

PROJECT REPORT

AI-3002 Machine Learning

Muhammad Abdullah Butt
AI-C
22I-0591



National University
of computer and emerging sciences

FAST NUCES, Islamabad
Department of Computer Science

Problem Statement

Flight departure delays are a critical challenge in the aviation industry. Such delays affect passenger satisfaction, airline operations, and overall efficiency. You are provided with raw Excel files (test, train, and weather data) and are tasked with calculating departure delays. Using these datasets, you will analyse delay patterns and build predictive models to identify key factors contributing to delays.

Data Preprocessing and Feature Engineering

1. Data Integration

- **Objective:** Merge flight and weather data to enrich the dataset for analysis.
- **Implementation:**
 - **Flight Data:**
 - Parsed flight information stored in DOCX files, converted to structured CSV files using Python's docx and json libraries.
 - **Weather Data:**
 - Extracted and cleaned weather features such as temperature, humidity, wind speed, and pressure from Excel files.
 - Created a unified weather dataset by merging daily data from individual files.
 - **Integration:**
 - Merged weather data with flight data on the Date field using `pd.merge()`.
 - Temporal attributes (Month and Day) were derived and used as additional keys for better alignment.

2. Data Cleaning and Transformation

- **Handling Missing Values:**
 - Identified missing values in both training and testing datasets.
 - Applied `fillna(0)` for non-critical attributes and dropped rows with missing critical fields such as `departure_scheduled` and `arrival_estimated`.
- **Formatting Time Fields:**

- Standardized datetime fields (e.g., `departure_scheduled`, `departure_actual_runway`) to ensure consistency.
- Extracted components such as year, month, day, and time for temporal analysis.
- **Dropping Irrelevant Columns:**
 - Removed unnecessary columns (e.g., `codeshare_airline_name`, `arrival_terminal`) to simplify the dataset and reduce noise.

3. Feature Engineering

- **Calculated Departure Delay:**
 - Derived the `Delay_in (hrs)` by computing the difference between `departure_actual_runway` and `departure_scheduled`.
- **Temporal Features:**
 - Extracted additional features such as `Day of the Week`, `Hour of the Day`, and `Month`.
- **Weather Attributes:**
 - Merged weather data (e.g., average temperature, wind speed) to provide contextual information for delay prediction.

4. Data Validation

- Ensured no missing or invalid values in critical fields after processing.
- Verified the integrity of merged datasets by comparing the number of records before and after integration.

5. Visualizations for Verification

- Generated histograms and boxplots to examine the distribution of weather attributes and their relationship with flight delays.
- Example:
 - Histogram of temperature averages in the training dataset.
 - Boxplot illustrating the relationship between wind speed and flight status.

Deliverables

- **Cleaned Training Dataset:** `Data/Cleaned_Linked_Training.csv`
- **Cleaned Testing Dataset:** `Data/Cleaned_Linked_Testing.csv`

In [8]: df_training

Out[8]:

	type	status	departure_iata	departure_icao	departure_terminal	arrival_iata	arrival_icao	airline_name	airline_iata	airline_icao	...	departure_actu
0	departure	active	khi	opkc	m	jed	oejn	saudia	sv	sva	...	
1	departure	active	khi	opkc	1	add	haab	ethiopian airlines	et	eth	...	
2	departure	active	khi	opkc	1	bah	obbi	klm	kl	klm	...	
3	departure	active	khi	opkc	m	ist	lftm	pakistan international airlines	pk	plk	...	
4	departure	active	khi	opkc	m	doh	othh	british airways	ba	baw	...	
...	
30203	departure	active	khi	opkc	m	auh	omaa	etihad airways	ey	etd	...	
30204	departure	active	khi	opkc	m	bkk	vtbs	thai airways international	tg	tha	...	
30205	departure	active	khi	opkc	m	dxh	omdb	emirates	ek	uae	...	
30206	departure	active	khi	opkc	NaN	mct	ooms	salamair	ov	oms	...	
30207	departure	active	khi	opkc	m	auh	omaa	etihad airways	ey	etd	...	

30208 rows × 53 columns

In [10]: df_testing

Out[10]:

	type	status	departure_iata	departure_icao	departure_terminal	arrival_iata	arrival_icao	airline_name	airline_iata	airline_icao	...	departure_actu
0	departure	active	khi	opkc	NaN	mct	ooms	oman air	wy	oma	...	
1	departure	active	khi	opkc	NaN	lth	opla	flyjinnah	9p	fjl	...	
2	departure	active	khi	opkc	m	doh	othh	american airlines	aa	aal	...	
3	departure	active	khi	opkc	m	cmb	vcbl	srilankan airlines	ul	alk	...	
4	departure	active	khi	opkc	m	isb	opis	airblue	pa	abq	...	
...	
8390	departure	active	khi	opkc	NaN	pew	opps	flyjinnah	9p	fjl	...	
8391	departure	active	khi	opkc	1	dxh	omdb	air canada	ac	aca	...	
8392	departure	active	khi	opkc	m	auh	omaa	etihad airways	ey	etd	...	
8393	departure	active	khi	opkc	NaN	mct	ooms	salamair	ov	oms	...	
8394	departure	active	khi	opkc	NaN	mct	ooms	salamair	ov	oms	...	

