TRƯỜNG ĐẠI HỌC MỞ HÀ NỘI

**KHOA CÔNG NGHỆ THÔNG TIN**

----------------------



**BÁO CÁO BÀI TẬP LỚN**

**MÔN: NHẬP MÔN KHAI PHÁ DỮ LIỆU VÀ MÁY HỌC**

**ĐỀ TÀI: Ứng dụng khai phá dữ liệu và máy học tích hợp vào hệ thống Đặt Tour để giải quyết bài toán dự đoán tỷ lệ khách hàng ngưng sử dụng dịch vụ**

***Giảng viên hướng dẫn*** **: *TS Lê Hữu Dũng***

***Sinh viên thực hiện* :** **Nhóm G04**

Nguyễn Gia Đạt - 2110A04

Hán Đức Giang - 2110A03

Nguyễn Tiến Khởi - 2110A01

***Hà Nội, năm 2024***

# Yêu cầu với BTL NHẬP MÔN KHAI PHÁ DỮ LIỆU VÀ MÁY HỌC

1. Tên đề tài có dạng  
   “**Ứng dụng khai phá dữ liệu và máy học tích hợp vào hệ thống <tên hệ thống> [giải quyết bài toán <tên bài toán>]**”
2. Sử dụng dữ liệu có nguồn gốc rõ ràng, tin cậy, đủ lớn và có liên quan đến bài toán định xử lý; được cung cấp sẵn trên các cộng đồng DMML hoặc SV tự thu thập.
3. Áp dụng Khai phá luật kết hợp (1) hoặc Phân lớp (2) hoặc Phân cụm (3) để xây dựng giải pháp ứng dụng cho bài toán cụ thể
4. Mô hình được lựa chọn sau huấn luyện (/khai phá) phải căn cứ trên kết quả đánh giá tin cậy, chính xác và có căn cứ khoa học.
5. Mô hình cần được tích hợp vào hệ thống cụ thể, ở chức năng liên quan tiêu biểu để chứng minh tính ứng dụng - giải quyết bài toán đã xác định (các chức năng khác không phải là mục tiêu của học phần này)
6. Đáp ứng đúng mục tiêu theo đề cương của học phần NHẬP MÔN KHAI PHÁ DỮ LIỆU VÀ MÁY HỌC
7. Từng SV cần thực hiện nhiệm vụ để chứng minh đạt được Chuẩn đầu ra của học phần.

## 

***Báo cáo BTL học phần này cần trình bày theo cấu trúc dưới đây (SV soạn thảo***

***trực tiếp vào file này), tuân thủ Quy định trình bày của khoa***

# 

[**Chương 1. Tổng quan về đề tài 3**](#_rzwy357qd4fs)

[1.1. Giới thiệu bài toán 3](#_hlekjk5eda4v)

[1.2. Đánh giá đề tài liên quan   
(vấn đề hiện đã được giải quyết như thế nào, hiệu quả đến đâu) 4](#_of77iq1bhvus)

[1.2.1 Đánh giá đề tài liên quan : 4](#_lnrle1tyxhe4)

[1.2.2 Mô tả hiện trạng của bài toán : 5](#_cwhzxgp3umsu)

[1.3. Mục đích đề tài: 7](#_lnafygkw5qt1)

[1.4. Mục tiêu đề tài: 7](#_oix68n853nmo)

[1.5. Phân công và Kế hoạch thực hiện 8](#_5jy67wkpvucs)

[**Chương 2. Tiền xử lý dữ liệu 8**](#_z5414x6psqma)

[2.1. Phân tích dữ liệu: 9](#_a34xvtpzc2lu)

[2.2. Tiến hành các bước tiền xử lý dữ liệu: 12](#_f4eujtrpbkcz)

[2.3. Phân tích dữ liệu sau khi làm sạch dữ liệu: 15](#_1rlhnklfw9y)

[**Chương 3. Xây dựng mô hình 15**](#_4a99ryaeqcid)

[3.1. Hướng giải quyết: 16](#_7wxsfyn9noau)

[3.2 Lựa chọn thuật toán : Decision Tree 17](#_e6rtemabmv1w)

[● Cơ sở lý thuyết: 17](#_3mhu5j21mkt4)

[● Xây dựng thuật toán 19](#_mn86qho7716g)

[3.2.1. Khai báo thư viện (Import library) : 19](#_qb0ux548zi7y)

[3.2.2. Load dữ liệu (Data set loading) : 20](#_tkgkf9lxxgjm)

[3.2.3. Phân tích giải thích dữ liệu (Explanatory data analysis) : 20](#_42adqjh5hton)

[3.2.4.Tiền xử lý dữ liệu (Data preprocessing) : 26](#_xsrek3nz3cti)

[3.2.5. Modeling : 27](#_c7v0cjrq4k9q)

[3.2.6. Model tuning (Random Forest) : 29](#_ni5ncg4vag4q)

[3.2.7. Kết quả mô hình tốt nhất (RandomForestClassifier) : 30](#_bt3sjvtvg0yq)

[3.2.8. Phân tích tính năng : 31](#_ss4zjmuc78as)

[3.3. Random forest: 34](#_tyda8vcqk95l)

[3.4. Đánh giá mô hình 37](#_ntbnfy7mgplb)

[Đánh giá mô hình random forest 38](#_3t0eaqkg185b)

[So sánh: 39](#_x222h4qwksjy)

[**Chương 4. Tích hợp mô hình vào hệ thống 40**](#_xdt9bbjf0vpg)

[Kết luận 42](#_t9xh8zh28ff)

[Tài liệu tham khảo: 42](#_yi52o0fd4kpj)

# Chương 1. Tổng quan về đề tài

## 1.1. Giới thiệu bài toán

Bài toán Dự đoán tỷ lệ khách hàng rời bỏ trong ngành Tour & Du lịch là một trong những thách thức quan trọng đối với các doanh nghiệp trong lĩnh vực này. Việc hiểu được nguyên nhân và dự đoán khách hàng sẽ rời bỏ trước khi họ thực sự làm điều đó có thể giúp công ty du lịch thực hiện các biện pháp phòng ngừa và tối ưu hóa chiến lược giữ chân khách hàng. Khi tích hợp AI để dự đoán khách hàng ngưng sử dụng dịch vụ sẽ mang đến cho công ty dự đoán về khách hàng để có các điều chỉnh chiến lược kinh doanh giữ chân khách hàng. Đối với khách hàng, việc dự đoán này gián tiếp giúp khách hàng có trải nghiệm dịch vụ tốt hơn vì công ty phải nâng cao dịch vụ để giữ chân khách hàng

Các yếu tố có thể ảnh hưởng đến quyết định của khách hàng rời bỏ một dịch vụ du lịch có thể bao gồm:

* Chất lượng dịch vụ: Sự hài lòng của khách hàng với chất lượng dịch vụ cung cấp, bao gồm cả về chất lượng của tour, dịch vụ khách sạn, nhà hàng và các hoạt động giải trí.
* Giá cả: Sự hợp lý về giá cả so với chất lượng dịch vụ.
* Trải nghiệm khách hàng: Sự thoải mái và tiện nghi trong quá trình du lịch, bao gồm cả trải nghiệm mua sắm, đi lại và giao tiếp.
* Tính cá nhân hóa: Khả năng cung cấp các gói dịch vụ phù hợp với nhu cầu và sở thích cá nhân của từng khách hàng.
* Chính sách hỗ trợ và hoàn tiền: Sự linh hoạt và thái độ hỗ trợ từ phía doanh nghiệp trong trường hợp khách hàng gặp vấn đề hoặc không hài lòng.
* Cạnh tranh: Sự cạnh tranh từ các doanh nghiệp khác trong ngành du lịch.

## 1.2. Đánh giá đề tài liên quan

### 1.2.1 Đánh giá đề tài liên quan :

* **Đề tài: - TÌM HIỂU THUẬT TOÁN CÂY QUYẾT ĐỊNH VÀ ỨNG DỤNG VÀO BÀI TOÁN DỰ ĐOÁN THU NHẬP**
* Nguồn:<https://www.studocu.com/vn/document/truong-dai-hoc-cong-nghiep-ha-noi/khai-pha-du-lieu/bao-cao-ap-dung-cay-quyet-dinh-vao-du-doan-thu-nhap/43682722?origin=home-recent-1>
* Đánh giá: Đề tài đã cung cấp đầy đủ thông tin về khái niệm, cấu trúc và các ứng dụng của thuật toán Decision Tree. Đưa ra được các ví dụ thực tế áp dụng thuật toán Decision Tree và chỉ ra được ưu, nhược điểm của thuật toán.
* **Đề tài: - ỨNG DỤNG KHAI PHÁ DỮ LIỆU DỰ ĐOÁN KHÁCH HÀNG RỜI MẠNG VIỄN THÔNG**
* Nguồn:<https://luanvan.net.vn/luan-van/luan-van-ung-dung-khai-pha-du-lieu-du-doan-khach-hang-roi-mang-vien-thong-59949/>
* Đánh giá: Báo cáo đưa ra được đầy đủ lý do và mục tiêu của đề tài, cung cấp đầy đủ thông tin về các kỹ thuật dự báo dựa vào khai phá dữ liệu.

### 1.2.2 Mô tả hiện trạng của bài toán :

Để giải quyết vấn đề dự đoán tỷ lệ khách hàng rời bỏ trong ngành Tour & Du lịch, có một số bước và chiến lược cụ thể đang được thực hiện trong bài toán :

* **Thu thập dữ liệu:**
* Tìm kiếm tài liệu từ các nguồn khác nhau, để phục vụ mục đích nghiên cứu, học tập nên chọn chủ yếu các dataset từ kaggle.com. Nguồn dataset: <https://www.kaggle.com/datasets/tejashvi14/tour-travels-customer-churn-prediction>
* **Tiền xử lý dữ liệu:**
* Để đạt được hiệu quả tốt nhất và kết quả chính xác nhất nên cần phải xử lý lại các dữ liệu từ data thu thập được. Giai đoạn này cần phải loại bỏ các dữ liệu trùng lặp, thiếu và gây nhiễu do những loại dữ liệu này làm tăng thời gian xử lý và làm giảm độ tin cậy, tính chính xác khi nghiên cứu. Khi nghiên cứu cũng cần thêm chứng minh để có thể loại bỏ dữ liệu đó 1 cách chính xác.
* **Phân tích và gắn nhãn đặc trưng:**
* Kiểm tra tính chính xác của các thuộc tính trong dataset, xác định các thuộc tính(cột) liệu có cần thiết cho bài toán không. Khi không cần thiết cần loại bỏ đồng thời chứng minh được tại sao loại bỏ.
* Gắn nhãn cho các cột thuộc tính vì các tên cột không rõ ràng cho người khác khi đọc qua về đề tài.
* **Xây dựng mô hình trực quan hóa dữ liệu:**
* Để xác định rõ hơn về sự phân bố dữ liệu nên cần trực quan hóa dữ liệu thành các đồ thị trong không gian 2 3 chiều, biểu đồ. Hiện tại khi thu thập dữ liệu đã có sẵn các biểu đồ cột biểu thị số lượng trong từng đặc trưng tuy nhiên cần làm lại sau khi làm sạch dữ liệu.
* Giải pháp: Sử dụng công cụ Weka để hỗ trợ biểu diễn dữ liệu trực quan.
* **Xây dựng mô hình:**
* Dùng nhiều loại mô hình để thực hiện phân lớp dữ liệu như Ma trận nhầm lẫn, Decision Tree, VMS,... .Để đạt hiệu quả cao và chính xác nên sẽ thực hiện ít nhất 2 kiểu phân lớp dữ liệu khác nhau, sau mỗi lần thực hiện phân lớp dữ liệu sẽ thực hiện đánh giá, so sánh 2 mô hình với nhau, thực hiện việc này nhiều lần(tối thiểu 3 lần) để đạt được độ chính xác cao.
* Mỗi lần thực hiện lại cần điều chỉnh lại các tham số để tối ưu hóa mô hình.
* **Tích hợp và đánh giá mô hình:**
* Để đánh giá được kết quả nghiên cứu cần áp dụng vào thực tế. Xây dựng một chương trình mới có giao diện để sử dụng được kết quả nghiên cứu.
* Các công cụ để xây dựng chương trình phù hợp với kết quả nghiên cứu: + Weka + Java
* ML.net
* HTML, JS
* Java (Androids Studio)

## 1.3. Mục đích đề tài:

Mục đích của dự án là xác định và dự đoán tỷ lệ khách hàng sẽ ngừng sử dụng dịch vụ của doanh nghiệp trong tương lai. Bằng cách thu thập và phân tích dữ liệu về hành vi của khách hàng, dự án nhằm xây dựng các mô hình dự đoán và đưa ra dự báo về việc khách hàng có thể chuyển sang sử dụng dịch vụ của đối thủ hoặc dừng sử dụng dịch vụ hiện tại.

## 1.4. Mục tiêu đề tài:

* MT1 : Phân tích dữ liệu về hành vi khách hàng để xác định các yếu tố ảnh hưởng đến khả năng ngưng sử dụng dịch vụ.
* MT2 : Xác định thuật toán và mô hình huấn luyện phù hợp.
* MT3 : Xây dựng mô hình dự đoán và đưa ra dự báo có tính chính xác cao.
* MT4 : Tích hợp mô hình vào hệ thống.

## 1.5. Phân công và Kế hoạch thực hiện

**P :** chính, **S** : Phụ.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Nội dung** | **Nguyễn Gia Đạt** | **Hán Đức Giang** | **Nguyễn Tiến Khởi** |
| **1** | **Tổng quan đề tài** | **P** | **P** | **P** |
| **2** | **Tiền xử lý dữ liệu** | **P** | **S** | **S** |
| **3** | **Xây dựng mô hình** | **S** | **S** | **P** |
| **4** | **Tích hợp mô hình** | **S** | **P** | **S** |
| **5** |  |  |  |  |

# 

# 

# Chương 2. Tiền xử lý dữ liệu

Thực hiện phân tích về dữ liệu, tiến hành các bước tiền xử lý dữ liệu, chuẩn bị dữ liệu cho giai đoạn sau.

