Aviation Incident Analysis Report

# Introduction

This report outlines the steps taken to analyze aviation incidents, focusing on country-level data to explore factors influencing these events. The analysis integrates various datasets to predict damage levels using machine learning models.

# Data Preparation

* 1. **Library Importing**

Here are the used libraries:

* **pandas & numpy**: Used for data manipulation, handling missing values, and numerical operations.
* **sklearn.preprocessing (OrdinalEncoder, OneHotEncoder, SimpleImputer)**: Used for encoding categorical variables and imputing missing data.
* **sklearn.compose & sklearn.pipeline**: Used to streamline preprocessing steps into a single workflow.
* **seaborn & matplotlib**: Used for data visualization, correlation analysis, and exploratory data analysis (EDA).
* **sklearn.metrics (accuracy\_score, classification\_report, roc\_curve, auc)**: Used for evaluating model performance.
* **sklearn.model\_selection (train\_test\_split)**: Used to split data into training and testing sets.
* **RandomForestClassifier**: A tree-based ensemble learning model used for classification.
* **XGBClassifier & LGBMClassifier**: Gradient boosting models used to enhance prediction accuracy and efficiency.
  1. **Dataset Loading**

Each dataset was loaded and checked for integrity, including missing values and inconsistent data formats.

* **Aviation Details Dataset**: Contains information on incidents, including date, country, damage level, aircraft type, number of engines, and presence of a second pilot…
* **GDP Data**: Annual GDP per country.
* **Aircraft Departures/Air Traffic Data**: Annual aircraft departures per country.
* **Tech Export Data**: Annual technological exports per country.
* **Tourism Receipts Data**: Annual tourism revenue per country.
  1. **Feature Selection**

The following features were selected from the aviation dataset as most relevant for analysis:

* **Incident Date (cm\_eventDate)**: Represents the date of the aviation incident.
* **Country (cm\_country)**: Indicates the location where the incident occurred.
* **Damage Level (DamageLevel)**: Specifies the extent of damage to the aircraft as a result of the incident.
* **Aircraft Category (aircraftCategory)**: Defines the type of aircraft involved in the incident (e.g., commercial, private, military).
* **Amateur Built (amateurBuilt)**: Indicates whether the aircraft was built by an amateur or a certified manufacturer.
* **Number of Engines (numberOfEngines)**: Represents the total number of engines in the aircraft at the time of the incident.
* **Flight Operation Type (flightOperationType)**: Describes the nature of the flight operation (e.g., passenger, cargo, training).
* **Second Pilot Present (secondPilotPresent)**: Specifies whether a second pilot was present in the cockpit during the incident.
* **Presence of Hazardous Materials (presence\_hazardous\_materials)**: Indicates whether hazardous materials were onboard during the incident.
* **Safety Recommendation Present (cm\_hasSafetyRec)**: Specifies whether a safety recommendation was issued following the incident.
* **Accident Site Condition (accidentSiteCondition)**: Describes environmental and physical conditions at the accident site.
* **Phase of Flight (cicttPhaseSOEGroup)**: Identifies the phase of flight when the incident occurred (e.g., takeoff, cruise, landing).
* **Latitude (cm\_Latitude)**: Represents the geographical latitude where the incident occurred.
* **Longitude (cm\_Longitude)**: Represents the geographical longitude where the incident occurred.
  1. **Data Merging**
* The aviation dataset was merged with external datasets (GDP, aircraft departures, tech exports, and tourism receipts).
* Transformation from a wide dataset format to a long dataset format was performed where necessary to ensure consistency.
  1. **Handling Missing Values**
* Missing values were analyzed and handled using appropriate techniques.

# Exploratory Data Analysis (EDA)

* 1. **Data Distribution**
* Histograms were plotted for continuous variables to examine their distributions.

