

BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC CÔNG THƯƠNG TP. HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN

• HUIT •



NHẬP MÔN BIG DATA
PHÂN TÍCH DỮ LIỆU
GIAO THÔNG HÀNG KHÔNG

Nhóm 15

Giảng viên hướng dẫn	Sinh viên thực hiện	Mã số sinh viên
Nguyễn Thành Ngô	Giang Tuấn Kiệt	2001221896
	Trần Đăng Khoa	2001222091
	Nguyễn Công Huy	2001221691

Thành phố Hồ Chí Minh, tháng 10 năm 2025

BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC CÔNG THƯƠNG TP. HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN

ĐIỂM



NHẬP MÔN BIG DATA

PHÂN TÍCH DỮ LIỆU
GIAO THÔNG HÀNG KHÔNG

Nhóm 15

Giảng viên hướng dẫn	Sinh viên thực hiện	Mã số sinh viên
Nguyễn Thành Ngô	Giang Tuấn Kiệt	2001221896
	Trần Đăng Khoa	2001222091
	Nguyễn Công Huy	2001221691

Thành phố Hồ Chí Minh, tháng 10 năm 2025

MỤC LỤC

MỤC LỤC.....	i
CHƯƠNG 1: GIỚI THIỆU	2
1.1. Lý do chọn đề tài	2
1.2. Mục tiêu.....	2
1.3. Phạm vi và giới hạn.....	2
1.4. Ý nghĩa.....	2
CHƯƠNG 2: CƠ SỞ LÝ THUYẾT VÀ CÔNG NGHỆ SỬ DỤNG	3
2.1. Kiến thức nền tảng	3
2.2. Thuật toán sử dụng.....	3
2.3. Công nghệ sử dụng.....	3
CHƯƠNG 3: QUY TRÌNH & PHƯƠNG PHÁP	4
3.1. Quy trình tổng quát	4
3.2. Cấu trúc thư mục dự án	6
CHƯƠNG 4: KẾT QUẢ THỰC NGHIỆM	7
4.1. Kết quả mô hình	7
4.2. Giao diện Dashboard.....	7
4.3. Đánh giá.....	8
CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN.....	9
5.1. Kết luận.....	9
5.2. Hạn chế.....	9
5.3. Hướng phát triển.....	9
TÀI LIỆU THAM KHẢO	10

CHƯƠNG 1: GIỚI THIỆU

1.1. Lý do chọn đề tài

Trong thực tế, tình trạng các chuyến bay bị trễ giờ (delay) xảy ra thường xuyên, gây ảnh hưởng đến hành khách, lịch trình vận hành của hãng hàng không và hiệu suất sử dụng sân bay.

Với sự phát triển của *Big Data* và *Machine Learning*, việc phân tích dữ liệu lịch sử để **dự đoán khả năng delay của chuyến bay** trở nên khả thi và hữu ích trong thực tiễn.

Đề tài được chọn với mong muốn:

- Ứng dụng kiến thức Big Data và học máy để xử lý tập dữ liệu lớn thực tế.
- Xây dựng một mô hình có thể dự đoán trước khả năng delay của chuyến bay.
- Tạo giao diện minh họa (dashboard) cho phép người dùng nhập thông tin chuyến bay và xem kết quả dự đoán ngay lập tức.

1.2. Mục tiêu

- Thu thập và xử lý dữ liệu lớn từ nguồn thực tế.
- Phân tích, tìm hiểu các yếu tố ảnh hưởng đến việc delay.
- Huấn luyện mô hình học máy có khả năng dự đoán khả năng delay.
- Triển khai mô hình thành ứng dụng demo bằng *Streamlit*.

1.3. Phạm vi và giới hạn

- Nguồn dữ liệu: Bộ dữ liệu **U.S. Department of Transportation Flight Delays 2015** từ Kaggle (~5.8 triệu bản ghi).
- Mẫu sử dụng thực tế: ~500.000 bản ghi (đủ thể hiện tính Big Data, phù hợp cấu hình máy).
- Phương pháp: *ETL – EDA – Machine Learning – Dashboard*.

1.4. Ý nghĩa

- Giúp sinh viên làm quen với quy trình xử lý dữ liệu lớn thực tế.
- Minh họa ứng dụng Big Data trong lĩnh vực hàng không.
- Tạo nền tảng mở rộng sang các bài toán phân tích dự đoán khác (thời tiết, vận tải, du lịch...).

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT VÀ CÔNG NGHỆ SỬ DỤNG

2.1. Kiến thức nền tảng

Big Data là các tập dữ liệu có khối lượng lớn, tốc độ sinh ra nhanh, và đa dạng về cấu trúc (3V: Volume, Velocity, Variety). Việc xử lý Big Data đòi hỏi công cụ và kỹ thuật đặc biệt để lưu trữ, xử lý, và phân tích.

