Weather Forecasting for Smart Agriculture using Machine Learning

E.K.K.D.R. Edirisinghe

March 2025

1 Introduction

Accurate weather predictions are important for farmers in many ways. Farmers rely on weather predictions to plan irrigation, planting, and harvesting. Although there are many traditional weather forecasting systems, they are not reliable for hyper-local conditions. This project aims to develop a machine learning model based on historical weather data to predict rain. The final goal is to provide a 21-day rain probability forecast.

2 Dataset

The dataset contains daily weather observations for 300 days, ranging from 2023-01-01 to 2023-11-07. The weather observations include:

- avg_temperature: Average temperature in °C
- humidity: Humidity in percentage
- avg_wind_speed: Average wind speed in km/h
- rain_or_not: Binary label (1 = rain, 0 = no rain)
- cloud_cover
- date: Date of observation

3 Data Preprocessing

3.1 Data Types

'date' was converted to date-time format.

3.2 Duplicates

No duplicates were found

3.3 Handling Missing Values

The missing values in the dataset were as follows.

Variable	Missing Values	Percentage
avg_temperature	15	4.823151
humidity	15	4.823151
avg_wind_speed	15	4.823151
cloud_cover	15	4.823151

Table 1: Missing Value counts and the percentages

Overall, 15 records had null values. All of those records had missing avg_temperature, humidity, avg_wind_speed, and cloud_cover. Since the percentage of the null values was low and the records had all the features missing, I decided to delete those records. (Note that the imputation of the null values decreased the accuracy of the models.)

4 Methodology

4.1 Feature Creation

Since this is a time series, many features could be obtained using the 'date'. 'year', 'month', 'day', 'day_of_week', 'week_of_year', and 'quarter' features were created using the 'date'.

4.2 Exploratory Data Analysis (EDA)

4.2.1 Visualizations

The distribution of the target variable ('rain_or_not') was plotted using a count plot.
'Rain' and 'No Rain' are not balanced, but the difference is not much of a concern.

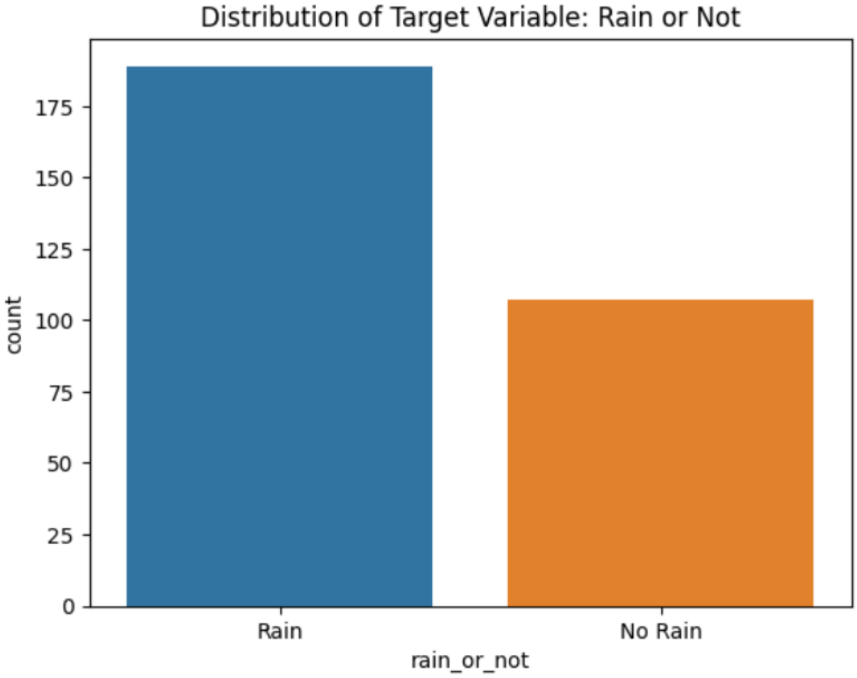


Figure 1: rain_or_not count plot distribution

'avg_temperature', 'avg_wind_speed', 'cloud_cover', 'pressure', 'month', and 'week_of_year' were plotted against 'rain_or_not' using scatter plots.

