|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Department of Data Science**  **Faculty of Computing and Information Technology**  **University of the Punjab**    **Semester Fall 2024**  **Final Term Examination**   |  |  | | --- | --- | | **Course Code: DS-302** | **Title: Introduction to Data Science** | | **Shift / Section:** | **Max Time: 120 minutes** | | **Student ID:** | **Student Name:** | |  |

**Instructions:**

|  |
| --- |
| **Sheet No.:** |
| **Invigilator’s Signature** |
| **Date: 15 Jan 2025** |

1. Do not forget to pray before starting to attempt the paper. Trust me it helps.  
   Remember! SOMEONE is always with you (Be Relaxed), and HE is also watching you (Be Honest)
2. Question Paper is **SELF EXPLANATORY**. Write your assumption. In case of error in question, credit goes to the student.☺
3. Solve your paper using Black/Blue Pen only strictly in the given space.
4. Check the all the 11 pages are present in answer sheet, if any page is missing, please inform the invigilator.
5. Attempt All Questions in a precise fashion.
6. No Answer sheet is required.

Good Luck

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Q3** | **Q4** | **Total** |
| **20** | **6** | **6** | **8** | **40** |
|  |  |  |  |  |

**Examiner’s Signature : \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Marks: \_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**DO NOT OPEN UNTIL YOU ARE TOLD TO DO SO.**

**Q#1: (20 marks)**

You are working as a data analyst for a retail chain. You are provided with a dataset containing columns: ProductID, Category, SalesDate, UnitsSold, Revenue, and Discount. The company wants you to preprocess and analyze this data to generate actionable insights.

For this question, **only use Pandas library functions** to complete the steps.

1. **Task:** Extract all rows where the Revenue is greater than the average Revenue of the entire dataset. Display only the ProductID and Revenue columns for these rows.
2. **Task:** Identify the columns in the dataset that have missing values. Write code to drop rows where the UnitsSold column has missing values.
3. **Task:** Create a new column, Net Revenue, by subtracting the Discount from the Revenue. Ensure rows with missing values in the Discount column are handled by assuming the discount is zero.
4. **Task:** Group the dataset by Category and calculate the total Revenue, average UnitsSold, and total number of products (ProductID) in each category.
5. **Task:** Perform a left join between the dataset and another DataFrame category\_info\_df containing Category and CategoryDescription columns. Retain all records from the original dataset and add the CategoryDescription to each row.
6. **Task:** Write code to reshape the dataset so that Category values become columns, with UnitsSold as the values. Fill missing values with the column's mean.
7. **Task:** Write code to rank products (ProductID) by Revenue within each Category. Display the top 3 products for each Category.
8. **Task:** Merge the dataset with another DataFrame containing ProductID and SupplierName. Retain only rows with matching ProductID values in both datasets.
9. **Task:** Use a custom aggregation method to calculate the percentage contribution of each Category to the total Revenue.
10. **Task:** Create a pivot table to show the total Revenue by Category, but exclude the Category values where the total revenue is below the average revenue across all categories. The pivot table should only display categories that meet this condition.

**Q#2: (6 Marks)**

You are working with a dataset that contains information about customers, such as their age, gender, and income. The target variable indicates whether the customer made a purchase (1) or not (0).

1. **Task**: Encode the 'gender' column so that the categorical values are converted into numerical labels. Provide the code for this transformation. Also briefly explain what this encoding process does and why it is used in machine learning, especially when dealing with categorical variables.
2. **Task**: Apply MinMax scaling to the 'age' and 'income' columns to bring them into the range [0, 1]. Explain the advantages of MinMax scaling compared to other scaling methods. Provide the code to apply the scaling.
3. **Task**: Using the preprocessed data, train a Random Forest classifier and evaluate the model's performance by computing the classification accuracy and generating the confusion matrix. Provide the code to train the model, calculate the accuracy, and display the confusion matrix

**Q#3: (6 Marks)**

You are provided with a dataset containing columns 'Salary', 'Age', and 'Department'. To use the PandasAI library, first, write how you would set up the environment to use PandasAI and then write a query that:

1. Calculates the average 'Salary' for employees in each department.
2. Handles any missing values in the 'Salary' column by filling them with the median value before the calculation.
3. Filters the result to only include departments where the average salary is greater than $50,000.

**Q#4: (8 Marks)**

Given the actual and predicted labels for a sample of 12 pictures, construct a confusion matrix and calculate the performance metrics such as accuracy, precision and F1-score. How would you interpret the results based on the confusion matrix?

**Dataset:**

* **Actual labels:** [1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0] (1 = cat, 0 = dog)
* **Predicted labels:** [0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1]

Also explain what are **Type 1 error** and **Type 2 error** in the context of model evaluation? Explain the difference between them with an example.