Intellihack 5.0

Task 01

Weather Forecasting Challenge – Part I

Team Scope

Introduction

Weather forecasting is essential for agricultural decision-making, helping farmers determine optimal timing for irrigation, planting, and harvesting. However, conventional weather prediction models often lack accuracy for specific local conditions.

This project's objective is to develop a machine learning model that can predict rainfall based on historical weather data. The dataset contains 300 daily weather observations with various meteorological features including temperature, humidity, wind speed, cloud cover, and pressure.

The specific aim is to accurately forecast whether rain will occur over the next 21 days by analyzing patterns in the historical weather data.

1. Data Preprocessing

Missing Value Treatment:

- Gaps identified in temperature, humidity, wind speed, and cloud cover measurements
- Addressed numerical gaps using average or middle values
- For any categorical variables, filled with most common values

Data Quality Corrections:

- Abnormal wind speed readings detected (peak of 56.6 km/h appeared suspicious)
- Applied interquartile range technique to identify and correct these anomalies

Feature Development:

- Calculated dew point using the formula: Temperature - ((100 - Humidity)/5)
- Added seasonal variables derived from date information

Data Transformation:

- Converted rain occurrence to numerical format (0 for dry days, 1 for rainy days)
- Normalized all features using standard scaling methods to ensure consistent ranges

2. Exploratory Data Analysis (EDA)

2.1. Correlation Matrix

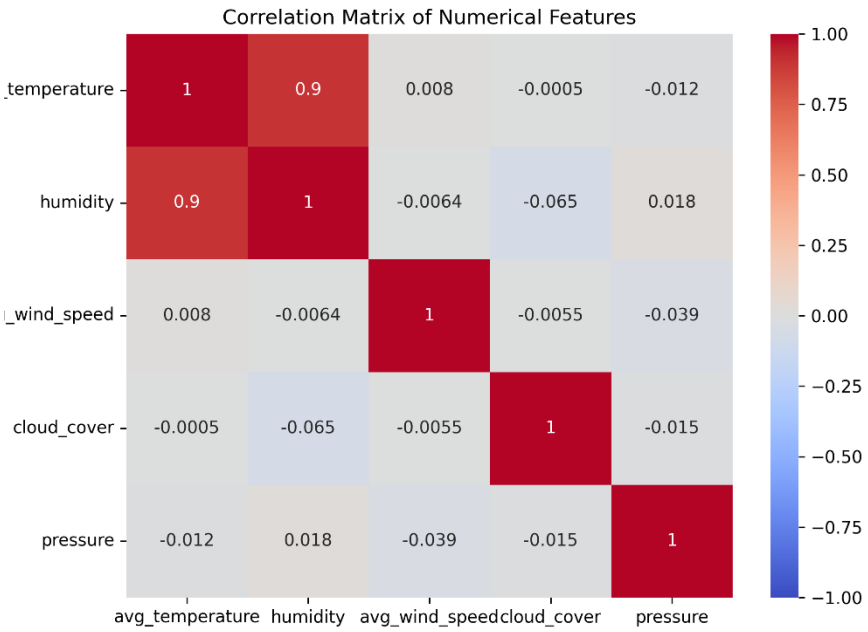


Figure 1 - Correlation Matrix

Main Findings

- The analysis reveals humidity as the most reliable indicator of rainfall
- Temperature exhibits a moderate relationship with precipitation events
- Neither cloud coverage nor atmospheric pressure demonstrates meaningful predictive value for rainfall occurrence

