TRƯỜNG ĐẠI HỌC KINH TẾ

**KHOA THỐNG KÊ – TIN HỌC**





**BÁO CÁO GIỮA KÌ**

**MÔN HỌC: MACHINE LEARNING**

**ỨNG DỤNG MÔ HÌNH HỌC MÁY DỰ ĐOÁN SỰ HÀI LÒNG CỦA KHÁCH HÀNG HÃNG HÀNG KHÔNG**

Sinh viên thực hiện : **Trần Thị Hằng**

: **Hoàng Nghĩa Đức**

: **Nguyễn Đức Thắng**

Lớp : **47K21.1**

Giảng viên bộ môn : **Ts. Nguyễn Văn Chức**

**Đà Nẵng, 202**

**BẢNG PHÂN CÔNG NHIỆM VỤ**

| **STT** | **Họ và tên** | **Lớp** | **Công việc đảm nhận** | **Phần trăm đóng góp** |
| --- | --- | --- | --- | --- |
| 1 | Trần Thị Hằng (**Nhóm trưởng**) | 47K21.1 | Tiền xử lý, Xây dựng mô hình | 33.3% |
| 2 | Hoàng Nghĩa Đức | 47K21.1 | Tiền xử lý, Xây dựng mô hình | 33.3% |
| 3 | Nguyễn Đức Thắng | 47K21.1 | Trực quan hóa mô hình, Xây dựng mô hình | 33.3% |

DANH MỤC HÌNH ẢNH

[Hình 1: Mô hình Random Forest 2](#_heading=h.17dp8vu)

[Hình 2: Mô hình XGboost 4](#_heading=h.26in1rg)

[Hình 3: Thông tin dữ liệu 7](#_heading=h.2jxsxqh)

[Hình 4: Thống kê mô tả dữ liệu 7](#_heading=h.z337ya)

[Hình 5: Số lượng khách hàng theo giới tính 9](#_heading=h.2xcytpi)

[Hình 6: Sự hài lòng theo giới tính của khách hàng 10](#_heading=h.1ci93xb)

[Hình 7: Sự hài lòng theo loại khách hàng 11](#_heading=h.3whwml4)

[Hình 8: Sự hài lòng theo hạng vé 12](#_heading=h.2bn6wsx)

[Hình 9: Framework cho dự án 13](#_heading=h.1pxezwc)

[Hình 10: Tiền xử lý dữ liệu 13](#_heading=h.2p2csry)

[Hình 11: Đọc dữ liệu từ file CSV 14](#_heading=h.3o7alnk)

[Hình 12: Thông tin kiểm tra của các cột trong dữ liệu 14](#_heading=h.23ckvvd)

[Hình 13: Kiểm tra dữ liệu có thiếu giá trị hay không 15](#_heading=h.32hioqz)

[Hình 14: Kết quả xử lý giá trị thiếu 16](#_heading=h.1hmsyys)

[Hình 15: Biểu đồ tương quan của num\_features 19](#_heading=h.2grqrue)

[Hình 16: Biểu đồ tương quan của ob\_features 19](#_heading=h.vx1227)

[Hình 17: Kiểu dữ liệu đầu vào 21](#_heading=h.1v1yuxt)

[Hình 18: Kết quả đánh giá của mô hình Random Forest 22](#_heading=h.3tbugp1)

[Hình 19: Biểu đồ Roc của mô hình Random Forest 23](#_heading=h.28h4qwu)

[Hình 20: Kết quả đánh giá của mô hình XGboost 24](#_heading=h.46r0co2)

[Hình 21: Biểu đồ Roc của mô hình XGboost 25](#_heading=h.2lwamvv)

[Hình 22: Mối tương quan giữa các biến đầu vào và biến mục tiêu 'satisfaction' 27](#_heading=h.2zbgiuw)

# DANH MỤC BẢNG BIỂU

[Bảng 1: Bảng dữ liệu 9](#_heading=h.1y810tw)

[Bảng 2: Bảng so sánh các chỉ sổ các các mô hình 26](#_heading=h.3l18frh)

# MỤC LỤC

[**DANH MỤC HÌNH ẢNH ii**](#_heading=h.gjdgxs)

[**DANH MỤC BẢNG BIỂU iii**](#_heading=h.30j0zll)

[**MỤC LỤC iv**](#_heading=h.3znysh7)

[**CHƯƠNG 1.**](#_heading=h.2et92p0) **TỔNG QUAN VỀ DỰ ÁN VÀ CƠ SỞ LÝ THUYẾT 1**

[**1.1.**](#_heading=h.tyjcwt) **Tổng quan về dự án 1**

[*1.1.1.*](#_heading=h.3dy6vkm) ***Đặt vấn đề*** *1*

[*1.1.2.*](#_heading=h.1t3h5sf) ***Mô tả dự án*** *1*

[**1.2.**](#_heading=h.4d34og8) **Lý thuyết các mô hình dự đoán 2**

[*1.2.1.*](#_heading=h.2s8eyo1) ***Mô hình Random Forest*** *2*

[*1.2.2.*](#_heading=h.3rdcrjn) ***Mô hình XGBoost*** *3*

[*1.2.3.*](#_heading=h.lnxbz9) ***GridSearchCV là gì*** *4*

[**1.3.**](#_heading=h.35nkun2) **Các chỉ số đánh giá mô hình 5**

[**CHƯƠNG 2.**](#_heading=h.1ksv4uv) **MÔ TẢ VÀ TRỰC QUAN HÓA DỮ LIỆU 7**

[**2.1.**](#_heading=h.44sinio) **Thông tin bộ dữ liệu 7**

[**2.2.**](#_heading=h.3j2qqm3) **Giải thích dữ liệu 8**

[**2.3.**](#_heading=h.4i7ojhp) **Trực quan hóa dữ liệu 10**

[**CHƯƠNG 3.**](#_heading=h.qsh70q) **XÂY DỰNG MÔ HÌNH DỰ ĐOÁN SỰ HÀI LÒNG CỦA KHÁCH HÀNG 14**

[**3.1.**](#_heading=h.3as4poj) **Frame work 14**

[**3.2.**](#_heading=h.49x2ik5) **Tiền xử lý dữ liệu 14**

[*3.2.1.*](#_heading=h.147n2zr) ***Kiểm tra dữ liệu tổng quát*** *14*

[*3.2.2.*](#_heading=h.ihv636) ***Làm sạch dữ liệu*** *15*

[*3.2.3.*](#_heading=h.41mghml) ***Giảm chiều dữ liệu*** *19*

[*3.2.4.*](#_heading=h.3fwokq0) ***Chuyển đổi dữ liệu*** *21*

[**3.3.**](#_heading=h.4f1mdlm) **Mô hình Random Forest 22**

[*3.3.1.*](#_heading=h.2u6wntf) ***Xây dựng mô hình*** *22*

[*3.3.2.*](#_heading=h.19c6y18) ***Đánh giá mô hình*** *23*

[**3.4.**](#_heading=h.nmf14n) **Mô hình XGBoost 24**

[*3.4.1.*](#_heading=h.37m2jsg) ***Xây dựng mô hình*** *24*

[*3.4.2.*](#_heading=h.1mrcu09) ***Đánh giá mô hình*** *25*

[**3.5.**](#_heading=h.111kx3o) **Đánh giá các mô hình 27**

[**CHƯƠNG 4.**](#_heading=h.4k668n3) **CHIẾN LƯỢC DỰ ÁN 28**

[**TÀI LIỆU THAM KHẢO 30**](#_heading=h.1egqt2p)

[**PHỤ LỤC 31**](#_heading=h.3ygebqi)

1. **TỔNG QUAN VỀ DỰ ÁN VÀ CƠ SỞ LÝ THUYẾT**

## Tổng quan về dự án

* + 1. **Đặt vấn đề**

Trong ngành công nghiệp hàng không, sự hài lòng của khách hàng có ảnh hưởng rất lớn tới việc khách hàng có tiếp tục sử dụng các dịch vụ của hãng hàng không đó nữa hay không. Việc khách hàng rời đi không chỉ làm giảm doanh thu mà còn ảnh hưởng đến hình ảnh và uy tín của hãng hàng không. Để giảm thiểu tình trạng này, các hãng hàng không ngày càng chú trọng vào việc phân tích và dự đoán sự hài lòng của khách hàng. Bằng cách xác định sớm, hãng hàng không có thể triển khai các chiến lược cải thiện chất lượng chuyến bay như cung cấp các gói dịch vụ ưu đãi mới .... Điều này giúp giảm thiểu tỷ lệ không hài lòng và duy trì lượng khách hàng ổn định.

AI và học máy là những công cụ mạnh mẽ giúp tối ưu hóa quá trình phân tích và dự đoán. Các mô hình học máy có thể học từ dữ liệu lịch sử để nhận diện các yếu tố ảnh hưởng đến đánh giá cảm nhận của khách hàng. Chúng có thể dự đoán chính xác nhóm khách hàng có nguy cơ cao và những yếu tố ảnh hưởng chính đến khách hàng nhằm đưa ra các dịch vụ, chính sách cải thiện kịp thời.