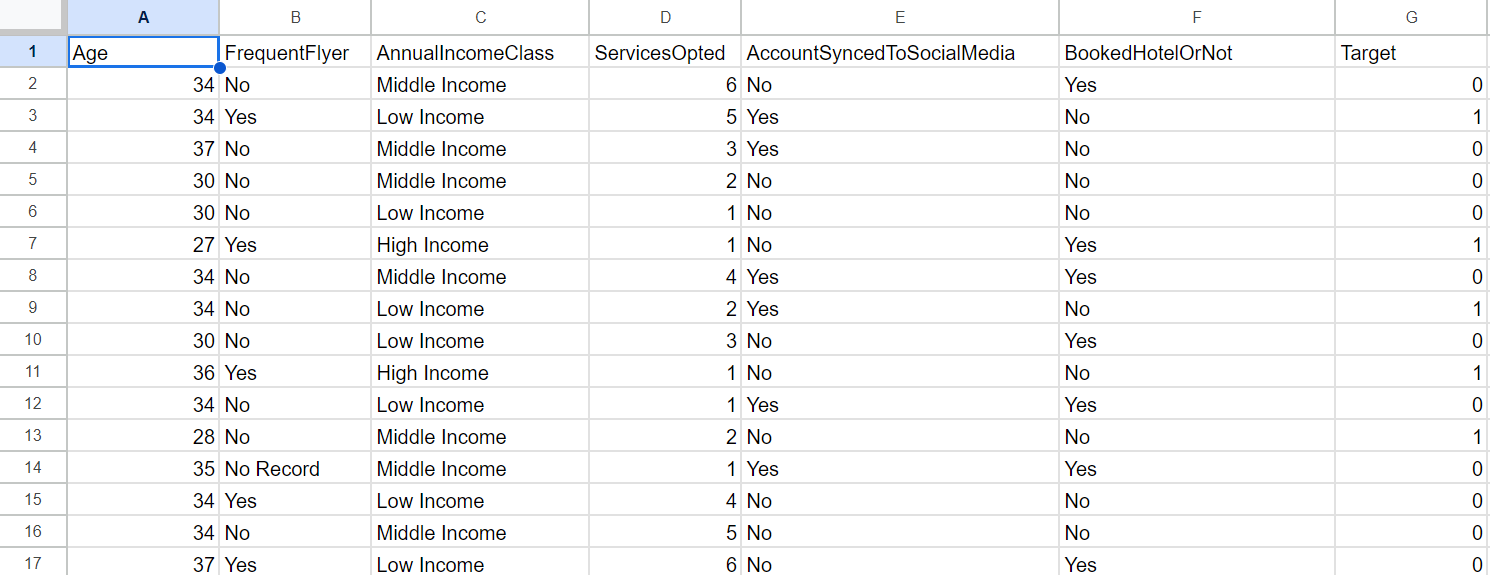
Quy trình thực hiện :

* Thu thập dữ liệu: Thu thập dữ liệu từ các nguồn như hệ thống đặt tour, cơ sở dữ liệu khách hàng, các trang web đánh giá du lịch, mạng xã hội, và các cuộc khảo sát.
* Khám phá dữ liệu (EDA - Exploratory Data Analysis): Thực hiện EDA để hiểu cấu trúc và tính chất của dữ liệu. Kiểm tra các biến, xác định giá trị thiếu, phân phối của các biến, và mối tương quan giữa chúng.
* Xử lý dữ liệu thiếu: Đối mặt với dữ liệu thiếu, điền giá trị thiếu bằng các phương pháp như điền giá trị trung bình, trung vị, hoặc sử dụng các mô hình học máy để dự đoán giá trị thiếu.
* Chuẩn hóa dữ liệu: Chuẩn hóa các biến nếu cần thiết để đưa chúng về cùng một phạm vi, ví dụ như chuẩn hóa các biến số liên tục để có trung bình bằng 0 và độ lệch chuẩn bằng 1.
* Mã hóa biến phân loại: Mã hóa các biến phân loại thành biến số để có thể sử dụng trong mô hình học máy. Các phương pháp mã hóa bao gồm One-Hot Encoding, Label Encoding, hoặc Target Encoding.
* Loại bỏ biến không cần thiết: Xác định và loại bỏ các biến không cần thiết hoặc không ảnh hưởng đến việc dự đoán tỷ lệ khách hàng rời bỏ.
* Lưu trữ dữ liệu đã tiền xử lý: Lưu trữ dữ liệu đã tiền xử lý vào các tập tin hoặc cơ sở dữ liệu để sử dụng cho các bước tiếp theo trong quá trình dự đoán tỷ lệ khách hàng rời bỏ.

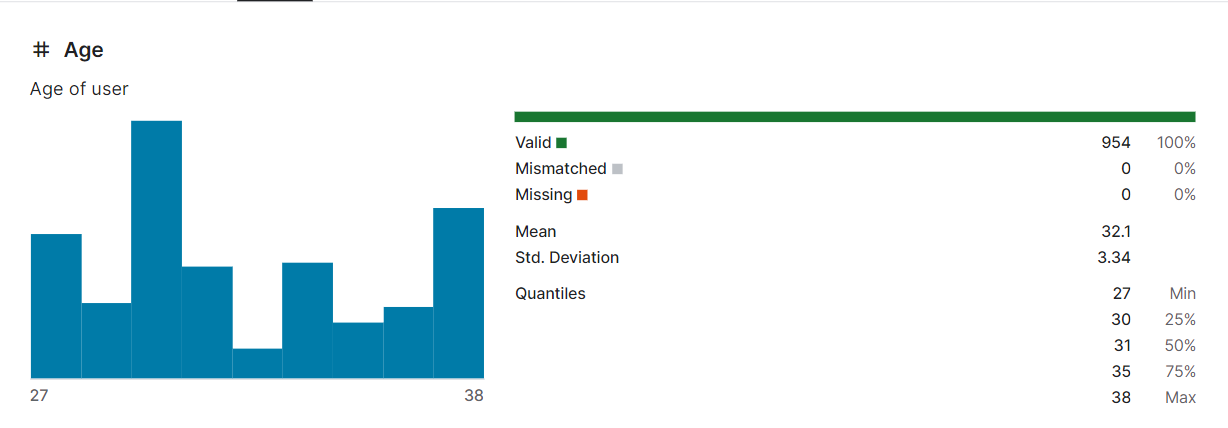
## 

## 2.1. Phân tích dữ liệu:

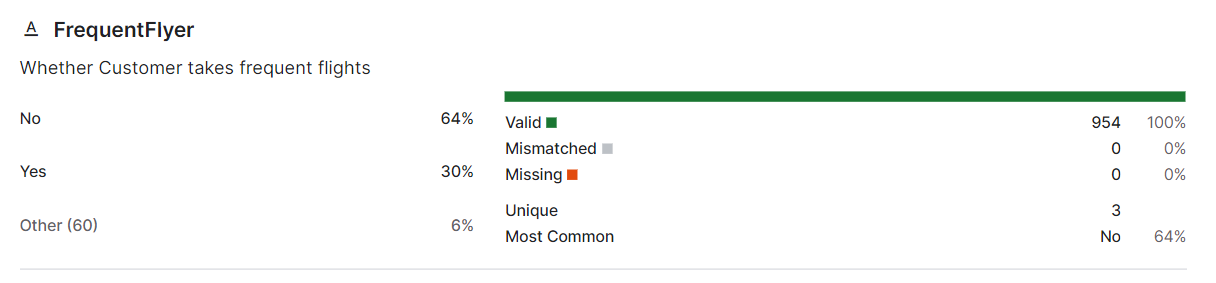
* Giới thiệu về dữ liệu: Dữ liệu thu thập từ nguồn:<https://www.kaggle.com/datasets/tejashvi14/tour-travels-customer-churn-prediction>
* Dataset gồm 7 trường dữ liệu:
* Age(Tuổi của người dùng);
* FrequentFlyer (Khách hàng có thường xuyên sử dụng dịch vụ hay không?);
* AnnuallIncomeClass ( Mức thu nhập của người dùng);
* ServiceOpted (Số lần khách hàng sử dụng dịch vụ của công ty);
* AccountSyncedToSocialMedia (Khách hàng có kết nối tài khoản mạng xã hội cá nhân với tài khoản sử dụng dịch vụ của công ty hay không?);
* BookedHotelOrNot (Khách hàng có sử dụng dịch vụ công ty để đặt phòng không?);
* Target (Khách hàng có ngưng sử dụng dịch vụ hay không? 1 - khách hàng đã ngưng sử dụng dịch vụ, 0 - khách hàng vẫn sử dụng dịch vụ);
* Dataset trên bao gồm 954 bản ghi.
* Đi kèm với dataset thu thập được là đồ thị biểu diễn phần trăm các giá trị tuy nhiên các giá trị này chưa qua xử lý.
* Dữ liệu thu thập được :



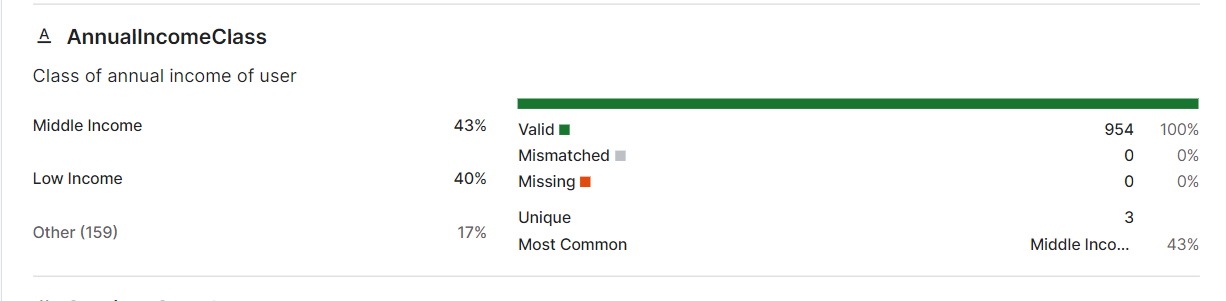
* Các columns trong dữ liệu :
* Age: Trường dữ liệu này hiển thị độ tuổi của khách hàng sử dụng dịch vụ



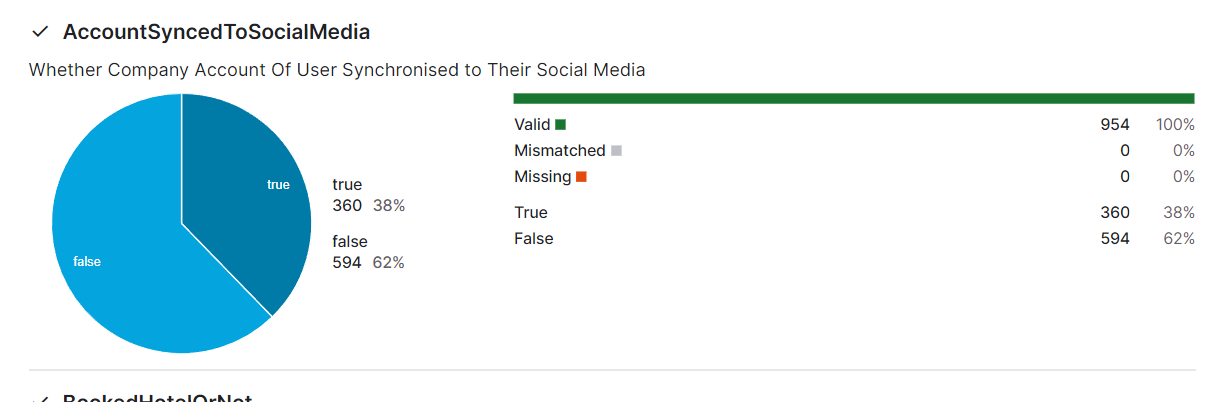
* FrequentFlyer : Trường dữ liệu này gồm 3 kiểu: Yes(Khách hàng sử dụng dịch vụ thường xuyên); No(Khách hàng không sử dụng thường xuyên)



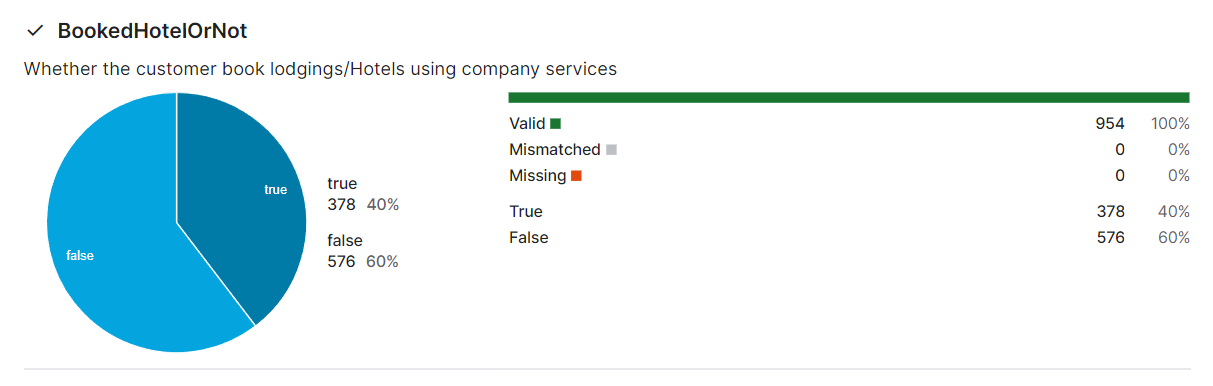
* AnnuallncomeClass :

