College of Professional Studies: Northeastern University

ALY 6020 – Predictive Analytics

Instructor: Dr. Shahram Sattar

Academic Term: Winter 2025

Rain Prediction in Australia

Predicting Tomorrow's Rain using Weather Data

Submitted By:

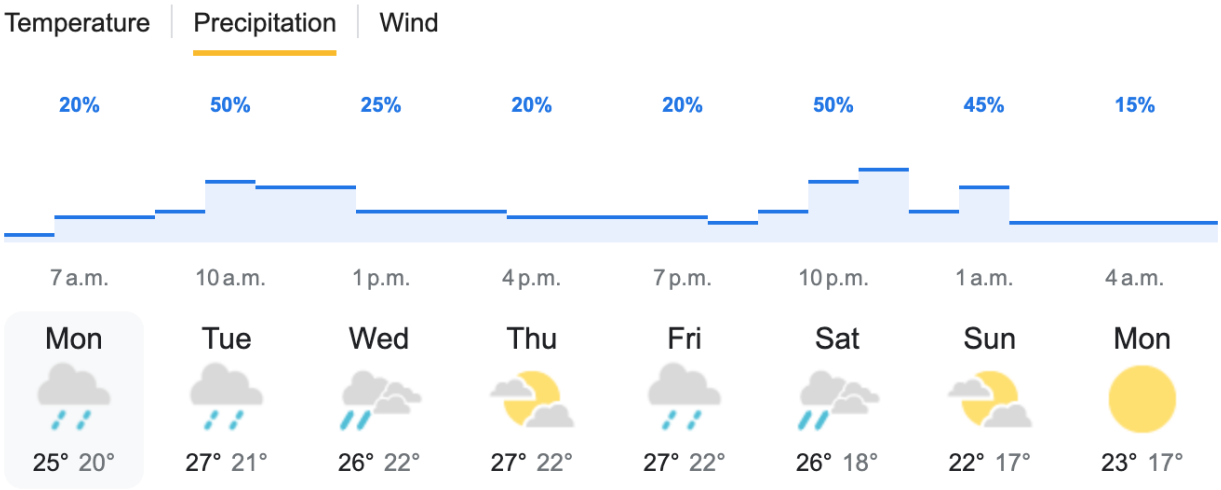
Group 6

Akanksha Chapra
Muhammad Hassan Zahoor
Rahul Hiteshkumar Prajapati
Saif Wasay Mohammed
Sheila Kwartemaa Boateng

February 9, 2025

Introduction

Weather, particularly rainfall, plays a critical role in shaping various aspects of life, from agricultural productivity to disaster preparedness. In Australia, a country known for its climatic extremes, weather patterns vary significantly across different locations and times of the year, making accurate rainfall prediction essential. The ability to anticipate rainfall can help farmers optimize crop planning, assist emergency services in disaster response, and support infrastructure planning for extreme weather events. However, predicting rainfall remains a challenge due to the dynamic and complex nature of meteorological conditions.



Recent years have seen an increase in extreme weather events such as floods, droughts, and heatwaves, disrupting farming operations, damaging crops, and affecting food security. This highlights the importance of accurately predicting rainfall to enable better planning and mitigation strategies. The Rain in Australia dataset on Kaggle provides a comprehensive collection of meteorological observations from across the continent, making it a valuable resource for building predictive models.

This study explores the effectiveness of various machine learning models—including Gradient Boosting, Support Vector Machines (SVM), Logistic Regression, Decision Trees, and Random Forests in forecasting rainfall based on key weather parameters. By evaluating these models using accuracy, precision, recall, and ROC-AUC scores, we aim to determine the best approach for minimizing false negatives—cases where rain is not predicted but occurs—which is crucial for timely preparedness and risk reduction.

Business Questions

The primary objective of this study is to develop an accurate machine learning model for predicting rainfall in Australia.

Other Objective

In addition to predicting rainfall, this study aims to:

- Identify key meteorological factors that contribute to rainfall prediction.
- Compare different machine learning models (e.g., Decision Tree, Random Forest, SVM, Gradient Boosting) based on performance metrics such as accuracy, precision, recall, and ROC-AUC.
- Minimize false negatives (i.e., cases where rain is not predicted but occurs) to enhance preparedness and risk mitigation.

Data Overview

Your dataset consists of 145,460 rows and 23 columns, capturing timely rainfall data from 2007 to 2017 across different locations in Australia. Each row represents weather measurements for a specific date and location, with the target variable being "RainTomorrow", which indicates whether it will rain the next day. The variables includes;

Temperature Variables: MinTemp, MaxTemp, Temp9am, Temp3pm

Rainfall & Atmospheric Conditions: Rainfall, Evaporation, Sunshine

Wind Features: WindGustDir, WindGustSpeed, WindDir9am, WindDir3pm, WindSpeed9am, WindSpeed3pm

Humidity & Pressure: Humidity9am, Humidity3pm, Pressure9am, Pressure3pm

Cloud Coverage: Cloud9am, Cloud3pm

Rain Indicators: RainToday, RainTomorrow (Target variable)

Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrow
2008-09-21	Melbourne	6.5	19.8	0.4	4.2	10.6	3.0	13.0	19.4	No	No
2009-07-06	Sale	4.9	13.0	0.0	2.0	6.8	6.0	8.6	11.7	No	No
2010-11-20	GoldCoast	18.8	26.4	2.0	NaN	NaN	NaN	24.0	22.1	Yes	No
2010-11-22	PearceRAAF	19.4	27.4	1.8	NaN	10.7	3.0	24.4	25.8	Yes	No
2012-04-26	Nuriootpa	5.1	16.6	0.0	1.4	1.4	7.0	12.1	15.7	No	No
2013-07-06	Sydney	7.8	17.4	0.0	4.2	9.8	0.0	10.2	17.1	No	No
2014-04-22	Perth	7.7	23.7	0.0	4.0	10.5	1.0	16.7	21.8	No	No
2014-06-08	Wollongong	11.1	16.8	0.0	NaN	NaN	1.0	14.0	15.9	No	No
2016-04-13	Sale	10.8	19.0	0.0	NaN	NaN	1.0	16.1	18.1	No	No
2017-04-11	Albany	13.0	NaN	0.0	4.0	NaN	NaN	17.8	NaN	No	NaN

Data Preparation

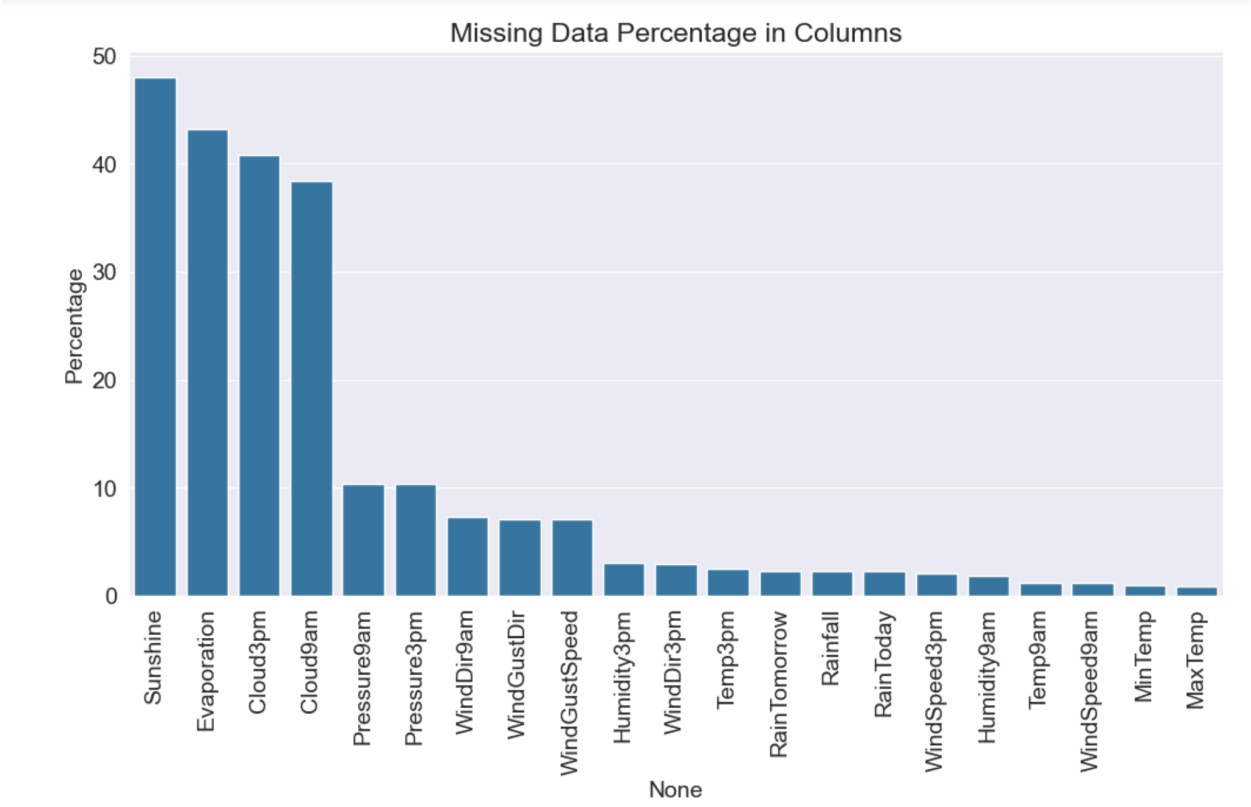
Before data exploratory and other analysis, we converted the Date column to the correct datetime format to ensure proper time-based analysis. And also ensure all other variables were in correct data format

Missing Values

Several variables in the dataset, including **Evaporation, Sunshine, and Cloud Cover**, have a significant number of missing values. To ensure data quality and model reliability, the following strategies were applied:

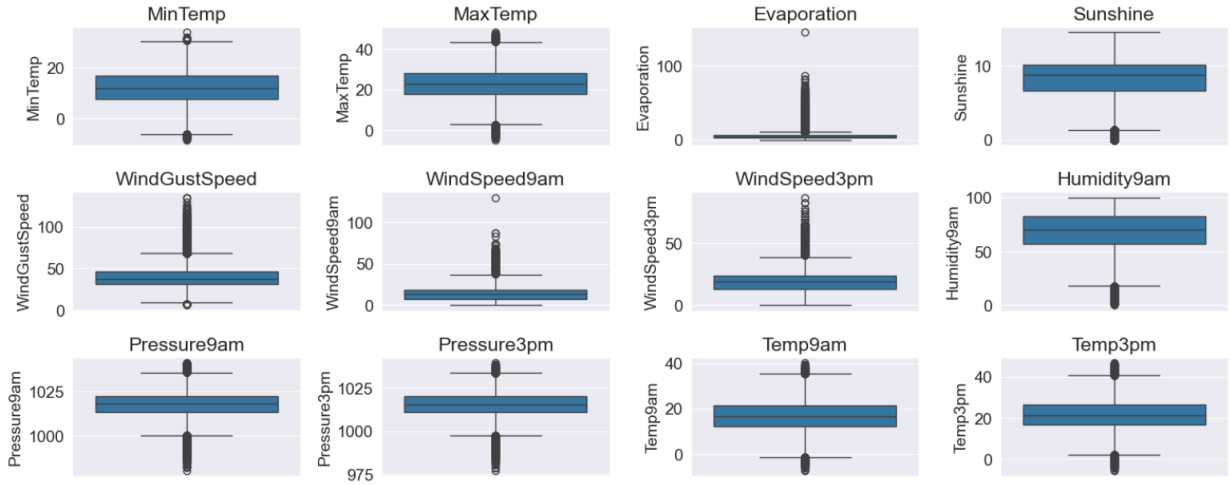
- **Imputation:**
  - **Numerical Variables** (e.g., **Evaporation, Sunshine, Cloud Cover**) were imputed using **KNN imputation**, which estimates missing values based on similar data points.
  - **Categorical Variables** (e.g., **WindGustDir, WindDir9am, WindDir3pm**) were filled using **mode imputation**, where the most frequent category was used to replace missing values.

