

THE KAVERY ENGINEERING COLLEGE*(An Autonomous Institution)***B.E/B.TECH DEGREE END SEMESTER THEORY EXAMINATIONS, APRIL-MAY 2025****Sixth Semester****Agricultural Engineering****OCS353 - Data Science Fundamentals****(Common to Chemical Engineering)****Regulations - 2021****Duration: 3 Hrs****Maximum: 100 Marks****PART – A (ANSWER ALL QUESTIONS) (Marks 10*2=20)**

Q.No	Questions	BLT	CO	Marks
1.	List any two benefits of data science.	L1	1	2
2.	State the purpose of data preparation in the data science process.	L2	1	2
3.	Define hierarchical indexing in Pandas.	L1	2	2
4.	How does time series data differ from regular data?	L2	2	2
5.	Differentiate between classification and regression.	L4	3	2
6.	Identify one difference between supervised and semi-supervised learning.	L2	3	2
7.	State the purpose of matplotlib. pyplot in Python.	L1	4	2
8.	List two features of Seaborn that enhance data visualization.	L1	4	2
9.	Name two tools commonly used for processing large datasets.	L1	5	2
10.	What is the purpose of model presentation in data projects?	L2	5	2

PART – B (ANSWER ALL QUESTIONS) (Marks 5*16 = 80)

Q.No	Questions	BLT	CO	Marks
11.	a i Explain the complete Data Science Process with suitable examples.	L2	1	8
	ii Describe the steps involved in data preparation and exploratory data analysis.	L2	1	8
	Or			
	b Explain how the Data Science process can be applied in the healthcare industry to improve patient diagnosis and treatment outcomes.	L2	1	16
12.	a Given a CSV file containing customer transaction data with missing values, string-based columns (like customer names and product categories), and timestamped entries, write a Python program using Pandas to: - Clean the missing data appropriately - Perform aggregation on total transaction amount by category	L4	2	16

- Extract and visualize monthly sales trends using time series analysis

Or

- b You are provided with a dataset retail_sales.csv containing daily sales data across different product categories. Perform the following operations using Python and Pandas:
- Import the dataset and convert the 'Date' column into a proper DateTime index
 - Handle the missing Sales_Amount values using an appropriate imputation technique
 - Resample the data to calculate weekly total sales for each category
 - Plot the resampled time series data to visualize sales trends
 - Analyze the observed trends, seasonality, or irregularities in the time series

Justify each data manipulation step with code and brief explanations.

Table: retail_sales.csv

Date	Product_Category	Sales_Amount
2024-01-01	Electronics	1200
2024-01-02	Clothing	800
2024-01-04	Electronics	
2024-01-05	Grocery	600
2024-01-06	Clothing	900
2024-01-08	Electronics	1100
2024-01-09	Grocery	
2024-01-10	Clothing	950

13. a Given a labeled dataset of email messages marked as 'spam' or 'not spam', apply a supervised machine learning classification algorithm using Python. Summarize the steps involved in data preprocessing, model training, testing, and evaluation.

Or

- b You are provided with a dataset customer_data.csv containing the following features:

- Customer_ID
- Annual_Income (in thousands)
- Spending_Score (1-100)
- Age

Perform the following tasks using Python and suitable libraries (e.g., Pandas, Scikit-learn, Matplotlib/Seaborn):

- Preprocess the data by removing irrelevant columns and scaling the necessary features
- Apply K-Means clustering to group customers into meaningful segments (choose an appropriate value of k using the Elbow Method)
- Evaluate the performance of the clustering using the Silhouette Score

		iv.	Visualize the clusters using a scatter plot (e.g., Income vs Spending Score)			
		v.	Interpret the results to provide business insights, such as identifying high-value customer groups or potential marketing targets			
14.	a	i	Explain the process of creating simple line plots and scatter plots using Matplotlib. Include examples with annotations, colors, and legends.	L2	4	8
		ii	Outline various techniques for visualizing distributions in Python, including histograms, density plots, and contour plots. Illustrate with suitable examples.	L2	4	8
		Or				
	b	i	Illustrate the use of subplots and customization options (like labels, titles, styles, and ticks) in Matplotlib. Show how to plot multiple graphs in a grid.	L2	4	8
		ii	Explain how Seaborn is different from Matplotlib in terms of ease of use and visual appeal. Using a multivariate dataset such as the built-in tips dataset in Seaborn, create multiple plots (such as bar plot, box plot, and scatter plot) to analyze the relationship between total bill, tip, day, and time.	L2	4	8
15.	a	i	Explain the challenges of handling large volumes of data. Outline at least four techniques or best practices to manage large datasets effectively.	L2	5	8
		ii	Summarize the key steps in developing a recommender system using a large dataset. Discuss how to handle memory and performance issues.	L2	5	8
		Or				
	b		With a case study of predicting malicious URLs, explain the process of data preparation, feature selection, model building, and result automation.	L2	5	16