2.2. Data Visualization

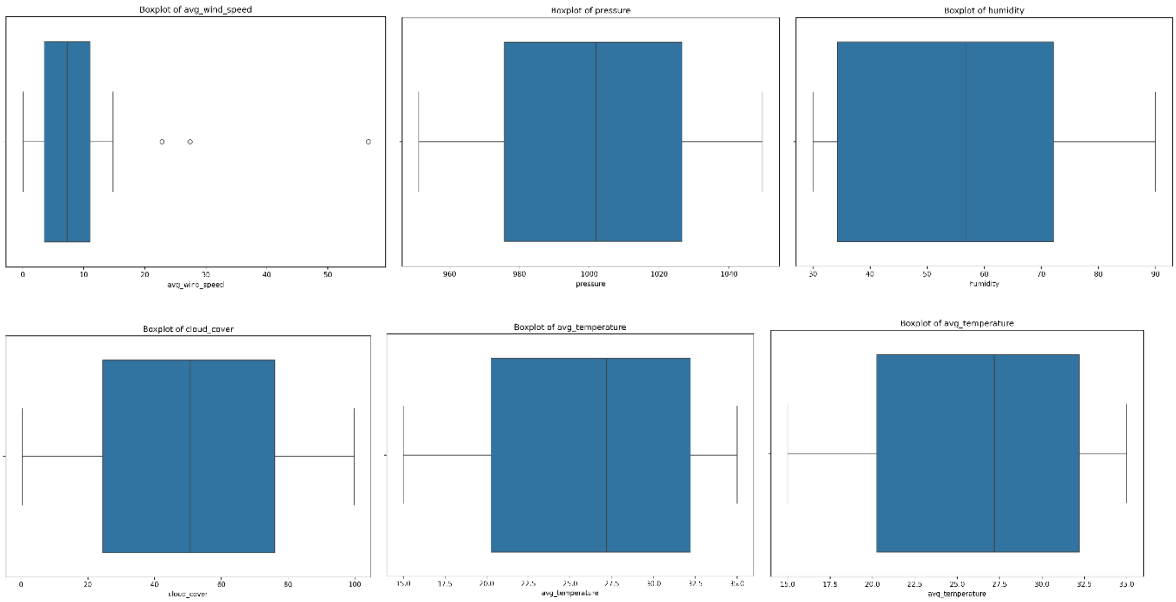


Figure 2 - Box Plots

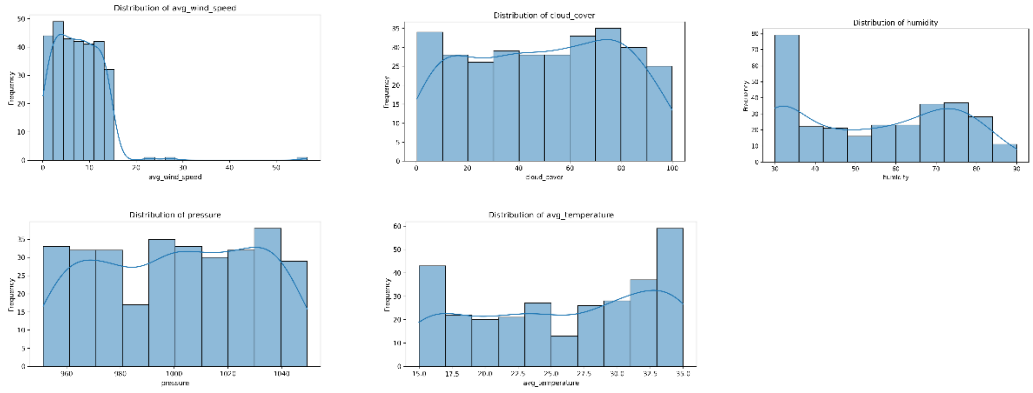
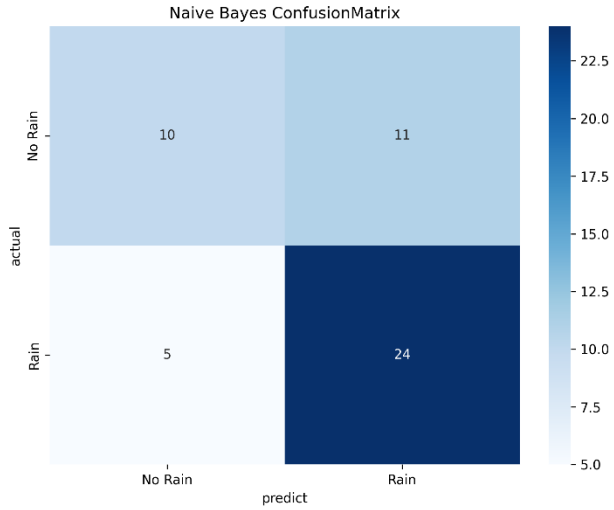
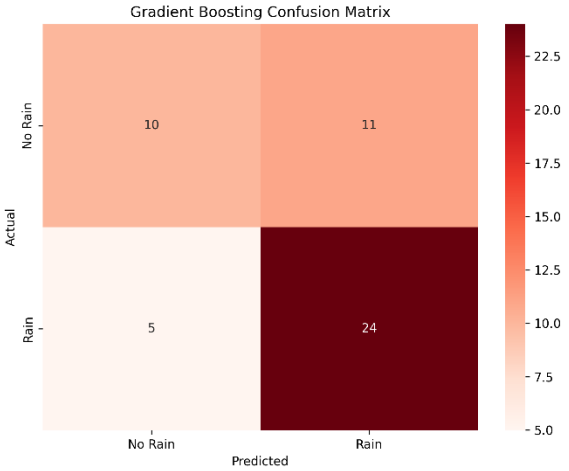
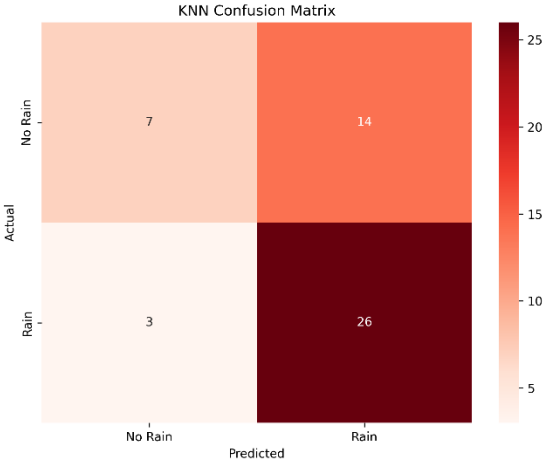


