Rainfall Predictor for Pakistan: A Machine Learning Approach

Maryam Khan(22F-3238), Muhammad Awais(22F-3233), Mujahid(22F-3215)

Course: Machine Learning Instructor: Shehbaz Ayaz

Date: May 11, 2025

**Abstract**

This project develops a machine learning model to predict rainfall in Pakistan, address- ing needs in agriculture and disaster preparedness. Using the weatherAUS.csv dataset, we conducted exploratory data analysis, preprocessed the data, and trained four models: Lo- gistic Regression, Random Forest, XGBoost, and a Neural Network. After hyperparameter tuning, XGBoost achieved the highest performance with an AUC of 0.89 and recall of 0.77. The model was deployed as a web application on Hugging Face. Limitations include the use of Australian data, suggesting future work with Pakistan-specific datasets.

# Introduction

Rainfall prediction is vital for Pakistan, where agriculture drives the economy and monsoons pose flood risks. Accurate forecasts support farming decisions and disaster mitigation. This project builds a rainfall prediction model using the weatherAUS.csv dataset from Kaggle, containing 145,460 rows and 23 features (e.g., temperature, humidity, rainfall) from Australian stations. Despite the mismatch, the dataset’s climate similarities with Pakistan make it a starting point.

Objectives include:

* Analyze data characteristics through EDA.
* Preprocess and engineer features for modeling.
* Compare multiple machine learning models.
* Deploy the best model on Hugging Face.

# Dataset Exploration

The dataset comprises 145,460 observations with 23 features, including MinTemp, MaxTemp, Rainfall, and RainTomorrow (Yes/No). Missing values were notable in Sunshine (48%) and Cloud9am (38%).

EDA revealed a 77% "No" and 23% "Yes" imbalance, skewed Rainfall and WindGustSpeed, and high correlation (0.96) between Pressure9am and Pressure3pm.

# Methodology

## Preprocessing

Missing values were imputed with medians for numerical features (e.g., Sunshine) and "Missing" for categorical features (e.g., WindGustDir). Outliers in Humidity9am and Humidity3pm were clipped to [1, 100]. Features were normalized with StandardScaler. Engineered features included TempDiff and HumidityDiff. Categorical variables were one-hot encoded, yielding 111 features. Class imbalance was addressed with weights (balanced or scale\_pos\_weight).

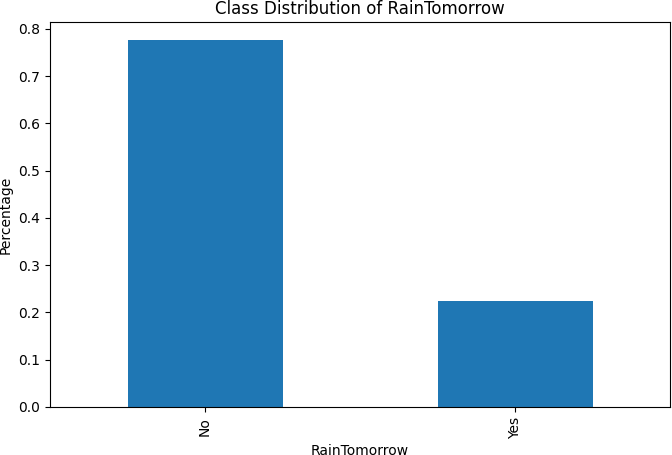


Figure 1: Class Distribution of RainTomorrow (77% "No," 23% "Yes").

