# Predicting Flight Delays - An Exploratory Data Analysis
# CS-675 Group Project

## Project Aim and Overview:

The project's goal was to create models to predict flight delays, a vital task for both airlines and passengers. Airlines can optimize their operations with accurate predictions, and passengers can improve their travel experience by anticipating delays.

A 2008 flight data set, comprising various flight-related features, was used. Three machine learning models: Logistic Regression, Decision Tree, and Deep Neural Networks, were employed to achieve the project's goal.


Each model was trained, tested, and evaluated based on its accuracy, precision, recall, and F1-score. The comparison of these models helped identify the most effective approach for flight delay prediction.

## Data Requirements and Description:

The dataset used for this project was sourced from Kaggle, a popular online community for data scientists and machine learning practitioners. The specific dataset, titled "Airline Delay Causes," is publicly available and can be accessed here. https://www.kaggle.com/datasets/giovamata/airlinedelaycauses/code

The dataset includes flight information from the year 2008, encompassing over 1.9 million records, which provides a substantial amount of data for training and testing our machine learning models.

Each record in the dataset represents a single flight, with 29 variables providing information about each flight. These variables include:

- Flight number: a unique identifier for each flight.
- Carrier: an identifier for the airline operating the flight.

- Departure and arrival delays: the number of minutes the flight was delayed for departure and arrival, respectively. These are our target variables for predicting flight delays.
- Other variables: additional variables provide further information about each flight, such as the date of the flight, scheduled and actual elapsed time, tail number, origin and destination airports, and the distance of the flight.

## Data Processing:

The data processing was carried out using Python and its data handling libraries like pandas and numpy. The dataset was cleaned, missing values were treated, and categorical variables were encoded before modeling.

**Data Loading and Cleaning:**

The project began by loading the required Python libraries (pandas, numpy, matplotlib, seaborn, and warnings) to handle data manipulation, analysis, and visualization tasks.

The flight dataset was loaded into a pandas DataFrame using the pd.read_csv() function.

Upon inspecting the data, it was noted that there was a superfluous column ('Unnamed: 0'), which was promptly dropped. Further data exploration was conducted to understand the shape of the dataset and to identify any missing values.

Certain rows with missing values in important columns were dropped to maintain data integrity. Additionally, some columns were deemed unnecessary for the analysis and were consequently removed. These include 'Year', 'Month', 'DayofMonth', 'SecurityDelay', 'ArrTime', and 'CancellationCode'.

**Correlation Analysis:**

To gain insights into the relationships between different variables, a correlation analysis was performed using the corr() function. This analysis helps in understanding the linear relationships between the features, which can be crucial in the feature selection process for machine learning models.

| | DayOfWeek | DepTime | CRSDepTime | CRSArrTime | UniqueCarrier | FlightNum | TailNum | ActualElapsedTime | CRSElapsedTime | AirTime | ... | TaxiOut | Cancelled | Diverted | CarrierDelay | WeatherDelay | NASDelay | LateAircraftDelay | TotalDelay | DepHour | TimeOfDay |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 4 | 1829.0 | 1755 | 1925.0 | WN | 3920.0 | N464WN | 90.0 | 90.0 | 77.0 | ... | 10.0 | 0.0 | 0.0 | 2.0 | 0.0 | 0.0 | 32.0 | 68.0 | 17 | Afternoon |
| 5 | 4 | 1937.0 | 1830 | 1940.0 | WN | 509.0 | N763SW | 240.0 | 250.0 | 230.0 | ... | 7.0 | 0.0 | 0.0 | 10.0 | 0.0 | 0.0 | 47.0 | 124.0 | 18 | Evening |
| 7 | 4 | 1644.0 | 1510 | 1725.0 | WN | 1333.0 | N334SW | 121.0 | 135.0 | 107.0 | ... | 8.0 | 0.0 | 0.0 | 8.0 | 0.0 | 0.0 | 72.0 | 174.0 | 15 | Afternoon |
| 9 | 4 | 1452.0 | 1425 | 1625.0 | WN | 675.0 | N286WN | 228.0 | 240.0 | 213.0 | ... | 8.0 | 0.0 | 0.0 | 3.0 | 0.0 | 0.0 | 12.0 | 42.0 | 14 | Afternoon |
| 11 | 4 | 1323.0 | 1255 | 1510.0 | WN | 4.0 | N674AA | 123.0 | 135.0 | 110.0 | ... | 9.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 16.0 | 44.0 | 12 | Afternoon |