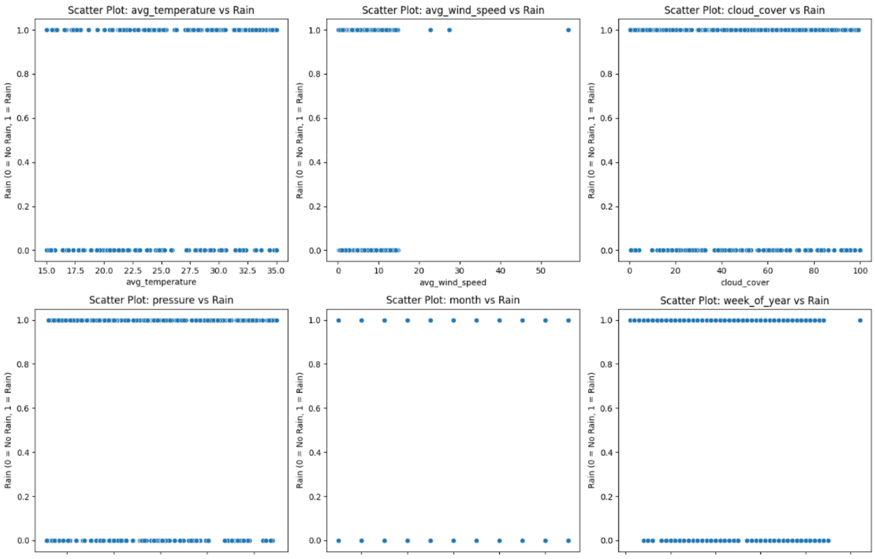
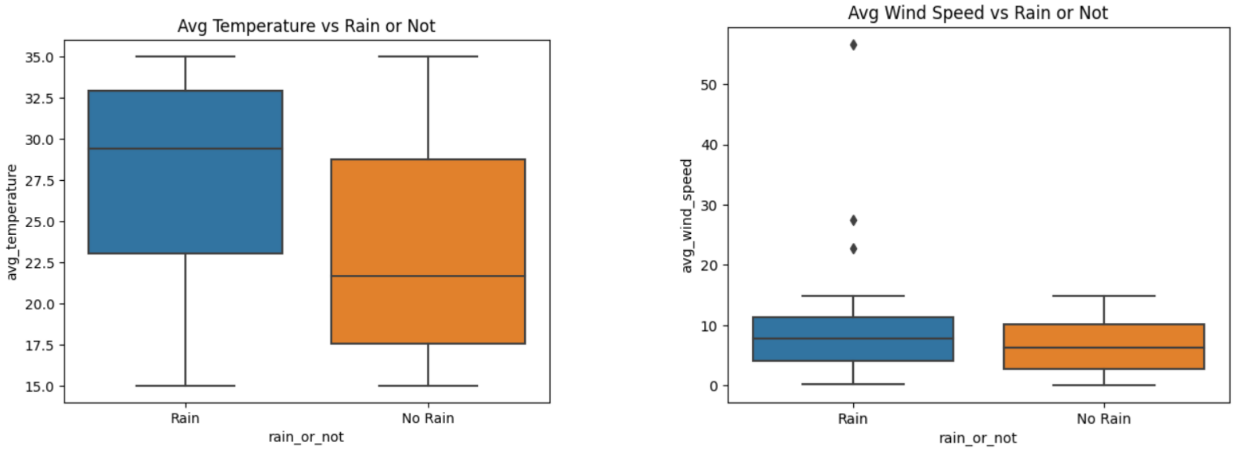


Figure 2: Scatter plots

Conclusions from the scatter plots:

- Cloud cover and pressure seem to have the strongest association with rainfall.
- Temperature shows some trend. Lower values have a higher probability of rain. Wind speed, month, and week of year may have some influence. But the relationship isn't clear from the scatter plots alone.

Box plots were also plotted for 'avg_temperature', 'avg_wind_speed', and 'humidity'.