No variable had **50% or more** missing values, so **all features were retained** in the dataset.



Outliers

The box plots reveal significant outliers in several meteorological variables within the dataset. Evaporation, WindGustSpeed, WindSpeed9am, and WindSpeed3pm all exhibit numerous high outliers, suggesting occasional extreme weather events. MinTemp, MaxTemp, Sunshine, Humidity9am, Pressure9am, Pressure3pm, Temp9am, and Temp3pm show fewer outliers, typically corresponding to extreme temperature, pressure, or dry conditions. To address these outliers, we applied a log transformation. Additionally, we removed extreme values where evaporation reached 100 or WindSpeed9am exceeded 100 to ensure more reliable data analysis.



Feature Engineering

1. **Date Features Extraction**
   We extracted the **year**, **month**, and **day of the year** from the **Date** column to capture seasonal patterns in rainfall.
2. **Location Categorization**
   We categorized the locations based on Australian states to visualize how rainfall patterns differ across regions and states.
3. **Wind Direction Conversion**
   We converted categorical wind directions into numerical angles using the following function:

```
def wind_dir_to_angle(direction):
    dir_to_angle = {
        'N': 0, 'NNE': 22.5, 'NE': 45, 'ENE': 67.5,
        'E': 90, 'ESE': 112.5, 'SE': 135, 'SSE': 157.5,
        'S': 180, 'SSW': 202.5, 'SW': 225, 'WSW': 247.5,
        'W': 270, 'WNW': 292.5, 'NW': 315, 'NNW': 337.5
    }
```

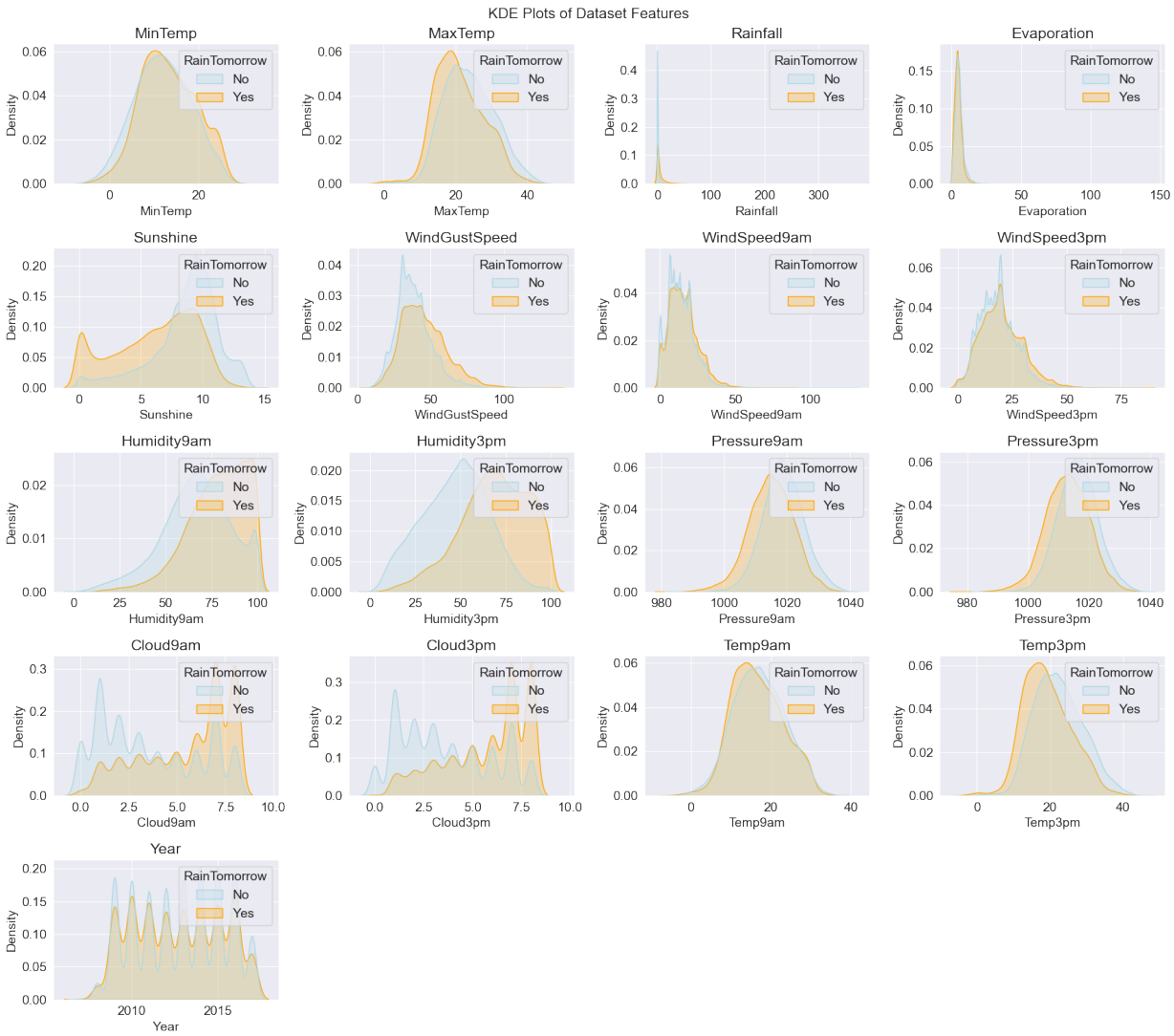
Additionally, we created new features for **WindDirChange**, **MonthSin**, **MonthCos**, **DayOfYearSin**, and **DayOfYearCos** to capture cyclical patterns in the data.

Data Exploratory

Feature Distributions - KDE (Kernel Density Estimation)

The KDE plots show how features are distributed, between RainTommorrow Yes vs No. Upon examining the "Rain in Australia" dataset, these key observations can be made

1. **Extreme Outliers in Rainfall and Evaporation**
   - The dataset contains extreme outliers, particularly in the variables **Rainfall** and **Evaporation**. These outliers may represent rare weather events or data anomalies and could significantly influence the statistical results and performance of machine learning models. Therefore, it is essential to decide whether to treat these outliers by capping or removing them.
2. **MinTemp, MaxTemp, and Temp3pm**:
   - Follow a near-normal distribution.Rainfall, Evaporation, and WindGustSpeed: Are heavily right-skewed.Humidity at 9 AM & 3 PM: Displays a bimodal distribution, indicating different weather conditions.
3. **Distinctive Distribution Patterns for Rain vs. No Rain**
   - **Density plots** reveal notable differences between variables when comparing rainy days to non-rainy days. The distribution patterns for certain variables, such as **Cloud Cover**, **Temperature**, **Humidity**, **Pressure**, **Sunshine**, and **Wind Gust Speed**, differ significantly when it rains versus when it does not. Some days have low humidity (clear days), while others have high humidity (rainy days). These variables exhibit clear patterns that are important for accurate rainfall prediction.



And these observations are supported by the table below.

1. Cloud Cover: Significantly higher cloud coverage on rainy days (Cloud3pm: 5.72 vs 3.60; Cloud9am: 5.49 vs 3.54).
2. Humidity: Notably higher humidity levels, especially in the afternoon (Humidity3pm: 68.50% vs 46.81%; Humidity9am: 77.97% vs 66.42%).
3. Temperature: Lower maximum and afternoon temperatures on rainy days (MaxTemp: 21.12°C vs 23.81°C; Temp3pm: 19.23°C vs 22.38°C).
4. Pressure: Lower atmospheric pressure associated with rainy days (Pressure3pm: 1012.70 hPa vs 1016.16 hPa; Pressure9am: 1014.77 hPa vs 1018.60 hPa).
5. Wind: Higher wind speeds, particularly gust speeds, on rainy days (WindGustSpeed: 45.21 km/h vs 38.22 km/h).
6. Sunshine: Fewer hours of sunshine recorded on days before rainfall (6.08 hours vs 8.61 hours).
7. Previous Day's Rainfall: Higher rainfall on the previous day is linked to rain the next day (6.05 mm vs 1.31 mm).

