Dawoud Tormos, Ghandi Mahmoud

Machine Learning Approach for detecting Rain in the second day of prediction.

Rainfall Prediction: Machine Learning Approach

Rainfall Prediction: Machine Learning Approach

Table of Contents

1. Introduction
2. Problem Definition
3. How Rain is Predicted Usually
   * Atmospheric Water Vapor
   * Cloud Formation Triggers
   * Weather Models
   * Human Expertise
   * Measurement Tools
4. Our Machine Learning Approach
5. Dataset Overview
6. Source & Acknowledgements
7. Feature Overview
8. Data Preprocessing
   * Missing Value Handling
   * Scaling / Normalization
   * Encoding Techniques
9. Exploratory data analysis (EDA)
10. Model Selection and Evaluation
    * Choosing Best K-Value for K-NN Model
    * Confusion Matrix for K-NN Results
    * Confusion Matrix for Logistic Model Results
    * Choosing Best Max-Depth for Decision Tree Model
    * Confusion Matrix for Decision Tree Model Results
    * Choosing Best Max-Depth Value for Random Forest Model
    * Confusion Matrix for Random Forest Model Results
    * Confusion Matrix for Naïve Bayes Model Results
11. Accuracy Comparison
12. **Introduction**

The ability to predict rainfall accurately is vital for various sectors, including agriculture, disaster management, and urban planning. This report outlines a machine learning approach to predict whether it will rain the following day based on meteorological data collected from previous days.

1. **Problem Definition**

The central challenge addressed in this project is determining whether it will rain tomorrow based on various measurements and factors recorded from the previous day. The goal is to develop a model that can generalize well to unseen data and provide reliable predictions.

1. **How Rain is Predicted Usually**

Atmospheric Water Vapor

Atmospheric water vapor is a critical component in predicting rainfall, measured using precipitable water content to assess moisture availability in the atmosphere.

Cloud Formation Triggers

Meteorologists analyze cloud stability, height, and type (e.g., stratiform or convective) to predict rain likelihood, as these factors directly influence precipitation patterns.

Weather Models

Advanced weather models utilize satellite imagery and data from weather balloons to simulate atmospheric conditions and provide forecasts of precipitation.

Human Expertise

Local knowledge plays an essential role in refining predictions, especially during unpredictable weather events like thunderstorms, where human forecasters can make informed judgments based on experience.

Measurement Tools

Rainfall is recorded using various measurement tools, including traditional rain gauges and modern automated systems that provide real-time data.

1. **Our Machine Learning Approach**

This project leverages machine learning techniques to enhance rainfall prediction accuracy. By utilizing historical meteorological data, the model aims to identify patterns that can indicate future rainfall events. Although machine learning models may not achieve perfect accuracy, they significantly reduce the reliance on human expertise and traditional forecasting methods.

1. **Dataset Overview**

The dataset used in this analysis was collected from multiple weather stations across Australia and is publicly available through the Bureau of Meteorology's Climate Data Online platform. The dataset includes daily observations over several years, providing a comprehensive view of weather patterns.

1. **Source & Acknowledgements**

The observations were gathered from numerous weather stations operated by the Bureau of Meteorology, Australia’s national weather agency. Daily observations can be accessed online, allowing researchers and practitioners to analyze historical weather data effectively.

1. **Feature Overview**

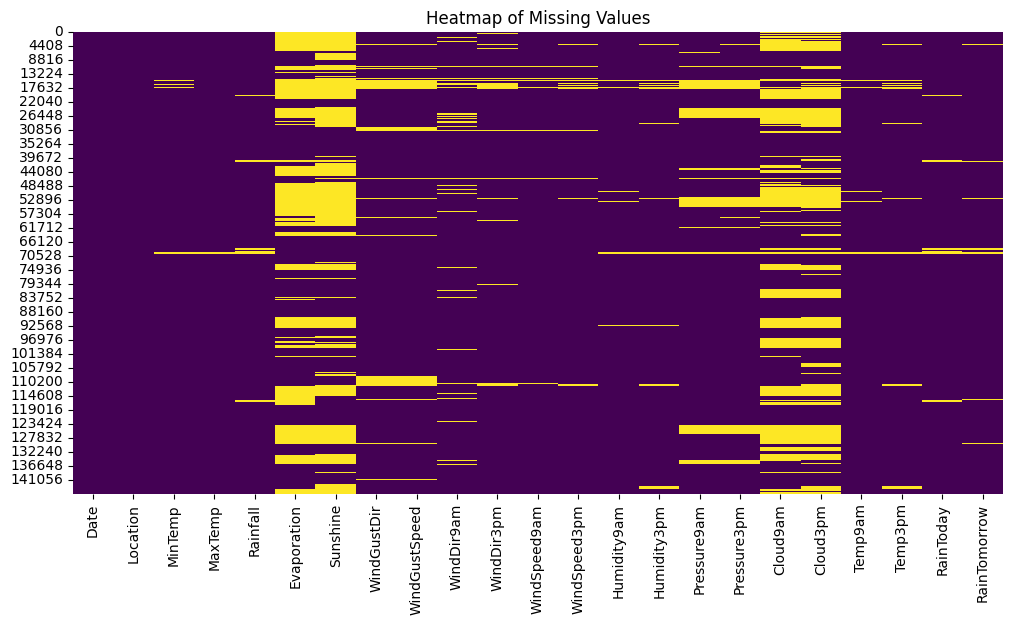
All features in the dataset are pertinent to predicting rainfall, including:

