Program:	**T.Y.B.Sc CS**	**Semester:**	**VI**	**Program Code:**	**1S00196**
Course:	**Data Science**			**Course Code:**	**USCS601**

Duration: 2 ½ Hours **Examination Pattern:** REV23 - Autonomous - External **Max. Marks:** 75

Instructions:

1. All questions are compulsory.
2. Figures to the right indicate full marks.
3. Draw neat diagrams wherever necessary.
4. Use of simple calculator is allowed

Examination:

REGULAR

Q. 1 Attempt ANY FOUR from the following: **(20M)**

(a) What are real-world applications where data science is commonly utilized, and how does it contribute to decision-making processes?

(b) Differentiate structured and unstructured data with suitable examples.

(c) What are the major sources of data collection, and which sources are most reliable for business analytics projects?

(d) What is data cleaning? Explain the different methods available for data cleaning.

(e) Define feature selection, and why is it important in machine learning?

(f) Discuss the significance of data merging in data science.

Q. 2 Attempt ANY FOUR from the following: **(20M)**

(a) How do descriptive statistics such as mean, median, mode, and standard deviation help in summarizing and understanding the distribution of data?

(b) Write a short note on Exploratory Data Analysis

(c) What is a hypothesis? What are the types of hypothesis testing?

(d) Two training programs are conducted for two groups of employees.

- Group A scores: 70, 75, 80, 85, 90
- Group B scores: 65, 68, 72, 74, 71

Test at 5% significance level whether there is a significant difference between the two groups.(Use T Test)

(e) Discuss Linear Regression in detail, including its assumptions, working principle, and applications.

(f) Explain the concept of hyperparameter tuning in machine learning.

Q. 3 Attempt ANY FOUR from the following: **(20M)**

(a) A credit card fraud detection system predicts transactions as Fraud (Positive) or Legitimate (Negative).

After evaluating 500 transactions, the results are:

	Predicted Positive	Predicted Negative
Actual Positive	40	10
Actual Negative	30	420

Define a confusion matrix, calculate the Accuracy, Precision, Recall, and F1-Score of the model.

(b) What is data storytelling, and why is it important in communicating insights from data to different stakeholders?

6 2 2 3 4 1

(c) Write a note on the bar chart and Box plot.

(d) Define the ROC curve, and explain why AUC is important in evaluating classification models.

(e) What are the major data privacy concerns in data management, and how can organizations address them?

(f) What is data Governance? How does data governance contribute to ensuring data quality, consistency, and compliance within an organisation?

Q. 4 Attempt <u>ANY FIVE</u> from the following: **(15M)**

(a) Define terms:

i) Categorical Data

ii) Ordinal Data

(b) What is under fitting and overfitting of the data?

(c) Mention two differences between Matplotlib and Seaborn.

(d) Write a short note on Chi-Square Test.

(e) Explain bias, variance and trade-off.

(f) Marks of 10 students: 12, 15, 18, 20, 15, 22, 25, 18, 15, 20
Calculate mean, median and mode.

----------x----------

One Sample T Table

Sample Size (n)	df = n − 1	10% (0.10)	5% (0.05)	1% (0.01)
2	1	6.314	12.706	63.657
3	2	2.920	4.303	9.925
4	3	2.353	3.182	5.841
5	4	2.132	2.776	4.604
6	5	2.015	2.571	4.032
7	6	1.943	2.447	3.707
8	7	1.895	2.365	3.499
9	8	1.860	2.306	3.355
10	9	1.833	2.262	3.250

Two Tailed T Table

df	10% (0.10)	5% (0.05)	1% (0.01)
1	6.314	12.706	63.657
2	2.920	4.303	9.925
3	2.353	3.182	5.841
4	2.132	2.776	4.604
5	2.015	2.571	4.032
6	1.943	2.447	3.707
7	1.895	2.365	3.499
8	1.860	2.306	3.355
9	1.833	2.262	3.250
10	1.812	2.228	3.169