5 rows × 26 columns

# Predicting Flight Delays - An Exploratory Data Analysis
# CS-675 Group Project

## Algorithm Description:

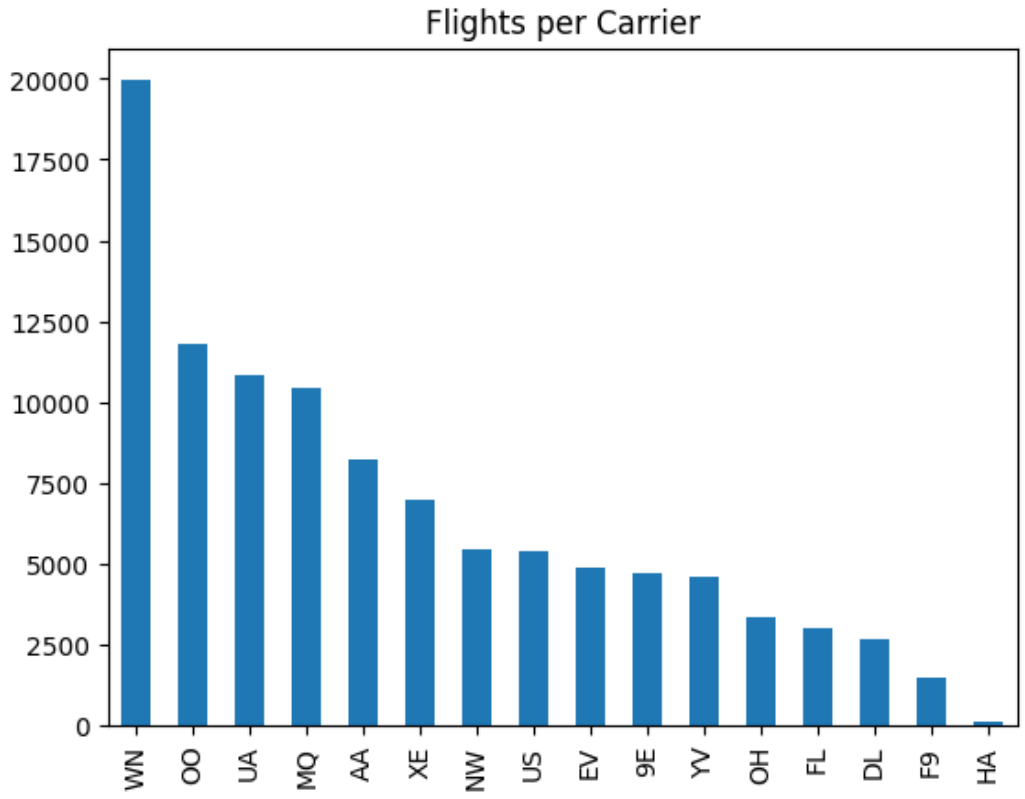
Logistic Regression, Decision Tree, and Deep Neural Networks were chosen for this project. Logistic Regression is a simple yet powerful classification algorithm that works well with binary classification problems. Decision Trees are versatile and interpretable models. Deep Neural Networks are powerful models that can learn complex patterns.

