Department of Artificial Intelligence, SVNIT,SURAT
B.Tech-III ,SEM-V
Subject- Machine Learning(AI301)

LAB ASSIGNMENT-4

1.

Objectives: Handling Class Imbalance and Hypothesis Testing

Dataset: Use the **Credit Card Fraud Detection dataset** (available on Kaggle).

- 284,807 transactions (rows).
- "Class" column = 0 (non-fraud) or 1 (fraud).
- Fraud cases are ~0.17% → strong class imbalance.

Part A : Perform EDA on dataset

Part B : **Exploring Class Imbalance**

1. Plot class distribution (pie chart / bar plot).
2. Discuss why class imbalance is problematic for machine learning models.

Part C: Under-sampling and Over-sampling

1. Apply **Random Under-Sampling (RUS)**: balance classes by downsampling majority class.
2. Apply **Random Over-Sampling (ROS)** or **SMOTE** (Synthetic Minority Over-sampling Technique).

Part D: Discussion & Insights

1. Which method (under/over-sampling) worked best and why?
2. Which metrics (precision, recall, F1) are most useful for fraud detection?
3. How does hypothesis testing support feature selection in imbalanced datasets?

2.

Objectives:

- Understand the concept of **simple and multiple linear regression**.
- Perform **exploratory data analysis (EDA)** before modeling.
- Train and evaluate **linear regression models**.
- Interpret coefficients and residuals.

Dataset : Boston housing

Part A: Exploratory Data Analysis

1. Load the dataset (e.g., Boston Housing or any regression dataset).
2. Plot the distribution of the target variable. Comment on skewness.
3. Compute and visualize the correlation matrix (heatmap). Which features are most correlated with the target?

Part B: Simple Linear Regression

	1. Select **one predictor feature** (e.g., "average number of rooms" in housing dataset). 2. Formulate the hypothesis equation: $y=\beta 0+\beta 1x+\epsilon$ 3. Fit a simple linear regression model. Report: o Intercept ($\beta 0$) o Slope ($\beta 1$) 4. Plot scatter plot of feature vs target with the regression line. 5. Interpret the slope: What does a unit increase in x mean for y? **Part C: Multiple Linear Regression** 1. Select at least **3 predictor variables**. 2. Fit a multiple linear regression model. Write down the fitted equation. 3. Report model coefficients and interpret any one of them. 4. Evaluate the model using: o R^2 score o Mean Squared Error (MSE) o Root Mean Squared Error (RMSE) 5. Compare performance with the simple regression model. Which one is better? Why?