2025

M.Sc. 1st Semester Examination

COMPUTER SCIENCE

Paper : COSE405A8

(Data Science Lab)

(Practical)

Full Marks : 15 Time : One Hour

The figures in the margin indicate full marks.

Answer any ***one*** question : 10×1=10

(Choose the question on the lottery basis)

1. Write Python code in a Jupyter Notebook to create a Pandas DataFrame with the following data :

Name	Age	Score
Alice	23	85
Bob	25	90
Carol	22	88

(a) Display the first two rows of the DataFrame.

(b) Calculate the mean of the Score column.

(c) Display the age-wise highest scorer.

(d) Eliminate all the rows with missing values.

2. Create a Python dictionary with the following data and save it in the drive :

   data = [{"Name": "Alice", "Age": 23, "Score": 85},

   {"Name": "Bob", "Age": 25, "Score": 90}

   Add at least 10 data with similar properties.

   (a) Read students.csv.

   Add a new column Passed which is True if Score >= 50 else False.

   (b) Save the modified DataFrame as students_ updated.csv.

   (c) Load the CSV file

   (d) Filter and display all rows where Score > 80.

   (e) Add a new column Passed which is True if Score >= 50 else False.

   (f) Save the modified DataFrame as students_updated.csv.

3. (a) Write Python code to : Connect to a MySQL/ PostgreSQL database named school_db using mysql.connector or psycopg2.

   (b) Print a message confirming if the connection was successful. Close the connection properly. Display the resulting DataFrame sorted by Score in descending order.

   (c) Execute a parameterized SQL query to fetch all students with Score >=minimum_score from the students table.

4. (a) Load a CSV file students.csv with columns : Name, Age, Score, Grade.

   (b) Write Python code to check for missing values in each column and display the count of missing values.

   (c) Remove all rows where any column has missing values.

   (d) Display the shape of the DataFrame before and after removing missing rows.

   (e) Replace missing values in the Score column with the mean of the column.

   (f) Replace missing values in the Grade column with the most frequent value.

   (g) Display the updated DataFrame.

   (h) Check if the DataFrame contains duplicate rows.

   (i) Remove duplicates and display the number of rows removed.

   (j) Display the cleaned DataFrame.

5. You are provided with a CSV file sales_data.csv containing the following columns :

   Date, Product, Category, Units_Sold, Revenue.

   (a) Write Python code to create a new column Unit_Price by dividing Revenue by Units_Sold.

   (b) Display the first 5 rows of the updated DataFrame.

   (c) Group the data by Category and calculate :

      (i) Total Revenue

      (ii) Average Units_Sold

   (d) Display the aggregated results.

   (e) Create a new column Revenue_Level based on Revenue :

      (i) "High" if Revenue > 1000

      (ii) "Medium" if Revenue is between 500 and 1000

      (iii) "Low" if Revenue < 500

   (f) Display the updated DataFrame.

6. You are provided with a CSV file employee_data.csv containing the following columns :

   Employee_ID, Name, Department, Age, Salary.

   (a) Write Python code to calculate and display the mean, median, and standard deviation for the Age and Salary columns.

   (b) Use Pandas and/or Matplotlib/Seaborn to plot the histogram and boxplot for the Salary column.

   (c) Identify if there are any outliers in the Salary data.

   (d) Group the data by Department and calculate :

      (i) Mean Salary per department

      (ii) Maximum Age per department

   (e) Display the results in a tabular format.

7. You are provided with a CSV file sales_data.csv containing the following columns :

   Product, Category, Units_Sold, Revenue.Mm

   (a) Write Python code to plot a histogram of the Revenue column using Matplotlib.

   (b) Set appropriate title, x-label, y-label, and number of bins.

   (c) Create a box plot of the Units_Sold column to visualize its distribution and identify outliers.

   (d) Add labels and a title for clarity.

   (e) Create a box plot of Revenue grouped by Category to compare distributions across categories.

   (f) Ensure the plot is well-labeled with title, x-label, and y-label.

8. You are provided with a CSV file student_scores.csv containing the following columns: Student_ID, Math, Science, English, History, Total_Score.

   (a) Write Python code to calculate the correlation matrix of the numeric columns (Math, Science, English, History, Total_Score).

   (b) Plot a heatmap of the correlation matrix using Seaborn.