| **Feature** | **Description** |
| --- | --- |
| Date | The date of observation |
| Location | The weather station location |
| MinTemp | Minimum temperature recorded |
| MaxTemp | Maximum temperature recorded |
| Rainfall | Amount of rainfall measured |
| Evaporation | Evaporation rate |
| Sunshine | Hours of sunshine |
| WindGustDir | Direction of wind gusts |
| WindGustSpeed | Speed of wind gusts |
| WindDir9am | Wind direction at 9 AM |
| WindDir3pm | Wind direction at 3 PM |
| WindSpeed9am | Wind speed at 9 AM |
| WindSpeed3pm | Wind speed at 3 PM |
| Humidity9am | Humidity level at 9 AM |
| Humidity3pm | Humidity level at 3 PM |
| Pressure9am | Atmospheric pressure at 9 AM |
| Pressure3pm | Atmospheric pressure at 3 PM |
| Cloud9am | Cloud cover at 9 AM |
| Cloud3pm | Cloud cover at 3 PM |
| Temp9am | Temperature at 9 AM |
| Temp3pm | Temperature at 3 PM |
| RainToday | Indicator if it rained today (Yes/No) |
| RainTomorrow | Indicator if it will rain tomorrow (Yes/No) |

1. **Data Preprocessing**

* Missing Value Handling

Handling missing values is crucial for maintaining dataset integrity. Numerical missing values were replaced with their mean values, while categorical missing values were filled with the most frequently occurring value within that column. Rows with more than eight missing values were dropped from the dataset to ensure quality.



* Scaling / Normalization

To ensure that numerical features are comparable and contribute equally to model training, all numerical values were normalized using MinMaxScaler, which scales features to a range between 0 and 1.

* Encoding Techniques

**Label Encoding** was employed to convert categorical variables into a numerical format suitable for machine learning algorithms. This technique allows algorithms to interpret categorical data as numerical values while preserving the order among categories when applicable.