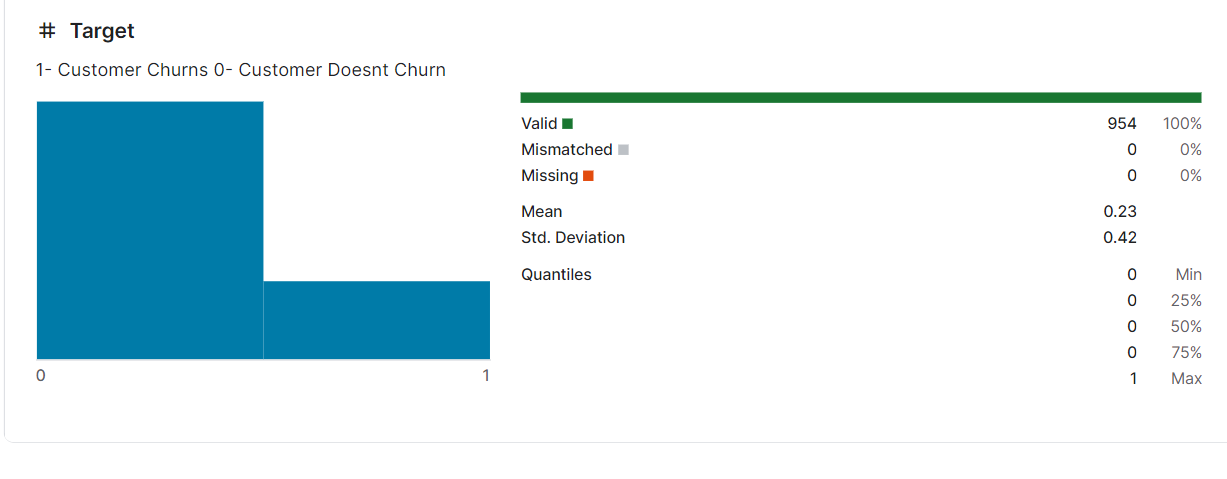


* AccountSyncedToSocialMedia:



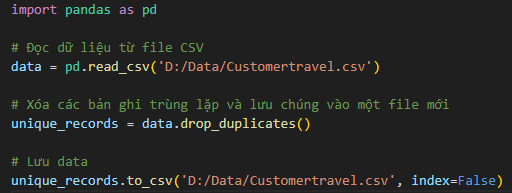
* Booked Hotel Or Not :

- Target : xác định khách hàng có ngưng sử dụng dịch vụ hay không

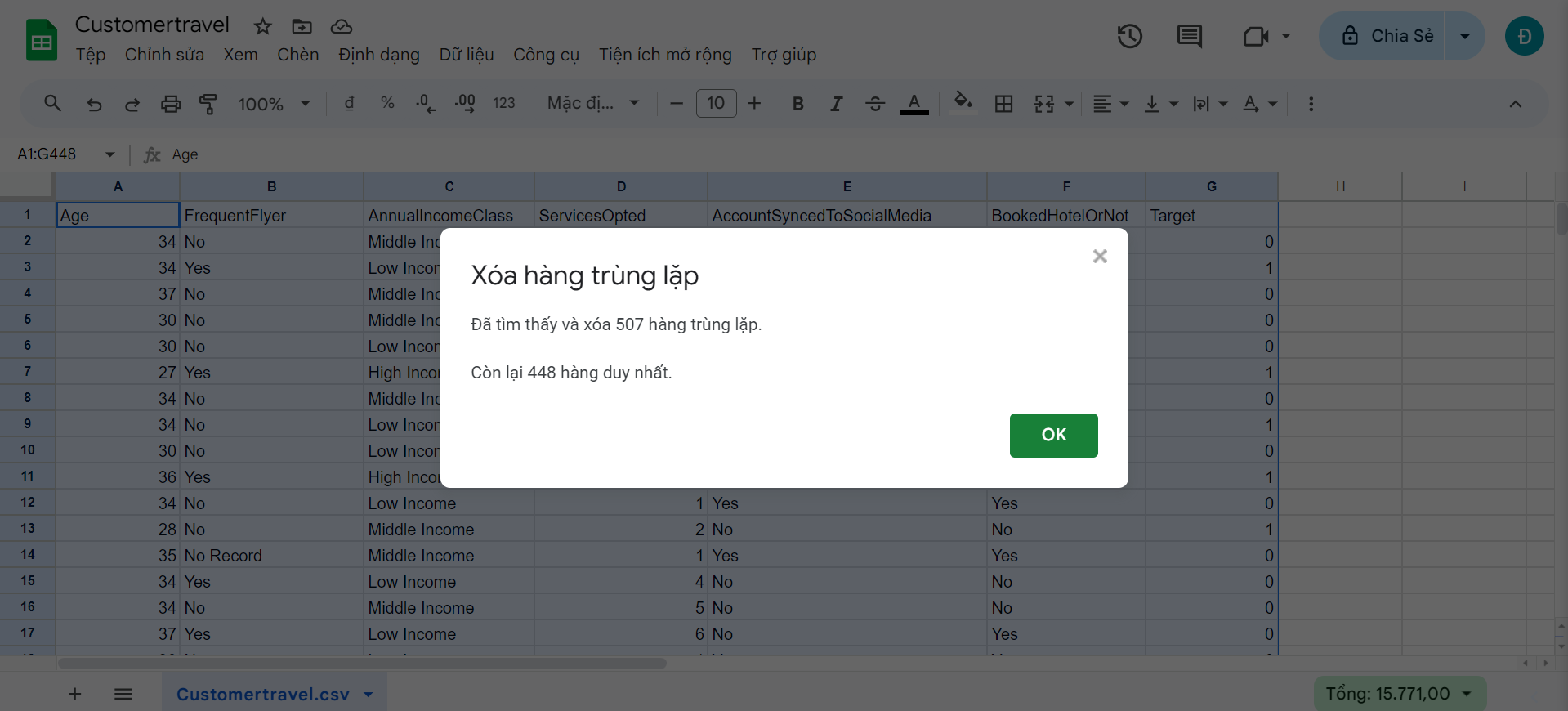


## 2.2. Tiến hành các bước tiền xử lý dữ liệu:

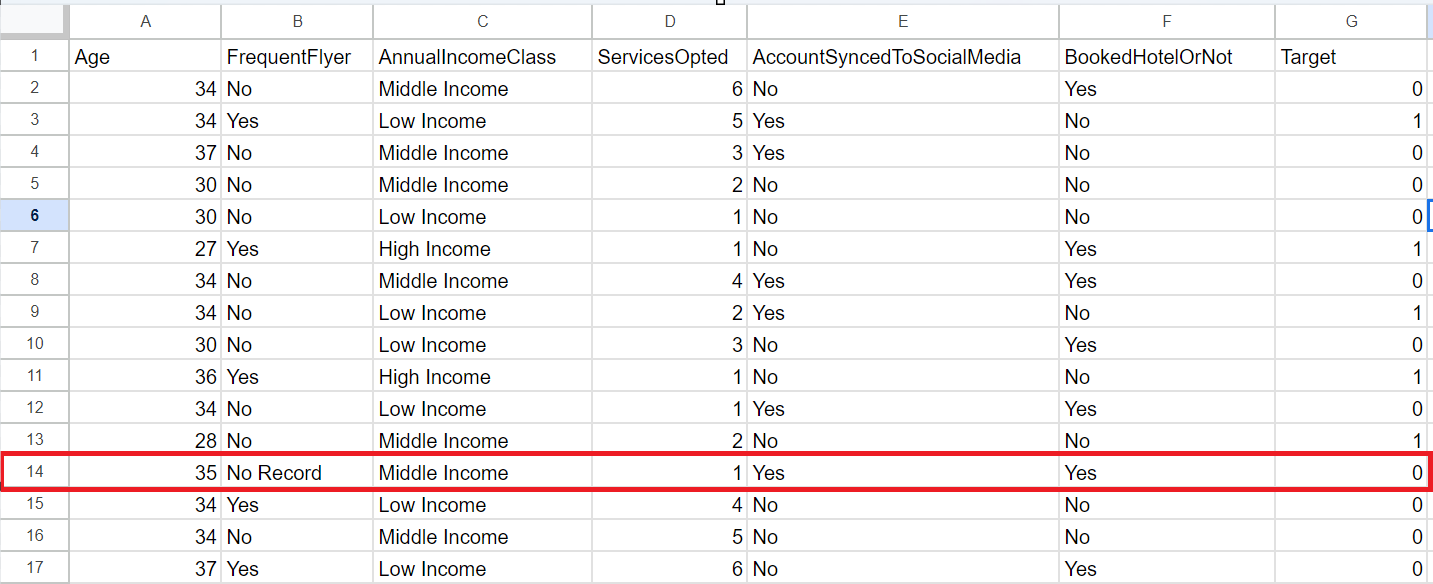
* Xử lý dữ liệu trùng lặp:
* Dữ liệu trùng lặp có thể là các hàng hoàn toàn giống nhau hoặc các hàng có các giá trị trùng lặp trong một số cột cụ thể. Sau khi xác định dữ liệu trùng lặp, loại bỏ các bản ghi trùng lặp hoặc chỉ giữ lại một bản ghi duy nhất từ mỗi nhóm bản ghi trùng lặp. Việc loại bỏ các dữ liệu dư thừa này giúp cho:
* Tệp dữ liệu trở lên gọn nhẹ hơn
* Tránh trường hợp overfitting khi tiến hành nghiên cứu.
* Giúp khi tiến hành huấn luyện trở nên hiệu quả hơn, giảm thời gian khi huấn luyện
* Sử dụng hàm drop\_duplicates() để xác định các bản ghi trùng lặp và loại bỏ các bản ghi đó. Drop\_duplicates() so sánh giá trị của từng cột để xác định các bản ghi trùng lặp. Nếu tất cả các giá trị trong mỗi cột của hai hàng là giống nhau, thì các hàng được coi là trùng lặp và một trong số chúng sẽ được loại bỏ.



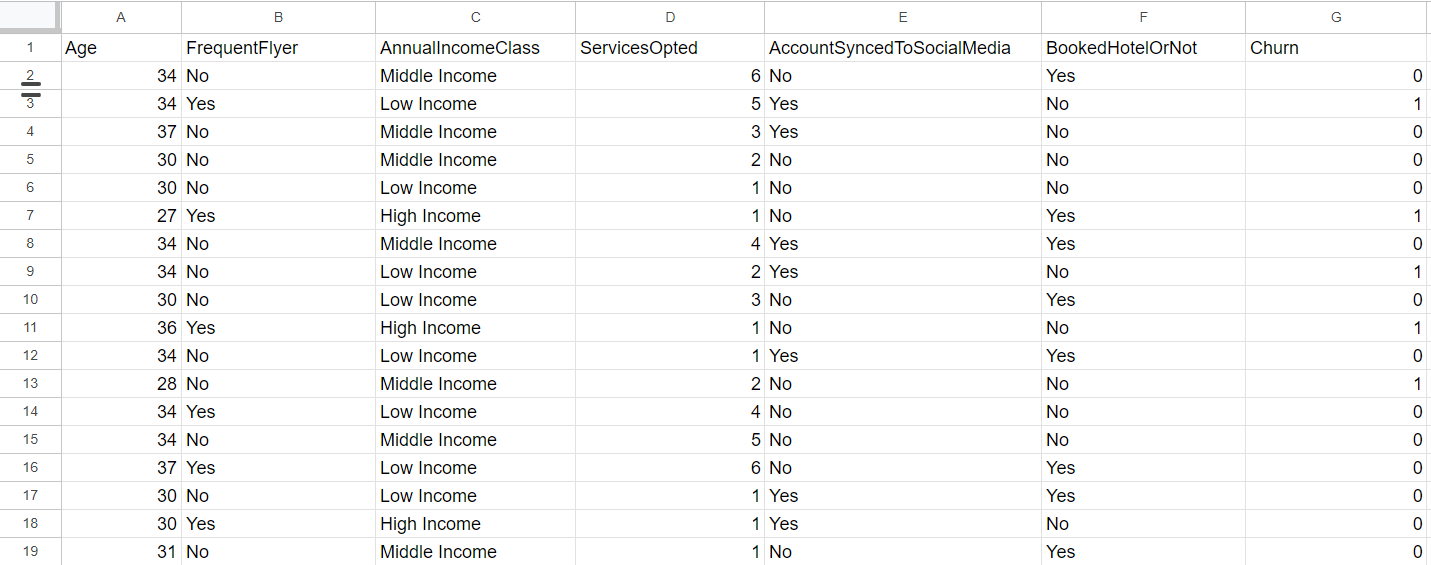
* Loại bỏ dữ liệu bằng phương pháp thủ công trên excel.



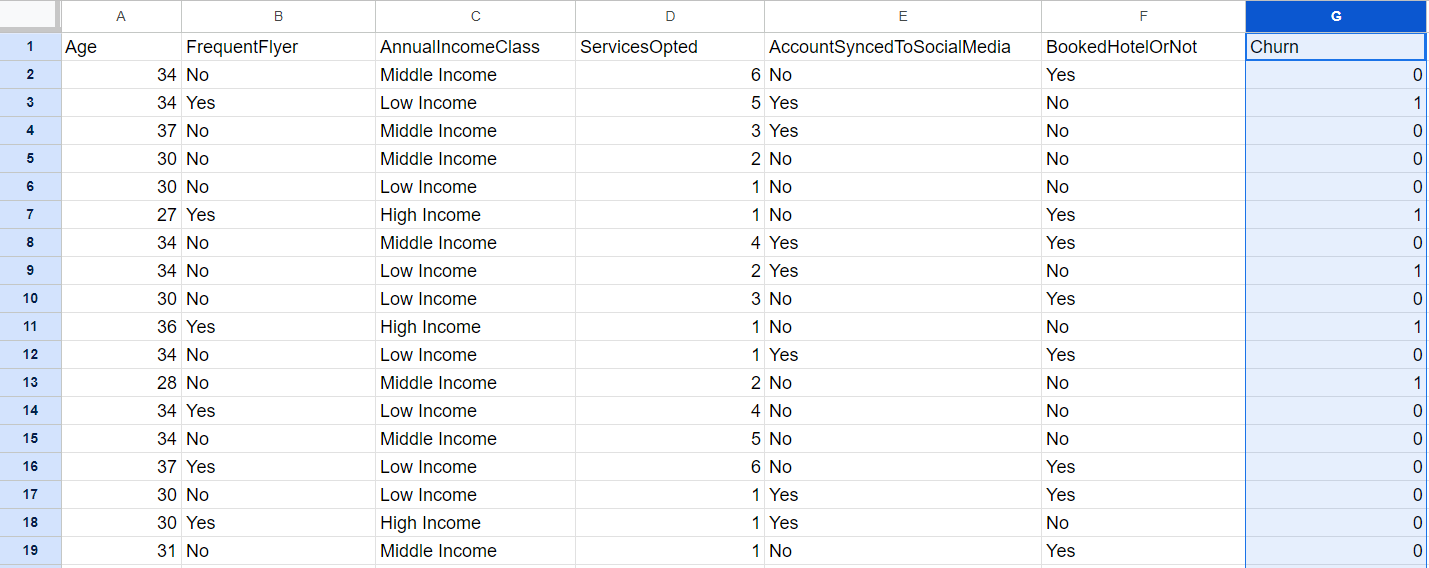
* Loại bỏ dữ liệu nhiễu: No Record:
* Tại trường dữ liệu “Frequent Flyer” ngoài các dữ liệu ‘Yes’ và ‘No’ xuất hiện các dữ liệu ‘NO RECORD’. Vì trường dữ liệu Frequent Flyer để xác định xem hành khách có thường xuyên i du lịch hay không nên chỉ xét YES hoặc NO.