* + 1. **Mô tả dự án**

**Mục đích:** Ứng dụng các mô hình machine learning dự đoán sự hài lòng của khách hàng dựa trên các yếu tố liên quan được cung cấp trong tập dữ liệu.

**Mục tiêu:** Dự đoán khả năng một khách hàng tương lai sẽ hài lòng hay không với dịch vụ của hãng dựa trên các thông tin chi tiết từ các thông số khác nhau. Bên cạnh đó, hãng hàng không cũng cần xác định những khía cạnh nào trong dịch vụ cần được chú trọng hơn để nâng cao mức độ hài lòng của khách hàng.

**Đầu vào:** tập dữ liệu đầu vào cho bài toán này bao gồm các thông tin sau đây về khách hàng: Satisfaction, Gender, Customer Type, Age, Type of Travel, 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.

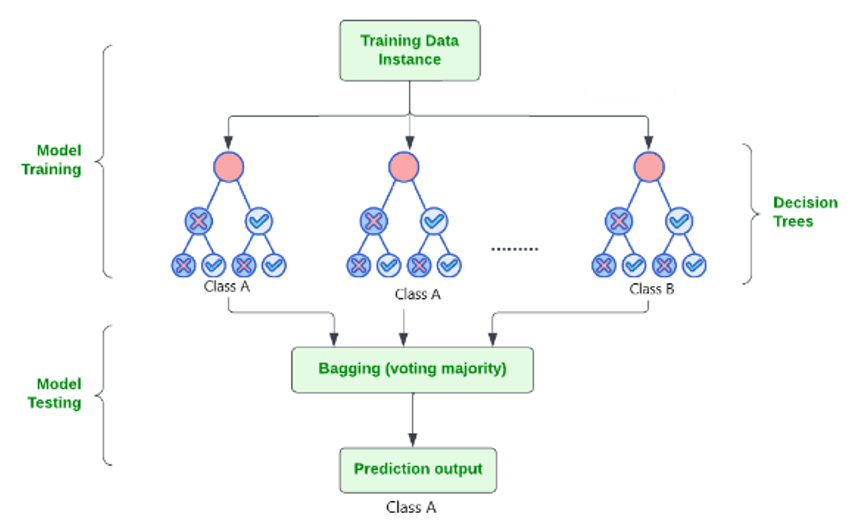
**Đầu ra:** Dự đoán sự hài lòng của mỗi khách hàng với hãng hàng không trong tương lai.

## Lý thuyết các mô hình dự đoán

* + 1. **Mô hình Random Forest**

1. **Khái niệm mô hình**

Random Forest là một thuật toán học máy ensemble sử dụng nhiều cây quyết định (decision tree) để thực hiện các nhiệm vụ phân loại và hồi quy. Thuật toán này hoạt động bằng cách xây dựng một tập hợp lớn các cây quyết định được huấn luyện trên các tập dữ liệu con ngẫu nhiên được lấy từ tập dữ liệu gốc. Mỗi cây quyết định được huấn luyện để phân loại hoặc dự đoán một giá trị cho một điểm dữ liệu mới, sau đó kết quả dự đoán của tất cả các cây quyết định được kết hợp để đưa ra dự đoán cuối cùng. [1]



Hình 1: Mô hình Random Forest

1. **Thuật toán hoạt động của Random Forest**

Bước 1: Chọn ngẫu nhiên K mẫu dữ liệu từ tập dữ liệu huấn luyện.

Bước 2: Xây dựng các cây quyết định tương ứng với các mẫu dữ liệu được chọn (các tập con).

Bước 3: Quyết định số lượng N cây quyết định bạn muốn xây dựng.

Bước 4: Lặp lại Bước 1 và 2 N lần.

Bước 5: Đối với các mẫu dữ liệu mới, tìm dự đoán của từng cây quyết định và gán các mẫu dữ liệu mới vào nhóm nhận được nhiều phiếu bầu (dự đoán) nhất.

Lưu ý:

* K là một số được chọn ngẫu nhiên mà không cần đặt trước, thường nhỏ hơn tổng số mẫu dữ liệu trong tập huấn luyện.
* Bước lặp lại (Bước 4) giúp tạo ra sự đa dạng giữa các cây quyết định, cải thiện độ chính xác của mô hình Random Forest.

1. **Ưu và nhược điểm của mô hình**

**+ Ưu điểm:**

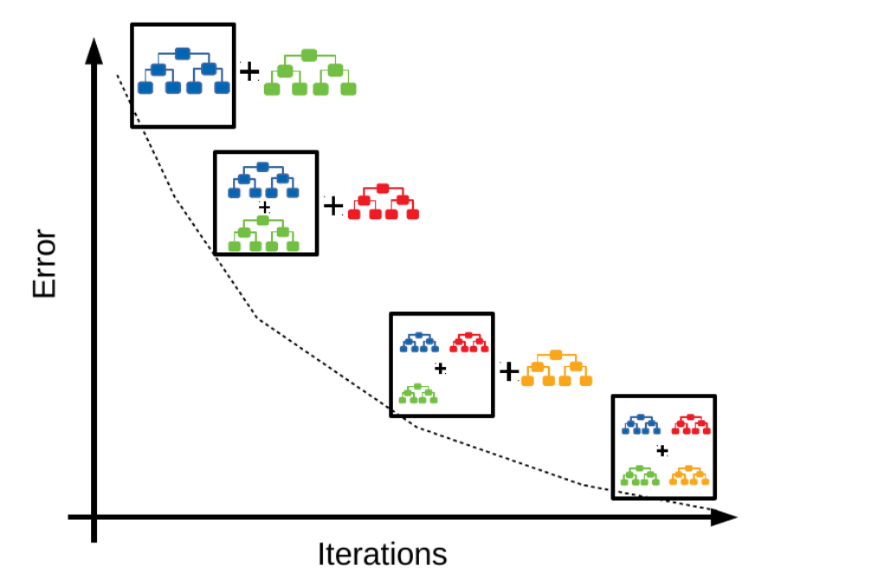
* Random Forest đạt được độ chính xác và tính mạnh mẽ nhờ việc kết hợp nhiều cây quyết định. Các dự đoán trung bình từ các cây này giúp loại bỏ các sai lệch của từng cây riêng lẻ, dẫn đến kết quả tổng thể chính xác hơn.
* Random Forest có khả năng chống overfitting do tính chất trung bình các dự đoán từ nhiều cây quyết định khác nhau.
* Random Forest có thể xử lý tốt dữ liệu thiếu bằng cách tận dụng các thuộc tính có sẵn để dự đoán.
* Đánh giá tầm quan trọng của biến: Điều này hỗ trợ trong việc lựa chọn thuộc tính và diễn giải vai trò của các yếu tố ảnh hưởng tới dự đoán.

**+ Nhược điểm:**

* Tốc độ tính toán chậm hơn so với thuật toán sử dụng một mô hình duy nhất.
* Khó lý giải cách thức đưa ra dự đoán, hạn chế hiểu biết về mô hình và ảnh hưởng đến độ tin tưởng kết quả.
* Lựa chọn tham số không phù hợp ảnh hưởng đến hiệu suất mô hình.
  + 1. **Mô hình XGBoost**

**(a) Khái niệm mô hình**

XGBoost hoạt động dựa trên việc kết hợp nhiều mô hình cây quyết định (decision tree) yếu thành một mô hình mạnh hơn. Mỗi mô hình cây được xây dựng dựa trên lỗi dự đoán của các mô hình trước đó, giúp cải thiện dần dần độ chính xác của mô hình tổng thể. [2]



Hình 2: Mô hình XGboots

**(b) Quy trình hoạt động**

1. Khởi tạo: Bắt đầu với một mô hình cây quyết định đơn giản.
2. Lặp:

* Huấn luyện: Huấn luyện một mô hình cây quyết định mới để sửa chữa lỗi dự đoán của mô hình hiện tại.
* Cập nhật: Cập nhật trọng số của các mô hình cây quyết định để mô hình mới có ảnh hưởng lớn hơn.

1. Dự đoán: Dự đoán cho một ví dụ mới bằng cách kết hợp dự đoán của tất cả các mô hình cây quyết định, với trọng số tương ứng.

**(c) Ưu nhược điểm của mô hình**

**+ Ưu điểm:**

* Độ chính xác cao: XGBoost thường đạt được độ chính xác cao hơn so với các thuật toán học máy khác.
* Hiệu quả tính toán: XGBoost có thể được tối ưu hóa để chạy hiệu quả trên cả CPU và GPU.
* Khả năng mở rộng: XGBoost có thể xử lý được tập dữ liệu lớn và phức tạp.
* Dễ sử dụng: XGBoost có nhiều thư viện và API dễ sử dụng cho nhiều ngôn ngữ lập trình khác nhau.

**+ Nhược điểm:**

* Có thể bị quá khớp: XGBoost có thể học quá tốt dữ liệu huấn luyện dẫn đến hiệu suất kém trên dữ liệu mới.
* Khó giải thích: Mô hình XGBoost có thể khó giải thích do cấu trúc phức tạp của nó.
  + 1. **GridSearchCV là gì**

GridSearch là một kỹ thuật trong máy học dùng để tìm kiếm và tối ưu hóa các siêu tham số (hyperparameters) của một mô hình học máy. Mục tiêu của GridSearch là chọn ra tổ hợp các siêu tham số tốt nhất nhằm cải thiện hiệu suất của mô hình.