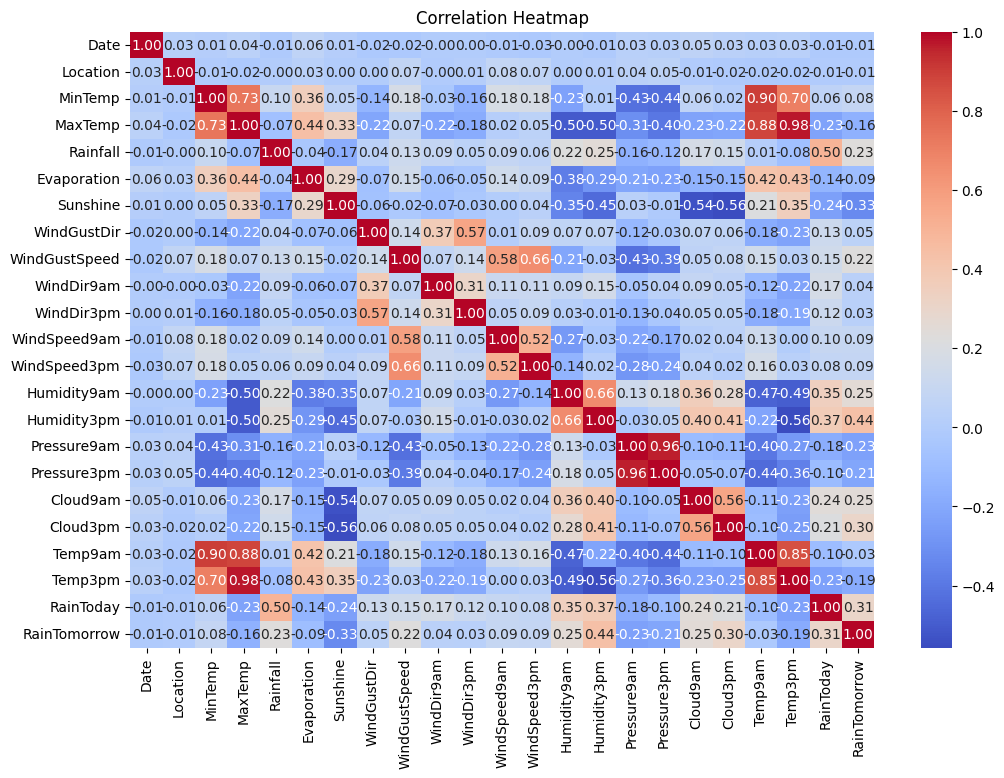
1. **EDA**

**Features Used**

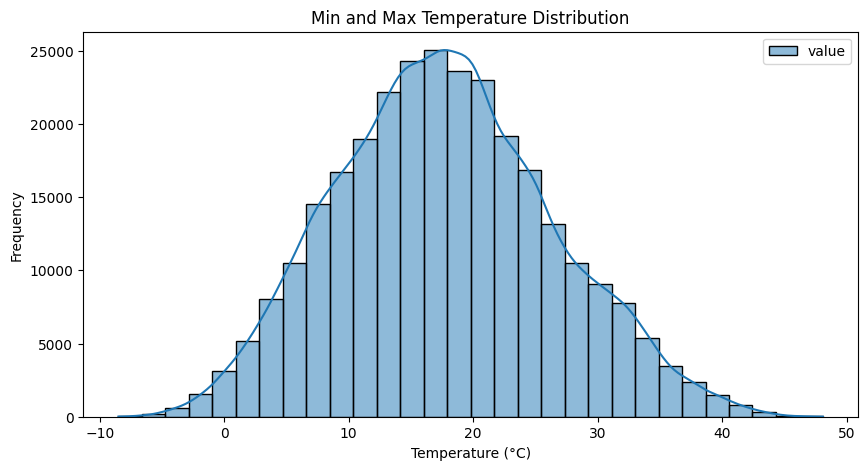
All Features in the dataset where used as they all where relevant to the target.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Date | Location | MinTemp | MaxTemp | Rainfall | Evaporation |
| Sunshine | WindGustDir | WindGustSpeed | WindDir9am | WindDir3pm | WindSpeed9am |
| WindSpeed3pm | Humidity9am | Humidity3pm | Pressure9am | Pressure3pm | Cloud9am |
| Cloud3pm | Temp9am | Temp3pm | RainToday | RainTomorrow (Target) |  |

