ĐẠI HỌC HUẾ



**KHOA KỸ THUẬT VÀ CÔNG NGHỆ**

🙠🙟🕮🙝🙢

Icon

Description automatically generated

**BÁO CÁO ĐỒ ÁN**

**HỌC KÌ I, năm học 2023-2024**

**Học phần:**

**HỌC MÁY 1**

**Số phách**

**(**Do hội đồng chấm ghi thi)

***Thừa Thiên Huế, tháng 12 năm 2023.***

ĐẠI HỌC HUẾ



**KHOA KỸ THUẬT VÀ CÔNG NGHỆ**

🙠🙟🕮🙝🙢

Icon

Description automatically generated

**BÁO CÁO ĐỒ ÁN**

**HỌC KÌ I, năm học 2022-2023**

**Học phần:**

**HỌC MÁY 1**

**Giảng viên hướng dẫn:** T.s Hoàng Hữu Trung.

**Lớp:** Khoa học dữ liệu và Trí tuệ nhân tạo khóa 3.

**Sinh viên thực hiện:** Phạm Phước Bảo Tín\_22E1020021

Trần Tùng Dương\_22E1010001 .

(ký và ghi rõ họ tên)

**Số phách**

**(**Do hội đồng chấm ghi thi)

***Thừa Thiên Huế, tháng 12 năm 2023.***

# LỜI CẢM ƠN

Được trở thành sinh viên Khoa Kỹ Thuật và Công Nghệ - Đại học Huế em rất hạnh phúc và biết ơn. Hạnh phúc vì mình đã đạt được mục tiêu mong muốn và biết ơn sự cống hiến, chỉ bảo tận tình sâu sắc của quý thầy cô trong khoa đồng thời đã tạo điều kiện học tập lí tưởng cho chúng em. Để hoàn thành đồ án một cách chỉnh chu nhất có thể em xin gửi lời cảm ơn đến thầy giáo bộ môn - TS. Hoàng Hữu Trung đã hướng dẫn tận tình, chi tiết cho chúng em trong quá trình hoàn thành đồ án sinh viên chúng em. Hi vọng rằng thời gian sắp tới em sẽ luôn cố gắng, nổ lực hơn nữa trong học tập chuyên nghành của mình.

Trong quá trình hoàn thành đồ án mặc dù em đã chuẩn bị kĩ nhưng không thể tránh khỏi những sai sót, em mong nhận được sự góp ý từ quý thầy, cô. Lời cuối cùng em xin kính chúc quý thầy, cô thật nhiều sức khỏe để tiếp tục dẫn dắt chúng em và những thế hệ tiếp theo thành người.

# DANH MỤC HÌNH ẢNH

[Hình 1: Dữ liệu trước tiền xử lí 7](#_Toc155634264)

[Hình 2: Kết quả kiểm tra giá trị trùng lặp 8](#_Toc155634265)

[Hình 3: Kiểm tra giá trị thiếu 8](#_Toc155634266)

[Hình 4: Loại bỏ cột không cần thiết 9](#_Toc155634267)

[Hình 5: Thống kê cơ bản các giá trị trong cột 9](#_Toc155634268)

[Hình 6: Mã nguồn biểu đồ hộp kiểm tra giá trị ngoại lai 9](#_Toc155634269)

[Hình 7: Biểu đồ hộp kiểm tra giá trị nhiễu 10](#_Toc155634270)

[Hình 8: Chuyển đổi dữ liệu từ dạng phân loại sang dạng số 10](#_Toc155634271)

[Hình 9: Chuẩn hóa dữ liệu bằng loại Min-Max Scaling 11](#_Toc155634272)

[Hình 10: Kết quả tiền xử lí dữ liệu 11](#_Toc155634273)

[Hình 11: Mã nguồn biểu đồ cột 11](#_Toc155634274)

[Hình 12: Kết quả biểu đồ cột 12](#_Toc155634275)

[Hình 13: Biểu đồ cột về hệ số tương quan 12](#_Toc155634276)

[Hình 14: Kết quả biểu đồ cột hệ số tương quan 13](#_Toc155634277)

[Hình 15: Mã nguồn biểu đồ đường 13](#_Toc155634278)

[Hình 16: Mã nguồn biểu đồ tròn 14](#_Toc155634279)

[Hình 17: Kết quả biểu đồ tròn phân phối giới tính 14](#_Toc155634280)

[Hình 18: Mã nguồn biểu đồ tần suất 15](#_Toc155634281)