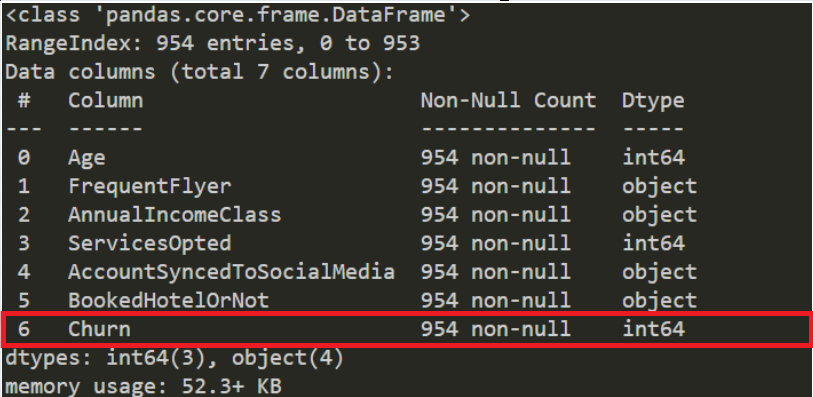
* Xét 1 đoạn các bản ghi ‘No record’ tại cột ServicesOpted - số lần khách hàng sử dụng dịch vụ của công ty. Một số bản ghi ‘No record’ lại có số lần dịch vụ từ 2 trở nên, trong thực tế nếu không có bản ghi không có dữ liệu nào mà lại có số lần sử dụng dịch vụ nhiều lần, vì vậy các bản ghi ‘No record’ là dữ liệu nhiễu nên việc loại bỏ nó là cần thiết, việc loại bỏ nó sẽ làm dữ liệu thêm chính xác hơn.
* 1 phần dữ liệu sau khi loại bỏ các bản ghi ‘No record’:



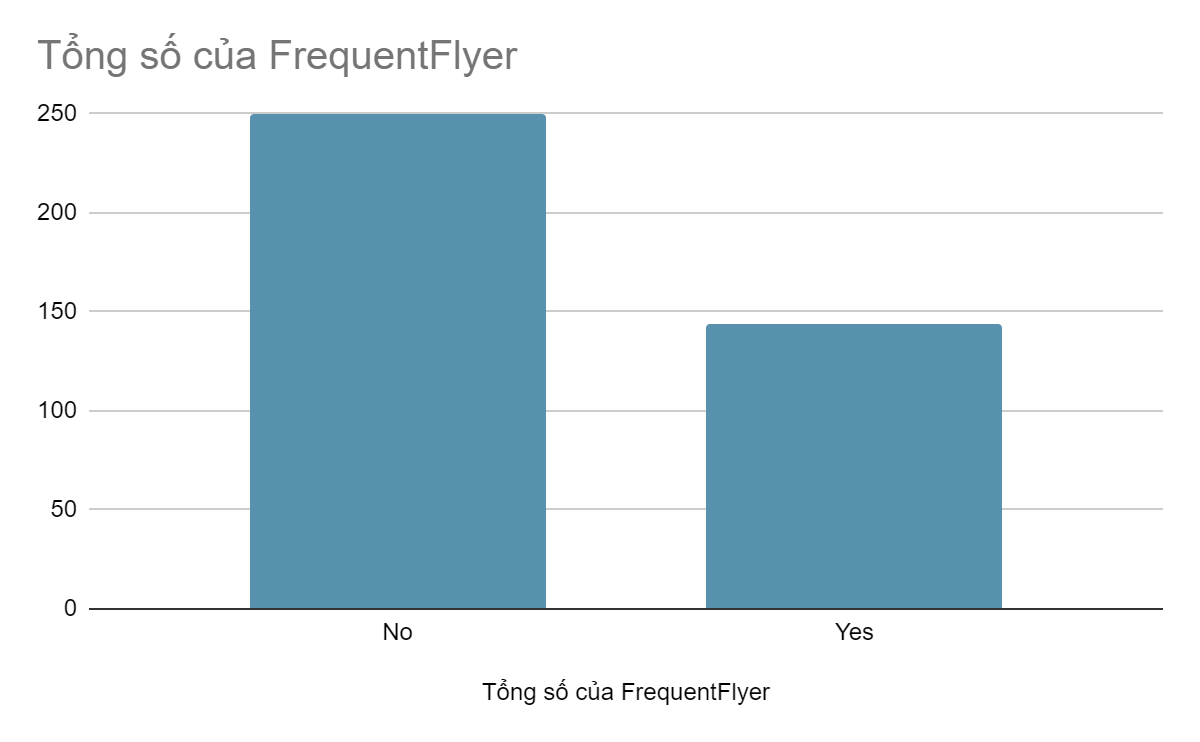
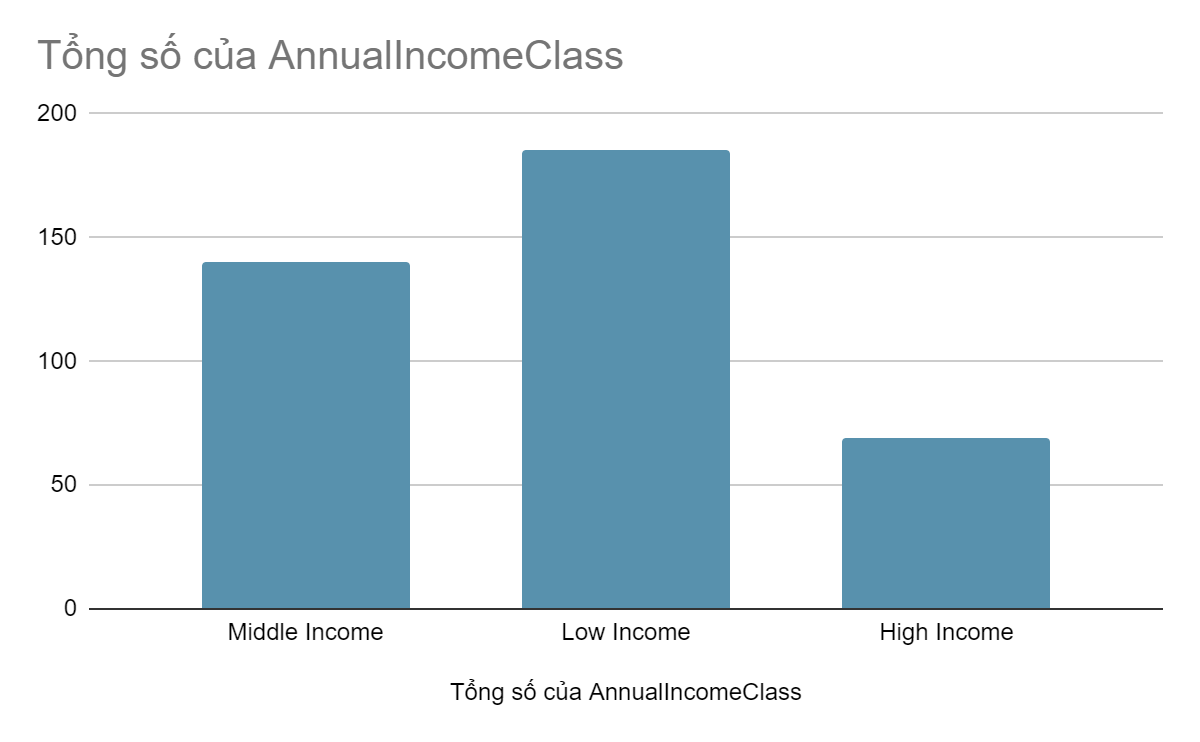
* Gán nhãn dữ liệu
* Trường dữ liệu cuối cùng ‘Target’ với ý nghĩa xác định khách hàng có hoặc không ngừng sử dụng dịch vụ của công ty tuy nhiên với nhãn này khó có thể hiểu đúng ý nghĩa nếu không có chú thích. Giải pháp là thay đổi nhãn để có thể dễ hiểu hơn ‘Target’ -> ‘Churn’

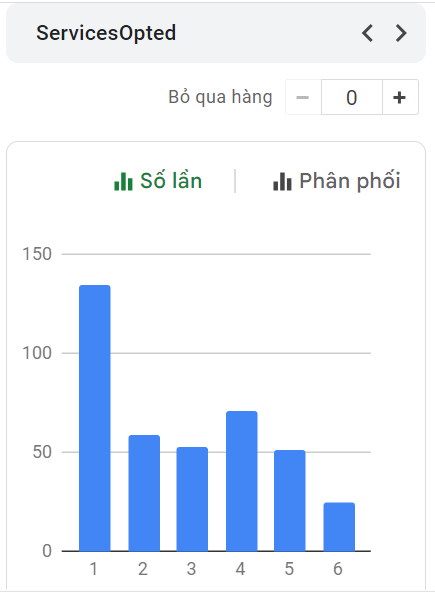


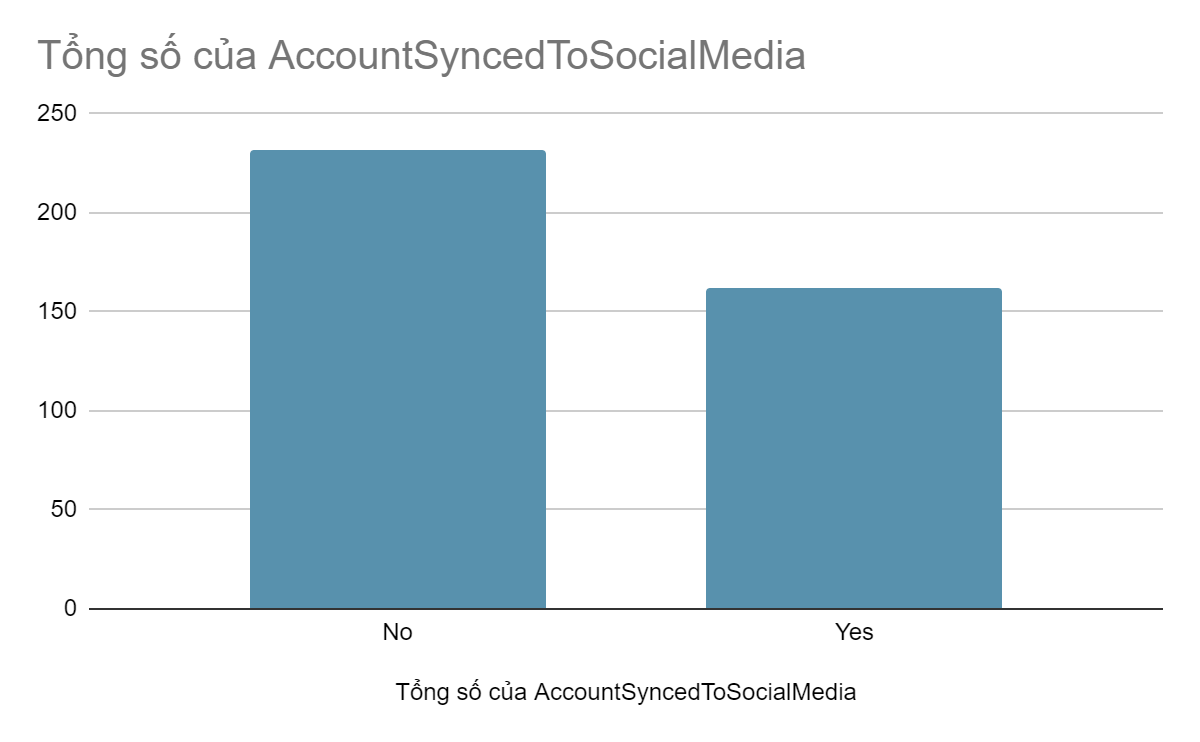


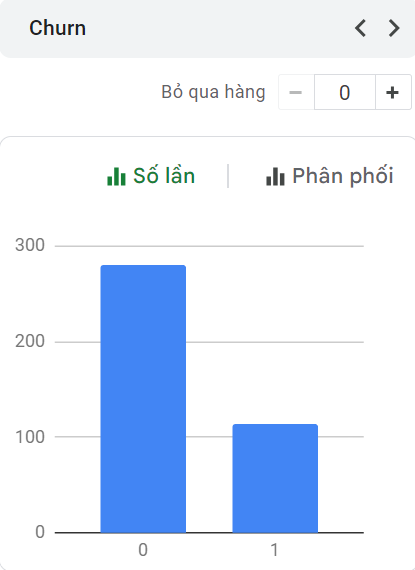


## 2.3. Phân tích dữ liệu sau khi làm sạch dữ liệu:

* Frequent Flyer: 
* Annual Income Class:
* Services Opted:



* Account Synced To Social Media:
* Booked Hotel Or Not: 
* Churn:



# Chương 3. Xây dựng mô hình

Lựa chọn hướng giải quyết, lựa chọn thuật toán, công nghệ sử dụng cần căn cứ theo tính chất của dữ liệu và mục tiêu của đề tài, kết hợp với việc tham khảo hướng giải quyết từ các đề tài liên quan.