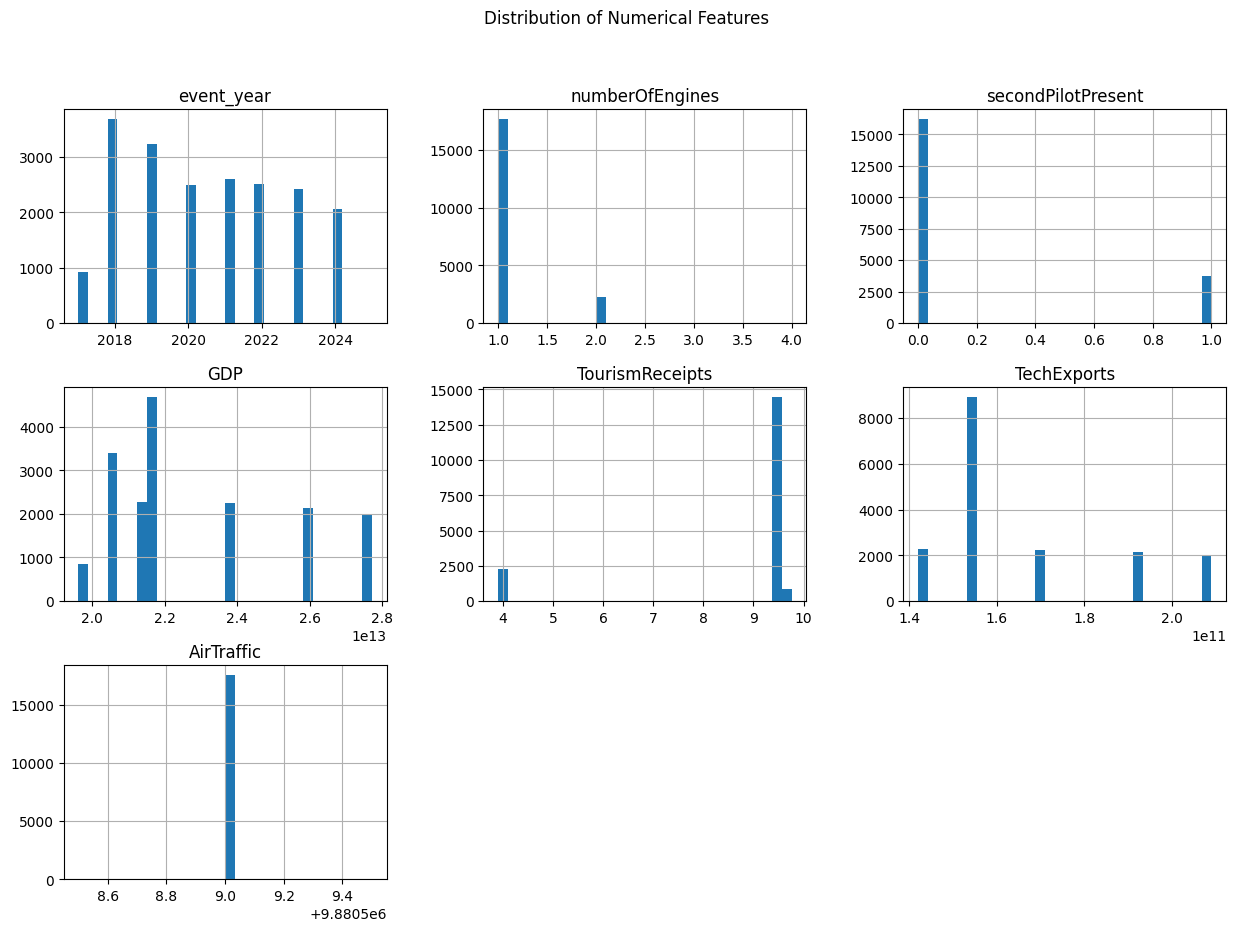


Figure 1: Data Distribution for numerical features

This step helps us with the following points:

* **Understand Data Spread & Patterns** : Helps identify how values are distributed (e.g., normal, skewed, or multimodal).
* **Detect Outliers & Anomalies** : Highlights extreme values that might affect modeling or require special handling.
* **Identify Data Imbalance** : Reveals whether certain categories or ranges dominate, which could bias analysis.
* **Set the Stage for Relationship Analysis** : Understanding each variable individually helps in later comparisons (e.g., GDP vs. incidents).
  1. **Categorical Variables**
* **The top 10 countries with the most incidents in our dataset**