[Hình 19: Biểu đồ tần suất 15](#_Toc155634282)

[Hình 20: Mã nguồn biểu đồ mật độ 15](#_Toc155634283)

[Hình 21: Biểu đồ mật độ 16](#_Toc155634284)

[Hình 22: Biểu đồ nhiệt hệ số tương quan 17](#_Toc155634285)

[Hình 23: Đánh giá hiệu suất mô hình bằng Lazy Preditct 18](#_Toc155634286)

[Hình 24: Thời gian thực hiện từng mô hình 19](#_Toc155634287)

[Hình 25: Chọn các đặc trưng đưa vào mô hình 19](#_Toc155634288)

[Hình 26: Chia ngẫu nhiễn tập huấn luyện và tập kiểm tra 20](#_Toc155634289)

[Hình 27: Mã nguồn huấn luyện mô hình LGBM 20](#_Toc155634290)

[Hình 28: Kết quả sau khi huấn luyện mô hình LGBM 21](#_Toc155634291)

[Hình 29: Mã nguồn huấn luyện mô hình hồi quy Logistic 22](#_Toc155634292)

[Hình 30: Kết quả sau khi huấn luyện mô hình hồi quy Logistic 23](#_Toc155634293)

[Hình 31: Ranh giới quyết định mô hình Logistic 24](#_Toc155634294)

[Hình 32: Mã nguồn huấn luyện mô hình KNN 25](#_Toc155634295)

[Hình 33: Kết quả huấn luyện mô hình KNN 26](#_Toc155634296)

[Hình 34: Mã nguồn huấn luyện mô hình SVM 27](#_Toc155634297)

[Hình 35: Kết quả huấn luyện mô hình SVM 28](#_Toc155634298)

[Hình 36: Kiểm tra giá trị trùng lặp 32](#_Toc155634299)

[Hình 37: Kiểm tra giá trị thiếu 32](#_Toc155634300)

[Hình 38: Xử lí giá trị thiếu 32](#_Toc155634301)

[Hình 39: Loại bỏ cột không cần thiết 33](#_Toc155634302)

[Hình 40: Xử lí giá trị nhiễu 34](#_Toc155634303)

[Hình 41: Tích hợp dữ liệu 34](#_Toc155634304)

[Hình 42: Chuyển hóa dữ liệu 34](#_Toc155634305)

[Hình 43: Chuẩn hóa dữ liệu 35](#_Toc155634306)

[Hình 44: Kết quả tiền xử lí dữ liệu 35](#_Toc155634307)

[Hình 45: Biểu đồ cột tình trạng học vấn 36](#_Toc155634308)

[Hình 46: Biểu đồ đường dao động ngày cuối cùng mua với thu nhập 37](#_Toc155634309)

[Hình 47: Biểu đồ tròn phân phối số lượng trẻ trong gia đình 38](#_Toc155634310)

[Hình 48: Biểu đồ phân tán tiền mua rượu và tiền mua trái cây 38](#_Toc155634311)

[Hình 49: Biểu đồ nhiệt 39](#_Toc155634312)

[Hình 50: Triển khai thuật toán PCA 40](#_Toc155634313)

[Hình 51: Mã nguồn biểu đồ điểm sau khi giảm chiều 40](#_Toc155634314)

[Hình 52: Biểu đồ phân tán sau khi áp dụng PCA 41](#_Toc155634315)

[Hình 53: Chuẩn bị dữ liệu cho K-Means 42](#_Toc155634316)

[Hình 54: Mã nguồn tìm hệ số k phù hợp cho model K-Means 42](#_Toc155634317)

[Hình 55: Kết quả tìm hệ số K cho mô hình K-means 43](#_Toc155634318)

[Hình 56: Mã nguồn huấn luyện mô hình K-Means 43](#_Toc155634319)

[Hình 57: Kết quả phân cụm bằng K-means 44](#_Toc155634320)

[Hình 58: Mã nguồn trực quan hóa phân cụm sau khi dùng K-means 44](#_Toc155634321)

[Hình 59: Biểu đồ phân cụm bằng K-means 45](#_Toc155634322)

[Hình 60: Khởi tạo mô hình phân cụm bằng DBSCAN 46](#_Toc155634323)

[Hình 61: Kết quả phân cụm bằng DBSCAN 46](#_Toc155634324)

[Hình 62: Mã nguồn trực quan hóa phân cụm bằng DBScan 47](#_Toc155634325)