## Approach and Process:

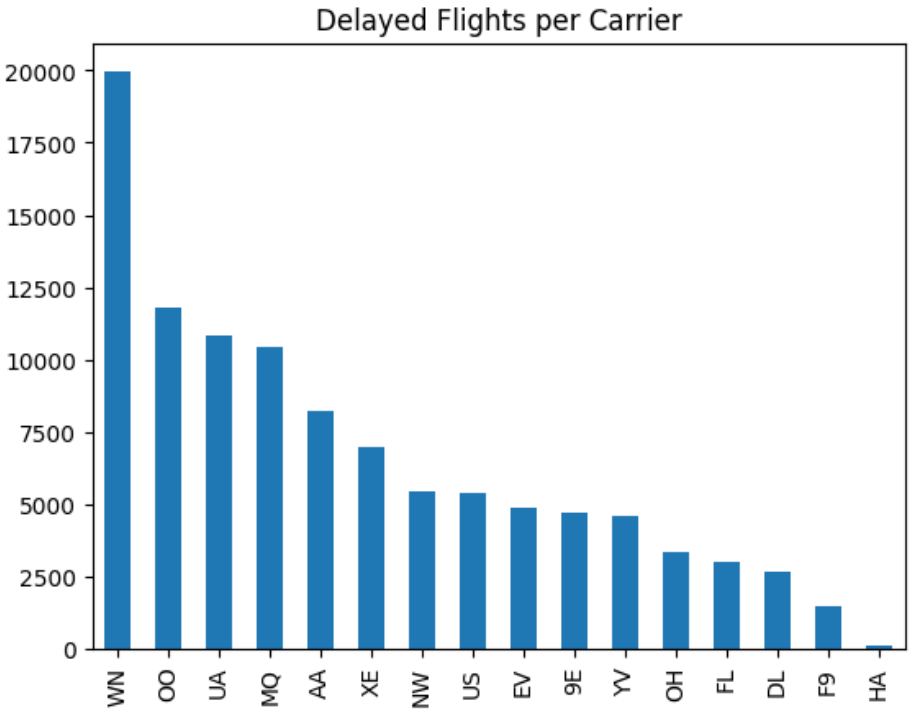
The dataset was first explored and cleaned. Then, the models were trained using the training dataset and evaluated on the test dataset. The performance of the models was compared to selecting the best one.

### High-Level Diagram:

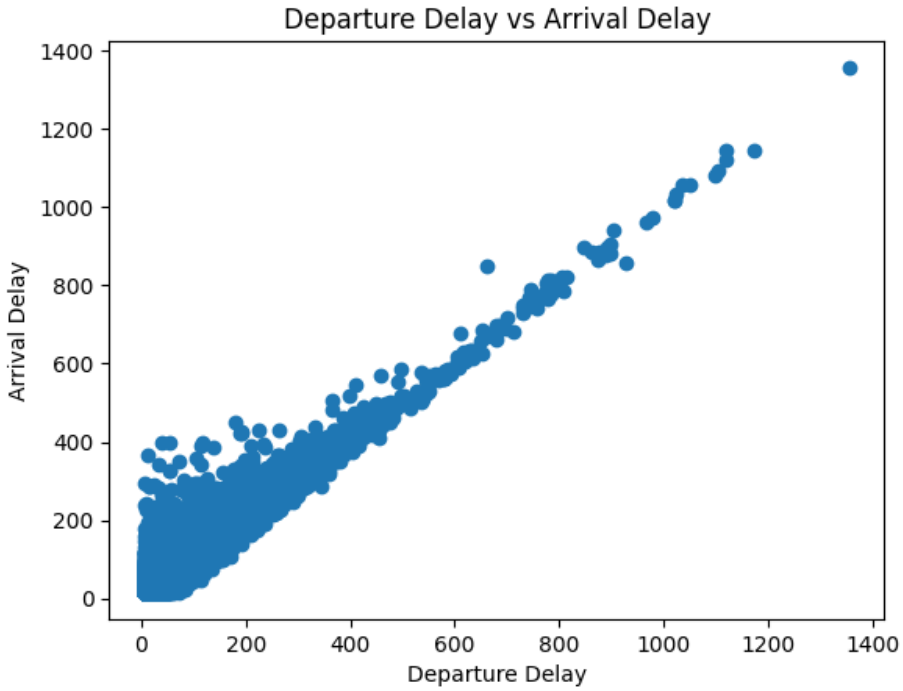
**Flights per Carrier**: A bar plot was used to visualize the number of flights per unique carrier. This gave an insight into which airlines have the most flights.