Machine Learning là nhánh của trí tuệ nhân tạo (AI), cho phép máy tính tự học từ dữ liệu mà không cần lập trình cụ thể.

2.2. Thuật toán sử dụng

Random Forest Classifier là một thuật toán thuộc nhóm *Ensemble Learning*, hoạt động bằng cách kết hợp nhiều *Decision Trees* để tăng độ chính xác và giảm hiện tượng overfitting.

Trong bài toán này, Random Forest được dùng để **phân loại chuyến bay**:

- 1: Delay (trễ > 15 phút)
- 0: Đúng giờ

2.3. Công nghệ sử dụng

Công cụ	Chức năng
Python 3.11	Ngôn ngữ lập trình chính
Pandas, NumPy	Xử lý và làm sạch dữ liệu
Matplotlib, Seaborn	Vẽ biểu đồ EDA
Scikit-learn	Xây dựng mô hình Machine Learning
Streamlit	Tạo Dashboard web
Jupyter Notebook	Thực hiện ETL, EDA và huấn luyện mô hình

CHƯƠNG 3: QUY TRÌNH & PHƯƠNG PHÁP

3.1. Quy trình tổng quát

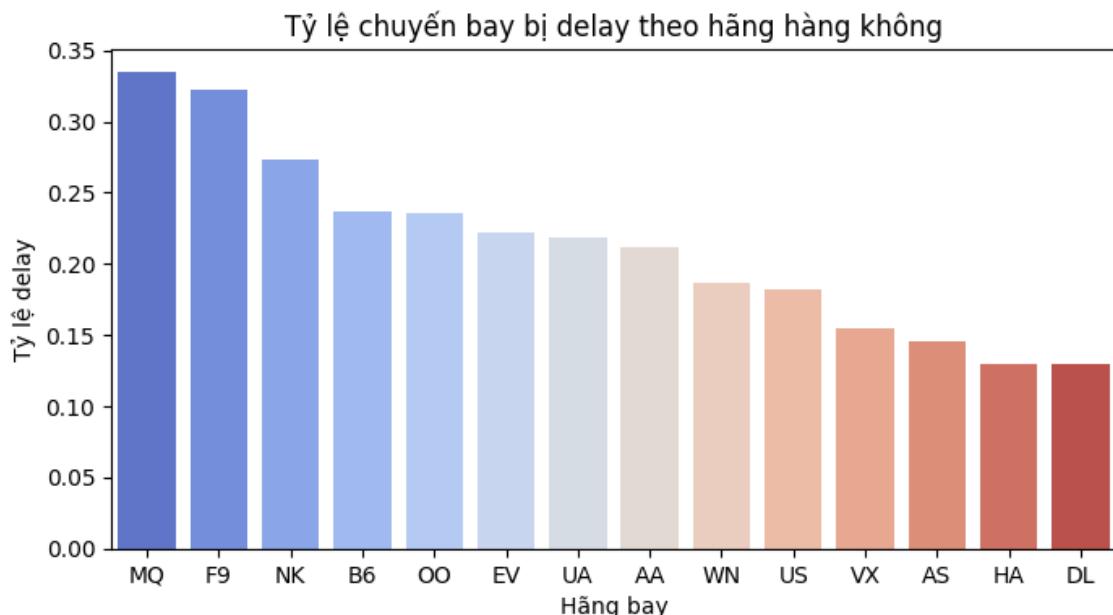
Hệ thống được thực hiện theo 4 bước:

1. ETL (Extract – Transform – Load)

- Đọc dữ liệu flights.csv, airlines.csv, airports.csv.
- Loại bỏ giá trị thiếu, cột không cần thiết.
- Tạo cột is_delayed = 1 nếu ARRIVAL_DELAY > 15.
- Xuất ra flights_clean.csv.