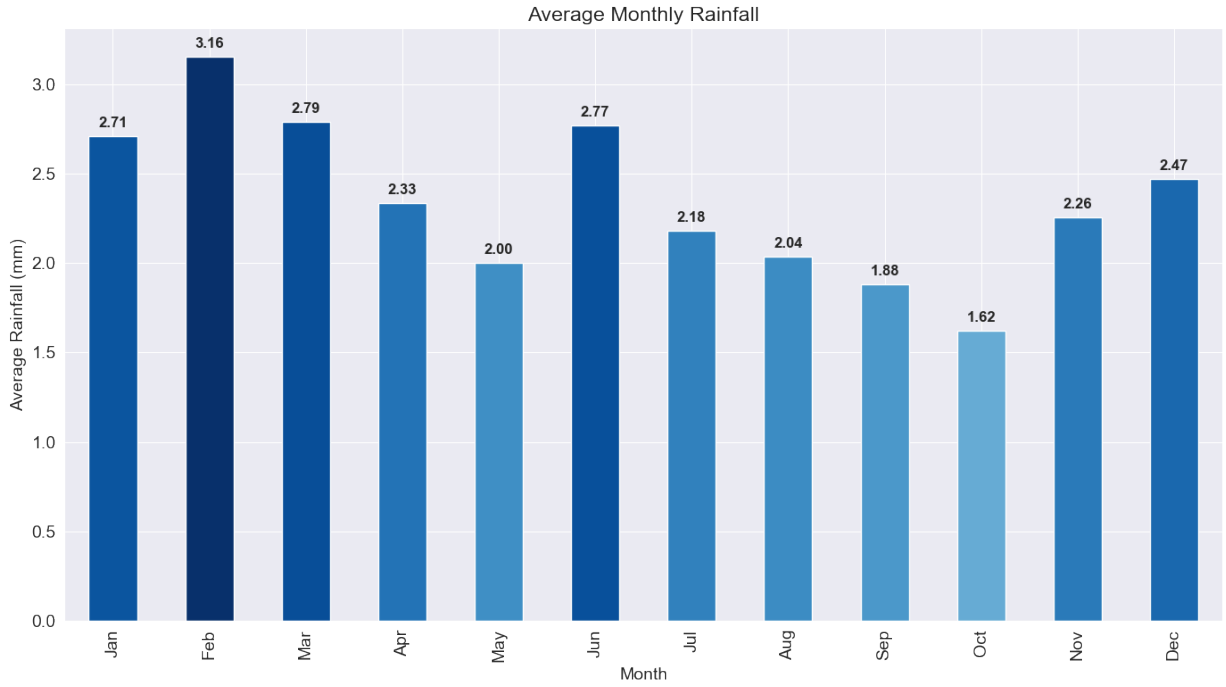
These observations align with the findings from the search results, particularly the importance of variables such as temperature, humidity, pressure, wind speed, and cloud cover in rainfall prediction models

The scatter plot in the appendix show that various numerical features correlate with Rainfall

- WindGustSpeed vs Rainfall.: Higher wind gust speeds tend to correlate with higher rainfall. **Strong win gusts** often accompany **storm systems**, which bring rain.
- Humidity vs. Rainfall: As humidity increases, rainfall tends to increase. o Temperature vs. Rainfall: There is no strong correlation.
- **Temperature vs. Rainfall:** There is no strong correlation. Rain occurs under **various temperature conditions**, so it isn't a strong independent predictor

Temporal Analysis

The average rainfall per month aligns with the typical rainfall patterns observed in Australia, which experiences significant variation in rainfall across different regions. Some months having **significantly higher average rainfall**. The higher values in February (3.16) and March (2.79) indicate a peak in rainfall during late summer and early autumn, a period associated with the wet season in northern Australia, particularly in regions like Darwin and Cairns, where monsoonal rains are common from November to April. In contrast, the drop in rainfall from May (1.99) to October (1.62) reflects the drier conditions typical of southern and central Australia during the winter and early spring months, when rainfall is generally lower outside the wet season. Interestingly, the slight increase in December (2.47) suggests the onset of the wet season in tropical areas, marking the transition to heavier rainfall as the monsoon season begins. These trends mirror the seasonal weather patterns in countries with similarly diverse climates, where regional variations influence the timing and intensity of rainfall throughout the year.



Rainfall Distribution by Australia States

When analyzing the rainfall pattern across Australian states by month, several key observations emerge:

Queensland (3.98 mm):

Queensland experiences the highest average rainfall among the regions. This is consistent with its tropical climate, which sees significant rainfall during the monsoon season (roughly November to April). Tropical regions like Cairns and Townsville are affected by heavy rains, especially in summer, which supports the higher rainfall value for Queensland.

New South Wales (2.72 mm):

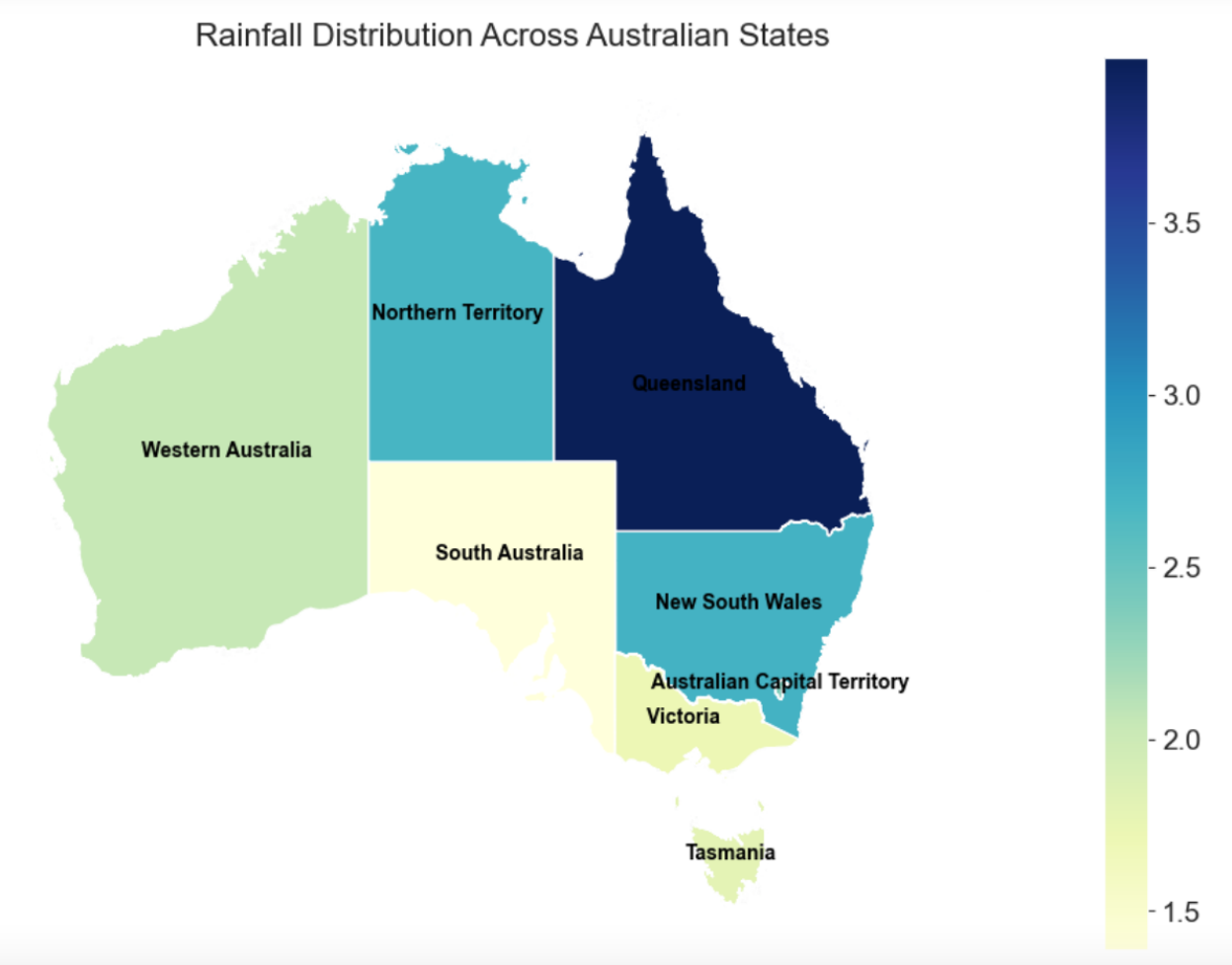
New South Wales falls into the mid-range for rainfall, with values higher than regions like Victoria and South Australia. The state's climate varies significantly, from the temperate coastal areas (Sydney, Newcastle) to the drier inland. The average rainfall here reflects the variability between these different areas, with coastal areas receiving more rainfall, especially during the summer.

Northern Territory (2.69 mm):

The Northern Territory, including cities like Darwin, experiences a monsoonal climate, with wet summers and dry winters. The high rainfall is consistent with the wet season from November to April, contributing to the average rainfall value.

Western Australia (2.05 mm):

Western Australia has a range of climates, from tropical in the north (Broome) to arid in the south (Perth). This region is heavily influenced by seasonal rainfall, with higher rainfall during the winter months in the south and dry conditions during summer. The value of 2.05 mm is likely reflective of these seasonal patterns and the combination of arid and tropical zones.



Australian Capital Territory (2.37 mm):

The ACT's rainfall is typical for its temperate climate. Canberra and surrounding areas receive moderate rainfall throughout the year, with a peak in spring and summer. The figure of 2.37 mm fits well with the state's known rainfall characteristics.

Tasmania (1.80 mm):

Tasmania, with its cooler, more temperate climate, receives consistent rainfall throughout the year. The figure of 1.80 mm is slightly lower than the mainland states but still reflects Tasmania's relatively high and consistent rainfall, especially on the west coast.

Victoria (1.71 mm):

Victoria is a state with a Mediterranean climate, characterized by relatively dry summers and wetter winters. The average rainfall figure of 1.71 mm is reflective of this, with the coastal regions, including Melbourne, receiving more rain than the inland areas.

South Australia (1.39 mm):

South Australia has the lowest rainfall, particularly in the interior regions, which are classified as semi-arid or arid. The value of 1.39 mm is consistent with the generally dry conditions across the state, especially during the summer months when much of the state receives little rainfall.

To better understand these diverse rainfall patterns, it's important to note that the majority of rainfall occurs from November to March. When we analyzed the rainfall patterns across Australia's states by month, we identified that the regions (states) experiencing the highest rainfall are concentrated in the months when rainfall is most frequent.

Tropical regions (like Queensland and the Northern Territory) naturally have high rainfall during the wet season (January to March and a bit in April), with rainfall levels peaking in these months due to the monsoon seasons and humid conditions and drops after the wet season. November and December also show considerable rainfall, as the wet season extends into the later months of the year.Monsoom seasons starts from Novemeber and ends in March

Coastal regions (like New South Wales, and Tasmania) typically see more consistent, moderate rainfall, with summer(December-February) and autumn(March-May) being wetter.