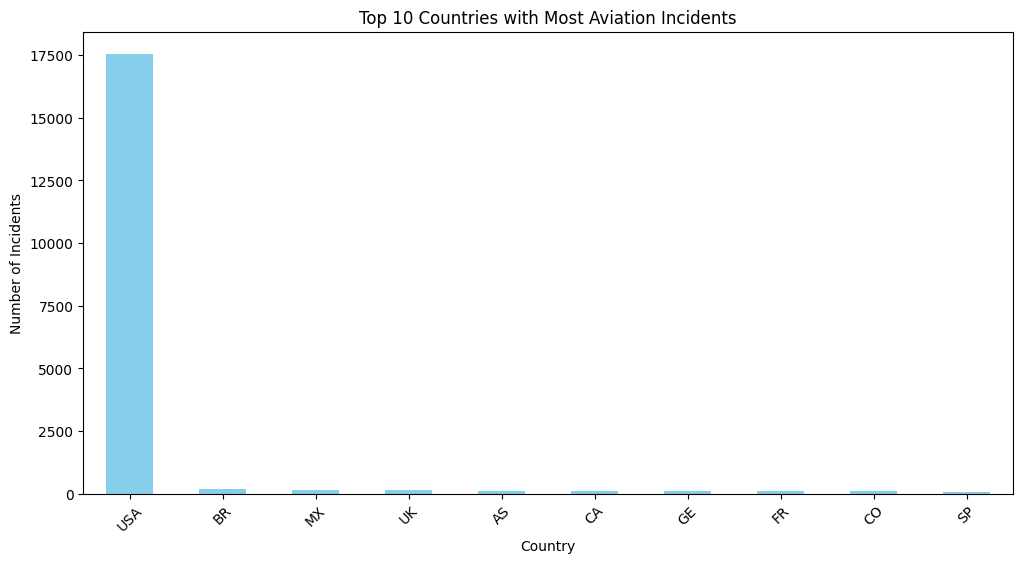


Figure 2: top 10 countries with the most incidents.

Key Insights from the Graph:

1. The USA has the highest number of incidents, but this could be due to its high air traffic volume rather than poor safety.
2. Other countries have significantly fewer incidents, which might indicate differences in reporting standards, safety regulations, or air traffic volume.

* **Distribution of damage levels was analyzed**

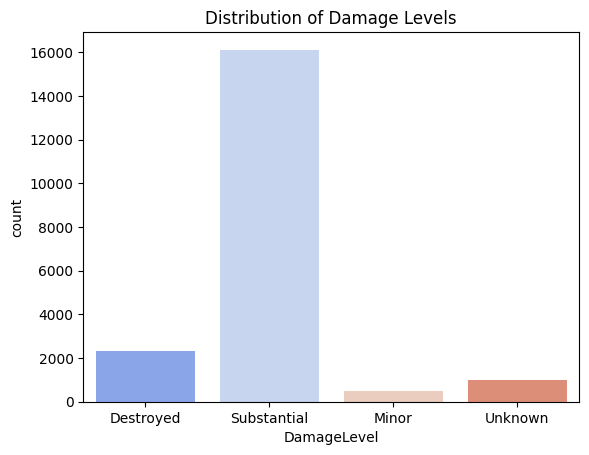


Figure 3: Distribution of damage levels

* The majority of incidents fall under the "Substantial" damage category, significantly outnumbering the other categories.
* While much lower than "Substantial," the number of incidents where aircraft were "Destroyed" is still notable.
* These categories have a much smaller count, possibly due to classification issues or reporting variations.
  1. **Relationship Between Variables**
* Boxplots were created to visualize the relationship between GDP, Tech Exports vs Damage level.

