



# Vivekanand Education Society's Institute of Technology

(Autonomous Institute Affiliated to University of Mumbai, Approved by AICTE & Recognised by Govt. of Maharashtra)  
*NAAC accredited with 'A' grade*

## Semester: VI Review

Title of the Project: **Flight Delay Analysis Project**

**Domain:** AIML, Data Science

### **Group Members:**

Member 1: Anish Kulkarni (29)

Member 2: Aditya Lalwani (30)

Member 3: Shivam Prajapati (41)

**Guide:** Dr. Ravita Mishra



# Content

- Introduction
- Problem Statement
- Requirements
- Literature Survey
- Dataset Overview
- Implementation
- Results and Analysis
- Conclusion
- References



# Introduction to Project

- This project explores **flight delay prediction** using machine learning techniques.
- It analyzes **factors like weather, departure time, and flight distance** to identify delay patterns.
- By leveraging **Decision Tree and Random Forest algorithms**, the model classifies flights as **delayed** or **on time**.
- The goal is to provide **better scheduling insights** and minimize disruptions for passengers and airlines.



# Problem Statement

Flight delays are hard to predict as they depend on many factors like weather, time of day, and date. These factors create complex dependencies, making it difficult to manage flight schedules. Unexpected delays cause inconvenience to passengers, disrupting their travel plans. Airlines also face challenges, as delays lead to inefficient scheduling and financial losses. Managing these delays requires a smart approach that considers multiple factors to improve accuracy and reduce disruptions.



# Requirements

- **Data Sources :**
  - **Flight dataset** with flight date, arrival & departure time, weather conditions, and delays.
  - **Historical flight records** to analyze past trends and patterns.
- **Software Requirements :**
  - **Programming Language:** Python (for model development and data processing).
  - **Development Environment:** Google Colab (for cloud-based execution).
- **Libraries & Frameworks**
  - **Data Processing & Analysis:** Pandas, NumPy.
  - **Machine Learning:** Scikit-learn, TensorFlow/Keras.
  - **Data Visualization:** Matplotlib, Seaborn



# Requirements

- **Hardware Requirements**
  - Multi-core CPUs/GPU for parallel computations
  - **Sufficient RAM and storage** to handle large flight datasets.



# Literature Survey

Title	Published	Author	Description
Probabilistic Flight Delay Predictions Using Machine Learning and Applications to the Flight-to-Gate Assignment Problem	2021	Micha Zoutendijk and Mihaela Mitici	This paper presents a machine learning-based approach to predicting flight delays probabilistically, focusing on improving airport operations' robustness. Traditional flight delay prediction methods provide either binary (delayed/not delayed) or point estimates, but this study introduces probabilistic forecasting using Mixture Density Networks (MDNs) and Random Forest Regression (RFR). These methods estimate the probability distributions of flight delays rather than just a single predicted delay value, allowing for a more informed decision-making process.



# Literature Survey

Title	Published	Author	Description
Flight Delay Classification Prediction Based on Stacking Algorithm	2021	Jia Yi, Honghai Zhang, Hao Liu, Gang Zhong and Guiyi Li	<p>This paper presents a machine learning-based approach to predicting flight delays using a Stacking algorithm, a form of ensemble learning. The study addresses the challenge of selecting the best algorithm for flight delay prediction by combining multiple machine learning models into a two-level Stacking classifier. The research employs five first-level classifiers—K-Nearest Neighbors (KNN), Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), and Gaussian Naive Bayes (GNB)—while the second-level learner is Logistic Regression. The Synthetic Minority Oversampling Technique (SMOTE) is used to handle imbalanced datasets, and the Boruta algorithm is applied for feature selection.</p>