**3.1 Hướng giải quyết :**

Phân tích dữ liệu: Bắt đầu bằng việc thu thập và phân tích dữ liệu khách hàng đã chia tay (churned customers) và các khách hàng hiện tại. Điều này có thể bao gồm việc phân tích lịch sử, hành vi sử dụng dịch vụ, lịch sử đặt hàng, thông tin cá nhân, v.v.

Xây dựng mô hình học máy: Sử dụng các thuật toán học máy như Random Forest, DecisionTree, GradientBoosting, XGBoost, LightGBM, CatBoost, để xây dựng mô hình dự đoán khách hàng rời bỏ dịch vụ. Đối với mô hình này, cần phải chia tách dữ liệu thành tập huấn luyện và tập kiểm tra để đánh giá hiệu suất của mô hình.

Tinh chỉnh mô hình: Tinh chỉnh các tham số của mô hình để tối ưu hóa hiệu suất. Điều này có thể bao gồm việc sử dụng kỹ thuật tối ưu hóa siêu tham số (hyperparameter optimization) hoặc kiểm tra các biến thể của mô  hình.

Sử dụng kỹ thuật tăng cường dữ liệu: Nếu dữ liệu thu thập ban đầu không đủ, sử dụng kỹ thuật tăng cường dữ liệu (data augmentation) để tạo ra thêm dữ liệu mô phỏng hoặc tổng hợp.

Phân tích đặc trưng (feature engineering): Tìm kiếm và chọn lọc các đặc trưng quan trọng từ dữ liệu, hoặc tạo ra các đặc trưng mới từ dữ liệu hiện có để cải thiện hiệu suất của mô hình.

Kiểm tra và đánh giá: Kiểm tra hiệu suất của mô hình trên tập kiểm tra và sử dụng các độ đo như độ chính xác (accuracy), độ phủ (recall), và F1-score để đánh giá mô hình.

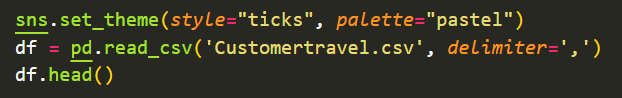
Triển khai và theo dõi: Sau khi xây dựng và kiểm tra mô hình, triển khai nó vào môi trường thực tế và theo dõi hiệu suất của nó thường xuyên. Điều này giúp bạn cập nhật và cải thiện mô hình theo thời gian để duy trì tính khả dụng của nó.

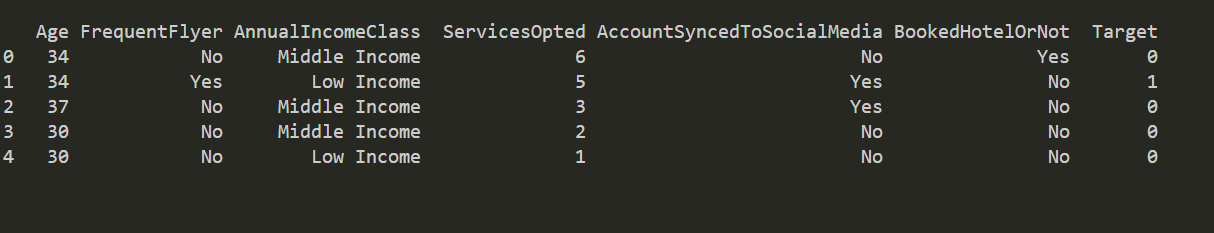
### Xây dựng thuật toán

#### 3.2.1. Khai báo thư viện (Import library) :



#### 3.2.2. Load dữ liệu (Data set loading) :

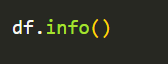


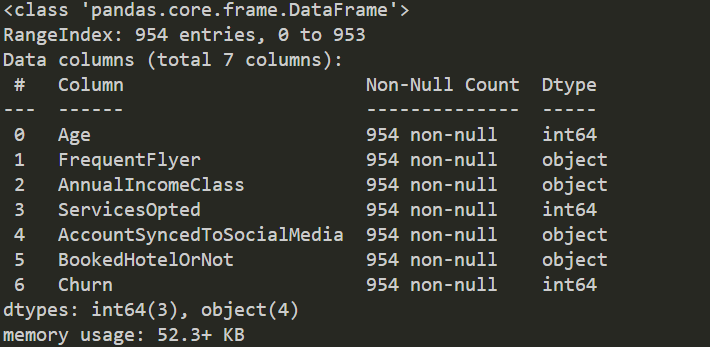


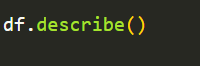


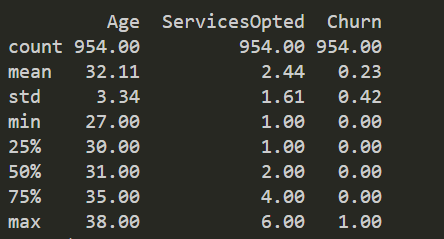
#### 3.2.3. Phân tích giải thích dữ liệu (Explanatory data analysis) :

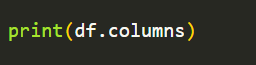
* Tóm tắt tập dữ liệu (Summarize dataset) :

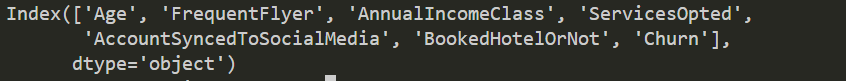




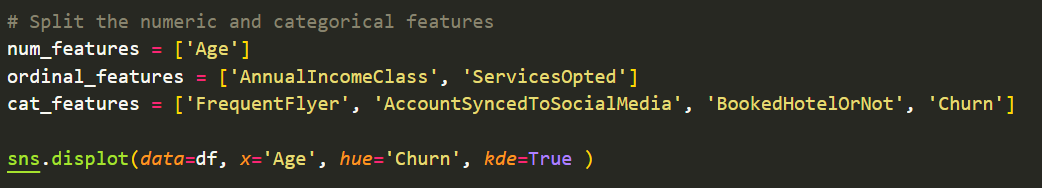


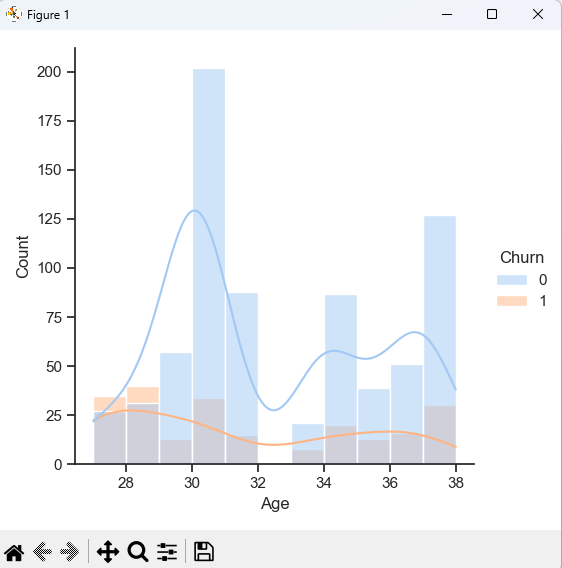


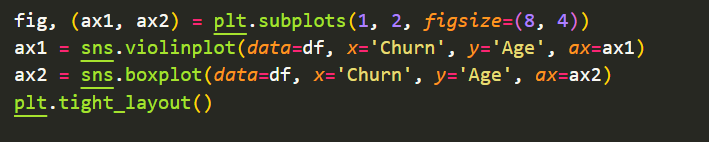


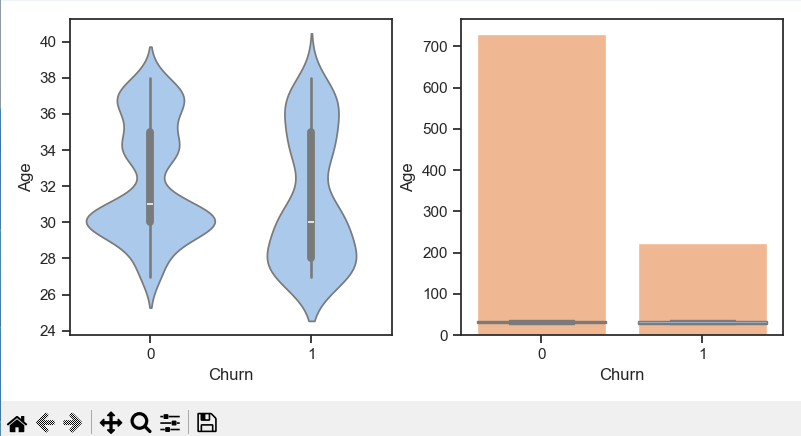


* Visualization:



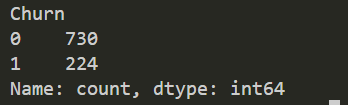




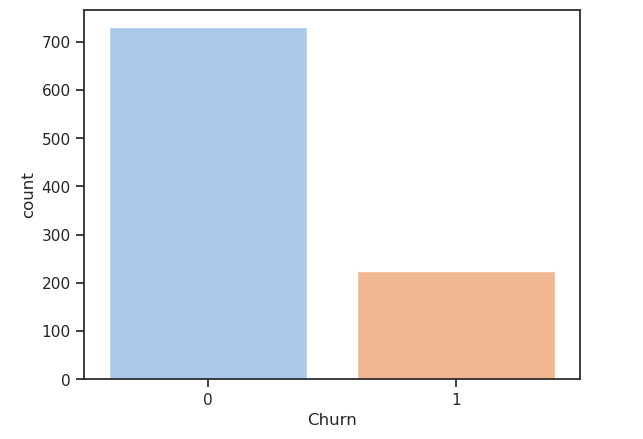


* Trực quan hóa biến phân loại( Visualizing categorical variables) :