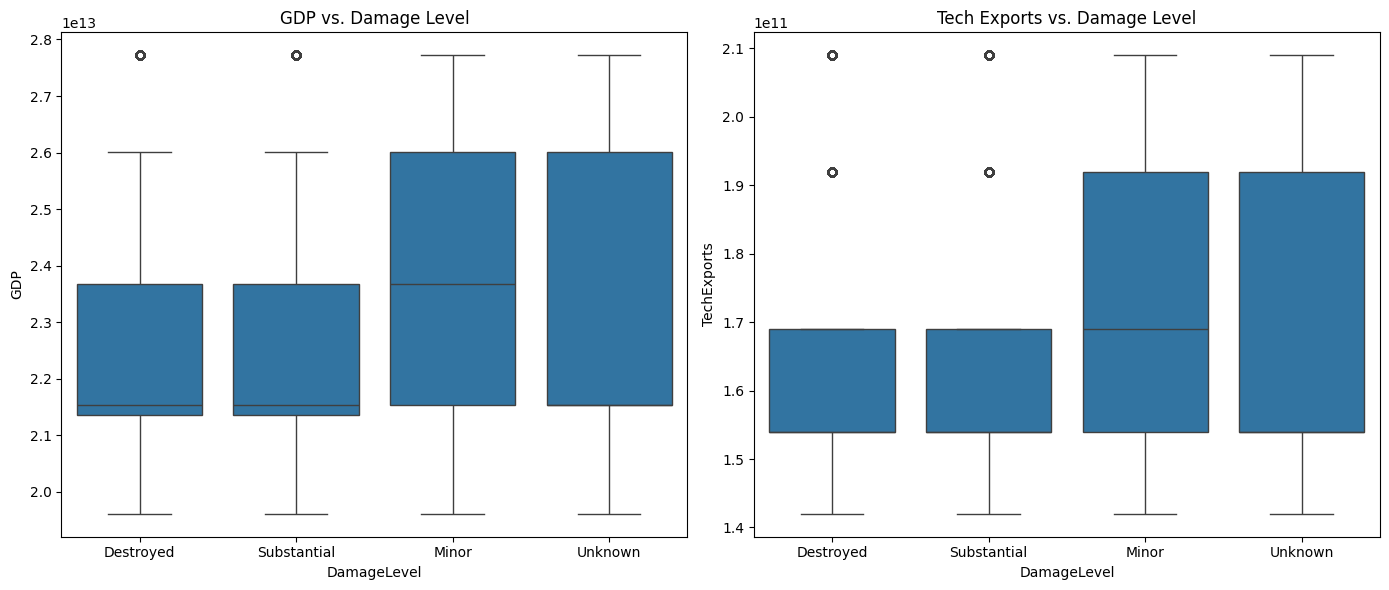


Figure 4: GDP vs. Damage Level and Tech Exports vs. Damage Level

**GDP vs. Damage Level (Left Plot)**

* Countries with higher GDP or tech exports might have more robust aviation industries, leading to more incidents recorded.
* The presence of higher GDP outliers suggests some wealthy countries still experience aviation damage events.
* The "Minor" and "Unknown" categories may contain more data points from diverse economies, leading to broader distributions.
  1. **Time Trends**
* The number of incidents over the years was plotted to analyze trends.

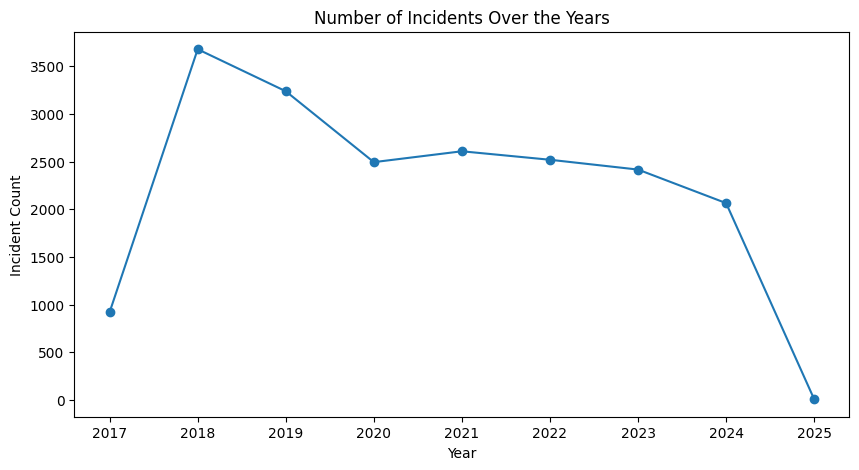


Figure 5: number of incidents per year

**Peak in 2018**: The highest number of incidents occurred in 2018, followed by a decline.

**- This peak can be due to Statistical Variability**: Aviation incidents, while generally rare, can exhibit year-to-year fluctuations. The increase in 2018 may partly reflect this natural variability rather than a systemic decline in safety standards.

**Fluctuations between 2019 and 2024**: There was a general downward trend with some stability in the middle years.

**Sharp drop in 2025**: This is likely due to incomplete data rather than an actual decrease in incidents since the year had just begun.

# Feature Encoding

Feature encoding transforms categorical data into numerical formats, enabling machine learning algorithms to process and interpret the data effectively.

* 1. **Boolean Features**

The following Boolean variables were encoded, 1 for True and 0 for False:

"amateurBuilt",

"presence\_hazardous\_materials"

"cm\_hasSafetyRec"

* 1. **Damage Level Encoding**

Damage levels were mapped to numerical values using the python map function after identifying unique values within the feature column.