# Dataset Overview

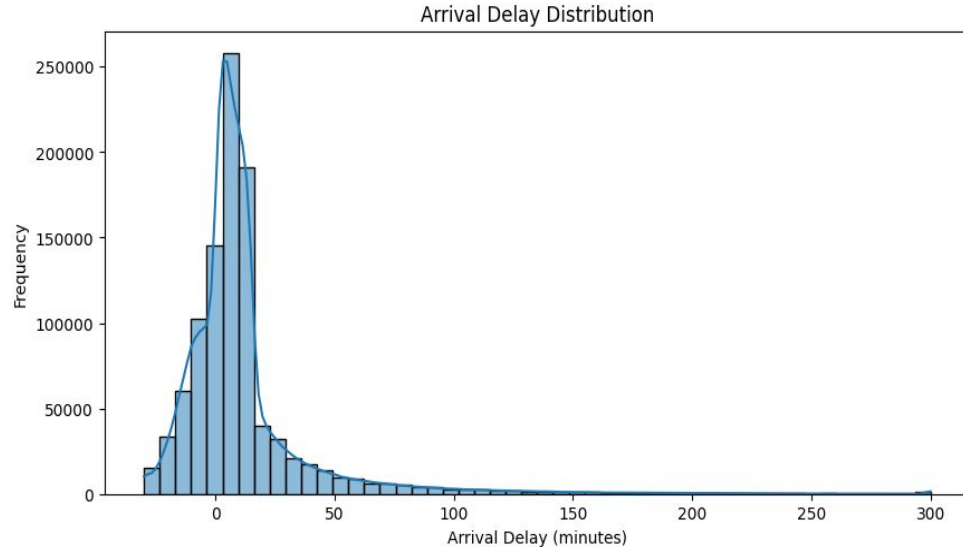
- **Total Entries:** 1,000,000
- **Key Features:** Flight date, delays, air time, distance, weather.
- **Target Variable:** Arrival
- **Weather Impact:** Conditions like **rain and storms** affect delays.
- **Data Type:** Mix of **numerical**

```
>>> <class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000000 entries, 0 to 999999
Data columns (total 8 columns):
#   Column      Non-Null Count  Dtype
---  -
0   FL_DATE     1000000 non-null  object
1   DEP_DELAY   1000000 non-null  int64
2   ARR_DELAY   1000000 non-null  int64
3   AIR_TIME    1000000 non-null  int64
4   DISTANCE    1000000 non-null  int64
5   DEP_TIME    1000000 non-null  float64
6   ARR_TIME    1000000 non-null  float64
7   Weather     1000000 non-null  object
dtypes: float64(2), int64(4), object(2)
memory usage: 61.0+ MB
```



# Implementation (EDA)

- Most flights have **small or no delays**, with a peak around **0 minutes**.
- A few flights experience **long delays**, but they are less common.
- The distribution is **right-skewed**, meaning delays are usually short, but some are very long.
- Helps in understanding **delay patterns** and predicting future delays.

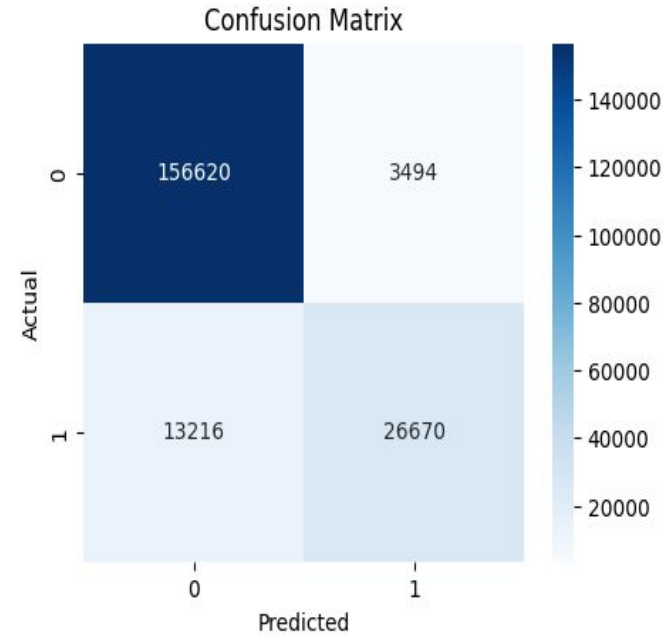




# Implementation (Decision Tree)

- **Confusion Matrix:** Correctly predicted **156,620 (0)** and **26,670 (1)**
- **Accuracy:** **91.65%** overall correctness.
- **Precision & Recall:** High for **class 0**, lower recall (67%) for **class1**.
- **F1-score:** **0.95 (0)**, **0.76 (1)** – class 1 is

Accuracy: 0.91645					
Classification Report:					
	precision	recall	f1-score	support	
0	0.92	0.98	0.95	160114	
1	0.88	0.67	0.76	39886	
accuracy			0.92	200000	
macro avg	0.90	0.82	0.86	200000	
weighted avg	0.91	0.92	0.91	200000	