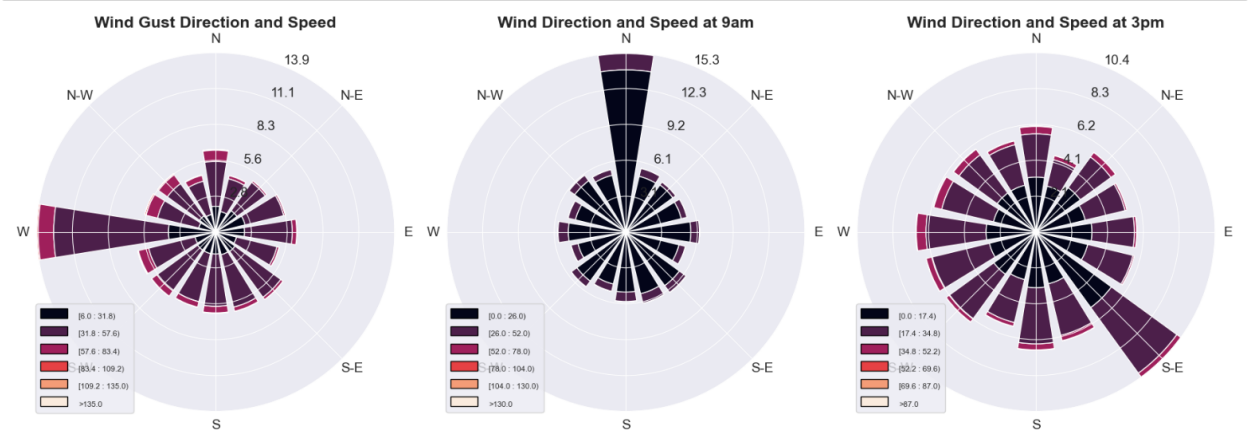
Desert and semi-arid regions (like South Australia and the central parts of Western Australia) have much lower rainfall, especially during the summer months. Rainfall tends to peak in the autumn and winter months (April to July) especially for Western Australia

The data suggests that the **Australian Capital Territory** follows a typical temperate climate pattern with moderate rainfall during the summer months and drier conditions in winter. The region does not show signs of monsoon rainfall, which is characteristic of tropical climates. Instead, it experiences sporadic rainfall throughout the year, with a slight peak during the summer.

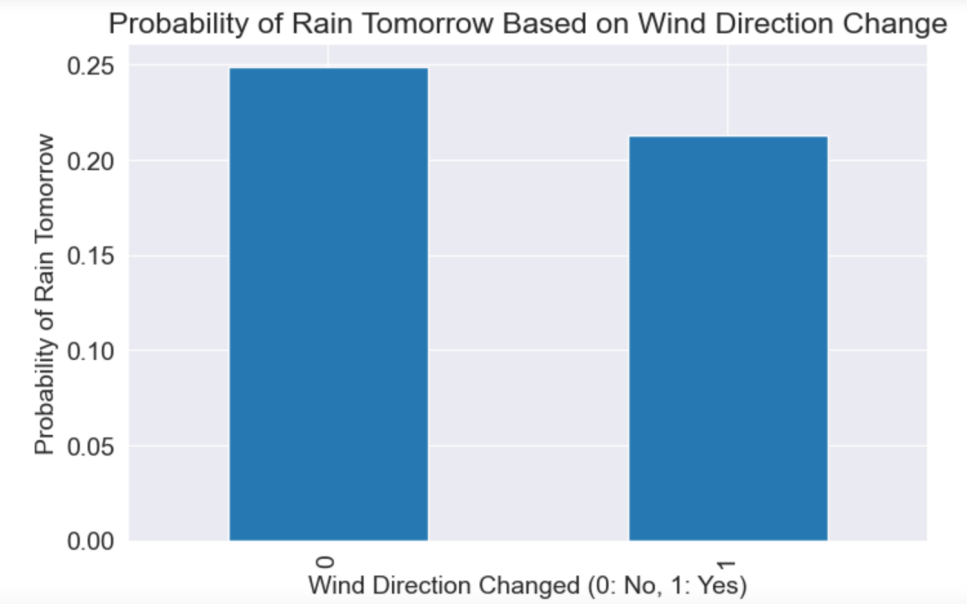
Analysis of Wind Patterns

The wind rose plots indicate distinct wind patterns throughout the day. Wind gusts are mainly from the West (W), with occasional high speeds. At 9 am, winds are primarily from the North (N) and tend to be lighter. By 3 pm, wind directions are more varied, with a significant portion coming from the South-East (S-E), and speeds are moderate.

Strong westerly gusts may suggest incoming weather systems from the west, potentially bringing rainfall, especially in specific seasons. The shift from northerly winds in the morning to southeasterly winds in the afternoon could be linked to local weather phenomena, like sea breezes, which often contribute to afternoon showers or thunderstorms.



Regarding wind direction stability, when the wind direction remains consistent, the chance of rainfall is about 24.88%. However, with changing wind directions, the probability decreases slightly to 21.31%. The weak correlation between wind direction stability and rainfall (-0.032) indicates that wind direction alone is not a reliable predictor of rainfall.



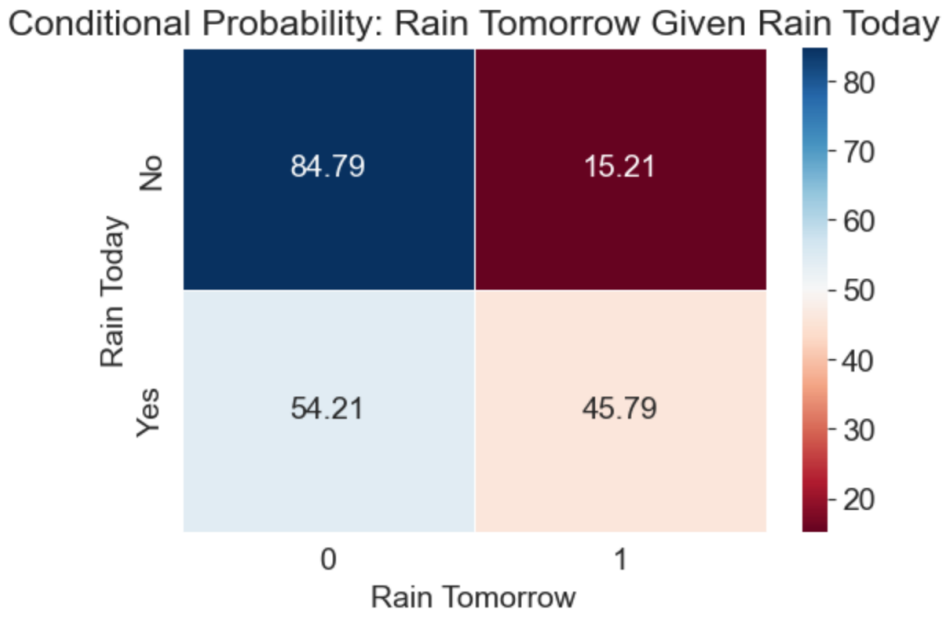
Probability of Rain Tomorrow Based on Today's RainFall

If it did not rain today (RainToday = No):

- There is an 84.79% chance that it will not rain tomorrow.
- There is a 15.21% chance that it will rain tomorrow.

If it rained today (RainToday = Yes):

- There is a 54.21% chance that it will not rain tomorrow.
- There is a 45.79% chance that it will rain tomorrow.

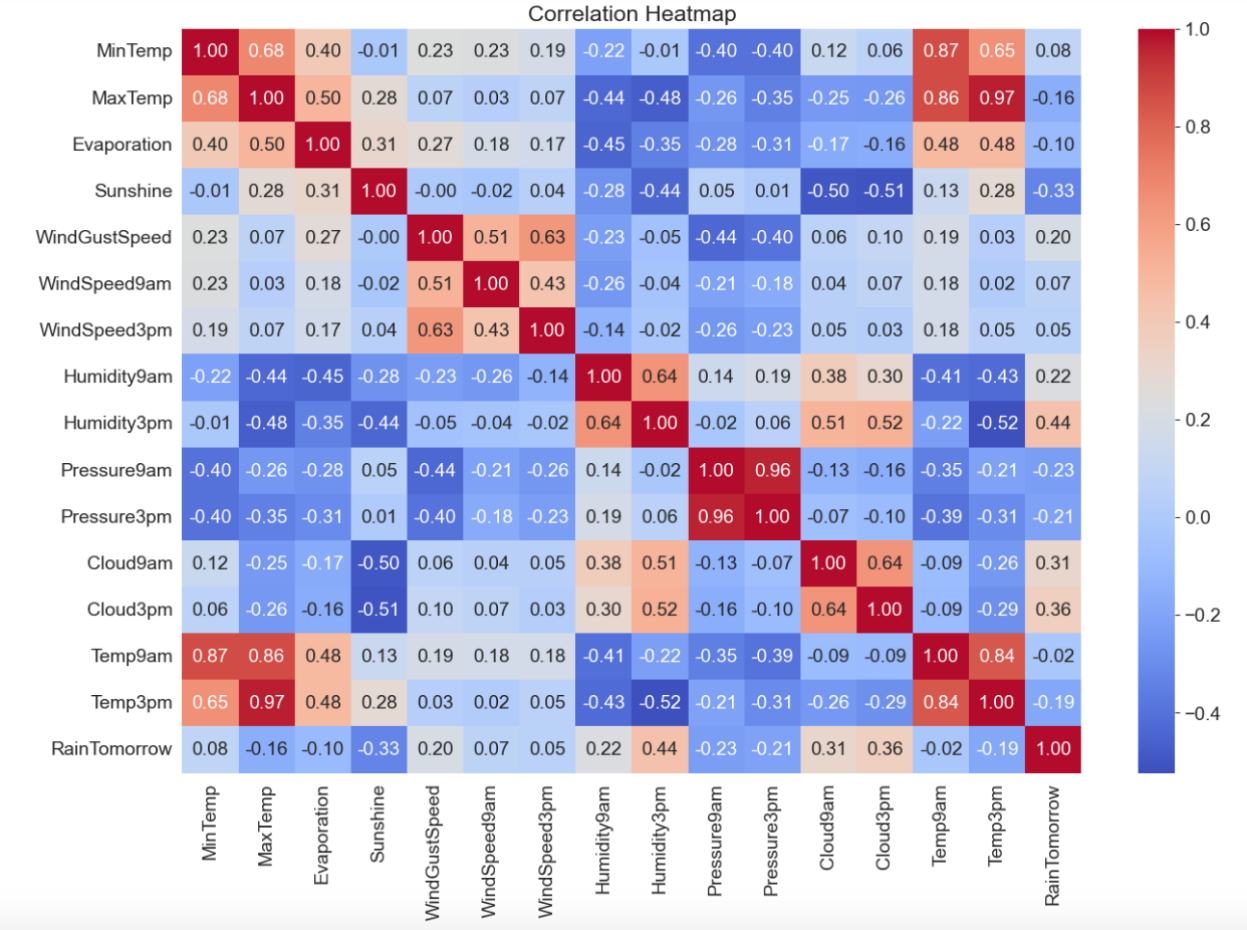


Hence we can say that, the probability of rain tomorrow increases significantly if it rained today (from 15.21% to 45.79%). This suggests that rainfall tends to persist, meaning if it rains one day, the next day is more

likely to also have rain. However, rain today does not guarantee rain tomorrow, as the probability is still below 50%.

