Weather in AUS project

Name: Kirlos Ayad Libeb
Course: AI&Data science - Machine Learning Engineer

GitHub Repository:
https://github.com/kirlosAyad/Weather-AUS-Logistic-Regression.git

---

1. Project Overview

This project aims to build a Logistic Regression model to predict whether it will rain tomorrow (RainTomorrow) using the Weather Australia dataset.

The project follows a complete Machine Learning workflow including:

- Exploratory Data Analysis (EDA)
- Data Cleaning
- Feature Engineering
- Encoding
- Feature Scaling
- Model Training
- Model Evaluation using Classification Metrics

2. Dataset Description

The dataset contains daily weather observations from multiple locations in Australia.
Some important features include:

- MinTemp
- MaxTemp
- Rainfall
- WindGustSpeed
- Humidity9am
- Humidity3pm
- Pressure9am
- Pressure3pm
- RainToday
- RainTomorrow (Target Variable)

The target variable RainTomorrow is binary:

- 1 → Rain
- 0 → No Rain

3. Data Preprocessing

3.1 Handling Missing Values

- Columns with excessive missing values were removed.
- Numerical features were filled using the median.
- Categorical features were filled using the mode.

3.2 Feature Engineering

The Date column was converted into datetime format and new features were extracted:

- Year
- Month
- Day

The original Date column was then removed.

3.3 Encoding

- RainTomorrow was converted to binary values (Yes = 1, No = 0).
- RainToday was converted to binary values.
- Other categorical features were encoded using One-Hot Encoding.

3.4 Feature Scaling

Standardization (Z-score scaling) was applied to ensure that all features are on the same scale:

X_scaled = (X - mean) / standard_deviation

This improves model convergence and stability.

4. Logistic Regression Model

Logistic Regression is a classification algorithm that predicts probabilities using the Sigmoid function:

sigmoid(z) = 1 / (1 + e^(-z))

Where:
z = w0 + w1x1 + w2x2 + ... + wnxn

The output is a probability between 0 and 1.

If probability ≥ 0.5 → Predict Rain
If probability < 0.5 → Predict No Rain

5. Model Evaluation

The model was evaluated using several classification metrics:

5.1 Confusion Matrix

The confusion matrix shows:

- True Positives (TP)
- True Negatives (TN)
- False Positives (FP)
- False Negatives (FN)

5.2 Accuracy

Accuracy measures the percentage of correct predictions out of total predictions.

Accuracy = (Correct Predictions) / (Total Predictions)

5.3 Precision

Precision measures how many predicted positive cases were actually correct.

Precision = TP / (TP + FP)

5.4 Recall

Recall measures how many actual positive cases were correctly identified.

Recall = TP / (TP + FN)

5.5 F1-Score

F1-Score is the harmonic mean of Precision and Recall.

F1 = 2 × (Precision × Recall) / (Precision + Recall)

5.6 ROC Curve and AUC

The ROC Curve plots:

- True Positive Rate (TPR)
- False Positive Rate (FPR)

The AUC (Area Under the Curve) measures the model's ability to distinguish between classes.

A higher AUC indicates better classification performance.

6. Conclusion

In this project:

- A complete Machine Learning pipeline was implemented.
- Logistic Regression was applied to predict rainfall.
- The model was evaluated using multiple classification metrics.
- ROC-AUC was used to assess model performance.

The results demonstrate that Logistic Regression is effective for binary classification problems such as rainfall prediction.