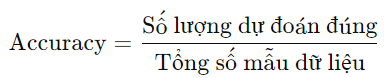
Cách thức hoạt động của GridsearchCv:

* Xác định bộ siêu tham số của mô hình cần tối ưu
* Truyền danh sách các giá trị thử ngiệm
* Thử nghiệm dựa vào đầu vào cung cấp và đánh giá
* Lựa chọn bộ siêu tham số tối ưu nhất

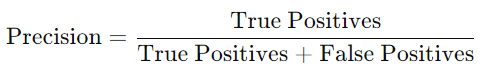
Trong Python, GridSearch thường được thực hiện bằng cách sử dụng GridSearchCV từ thư viện scikit-learn. Hàm này tự động kết hợp GridSearch với cross-validation để đảm bảo rằng mô hình không bị overfitting và các siêu tham số được tối ưu hóa một cách chính xác. [3]

## Các chỉ số đánh giá mô hình

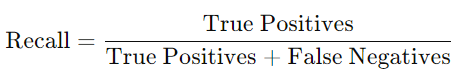
 **Accuracy (Độ chính xác)**: Tỷ lệ dự đoán đúng trên tổng số mẫu dữ liệu. Công thức:



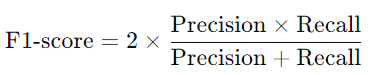
 **Precision (Độ chính xác của lớp dương)**: Tỷ lệ khách hàng được dự đoán là sẽ rời bỏ dịch vụ thực sự rời bỏ dịch vụ. Công thức:

​

 **Recall (Độ phủ của lớp dương ):** Tỷ lệ khách hàng thực sự rời bỏ dịch vụ được dự đoán là sẽ rời bỏ dịch vụ. Công thức:



 **F1-score**: Trung bình điều hòa giữa Precision và Recall, là một chỉ số tổng hợp giữa hai chỉ số trên. Công thức:



 **AUC (Area Under the ROC Curve)**: Diện tích dưới đường cong ROC. Một chỉ số tổng hợp cho biết mức độ mà mô hình phân loại phân biệt được giữa các lớp. [4]

1. **MÔ TẢ VÀ TRỰC QUAN HÓA DỮ LIỆU**

## Thông tin bộ dữ liệu

Tên bộ dữ liệu: Invistico\_Airline

Dữ liệu này được cung cấp bởi một tổ chức hàng không. Tên thật của công ty không được nêu do nhiều mục đích khác nhau, đó là lý do tại sao có tên hãng hàng không Invistico.

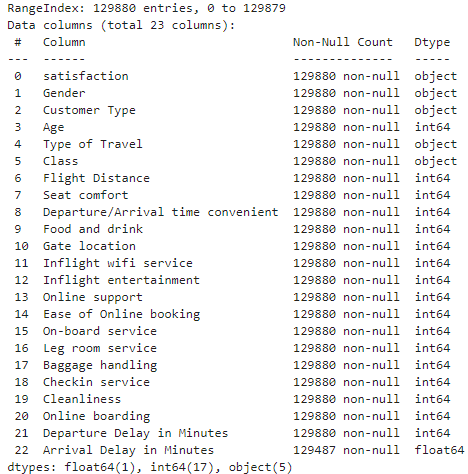
Tập dữ liệu bao gồm thông tin chi tiết về những khách hàng đã bay cùng họ. Phản hồi của khách hàng về các bối cảnh khác nhau và dữ liệu chuyến bay của họ đã được tổng hợp.

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

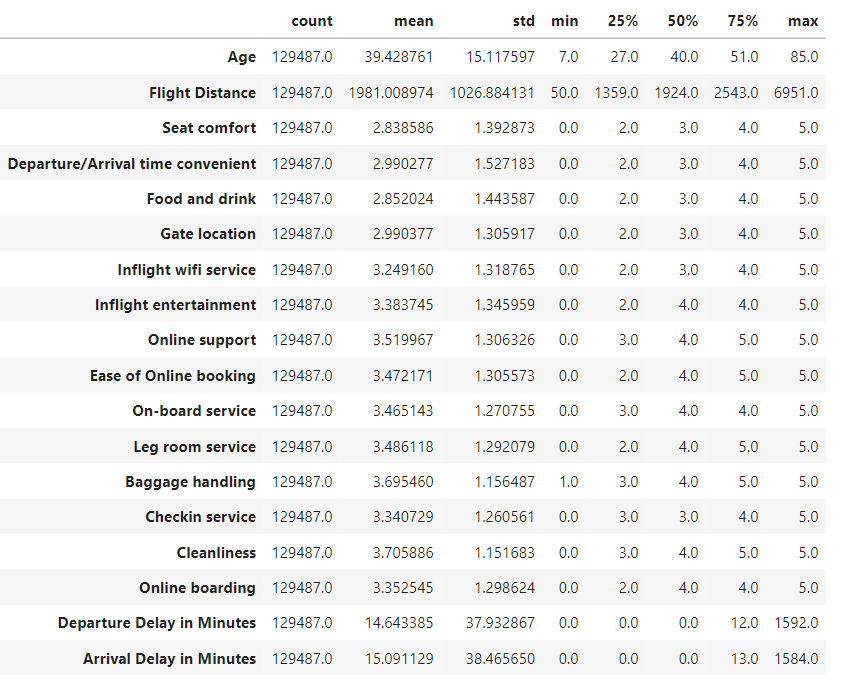
Ngoài ra, các hãng hàng không cần biết khía cạnh nào của dịch vụ mà họ cung cấp phải được chú trọng hơn để tạo ra nhiều khách hàng hài lòng hơn. [5]

Thông tin chung:

* Số cột: 23 cột
* Số dòng: 129880 dòng
* Các kiểu dữ liệu: float64(1), int64(17), object(5)
* Số lượng giá trị thiếu: Cột “Arrival Delay in Minutes” thiếu 393 dòng