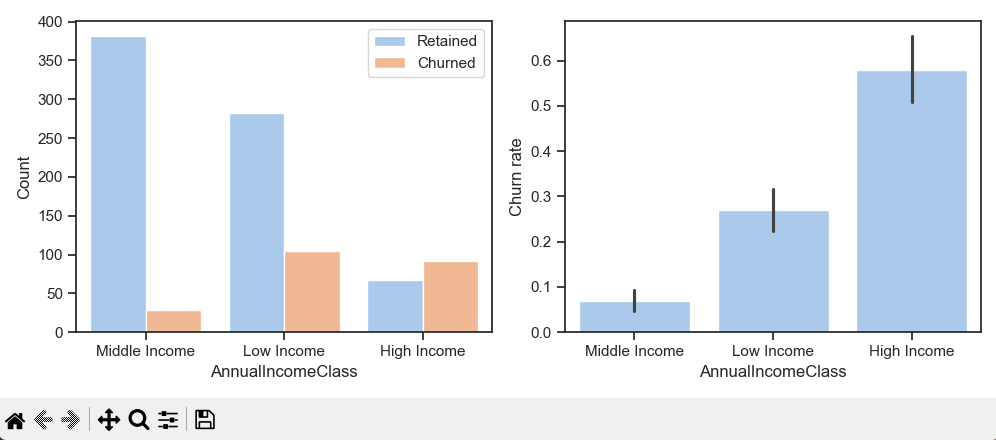


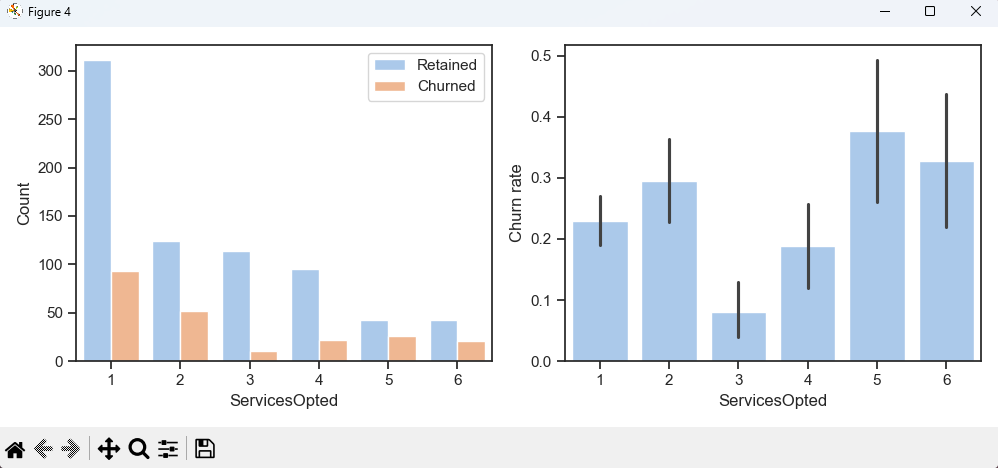


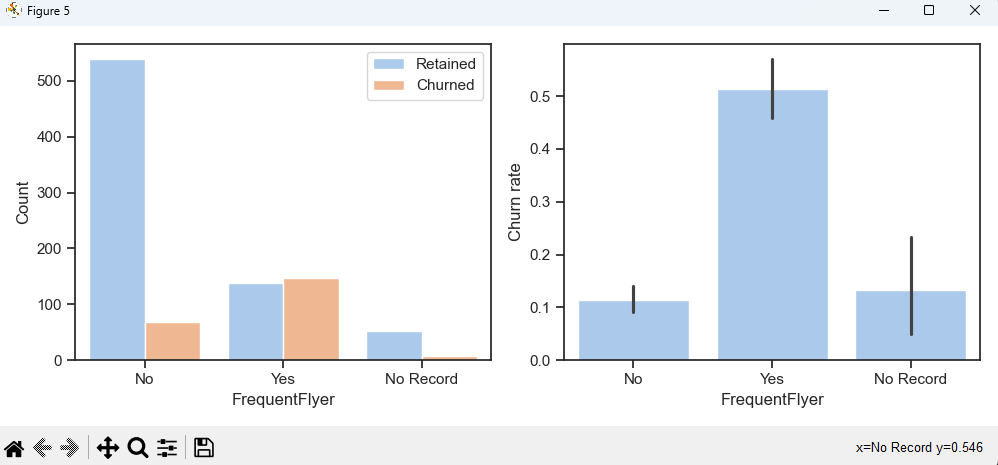


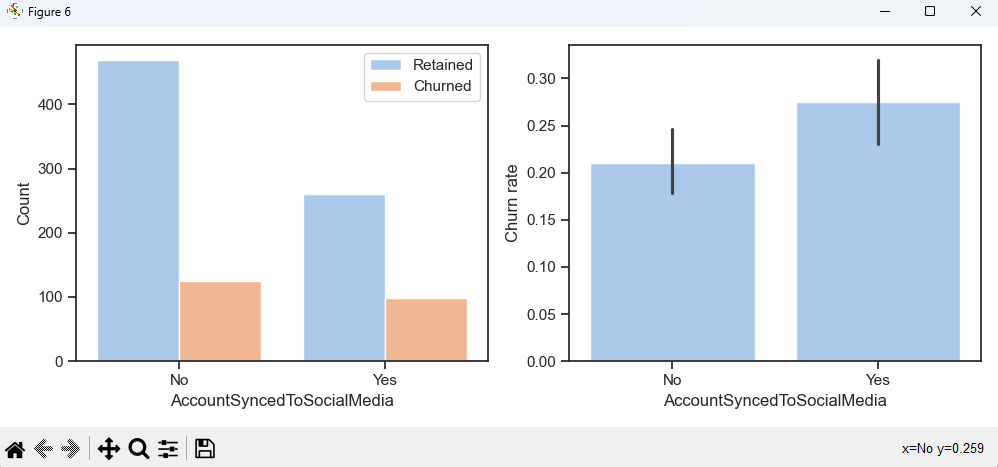


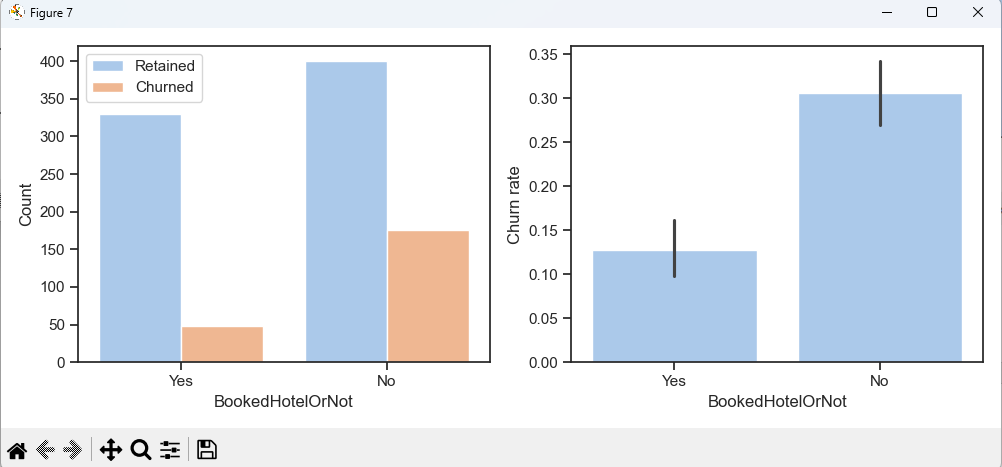






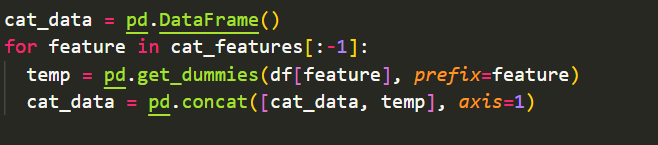




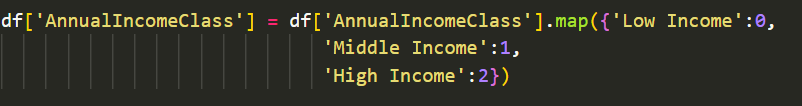


#### 3.2.4.Tiền xử lý dữ liệu (Data preprocessing) :

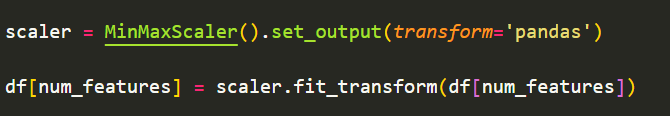
* Mã hóa một lần các biến phân loại (One-hot encoding of categorical variables):



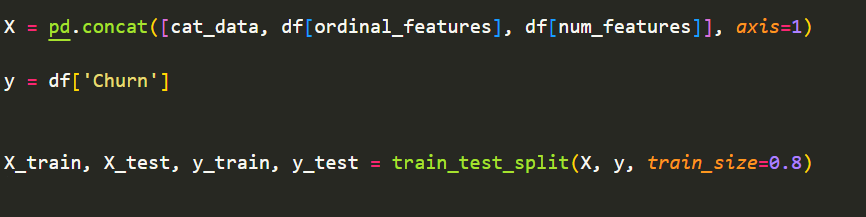
* Mã hóa thứ tự của các biến thứ tự (Ordinal-encoding of ordinal variables):



* Thang đo tối thiểu – tối đa cho các biến số (Min-max scale for numerical variables) :

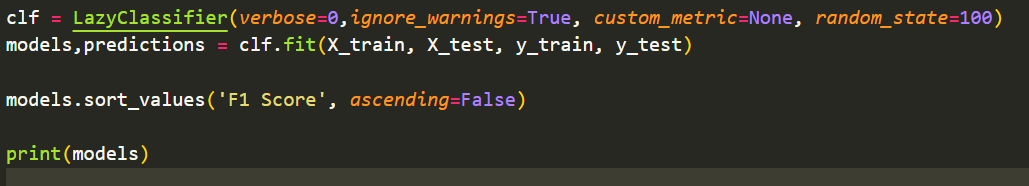


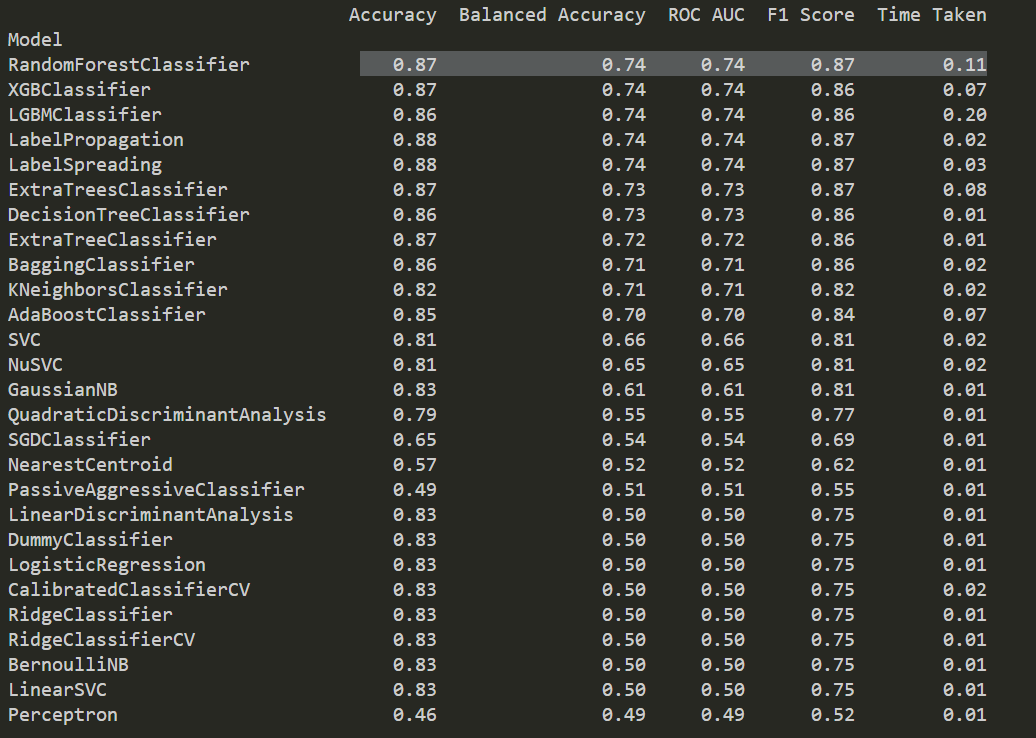
* Train test split :



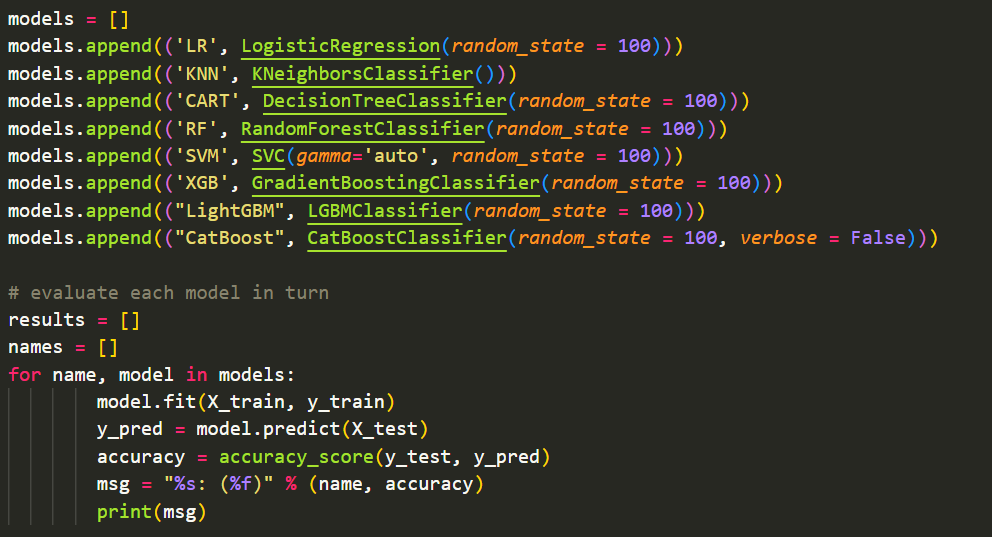
#### 3.2.5. Modeling :

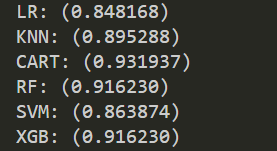
* AutoML model( lazy predict) :





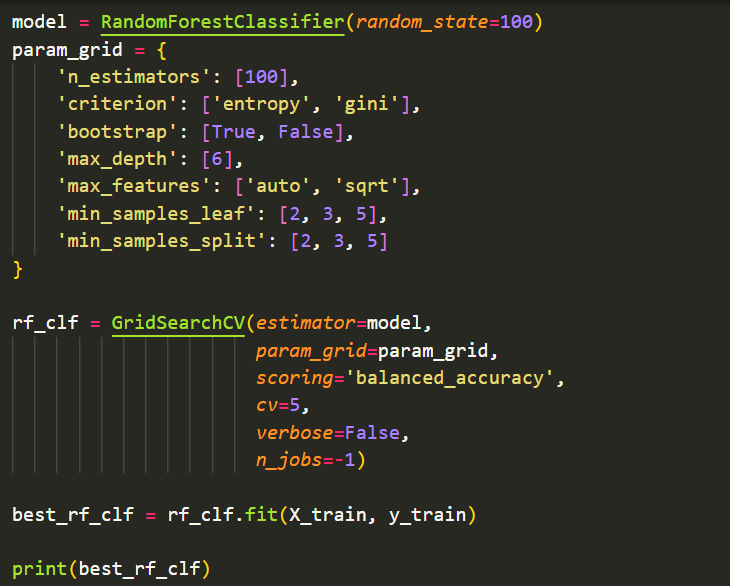
* Sklearn models

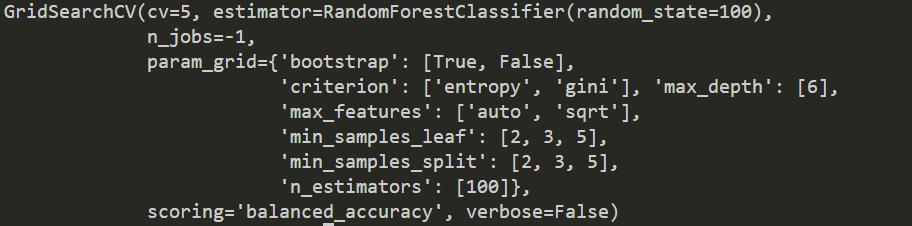




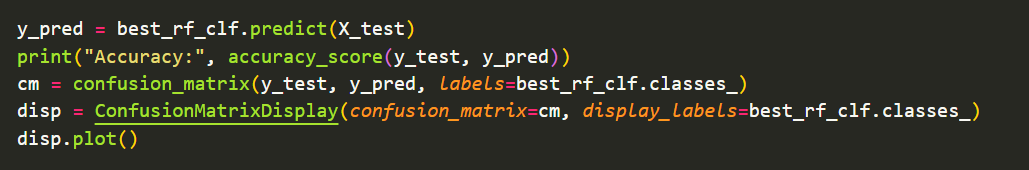


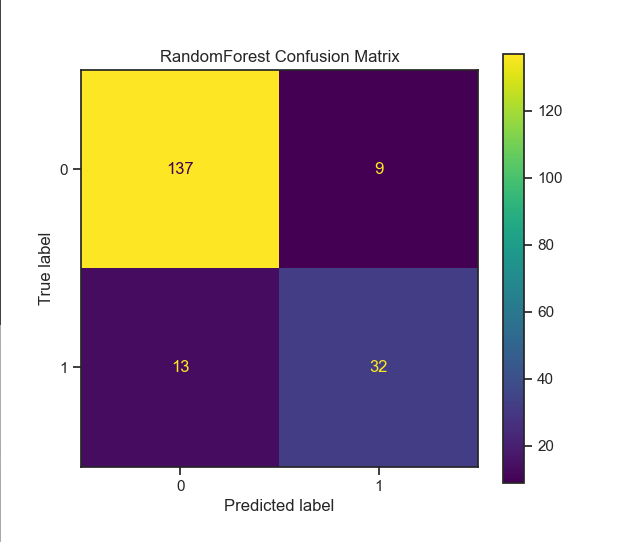
#### 3.2.6. Model tuning (Random Forest) :