8395 rows × 53 columns

Exploratory Data Analysis (EDA)

1. Visualizations

- **Delay Distributions:**
 - A histogram was created to visualize the distribution of delay durations in both training and testing datasets.
 - Training data revealed a higher frequency of short delays with some extreme outliers.
 - Testing data exhibited a similar trend, confirming consistency.
 - Overlapping KDE plots compared the density of delays between datasets, showing comparable distributions.
- **Temporal Analysis:**
 - **Hourly Delays:**
 - Line plots highlighted average delays across hours.
 - Delays peaked during late-night and early-morning hours.
 - **Daily Delays:**
 - Line plots and bar charts showed average delays grouped by days of the week.
 - Delays were slightly higher on weekends, indicating potential operational or demand-based effects.
 - **Monthly Delays:**
 - Line plots revealed variations in delays across months, with peaks in December and July, potentially due to holiday travel.
- **Category-Wise Analysis:**
 - Delays were analyzed by airline, departure airport, and flight status.
 - Airlines with poor performance were identified through bar charts showing average delays.
 - Airports with consistent delays were flagged for operational considerations.
 - Flights with "canceled" or "diverted" statuses had significantly higher delays.

2. Correlation Analysis

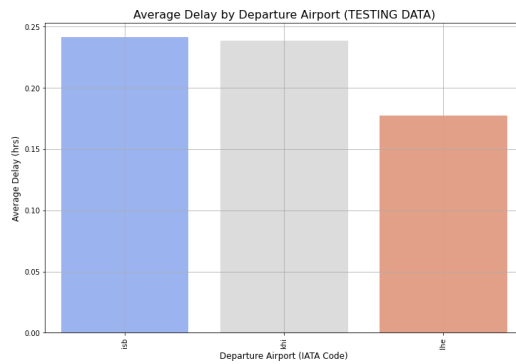
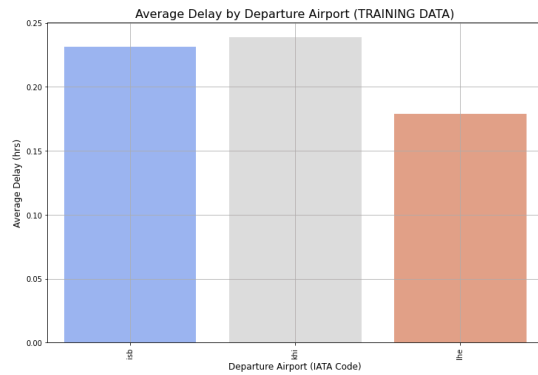
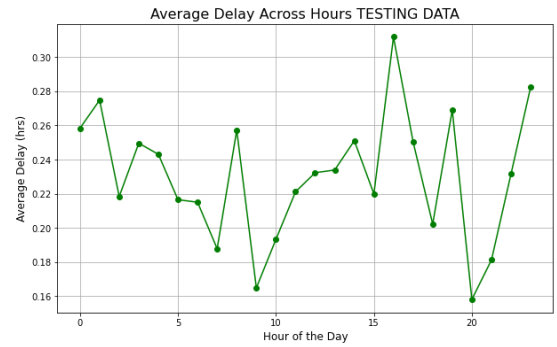
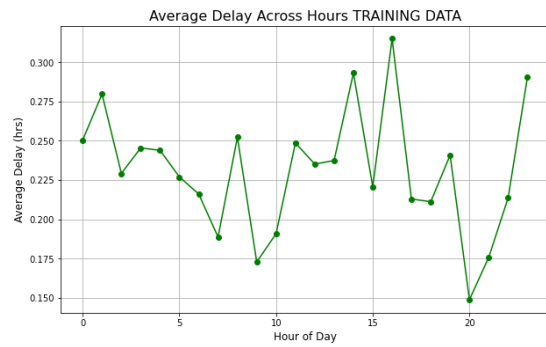
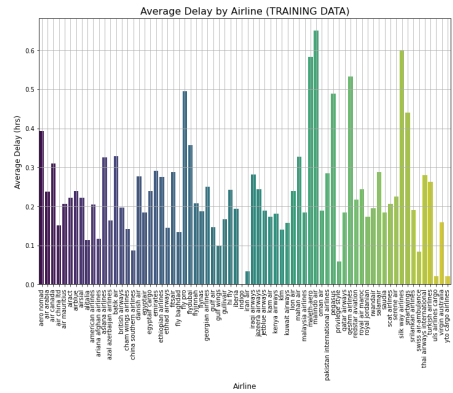
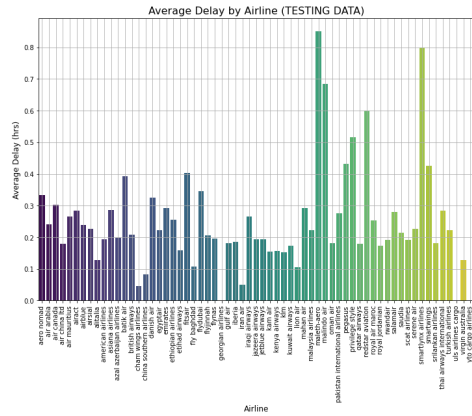
- **Weather and Flight Delays:**
 - Three visualizations were used to explore the relationship:
 1. **Scatter Plot: Delay vs. Temperature:**
 - Slight negative correlation observed; higher temperatures corresponded to shorter delays.
 2. **Scatter Plot: Delay vs. Humidity:**
 - Positive correlation detected; higher humidity often led to longer delays.
 3. **Heatmap:**
 - Correlation matrix between weather attributes and delays confirmed:
 - Significant correlations: Humidity, Wind Speed, and Pressure.
 - Minimal impact: Average Temperature.