Hình 3: Thông tin dữ liệu



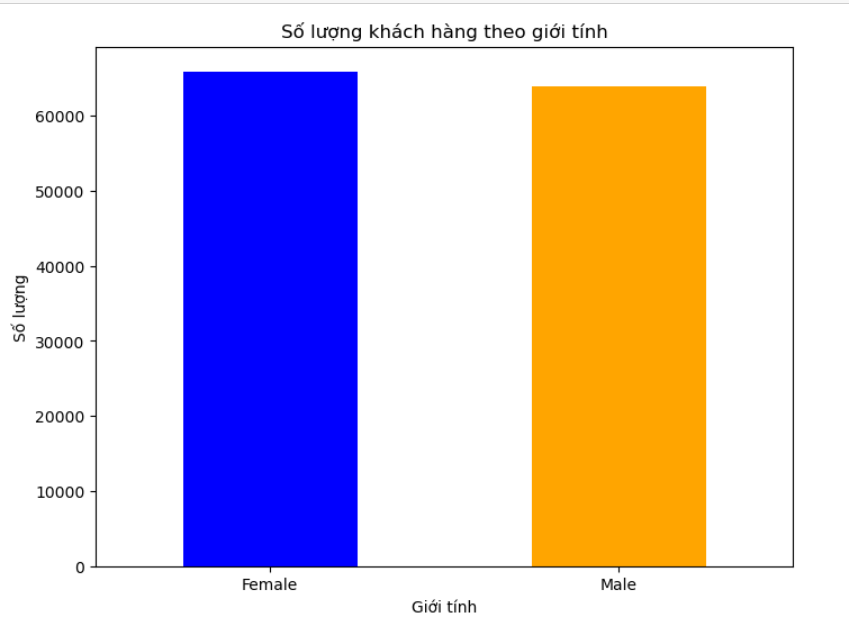
Hình 4: Thống kê mô tả dữ liệu

## Giải thích dữ liệu

Bảng 1: Bảng dữ liệu

| **Stt** | **Tên thuộc tính** | **Kiểu dữ liệu** | **Giải thích** |
| --- | --- | --- | --- |
| 1 | satisfaction | object | Sự hài lòng của khách hàng (satisfied hoặc dissatisfied) |
| 2 | Gender | object | Giới tính của khách hàng (Male hoặc Female) |
| 3 | Customer Type | object | Loại khách hàng (Loyal Customer hoặc Disloyal Customer) |
| 4 | Age | int64 | Độ tuổi của khách hàng |
| 5 | Type of Travel | object | Loại hình du lịch (Personal Travel hoặc Business Travel) |
| 6 | Class | object | Hạng vé đã mua (Eco, Eco Plus, hoặc Business) |
| 7 | Flight Distance | int64 | Khoảng cách của chuyến bay tính bằng dặm |
| 8 | Seat comfort | int64 | Đánh giá về độ thoải mái của ghế (thang số). |
| 9 | Departure/Arrival time convenient | int64 | Đánh giá về sự thuận tiện của thời gian đi và đến |
| 10 | Food and drink | int64 | Đánh giá chất lượng đồ ăn, đồ uống được phục vụ trên chuyến bay |
| 11 | Gate location | int64 | Đánh giá về sự thuận tiện của vị trí cổng |
| 12 | Inflight wifi service | int64 | Đánh giá chất lượng dịch vụ wifi trên máy bay |
| 13 | Inflight entertainment | int64 | Xếp hạng cho các lựa chọn giải trí trên chuyến bay |
| 14 | Online support | int64 | Đánh giá chất lượng hỗ trợ trực tuyến |
| 15 | Ease of Online booking | int64 | Đánh giá mức độ thuận tiện của quá trình đặt vé trực tuyến |
| 16 | On-board service | int64 | Đánh giá về chất lượng tổng thể của dịch vụ trên tàu |
| 17 | Leg room service | int64 | Đánh giá về không gian và sự thoải mái của chỗ để chân |
| 18 | Baggage handling | int64 | Đánh giá dịch vụ vận chuyển hành lý |
| 19 | Checkin service | int64 | Đánh giá cho quá trình phục vụ check-in |
| 20 | Cleanliness | int64 | Đánh giá độ sạch của máy bay |
| 21 | Online boarding | int64 | Đánh giá về sự thuận tiện của việc check in trực tuyến |
| 22 | Departure Delay in Minutes | int64 | Thời gian trì hoãn khởi hành tính bằng phút |
| 23 | Arrival Delay in Minutes | float64 | Thời gian trễ đến tính bằng phút |

## Trực quan hóa dữ liệu

****

Hình 5: Số lượng khách hàng theo giới tính

Nhận xét:

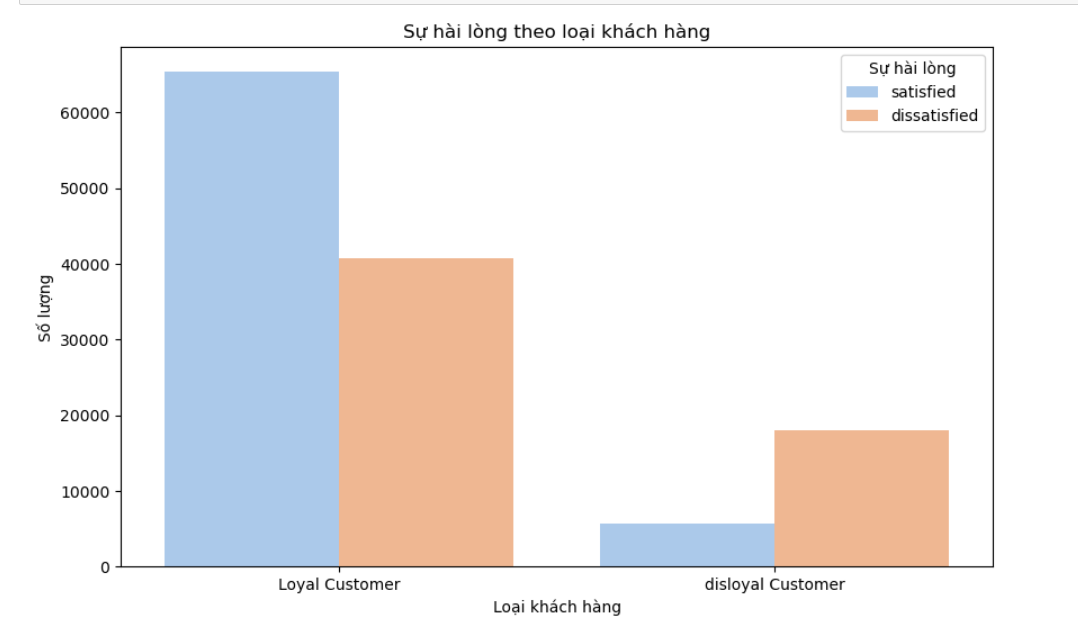
* Có hơn 65000 khách hàng là nữ, chiếm hơn 50% bộ dữ liệu.
* Số lượng khách hàng nam gần ngang bằng số lượng khách hàng nữ với hơn 60000 khách hàng và chiếm gần 50%.
* Phân phối khách hàng theo giới tính khá cân bằng nhưng nữ có lớn hơn một chút không đáng kể so với độ lớn bộ dữ liệu.



Hình 6: Sự hài lòng theo giới tính của khách hàng

Nhận xét:

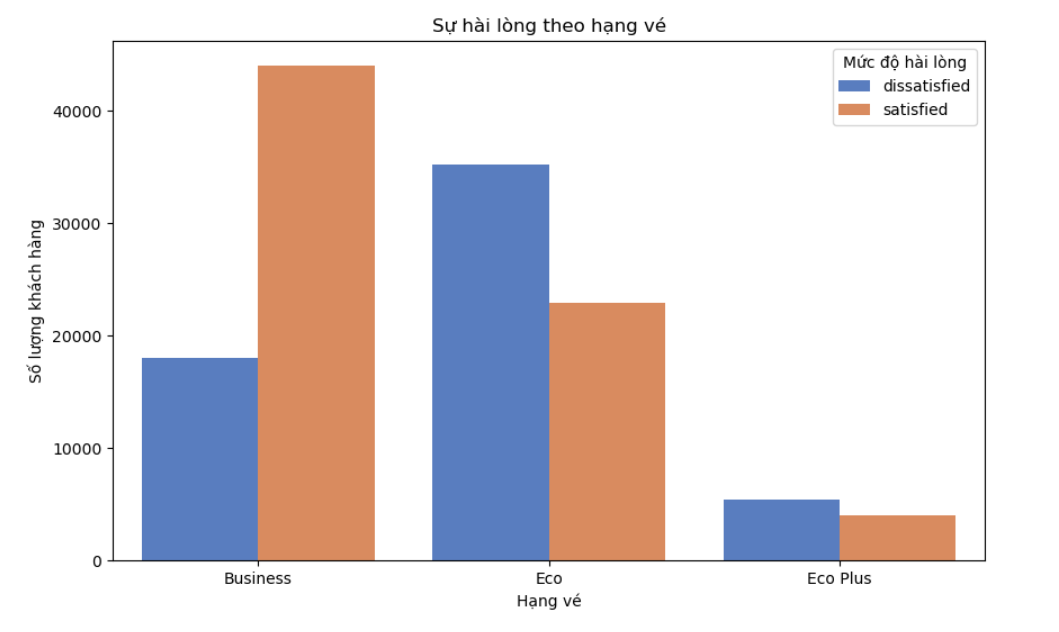
* Với khách hàng có giới tính nữ, số lượng khách hàng hài lòng (khoảng hơn 40000 khách hàng) cao hơn gần gấp đôi so với số lượng khách hàng không hài lòng  (khoảng hơn 20000 khách hàng).
* Với khách hàng có giới tính nam, số lượng khách hàng không hài lòng (khoảng hơn 30000 khách hàng) cao hơn số lượng khách hàng hài lòng (khoảng gần 30000 khách hàng).
* Nhóm khách hàng có giới tính nữ có tỷ lệ hài lòng cao hơn rõ rệt so với nhóm khách hàng có giới tính nam.
* Nhóm khách hàng có giới tính nam có xu hướng không hài lòng nhiều hơn so với nhóm khách hàng có giới tính nữ.



Hình 7: Sự hài lòng theo loại khách hàng

Nhận xét:

* Với nhóm Loyal customer (Khách hàng trung thành), số lượng hài lòng rất cao (Khoảng hơn 60000 khách hàng) và số lượng không hài lòng thấp hơn nhưng vẫn tương đối lớn (Khoảng hơn 40000 khách hàng)
* Với nhóm Disloyal customer (Khách hàng không trung thành), hầu hết khách hàng không hài lòng (Khoảng gần 20000 khách hàng) và số lượng hài lòng không đáng kể (Chưa đến 10000 khách hàng)
* Nhóm khách hàng trung thành có tỷ lệ hài lòng cao nhưng vẫn có một lượng lớn đáng kể khách hàng không hài lòng
* Nhóm khách hàng không trung thành chỉ có một số ít hài lòng còn lại phần lớn đều thấy không hài lòng, tỷ lệ không hài lòng khá cao



