Project: Passenger Inconvenience Classification for Flights

1. Problem Understanding

The objective of the project was to **classify flights into Easy, Medium, and Difficult** based on the **level of inconvenience caused to boarding passengers**.
Inconvenience was driven by operational factors such as:

- Delay in takeoff vs scheduled time
- Slow boarding
- Maintenance turnaround time
- Arrival-to-departure buffers

The classification was intended to **help operations teams prioritize actions**, especially for flights that would cause **maximum passenger dissatisfaction**.

2. Framing the Problem

- **Supervised ML approach**, provided a meaningful target variable could be defined

To capture passenger inconvenience objectively, we needed a **quantifiable signal** that directly reflects passenger impact.

3. Target Variable Identification

We identified **Delay Delta** as the most relevant indicator:

Delay Delta = Actual Takeoff Time − Scheduled Takeoff Time

This metric directly correlates with:

- Passenger waiting time
- Missed connections
- Overall dissatisfaction

However, the raw delay values ranged from **−100 minutes (early departures)** to **400+ minutes (severe delays)**, resulting in:

- High variance
- Skewed distribution
- Poor model learnability if treated as a raw regression target

4. Target Transformation Using Percentiles

To stabilize the target and make it **relative rather than absolute**, we transformed the delay into **percentile ranks (0–100)**.

So instead of predicting *exact delay minutes*, the model predicts **where a flight stands relative to others**.

5. Business-Driven Class Labeling

The final requirement was categorical classification:

- **Easy**
- **Medium**
- **Difficult**

This was intentionally handled through **business logic**, not model logic.

Example mapping:

- **0–25 percentile → Easy**
- **25–75 percentile → Medium**
- **75–100 percentile → Difficult**

This approach ensures:

- Labels reflect **operational tolerance**, not statistical artifacts
- Flexibility to change thresholds based on business policy
- Alignment with actionability rather than prediction accuracy alone

6. Data Preparation & Feature Engineering

- Aggregated data at **flight level**, with **Flight ID as the primary key**
- Joined multiple datasets (arrival, boarding, maintenance, turnaround metrics)
- Ensured one row per flight with all relevant operational features
- Cleaned missing values and standardized time-based features

This created a **single modeling table** suitable for supervised learning.

7. Modeling Approach

Since the transformed target (percentile) is continuous:

- We treated this as a **regression problem**
- Trained multiple regression models
- Used **R^2 score** to evaluate how well the model explained variance in relative inconvenience

After achieving stable performance on the training set, the model was validated on the **test set** to check generalization.

8. Evaluation Philosophy (Beyond R^2)

We explicitly acknowledged that:

- **Exact percentile prediction is not critical**
- Some deviation between actual and predicted percentiles is expected

What truly mattered was:

- **Correct classification into Easy / Medium / Difficult**
- Especially **high precision for the "Difficult" class**

9. Business-Focused Evaluation (Confusion Matrix Insight)

From a business standpoint:

- **Easy flights** → No action required
- **Medium flights** → Monitor / limited action
- **Difficult flights** → Immediate intervention

Hence, the most important evaluation criterion was:

- **High precision and low overlap between Easy and Difficult classes**

Even if:

- Medium overlaps slightly with Easy or Difficult → acceptable

- Easy misclassified as Difficult or vice versa → **not acceptable**

We used **confusion matrix analysis** to ensure:

- Difficult flights are correctly identified
- Operational resources are not wasted on low-impact flights

10. Final Outcome & Key Insight

- The percentile-based target made the model **stable and interpretable**
- Business-driven thresholds ensured **real-world usability**
- The solution balanced:
    - **Statistical robustness**
    - **Operational decision-making**
    - **Passenger experience impact**

This project demonstrates an end-to-end workflow where **modeling choices are guided by business intent**, not just metrics.