Correlation Matrix

Our analysis found strong positive correlations between **Temp3pm & MaxTemp (~0.98)** and **Pressure9am & Pressure3pm (~0.99)**, indicating warmer afternoons lead to higher max temperatures and stable pressure throughout the day. A strong negative correlation was observed between **Humidity3pm & Temp3pm (-0.63)**, showing higher humidity lowers afternoon temperatures. Rainfall, however, did not strongly correlate with any single variable, suggesting it is influenced by multiple factors rather than one dominant predictor.



Variables such as **Humidity3pm and 9am, Sunshine, Cloud3pm and 9am, and Pressure3pm and 9am** showed a strong correlation with **RainTomorrow**. This may suggests that these factors play a significant role in determining whether it will rain the next day.

Modeling

The objective of this study is to develop a **predictive model** that can determine the likelihood of rainfall on the following day based on past weather conditions. This is a **binary classification problem**, where the target variable **"RainTomorrow"** takes two values:

- **Yes (1):** It will rain the next day.
- **No (0):** It will not rain the next day.

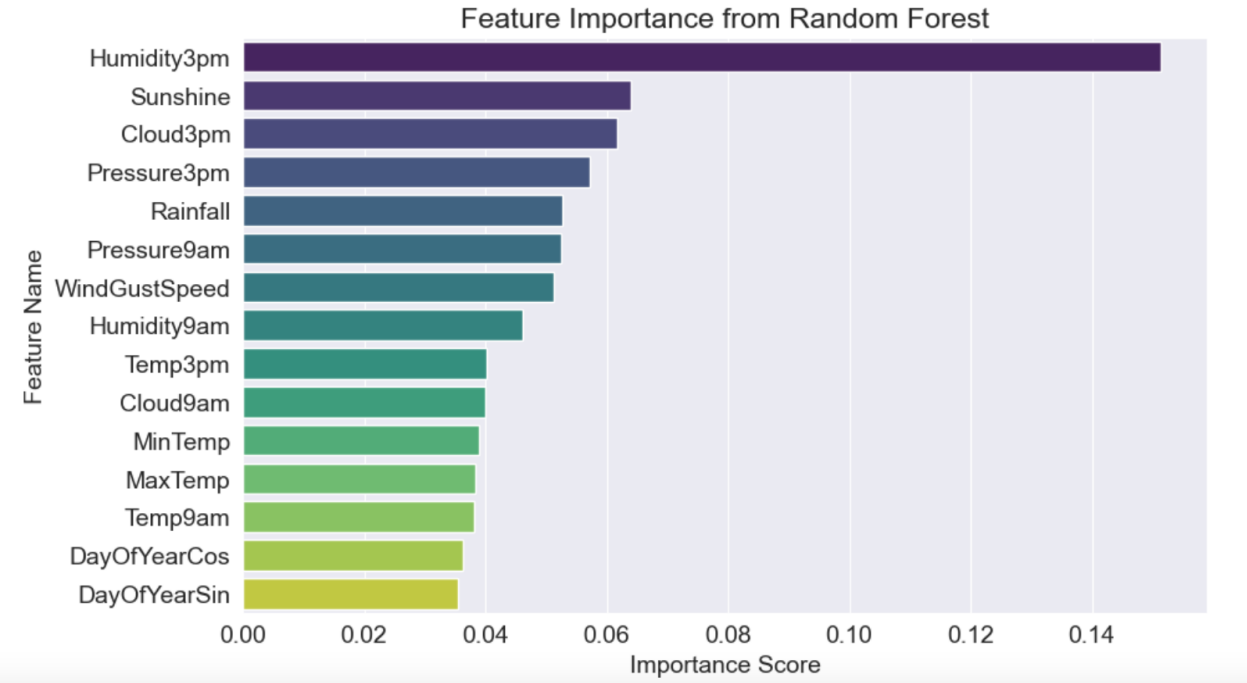
Given the **imbalanced class distribution (78% "No" vs. 22% "Yes")**, the challenge lies in building a model that not only achieves high accuracy but also performs well in distinguishing between the two classes. Accuracy alone may not be sufficient in this case, as a model that simply predicts "No" most of the time would still achieve high accuracy, but it would fail to capture the critical "Yes" predictions (rain events).

In weather forecasting, **recall** is particularly important. Recall refers to the model's ability to correctly identify instances of rain (true positives). A high recall value means the model can effectively predict rain on days when it occurs. Failing to predict rain (a false negative) can have more serious consequences, especially for applications where accurate rainfall predictions are crucial, such as agriculture, event planning, and emergency management. On the other hand, predicting rain when it doesn't occur (false positives) may lead to inconvenience but generally carries fewer severe consequences.

To address the challenge posed by class imbalance and to improve model performance, we employed **SMOTEENN (Synthetic Minority Over-sampling Technique and Edited Nearest Neighbors)**. SMOTEENN is an advanced technique that generates synthetic samples for the minority class (rain) while also removing noisy instances from both classes. This method enhances the model's ability to predict rainfall events without significantly altering the overall distribution of the classes. By using SMOTEENN, we aim to improve recall and the F1 score, ensuring that the model is robust and reliable for predicting rain events.

Feature Selection

The features selected for the model were identified using a Random Forest Classifier, which helps determine the most significant variables for predicting rainfall. The selected features are: Humidity3pm, Sunshine, Cloud3pm, Pressure3pm, Rainfall, Pressure9am, WindGustSpeed, Humidity9am, Temp3pm, Cloud9am, MinTemp, MaxTemp, Temp9am, DayOfYearCos, and DayOfYearSin. Among these, Humidity3pm appears to be the most significant, followed by Sunshine, Cloud3pm, and Pressure3pm.



Data Splitting & Preprocessing

For this analysis, we employed a time-based data splitting approach to ensure the model is trained on historical data and tested on future data. Instead of the usual random splitting, we divided the dataset based on the date column:

- **Training Data**: The data from years before **2015-01-01** was used to train the model. This ensured that the model was built using past weather conditions, which mirrors how weather forecasting models would typically function in real-world scenarios, where predictions for the future are made based on past observations.
- **Testing Data**: The data from **2015-01-01** onwards was used for model evaluation. This period serves as our test set to assess how well the model generalizes to unseen data, simulating a forecasting scenario for future dates.

Standard scaling was applied to the train and test data.

Machine Learning Models Used

To build an effective rain prediction system, we implemented and evaluated **four machine learning models**:

Gradient Boosting Classifier

Gradient Boosting is a powerful algorithm, well-regarded for its ability to build accurate models by combining multiple weak learners (typically decision trees). It excels in handling complex problems like rainfall prediction by capturing intricate patterns and interactions within the data.

Before balancing, the Gradient Boosting Classifier achieved an accuracy of 0.8447, with an F1-score of 0.56, recall of 0.46, and precision of 0.72 for class 1 (rain occurring). It correctly predicted 4,428 instances of rainfall, while 5,166 instances were misclassified as no rain. This suggests that while the model was effective, there was significant room for improvement, especially in its ability to identify rain events.

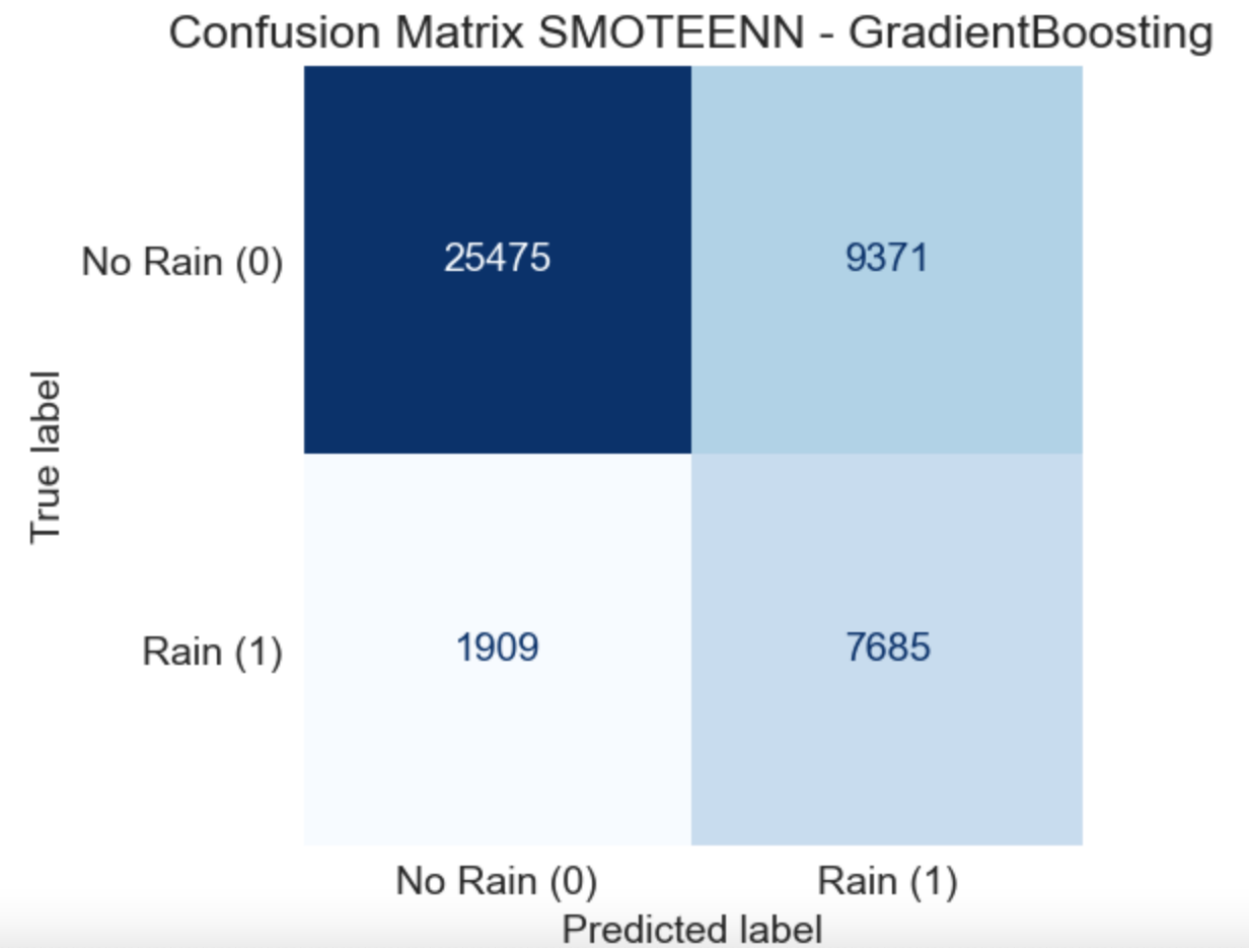
After applying SMOTEENN for balancing, accuracy dropped to 0.7462, but the ROC score improved to 0.8501, reflecting better model discrimination between the two classes. The recall for class 1 increased significantly to 0.80, and the F1-score improved to 0.58, highlighting a stronger capacity to correctly identify rainfall occurrences. The number of correct predictions for rain increased to 7,685, although 1,909 rain instances were still missed.