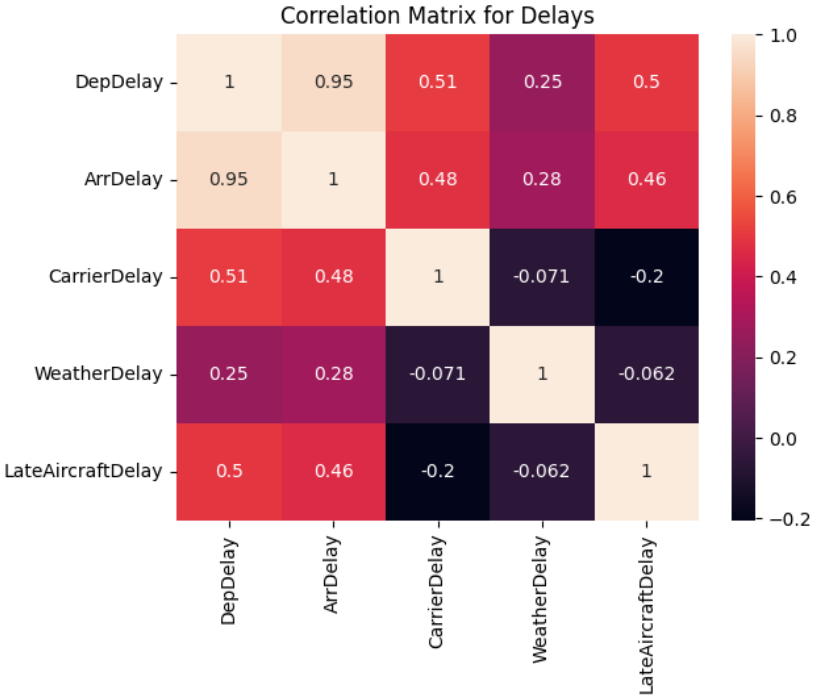


**Average delay per Carrier:** The average delay per carrier was calculated and visualized using a bar plot. This helped identify the carriers that have the highest average delays.