(a) Average temperature vs rain

(b) Average wind speed vs rain

Figure 3: Box plots for avg_temperature and avg_wind_speed vs rain

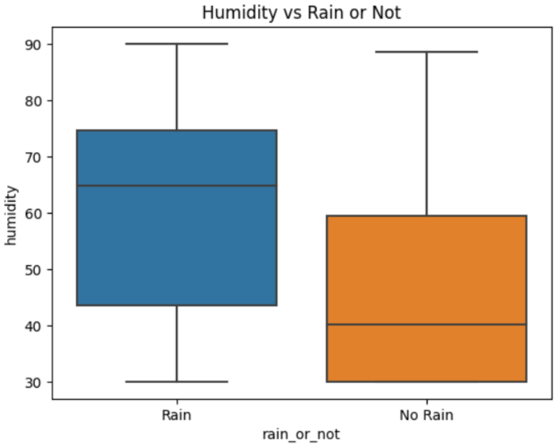
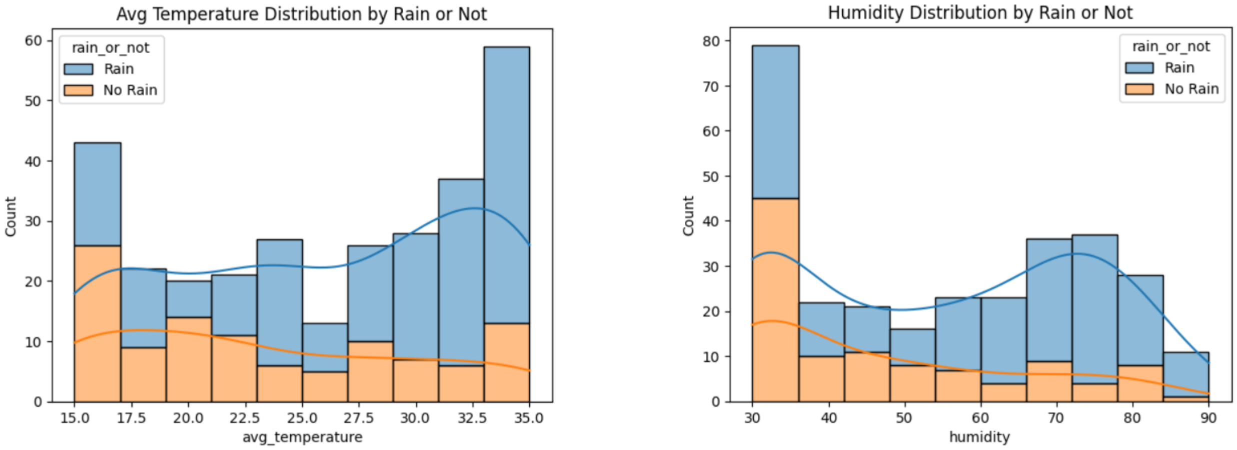


Figure 4: Humidity vs rain

It seems like 'avg_wind_speed' is not a differentiating factor between rain and no rain. However, the outliers might suggest that very high winds can predict rain. Therefore, it can not be neglected.

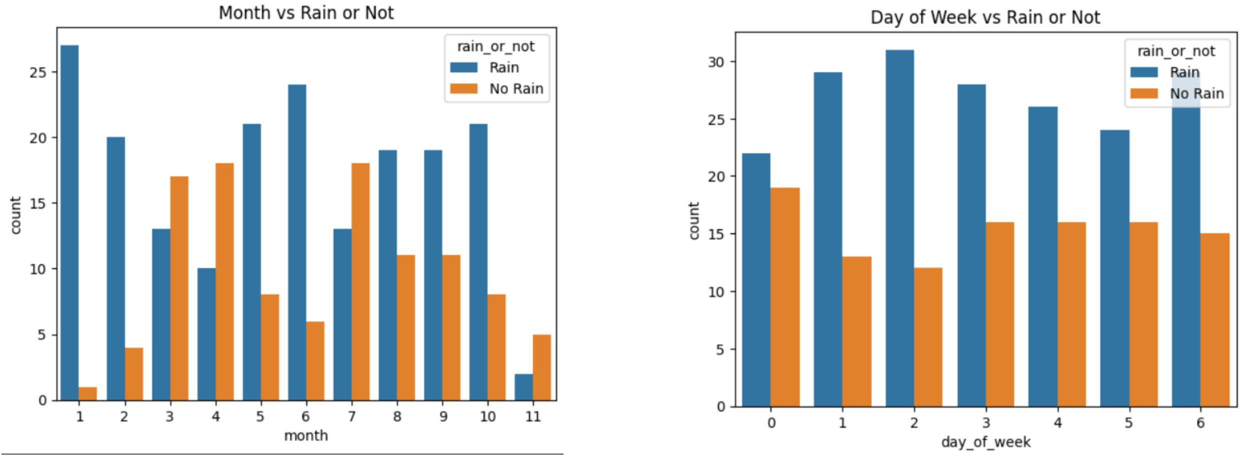
More histograms and count plots were also plotted as follows for further analysis.