"DamageLevel"

Before and after Encoding:

|  |
| --- |
| DamageLevel |
| 'Destroyed' |
| 'Substantial' |
| 'Minor' |
| 'Unknown' |

|  |
| --- |
| DamageLevel |
| 3 |
| 2 |
| 1 |
| 0 |

* 1. **Target Encoder**

The **Target Encoder** is used for encoding categorical variables by replacing each category with the mean (or another statistic) of the target variable. It is commonly used in **regression and classification tasks** when working with categorical data, it was applied on the following columns.

"aircraftCategory", "flightOperationType", "accidentSiteCondition", "cicttPhaseSOEGroup"

# Predictive Modeling

* 1. **Model Training and Score Evaluation**

The following models were used for further evaluation and comparisons based on scores.

### Random Forest Classifier

The Random Forest Classifier was used due to its ability to handle high-dimensional data and capture complex interactions between features. It works by constructing multiple decision trees and aggregating their predictions, reducing overfitting while maintaining high accuracy. Given the categorical nature of damage levels, this model was effective in classifying incidents based on various aviation and economic factors.

### Random Forest Regressor

The Random Forest Regressor was tested to explore whether damage severity could be treated as a continuous variable instead of discrete categories. However, its performance was significantly lower (0.43), indicating that a regression approach was not suitable for this problem. This suggests that damage levels have more distinct boundaries rather than forming a smooth numerical progression.

### XGBoost

XGBoost was included for its efficiency and strong performance in structured data problems. It uses gradient boosting to iteratively refine decision trees, improving predictive accuracy by minimizing errors in a sequential manner. Its accuracy (0.8506) was slightly higher than LightGBM but comparable to Random Forest, confirming its reliability in classification tasks.

### LightGBM

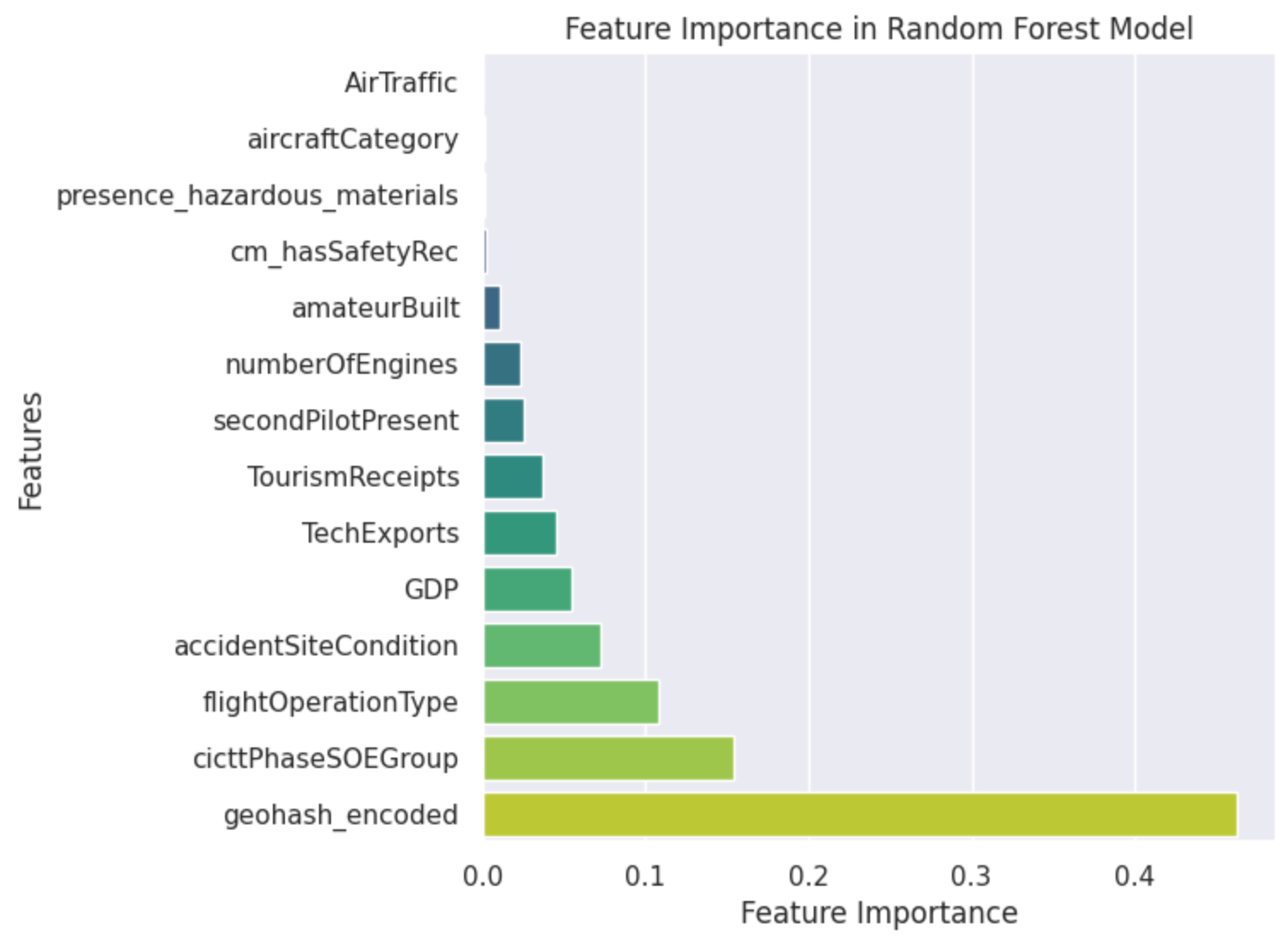
LightGBM, another gradient boosting model, was tested for its speed and ability to handle large datasets with categorical variables. It builds trees in a leaf-wise manner rather than level-wise, which allows for faster training and better handling of imbalanced data. Its accuracy (0.8438) was slightly lower than XGBoost but still competitive, making it a viable alternative for this task.