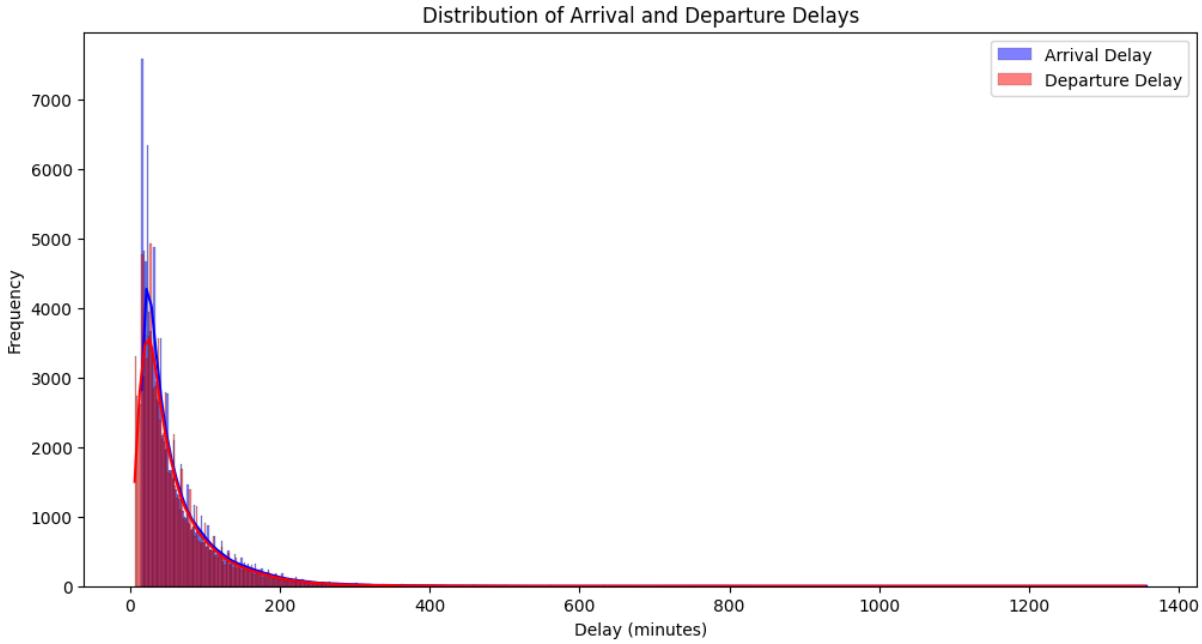

**Scatter plots and Correlation matrix:**





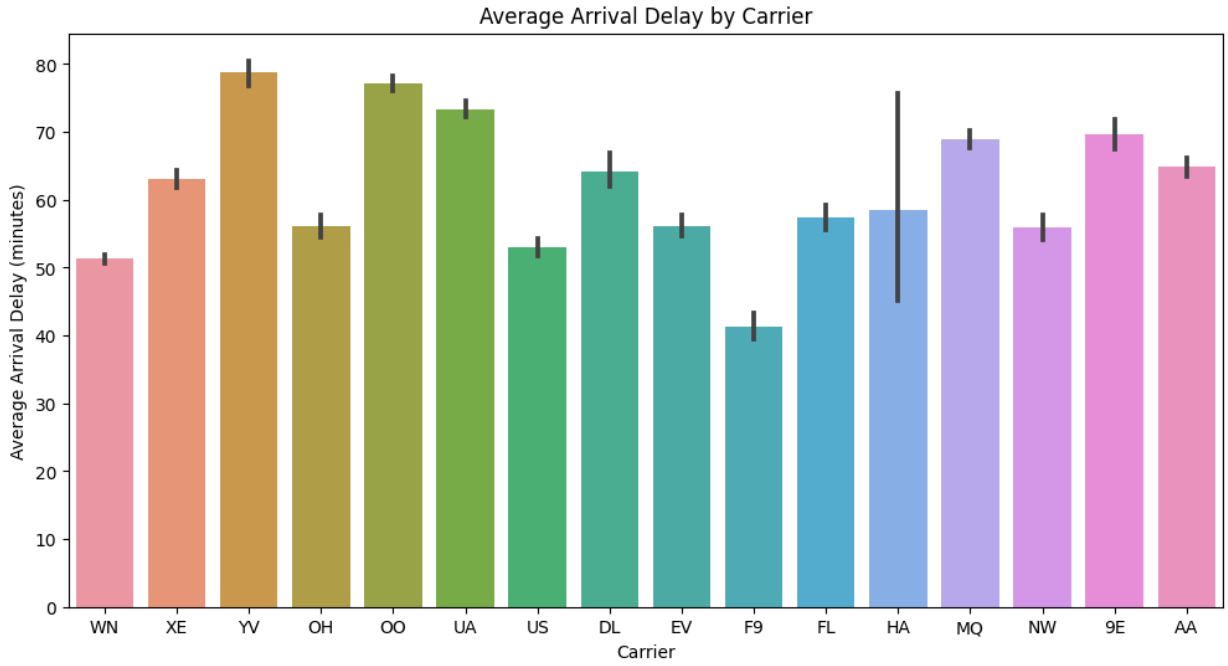

**Distribution of Arrival and Departure Delays:**

Histograms were used to visualize the distribution of arrival and departure delays. This helped in understanding the common delay times and identify any skewness in the data.