(a) Average temperature vs rain

(b) Humidity vs rain

Figure 5: Histograms for avg_temperature and humidity vs rain



(a) Month vs rain

(b) Day of week vs rain

Figure 6: Count plots for Month and Day of week vs rain

4.2.2 Numerical Analysis

The correlation matrix was computed for the numerical features ('avg_temperature', 'avg_wind_speed', 'humidity', 'cloud_cover', and 'pressure'), and the correlation heatmap was plotted.
The correlation matrix suggested that average temperature and humidity were correlated with each other. Therefore, it was necessary to either remove a feature or create a new feature using them.

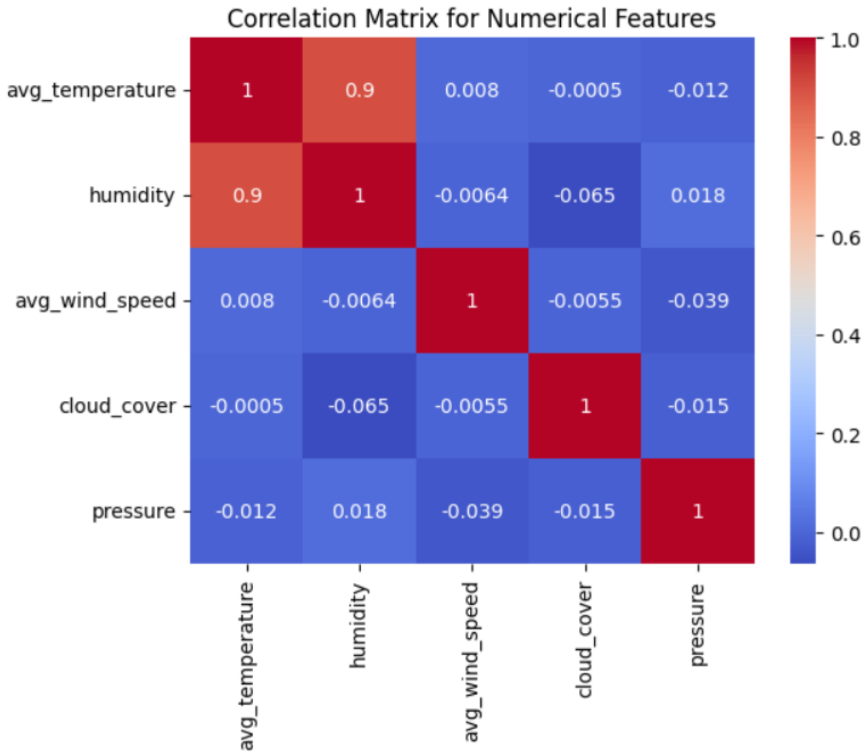


Figure 7: Correlation Matrix

The chi-square test was carried out for the categorical features ('rain_or_not', 'month', 'day_of_week', 'week_of_year', 'day', and 'quarter')
The results from the chi-square test were as follows.

Variable	**p-value**	**Analysis**
Month	8.1168e-07	Significant relationship between month and rain_or_not.
Day of Week	0.6753	No significant relationship between day_of_week and rain_or_not.
Week of Year	0.0011	Significant relationship between week_of_year and rain_or_not.
Day	0.5767	No significant relationship between day and rain_or_not.
Quarter	0.1384	No significant relationship between quarter and rain_or_not.

Table 2: Chi-Square Test Results

4.3 Feature Selection

The 'year' was common to all records. Therefore, it was removed. The results from the visualizations and the numerical analysis suggested 'day', 'day_of_week', and 'quarter' were not useful for the model. Hence, these features were also dropped. 'date' was also not useful as the required features were extracted from it. Although the correlation matrix suggested that 'humidity' and 'avg_temperature' are correlated with each other, both were used for the modeling at first, as both features are useful for weather prediction.

4.4 Binning

Binning was carried out for the 'avg_temperature' and 'wind_speed'.

4.5 Scaling and Standardization

'week_of_year' was scaled using min max scaling(or normalization).
'avg_wind_speed', 'cloud_cover', and 'pressure' were also scaled using standard scaling (Z-score normalization).

