JAIPURIA INSTITUTE OF MANAGEMENT
LUCKNOW NOIDA JAIPUR INDORE

JAIPURIA INSTITUTE OF MANAGEMENT, NOIDA

PGDM / PGDM (M) / PGDM (SM)

FOURTH TRIMESTER (Batch 2024-26)

END TERM EXAMINATIONS, OCTOBER 2025

SET-2

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_2".
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 PGFA2203, use random_state = 3).
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 as a **single zipped folder**.
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.

Transforming Placements at PIMN

In the heart of Noida, Premium Institute of Management, Noida (PIMN) has built a reputation as one of Uttar Pradesh's top private B-schools since its founding in 1995. PIMN's flagship Post Graduate Diploma in Management (PGDM) attracts students from across India. By the second year, learners choose dual specializations such as Marketing, Finance, Operations, Human Resources, and Business Analytics.

Over the years, PIMN has delivered 100% placement, with an average salary package of ₹10 lakhs per annum. But the Institute's Career Development Cell (CDC) is determined to aim higher, its new goal is to raise the average package to ₹12 lakhs per annum.

To achieve this, CDC has rolled out several initiatives under its Integrated Development Program (IDP):

- Corporate speaker sessions and industry interactions
- Communication and group discussion workshops
- Rigorous quantitative, verbal, and aptitude test training
- Personalized interview practice and feedback

Now, with the 2024–25 graduating batch placed, CDC wants to know:

- Which of its efforts actually impact salary outcomes?
- Are quantitative and aptitude scores helping?
- Do communication, group discussion, and interview ratings really matter?
- Are students with work experience getting better packages?

To answer these questions, CDC has collected detailed placement data from the latest batch and has asked you, a data analyst, to uncover insights and predictors of salary. Your task is to explore the dataset, clean and analyze it, and build a multiple regression model that can explain what drives MBA salaries.

Dataset Description

Column Name	Description
Student_ID	Unique student identifier
SIP_Performance	Summer Internship Project score (out of 100)
Comm_Rating	Communication skills rating
GD_Rating	Group discussion performance rating
Interview_Rating	Personal interview performance rating during IDP
Quant_Skills_Score	Quantitative skills test score (out of 100)
Verbal_Skills_Score	Verbal skills test score (out of 100)
Aptitude_Test_Score	Aptitude test score (out of 100)
PGDM_CGPA	Final CGPA in PGDM
Gender	Gender of the student
Dual_Specialization	Dual specialization track (e.g., MKT_HR, FIN_BA)
Work_Experience_Flag	Whether the student had prior work experience (Yes/No)
CTC_Offered_LPA *(Target)*	Final salary package offered (in ₹ Lakhs per Annum)

Q1 Data Understanding & Preprocessing **(8 marks)**

a) Load the dataset and inspect its structure: data types, null values, and basic statistics. (4 marks)

b) Clean the data. Clearly mention your steps and reasoning. (3 marks)

c) Encode categorical for modeling. (1 marks)

Q2 Exploratory Data Analysis (EDA) **(11 marks)**

a) Present univariate statistics and visualizations for key numeric variables. **(2 marks)**

b) Explore and visualize CTC_Offered_LPA across Work_Experience_Flag and Gender. **(2 marks)**

c) Plot a correlation heatmap for numeric features and discuss the key relationships with salary. **(3 marks)**

d) Analyze and visualize how PGDM_CGPA and SIP_Performance relate to CTC_Offered_LPA. **(2 marks)**

e) Summarize two business insights about salary patterns that the CDC can act on. **(2 marks)**

Q3 Multicollinearity Check **(5 marks)**

a) Build a baseline regression and compute Variance Inflation Factor (VIF) for predictors. **(3 marks)**

b) Identify any highly correlated features and explain how you would address multicollinearity. **(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 using statsmodels or sklearn. **(3 marks)**

c) Evaluate the model using R^2 and RMSE on the test set. **(3 marks)**

d) Perform residual analysis (normality & heteroscedasticity). **(2 marks)**

e) Identify and interpret the top 2–3 predictors of salary and explain their business implications. **(2 marks)**

Q5 Executive Summary **(4 marks)**

Create a summary report for the CDC:

- What patterns affect MBA salary packages
- How reliable your model is
- How CDC can act on these insights to improve average CTC