Assignment 4

Name: Dhananjay Siddhanath Jagtap

Roll No: 282057

Batch: B3

# Statement:

Q. Apply an appropriate machine learning algorithm on a dataset collected from a cosmetics shop showing details of customers, to predict customer response to a special offer.  
Tasks to perform:  
a) Preprocess the dataset – handle missing values, encode categorical variables, and normalize data  
b) Apply logistic regression  
c) Generate a confusion matrix  
d) Calculate accuracy, precision, recall, and F1-score  
e) Visualize model performance

# Objective:

1. To apply supervised machine learning (Logistic Regression) to a real-world dataset.  
2. To predict whether a customer will respond to a special offer based on their demographic and purchase behavior.  
3. To evaluate model performance using classification metrics.  
4. To practice data preprocessing, feature selection, and data visualization using pandas, matplotlib, seaborn, and scikit-learn.

# Resources Used:

Software: Jupyter Notebook

Libraries: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn

# Introduction to Logistic Regression:

Logistic Regression is a classification algorithm used to predict binary outcomes (e.g., yes/no, 0/1, true/false). It models the probability that a given input belongs to a particular category.

# Methodology:

1. Dataset Used:  
\*marketing\_campaign.csv\* — customer data from a cosmetics shop, including demographics, product purchases, and marketing responses.  
  
2. Data Preprocessing:  
- Removed rows with missing values  
- Converted the Dt\_Customer column to datetime  
- Encoded categorical columns (Education, Marital\_Status) using One-Hot Encoding  
- Scaled numerical features using StandardScaler  
  
3. Feature Selection:  
Selected relevant features like: Income, Kidhome, Teenhome, Recency, MntWines, NumWebPurchases, etc.  
Target variable: Response (1 if customer responded to offer, else 0)  
  
4. Model Building:  
Applied Logistic Regression using scikit-learn's LogisticRegression with max\_iter=1000  
  
5. Model Evaluation:  
- Predicted on the test set  
- Evaluated using:  
 \* Confusion Matrix  
 \* Accuracy  
 \* Precision  
 \* Recall  
 \* F1-Score  
  
6. Visualization:  
- Confusion matrix was visualized using seaborn.heatmap

# Sample Output:

Confusion Matrix:

[[372 5]

[ 53 14]]

Accuracy: 0.87

Precision: 0.74

Recall: 0.21

F1 Score: 0.33

# Advantages:

1. Logistic Regression is simple and interpretable.  
2. It works well when the relationship between features and target is linear.  
3. Efficient for binary classification problems.

# Disadvantages:

1. May underperform on complex or non-linear problems.  
2. Sensitive to multicollinearity among features.  
3. Requires data to be preprocessed properly (e.g., scaled and encoded).

# Conclusion:

This assignment introduced the application of Logistic Regression to predict marketing response. Through real-world data preprocessing, feature selection, model training, and performance evaluation, we learned how a simple ML model can provide insights into customer behavior. While the model’s performance was limited by class imbalance, it demonstrated the practical application of ML in marketing analytics. This forms a strong base for future improvements using more advanced algorithms.