FLIGHT DELAY PREDICTION

Abstract: Flight delay is vexatious for passengers and incurs an agonizingly high financial loss to airlines and countries. A structured prediction system is an indispensable tool that can help aviation authorities effectively alleviate flight delays. This project aims to build a two-stage machine learning engine to effectively predict the arrival delay of a flight after departure based on real-time flight and weather data.

GROUP MEMBERS:

1) Siddhesh Pisal : 201080076

2) AliRaza Malik : 201080078

Aim: The aim of this project is to develop a machine learning model that accurately predicts flight delays. The project seeks to leverage historical flight data and relevant variables to create a predictive model that can assist airlines, passengers, and other stakeholders in anticipating and managing flight delays more effectively. By accurately predicting flight delays, the project aims to improve customer satisfaction, optimize airline operations, and enhance overall efficiency in the aviation industry.

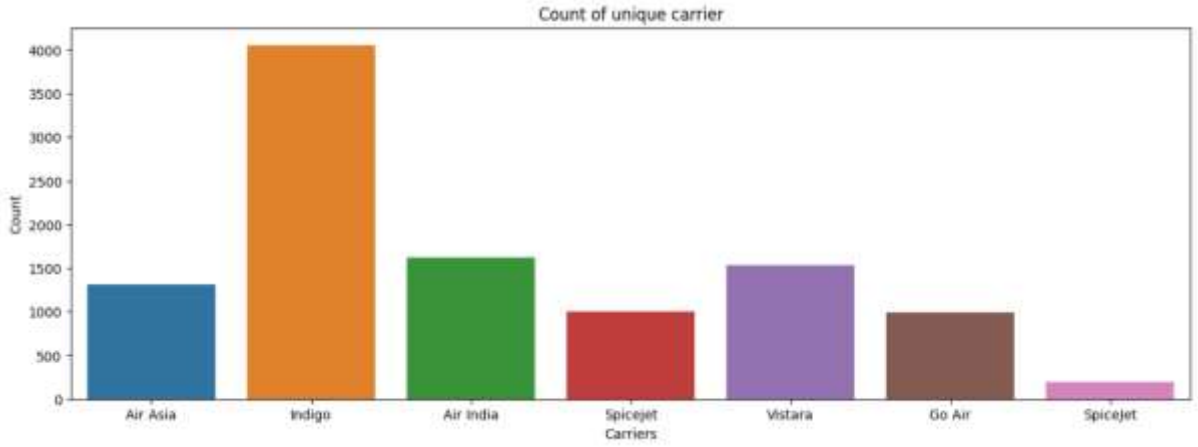
Problem Statement:

The problem addressed in this project is to develop a hybrid flight delay prediction system that combines classification and regression models to improve the accuracy and efficiency of flight delay predictions. The objective is to create a two-step approach where a classification model is applied first to determine whether a flight is likely to experience a delay, and if the classification model indicates a delay, a regression model is then used to estimate the duration of the delay.

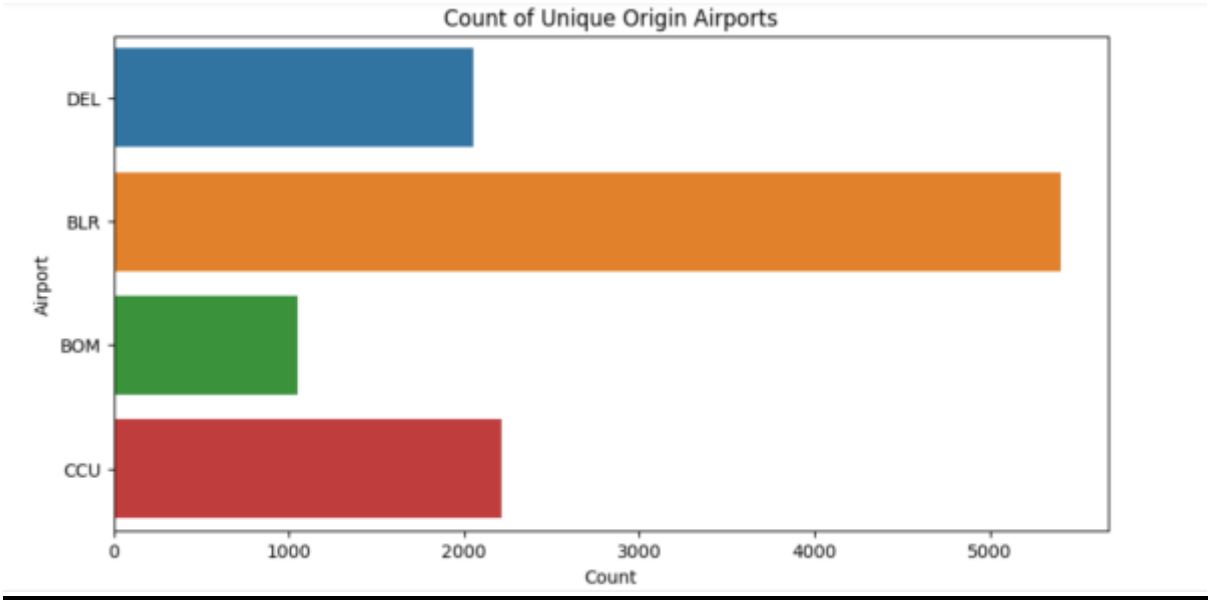
Graphs:

1. Carriers vs Counts:

a graph was plotted to analyze the distribution of flight counts across different carriers. This analysis provides insights into the frequency and distribution of flights operated by each carrier, which can be helpful in understanding the airline landscape and identifying potential patterns or trends.

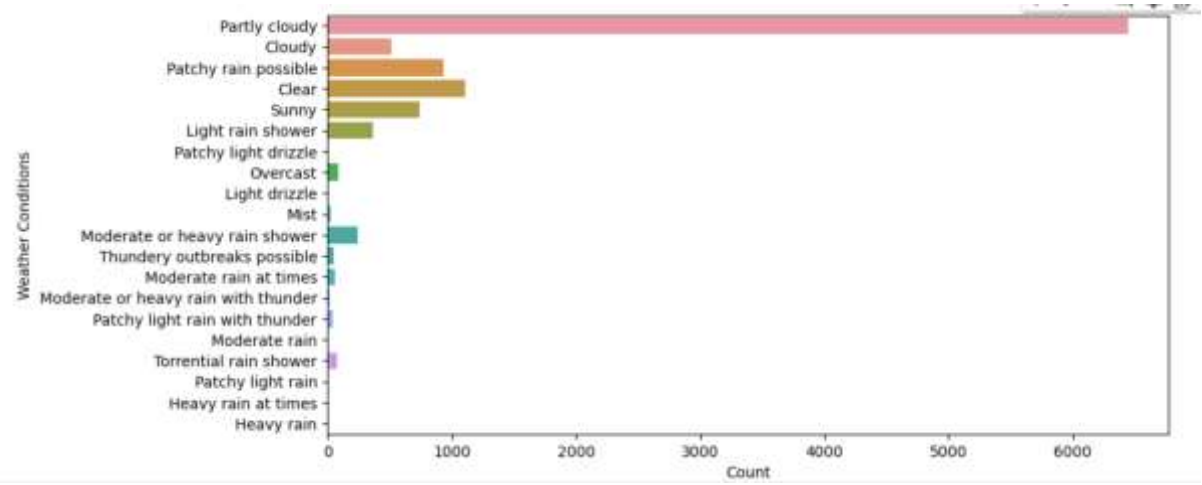


2. Count of Unique Origin Airports:



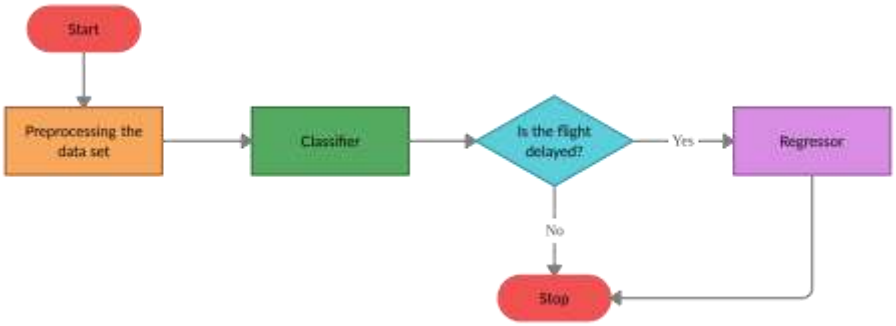
3. Weather Graph:

a count plot was created to visualize the distribution of different weather conditions in the dataset. The purpose of this graph is to gain insights into the frequency of various weather conditions recorded during the flights.