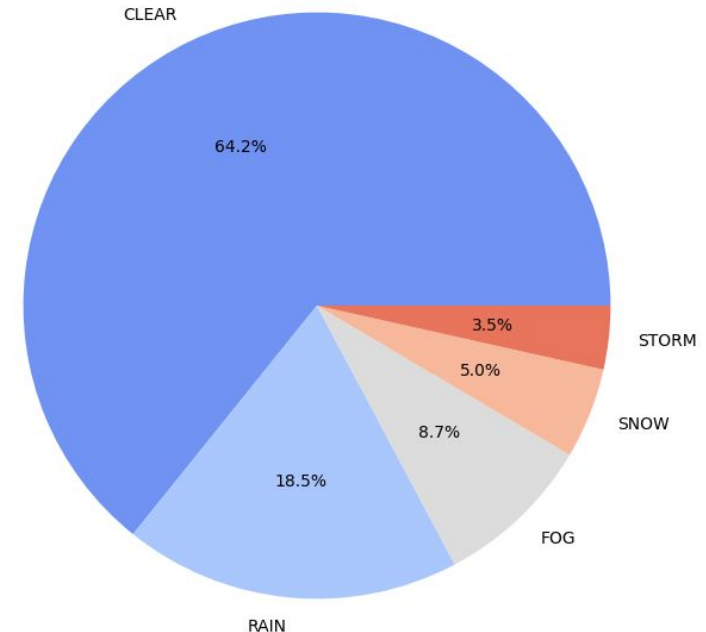




# Implementation (Weather Impact Analysis)

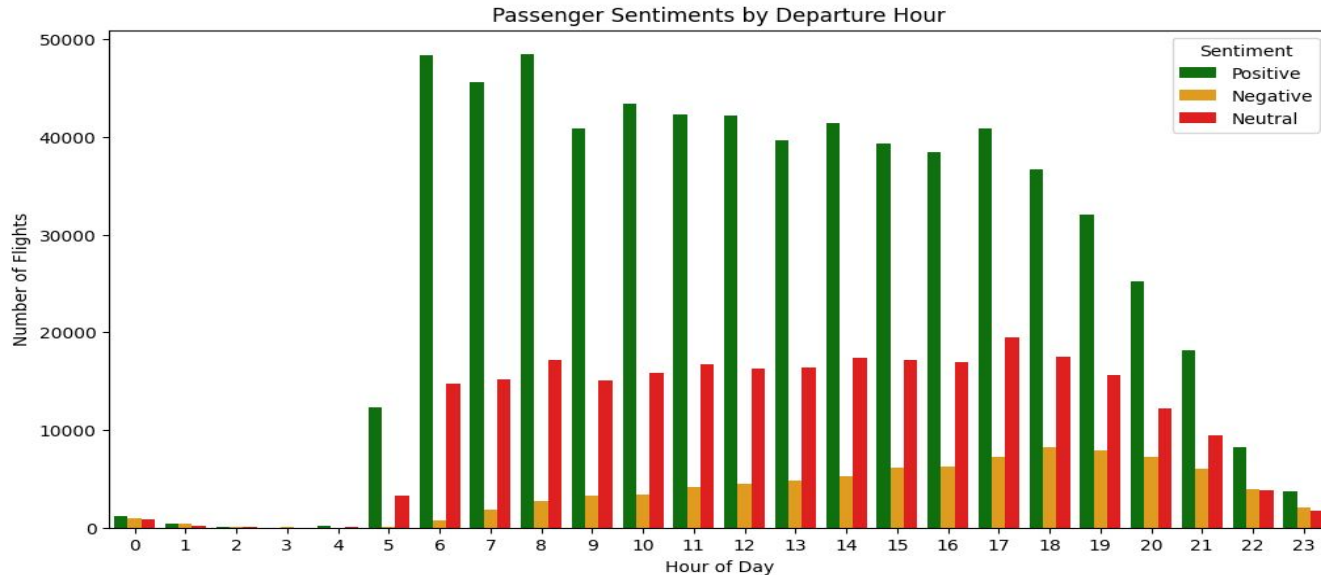
- **Clear weather** is the most common, occurring **64.2%** of the time.
- **Rainy conditions** affect
- **Fog (8.7%), snow (5.0%), and storms (3.5%)** occur less frequently.
- Weather conditions play a role in **flight delays and disruptions**.