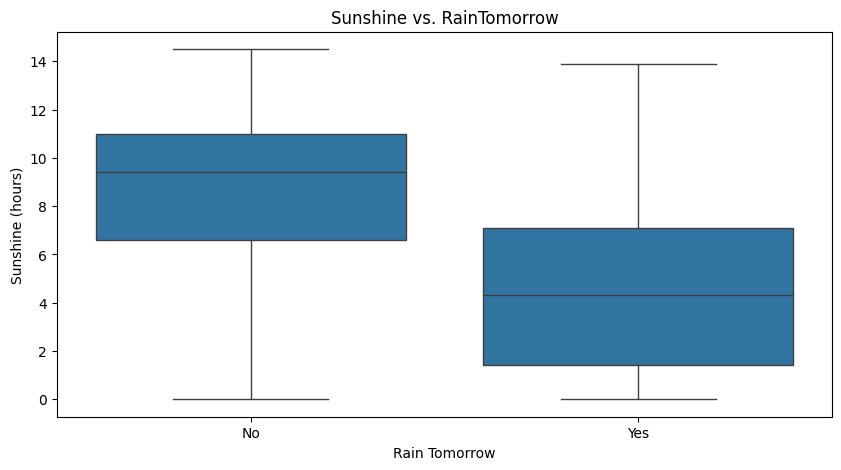
Correlation matrix heatmap



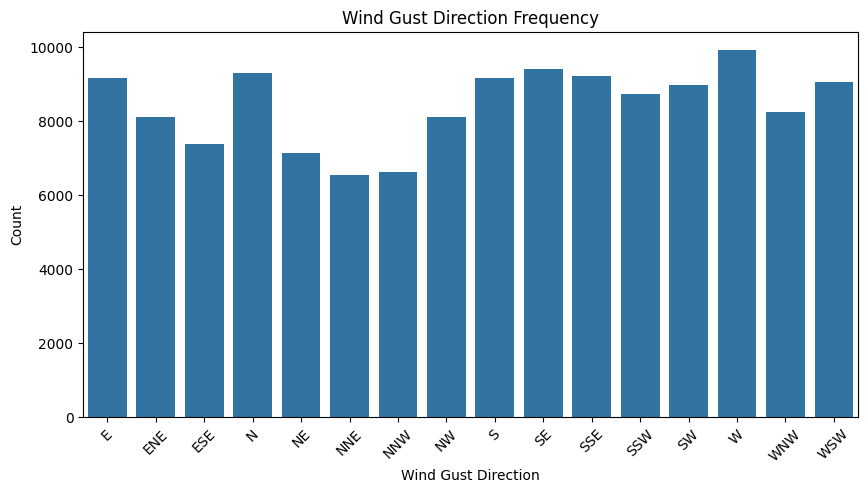
Min and Max Distribution



Sunshine vs RainTomorrow

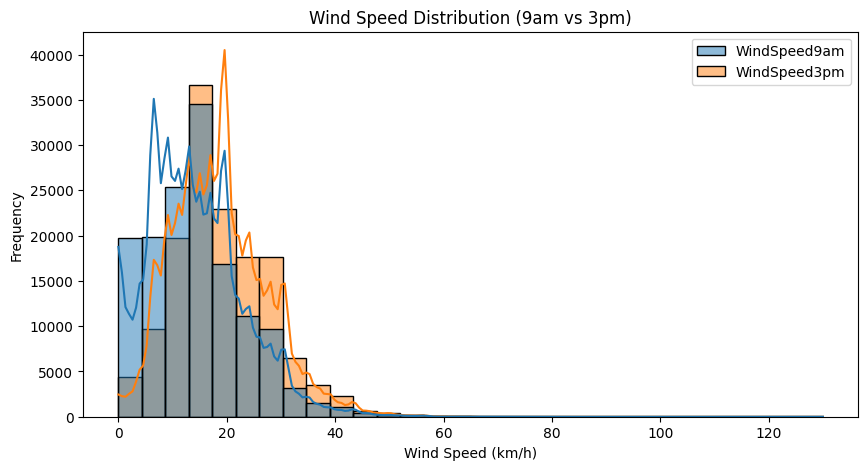


Wind Gust Direction frequency



Wind Speed Distribution (9am vs 3pm)

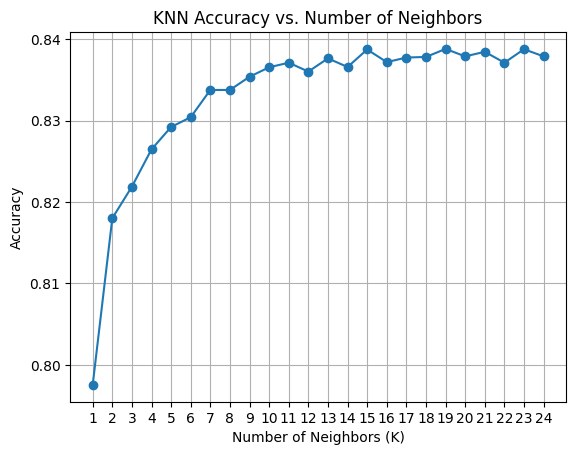
.



1. **Model Selection and Evaluation**

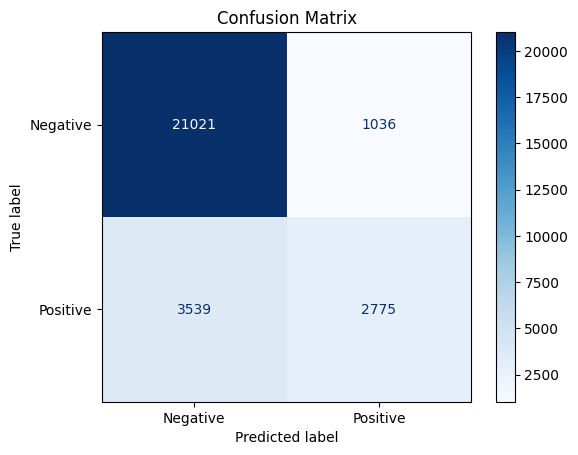
Choosing Best K-Value for K-NN Model

The K-Nearest Neighbors (K-NN) algorithm's performance was evaluated by testing various values of K (the number of neighbors). Cross-validation techniques were employed to determine which K-value yielded the highest accuracy without overfitting.