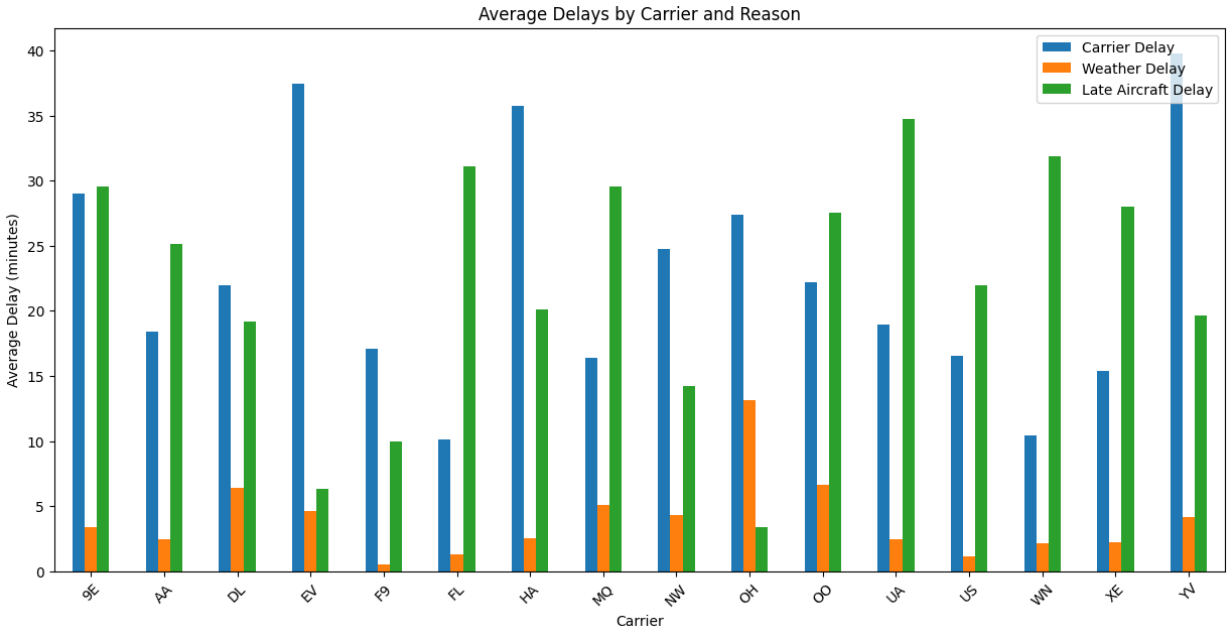
**Average Arrival Delay by Carrier:**

A bar plot was used to visualize the average arrival delay by carrier. This helped in identifying which carriers generally have more arrival delays.




**Average Arrival Delay by Carrier:**

A bar plot was used to visualize the average arrival delay by carrier. This helped in identifying which carriers generally have more arrival delays.



## Review and Validation:

### Machine Learning Models

Three machine learning models were used in this study, namely Logistic Regression, Decision Tree, and Deep Neural Networks (DNN).

- **Logistic Regression:** Preprocessing was done using a column transformer, the model was trained, predictions were made, and the model's performance was evaluated. It had an accuracy of 66.4%.

```
Logistic Regression Accuracy: 0.6643652102225887
Logistic Regression Classification Report:
              precision    recall  f1-score   support

           0       0.53      0.07      0.12      6566
           1       0.67      0.97      0.79     12842

    accuracy                           0.66     19408
   macro avg       0.60      0.52      0.46     19408
weighted avg       0.62      0.66      0.57     19408
```