[Hình 63: Kết quả trực quan hóa phân cụm bằng DBScan 47](#_Toc155634326)

[Hình 64: Mã nguồn tìm hệ số K cho MiniBactch Kmeans 48](#_Toc155634327)

[Hình 65: Kết quả tìm hệ số K cho MinniBatch Kmeans 49](#_Toc155634328)

[Hình 66: Mã nguồn mô hình MiniBatchKmeans 49](#_Toc155634329)

[Hình 67: Kết quả phân cụm bằng MiniBatch Kmeans 50](#_Toc155634330)

[Hình 68: Kết quả trực quan hóa bằng MiniBatchKmeans 51](#_Toc155634331)

[Hình 69: Mã nguồn và kết quả kiểm tra của các mô hình 52](#_Toc155634332)

[Hình 70: Mã nguồn trực quan phân cụm dựa trên tổng thu và chi 53](#_Toc155634333)

[Hình 71: Trực quan phân cụm dựa vào 2 tổng thu và tổng chi 53](#_Toc155634334)

[Hình 72: Biểu đồ hộp về số lượng trẻ em của mỗi cụm 54](#_Toc155634335)

# MỤC LỤC

[LỜI CẢM ƠN i](#_Toc155639379)

[DANH MỤC HÌNH ẢNH ii](#_Toc155639380)

[MỤC LỤC iv](#_Toc155639381)

[DANH MỤC BẢNG BIỂU vi](#_Toc155639382)

[PHẦN 1: HỌC CÓ GIÁM SÁT (SUPERVISED LEARNING) 7](#_Toc155639383)

[1.1 Giới thiệu dữ liệu và mục đích bài toán 7](#_Toc155639384)

[1.1.1 Giới thiệu dữ liệu 7](#_Toc155639385)

[1.1.2 Mục đích bài toán 8](#_Toc155639386)

[1.2 Tiền xử lí dữ liệu 8](#_Toc155639387)

[1.2.1 Làm sạch dữ liệu 8](#_Toc155639388)

[1.2.2 Chuyển đổi dữ liệu 10](#_Toc155639389)

[1.3 Trực quan hóa dữ liệu trước phân tích dữ liệu 11](#_Toc155639390)

[1.2.1 Biểu đồ cột 11](#_Toc155639391)

[1.2.2 Biểu đồ đường 13](#_Toc155639392)

[1.2.3 Biểu đồ tròn 14](#_Toc155639393)

[1.2.4 Biểu đồ phân tán 14](#_Toc155639394)

[1.2.5 Biểu đồ phân phối tần suất 14](#_Toc155639395)

[1.2.6 Biểu đồ mật độ 15](#_Toc155639396)

[1.2.7 Biểu đồ nhiệt 16](#_Toc155639397)

[1.4 Ứng dụng các mô hình học giám sát vào bài toán 17](#_Toc155639398)

[1.3.1 Mục đích bài toán 17](#_Toc155639399)

[1.3.2 Thư viện Lazy Predict 18](#_Toc155639400)

[1.3.2 Mô hình Light Gradient Boosting Machine (LGBM) 20](#_Toc155639401)

[1.3.3 Mô hình hồi quy Logistic 21](#_Toc155639402)

[1.3.4 Mô hình K-Nearest Neighbors 24](#_Toc155639403)

[1.3.5 Mô hình Supoort Vecto Machine 26](#_Toc155639404)

[1.5 Thảo luận, phân tích, đánh giá và kết luận về kết quả nhận được sau khi tích dữ liệu 29](#_Toc155639405)

[PHẦN 2: HỌC KHÔNG GIÁM SÁT (UNSUPERVISED LEARNING) 31](#_Toc155639406)

[2.1 Giới thiệu mục đích bài toán 31](#_Toc155639407)

[2.1.1 Giới thiệu dữ liệu 31](#_Toc155639408)

[2.2 Tiền xử lí dữ liệu 32](#_Toc155639409)

[2.2.1 Làm sạch dữ liệu 32](#_Toc155639410)

[2.2.2 Tích hợp dữ liệu 34](#_Toc155639411)

[2.2.3 Chuyển đổi dữ liệu 34](#_Toc155639412)

[2.3 Trực quan hóa dữ liệu 36](#_Toc155639413)

[2.3.1 Biểu đồ cột 36](#_Toc155639414)

[2.3.2 Biểu đồ đường 37](#_Toc155639415)

[2.3.3 Biểu đồ tròn 37](#_Toc155639416)