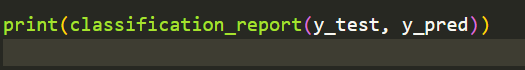


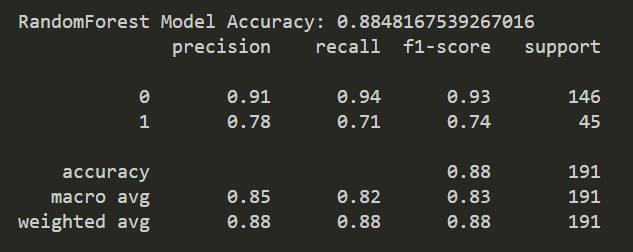


#### 3.2.7. Kết quả mô hình tốt nhất (RandomForestClassifier) :

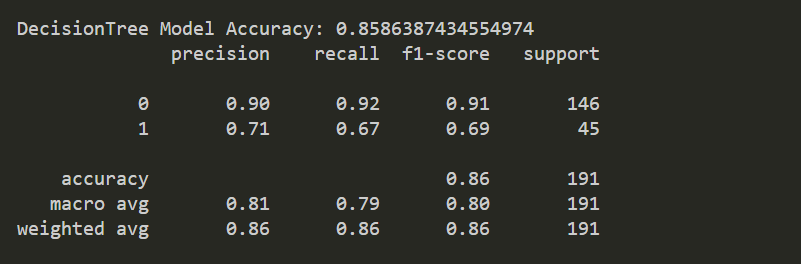


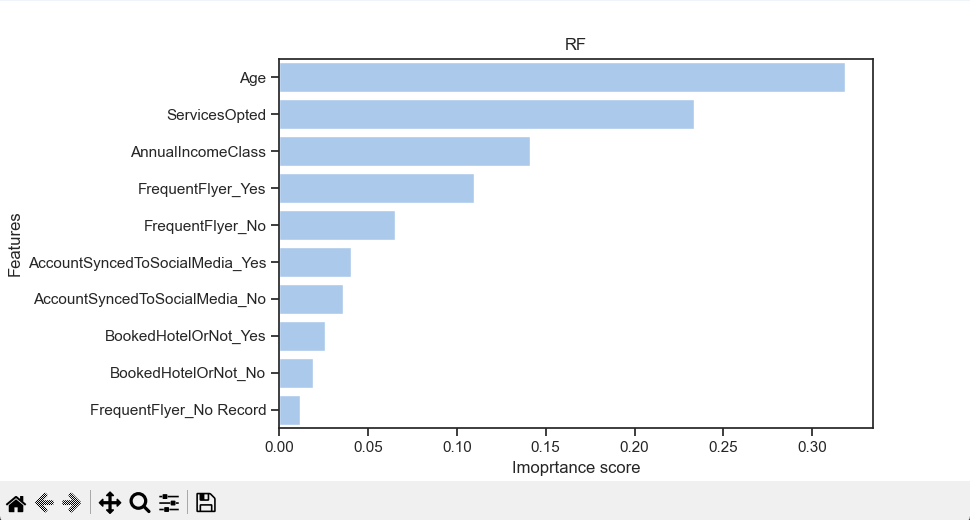


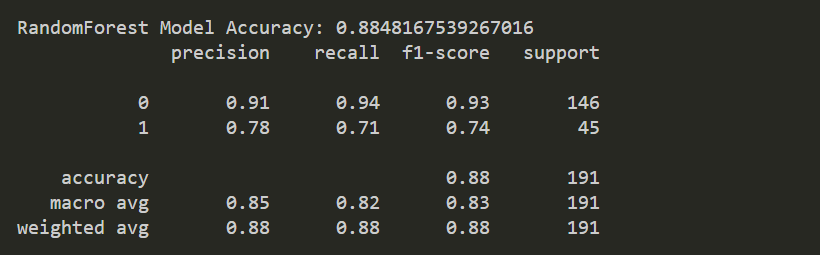


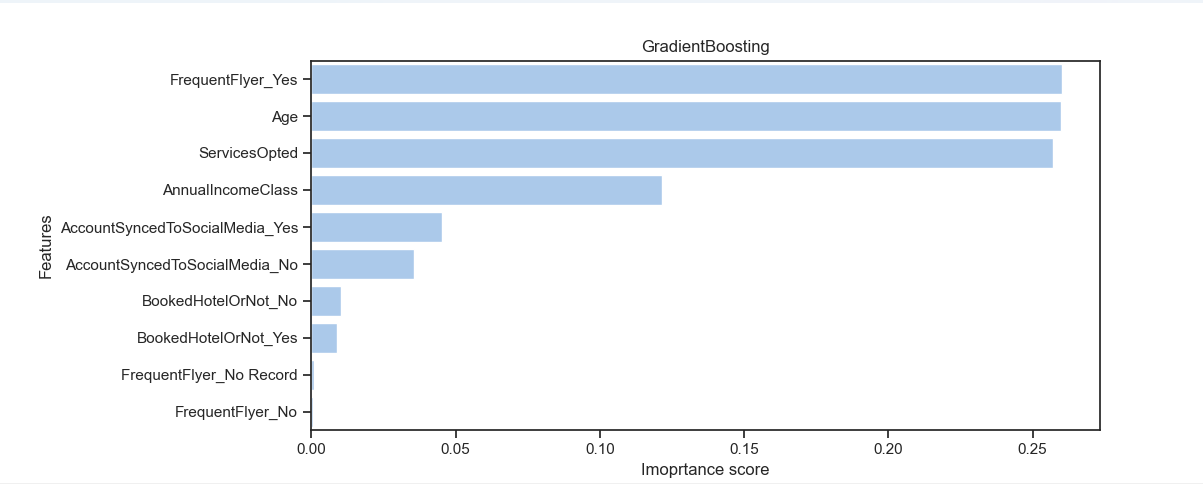


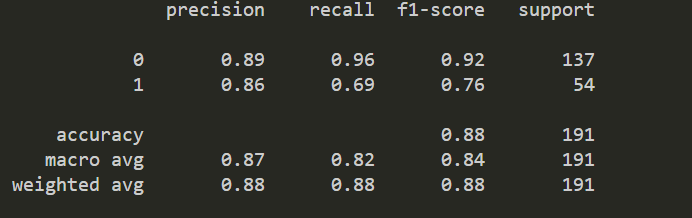
#### 3.2.8. Phân tích tính năng :

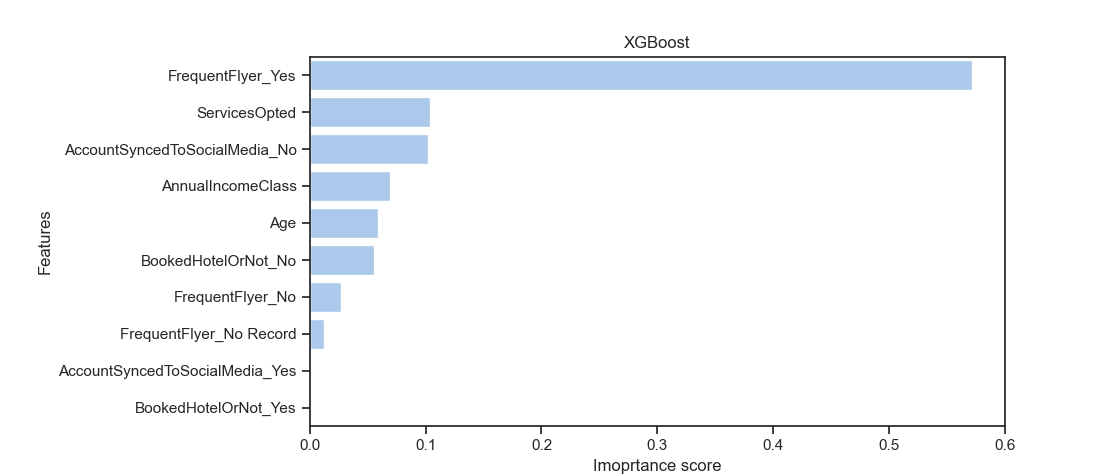


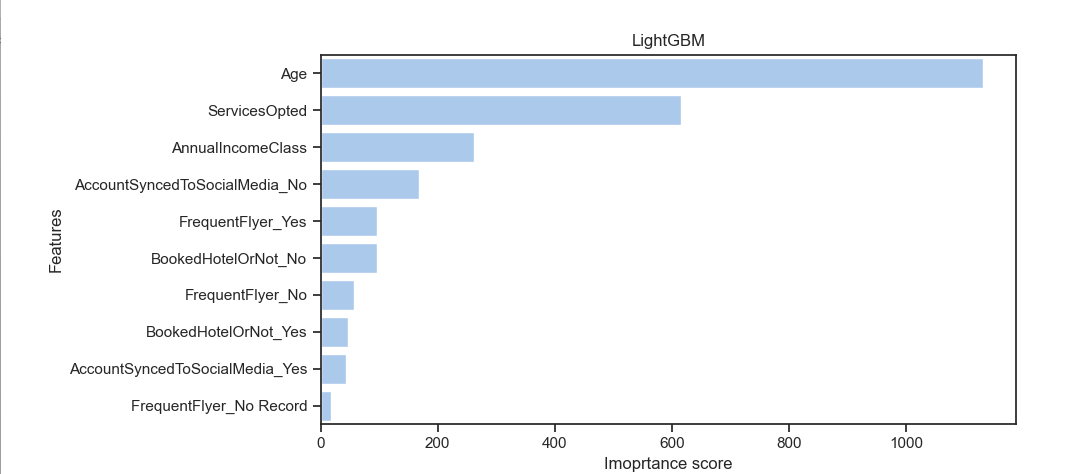
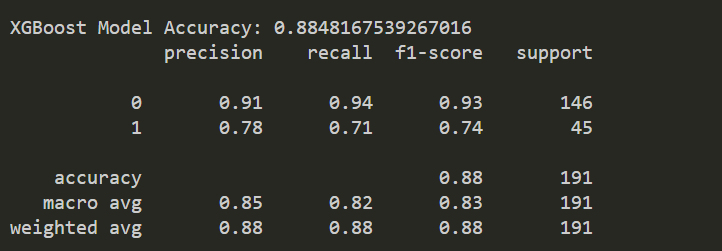


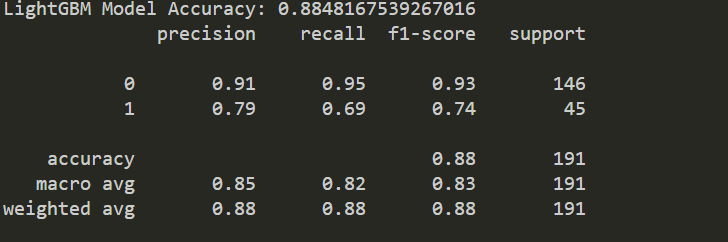


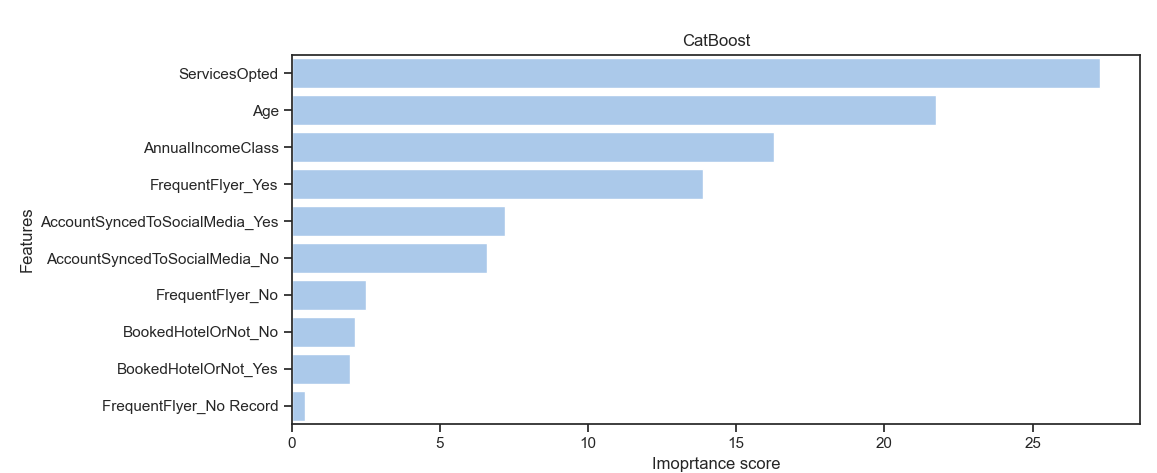


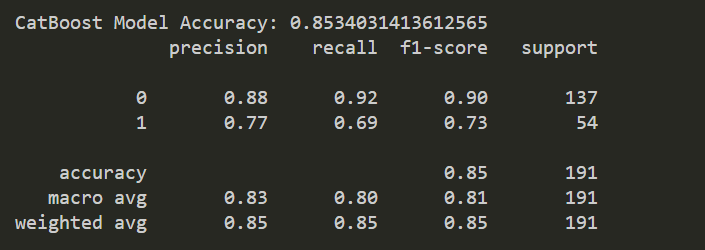












## 3.3. Lựa chọn Random forest:

* **Cơ sở lý thuyết**
* Random Forest là một phương pháp học máy dựa trên cả cây quyết định và sự kết hợp của chúng. Phương pháp này hoạt động bằng cách xây dựng một loạt các cây quyết định tại các điểm dữ liệu ngẫu nhiên, sau đó kết hợp kết quả từ các cây này để tạo ra dự đoán cuối cùng. Quan trọng nhất, Random Forest có khả năng giảm overfitting, một vấn đề thường gặp trong việc xây dựng các mô hình phức tạp.
* Thuật toán:
* Input: Bộ dữ liệu gồm n dữ liệu (sample) và mỗi dữ liệu có k thuộc tính (feature).
* Output: Dự đoán cho dữ liệu mới dựa trên tổng hợp kết quả từ các cây quyết định.
* Khởi tạo: chọn số lượng cây quyết định và chọn ngẫu nhiên một tập con của dữ liệu huấn luyện cho mỗi cây.
* Xây dựng cây quyết định:

Đối với mỗi cây:

Chọn ngẫu nhiên một tập con của các thuộc tính.