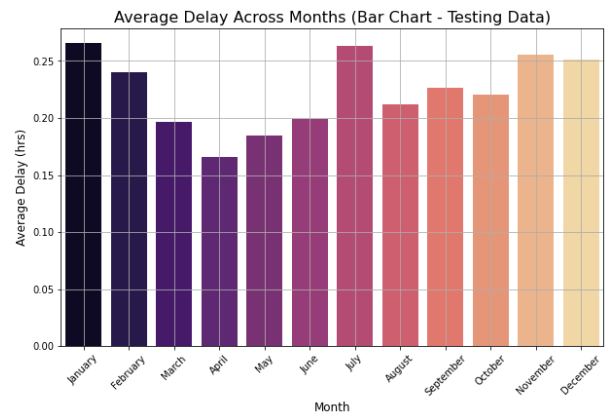
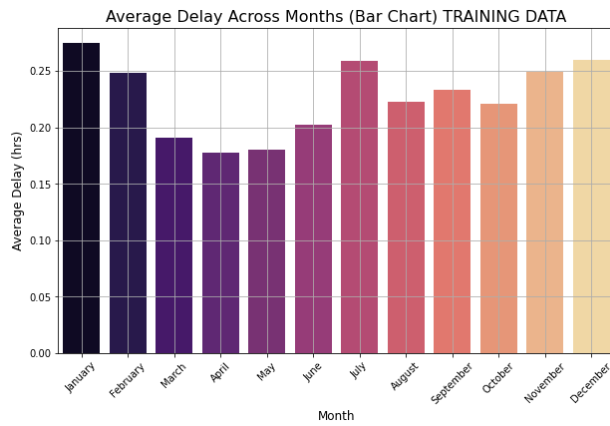
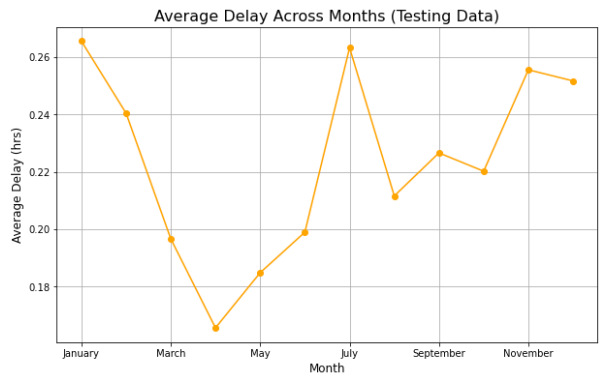
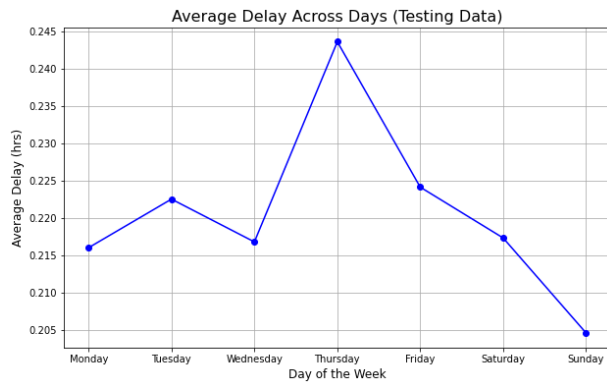
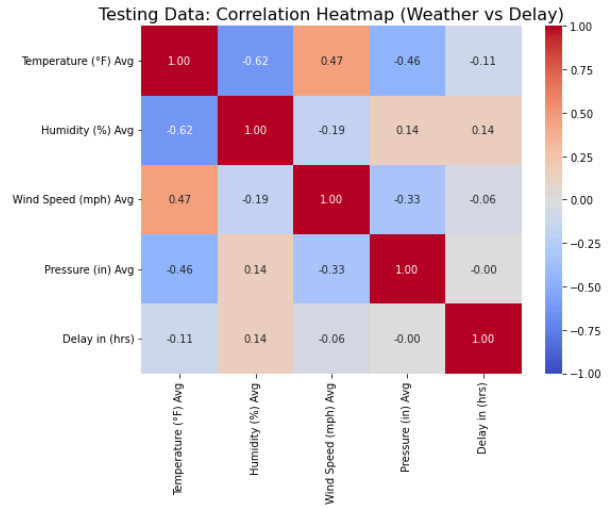
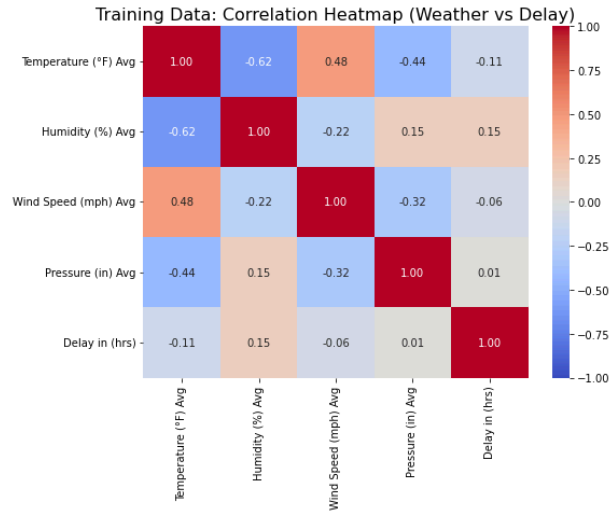
3. Comparison