This improvement in recall and F1-score after balancing demonstrates that Gradient Boosting, despite the trade-off in accuracy, is a strong model choice for predicting rain occurrences, as it prioritizes minimizing false negatives and enhances overall prediction reliability.

```
Model Accuracy: 0.7462
ROC-AUC Score: 0.8501
Decision tree
              precision    recall  f1-score   support

           0       0.93      0.73      0.82     34846
           1       0.45      0.80      0.58      9594

    accuracy                           0.75     44440
   macro avg       0.69      0.77      0.70     44440
weighted avg       0.83      0.75      0.77     44440
```



Logistic Regression

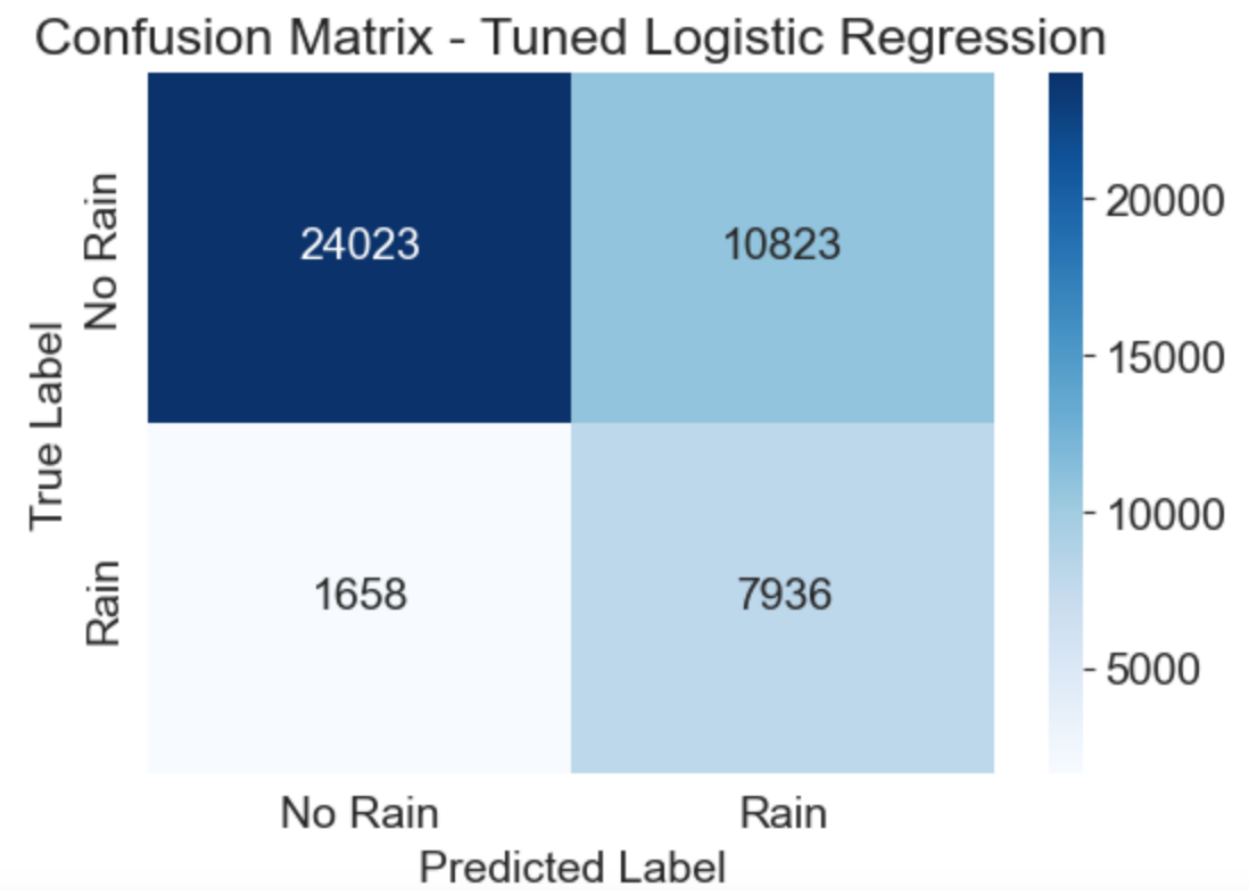
A simple, interpretable **linear model** that predicts the probability of rain. Works well when relationships between features and output are **linear**. Often serves as a **baseline model** in classification problems.

For this model, we performed grid search to find the optimal parameters, which were {'C': 0.01, 'penalty': 'l2', 'solver': 'saga'}. Initially, the model showed an accuracy of 0.84 and a recall for class 1 (rain) of 0.44. This indicated that only a small portion (4,285 instances) of the actual rainfall occurrences were correctly predicted. After applying a balancing technique to address the class imbalance, the recall for class 1 increased significantly to 0.83, although this came at the expense of precision. As a result, the model correctly predicted 7,936 rainfall occurrences, and the ROC score improved to 0.8479. The balancing approach helped the model become more sensitive to detecting rainfall events, which is crucial for weather forecasting, but it did lead to a trade-off in precision.

```
Model Accuracy: 0.7191
ROC-AUC Score: 0.8479
Logistic Regression
              precision    recall  f1-score   support

           0       0.94      0.69      0.79     34846
           1       0.42      0.83      0.56      9594

    accuracy                           0.72     44440
   macro avg       0.68      0.76      0.68     44440
weighted avg       0.82      0.72      0.74     44440
```



Decision Tree Classifier

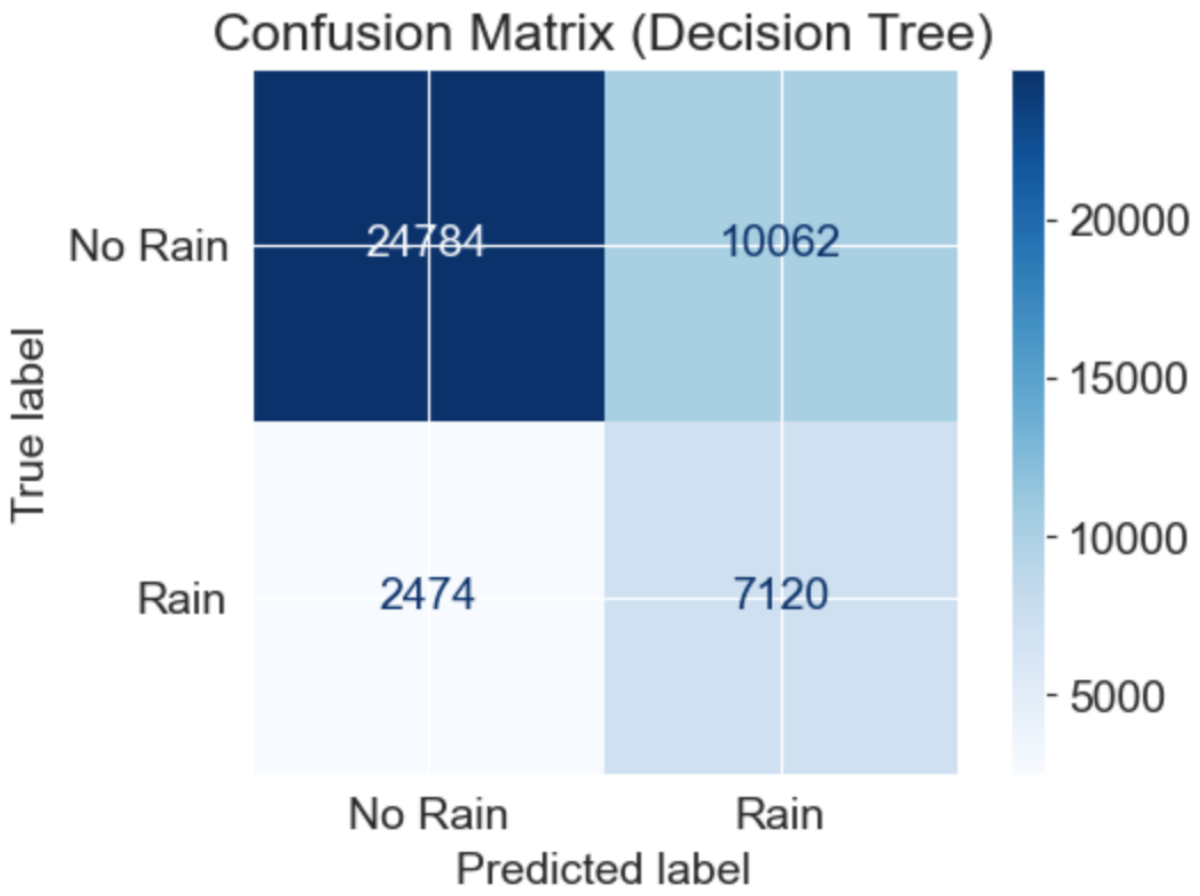
A rule-based model that splits data based on feature values to make decisions. Easily interpretable, but prone to overfitting if not properly tuned.

A Decision Tree model was trained with a maximum depth of 3 to help understand the decision-making process. Without data balancing, the model achieved an accuracy of 0.8280, correctly predicting 3,259 instances of rain occurrences out of 44,440. However, due to the imbalanced nature of the dataset, the model struggled to predict the rare event of rainfall. After applying data balancing techniques, the accuracy decreased to 0.7179, with an ROC score of 0.7873. This trade-off led to a substantial improvement in recall for the "Rain" class (1), rising from 0.34 to 0.74, and correctly predicting 7,120 rain occurrences. The increase in recall is significant for practical applications, as it reduces the chance of missing rainfall predictions, which can be critical for weather forecasting and decision-making.