Xây dựng cây quyết định bằng cách sử dụng tập con dữ liệu và tập con thuộc tính đã chọn.

Tại mỗi nút trong cây:

Chọn thuộc tính tốt nhất để chia dữ liệu.

Chia dữ liệu thành hai phần dựa trên giá trị của thuộc tính được chọn.

Tiếp tục xây dựng cây con cho mỗi phần dữ liệu.

Dự đoán:

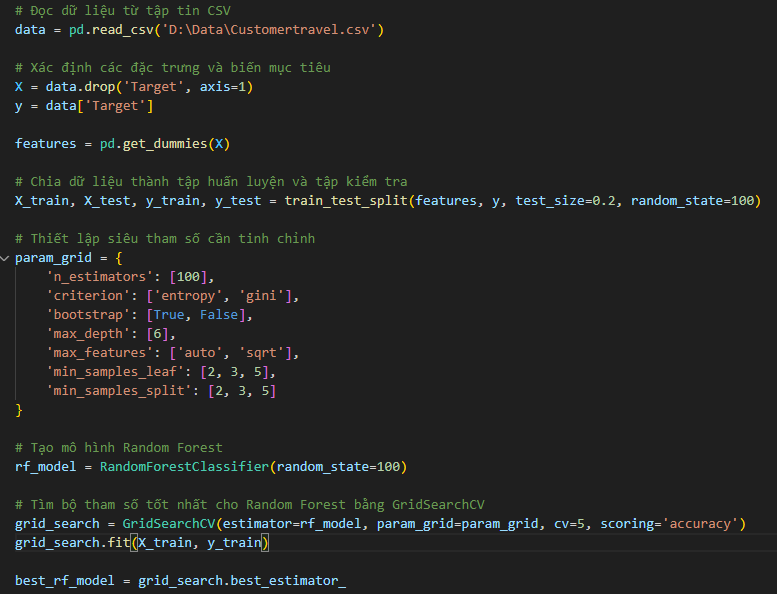
Đối với mỗi cây: Dự đoán kết quả bằng cách đi qua cây từ gốc đến lá.

Đối với bài toán phân loại: Đưa ra dự đoán cuối cùng bằng cách bỏ phiếu.

Kết quả cuối cùng: Kết quả đa số các cây.

* **Lý do chọn Random forest cho bài toán**
* Khả năng xử lý dữ liệu lớn: Random Forest có thể xử lý hiệu quả các tập dữ liệu lớn mà không cần phải giảm kích thước hoặc chọn lọc đặc trưng trước. Điều này phù hợp với các bài toán có tập dữ liệu lớn như trong lĩnh vực du lịch.
* Độ chính xác cao: Random Forest thường có độ chính xác cao trên các tập dữ liệu đa dạng. Với việc kết hợp nhiều cây quyết định, mỗi cây được huấn luyện trên một tập dữ liệu con khác nhau, giúp giảm thiểu overfitting và cải thiện khả năng tổng quát hóa của mô hình.
* Khả năng xử lý đặc trưng lớn: Random Forest không yêu cầu phải chọn lọc đặc trưng trước khi huấn luyện mô hình. Điều này làm cho thuật toán này phù hợp với các bài toán có nhiều đặc trưng, như trong bài toán dự đoán hành vi của khách hàng trong du lịch, nơi có thể có nhiều yếu tố ảnh hưởng.
* Khả năng xử lý dữ liệu không cân bằng: Random Forest có khả năng xử lý dữ liệu không cân bằng, tức là dữ liệu có một số lớp hoặc nhóm dữ liệu chiếm tỷ lệ cao hơn so với các lớp hoặc nhóm khác.
* Dễ dàng điều chỉnh siêu tham số: Random Forest có các siêu tham số có thể điều chỉnh để tinh chỉnh hiệu suất của mô hình, bao gồm số lượng cây quyết định, độ sâu của cây, và các tham số liên quan khác.
* **Xây dựng mô hình:**

Sử dụng GridSearchCV để tìm ra bộ tham số tốt nhất cho mô hình Random forest



## 

## 3.4. Đánh giá mô hình

* Confusion Matrix (Ma trận nhầm lẫn): là một công cụ quan trọng trong việc đánh giá hiệu suất của các mô hình phân loại. Trong đề tài này, Confusion Matrix sẽ giúp phân tích cách mà mô hình phân loại (Decision Tree và Random Forest) dự đoán các khách hàng rời bỏ tour du lịch (lớp 1) và không rời bỏ (lớp 0).
* Accuracy (Độ chính xác): Accuracy là tỉ lệ giữa số điểm dữ liệu được dự đoán đúng và tổng số điểm dữ liệu.



Trong đó:

* TP là số lượng true positives (số lượng dữ liệu positive dự đoán đúng).
* TN là số lượng true negatives (số lượng dữ liệu negative dự đoán đúng).
* FP là số lượng false positives (số lượng dữ liệu negative dự đoán sai).
* FN là số lượng false negatives (số lượng dữ liệu positive dự đoán sai).
* Specificity (Độ đặc hiệu):

Specificity là tỷ lệ phần trăm của các mẫu thực sự thuộc nhóm negative mà lại được dự đoán đúng.



* Precision (Độ chính xác):

Precision là tỉ lệ số điểm true positive trong số những điểm được phân loại là positive.



* Recall

Recall (còn được gọi là Sensitivity hoặc True Positive Rate) là tỷ lệ số điểm true positive trong số những điểm thực sự là positive.

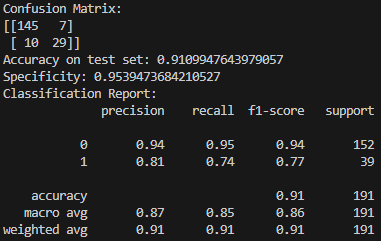


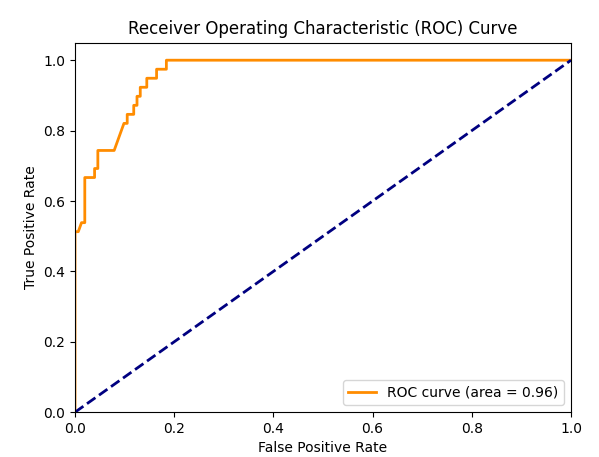
* F1-Score

F1-Score là một số liệu kết hợp giữa Precision và Recall, là trung bình điều hòa của hai giá trị này.



### Đánh giá mô hình random forest





* Accuracy: Độ chính xác của Random Forest là 91.10%, tức là khoảng 91.10% các dự đoán của mô hình là chính xác.
* Specificity: Specificity của Random Forest là 0.954, vẫn cho thấy khả năng phân loại tốt của mô hình cho các trường hợp không rời bỏ.
* Classification Report:
* Precision cho lớp 0 là 94%, recall là 95%, và F1-score là 0.94.
* Precision cho lớp 1 là 81%, recall là 74%, và F1-score là 0.77.
* ROC Curve (Area = 0.96): Random Forest có diện tích dưới đường cong ROC (AUC) là 0.96.

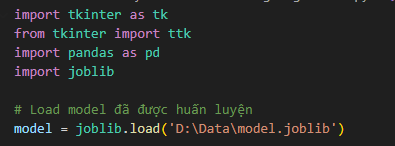
# Chương 4. Tích hợp mô hình vào hệ thống

Mô hình cần được tích hợp vào chức năng cụ thể, thỏa mãn yêu cầu của đề bài.

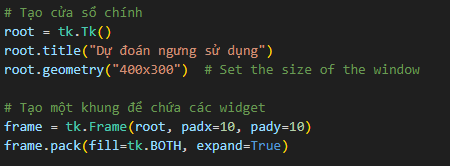
* Load mô hình đã được huấn luyện

○ Bước 1: Cài đặt thư viện joblib

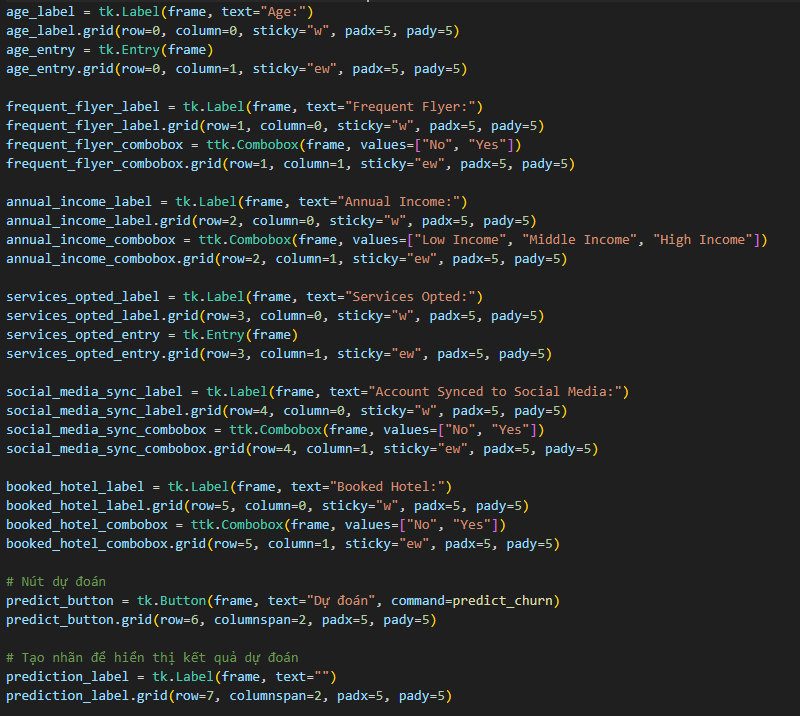
○ Bước 2: Tải mô hình: tải mô hình đã được lưu bằng joblib bằng cách sử dụng hàm load() từ thư viện này



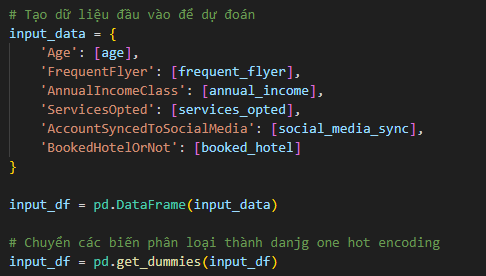
* Tạo giao diện:dụng thư viện Tkinter để tạo giao diện.



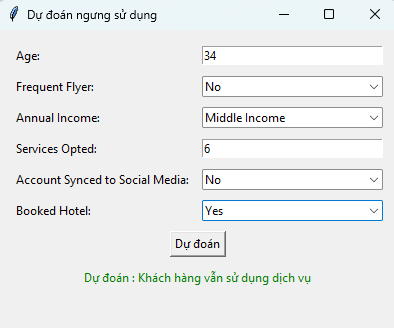
* Tạo nhãn, tạo ô nhập liệu, tạo nút dự đoán và kết quả dự đoán

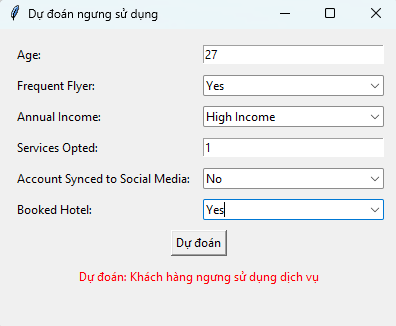


- Đưa dữ liệu vào mô hình huấn luyện



* Kết quả dự đoán





# Kết luận

Kết luận theo từng mục tiêu của đề tài :

* Mục tiêu 1: Phân tích dữ liệu về hành vi khách hàng để xác định các yếu tố ảnh hưởng đến khả năng ngưng sử dụng dịch vụ :
  + Đã đạt được. Bằng cách thu thập và phân tích dữ liệu hành vi khách hàng, ta đã có thể xác định các yếu tố quan trọng có thể ảnh hưởng đến quyết định của khách hàng về việc tiếp tục sử dụng hoặc ngưng sử dụng dịch vụ.
* Mục tiêu 2: Xác định thuật toán và mô hình huấn luyện phù hợp.
  + Đã đạt được. Dựa trên phân tích dữ liệu, ta đã có thể chọn ra các thuật toán và mô hình học máy phù hợp như hồi quy logistic, cây quyết định, hoặc mạng nơ-ron để dự đoán khả năng ngưng sử dụng dịch vụ của khách hàng.
* Mục tiêu 3: Xây dựng mô hình dự đoán và đưa ra dự báo có tính chính xác cao.
  + Đã đạt được. Bằng cách tiền xử lý dữ liệu, huấn luyện mô hình trên dữ liệu đã được làm sạch và đánh giá hiệu suất của mô hình trên tập dữ liệu kiểm tra, ta có thể đảm bảo rằng mô hình đưa ra dự báo với độ chính xác cao và có khả năng áp dụng trong thực tế.
* Mục tiêu 4: Tích hợp mô hình vào hệ thống.
  + Đã đạt được. Tích hợp mô hình vào hệ thống, triển khai mô hình đã huấn luyện vào hệ thống thực tế và tích hợp với giao diện người dùng để người dùng có thể tương tác với mô hình một cách dễ dàng.

# Tài liệu tham khảo:

<https://www.kaggle.com/datasets/tejashvi14/tour-travels-customer-churn-prediction/data>