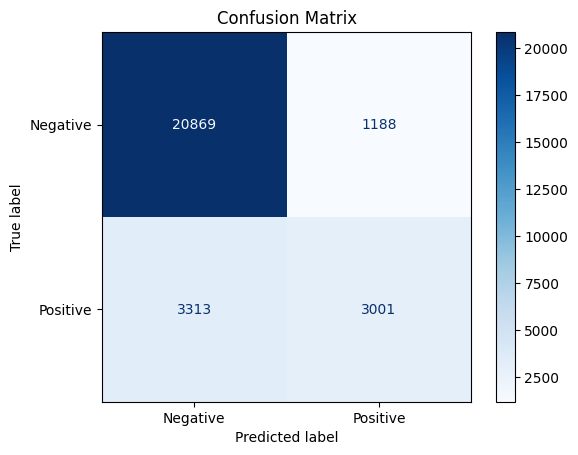
Confusion Matrix for K-NN Results

A confusion matrix was generated to visualize the performance of the K-NN model, showcasing true positives, false positives, true negatives, and false negatives.



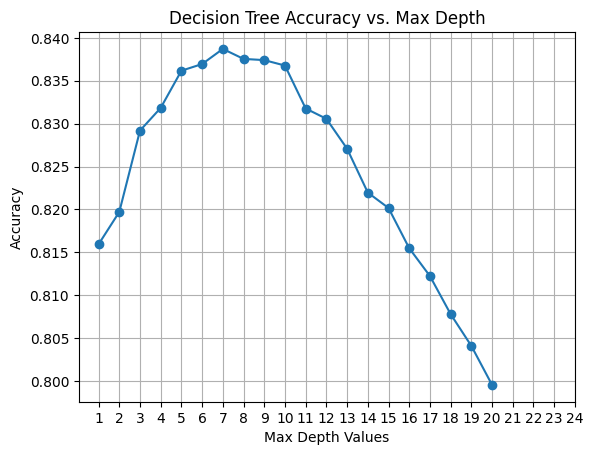
Confusion Matrix for Logistic Model Results

Similarly, logistic regression results were assessed using a confusion matrix to evaluate classification performance against actual outcomes.



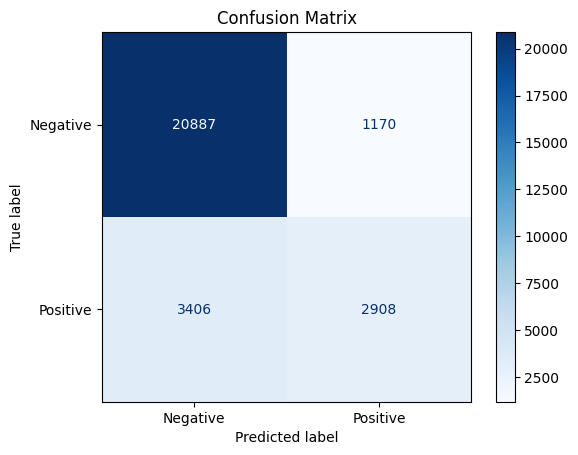
Choosing Best Max-Depth for Decision Tree Model

For the decision tree model, different maximum depth values were tested to find an optimal depth that balances bias and variance effectively.