[2.3.4 Biểu đồ phân tán 38](#_Toc155639417)

[2.3.5 Biểu đồ nhiệt 39](#_Toc155639418)

[2.4 Ứng dụng các mô hình vào bài toán 39](#_Toc155639419)

[2.4.1 Thuật toán PCA (Principal component analysis) 39](#_Toc155639420)

[2.4.2 Mô hình Kmeans 41](#_Toc155639421)

[2.4.3 Mô hình DBSCAN 45](#_Toc155639422)

[2.4.4 Mô hình MiniBatchKMeans 48](#_Toc155639423)

[2.5 Thảo luận, phân tích, đánh giá và kết luận về kết quả nhận được sau khi tích dữ liệu 51](#_Toc155639424)

[TÀI LIỆU THAM KHẢO 55](#_Toc155639425)

# DANH MỤC BẢNG BIỂU

# PHẦN 1: HỌC CÓ GIÁM SÁT (SUPERVISED LEARNING)

Mã nguồn toàn bộ đồ án: [Mã nguồn Github](https://github.com/BAOTIN2004/Do_an_ML).

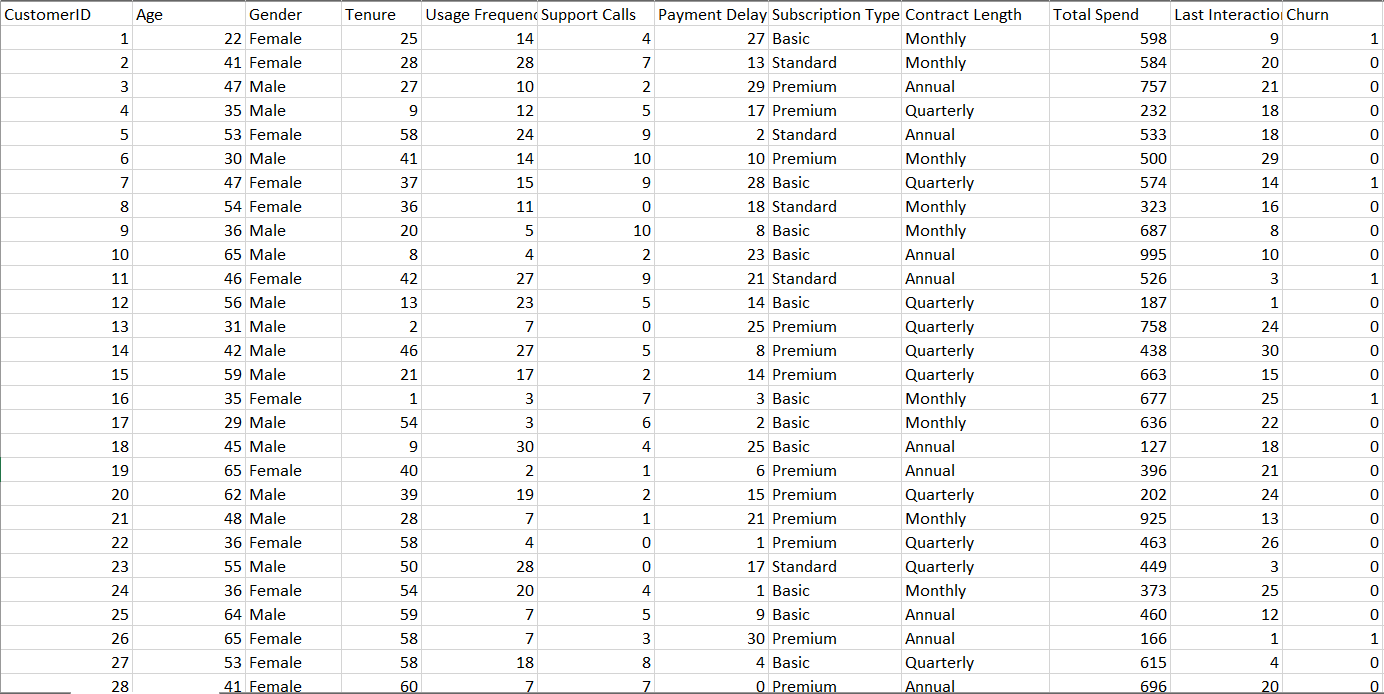
## Giới thiệu dữ liệu và mục đích bài toán

### Giới thiệu dữ liệu

Dữ liệu được dùng trong phần học có giám này là datasets về hiện tượng khách hàng chấm dứt hoặc tiếp tục đăng kí với nhà cung cấp dịch vụ. Hiểu được lí do dừng sử dụng dịch vụ của khách khách là yếu tố quan trọng giúp các nhà cung cấp dịch vụ phát triển các chiến lược để giữ chân khách hàng hiện tại và thu hút khách hàng tiềm năng.

