**NHÓM 2**

**Phân tích xu hướng ngành hàng không tại Mỹ (1993–2024)**

# Giới thiệu về bài toán và dữ liệu

**Bài toán:**

Nhóm tập trung phân tích xu hướng ngành hàng không tại Mỹ trong giai đoạn 1993–2024. Mục tiêu là tìm hiểu sự thay đổi của giá vé, lượng hành khách, thị phần các hãng hàng không và mức độ cạnh tranh giữa các hãng lớn và hãng giá rẻ. Ngoài ra, nhóm cũng hướng đến xây dựng các mô hình dự báo, phân loại, phát hiện bất thường và tối ưu hóa nhằm hỗ trợ ra quyết định trong ngành.

**Câu hỏi nghiên cứu chính:**

* Xu hướng giá vé trung bình và số lượng hành khách thay đổi ra sao theo thời gian?
* Ảnh hưởng của khoảng cách bay đến giá vé và hành khách thế nào?
* Thị phần và chiến lược cạnh tranh giữa các hãng hàng không biến động ra sao?
* Tuyến bay nào có tiềm năng phát triển và có thể được phân loại thành nhóm tương đồng?
* Có thể dự đoán giá vé và lượng hành khách trong tương lai không?
* Có tuyến bay hoặc giá vé nào bất thường không?
* Làm sao tối ưu hóa giá vé hoặc lịch bay để tăng lợi nhuận?

**Dữ liệu:**

**Nguồn:** [*US Airline Flight Routes and Fares 1993–2024*](https://www.kaggle.com/datasets/bhavikjikadara/us-airline-flight-routes-and-fares-1993-2024).

Bộ dữ liệu cung cấp thông tin chi tiết về các tuyến bay, giá vé, lượng hành khách và hãng vận hành tại Mỹ trong 3 thập kỷ.

Dữ liệu bao gồm các chỉ số như thành phố xuất phát và điểm đến, khoảng cách giữa các sân bay, số lượng hành khách, và thông tin giá vé được phân loại theo từng hãng hàng không. Đây là một nguồn dữ liệu toàn diện để phân tích xu hướng trong ngành hàng không, chiến lược định giá và cạnh tranh giữa các hãng trong suốt ba thập kỷ.

**Biến:**

* Thông tin chuyến bay: city1, city2, airport\_1, airport\_2, nsmiles.
* Thông tin hành khách: passengers.
* Thông tin giá vé: fare, fare\_lg, fare\_low.
* Thông tin hãng: carrier\_lg, carrier\_low, large\_ms, lf\_ms.
* Thông tin thời gian: Year, quarter.

**Quan sát:** Mỗi bản ghi trong dữ liệu tương ứng với một tuyến bay cụ thể trong một khoảng thời gian nhất định, kèm các chỉ số hành khách, giá vé và hãng vận hành.

# Kế hoạch phân tích dữ liệu

## 2.1 Các phương pháp dự kiến sử dụng

| **Câu hỏi nghiên cứu** | **Phương pháp phân tích** | **Mô hình/ Kỹ thuật áp dụng** |
| --- | --- | --- |
| Xu hướng giá vé trung bình và số lượng hành khách thay đổi ra sao theo thời gian? | Phân tích chuỗi thời gian (Time-series analysis) | Line chart, Moving average, ARIMA, Prophet |
| Ảnh hưởng của khoảng cách bay đến giá vé và hành khách thế nào? | Phân tích khám phá dữ liệu (Exploratory Data Analysis – EDA) | Scatter plot, Correlation analysis, Linear Regression |
| Thị phần và chiến lược cạnh tranh giữa các hãng hàng không biến động ra sao? | Phân tích thị trường (Market analysis) | Pie chart, Stacked bar chart, Comparative analysis |
| Tuyến bay nào có tiềm năng phát triển và có thể được phân loại thành nhóm tương đồng? | Phân cụm (Clustering) | K-means, Hierarchical clustering |
| Có thể dự đoán giá vé và lượng hành khách trong tương lai không? | Mô hình dự báo (Predictive modeling) | Linear Regression, Random Forest, XGBoost |
| Có tuyến bay hoặc giá vé nào bất thường không? | Phát hiện bất thường (Anomaly detection) | Isolation Forest, Local Outlier Factor (LOF), One-Class SVM |
| Làm sao tối ưu hóa giá vé hoặc lịch bay để tăng lợi nhuận? | Tối ưu hóa (Optimization) | Linear/Integer Programming, Reinforcement Learning |

## 2.2.Input (X) và Output (Y)

Input (X): Year, quarter, city1, city2, airport\_1, airport\_2, nsmiles, carrier\_lg, carrier\_low, large\_ms, lf\_ms.  
Output (Y): fare, passengers, loại giá vé, loại tuyến bay, nhãn bất thường, nhãn tối ưu.