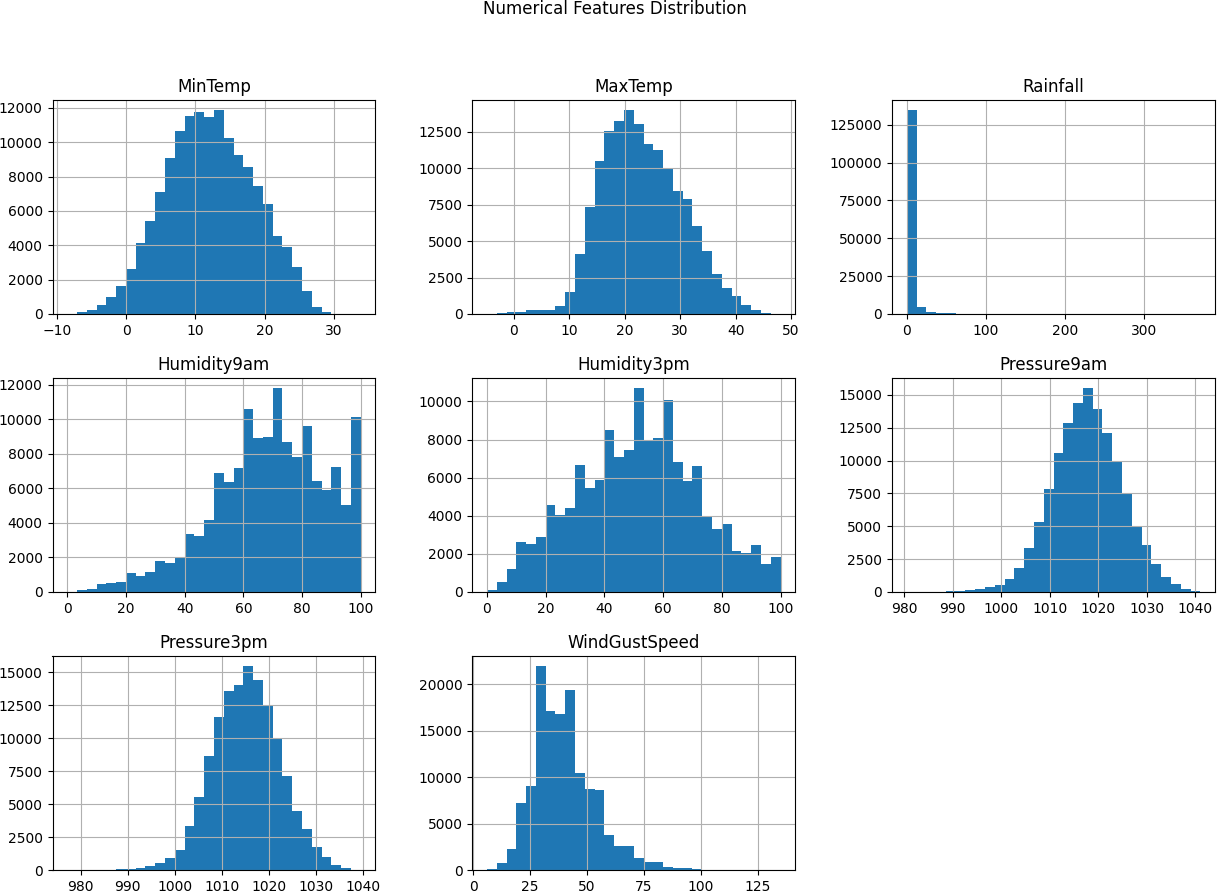


Figure 2: Distribution of Numerical Features (e.g., skewed Rainfall, normal Pressure9am).

## Model Development

Four models were implemented:

* + - **Logistic Regression**: Baseline for interpretability.
    - **Random Forest**: Ensemble for non-linear patterns.
    - **XGBoost**: Gradient boosting for complex relationships.
    - **Neural Network**: Three-layer model with dropout (0.2) and early stopping.

## Hyperparameter Tuning

XGBoost was tuned with Grid Search (5-fold CV) on max\_depth [3, 5, 7], learning\_rate [0.01, 0.1, 0.3], subsample [0.7, 0.9], and colsample\_bytree [0.7, 0.9], optimizing recall.