4.6 Encoding

'avg_wind_speed' and 'pressure' had extreme values. Therefore, log transformation (or logarithmic scaling) was carried out for those two features.
To capture the cyclic nature of 'week_of_year' and 'month', they were encoded using sin-cos encoding.

4.7 Model Evaluation and Comparison

Few models were used for this project, namely, Logistic Regression, Random Forest, XGB Classifier, Support Vector Machine (SVM), Gradient Boosting, and LightGBM Classifier. Each model was tested using various features, scaling, and standardization.
For model evaluation, test size was taken as 0.2
Model performance was analyzed as follows:

- With 'month' and 'day'
- Without 'humidity'
- With 'week_of_year'
- After sin-cos encoding of the 'month' and 'week_of_year'
- After standardization of 'avg_wind_speed', 'cloud_cover', and 'pressure'
- Without 'avg_wind_speed'

4.7.1 Results of the Model Testing

Table 3: Comparison of Model Performance

Model*	Evaluation Method	Precision (Rain)	Recall (Rain)	F1 (Rain)	Accuracy
LR	'month' + 'day'	0.76	0.52	0.62	0.55
LR	Without 'humidity'	0.78	0.60	0.68	0.60
LR	With 'week_of_year'	0.80	0.57	0.67	0.60
LR	After sin-cos encoding	0.70	0.50	0.58	0.50
LR	After Standardizing	0.80	0.57	0.67	0.60
LR	Without 'avg_wind_speed'	0.79	0.52	0.63	0.57
RF	'month' + 'day'	0.82	0.74	0.78	0.70
RF	Without 'humidity'	0.71	0.71	0.71	0.60
RF	With 'week_of_year'	0.74	0.69	0.72	0.62
RF	After sin-cos encoding	0.77	0.71	0.74	0.65
RF	After Standardizing	0.74	0.74	0.74	0.63
RF	Without 'avg_wind_speed'	0.70	0.74	0.72	0.60
XGB	'month' + 'day'	0.79	0.71	0.75	0.67
XGB	Without 'humidity'	0.81	0.71	0.76	0.68
XGB	With 'week_of_year'	0.76	0.76	0.76	0.67
XGB	After sin-cos encoding	0.77	0.71	0.74	0.65
XGB	After Standardizing	0.76	0.76	0.76	0.67
XGB	Without 'avg_wind_speed'	0.75	0.71	0.73	0.63
GB	'month' + 'day'	0.79	0.62	0.69	0.62
GB	Without 'humidity'	0.79	0.74	0.77	0.68
GB	With 'week_of_year'	0.79	0.79	0.79	0.70
GB	After sin-cos encoding	0.79	0.74	0.77	0.68
GB	After Standardizing	0.79	0.79	0.79	0.70
GB	Without 'avg_wind_speed'	0.75	0.71	0.73	0.63
LGBM	'month' + 'day'	0.78	0.70	0.74	0.66
LGBM	Without 'humidity'	0.77	0.77	0.77	0.69
LGBM	With 'week_of_year'	0.79	0.79	0.79	0.71
LGBM	After sin-cos encoding	0.73	0.72	0.73	0.63
LGBM	After Standardizing	0.77	0.72	0.75	0.66
LGBM	Without 'avg_wind_speed'	0.75	0.67	0.71	0.62

* LR - Logistic Regression, RF - Random Forest Classifier, XGB - XGB Classifier, GB - Gradient Boosting, LGBM - LightGBM Classifier The results of the SVM model is not shown here since it showed zero for precision, recall, and F1-score with 'No Rain'.

4.7.2 Analysis of the Results

- Although in most cases the precision (Rain), recall (Rain), and F1 (Rain) were high, the values for 'No Rain' were very low. Only LGBM and GB had a more balanced result for both 'Rain' and 'No Rain'.

- In most cases, standardization decreased the performance. Features like 'month', 'day', and 'quartile' did not have any effect on the performance. 'humidity' decreased the performance of models like Gradient Boosting and LightGBM. Although the EDA suggested 'avg_wind_speed' does not affect the final predictions, removing it caused poor performance across many models.

- LightGBM and Gradient Boosting consistently performed better.

- Random Forest had a decent performance with a maximum accuracy.

