1. Introduction

Logistic Regression is a supervised machine learning algorithm used for binary classification problems. It estimates the probability that a given input belongs to a particular category.

In this project, we use logistic regression to predict whether it will rain tomorrow in Australia based on historical weather features such as temperature, humidity, and rainfall.

2. Problem Statement

To build a binary classification model that predicts the value of RainTomorrow (Yes/No) using historical weather data.

3. Understanding the Dataset

- Source: WeatherAUS dataset from Kaggle
- Total Records: ~145,000
- Target Variable: RainTomorrow (Yes/No)
- Features Used:
  - MinTemp: Minimum temperature
  - MaxTemp: Maximum temperature
  - Rainfall: Amount of rainfall
  - Humidity3pm: Humidity at 3 PM
  - RainToday: Whether it rained today (Yes/No)

4. Data Preprocessing

- Dropped rows with missing values in selected features and target
- Encoded binary categorical features (RainToday and RainTomorrow) using LabelEncoder
- Standardized numerical features using StandardScaler

5. Exploratory Data Analysis (EDA)

- Countplot: Checked class distribution of the target variable RainTomorrow
- Boxplot: Observed humidity and rainfall distributions across the target classes
- Pairplot: Visualized relationships between numerical features and class labels
- Insights:
  - High humidity at 3pm strongly correlates with likelihood of rain
  - RainToday is a strong indicator for predicting RainTomorrow

6. Model Building

- Model Used: LogisticRegression from scikit-learn
- Train/Test Split: 80% training, 20% testing
- Max Iterations: 1000
- Scaled Data: Applied StandardScaler before training

7. Model Evaluation

- Accuracy Score: ~84%
- Confusion Matrix: Showed balanced classification between rain and no-rain
- Classification Report:
  - Provided precision, recall, F1-score for each class
  - Demonstrated good balance in prediction quality

8. Feature Importance

- Analyzed model coefficients to determine the influence of each feature
- Top contributing features:
  - Humidity3pm: Strongest positive correlation with rain
  - RainToday: Immediate predictor
  - Rainfall: Moderate positive impact

9. Sample Prediction

Predicted rain based on the following input:

[MinTemp=15.0, MaxTemp=25.0, Rainfall=2.5, Humidity3pm=70.0, RainToday=1]

Prediction: RainTomorrow = 1 -> It will rain tomorrow.

10. Conclusion

- Logistic Regression is a powerful and interpretable algorithm for binary classification
- Our model performed well with a relatively high accuracy (~84%)
- Advantages: Simplicity, speed, explainability
- Future Work:
  - Explore ensemble methods like Random Forest, XGBoost
  - Include more features like wind speed, pressure, evaporation
  - Handle class imbalance with oversampling or SMOTE