WORKING:

The flow chart depicted in Figure below represents the two-stage flight delay prediction machine learning engine. The pipelined model involves chaining the best performing classifier before the best regressor. The data was preprocessed and a model was trained to perform classification using the Random Forest Classifier. The Random Forest Classifier is chosen as it has the maximum F1 Score (0.78) and area under ROC (0.85). The flight delay needs to be calculated only for the flights that will be delayed. Thus, only those data points that were predicted to be delayed by the classifier are selected to perform regression and predict the flight arrival delay in minutes. The Random Forest Regressor is chosen as it has the highest R-squared score (0.89) and lower values of RMSE (38.22) and MAE (11.5076).



Results:

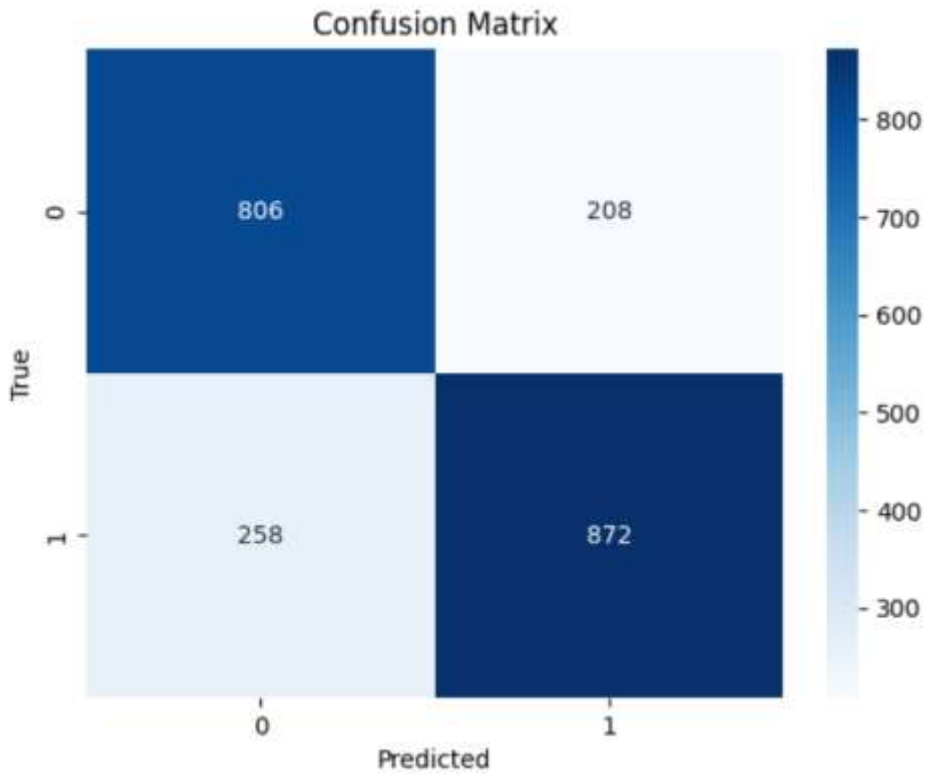
CLASSIFICATION:

1. Logistic Regression:

- The logistic regression model shows promising performance in predicting flight delays, with relatively high precision, recall, and F1-scores for both classes.
- The balanced F1-scores suggest that the model achieves a reasonable trade-off between identifying flights without delays and flights with delays.
- The accuracy of 78% indicates that the model provides reasonably accurate predictions overall, but it's essential to consider the specific goals and requirements of the flight delay prediction task.
- Further analysis and evaluation may be required to assess the model's performance on specific subsets of data or to compare it with other classification algorithms.

```
y_pred_log = logisticRegr.predict(X_test_scaled)
print(classification_report(y_test, y_pred_log))
```

```
              precision    recall  f1-score   support

           0       0.76      0.79      0.78      1014
           1       0.81      0.77      0.79      1130

    accuracy                           0.78      2144
   macro avg       0.78      0.78      0.78      2144
weighted avg       0.78      0.78      0.78      2144
```