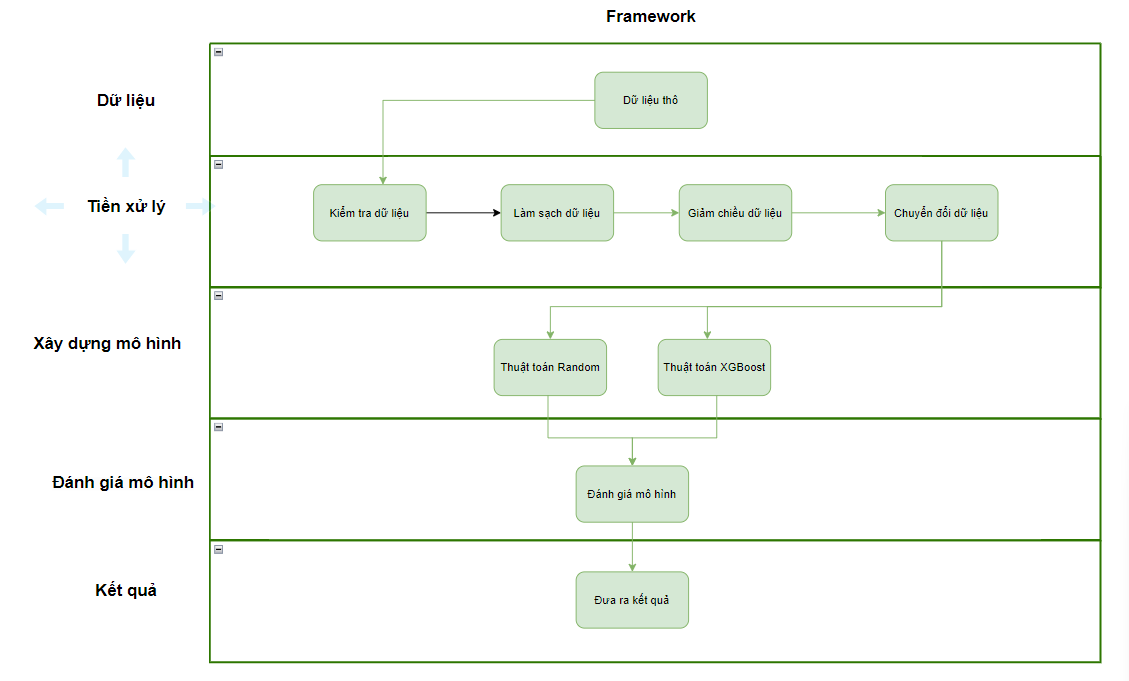
Hình 8: Sự hài lòng theo hạng vé

Nhận xét:

* Với hạng vé Business, số lượng khách hàng hài lòng khá cao (Khoảng hơn 40000 khách hàng) trong khi đó số lượng khách hàng không hài lòng chưa được một nửa số lượng hài lòng (Khoảng gần 20000 khách hàng).
* Với hạng vé Eco, số lượng khách hàng không hài lòng chiếm phần lớn (Khoảng gần 40000 khách hàng) và số lượng khách hàng hài lòng ít hơn (Khoảng hơn 20000 khách hàng).
* Với hạng vé Eco plus, số lượng khách hàng khá ít và khách hàng không hài lòng (Chưa tới 10000 khách hàng) cao hơn một chút so với số lượng khách hàng hài lòng.
* Hạng vé Business có tỷ lệ khách hàng hài lòng cao nhất trong khi đó  hạng vé Eco có số lượng khách hàng không hài lòng cao nhất, tỷ lệ sự hài lòng ở hạng vé Eco plus không có sự chênh lệch lớn.

1. **XÂY DỰNG MÔ HÌNH DỰ ĐOÁN SỰ HÀI LÒNG CỦA KHÁCH HÀNG**

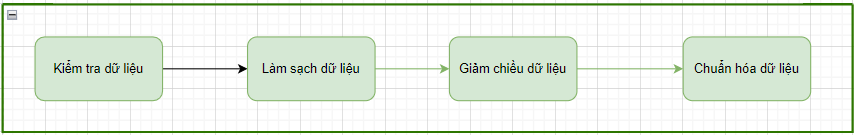
## Frame work



Hình 9: Framework cho dự án

## Tiền xử lý dữ liệu

Tiền xử lý dữ liệu (preprocessing data) là quá trình chuẩn bị và biến đổi dữ liệu từ dạng ban đầu thành dữ liệu có thể được sử dụng hiệu quả trong các mô hình học máy và phân tích dữ liệu. Quá trình này rất quan trọng và thường bao gồm các bước như sau:

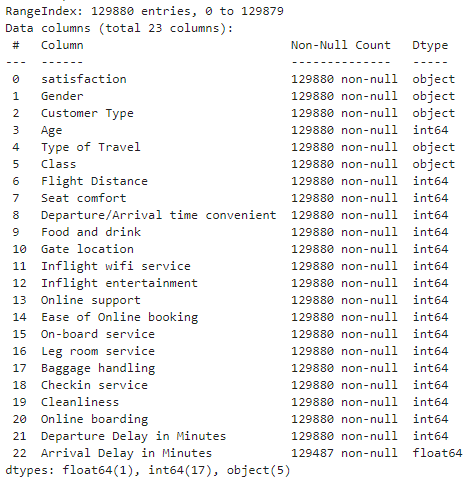


Hình 10: Tiền xử lý dữ liệu

* + 1. **Kiểm tra dữ liệu tổng quát**
* Kiểm tra dữ liệu



Hình 11: Đọc dữ liệu từ file CSV



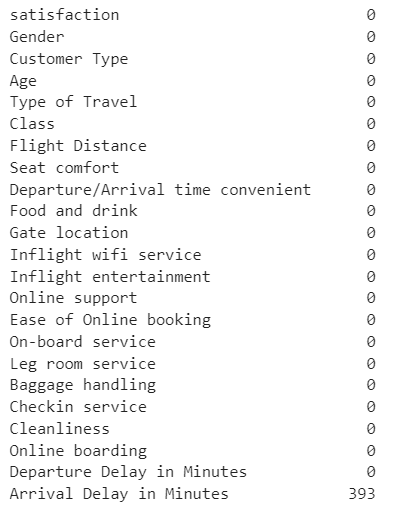
Hình 12: Thông tin kiểm tra của các cột trong dữ liệu

Sau khi đọc dữ liệu từ file csv chúng ta kiểm tra các thông tin cơ bản của bộ dữ liệu. Bộ dữ liệu gồm có 129880 dòng và 23 cột trong đó có 1 cột có kiểu dữ liệu là float64, 17 cột là int64, 5 cột là object.

* + 1. **Làm sạch dữ liệu**

Làm sạch dữ liệu là quá trình xử lý các giá trị bị thiếu, giá trị nhiễu, giá trị ngoại lai từ tập dữ liệu ban đầu nhằm tránh làm sai lệch kết quả và gây ra các quyết định sai lầm hoặc không thực tế.

* Kiểm tra dữ liệu

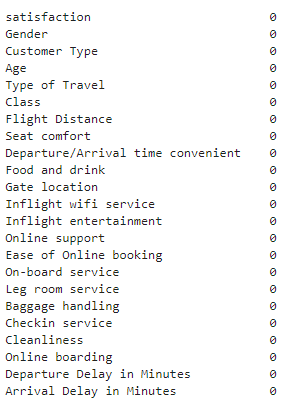
****

Hình 13: Kiểm tra dữ liệu có thiếu giá trị hay không

Cột Arrival Delay in Minutes thiếu 393 giá trị chiếm 0.3% so với bộ dữ liệu.

* **Xử lý giá trị bị thiếu**

Dữ liệu bị thiếu của cột Arrival Delay in Minutes chiếm 0.3 % so với bộ dữ liệu. Phần dữ liệu này khá nhỏ nên chúng ta có thể lựa chọn việc xử lý là xóa để không ảnh hưởng đến các thông tin khác.

****

Hình 14: Kết quả xử lý giá trị thiếu

Sau khi xóa dữ liệu thiếu, chúng ta kiểm tra lại bộ dữ liệu đã đủ.

* **Xử lý giá trị ngoại lai**

Outliers là một hoặc nhiều cá thể khác hẳn đối với các thành viên còn lại của nhóm. Sự khác biệt này có thể dựa trên nhiều tiêu chí khác nhau như giá trị hay thuộc tính. Giá trị outliers có thể gây ra ảnh hưởng đến kết quả dự đoán sai lệch, vậy nên chúng ta cần loại bỏ giá trị outliers ra khỏi bộ dữ liệu trước khi xây dựng mô hình.

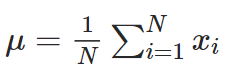
Trong machine learning có rất nhiều phương pháp để loại bỏ outliers nhưng ở đây chúng ta đề cập đến 2 phương pháp phổ biến nhất là:

* Sử dụng điểm Z để phát hiện giá trị ngoại lai liên quan đến việc đo khoảng cách giữa điểm dữ liệu và giá trị trung bình của một tập dữ liệu theo độ lệch chuẩn [6]. Sau đây là hướng dẫn từng bước về cách sử dụng điểm Z cho mục đích này:

**Các bước sử dụng Z-Scores để phát hiện giá trị ngoại lệ**

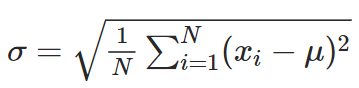
**Bước 1: Tính giá trị trung bình (μ)**:

* Tìm giá trị trung bình của tập dữ liệu:



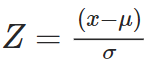
**Bước 2: Tính độ lệch chuẩn (σ)**:

* Đo độ phân tán của tập dữ liệu so với giá trị trung bình.



**Bước 3: Tính Điểm Z cho từng Điểm dữ liệu**:

* Điểm Z cho biết điểm dữ liệu cách giá trị trung bình bao nhiêu độ lệch chuẩn.