Datasets về các yếu tố ảnh hưởng đến việc khách hàng tiếp tục hay ngừng sử dụng dịch vụ của nhà cung cấp:

* CustomerID: Mã khách hàng.
* Age: Tuổi khách hàng.
* Gender: Giới tính khách hàng.
* Tenure: Số lần kí hợp đồng.
* Usage Frequency: Số lần sử dụng.
* Support Calls: Số cuộc gọi hỗ trợ.
* Payment Delay: Số ngày trễ hạn thanh toán.
* Subscription Type: Loại đăng kí.
* Contract Lenght: Loại thanh toán (theo thàng, năm,..)
* Total Spend: Tổng chi tiêu trong quá trình sử dụng dịch vụ (usd)
* Last Interaction: Số ngày tương tác cuối cùng.
* Churn: Kết quả đã ngưng sử dụng dịch vụ hay chưa (0: tiếp tục,1: đã ngưng)



Hình 1: Dữ liệu trước tiền xử lí

### 1.1.2 Mục đích bài toán

## Tiền xử lí dữ liệu

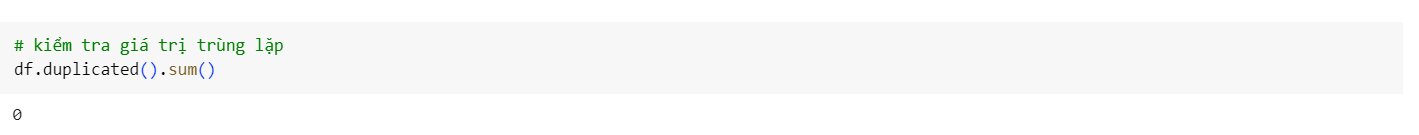
Dữ liệu ban đầu gồm loại dữ liệu số và dữ liệu phân loại không nhất quán, để phục vụ cho việc trực quan hóa dữ liệu và thực hiện bài toán phân loại việc tiền xử lí dữ liệu đóng vai trò rất quan trọng.

### 1.2.1 Làm sạch dữ liệu

Công việc làm sạch dữ liệu là một phần của tiền xử lí dữ liệu, gồm những việc dưới đây:

1. Kiểm tra giá trị trùng lặp

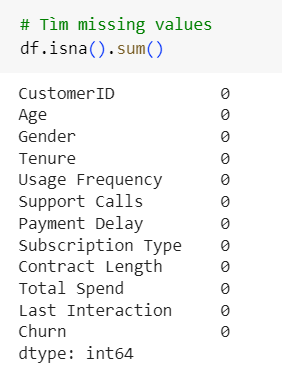
Sử dụng phương thức duplicated trong pandas để kiểm tra số hàng trùng lặp trong datasets, kết quả thu được như Hình 2.