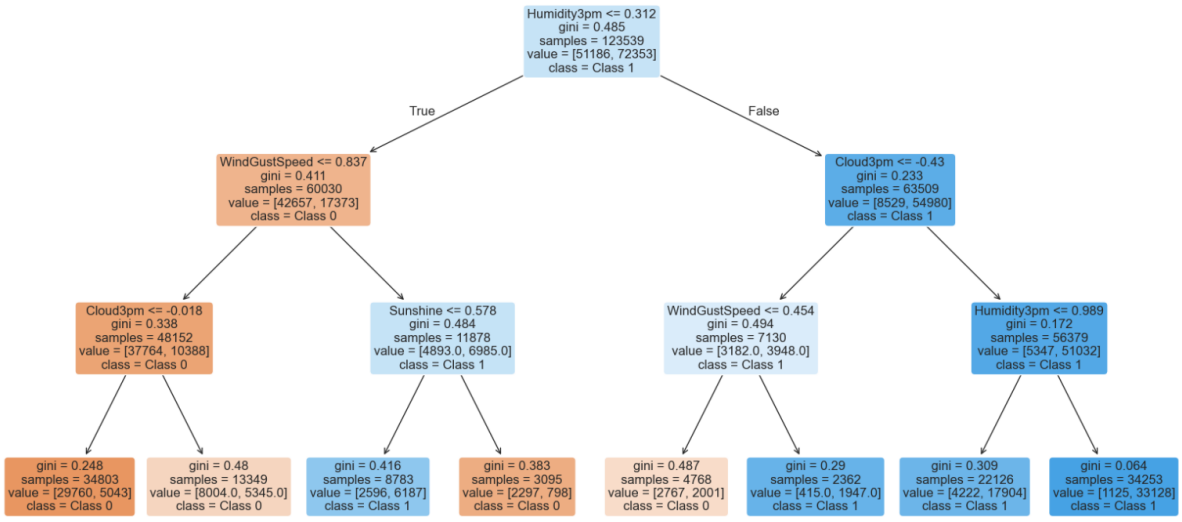
```
Accuracy Score: 0.7179
ROC-AUC Score: 0.7873
              precision    recall  f1-score   support

           0       0.91      0.71      0.80     34846
           1       0.41      0.74      0.53      9594

    accuracy                           0.72     44440
   macro avg       0.66      0.73      0.66     44440
weighted avg       0.80      0.72      0.74     44440
```



The decision tree visualizes key features and conditions for predicting rain. The most important feature is Humidity3pm, with a threshold of 31.2%. If the humidity is lower than or equal to this value, the tree follows the left branch, predicting "no rain" in most cases unless other factors like WindGustSpeed or Cloud3pm are involved. For higher humidity levels (above 31.2%), Cloud3pm becomes the next crucial factor in the prediction. When Cloud3pm is low, the tree predicts "rain," but if it's higher, Humidity3pm takes precedence.



Random Forest Classifier

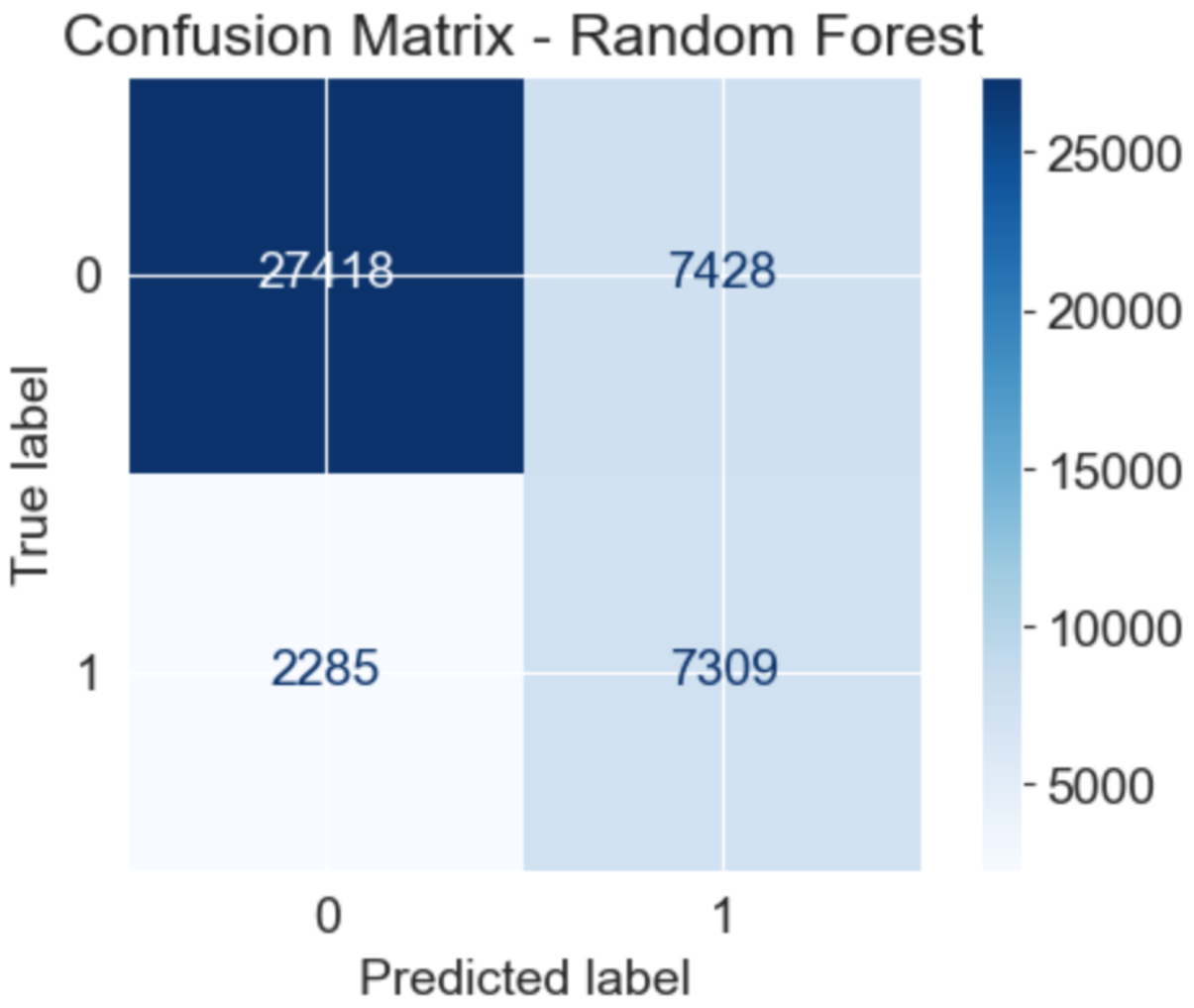
An **ensemble learning method** that combines multiple decision trees to improve accuracy. More robust and less prone to overfitting than a single decision tree.

Before balancing, the Random Forest model achieved an accuracy of 0.8454, with a recall of 0.45, precision of 0.73, and an F1-score of 0.56 for class 1 (rain occurring), correctly predicting 4,321 out of 9,594 rainfall occurrences. After applying SMOTEENN for balancing, the precision for class 1 decreased to 0.50, while recall increased to 0.76, and the F1-score improved to 0.60. However, accuracy dropped to 0.7814, and the ROC score improved to 0.8542. The model correctly predicted 7,309 out of 9,594 instances of rainfall, which makes up approximately 76.18% of the total rainfall occurrences. This improvement reflects the model's enhanced ability to identify rain events after applying the SMOTEENN technique, balancing the class distribution. While the accuracy decreased slightly, the increased recall and F1-score highlight the model's improved performance in detecting rainfall, reducing the number of false negatives, and offering a more reliable forecast for rain events.

```
Accuracy Score: 0.7814
ROC-AUC Score: 0.8542
              precision    recall  f1-score   support

           0       0.92      0.79      0.85     34846
           1       0.50      0.76      0.60      9594

    accuracy                           0.78     44440
   macro avg       0.71      0.77      0.73     44440
weighted avg       0.83      0.78      0.80     44440
```



Support Vector Machine (SVM)

A powerful **classification algorithm** that finds an optimal hyperplane to separate classes. Works well in **high-dimensional spaces** and with complex data distributions. Can be enhanced with different **kernel functions** (linear, polynomial, RBF).

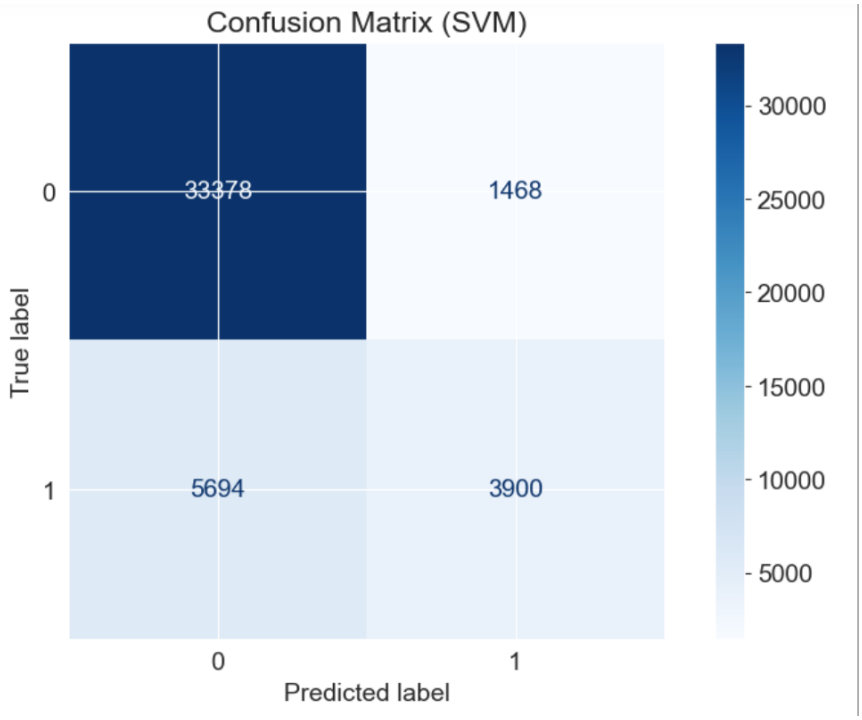
Due to the high computational cost, balancing techniques (like SMOTEENN) could not be applied to the model. The training process took 76 minutes to complete on the unbalanced dataset. Despite the class imbalance, the model achieved an accuracy of 0.8388. For class 1 (rain), the F1-score was 0.52, recall was 0.41, and precision was 0.73. The model correctly predicted 3,900 out of 9,564 instances of rainfall, highlighting the difficulty of predicting the minority class without balancing. The ROC score of 0.8475 indicates a decent model performance overall, but the relatively low recall suggests room for improvement in detecting true rain events (minimizing false negatives).