**Bước 4: Xác định ngưỡng cho các giá trị ngoại lệ**:

* Ngưỡng phổ biến để xác định các giá trị ngoại lệ là điểm Z lớn hơn +3 hoặc nhỏ hơn -3. Điều này có nghĩa là điểm dữ liệu cách giá trị trung bình hơn 3 độ lệch chuẩn.

**Bước 5: Xác định giá trị ngoại lệ**:

* Bất kỳ điểm dữ liệu nào có điểm Z vượt quá ngưỡng đều được coi là giá trị ngoại lệ.

Phương pháp loại bỏ outline bằng z-score áp dụng cho mô hình chúng ta đang xây dựng bằng python sẽ sử dụng 2 thư viện là **scipy.stats** để tính toán Z-score cho các giá trị số trong DataFrame và numpy để tính giá trị tuyệt đối của Z-score.

* Biểu đồ Boxplot là một công cụ hữu ích để trực quan hóa và phát hiện giá trị ngoại lai (outliers) trong dữ liệu. Boxplot cho thấy phạm vi của các giá trị, trung vị (median), và các giá trị ngoại lai của một biến số.

Các bước sử dụng boxplot để phát hiện ngoại lai:

**Bước 1:** Tính toán các giá trị Q1, Q2, Q3 lầ lượt là tứ phân vị thứ nhất, tứ phân vị thứ hai và tứ phân vị thứ 3.

**Bước 2:** Dùng công thức IQR = Q3 - Q1 để tính khoảng tứ phân vị (IQR). Sau đó, xác định giới hạn trên và dưới cho các giá trị ngoại lai theo công thức:

* Giới hạn dưới: Q1 - 1.5 \* IQR
* Giới hạn trên: Q3 + 1.5 \* IQR

**Bước 3:** Loại bỏ các giá trị nằm ngoài vùng giới hạn

Chúng ta thực hiện kiểm tra giá trị outliers bằng 2 phương pháp trên và cho ra được kết quả như sau:

* 43230 là kết quả outlier được tìm kiếm bằng phương pháp dùng biểu đồ boxplot
* 3715 là kết quả outlier được tìm kiếm bằng phương pháp Z-score

Qua kết quả ta thấy việc loại bỏ outliers bằng phương pháp biểu boxplot sẽ làm mất quá nhiều dữ liệu và outlier chiếm số lượng lớn thì chưa chắc đã là outlier, nên chúng ta sẽ sử dụng phương pháp Z-score để loại bỏ outlier mà không ảnh hưởng đến bộ dữ liệu.

Quá trình loại bỏ outline của các cột có dtype int64 và float64 bằng phương pháp Z-score với ngưỡng xác định Z = 3 đã cho ta kết quả:

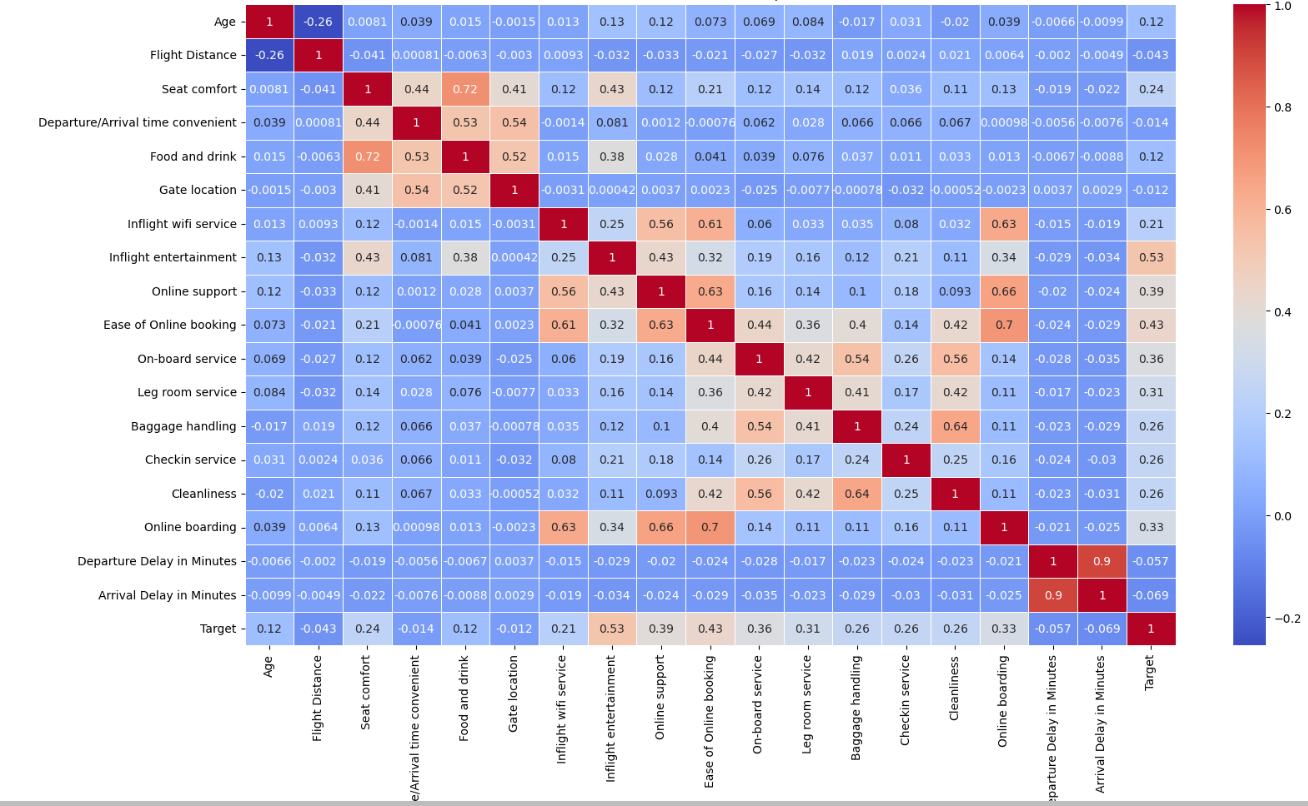
* Ban đầu bộ dữ liệu có 129487 dòng dữ liệu.
* Số lượng dữ liệu outline tìm thấy là 3715 dòng.
* Sau khi loại bỏ outline bộ dữ liệu còn lại 125772 dòng.
  + 1. **Giảm chiều dữ liệu**

Đôi khi dữ liệu có quá nhiều đặc trưng (features) không cần thiết, có thể dẫn đến hiện tượng overfitting và tăng chi phí tính toán. Vậy nên chúng ta cần giảm bớt các chiều không cần thiết và giữ lại các chiều quan trọng để xây dựng mô hình nhằm giảm bớt chi phí và thời gian xây dựng mô hình.

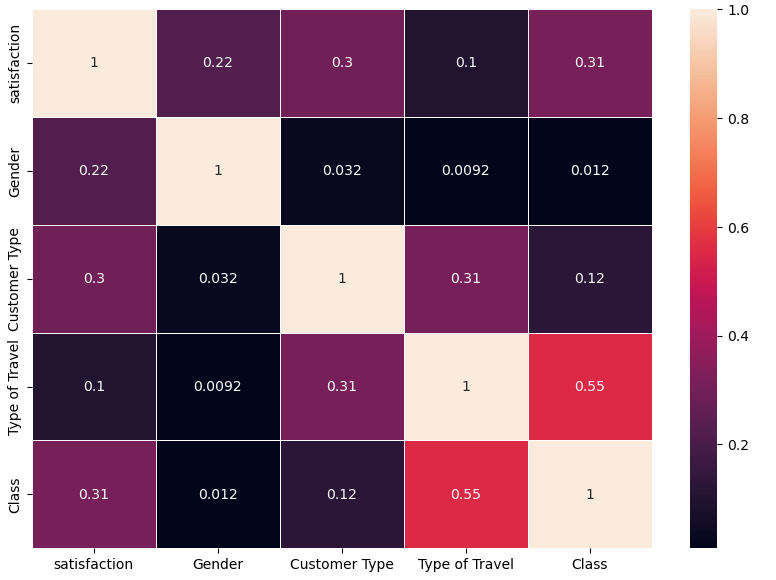
Sử dụng ma trận tương quan (correlation matrix) và lựa chọn ngưỡng (threshold) để loại bỏ các biến có mức độ tương quan thấp.

***Bước 1:*** Tạo num\_features chứa các features có kiểu dữ liệu là mumber, ob\_features sẽ lưu các cột có kiểu dữ liệu là object.

***Bước 2:*** Xem biểu đồ tương quan của num\_features và ob\_features



Hình 15: Biểu đồ tương quan của num\_features



Hình 16: Biểu đồ tương quan của ob\_features

Biểu đồ tương quan:

* Giá trị dương biểu thị mỗi tương quan thuận.
* Giá trị âm biểu thị mỗi tương quan nghịch.
* Giá trị gần bằng 0 cho thấy các feature không có mỗi tương quan với nhau

***Bước 3:*** Lựa chọn num\_features và ob\_feature quan trọng:

Để xác định được các features quan trọng mình cần chọn được ngưỡng xác định phù hợp với bộ dữ liệu. Nếu chọn ngưỡng xác định quá thấp, bộ dữ liệu vẫn còn lại rất nhiều feature và những feature không quan trọng có thể chưa được loại bỏ. Nếu chọn ngưỡng xác định quá cao, bộ dữ liệu sẽ có thể mất đi nhiều feature quan trong và còn quá ít feature để xây dựng mô hình.