Hình 2: Kết quả kiểm tra giá trị trùng lặp

1. Xử lí giá trị thiếu

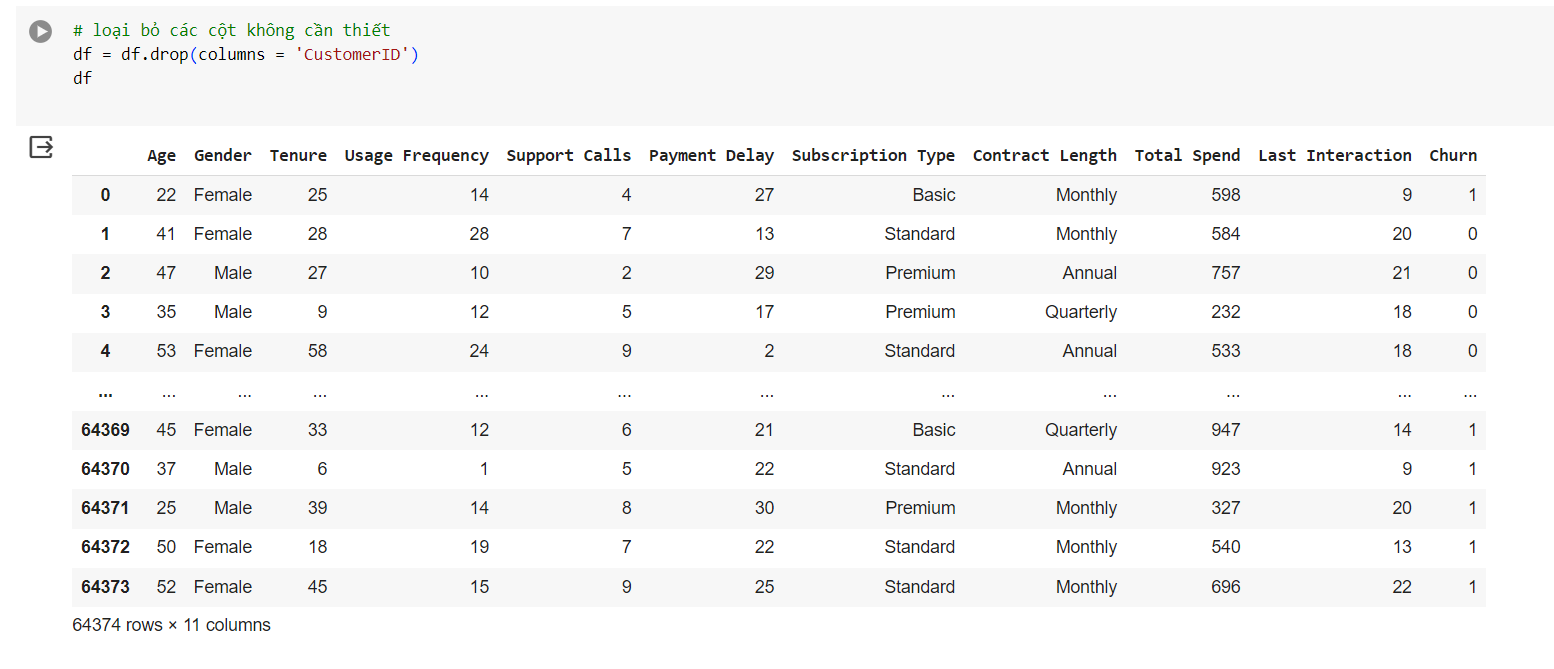
Sử dụng phương thức isna để kiểm tra số giá trị thiếu trong mỗi cột có trong datasets, kết quả thu được như Hình 3.



Hình 3: Kiểm tra giá trị thiếu

1. Loại bỏ cột không cần thiết

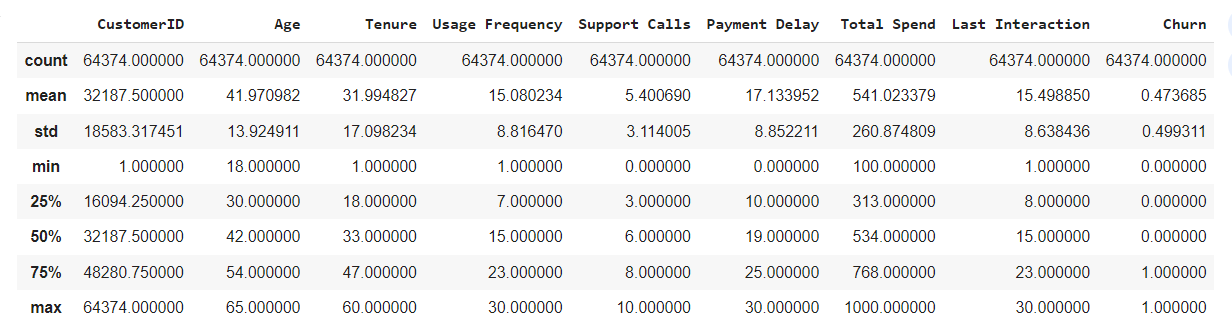
Cột mã số khách hàng không có vai trò trong mô hình chúng tôi sẽ sử dụng vì thế có thể loại bỏ cột này để giảm tải dữ liệu, thu kết quả như Hình 4.