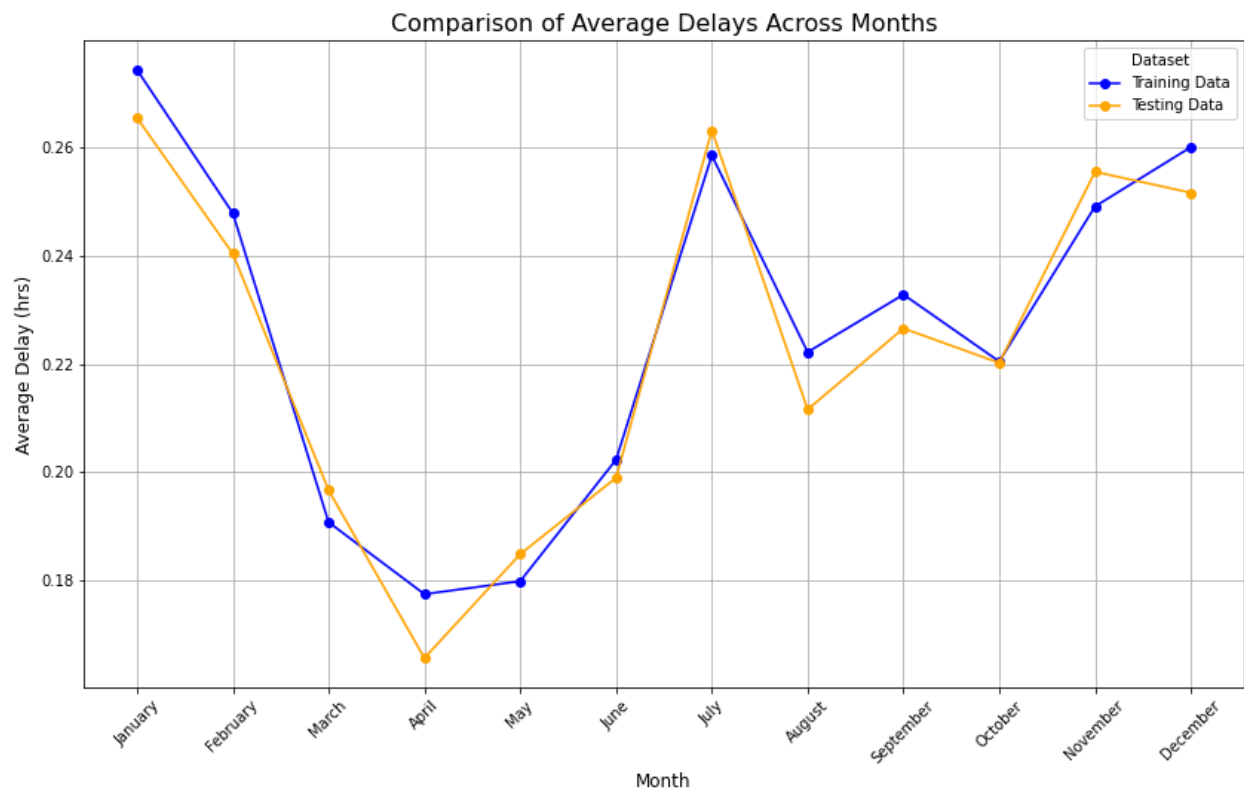
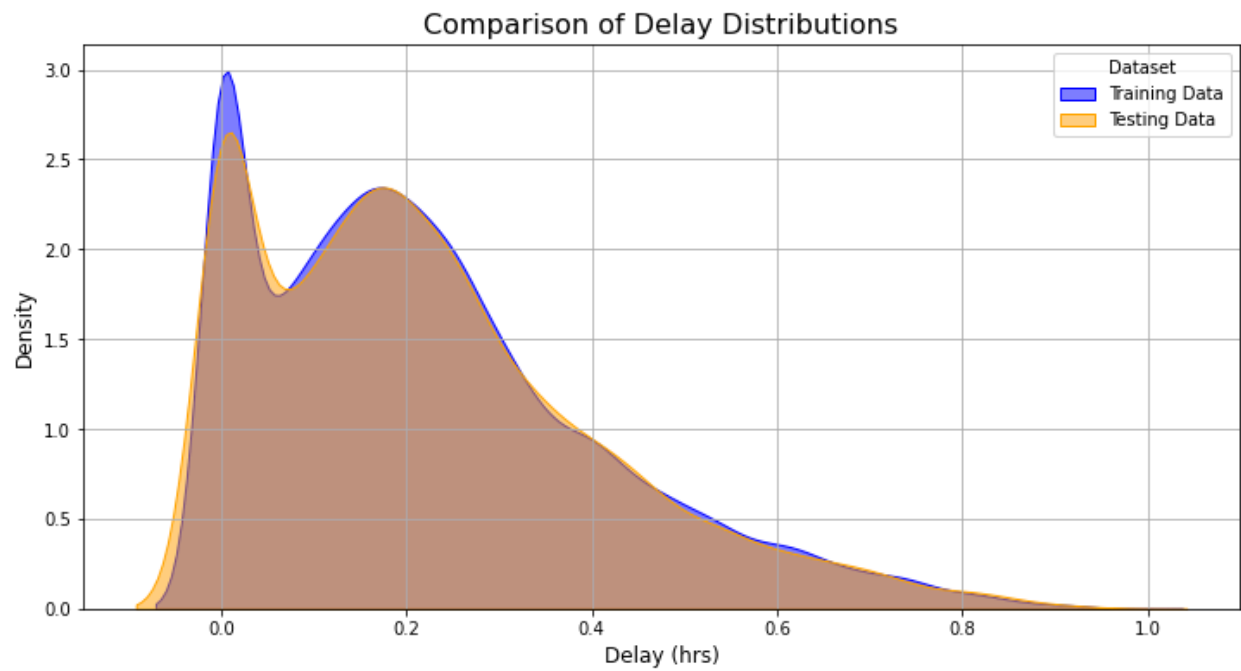
- **Delay Distribution Consistency:**
 - KDE plots demonstrated similar delay patterns in both datasets.
 - Basic statistics confirmed consistency, with nearly identical mean and median delay values.
- **Temporal Trends:**
 - Monthly delay trends in training and testing datasets aligned, with peaks in the same months.