- XGBoost showed moderate results, but it lagged behind LightGBM and Gradient Boosting in most cases.
- Logistic Regression had lower performance compared to tree-based models.
- SVM was neglected due to its poor performance across all metrics.
- LightBGM and Gradient Boosting were used as the final models for further testing. Binning and keeping 'humidity' reduced the accuracy of both models. Therefore, 'humidity' was removed. Since standardization, sin-cos encoding also decreased the performance, they were also not carried out for these two models.

4.8 Feature Engineering

A new feature was created by multiplying 'humidity' and 'avg_temperature' with each other. But this reduced the performance of the models.

4.9 Cross Validation

Both models were cross validated using 5 folds. Cross validation scores for the two models using Stratified K-Fold was as follows:

4.9.1 Cross-Validation Results

	LightGBM	**Gradient Boosting**
Cross-validation scores	[0.6167, 0.5424, 0.5254, 0.6441, 0.7119]	[0.6545, 0.4909, 0.6000, 0.4545, 0.6545]
Average cross-validation score	0.6081	0.5709
Standard deviation	0.0682	0.0834

Table 4: Cross-validation results for LightGBM and Gradient Boosting

4.10 Hyperparameter Tuning

LightGBM and Gradient Boosting were subjected to RandomizedSearchcv and Grid Search. LightGBM had better results in hyperparameter tuning, but it may have been subject to overfitting due to the lower number of test records

4.10.1 Hyperparameter Tuning Results

Method	**LightGBM**	**Gradient Boosting**
Randomized Search Best Hyperparameters		
num_leaves	31	-
min_data_in_leaf	100	-
max_depth	-1	7
learning_rate	0.05	0.3158
lambda_l2	0	-
lambda_l1	0.1	-
feature_fraction	0.8	-
bagging_freq	5	-
bagging_fraction	0.8	-
max_features	-	None
min_samples_leaf	-	17
min_samples_split	-	11
n_estimators	-	64
subsample	-	0.8368
Grid Search Best Hyperparameters		
learning_rate	0.3	-
max_depth	-1	-
n_estimators	100	50
num_leaves	31	-
max_features	-	sqrt
subsample	-	0.8

Table 5: Best hyperparameters found using Randomized Search and Grid Search

4.11 Final Prediction

The models were tested for the future 21 days.
The final predictions of LightGBM and Gradient Boosting are as follows:

4.11.1 Prediction Results

Date	LGBM Prediction	GB Prediction	Actual
2023-10-18	Rain	Rain	Rain
2023-10-19	Rain	Rain	Rain
2023-10-20	Rain	Rain	Rain
2023-10-21	Rain	Rain	Rain
2023-10-22	Rain	Rain	No Rain
2023-10-23	Rain	Rain	Rain
2023-10-24	No Rain	Rain	Rain
2023-10-25	Rain	Rain	Rain
2023-10-26	No Rain	Rain	No Rain
2023-10-27	No Rain	Rain	Rain
2023-10-28	Rain	Rain	Rain
2023-10-29	Rain	No Rain	No Rain
2023-10-30	No Rain	No Rain	Rain
2023-10-31	Rain	Rain	No Rain
2023-11-01	No Rain	Rain	Rain
2023-11-02	No Rain	No Rain	No Rain
2023-11-03	No Rain	No Rain	No Rain
2023-11-04	Rain	Rain	No Rain
2023-11-05	Rain	Rain	Rain
2023-11-06	No Rain	Rain	No Rain
2023-11-07	No Rain	No Rain	No Rain

Table 6: Predicted vs Actual Values for LightGBM and Gradient Boosting

4.11.2 Classification Reports

Class	Precision	Recall	F1-score	Support
No Rain (0)	0.56	0.56	0.56	9
Rain (1)	0.67	0.67	0.67	12
Accuracy	0.62			21

Table 7: LightGBM Classification Report

Class	Precision	Recall	F1-score	Support
No Rain (0)	0.80	0.44	0.57	9
Rain (1)	0.69	0.92	0.79	12
Accuracy	0.71			21

Table 8: Gradient Boosting Classification Report

5 Conclusion

Gradient Boosting performs better overall than LightGBM, particularly in identifying rain with a higher accuracy and recall. However, there is still room for improvement in predicting No Rain events. Both models struggle to identify No Rain with high recall. The F1-score for Rain is higher in the Gradient Boosting model. Therefore, it has a better balance between precision and recall for the prediction of rain.

Therefore, Gradient Boosting is the more accurate model for this task. Further improvements can be made by tuning the model for better performance on the No Rain class and using a large training dataset.