Distribution of Weather Conditions in Flights





# Implementation (Passengers sentiment Analysis)



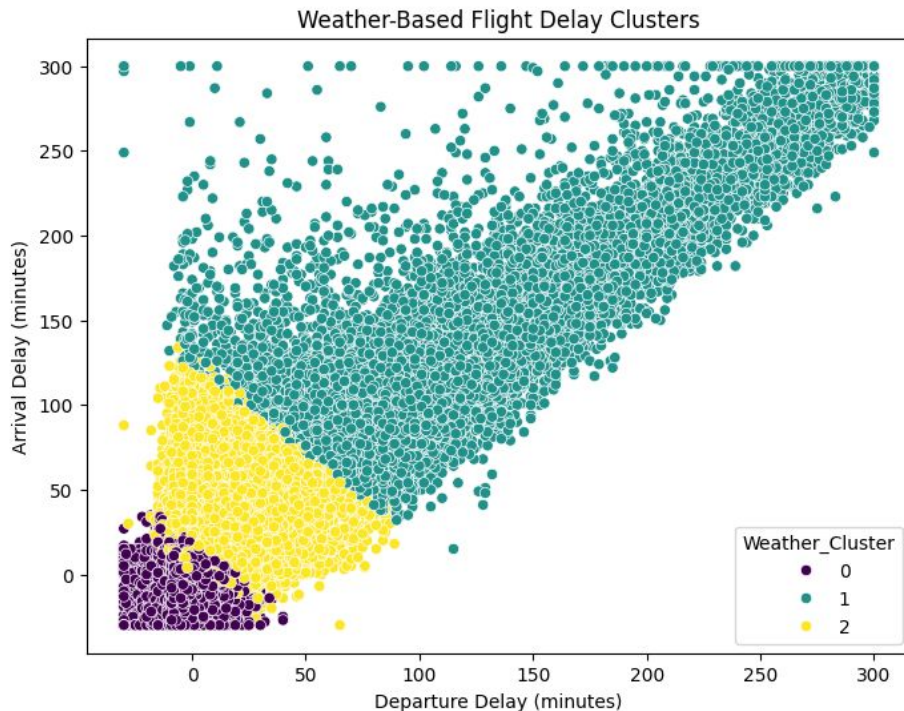
- Most flights depart between **6 AM and 10 AM**, with high passenger volume and peak positive sentiment in the early morning and afternoon.
- **Neutral sentiment remains steady**, while negative sentiment rises in the evening, likely due to **delays or passenger fatigue**.



# Implementation

## (Weather-Based Flight Delay Clustering)

- **Clusters show different delay patterns** based on weather conditions.
- **Low delays (Cluster 0 - Purple)** occur in **good** weather.
- **Moderate delays (Cluster 2 - Yellow)** are seen in **mild** weather disruptions.
- **High delays (Cluster 1 - Teal)** occur under **severe** weather conditions.





# Result and Analysis

- **Flight delays are predictable** using historical data and machine learning.
- **Weather and departure delays impact arrival delays** significantly.
- **Decision Tree provides higher accuracy (91%)** but may overfit.
- **Random Forest (85% accuracy) generalizes better** for unseen flight data.

## Result and Analysis

- **Decision Tree (91% Accuracy):** Higher accuracy but may overfit; strong recall for delayed flights (98%).
- **Random Forest (85% Accuracy):** More generalized; better recall for on-time flights (92%).
- **Precision vs. Recall:** Decision Tree prioritizes recall, while Random Forest balances both.
- **Model Choice:** Decision Tree for detecting delays; Random Forest for stable predictions.

```

➡➡ Accuracy: 0.8585408080694029
Classification Report:
              precision    recall  f1-score   support

     0           0.82       0.92       0.87       159078
     1           0.91       0.80       0.85       159756

 accuracy              0.86       318834
  macro avg           0.86       0.86       0.86       318834
 weighted avg           0.86       0.86       0.86       318834

```

# Random Forest

```

➡ Accuracy: 0.91645
Classification Report:

```

	precision	recall	f1-score	support
0	0.92	0.98	0.95	160114
1	0.88	0.67	0.76	39886
accuracy			0.92	200000
macro avg	0.90	0.82	0.86	200000
weighted avg	0.91	0.92	0.91	200000

# Decision Tree





# Conclusion

- **Machine learning can effectively predict flight delays based on historical data.**
- **Weather and departure delays are key factors influencing arrival delays.**
- **Decision Tree achieves higher accuracy (91%) but may overfit the data.**
- **Random Forest (85% accuracy) is more robust and generalizes better.**



# References

- [1] Flight Delay Classification Prediction Based on Stacking Algorithm, Jia Yi, Honghai Zhang, Hao Liu, Gang Zhong, Guiyi Li, <https://onlinelibrary.wiley.com/doi/full/10.1155/2021/4292778>
- [2] Probabilistic Flight Delay Predictions Using Machine Learning and Applications to the Flight-to-Gate Assignment Problem, Micha Zoutendijk, Mihaela Mitici, <https://www.mdpi.com/2226-4310/8/6/152>
- [3] Dataset: [Free CSV Sample Files - Download Example CSV Datasets | TabLab](#) - Flight Data (1M rows)