

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Roll No |  |  |  |  |  |  |  |  |  |  |  |

PRESIDENCY UNIVERSITY BENGALURU

**SET-B**

SCHOOL OF COMMERCE

**END TERM EXAMINATION – MAY/JUNE 2024**

**Semester :** Semester IV - 2022

**Course Code :** BSE2055

**Course Name :** Data Mining and Forecasting

**Program :** B.Sc. Economics

**Date :** June 21, 2024

**Time :** 9.30 AM – 12.30 PM

**Max Marks :** 100

**Weightage :** 50%

**Instructions:**

1. *Read all questions carefully and answer accordingly.*
2. *Question paper consists of 3 parts.*
3. *Scientific and non-programmable calculator are permitted.*
4. *Do not write any information on the question paper other than Roll Number.*

**PART - A**

**ANSWER ANY 5 QUESTIONS 5Q X 2M = 10M**

# What is the primary objective of Exploratory Data Analysis (EDA)?

(CO4,CO3) [Knowledge]

# What function is used to calculate the mean of student scores in the DataFrame?

(CO3,CO4) [Knowledge]

# How is the coefficient of variation calculated and what does it represent?

* 1. Explain why data mining is important for making smarter business decisions.
  2. Define the term "Knowledge Discovery in Databases (KDD)".

# What role does EDA play in the data mining process?

(CO4,CO3) [Knowledge] (CO4,CO3) [Knowledge] (CO4,CO3) [Knowledge] (CO4,CO5) [Knowledge]

# Which visual tool in EDA is best for examining the relationship between two variables?

(CO5,CO4) [Knowledge]

**PART - B**

**ANSWER ANY 5 QUESTIONS 5Q X 10M = 50M**

Evaluate the strengths of using statsmodels for econometric analyses compared to other Python libraries.

(CO4,CO3) [Comprehension]

# Split a dataset into training and test sets, then explain the importance of this process in machine learning model validation.

Dataset: features = np.arange(10).reshape(5, 2), labels = [0, 1, 0, 1, 0]

(CO3,CO4) [Comprehension]

# Create a regression model using the statsmodels library to predict housing prices based on the number of rooms. Analyze the significance of the number of rooms on price.

Dataset: rooms = [3, 4, 5, 6, 7], prices = [200000, 250000, 300000, 350000, 400000]

(CO3,CO4) [Comprehension]

# Propose a preprocessing pipeline for a dataset with categorical data, using both One-Hot Encoding and Label Encoder where appropriate. Justify your choices for each type of encoding.

Dataset: data = {'Brand': ['Apple', 'Samsung', 'Apple'], 'Model': ['X', 'S', 'Y'], 'Condition': ['New', 'Used', 'New']}

(CO4,CO3) [Comprehension]

# Critically analyze the application of scikit-learn in industrial contexts for predictive modeling tasks.

(CO5,CO4) [Comprehension]

# Develop a multiple linear regression model using the dataset provided and assess the impact of each variable on sales.

 Dataset: R&D Spend = [150000, 120000, 130000],

 Administration = [50000, 60000, 55000],

 Marketing Spend = [200000, 210000, 205000],

 Sales = [400000, 370000, 390000]

(CO4,CO5) [Comprehension]

* 1. Implement simple linear regression on a dataset of annual sales based on advertising spend using Sklearn Package. Evaluate the effectiveness of advertising spend on sales growth. Dataset: advertising spend = [100, 200, 300, 400, 500], annual sales = [10, 20, 25,

40, 50]

(CO4,CO5) [Comprehension]

**PART - C**

**ANSWER ANY 2 QUESTIONS 2Q X 20M = 40M**

* 1. You are a data analyst working for a retail company that has several stores across the city. The company wants to open two new distribution centers to serve these stores more efficiently. The locations of the stores are represented by the given data points (x, y) coordinates. Using the K-Means clustering algorithm with two centers, find the optimal locations for the two distribution centers.

Data:

 Point 1: (3, 5)

 Point 2: (2, 8)

 Point 3: (6, 4)

 Point 4: (7, 6)

 Point 5: (4, 3)

 Point 6: (5, 7)

 Point 7: (1, 2)

 Point 8: (8, 5)

(CO5,CO4) [Application]

* 1. Data: Customers' satisfaction vs. service usage: [(3, 5), (3, 4), (2, 8), (8, 7), (6, 4),

(7, 3)]

Question: Given the initial cluster centers at (2, 5) and (5, 5), perform one iteration of K-Means clustering. Assign each point to the nearest center and compute the new centers.

(CO4,CO5) [Application]

* 1. Data: Locations of community centers on a map: [(1, 2), (2, 1), (3, 2), (5, 4), (6, 5)] Question: Starting with initial centers at (2, 2) and (5, 5), classify each location to its nearest center and calculate the new cluster centers after the first iteration.

(CO5,CO4) [Application]