Practical Assignment No. 5	
Title:	**Binary Classification Model**
Problem Statement:	Demonstrate binary classification model using Logistic Regression for Rain in Australia dataset from the Kaggle repository.
Objective:	To understand application development using techniques of data science and machine learning.
Outcome:	CO606.3: Apply machine learning techniques to develop real world applications.
Software or Hardware Requirements:	Anaconda/Java/GCC
Theory:	**Binary Classification** Binary classification is a type of supervised learning problem where the output variable has only two possible outcomes, typically represented as 0 and 1. Examples include: - Spam vs. Non-spam emails - Disease vs. No disease - Rain vs. No rain In this experiment, the goal is to classify whether it will rain tomorrow (“Yes” or “No”) using the historical weather data from the Rain in Australia dataset available on Kaggle. **Logistic Regression Overview** **Binary Logistic Regression** Logistic Regression is a statistical and machine learning technique used for predicting the probability of a binary outcome. Unlike linear regression, which predicts continuous values, logistic regression predicts probabilities that are mapped to two possible classes using a sigmoid (logistic) function. The logistic (sigmoid) function is defined as:

$$P(y=1|x)=\frac{1}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n)}}$$

Where,

- $P(y=1|x)$ = probability of the positive class (RainTomorrow = Yes)
- $\beta_0,\beta_1,\ldots,\beta_n$ = model coefficients
- $x_1,x_2,\ldots,x_n$ = input features

The model makes predictions based on a decision boundary:

$$\text{If } P(y=1|x)\geq 0.5 \text{ then } y=1,\ \text{else } y=0$$

Multinomial Logistic Regression

- **Definition:**
  Used when the dependent variable has more than two categories that are not ordered.
- **Example:**
  - **Classifying types of fruits:** ***apple, orange, banana***
  - **Predicting the mode of transport:** ***bus, car, bike***
- **Model:**
  One category is treated as a reference class, and separate logistic equations are estimated for the others.

Ordinal Logistic Regression

- **Definition:**
  Used when the dependent variable has more than two categories that are ordered or ranked.
- **Example:**
  - **Customer satisfaction:** ***poor, average, good, excellent***
  - **Movie ratings:** ***1 star, 2 stars, 3 stars, 4 stars, 5 stars***
- **Model:**
  Assumes that there is an underlying continuous variable

determining the ordered response.
It uses the proportional odds model or cumulative logit model.

Data Preprocessing Steps

Before building the logistic regression model, the dataset must be preprocessed:

1. **Handling Missing Values:**
   Replace or drop missing data using methods like mean/median imputation or deletion.

2. **Encoding Categorical Variables:**
   Convert categorical columns like RainToday and RainTomorrow into numerical form (Yes → 1, No → 0).

3. **Feature Selection:**
   Select important features such as humidity, rainfall, temperature, wind speed, etc.

4. **Data Normalization/Standardization:**
   Scale numeric values to improve convergence during model training.

5. **Train-Test Split:**
   Divide dataset into training and testing sets (e.g., 80% train and 20% test).

Model Building and Training

- Use **Logistic Regression** from the sklearn.linear_model library in Python.

- Fit the model on training data (X_train, y_train).

- The model learns coefficients (βi\beta_iβi) that best separate the two classes.

Model Evaluation

After training, predictions are made on the test data and evaluated using metrics such as:

- **Accuracy:**

	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$ **Precision, Recall, and F1-Score:** To measure the quality of predictions. **Confusion Matrix:** Shows true vs. predicted classifications. **ROC Curve and AUC:** Assess the model’s discriminative ability between classes. **Visualization** Visualizations help understand model performance: ● Confusion matrix heatmap ● ROC curve ● Feature importance (using model coefficients)
Input/Datasets/Test Cases:	(write details from UCI repository) Dataset- Name of the Dataset: Description of the Dataset: Dataset Characteristics: Subject Area: Associated Tasks: Feature Type: # Instances: # Features:
Results:	Execute code for Logistic Regression to perform binary classification. Take a print of this code with output for submission as part of results.
Analysis and conclusion:	Write your own analysis of output and conclusion(Minimum 1 statement of Analysis, Minimum 1 Statement Conclusion)
References:	Reference /Links(min Any 2, include dataset ref.). Write references in IEEE format.