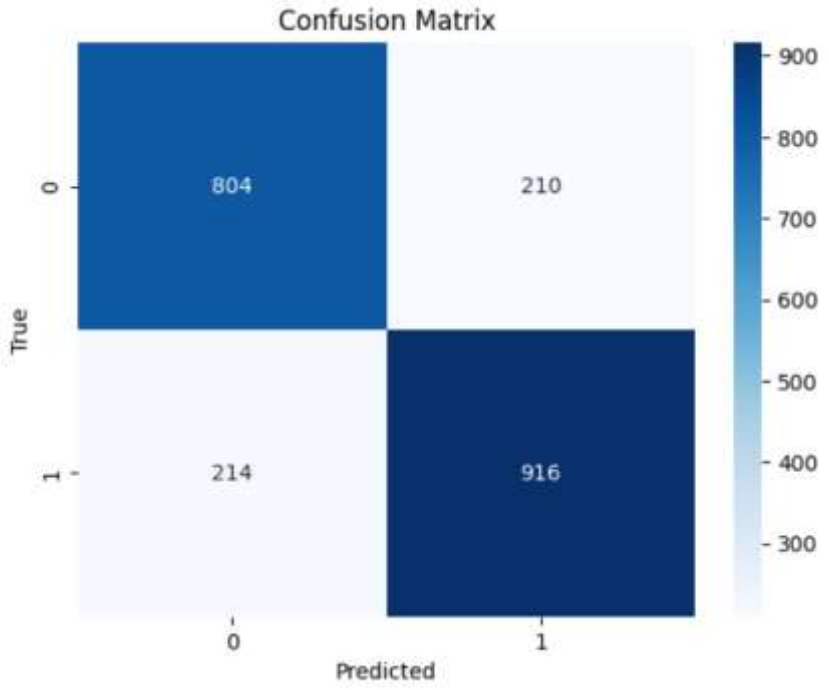


2. SVM Classification:

- The SVM classification model shows promising performance in predicting flight delays, with reasonably high precision, recall, and F1-scores for both classes.
- Both models demonstrate similar precision, recall, and F1-scores for both classes, indicating comparable performance in identifying flight delays.
- However, the SVM model achieves a slightly higher accuracy of 80% compared to the logistic regression model's accuracy of 78%.
- The marginal improvement in accuracy suggests that the SVM model may better capture the underlying patterns in the flight data, resulting in slightly more accurate predictions.

```
[ ]  y_pred_svm = clf.predict(X_test_scaled)
     print(classification_report(y_test, y_pred_svm))
```

```
              precision    recall  f1-score   support

           0       0.79      0.79      0.79      1014
           1       0.81      0.81      0.81      1130

    accuracy                           0.80      2144
   macro avg       0.80      0.80      0.80      2144
weighted avg       0.80      0.80      0.80      2144
```

