

JAIPURIA INSTITUTE OF MANAGEMENT, NOIDA
PGDM / PGDM (M) / PGDM (SM)
FOURTH TRIMESTER (Batch 2024-26)

END TERM EXAMINATIONS, OCTOBER 2025

SET-1

Course Name	Machine Learning	Course Code	20827
Max. Time	2 hours	Max. Marks	40 MM

INSTRUCTIONS:

- a) Attempt all questions in a single Jupyter Notebook.
- b) The data set for this case is available in a csv file named "data_set_1".
- c) When you split the data into train and test sets, set the value of random_state to the last two digits of your Roll Number
(e.g., if your Roll Number is PGFA2213, use random_state = 13).
- d) Provide interpretation and explanation for all outputs in the same notebook using comments or Markdown cells.
- e) Save your Jupyter Notebook in two formats:
 - as an .ipynb file
 - and as a PDF file.
- f) Upload both files.
- g) This is an open book exam. You may refer to the sample code provided with the Question paper.

Read the case below and answer the questions given by analyzing the data using Python.

SpeedEx Logistics: Predicting Delivery Time

SpeedEx Logistics, founded in 2012, has become one of India's most trusted third-party logistics (3PL) partners. A 3PL company is hired by businesses to plan, store, transport, and deliver goods to their customers. Instead of maintaining their own trucks, warehouses, and shipping teams, companies outsource these activities to 3PL providers like SpeedEx.

SpeedEx's customers include:

- **E-commerce platforms:** handling the shipment of customer orders directly to homes and offices.

- **Manufacturers:** moving raw materials or finished goods between production plants and warehouses.
- **Retail and FMCG companies:** replenishing stock to distribution centers and stores.

SpeedEx operates a nationwide network and uses four modes of transport:

- **Road trucks:** flexible and widely used, but vulnerable to traffic and holiday delays.
- **Rail freight:** cost-effective for long distances but limited by train schedules.
- **Air cargo:** the fastest option, albeit expensive, and used for urgent or high-value shipments.
- **Coastal shipping:** inexpensive but slow, primarily used for bulky cargo along India's coasts.

During peak seasons like Diwali, Holi, Eid, and Christmas, SpeedEx struggles to meet delivery promises. Road shipments slow down due to traffic jams, air freight can be costly, warehouses often become congested with small e-commerce packages, and manufacturers express frustration when raw materials do not arrive on time.

To remain competitive, SpeedEx's leadership aims to leverage historical shipment data to predict delivery times in hours. You have been hired as a data analyst to develop and validate a predictive model for delivery times using actual shipment records.

Column	Type	Description
Shipment_ID	String	Unique shipment identifier (e.g., S1001).
Order_Date	Date	Date the shipment was booked. You may derive month/weekday/season.
Origin_Region	Categorical	Region where shipment starts — North, South, East, West.
Destination_Region	Categorical	Region where shipment is delivered — North, South, East, West.
Transport_Mode	Categorical	Mode of transport — Road, Rail, Air, Ship.
Distance_km	Numeric	Route distance (km).
Shipment_Weight_kg	Numeric	Total weight of shipment (kg).
Shipment_Volume_m3	Numeric	Volume (cubic metres).
Total_Shipment_Cost_INR	Numeric	Total transport cost (₹) including fuel, tolls, surcharges.
Warehouse_Handling_Time_hr	Numeric	Time spent in warehouse before dispatch (hours).
Number_of_Items	Integer	Number of individual packages.
Holiday_Season	Categorical	Yes if shipped during Diwali, Holi, Eid, Christmas, etc.; No otherwise.
Customer_Type	Categorical	Type of client — E-commerce, Manufacturer, Retail/FMCG.
Delivery_Time_hr	Numeric (Target)	Actual time taken for delivery (hours).

Q1 Data Understanding & Preprocessing**(9 marks)**

- a) Load the dataset and inspect its structure: data types, null values, and basic statistics. **(5 marks)**
- b) Handle missing values and outliers appropriately. Clearly mention the steps and reasoning. **(3 marks)**
- c) Encode categorical variables for modeling. **(1 marks)**

Q2 Exploratory Data Analysis (EDA)**(10 marks)**

- a) Create summary statistics and plots for numeric columns. **(4 marks)**
- b) Analyze Delivery_Time_hr by Transport_Mode & Holiday_Season. **(3 marks)**
- c) Plot a correlation heatmap for numeric features and evaluate key relationships. **(3 marks)**

Q3 Multicollinearity Check**(5 marks)**

- a) Build a baseline regression and compute Variance Inflation Factor (VIF) for predictors. **(3 marks)**
- b) Identify at least one pair of highly correlated features and explain how you would handle it with justification. **(2 marks)**

Q4 Model Building & Validation**(12 marks)**

- a) Split the data into train/test (80/20) with random_state = last two digits of your Roll Number. **(2 marks)**
- b) Build a Multiple Linear Regression model. **(3 marks)**
- c) Evaluate with R^2 , and RMSE on the test set. **(3 marks)**
- d) Perform residual analysis. **(2 marks)**
- e) Identify top 2–3 significant predictors of delivery time and explain their business meaning. **(2 marks)**

Q5 Executive Summary**(4 marks)**

- a) Create a summary report for SpeedEx leadership. It should include the following:
- What patterns affect delivery time?
 - How accurate the model is?
 - How it can help improve planning & cost optimization?