Hình 4: Loại bỏ cột không cần thiết

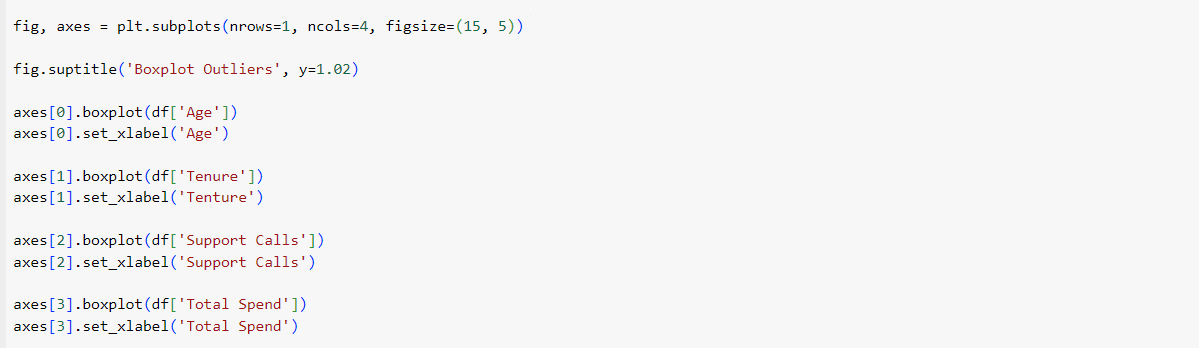
1. Xử lí giá trị nhiễu

Thống kê cơ bản các cột giá trị trong datasets như Hình 5.

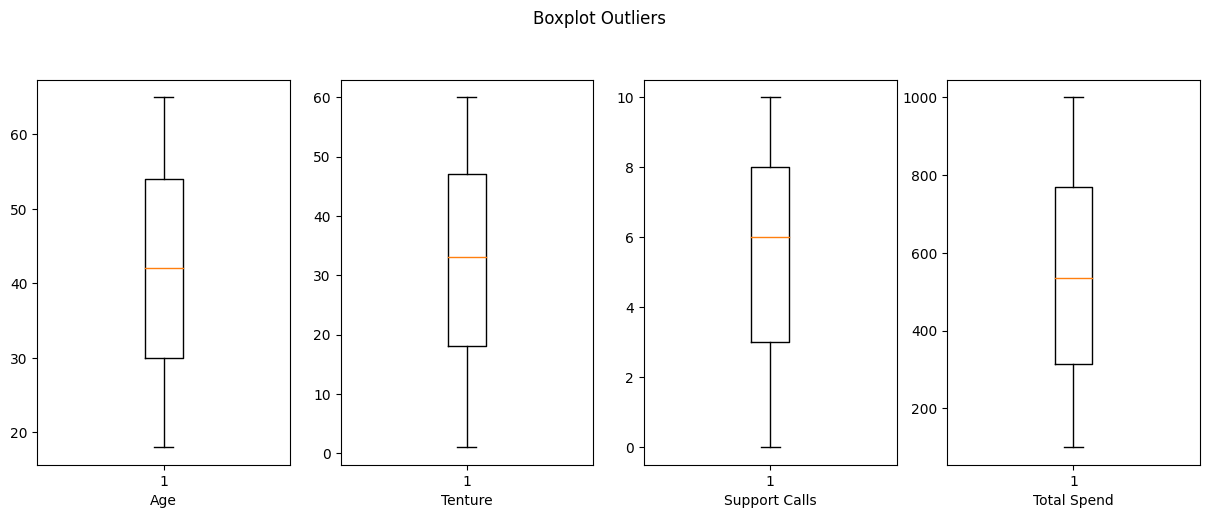


Hình 5: Thống kê cơ bản các giá trị trong cột

Kiểm tra giá trị ngoại lai bằng biểu đồ hộp



Hình 6: Mã nguồn biểu đồ hộp kiểm tra giá trị nhiễu

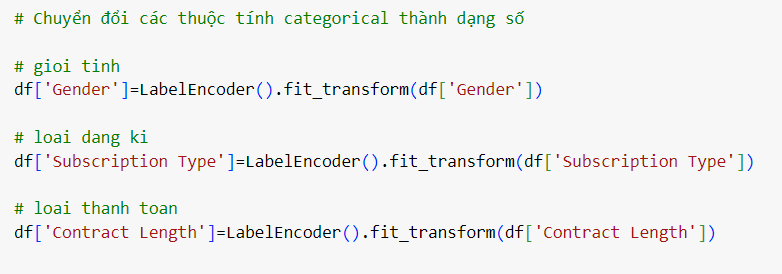


Hình 7: Biểu đồ hộp kiểm tra giá trị nhiễu

### 1.2.2 Chuyển đổi dữ liệu

1. Đánh giá thuộc tính độc lập ( số và phân loại):
2. Chuyển đổi các thuộc tính phân loại thành dạng số:

Từ kết quả của việc đánh giá thuộc tính độc lập ta thu được 3 cột cần chuyển đổi thuộc tính từ phân loại sang số như dưới đây.