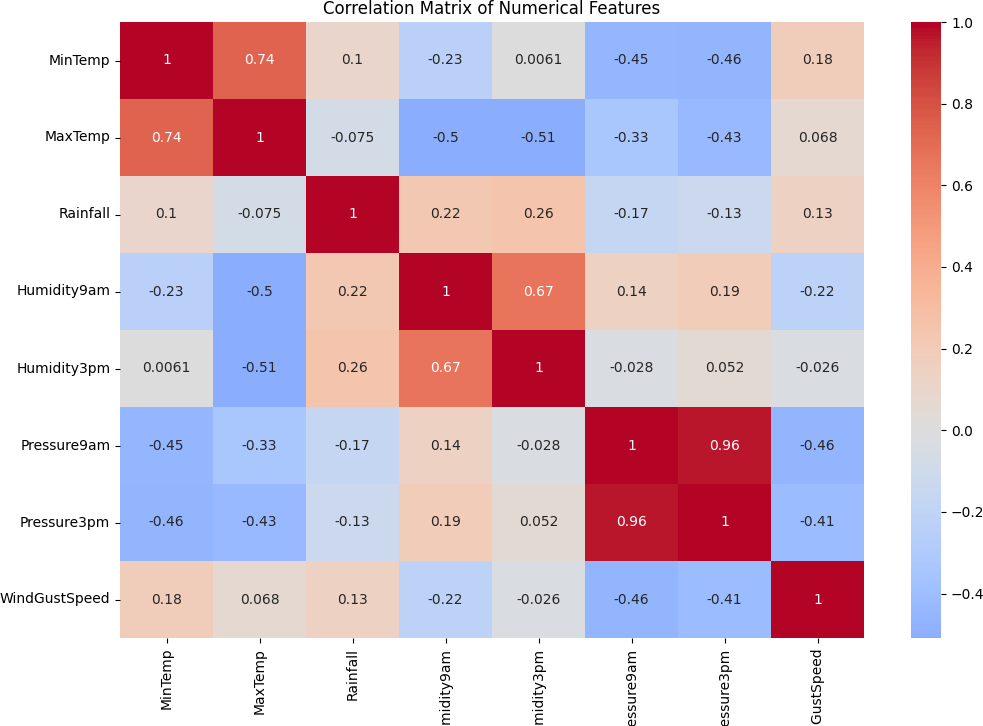
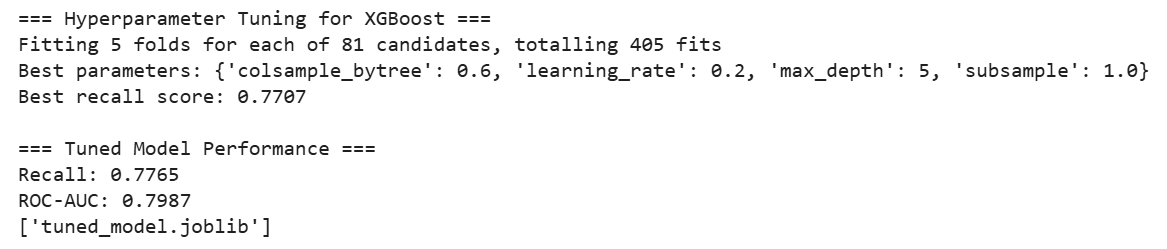


Figure 3: Correlation Matrix (e.g., 0.96 between Pressure9am and Pressure3pm).



## Evaluation Metrics

Metrics included:

* + - **Accuracy**, **Precision**, **Recall** (prioritized for imbalance).
    - **F1 Score**, **ROC-AUC**.

Confusion matrices and ROC curves were analyzed.

# Results and Discussion

## Model Performance

Table [1](#_bookmark0) summarizes performance. XGBoost led with AUC 0.89 and recall 0.77.

Table 1: Model Performance Comparison

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 Score | ROC-AUC |
| Logistic Regression | 0.81 | 0.52 | 0.77 | 0.62 | 0.86 |
| Random Forest | 0.84 | 0.78 | 0.47 | 0.59 | 0.88 |
| XGBoost | 0.82 | 0.57 | 0.77 | 0.65 | 0.89 |
| Neural Network | 0.80 | 0.51 | 0.62 | 0.56 | 0.85 |

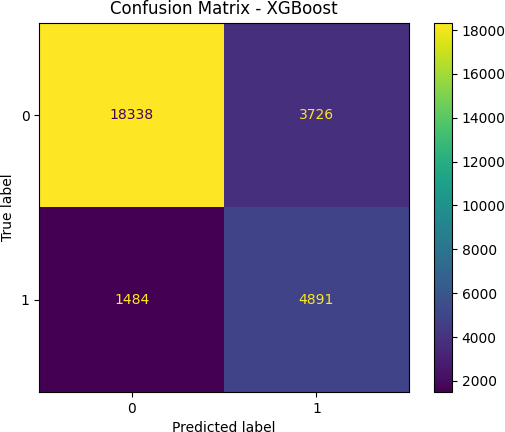


Figure 5: Confusion Matrix for XGBoost (TP = 4,891, FN = 1,484).

