



QP CODE: 26019598



Reg No :

Name :

M.Sc DEGREE (CSS) EXAMINATION, MARCH 2026

Fourth Semester

MSc Statistics with Data Science

Core Courses - ST040402 - APPLIED ALGORITHMS AND ANALYSIS OF MULTITYPE DATA

2020 Admission Onwards

1F62FC0D

Time: 3 Hours

Weightage: 30

(Answer any **THREE** questions. Each question carries a weight of **10**)

1. (a) Briefly explain the EM algorithm.
(b) Generate 2 univariate normal random variables, check the performance of the EM algorithm and comment on the findings.
2. (1) Consider the '**bill_authentication**' data set, which is extracted from images that were taken for the evaluation of an authentication procedure for banknotes. Create a model for classifying the variable – Class using SVM, based on the other variables of the dataset and find the efficiency of the model.
(2) Construct a logistic regression model as an alternative to SVM and compare their efficiencies.
3. (a) The following 'Animal Milk Data' contains 9 animals and the constituents of their milk. Conduct both metric and non-metric multi-dimensional scaling on the given data and visualize the similarity/dissimilarity between the samples by plotting points in two dimensional plots.

Animal	Water %	Protein %	Fat %	Lactose %
Horse	90.1	2.6	1.0	6.9
Donkey	90.3	1.7	1.4	6.2
Mule	90.0	2.0	1.8	5.5
Camel	87.7	3.5	3.4	4.8
Llama	86.5	3.9	3.2	5.6
Zebra	86.2	3.0	4.8	5.3
Sheep	82.0	5.6	6.4	4.7
Buffalo	82.1	5.9	7.9	4.7
Deer	65.9	10.4	19.7	2.6





(b) The following table takes the similarity index of eight brands of TV sets available in the Indian market. Both the rows and columns represent brands of TV. Determine how these 8 brands are perceived by Indian consumers and plot a positioning map of the eight brands.

TV Brand	LG	SAMSUNG	TCL	VIDEOCON	IMPEX	SONY	PANASONIC	MI
LG	0	3	6	8	1	2	7	8
SAMSUNG	3	0	4	6	4	5	2	5
TCL	6	4	0	3	2	4	6	1
VIDEOCON	8	6	3	0	3	5	4	7
IMPEX	1	4	2	3	0	2	8	5
SONY	2	5	4	5	2	0	3	6
PANASONIC	7	2	6	4	8	3	0	5
MI	8	5	1	7	5	6	5	0

4. Each of three football enthusiasts had devised his own system for predicting the outcomes of collegiate football games. Twelve games were selected at random, and each sportsman presented a prediction of the outcome of each game. After the games were played, the results were tabulated using '1' for successful prediction and '0' for unsuccessful prediction. Use Cochran's Q test to test whether each sportsman is equally effective in his ability to predict the outcomes of the football games.

Game	Sports Man 1	Sports Man 2	Sports Man 3	Total
1	1	1	1	3
2	1	1	1	3
3	0	1	0	1
4	1	1	0	2
5	0	0	0	0
6	1	1	1	3
7	1	1	1	3
8	1	1	0	2
9	0	0	1	1
10	0	1	0	1
11	1	1	1	3
12	1	1	1	3
Total	8	10	7	25

5. Use the cars dataset, a default dataset in R
- (1) Plot the scatter diagram and fit a regression line. Briefly explain the model.
 - (2) How well does this equation fit the data?
 - (3) Test the significance of the regression coefficients.
 - (4) Give the 95% confidence interval for coefficients of the linear model and error variance.





(5) Predict the stopping distance when the speed is 20 miles per hour. Also construct the 95% confidence interval for the same.

6. Consider the **Coronary Heart Disease Data**

(a) Fit a logistic regression model.

(b) Test the significance of the regression coefficients.

(c) Fit a new logistic model by removing the insignificant variables and compare the efficiencies of the models. Interpret the better model.

(d) Compute the efficiency of the model.

