A picture containing logo

Description automatically generated**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

~~~~~\*~~~~~

**BÁO CÁO DỰ ÁN CUỐI KÌ**

**BỘ MÔN: NHẬP MÔN KHOA HỌC DỮ LIỆU**

**Đề tài:** Nâng cấp dịch vụ hãng hàng không bằng xây dựng mô hình dự đoán khách hàng hài lòng và không hài lòng.

**Giảng viên hướng dẫn: Hà Văn Thảo**

**Nhóm 12:**

**Tên các thành viên MSSV**

Nguyễn Xuân Tuấn 20280112

Đào Thanh Nguyên 20280068

Lại Toàn Thắng 20280083

# Giới thiệu

Ngày nay, việc đi lại rất phổ biến với các công ty hàng không và vì nó là phổ biến nên sự hài lòng của khách hàng là một trong những yếu tố quyết định quan trọng nhất. Nó cũng là một yếu tố khác liên quan đến lợi nhuận kinh tế của các công ty và vì điều này rất quan trọng, các nghiên cứu về chủ đề này đã và đang phát triển từng ngày.

Bằng cách sử dụng kết quả khảo sát sự hài lòng của khách hàng, chúng ta có thể dự đoán các tính năng thiết yếu nhất dẫn đến sự hài lòng của khách hàng như sự thoải mái, dịch vụ nhận phòng, v.v. Nếu chúng ta có thể dự đoán xếp hạng của hành khách, chúng ta có thể dự đoán những tính năng nào là quan trọng và cần được cải thiện.

# Thông tin về bộ dữ liệu:

Tên bộ dữ liệu: Invistico\_Airline.csv.

Bộ dữ liệu được lấy từ Kaggle và người public là TJ Klein

Bộ dữ liệu bao gồm các kết quả của sự khảo sát hài lòng của hành khách hàng không Hoa Kỳ

Có 23 features và 129880 mẫu dữ liệu trong Dataset

Bộ dữ liệu cung cấp từ một tổ chức hàng không trên thế giới về thông tin bay của khách hàng, đánh giá của khách hàng về các dịch vụ, những sự cố khi xảy ra

Mục đích chính của bộ dữ liệu này là dự đoán liệu một khách hàng trong tương lai có hài lòng với dịch vụ của họ với các chi tiết của các giá trị tham số khác hay không.

Tên cột dự đoán là satisfaction và các cột còn lại lần lượt là Gender, Customer Type, Age, Type of Trave, Class, Flight Distance, Seat comfort, Departure/Arrival time convenient, Food and drink, Gate location, Inflight wifi service, Inflight entertainment, Online support, Ease of Online booking, On-board service, Leg room service,Baggage handling, Checkin service, Cleanliness, Online boarding, Departure Delay in Minutes, Arrival Delay in Minutes.

# Nội dung chính:

## Chuẩn bị dữ liệu.

* Nhập dữ liệu.
* Nhìn nhanh qua dữ liệu:

Text

Description automatically generated

Text

Description automatically generated

* Chuẩn hóa tên cột: Chuyển kí tự khoảng trống có trong tên cột thành kí tự \_
* Khám phá giá trị rỗng: Cột Arrival\_Delay\_in\_minutes có 393 dữ liệu rỗng, chiếm 0.3026%

Table

Description automatically generated with low confidence

* Chuyển các giá trị trong côt satisfaction : satisfied: 1, disatisfied: 0.