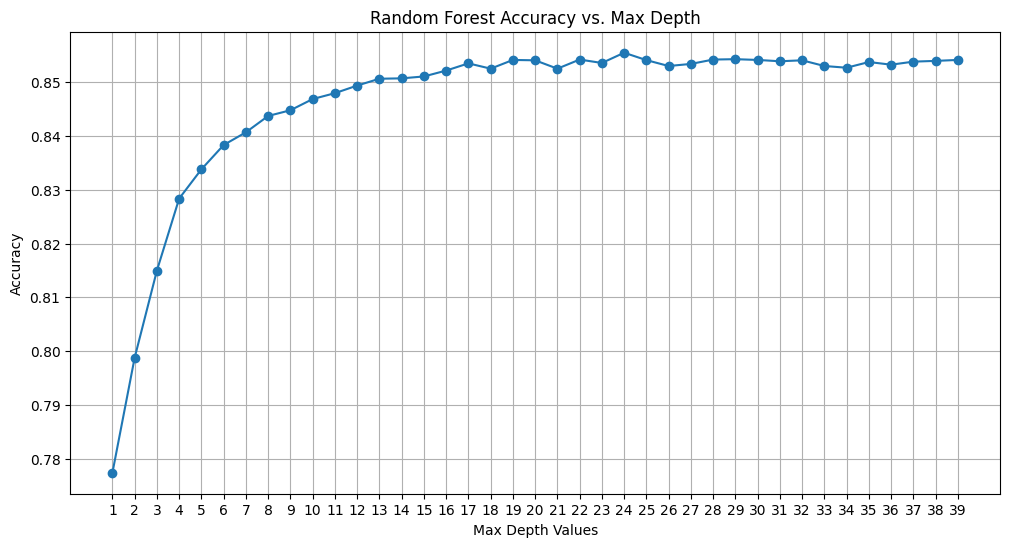
Confusion Matrix for Decision Tree Model Results

The decision tree's performance was evaluated through its confusion matrix, providing insights into its predictive capabilities.



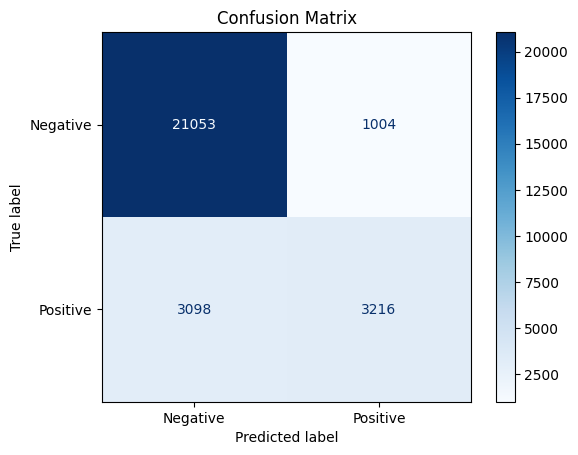
Choosing Best Max-Depth Value for Random Forest Model

In the random forest model, various max-depth values were analyzed to optimize tree depth while maintaining model robustness against overfitting.



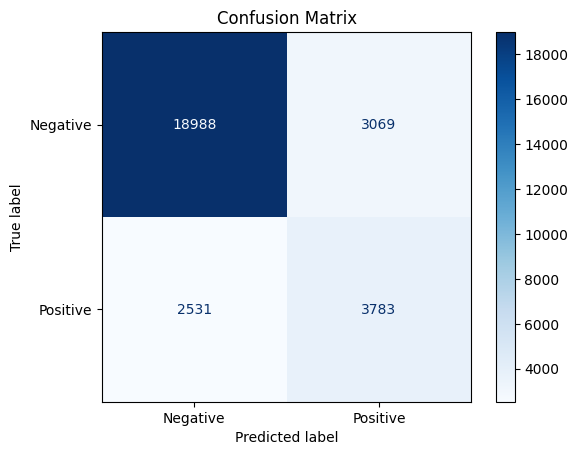
Confusion Matrix for Random Forest Model Results

Confusion matrices were created for random forest results as well, allowing comparison with other models regarding their predictive accuracy.



Confusion Matrix for Naïve Bayes Model Results

Naïve Bayes model performance was similarly evaluated through confusion matrices to assess its effectiveness in predicting rainfall outcomes.



1. **Accuracy Comparison**

An overall accuracy comparison across different models was conducted to identify which approach yielded the best predictive performance in rainfall forecasting. The results indicated varying levels of accuracy among models:

* **K-NN**: Achieved moderate accuracy but sensitive to outliers.(Accuracy: 0.838)
* **Logistic Regression**: Provided good baseline performance. (Accuracy: 0.8414)
* **Decision Tree**: Showed potential but prone to overfitting with high max-depth. (Accuracy: 0.8387)
* **Random Forest**: Outperformed other models due to ensemble learning capabilities. (Accuracy: 0.8554)
* **Naïve Bayes**: Offered competitive results despite its simplicity. (Accuracy: 0.8026)