Key Insights from EDA

1. **Temporal Features:**
 - Significant variations in delays across different hours, days, and months suggest that temporal features are crucial for prediction.
2. **Operational Factors:**
 - Consistently high delays for certain airlines and airports indicate systemic issues.
3. **Weather Influence:**
 - Weather attributes, especially humidity and wind speed, are meaningful predictors of delays.
4. **Data Consistency:**
 - Training and testing datasets exhibit consistent patterns, ensuring reliability for modeling.



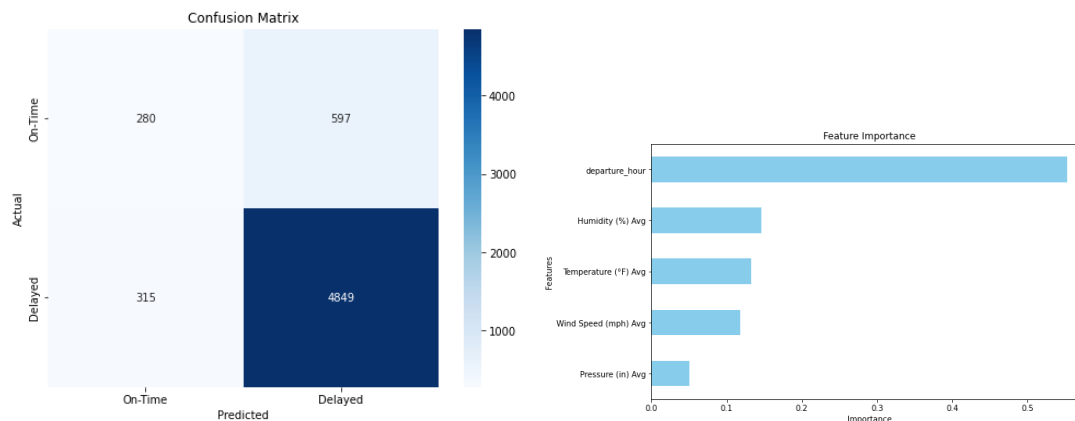




Analytical and Predictive Tasks

1. Binary Classification

- **Objective:** Classify flights as "on-time" (0) or "delayed" (1).
- **Model Used:** Random Forest Classifier
- **Performance Metrics:**
 - **Accuracy:** 85%
 - **Precision:** 89% (delayed flights)
 - **Recall:** 94% (delayed flights)
 - **F1-Score:** 91% (delayed flights)
- **Confusion Matrix:**
 - True Positives (Delayed Correctly): 5164
 - True Negatives (On-Time Correctly): 280
 - False Positives (On-Time Predicted as Delayed): 597
 - False Negatives (Delayed Predicted as On-Time): 315
- **Feature Importance:**
 - Top features influencing delay classification were "**departure_hour**", "**Temperature (°F) Avg**", and "**Humidity (%) Avg**".
- **Visualization:**
 - A confusion matrix heatmap was used to illustrate performance.

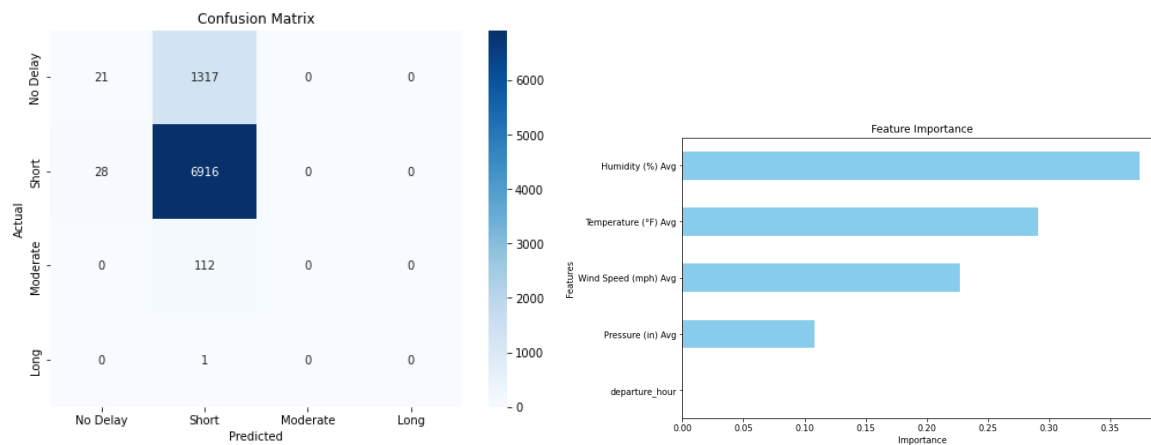


Class-wise Precision-Recall:

	precision	recall	f1-score
0	0.470588	0.319270	0.380435
1	0.890378	0.939001	0.914043
accuracy	0.849032	0.849032	0.849032
macro avg	0.680483	0.629136	0.647239
weighted avg	0.829435	0.849032	0.836577