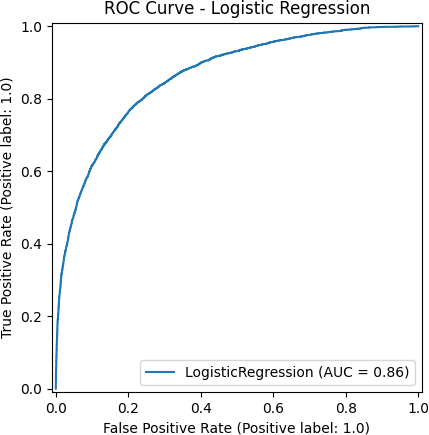
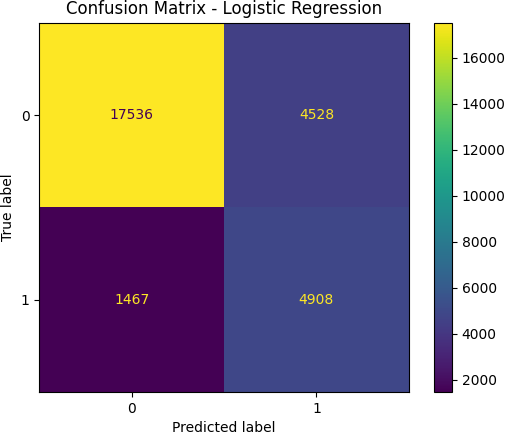
 

Figure 6: ROC Curve (AUC = 0.86) and Confusion Matrix for Logistic Regression.

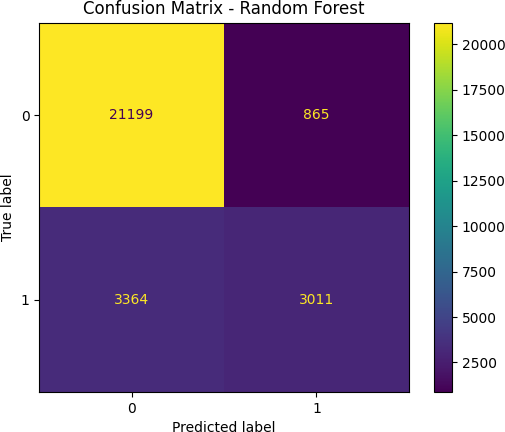
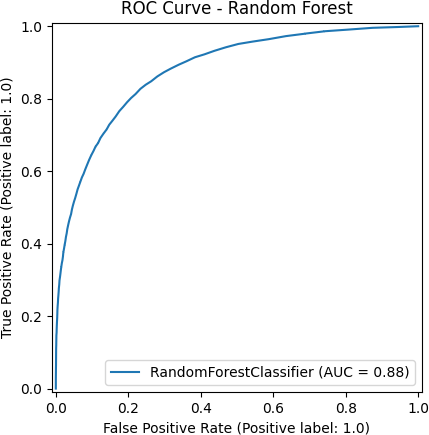


Figure 7: ROC Curve (AUC = 0.88) and Confusion Matrix for Random Forest.

## Feature Importance

XGBoost identified Humidity3pm, TempDiff, and Location\_Brisbane as key predictors.

## Discussion

XGBoost’s high recall (0.77) suits rainfall prediction, where false negatives are critical. The Australian dataset limits relevance to Pakistan. The Neural Network’s lower recall (0.62) may indicate overfitting.

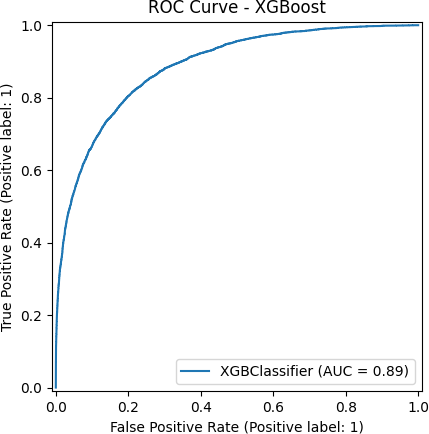


Figure 4: ROC Curve for XGBoost (AUC = 0.89).

# Deployment

The XGBoost model is deployed on Hugging Face as "Rainfall Predictor for Pakistan." Users input weather data to predict rain tomorrow. Limitations include manual input and dataset mismatch.

1. Visit the app link [App](https://huggingface.co/spaces/MAwaisM/Rainfall_Predictor)
2. Enter features (e.g., temperature, humidity).
3. Get a "Yes" or "No" prediction.

# Conclusion and Future Work

XGBoost achieved the best performance (AUC 0.89, recall 0.77), deployed on Hugging Face. The Australian dataset’s mismatch is a key limitation. Future work includes:

* Collecting Pakistan-specific data.
* Applying SMOTE for imbalance.
* Exploring ensemble stacking.

# References

* Kaggle. weatherAUS.csv dataset. [https://www.kaggle.com/datasets/jsphyg/weather-dataset-ratt](https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package)
* Scikit-learn. [https://scikit-learn.org](https://scikit-learn.org/).
* XGBoost. [https://xgboost.readthedocs.io](https://xgboost.readthedocs.io/).
* Keras. [https://keras.io](https://keras.io/)