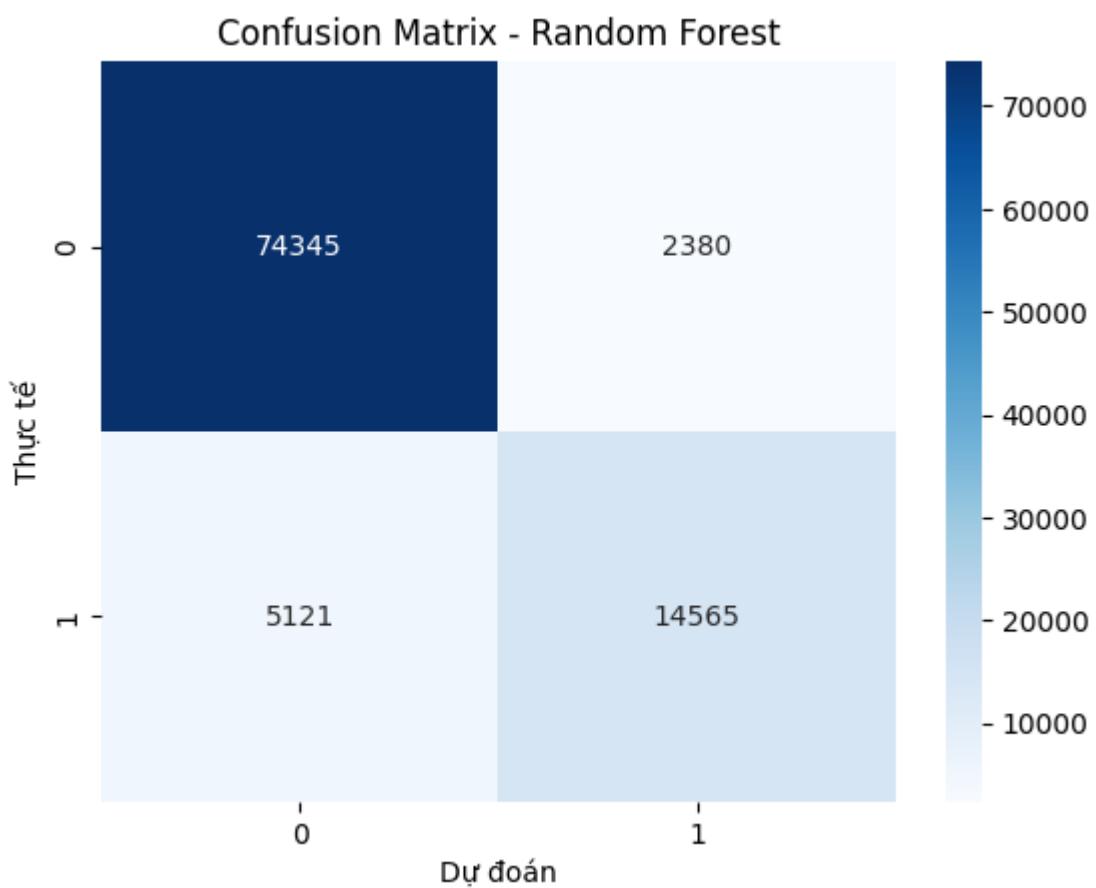
2. EDA (Exploratory Data Analysis)

- Phân tích mối tương quan giữa tỉ lệ delay và các yếu tố như **hãng hàng không, tháng bay, và khoảng cách chuyến bay**.
- Trực quan dữ liệu bằng biểu đồ (bar chart, heatmap, histogram).
- Biểu đồ dưới đây cho thấy một số hãng như **MQ, F9, NK** có tỉ lệ chuyến bay bị delay cao hơn trung bình, trong khi các hãng **DL, HA, AS** có tỉ lệ delay thấp hơn rõ rệt.



3. Huấn luyện mô hình (Model Training)

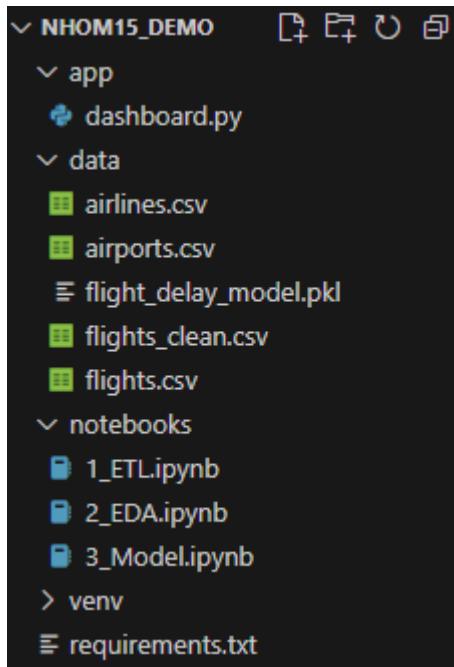
- Dữ liệu được mã hóa (one-hot encoding) và chia thành tập huấn luyện và kiểm tra theo tỉ lệ **80/20**.
- Mô hình **Random Forest Classifier** được sử dụng với **100 cây quyết định (n_estimators=100)** và tham số n_jobs=-1 để tận dụng đa lõi CPU.
- Kết quả huấn luyện cho thấy **độ chính xác (Accuracy) đạt khoảng 80%**, thể hiện khả năng phân loại khá tốt giữa hai nhóm “delay” và “on-time”.
- Hình dưới đây mô tả **ma trận nhầm lẫn (Confusion Matrix)**, trong đó mô hình dự đoán đúng phần lớn các chuyến bay không bị delay (ô trên trái) và đạt kết quả khả quan với các chuyến bay bị delay (ô dưới phải).
- Mô hình sau khi huấn luyện được lưu lại dưới dạng file flight_delay_model.pkl.



