSITY

ME - SEMESTER–1 (NEW) EXAMINATION – WINTER 2018

Subject Code: 3710219 **Date: 07/01/2019**

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 mark.**

			MARKS
Q.1*	**(a)**	The test for rare disease is conducted, where 1% of the population is infected. It is a highly sensitive and specific test, which is not quite perfect: • 99% of sick patients test positive. • 99% of healthy patients test negative. Given that a patient tests positive, what is the probability that the patient is actually sick?	**03**
	(b)	Consider spam filter. Nonspam is called "ham". There are 1500 spam versus 3672 ham. The word "meeting" occurs 16 times in spam folder. There are 153 occurrences of the word "meeting" in ham. Compute the chance that an Email is spam only knowing it contains the word "meeting"?	**04**
	(c)	Consider two sets of data S={(x,y) = (1,25),(10,250),(100,2500),(200,5000)} R={(restaurant,ranking)=("ABC","fivestars"), ("PQR","twostars"),("Z","zerostars"),("T","twostars"). How linear regression can be applied to find out trend and variation?	**07**
Q.2	**(a)**	For the prediction that student will be allowed to sit in the placement of an Infocom company. Identify the predictor variable, target variable, data type of variable.	**03**
	(b)	How data science process differ in case of (i) recommendation system (ii) predicting the weather.	**04**
	(c)	Why anyone working with data should do exploratory data analysis? Consider 31 datasets, Each dataset represents one day's worth of advertisements and clicks recorded on the Today Time's home page. Each row represents a single user. There are five columns : age, gender, image impression, number of clicks and logged in time. What kind of exploratory data analysis can be performed?	**07**
		OR	
	(c)	What is central limit theorem? Give example for following with respect to central limit theorem. (i) Variable is normally distributed (ii) Variable is greater than a certain number (iii) Variable is less than a certain number	**07**

Q.3	**(a)**	Find the variance for following set of numbers 28,29,30,31,32	**03**
	(b)	What are outliers? How to identify outliers in data?	**04**
	(c)	Consider the following sentences for sentiment analysis (i) *The weather is pleasant.* (ii) *The devotional movie is excellent* (iii) *The bicycle race is exciting.* What type of encoding can be used to represent sentiment data? Explain.	**07**

OR

Q.3	**(a)**	If a company wants to estimate growth in sales of a company based on current economic conditions. What kind of analysis company must do? What are the benefits?	**03**
	(b)	What is the difference between API and library files. Explain the use of API for data collection.	**04**
	(c)	The Music Timeline App illustrates a variety of music genres popular from 2010 to present day, based on how Music users have an artist or album in their library, and other data such as album release dates. Which data visualization techniques can be used to represent what kind of data? keep Music App in mind.	**07**

Q.4	**(a)**	What are the challenges for data storage and management?	**03**
	(b)	The analysis of age is to be performed for customer visiting the mall. Which visualization techniques can be used?	**04**
	(c)	Which methods can be used to fill the missing data? Explain the case of numerical and categorical data.	**07**

OR

Q.4	**(a)**	Which types of data are used in data science?	**03**
	(b)	In the analysis of product category, how SVM can be applied?	**04**
	(c)	What are retinal variables? How encoding of retinal variables is done?	**07**

Q.5	**(a)**	Which type of statistics can overcome the issue of outliers?	**03**
	(b)	How data form multiple sources can be handled?	**04**
	(c)	The task is to automate the assignment of new products to company's product categories, For example stereo is to be categorized as electronic system. This is which type of problem and what kind of learning can be applied? Which method best suits for this ? Justify.	**07**

OR

Q.5	**(a)**	"Significant skewness indicate that the mean and standard deviation are not good measures of distribution". True or False? Justify.	**03**
	(b)	How distribution of categorical data can be calculated?	**04**
	(c)	Explain the process of credit card transaction. How can it be verified that the transaction is fraudulent or not?	**07**

****