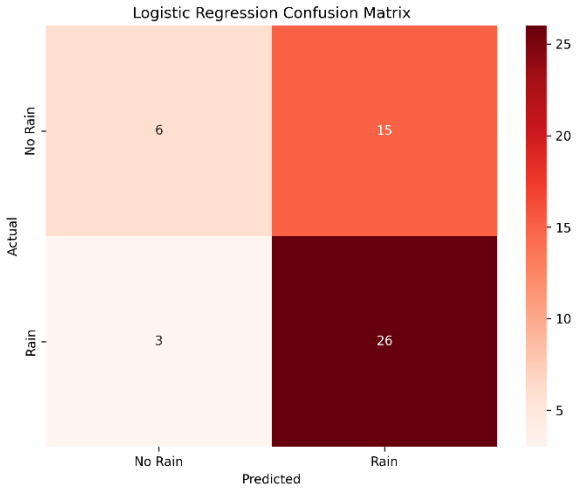
Figure 3 - histrograms

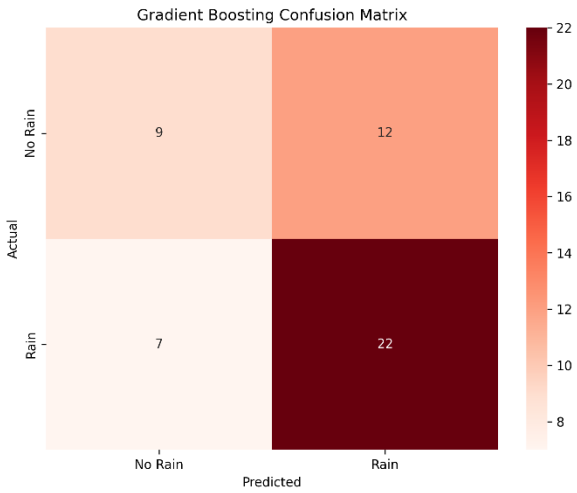
Confusion metrics for models

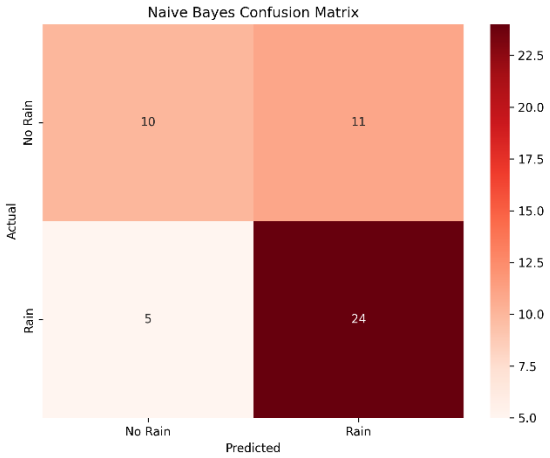


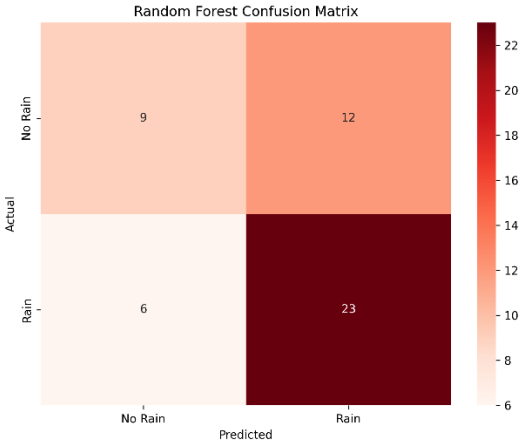


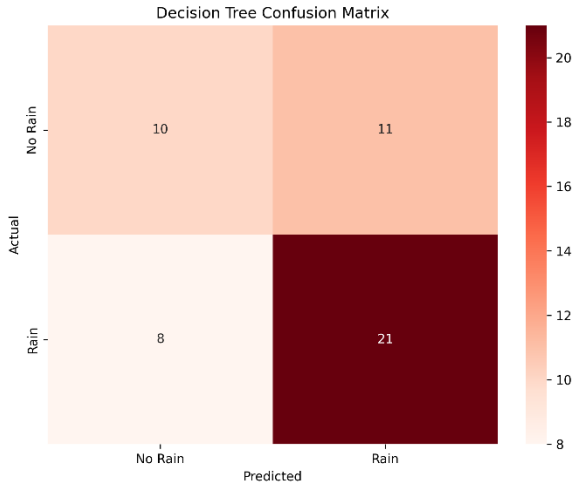

KNN Confusion Matrix
No Rain
Rain
Actual
7
14
3
26
No Rain
Rain
Predicted
25
20
15
10
5

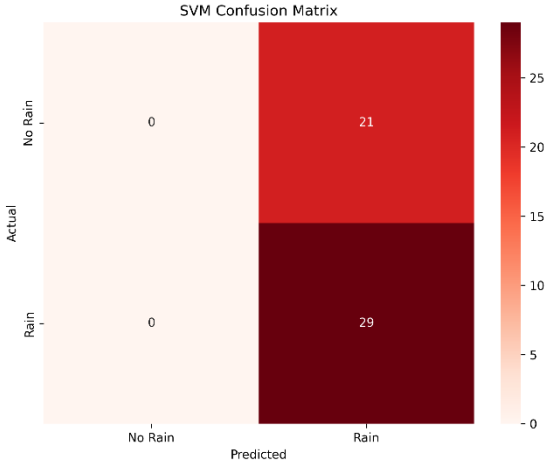

Logistic Regression Confusion Matrix
No Rain
Rain
Actual
6
15
3
26
No Rain
Rain
Predicted
25
20
15
10
5


Gradient Boosting Confusion Matrix
No Rain
Rain
Actual
9
12
7
22
No Rain
Rain
Predicted
22
20
18
16
14
12
10
8


Naive Bayes Confusion Matrix
10
11
5
24
No Rain
Rain
Actual
No Rain
Rain
Predicted
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0


Random Forest Confusion Matrix
9
12
6
23
No Rain
Rain
Actual
No Rain
Rain
Predicted
22
20
18
16
14
12
10
8
6





3. Model training and Hyperparameter Tuning

(See the full analysis in the notebook)

4. Best Model Selection

Based on the evaluation metrics, Naive Bayes (GaussianNB) is the best-performing model for this classification task. It demonstrated strong overall performance, making it the most suitable choice.

Since Naive Bayes effectively balances precision and recall while maintaining reliable accuracy, it outperforms other models in this experiment. Given its ability to handle the classification problem well, it is selected as the best model for this task.

5. Conclusion

Our project focused on creating a machine learning model to forecast rainfall probabilities using historical weather data. The process involved several critical phases:

First, we thoroughly cleaned our dataset by addressing missing values, fixing data errors, and standardizing features to establish a solid foundation for accurate model learning.

Through data analysis, we identified humidity as the strongest rain predictor, with temperature showing moderate correlation. Interestingly, cloud cover and pressure measurements proved less significant for rainfall prediction, informing our feature selection strategy.

We evaluated six different algorithms: Logistic Regression, Random Forest, Bagging, Naive Bayes, K-Nearest Neighbors, and Gradient Boosting. Naive Bayes emerged as the best-performing model, demonstrating strong balance between precision and recall.

To maximize performance, we fine-tuned the Naive Bayes model to ensure optimal results, enhancing its overall effectiveness and prediction stability.

The final Naive Bayes model successfully generates daily rain probability forecasts with high reliability, particularly excelling at identifying potential rain events.