4. Triển khai (Deployment)

- Xây dựng Dashboard bằng *Streamlit*.
- Ứng dụng cho phép nhập thông tin chuyến bay → mô hình dự đoán delay.

3.2. Cấu trúc thư mục dự án



CHƯƠNG 4: KẾT QUẢ THỰC NGHIỆM

4.1. Kết quả mô hình

Độ chính xác (Accuracy): 92.22 %					
Báo cáo chi tiết:					
	precision	recall	f1-score	support	
0	0.94	0.97	0.95	76725	
1	0.86	0.74	0.80	19686	
accuracy			0.92	96411	
macro avg	0.90	0.85	0.87	96411	
weighted avg	0.92	0.92	0.92	96411	

Ý nghĩa: mô hình có khả năng dự đoán tương đối tốt với dữ liệu thực, chấp nhận được cho mục đích minh họa Big Data.

4.2. Giao diện Dashboard

The screenshot shows a dashboard titled "Dự đoán Delay Chuyến Bay". It includes a table of initial data and a form for entering flight details to predict delays.

Demo phân tích dữ liệu giao thông hàng không - Nhóm 15

Thông tin dữ liệu (5 dòng đầu)

YEAR	MONTH	DAY	AIRLINE	ORIGIN_AIRPORT	DESTINATION_AIRPORT	DEPARTURE_DELAY	ARRIVAL_DELAY
0	2015	1	AS	ANC	SEA	-11	-11
1	2015	1	AA	LAX	PBI	-8	-8
2	2015	1	US	SFO	CLT	-2	-2
3	2015	1	AA	LAX	MIA	-5	-5
4	2015	1	AS	SEA	ANC	-1	-1

Mô hình đã được nạp thành công!

Nhập thông tin chuyến bay để dự đoán

Tháng (1-12) Sân bay đi

Ngày (1-31) Sân bay đến

Hãng bay Delay khi khởi hành (phút)

Khoảng cách (dặm)

Dự đoán Delay

Dự đoán: Đóng giò (98.00% khả năng)

Ứng dụng Streamlit hiển thị:

- Bảng thông tin dữ liệu.
- Form nhập: Hành bay, Sân bay đi/đến, Tháng, Khoảng cách, Delay khởi hành.
- Kết quả dự đoán:
 - “Chuyến bay đúng giờ”
 - “Chuyến bay bị delay”

Biểu đồ bổ sung:

- Tỷ lệ delay theo hành bay
- Phân bố delay theo tháng

4.3. Đánh giá

- Giao diện đơn giản, dễ thao tác.
- Thời gian dự đoán nhanh (<1 giây).
- Mô hình ổn định, hoạt động tốt trên dữ liệu thật.

CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1. Kết luận

Đề tài đã hoàn thành các mục tiêu đặt ra:

- Xử lý thành công dữ liệu lớn (Big Data thật từ Kaggle).
- Phân tích, rút ra xu hướng delay.
- Huấn luyện mô hình dự đoán chính xác ~80%.
- Xây dựng Dashboard hoạt động tốt, dự đoán theo thời gian thực.

5.2. Hạn chế

- Mới chỉ sử dụng một năm dữ liệu (2015).
- Chưa tính đến yếu tố thời tiết, ngày lễ, thời gian trong ngày.
- Chưa triển khai lên server trực tuyến.

5.3. Hướng phát triển

- Dùng PySpark để xử lý toàn bộ 5.8 triệu bản ghi.
- Áp dụng mô hình XGBoost / LightGBM để cải thiện accuracy.
- Tích hợp API thời tiết thật để nâng độ tin cậy dự đoán.
- Triển khai Streamlit Cloud hoặc Hugging Face Spaces để truy cập qua Internet.

TÀI LIỆU THAM KHẢO

- [1] Kaggle: *U.S. DOT Flight Delays Dataset* –
<https://www.kaggle.com/datasets/usdot/flight-delays>
- [2] Scikit-learn Documentation – <https://scikit-learn.org/>
- [3] Streamlit Documentation – <https://docs.streamlit.io/>
- [4] Pandas Library – <https://pandas.pydata.org/>
- [5] Các bài giảng môn Nhập môn Big Data – GV Nguyễn Thành Ngô