## Khám phá và trực quan hóa dữ liệu.

* 1. Vẽ biểu đồ countplot cho cột satisfaction (cột mục tiêu):

Chart, bar chart, treemap chart

Description automatically generated

***Nhận xét:*** *Tỉ lệ 0 và 1 khá ngang nhau, không có sự chênh lệch nhiều giữa 2 giá trị dự đoán.*

* 1. Trực quan những thuộc tính của khách hàng :

Chart, bar chart

Description automatically generated

2.3. Vẽ biểu đồ cho các thuộc tính đánh giá từ 0 đến 5:   
- Những thuộc tính có chỉ số corr cao so với satisfation:

Chart, treemap chart

Description automatically generated***Nhận xét:*** *Có sự phân tách rõ ràng, đánh giá 4,5 sao thì thường là hài lòng và 1,2,3 sao thì là không hài lòng.*

* Những biểu đồ có chỉ số corr thấp so với satisfaction:

Chart, bar chart

Description automatically generated

***Nhận xét:*** *Không có sự rõ ràng giữa những đánh giá 4,5 sao và 1,2,3 sao, nhiều khách hàng hài lòng với hãng hàng không nhưng lại đánh giá những dịch vụ trên 1,2,3 sao. Chứng minh rằng những dịch vụ này không ảnh hưởng nhiều đến sự đánh giá khách hàng.*

2.4. Vẽ corr\_matrix():

Chart

Description automatically generated Quan sát thấy Arrival\_Delay\_in\_Minutes tương quan cao Departure\_Delay\_in\_Minutes 0.97 ~ 1 kết hợp 2 cột Arrival\_Delay\_in\_Minutes và Departure\_Delay\_in\_Minutes thành cột Delay\_in\_Minutes

## Tiền xử lí dữ liệu.

1. Chia dữ liệu huấn luyện, đánh giá, kiểm tra.

* Dùng phương pháp lấy mẫu phân tần theo tuổi.
* Viết pipeline cho tiền xử lí dữ liệu:
* Dữ liệu số: Dùng MinMaxScaler() để đưa dữ liệu nằm trong khoảng (0,1).
* Dữ liệu phân loại:

+ OrdinalEncoder(): Class.

+ OneHotEncoder(): Gender, Customer\_Type, Type\_of\_Travel.

Graphical user interface, text, application, email

Description automatically generated

## Lựa chọn mô hình và huấn luyện một mô hình có khả năng nhất.

* 1. Logistic Regression

Mô hình không thể hiện độ tương thích với tập dữ liệu cao, nhưng cũng không quá chệch, tham số của mô hình vừa đủ thể hiện độ quan trọng của các tính năng so với các mô hình khác.

* 1. Neural Network

Với 3 hidden layers với số unit tương ứng la2 256, 256 và 128 ta thu được mô hình có độ chính xác khá cao. Tuy nhiên, các tham số của mô hình nhìn chung vẫn còn khó giải thích nên cần sử dụng thêm các mô hình khác để đối chiếu.

* 1. RandomForest

Sử dụng GridSearch để tìm các siêu tham số phù hợp với chương trình và trích xuất các tính năng quan trọng của mô hình tương ứng với các thuộc tính của tập dữ liệu để so sánh với mô hình Logistic.

## Tinh chỉnh mô hình.

Table

Description automatically generated

## Đánh giá mô hình và rút ra kết luận cho đề tài.

Chart

Description automatically generated

*Từ phần trực quan và tham số của mô hình, ta thấy rằng chưa làm được hài lòng các khách hàng mới và các dịch vụ chưa tốt làm ảnh hưởng đến độ hài lòng như đồ ăn, thức uống và độ delay của chuyến bay. Độ chênh lệch về sự hài lòng giữa các loại vé còn khá lớn và đang nghiêng về phần không hài lòng, cần xem xét và đưa ra các phương án cải thiện cho từng loại vé khác nhau.*

# KẾT LUẬN VÀ ĐÁNH GIÁ

Tập dữ liệu thể hiện khá chính xác và trung thực với độ hài lòng của khách hàng ứng với các đánh giá trên các dịch vụ được cung cấp, bằng chứng là các mô hình thể hiện độ chính xác khá cao trên cả 3 tập huấn luyện – phát triển và kiểm thử. Từ đó có thể đưa ra các nhận xét, đánh giá tổng quan và tìm ra được các vấn đề đang đối mặt cũng như giải pháp khắc phục cho doanh nghiệp.