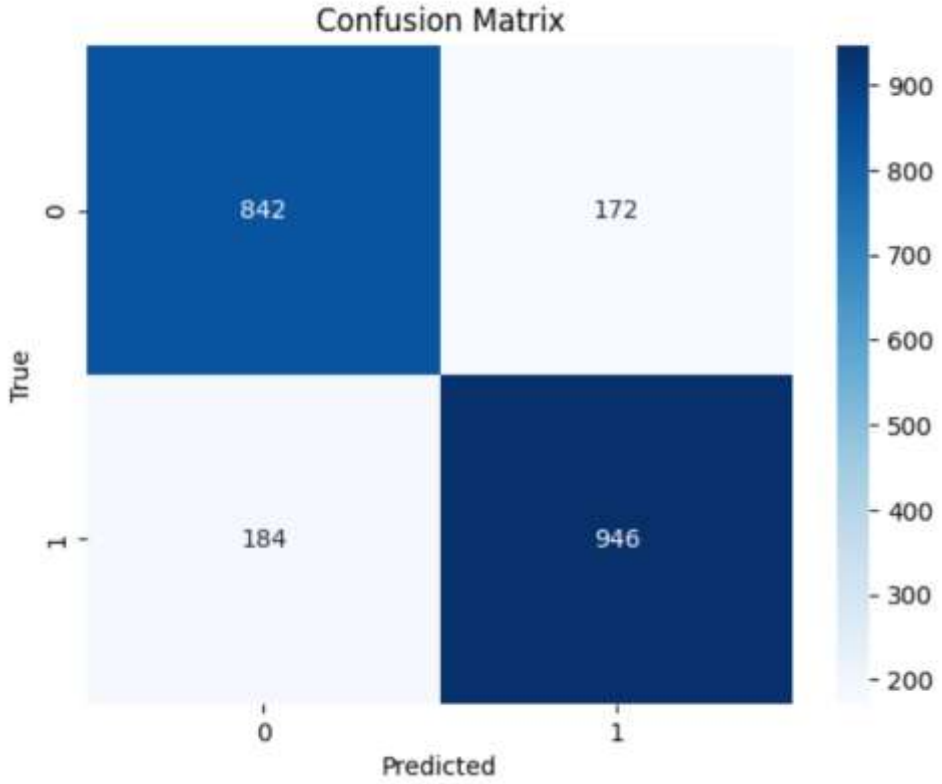


3. Random Forest:

- The Random Forest model outperforms both the logistic regression and SVM classification models in terms of precision, recall, F1-score, and accuracy for both classes.
- It achieves the highest precision, recall, and F1-scores, indicating a better ability to correctly identify flights without delays and flights with delays.
- The higher accuracy of 83% suggests that the Random Forest model captures the underlying patterns in the flight data more effectively, resulting in more accurate predictions.

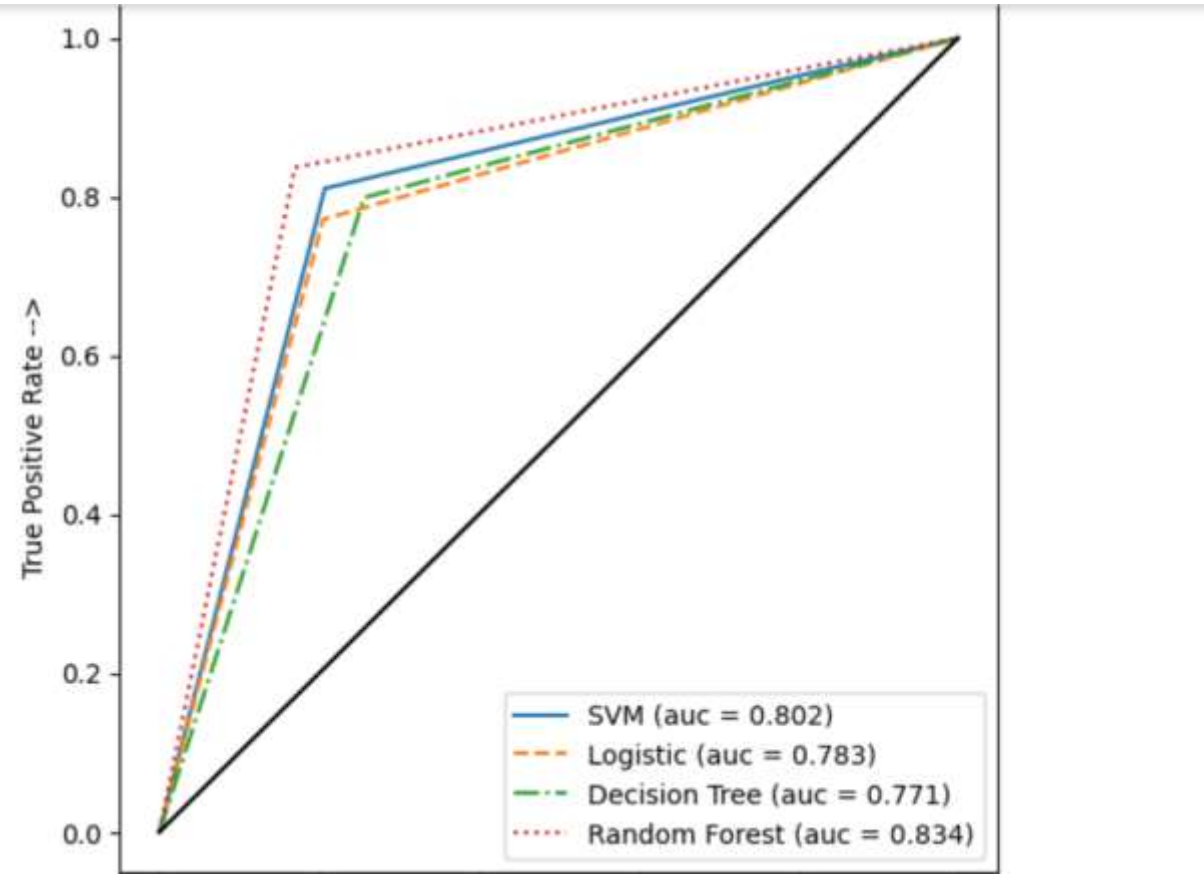
```
y_pred_forest = rf.predict(X_test_scaled)
print(classification_report(y_test, y_pred_forest))
```

```
              precision    recall  f1-score   support

           0       0.82      0.83      0.83      1014
           1       0.85      0.84      0.84      1130

    accuracy                           0.83      2144
   macro avg       0.83      0.83      0.83      2144
weighted avg       0.83      0.83      0.83      2144
```