## 2.2 Các độ đo và phương pháp đánh giá kết quả

#### **2.2.1 Mô hình dự báo (Prediction Models)**

1. Độ đo đánh giá:

* Mean Absolute Error (MAE): Đo mức sai số trung bình tuyệt đối giữa giá trị dự đoán và giá trị thực tế.
* Root Mean Squared Error (RMSE): Đo độ lệch chuẩn của sai số, phản ánh mức sai lệch lớn.
* R² Score (Coefficient of Determination): Đánh giá mức độ giải thích biến động dữ liệu của mô hình.

1. Phương pháp đánh giá:

* Chia dữ liệu thành tập huấn luyện và tập kiểm tra (Train/Test Split).
* Sử dụng Cross-validation để đảm bảo độ ổn định của mô hình.

#### **2.2.2 Mô hình phân loại (Classification Models)**

1. Độ đo đánh giá:

* Accuracy: Tỷ lệ dự đoán đúng trên tổng số dự đoán.
* Precision, Recall, F1-score: Đánh giá chi tiết hơn về hiệu quả phân loại, đặc biệt khi dữ liệu mất cân bằng.  
  Area Under Curve (AUC-ROC): Đo khả năng phân biệt giữa các lớp.

1. Phương pháp đánh giá:

* Chia dữ liệu thành tập huấn luyện và kiểm thử.
* Sử dụng Cross-validation để tối ưu tham số và đánh giá độ ổn định.

#### **2.2.3 Phát hiện bất thường (Anomaly Detection)**

1. Độ đo đánh giá:

* Precision và Recall: Đặc biệt quan trọng khi số lượng bất thường (outliers) ít.
* F1-score: Cân bằng giữa Precision và Recall.

1. Phương pháp đánh giá:

* So sánh kết quả với dữ liệu nhãn (nếu có).
* Sử dụng phương pháp trực quan (visualization) như scatter plot, box plot để đánh giá.

**2.2.4 Tối ưu hóa (Optimization)**

1. Độ đo đánh giá:

* Lợi nhuận tối đa: So sánh lợi nhuận hiện tại và lợi nhuận tối ưu.
* Hiệu quả chi phí: Đánh giá mức chi phí giảm được khi áp dụng mô hình tối ưu hóa.

1. Phương pháp đánh giá:

* So sánh kết quả mô hình với dữ liệu thực tế.
* Sử dụng mô phỏng (Simulation) để kiểm chứng.

## 

## 2.3 Nhiệm vụ các thành viên

| **STT** | **MSSV** | **Thành viên** | **Nhiệm vụ** | **Kỹ thuật/ Mô hình áp dụng** | **Output kỳ vọng** |
| --- | --- | --- | --- | --- | --- |
| 1 | 22133053 | Nguyễn Đức Cao Thăng | Phân tích xu hướng & Dự báo giá vé, hành khách | - Time-series analysis (trend, seasonality, so sánh trước/sau COVID)  - Linear Regression, Random Forest, XGBoost | - Biểu đồ xu hướng giá vé & hành khách theo năm/quý  - Dự báo giá vé & hành khách trong tương lai  - Bảng so sánh độ chính xác (MAE, RMSE, R²) |
| 2 | 22133018 | Dương Minh Hiếu | Ảnh hưởng của khoảng cách đến giá vé & hành khách | - Correlation analysis  - Linear Regression, Polynomial Regression | - Biểu đồ scatter khoảng cách – giá vé/hành khách  - Hệ số tương quan & mô hình hồi quy  - Kết luận định lượng về tác động khoảng cách |
| 3 | 22133002 | Nguyễn Thành An | Phân tích thị phần & cạnh tranh giữa các hãng.  Mô hình so sánh dự đoán | - Market share analysis (pie chart, stacked bar, time series)  - Comparative analysis giữa các hãng  - Logistic Regression | - Biểu đồ thị phần các hãng theo năm/quý  - So sánh chiến lược giá vé, lượng khách của các hãng  - Phân loại hành khách chọn hãng nào dựa trên giá vé & khoảng cách  - Kết luận về biến động cạnh tranh |
| 4 | 22133033 | Trương Trọng Đại Long | Phân loại tuyến bay theo tiềm năng phát triển | - Clustering (K-means, Hierarchical clustering)  - Classification (Logistic Regression, Decision Tree, Random Forest, Gradient Boosting) | - Nhóm tuyến bay tiềm năng theo đặc trưng (giá, khoảng cách, hành khách)  - Trực quan hóa cụm bằng biểu đồ/bản đồ  - Kết luận tuyến bay nào có tiềm năng phát triển |