The absence of balancing techniques in this case is a limitation, as it likely impacted the model's ability to better capture the minority class (rain occurrences).

```
Accuracy Score (SVM): 0.8388
ROC-AUC Score (SVM): 0.8475
Classification Report (SVM):
              precision    recall  f1-score   support
```

```
              0       0.85      0.96      0.90     34846
              1       0.73      0.41      0.52      9594

    accuracy                            0.84     44440
   macro avg        0.79      0.68      0.71     44440
weighted avg        0.83      0.84      0.82     44440
```



Model Performance Comparison

Model	Accuracy	F1-Score	Recall (Rain)	Precision (Rain)	ROC Score	True Positive	False Negative
Logistic Regression	0.7191	0.56	0.83	0.42	0.8479	7,936	1,658
Decision Tree	0.7179	0.53	0.74	0.41	0.7873	7,120	2,474
Random Forest	0.7814	0.60	0.76	0.50	0.8542	7,309	2,285
Gradient Boosting	0.7462	0.58	0.80	0.45	0.8501	7,685	1,909
SVM (Unbalanced data)	0.8388	0.52	0.41	0.73	0.8475	3900	5694

When comparing the performance of the different models in the context of rain prediction, the following observations can be made:

Logistic Regression:

- Strengths: Logistic Regression achieves the highest Recall (0.83) among all models, meaning it is the best at identifying actual rain events (true positives). This makes it suitable when missing a rain event (false negatives) is highly undesirable.
- Weaknesses: Its Precision (0.42) is low, meaning it generates a significant number of false positives (predicting rain when it doesn't occur). While its Accuracy (0.7191) and ROC Score (0.8479) are reasonable, they are not the best overall.

Decision Tree:

- Strengths: The Decision Tree provides an interpretable model with moderate performance across all metrics. It achieves a Recall of 0.74 and Precision of 0.41.
- Weaknesses: It has the lowest ROC Score (0.7873) and performs worse than Random Forest and Gradient Boosting in most metrics.

Random Forest:

- Strengths: Random Forest provides a good balance between Recall (0.76) and Precision (0.50), achieving the highest F1-Score (0.60) and the best overall ROC Score (0.8542). It also has high Accuracy (0.7814) and predicts 7,309 rain events correctly with 2,285 incorrect predictions.
- Weaknesses: While its Recall is slightly lower than Gradient Boosting, it still performs well across all metrics.

Gradient Boosting:

- Strengths: Gradient Boosting achieves a high Recall (0.80), second only to Logistic Regression, meaning it captures most rain events effectively. It also has a strong ROC Score (0.8501) and fewer incorrect predictions (1,909) compared to Random Forest.
- Weaknesses: Its Precision (0.45) is lower than Random Forest's, meaning it produces more false positives.

SVM (Unbalanced Data):

- Strengths: SVM achieves the highest Accuracy (0.8388) and Precision (0.73), meaning it minimizes false positives better than any other model.
- Weaknesses: Its Recall is extremely low (0.41), making it unsuitable for predicting rain events since it misses most actual rain occurrences.

Conclusion

In our quest to develop an accurate model for predicting rainfall in Australia, we found that machine learning can be highly effective for this task. While models like SVM and Decision Tree showed reasonable performance, more advanced ensemble methods such as Random Forest, Gradient Boosting, and Logistic Regression performed exceptionally well, each with its own strengths and weaknesses.

Logistic Regression emerged as the best at correctly identifying rain occurrences, achieving the highest recall. Gradient Boosting and Random Forest also demonstrated strong performance, with a good balance between recall and precision. SVM could have performed better with proper data balancing, but due to computational constraints, it was not further optimized.

Our analysis also revealed key factors influencing rainfall. Humidity showed the strongest correlation with rain occurrence, while other significant variables included sunshine, cloud cover at 3 PM, atmospheric pressure at 3 PM, wind gust speed, and the amount of prior rainfall. This indicates that weather conditions in the afternoon (3 PM) play a crucial role in predicting rainfall. Additionally, seasonal patterns captured through the cosine and sine transformations of the day of the year were also found to be significant.

From a geographical perspective, Queensland and the Northern Territory experience significantly higher rainfall compared to other regions in Australia, further reinforcing the influence of location on weather patterns.

Reference

Iqbal, R., Doctor, F., More, B., Mahmud, S., & Yousuf, U. (2022). Rainfall prediction system using machine learning fusion for smart cities. Future Internet, 14(5), 144. https://doi.org/10.3390/fi14050144

Rain in Australia. (2020, December 11). Kaggle. https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package

Kaya, A., & Yıldız, C. (2023). Rainfall prediction using an ensemble machine learning model based on K-Stars. Sustainability, 15(7), 5889. https://doi.org/10.3390/su15075889

Meinke, H., Howden, S. M., Struik, P. C., Nelson, R., Rodriguez, D., & Chapman, S. C. (2009). Adaptation science for agriculture and natural resource management—urgency and theoretical basis. Current Opinion in Environmental Sustainability, 1(1), 69-76. https://doi.org/10.1016/j.cosust.2009.07.007

Tao, Y., Hsu, K., Ihler, A., Gao, X., & Sorooshian, S. (2021). Statistical and machine learning methods applied to the prediction of 3-hourly rainfall over the tropical Pacific. Journal of Hydrometeorology, 22(12), 3079-3094. https://doi.org/10.1175/JHM-D-21-0110.1

Appendix

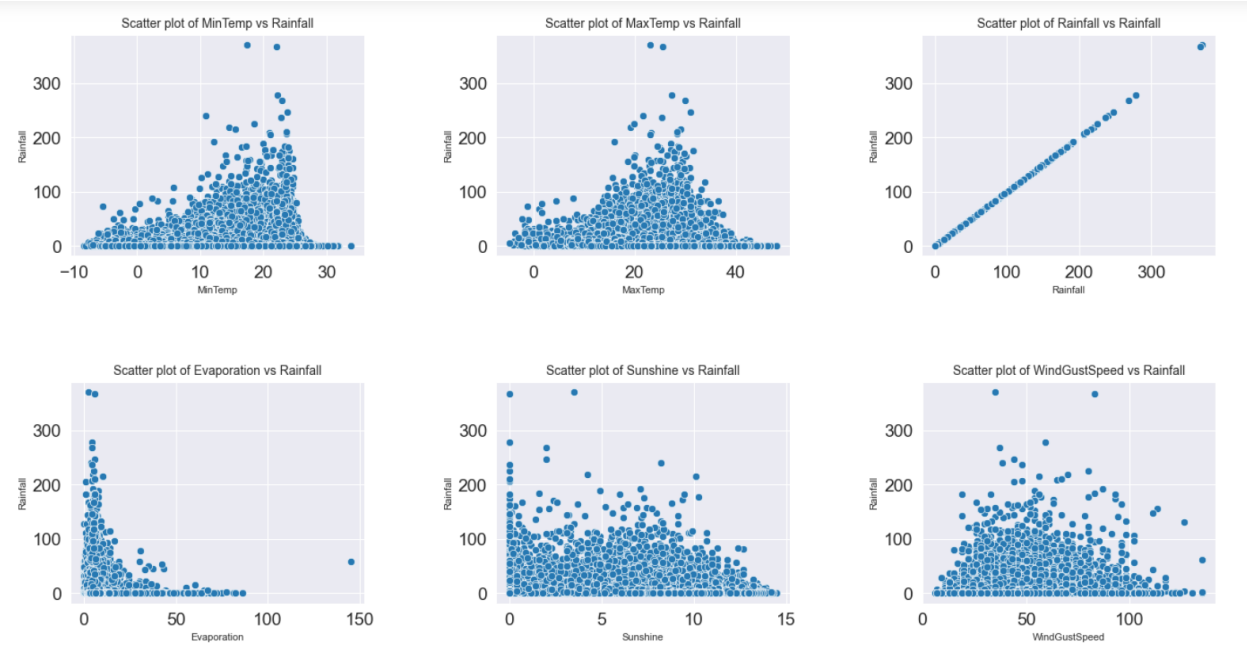
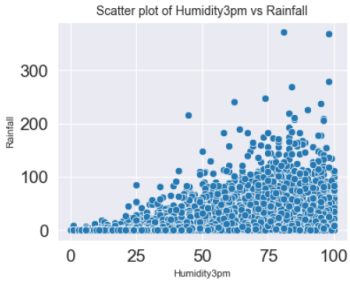
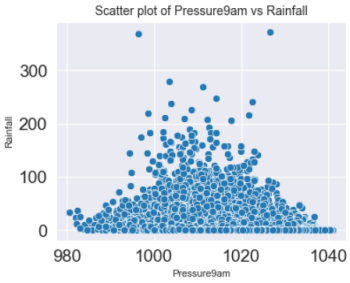
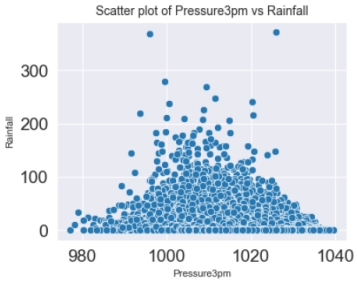
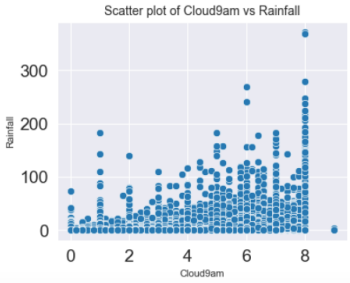


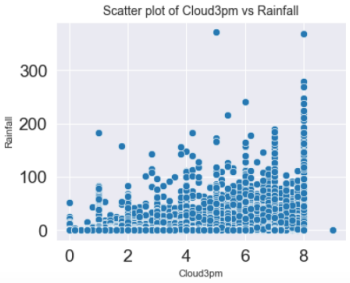
Figure 1: Scatterplot

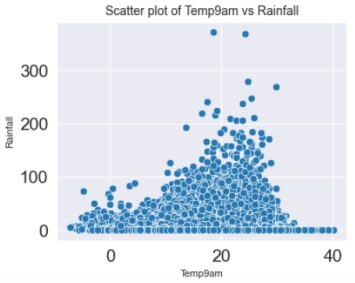

Scatter plot of Humidity3pm vs Rainfall
Rainfall
Humidity3pm
0
100
200
300
0
25
50
75
100


Scatter plot of Pressure9am vs Rainfall
Rainfall
Pressure9am
0
100
200
300
980
1000
1020
1040


Scatter plot of Pressure3pm vs Rainfall
Rainfall
Pressure3pm
0
100
200
300
980
1000
1020
1040


Scatter plot of Cloud9am vs Rainfall
Rainfall
Cloud9am
0
100
200
300
0
2
4
6
8


Scatter plot of Cloud3pm vs Rainfall
Rainfall
Cloud3pm
0
100
200
300
0
2
4
6
8


Scatter plot of Temp9am vs Rainfall
Rainfall
Temp9am
0
100
200
300
0
20
40