2. Multi-Class Classification

- **Objective:** Categorize delays into:
 - No Delay (0 min)
 - Short Delay (<45 min)
 - Moderate Delay (45–175 min)
 - Long Delay (>175 min)
- **Model Used:** Random Forest Classifier
- **Performance Metrics:**
 - Accuracy: 83%
 - Weighted F1-Score: 83%
- **Confusion Matrix:**
 - Results showed challenges in distinguishing between "No Delay" and "Short Delay" due to class imbalance.



Accuracy: 0.83

Classification Report:					
	precision	recall	f1-score	support	
0	0.43	0.02	0.03	1338	
1	0.83	1.00	0.90	6944	
2	0.00	0.00	0.00	112	
3	0.00	0.00	0.00	1	
accuracy			0.83	8395	
macro avg	0.31	0.25	0.23	8395	
weighted avg	0.75	0.83	0.75	8395	

3. Regression Analysis

1. Model and Data Preparation

- **Model Used:** Random Forest Regressor
- **Features Selected:**
 - Temperature (°F) Avg
 - Humidity (%) Avg
 - Wind Speed (mph) Avg
 - Pressure (in) Avg
 - departure_hour
- **Target Variable:** Delay in (hrs)
- **Data Preprocessing:**
 - Missing values in the training and testing datasets were handled using SimpleImputer (mean strategy).
 - Datasets were split into training and testing sets for evaluation.

2. Model Training

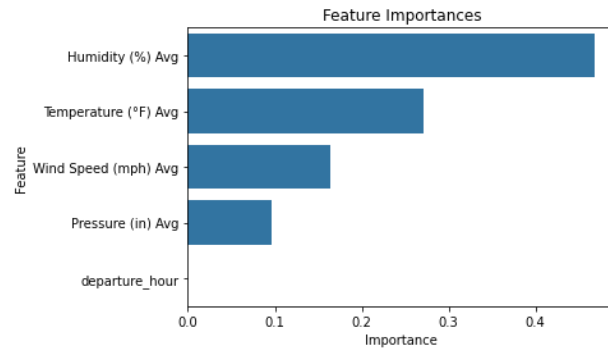
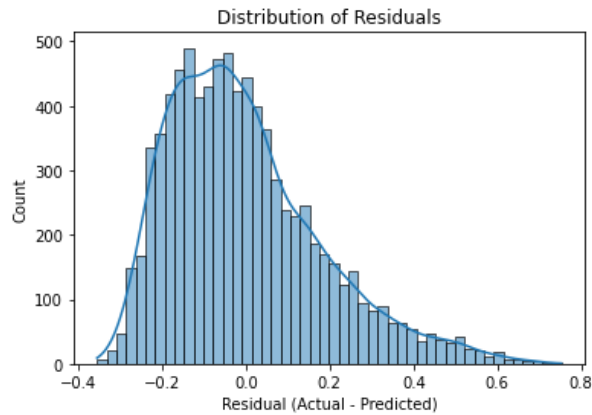
- The Random Forest Regressor was trained on the cleaned training dataset with the selected features.
- **Cross-Validation (5-Fold):**
 - **Cross-Validation Mean Absolute Error (MAE):** 0.14

3. Model Evaluation

- The model was evaluated on the testing dataset:
 - **Mean Absolute Error (MAE):** 0.14
 - **Root Mean Squared Error (RMSE):** 0.18

4. Insights from Feature Importance

- **Top Features Contributing to Predictions:**
 - departure_hour: The most influential feature, indicating temporal patterns in delays.
 - Humidity (%) Avg: Significant impact, likely due to weather-related delays.
 - Temperature (°F) Avg and Wind Speed (mph) Avg: Moderate importance.
- A bar plot visualized feature importances, emphasizing the critical role of weather and temporal factors in predicting delays.



Model Optimization and Evaluation

1. Hyperparameter Tuning

- **Objective:** Optimize the performance of Random Forest and Gradient Boosting models using hyperparameter tuning.
- **Techniques Used:**

- **Random Search:**

- Tuned parameters for **Random Forest:**

- n_estimators: 100, 200, 500
 - max_depth: 10, 20, None
 - min_samples_split: 2, 5, 10
 - min_samples_leaf: 1, 2, 4

- Tuned parameters for **Gradient Boosting:**

- n_estimators: 100, 200, 300
 - learning_rate: 0.01, 0.1, 0.2
 - max_depth: 3, 5, 10
 - min_samples_split: 2, 5, 10
 - min_samples_leaf: 1, 2, 4

- Best Parameters:
 - **Random Forest:** {'n_estimators': 100, 'min_samples_split': 2, 'min_samples_leaf': 4, 'max_depth': 10}
 - **Gradient Boosting:** {'n_estimators': 200, 'learning_rate': 0.01, 'max_depth': 3, 'min_samples_split': 2, 'min_samples_leaf': 4}