   (c) Ensure the plot has a title and annotations showing correlation values.

   (d) Create a pair plot of the numeric columns to visualize the relationships between different subjects.

   (e) Color the plots by a categorical variable, e.g., Total_Score bins (High/Medium/Low).

9. You are provided with a CSV file employee_data.csv containing the following columns :

   Employee_ID, Age, Experience_Years, Salary, Bonus.

   (a) Write Python code to calculate the correlation matrix for the numeric columns (Age, Experience_Years, Salary, Bonus) using Pandas.

   (b) Display the correlation matrix.

   (c) Identify the strongest positive and strongest negative correlations from the matrix.

   (d) Write a brief interpretation of what these correlations imply in a business or HR context.

10. You are tasked with simulating and analyzing probability distributions using Python.

    (a) Using NumPy, generate 1000 random samples from the following distributions :

        (i) Normal distribution with mean = 50 and standard deviation = 10

        (ii) Binomial distribution with n = 10 trials and probability p = 0.5

    (b) Display the first 10 samples from each distribution.

    (c) Plot histograms for both distributions to visualize their shape.

    (d) Calculate and display the mean, median, and standard deviation of the generated samples.

11. You are provided with a CSV file employee_salaries.csv containing the following columns :

    Department, Salary.

    (a) Write Python code to separate the salaries of employees from Department A and Department B.

    (b) Perform an independent two-sample t-test to determine if there is a significant difference between the mean salaries of the two departments.

    (c) Display the t-statistic and p-value.

    (d) Using a significance level of 0.05, interpret the results of the t-test.

    (e) Write a brief conclusion stating whether the null hypothesis is rejected or not, and explain what it implies in the context of employee salaries.

12. You are provided with a CSV file housing_data.csv containing the following columns :

    Area_sqft, Bedrooms, Age, Price.

    (a) Load the dataset using Pandas.

    (b) Split the data into features (X) and target (y = Price).

    (c) Split the data into training and testing sets (80% train, 20% test).

    (d) Build a Linear Regression model using Scikit-Learn and fit it on the training data.

    (e) Display the coefficients and intercept of the model.

13. You are provided with a CSV file housing_data.csv containing the following columns :

    Area_sqft, Bedrooms, Age, Price

    (a) Bb Predict prices for the test set.

    (b) Calculate and display the Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 score.

    (c) Plot a scatter plot of actual vs predicted prices.

    (d) Write a brief interpretation of the model performance based on the metrics.

14. You are provided with a CSV file customer_data.csv containing the following columns :

    Customer_ID, Age, Annual_Income, Spending_Score.

    (a) Load the dataset using Pandas.

    (b) Select the numeric features Age, Annual_Income, Spending_Score for clustering.

    (c) Apply K-Means clustering with 3 clusters using Scikit-Learn.

    (d) Display the cluster labels assigned to each customer.

    (e) Display the coordinates of cluster centroids.

15. You are provided with a CSV file employee_data.csv containing the following columns : Employee_ID, Department, Age, Salary, Experience_Years.

    (a) Select the numeric features Age, Salary, and Experience_Years.

    (b) Apply Min-Max Scaling to transform the features into the range [0, 1].

    (c) Display the first 5 rows of the scaled data.

16. You are provided with a CSV file monthly_sales.csv containing the following columns :

    Month, Sales.

    (a) Load the dataset using Pandas and set the Month column as the DateTime index.

    (b) Plot the time series of Sales to visualize trends and seasonality.

    (c) Check for stationarity using a rolling mean/variance plot or Augmented Dickey-Fuller test.

17. You are provided with a CSV file student_scores.csv containing the following columns :

    Hours_Studied, Previous_Score, Current_Score.

    (a) Load the dataset and split it into features (X = Hours_Studied, Previous_Score) and target (y = Current_Score).

    (b) Split the data into training and testing sets (80% train, 20% test).

    (c) Build a Sequential Neural Network with :

        (i) Input layer matching the number of features

        (ii) One hidden layer with 10 neurons and ReLU activation

        (iii) Output layer with 1 neuron (for regression)

    (d) Compile the model using mean squared error (MSE) loss and Adam optimizer.

    (e) Train the model for 50 epochs and display the training history.

    Practical Notebook and Viva-voce : 05