- **Deep Neural Networks (DNN):** The DNN model was built using the Sequential class from the Keras API. It was compiled and trained. The DNN model had an overall accuracy of 65%.

```
Deep Neural Networks:
              precision    recall  f1-score   support

           0       0.46      0.18      0.26      6566
           1       0.68      0.89      0.77     12842

    accuracy                           0.65     19408
   macro avg       0.57      0.54      0.52     19408
weighted avg       0.60      0.65      0.60     19408

Accuracy: 0.6499896949711459
```

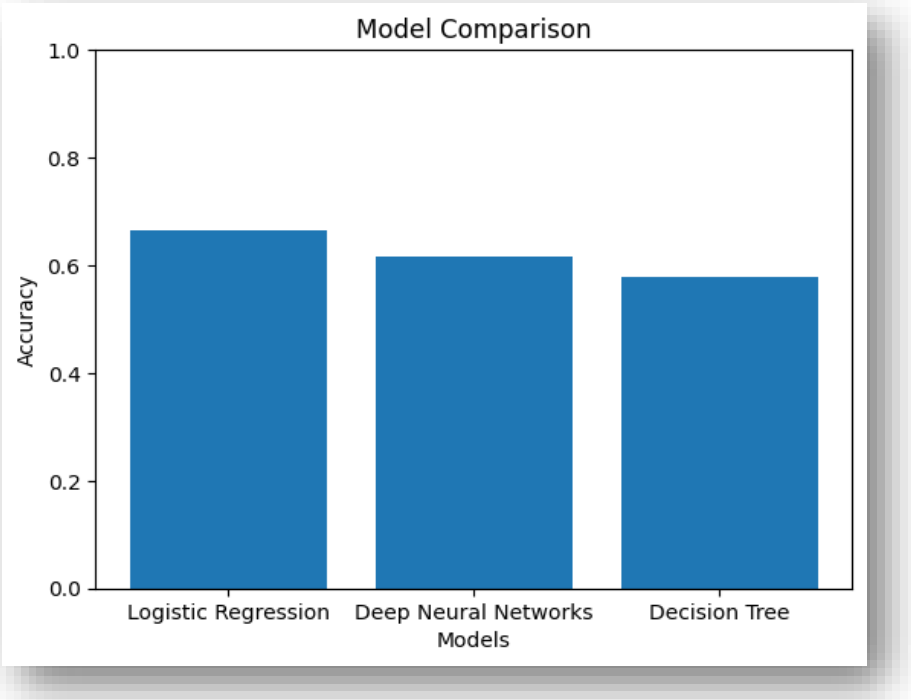
- **Decision Tree:** The Decision Tree Classifier from sklearn.tree was used. The model was created, trained, predictions were made, and its performance was evaluated. The decision tree model had an overall accuracy of 58%.

```
Decision Tree Accuracy: 0.5794002473206925
Decision Tree Classification Report:
              precision    recall  f1-score   support

           0       0.39      0.43      0.41      6563
           1       0.69      0.66      0.67     12845

    accuracy                           0.58     19408
   macro avg       0.54      0.54      0.54     19408
weighted avg       0.59      0.58      0.58     19408
```

The models were evaluated based on their accuracy, precision, recall, and F1-score. The logistic regression model outperformed the others with an accuracy of 66.44%. However, the dataset was imbalanced, which may have influenced the models' performances.



## Evaluation and Conclusion:

The models performed decently in predicting flight delays, with the Logistic Regression model performing the best. The results indicate that machine learning can be a viable approach to predicting flight delays.

## Future Work:

Given more time, the models could be improved by tuning their hyperparameters. With better programming skills, more advanced techniques could be utilized such as ensemble methods or more complex neural networks. Better data, especially incorporating weather conditions, could also improve the models. If better functions/algorithms were available, they could be used to enhance model performance.

## Future Enhancement Areas:

The models could be improved by using techniques to handle the imbalance in the dataset. More features could be engineered from the existing data. The project could also be expanded to include real-time prediction of flight delays.

## Prepared by:

Pruthvi Raj, Pudi
Vamshi Krishna, Bangaru
Othniel Adu Mensah.