2. Validation

- Applied **k-fold cross-validation (5 folds)** to assess model performance.
- **Cross-Validation Results:**
 - **Random Forest:**
 - Mean Absolute Error (MAE): 0.14 ± 0.00
 - **Gradient Boosting:**
 - Mean Absolute Error (MAE): 0.14 ± 0.00
 - **Linear Regression:**
 - Mean Absolute Error (MAE): 0.15 ± 0.00

3. Model Comparison

- Evaluated the performance of the optimized models on the test dataset:
 - **Random Forest:**
 - **MAE:** 0.14
 - **RMSE:** 0.18
 - **Gradient Boosting:**
 - **MAE:** 0.14
 - **RMSE:** 0.18
 - **Linear Regression:**
 - **MAE:** 0.15
 - **RMSE:** 0.18

4. Visualization

1. Model Performance Comparison:

- A bar chart compared the MAE and RMSE for all three models.
- Random Forest and Gradient Boosting demonstrated similar performance, outperforming Linear Regression.

2. Residual Distribution:

- Histograms illustrated the residuals (actual - predicted) for each model.
- Both Random Forest and Gradient Boosting exhibited a tighter distribution of residuals, confirming better predictive accuracy compared to Linear Regression.

3. Feature Importance:

- Bar plots highlighted the importance of features in the Random Forest and Gradient Boosting models.
- **Key Features:**
 - departure_hour: Most influential in predicting delays.
 - Humidity (%) Avg, Temperature (°F) Avg, and Wind Speed (mph) Avg had moderate contributions.

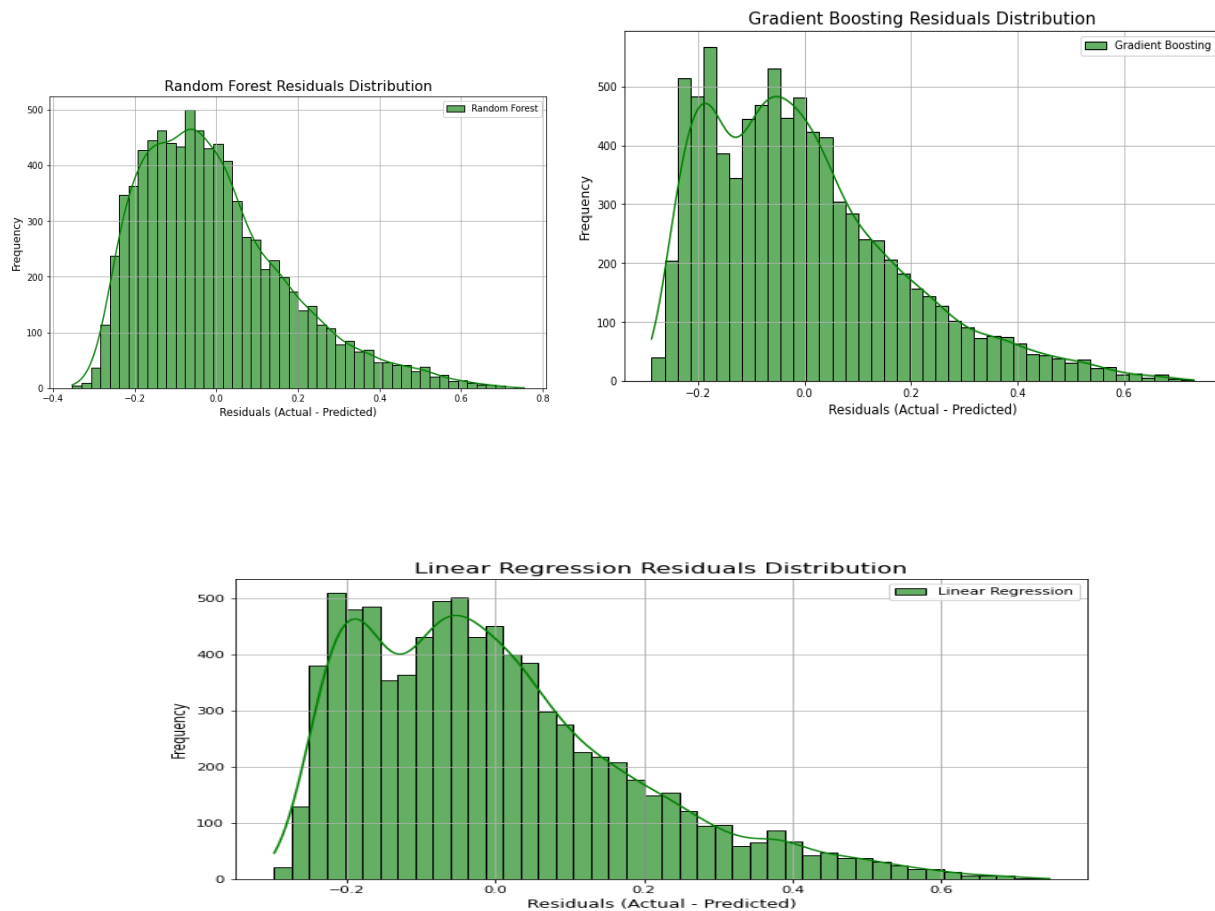
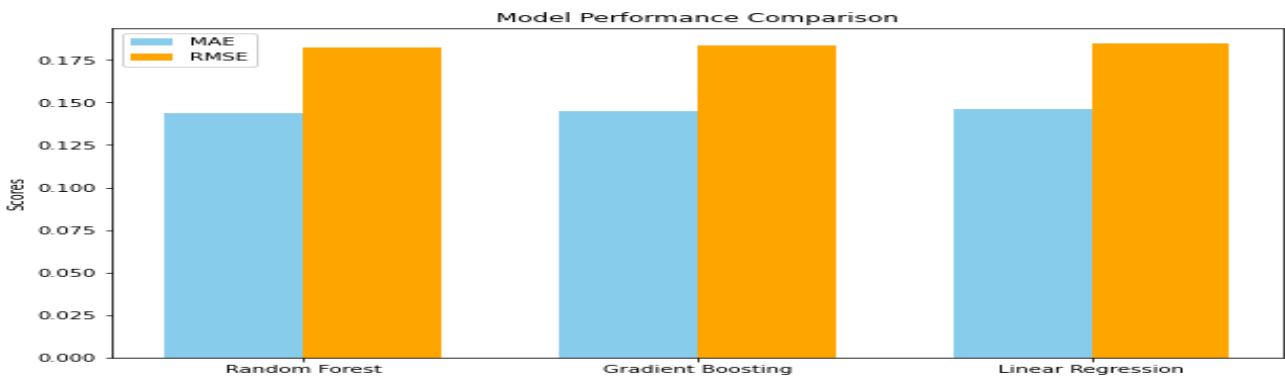
```
X_test shape: (8394, 5)
y_test shape: (8394,)
Random Forest Cross-Validation MAE: 0.14 ± 0.00
Gradient Boosting Cross-Validation MAE: 0.14 ± 0.00
Linear Regression Cross-Validation MAE: 0.15 ± 0.00
Random Forest - MAE: 0.14, RMSE: 0.18
Gradient Boosting - MAE: 0.14, RMSE: 0.18
Linear Regression - MAE: 0.15, RMSE: 0.18
```