ROC Curve:

AUC Value	**Inference**
AUC = 0	The classifier is predicting all negatives as positives, and all positives as negatives
AUC = 0.5	The classifier is unable to distinguish the positive and negative class pointsthereby predicting a random or constant class for all the data points
Between 0.5 and 1	There is a high chance that the classifier will be able to distinguishbetween the two classes
AUC = 1	The classifier is able to perfectly distinguish between all the positive andthe negative class points correctly



We can see that the Random Forest has highest AUC in the above graph and performed better than other models such as Decision Trees, SVM and Logistic.

Gradient Boosting and Random Forest yield the same accuracy score of around 0.834 However, the Random Forest Classifier is chosen as the best model as it has the highest area under ROC

REGRESSION:

1. Linear Regression:

The Linear Regression model shows relatively good performance in predicting flight delay durations, as indicated by the evaluation metrics.

The MSE of 2135.57 suggests that, on average, the predicted flight delay durations have a squared difference of approximately 2135.57 units from the actual durations.

The RMSE of 46.21 indicates that, on average, the model's predictions deviate by approximately 46.21 units from the actual durations. This suggests a reasonable level of accuracy in the predictions.

The MAE of 13.33 indicates that, on average, the model's predictions have an absolute difference of approximately 13.33 units from the actual durations.

The R2 score of 0.8419 implies that the linear regression model explains approximately 84.19% of the variance in the flight delay durations, indicating a good fit to the data.

```
MSE      : 2135.5693837084063

RMSE     : 46.21222115099432

MAE      : 13.330370558606768

R2 Score : 0.8418697177901076
```

2. Decision Tree:

The Linear Regression model outperforms the Decision Tree Regressor in terms of MSE, RMSE, MAE, and R2 score. It generally provides more accurate predictions and better fits the flight delay data.

The Decision Tree Regressor, however, still demonstrates reasonable performance with an R2 score of 0.8031 and can be considered as an alternative regression model for predicting flight delay durations.

```
MSE      : 2658.5452425373132

RMSE     : 51.56108263542682

MAE      : 15.193563432835822

R2 Score : 0.8031454689895507
```

3. Random Forest:

The Random Forest model outperforms both the Decision Tree Regressor and Linear Regression models in terms of MSE, RMSE, MAE, and R2 score. It consistently provides more accurate predictions and better fits the flight delay data.

The Random Forest model demonstrates superior performance in capturing the underlying patterns and relationships in the flight data, resulting in more accurate predictions of flight delay durations.

The Random Forest model's ability to leverage ensemble techniques and handle complex interactions between features contributes to its improved performance compared to the other models.

```
MSE      : 1461.2160489959747

RMSE     : 38.22585576538444

MAE      : 11.507815642768302

R2 Score : 0.8918028569054879
```

In summary, Random Forest Regression and Gradient Boosting Regression outperform both Linear Regression and Decision Tree Regression in terms of MSE, RMSE, MAE, and R2 score. They exhibit smaller prediction errors and better fit to the data, suggesting their suitability for predicting flight delay durations. Among the two top-performing models, Random Forest Regression achieves slightly lower MSE, RMSE, and MAE values and a marginally higher R2 score, indicating its potential as the most accurate and reliable model for this particular task.

Regressors	MSE	RMSE	MAE	R2
LinearRegression	2135.569384	46.212221	13.330371	0.841870
DecisionRegression	2658.545243	51.561083	15.193563	0.803145
Random Forest Regression	1461.216049	38.225856	11.507816	0.891803
GradientBoostingRegressor	1518.654716	38.969921	12.393579	0.887550

Conclusion:

The flight and weather data were combined into a single dataset and preprocessed to train a two-stage machine learning model that predicts flight arrival delay. Out of several classification algorithms, the Random Forest classifier gave the best F1 score (0.84) and Recall (0.84) for the delayed flights. Subsequently, the Random Forest regressor was pipelined, giving MAE 11.5078 minutes and RMSE 38.225 minutes with an R-squared score of 0.8919. In conclusion, the flight delay prediction was efficient and the Machine Learning model exhibited good performance.