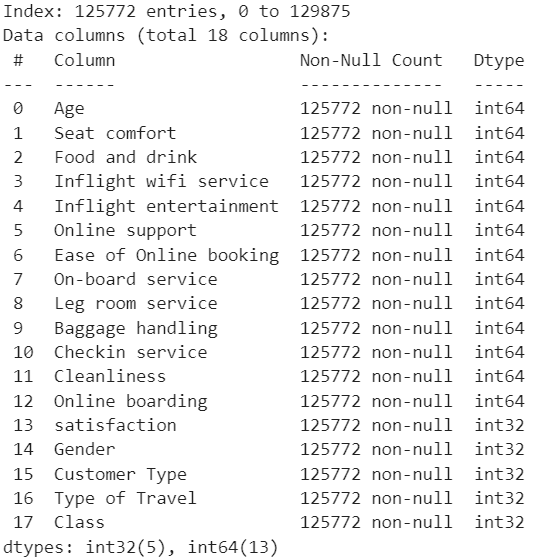
Với sự so sánh các hiệu suất của mô hình ở các ngưỡng xác định khác nhau cho ra hiểu suất mô hình khác nhau. Trong đó ngưỡng xác định = 0.1 cho ra hiệu suất sau khi xây dựng mô hình là cao nhất.

Thực hiện lựa chọn ngưỡng xác định = 0.1 lúc này bộ dữ liệu Invistico\_Airline sẽ còn lại 18 cột sau: 'Age', 'Seat comfort', 'Food and drink', 'Inflight wifi service', 'Inflight entertainment', 'Online support', 'Ease of Online booking', 'On-board service', 'Leg room service', 'Baggage handling', 'Checkin service', 'Cleanliness', 'Online boarding', 'satisfaction', 'Gender', 'Customer Type', 'Type of Travel', 'Class'.

* + 1. **Chuyển đổi dữ liệu**

Các mô hình học máy được xây dựng bằng python yêu cầu dữ liệu đầu vào thuộc kiểu numberic trong khi bộ dữ liệu của chúng ta còn có 1 số cột như Gender, Customer Type, Type of Travel, Class, satisfaction ở dạng object chưa phù hợp với đầu vào của mô hình. Vậy nên chúng ta cần đến bước chuyển đổi dữ liệu để chuyển dữ liệu của các cột biến phân loại về biến số.

Bằng cách sử dụng *LabelEncoder* từ thư viện *sklearn.preprocessing*. Các cột có kiểu dữ liệu là object trong Invistico\_Airline lúc này sẽ được chuyển về mã hóa int:



Hình 17: Kiểu dữ liệu đầu vào

## Mô hình Random Forest

* + 1. **Xây dựng mô hình**
* **Chia dữ liệu**

Mô hình được chia thành 2 phần là X\_train và  y\_train  chiếm 80% bộ dữ liệu dùng để training mô hình, X\_test và y\_test chiếm 20% bộ dữ liệu dùng để test mô hình và đánh giá hiệu quả mô hình.

* **Sử dụng GridSearchCV tìm ra bộ siêu tham số tối ưu cho mô hình Randomforest**

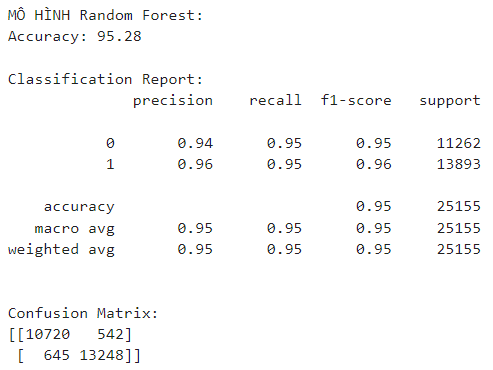
Các siêu tham số và giá trị sử dụng trong GridSearchCV bao gồm:

* **n\_estimators: [100, 200, 300]**
  + Đây là số lượng cây trong rừng ngẫu nhiên của mô hình Random Forest. Mô hình sẽ sử dụng các giá trị cây phổ biến và chọn ra số lượng tốt nhất.
* **max\_depth: [5, 10, 15]**
  + Độ sâu tối đa của mỗi cây quyết định độ phức tạp của mô hình. Độ sâu quá lớn có thể dẫn đến hiện tượng overfitting và cây phức tạp, trong khi độ sâu nhỏ có thể gây underfitting và cây quá đơn giản.
* **min\_samples\_split: [2, 5, 10]**
  + Đây là số lượng mẫu tối thiểu để phân chia một node. Giá trị nhỏ có thể tạo ra nhiều node nhỏ, làm tăng độ phức tạp, trong khi giá trị lớn giúp cây ngừng phân chia sớm, giảm độ phức tạp.
* **min\_samples\_leaf : [1, 2, 4]**
* Xác định số lượng mẫu tối thiểu mà một nút lá phải chứa để nó được coi là một nút lá hợp lệ.

Kết quả GridSearchCV cho thấy tổ hợp siêu tham số tối ưu là:

* **n\_estimators: 200**
* **max\_depth: 15**
* **min\_samples\_split: 2**
* **min\_samples\_leaf : 1**
  + 1. **Đánh giá mô hình**
* **Đánh giá các chỉ số**

Dựa vào kết quả dự báo nhãn của bộ dữ liệu test (y\_prep) và nhãn của bộ dữ liệu test ( y\_test) chúng ta sẽ đáng giá mô hình qua các chỉ số sau:

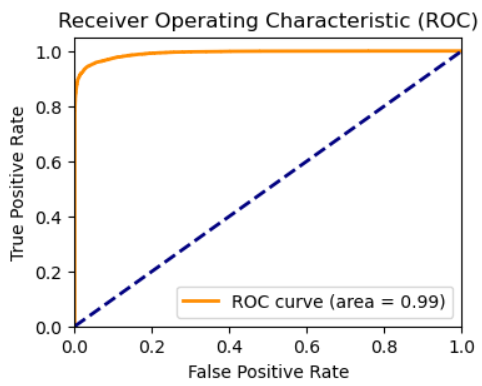


Hình 18: Kết quả đánh giá của mô hình Random Forest

Sau khi tìm ra siêu tham số tối ưu, mô hình Random Forest được huấn luyện với tổ hợp tham số trên và được áp dụng dự đoán trên tập kiểm tra. Độ chính xác đạt **95.28%**, thể hiện tính hiệu quả cao của mô hình trong việc dự đoán sự hài lòng của khách hàng.

Ma trận nhầm lẫn có thể được giải thích như sau:

* 10720 mẫu là số lượng khách hàng được dự đoán và thực tế đều là "dissatisfied".
* 542 mẫu là số lượng khách hàng được dự đoán là "dissatisfied" nhưng thực tế là "satisfied".
* 645 mẫu là số lượng khách hàng được dự đoán là "satisfied" nhưng thực tế là "dissatisfied".
* 13248 mẫu là số lượng khách hàng được dự đoán và thực tế đều là "satisfied".
* **Đánh giá đồ thị AUC-ROC**



Hình 19: Biểu đồ Roc của mô hình Random Forest

Với AUC = 0.99 mô hình Randomforest có khả năng phân biệt tốt giữa hai lớp hài lòng và không hài lòng. Cụ thể, mô hình có khoảng 99% cơ hội phân loại đúng các trường hợp khách hàng hài lòng hoặc không hài lòng.

## Mô hình XGBoost

* + 1. **Xây dựng mô hình**
* **Chia dữ liệu**

Mô hình được chia thành 2 phần là X\_train và  y\_train  chiếm 80% bộ dữ liệu dùng để training mô hình, X\_test và y\_test chiếm 20% bộ dữ liệu dùng để test mô hình và đánh giá hiệu quả mô hình.