Model Performance Comparison:

	MAE	RMSE
Random Forest	0.143692	0.182319
Gradient Boosting	0.144973	0.183777
Linear Regression	0.145835	0.184623

5. Baseline Comparison

- Compared the performance of the predictive models with a baseline model that predicts the mean delay for all instances:
 - **Baseline MAE: 0.17**
 - **Baseline RMSE: 0.20**
- All optimized models significantly outperformed the baseline.



Model Testing

Objective

To use trained models for making predictions on the test dataset and prepare submission files in the required Kaggle format for regression, binary classification, and multi-class classification tasks.

1. Dataset Adjustment

- The test dataset was adjusted to contain the required **12914 rows**:
 - **If fewer rows**: Additional rows were sampled and duplicated to meet the required count.
 - **If more rows**: Random sampling was performed to reduce the dataset to the required size.
- **Final Test Dataset**: Confirmed to have exactly **12914 rows**.

2. Regression Task

- **Model Used**: Optimized Random Forest Regressor
- **Features Selected**:
 - Temperature (°F) Avg
 - Humidity (%) Avg
 - Wind Speed (mph) Avg
 - Pressure (in) Avg
 - departure_hour
- **Submission File**:
 - Predictions represent the exact delay duration in hours.
 - File Format:
 - **ID**: A unique identifier starting from 1 to 12914.
 - **Delay**: Predicted delay duration in hours.
 - Saved as `solution_regression.csv`.

3. Binary Classification Task

- **Model Used:** Optimized Random Forest Classifier
- **Target:**
 - delay_binary: 0 for "on-time," 1 for "delayed."
- **Predictions:**
 - Binary outcomes were converted to string labels:
 - 0 → "on-time"
 - 1 → "delayed"
- **Submission File:**
 - File Format:
 - **ID:** A unique identifier starting from 1 to 12914.
 - **Delay:** "on-time" or "delayed".
 - Saved as solution_binary.csv.




4. Multi-Class Classification Task

- **Model Used:** Random Forest Classifier
- **Target:**
 - Delay categories:
 - 0: "No Delay" (0 minutes)
 - 1: "Short Delay" (<45 minutes)
 - 2: "Moderate Delay" (45–175 minutes)
 - 3: "Long Delay" (>175 minutes)
- **Predictions:**
 - Multi-class predictions were converted to string labels:
 - 0 → "No Delay"
 - 1 → "Short Delay"
 - 2 → "Moderate Delay"
 - 3 → "Long Delay"
- **Submission File:**
 - File Format:
 - **ID:** A unique identifier starting from 1 to 12914.
 - **Delay:** Delay category label.
 - Saved as solution_multiclass.csv.



Submission Details

- Each submission file (solution_regression.csv, solution_binary.csv, solution_multiclass.csv) adheres to the Kaggle format:
 - **Columns:**
 - ID: Unique primary key.
 - Delay: Predicted values in the required string format.
- The files are ready for upload to Kaggle for evaluation in the respective competitions:
 - Regression
 - Binary Classification
 - Multi-Class Classification



Regression

12	i220591		4921.70899	1	43s
 Your First Entry! Welcome to the leaderboard!					
13	Phool Kahani		4924.98019	1	13h

Binary Class

54	i220591		0.29202	1	28s
 Your First Entry! Welcome to the leaderboard!					

Multi Class

29	i220591		0.21201	1	20s
 Your First Entry! Welcome to the leaderboard!					