**Model Accuracy Scores**

* **Random Forest Classifier**: 0.85
* **Random Forrest Regressor**: 0.43
* **XGBoost Accuracy**: 0. 0.8506
* **LightGBM Accuracy**: 0. 0.8438

Using multiple models allowed for a comprehensive evaluation of different machine learning techniques. The comparison helped confirm that ensemble methods like Random Forest and gradient boosting approaches like XGBoost and LightGBM were well-suited for the classification problem, while regression-based methods were less effective.

* + 1. **Feature Importance Graph**



**Findings:**

* The strongest predictor is **geohash\_encoded**, indicating that location-related data plays a crucial role in determining accident severity. This suggests that certain geographical areas may have a higher likelihood of severe incidents, possibly due to environmental or infrastructural factors.
* Other significant contributors include **cicttPhaseSOEGroup** and **flightOperationType**, emphasizing that the phase of flight and operational conditions strongly influence accident outcomes. Additionally, **accidentSiteCondition** and economic indicators such as GDP and **TechExports** show some relevance, hinting at broader contextual factors affecting accident severity.
* In contrast, features like **AirTraffic**, **aircraftCategory**, and **presence\_hazardous\_materials** have minimal impact, suggesting that accident severity is not strongly linked to general traffic levels or aircraft classification. These findings provide valuable insights for risk assessment and preventive measures in aviation safety.

**Conclusion**

* The analysis successfully integrated multiple datasets to enhance aviation incident prediction.
* Feature Importance Identified the most important indicators of incident severity hinting at the effect these features have on causing aviation accidents.
* The random forest regressor model didn’t perform as good a the random forest classifier model.
* Geography has proven to play a major a role in incident occurrence, probably due to the existence of safety measures in wealthier countries.
* The performance of **XGBoost** and **LightGBM** was comparable but did not outperform Random Forest.

# References

bank, w. (2025). *air traffic.* Récupéré sur worldbank: https://data.worldbank.org/indicator/IS.AIR.PSGR

bank, w. (2025). *GDP .* Récupéré sur worldbank: https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

bank, w. (2025). *High Tech Exports.* Récupéré sur worldbank: https://data.worldbank.org/indicator/TX.VAL.TECH.MF.ZS

Brownlee, J. (2020, 08 27). *evaluate-gradient-boosting-models-xgboost-python.* Récupéré sur machinelearningmastery: https://machinelearningmastery.com/evaluate-gradient-boosting-models-xgboost-python/?utm\_source=chatgpt.com

Developers, S.-L. (2025). *Random Forests.* Récupéré sur Scikit-Learn: https://scikit-learn.org/stable/modules/ensemble.html#forest

*lightgbm-model-evaluation-metrics.* (2023, 10 06). Récupéré sur geeksforgeeks: https://www.geeksforgeeks.org/lightgbm-model-evaluation-metrics/?utm\_source=chatgpt.com

Sethi, A. (2024, 11 26). *One Hot Encoding vs Label Encoding in Machine Learning.* Récupéré sur analyticsvidhya: https://www.analyticsvidhya.com/blog/2020/03/one-hot-encoding-vs-label-encoding-using-scikit-learn/