* **Sử dụng GridSearchCV tìm ra bộ siêu tham số tối ưu cho mô hình Xgboost**

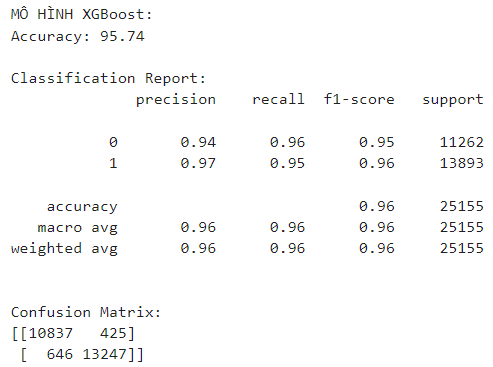
Các siêu tham số và giá trị sử dụng trong GridSearchCV bao gồm:

* **n\_estimators: [100, 200, 300]**
  + Đây là số lượng cây của mô hình XGboost. Mô hình sẽ sử dụng các giá trị cây phổ biến và chọn ra số lượng tốt nhất.
* **learning\_rate: [0.01, 0.1, 0.2]**
* Đây là chỉ số tốc độ học khi mỗi cây mới được tạo. Nếu tốc độ quá nhỏ sẽ phải tạo rất nhiều cây mới cạy thiện tốt mô hình nhưng nếu tốc đồ cây quá lớn sẽ dễ dẫn đến overfitting.
* **max\_depth: [5, 10, 15]**
  + Độ sâu tối đa của mỗi cây quyết định độ phức tạp của mô hình. Độ sâu quá lớn có thể dẫn đến hiện tượng overfitting và cây phức tạp, trong khi độ sâu nhỏ có thể gây underfitting và cây quá đơn giản.
* **subsample: [ 0.8, 0.9, 1.0 ]**
  + Đây là số lượng mẫu được sử dụng để huấn luyện mô hình.
* **colsample\_bytree: [ 0.8, 0.9, 1.0]**
  + Đây là số lượng feature sẽ được chọn để xây cây, nhằm tránh gây ra hiện tượng overfitting. Ví dụ 0.8 thì sẽ có 80% cột được chọn từ bộ dữ liệu đầu vào để xây cây.

Kết quả GridSearchCV cho thấy tổ hợp siêu tham số tối ưu là:

* **n\_estimators: 200**
* **learning\_rate: 0.1**
* **max\_depth: 9**
* **subsample: 1.0**
* **colsample\_bytree: 0.8**
  + 1. **Đánh giá mô hình**
* **Đánh giá các chỉ số**

Dựa vào kết quả dự báo nhãn của bộ dữ liệu test (y\_prep) và nhãn của bộ dữ liệu test ( y\_test) chúng ta sẽ đáng giá mô hình qua các chỉ số sau:

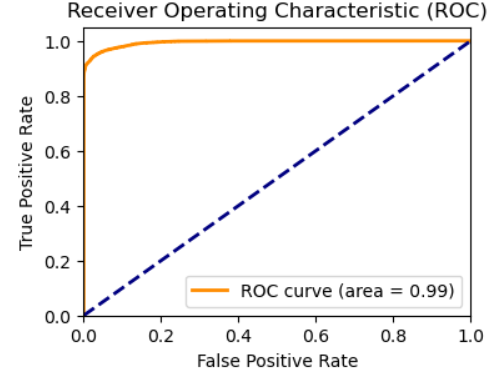


Hình 20: Kết quả đánh giá của mô hình XGboost

Độ chính xác (Accuracy) là tỷ lệ các dự đoán đúng so với tổng số dự đoán. Trong mô hình, độ chính xác là 95.74 %, nghĩa là mô hình đã dự đoán đúng 95.74 % nhãn của toàn bộ khách hàng. Đây là một tỷ lệ khá cao, cho thấy mô hình có khả năng dự đoán chính xác.

Ma trận nhầm lẫn có thể được giải thích như sau:

* + 10837 mẫu là số lượng khách hàng được dự đoán và thực tế đều là "dissatisfied".
  + 425 mẫu là số lượng khách hàng được dự đoán là "dissatisfied" nhưng thực tế là "satisfied".
  + 646 mẫu là số lượng khách hàng được dự đoán là "satisfied" nhưng thực tế là "dissatisfied".
  + 13247 mẫu là số lượng khách hàng được dự đoán và thực tế đều là "satisfied".
* **Đánh giá đồ thị ROC**



Hình 21: Biểu đồ Roc của mô hình XGboost

Với AUC = 0.99 mô hình Xgboost có khả năng phân biệt tốt giữa hai lớp hài lòng và không hài lòng. Cụ thể, mô hình có khoảng 99% cơ hội phân loại đúng các trường hợp khách hàng hài lòng hoặc không hài lòng.

## Đánh giá các mô hình

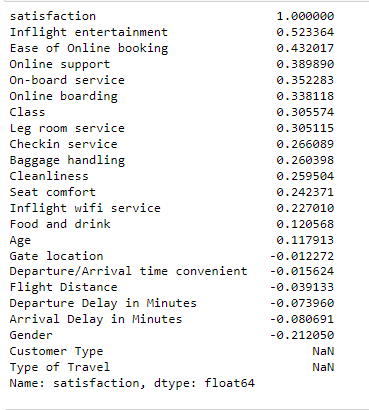
| Mô hình | Random Forest | XGBoost |
| --- | --- | --- |
| Accuracy  (TP + TN) / (TP + TN + FP + FN) | 95.28% | 95.74% |
| Precision  TP / (TP + FP) | 96% | 97% |
| Recall  TP / (TP + FN) | 95% | 95% |
| F1-score  2 \* (Precision \* Recall) / (Precision + Recall) | 96% | 96% |
| AUC | 0.99 | 0.99 |

Bảng 2: Bảng so sánh các chỉ sổ các các mô hình

Dựa trên các chỉ số đánh giá trên, Cả hai mô hình đều rất mạnh và đáng tin cậy, có khả năng cung cấp dự báo chính xác cao trong bài toán này. Nhận thấy mô hình XGBoost là mô hình có hiểu suất tổng thể nhỉnh hơn 1 chút so với mô hình Randomforest.

1. **CHIẾN LƯỢC DỰ ÁN**

Qua vấn đề của dự án, có thể thấy sự hài lòng của khách hàng rất quan trọng và từ đó việc hiểu rõ các yếu tố ảnh hưởng đến sự hài lòng của khách hàng càng quan trọng hơn. Thông qua việc đánh giá mức độ tương quan giữa các biến trong bộ dữ liệu chúng ta có thể đưa ra được những yếu tố quan trọng ảnh hưởng lớn đến sự hài lòng của khách hàng.



Hình 22: Mối tương quan giữa các biến đầu vào và biến mục tiêu 'satisfaction'

Tổng quan:

* Biến có mối tương quan mạnh nhất với biến mục tiêu là “Inflight entertainment” (0.523364), điều đó cho thấy các lựa chọn giải trí trên chuyến bay có ảnh hưởng rất lớn đối với sự hài lòng của khách hàng.
* Ngoài ra còn có một số biến có mối tương quan khá mạnh với biến mục tiêu là “Ease of Online booking” (0.432017), “Online support” (0.388298), “On-board service” (0.352283) và “Online boarding” (0.338318). Điều này cho thấy việc đặt vé trực tuyến, hỗ trợ trực tuyến, dịch vụ tổng thể và việc check in trực tuyến cũng có ảnh hưởng không nhỏ đến sự hài lòng của khách hàng.
* Các biến còn lại đều có mức độ tương quan thấp hoặc không có mối quan hệ rõ ràng và điều đó cho thấy các yếu tố ấy có ảnh hưởng không lớn đến sự hài lòng của khách hàng.

Từ việc trực quan hóa dữ liệu, phân tích mối tương quan giữa các biến với sự hài lòng của khách hàng và mô hình dự đoán chúng ta có thể đưa ra những giải pháp nhằm tăng tỷ lệ hài lòng của khách hàng cũng như tối ưu hóa hiệu quả của hãng hàng không như sau:

* Cải thiện chất lượng giải trí trên chuyến bay: Với việc lựa chọn giải trí trên chuyến bay có ảnh hưởng lớn nhất đến sự hài lòng của khách hàng thì việc nâng cấp cũng như cung cấp thêm nhiều dịch vụ giải trí đa dạng và tiện lợi cho khách hàng như phim, nhạc, trò chơi … là một vấn đề cần lưu tâm.
* Tăng cường trải nghiệm đặt vé online: Việc đặt vé trực tuyến có ảnh hưởng không nhỏ tới sự hài lòng của khách hàng, qua đó có thể cải thiện tính linh hoạt cũng như giao diện thân thiện dễ dùng để việc đặt vé của khách hàng trở nên dễ dàng hơn.
* Cải thiện chất lượng hỗ trợ trực tuyến: Đào tạo nhân viên chuyên nghiệp để kịp thời hỗ trợ  khách hàng.
* Nâng cấp các dịch vụ khác trên chuyến bay cũng như tối ưu hóa quy trình check in trực tuyến: Cải thiện các dịch vụ như đồ ăn thức uống và đơn giản hóa quy trình check in trực tuyến từ đó giúp khach hàng tiết kiệm thời gian và hài lòng hơn.

# TÀI LIỆU THAM KHẢO

| [1] | [Trực tuyến]. Available: https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-. |
| --- | --- |
| [2] | [Trực tuyến]. Available: https://vtitech.vn/xgboost-gioi-thieu-chuoi-bai-viet-ve-thuat-toan-xgboost/. |
| [3] | [Trực tuyến]. Available: https://www.mygreatlearning.com/blog/gridsearchcv/#what-is-gridsearchcv. |
| [4] | [Trực tuyến]. Available: https://tapit.vn/cac-phuong-phap-danh-gia-mot-mo-hinh-phan-loai/. |
| [5] | [Trực tuyến]. Available: https://www.kaggle.com/datasets/sjleshrac/airlines-customer-satisfaction. |
| [6] | [Trực tuyến]. Available: https://www.quora.com/How-did-you-use-a-Z-score-for-outlier-detection. |

# PHỤ LỤC