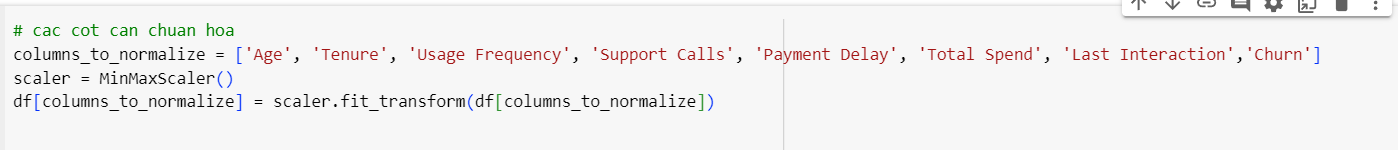


Hình 8: Chuyển đổi dữ liệu từ dạng phân loại sang dạng số

1. Chuẩn hóa dữ liệu

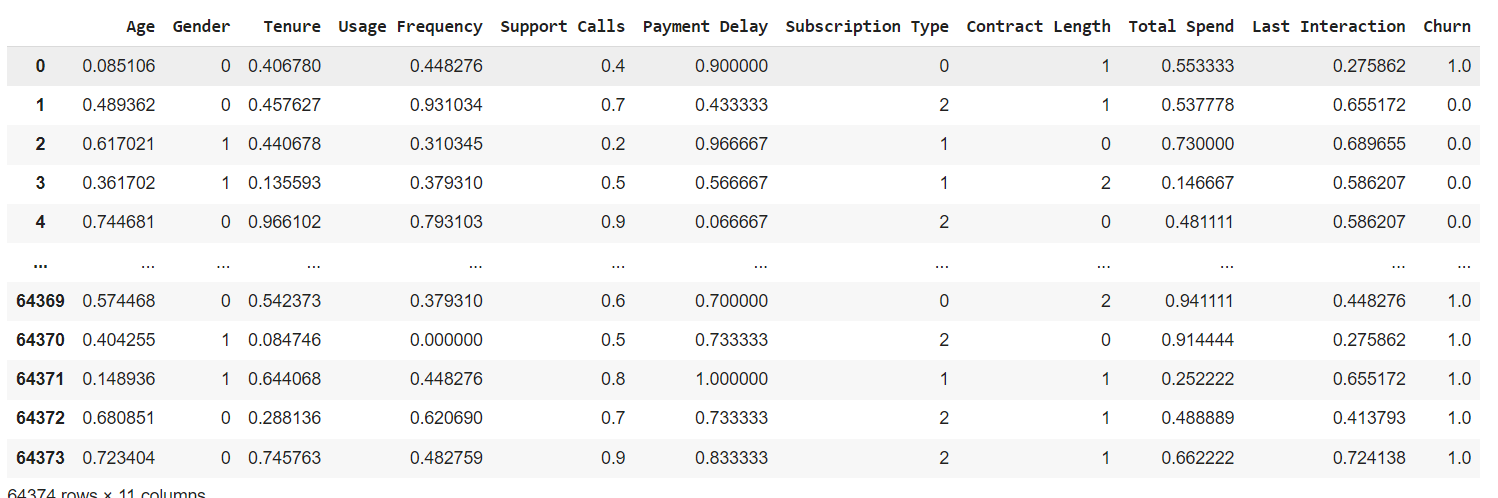
Chuẩn hóa dữ liệu là một bước quan trọng trong quá trình tiền xử lý dữ liệu, nhằm chuyển các đặc trưng định lượng (number) về cùng một thang đo chung, giúp cho việc biểu diễn dữ liệu dễ dàng và các mô hình phân tích nhất là các mô hình học máy (machine learning) hoạt động hiệu quả hơn.

Có hai loại chuẩn hóa dữ liệu thường dùng đó là Standardization (Z-Score Scaling) và Normalization (Min-Max Scaling). Lần này chúng tôi sử dụng phướng pháp Min-Max Scaling để chuẩn hóa dữ liệu.



Hình 9: Chuẩn hóa dữ liệu bằng loại Min-Max Scaling

Kết quả sau quá trình chuyển đổi và chuyển hóa dữ liệu đồng thời cũng là kết quả của quá trình tiền xử lí dữ liệu như Hình 10.



Hình 10: Kết quả tiền xử lí dữ liệu

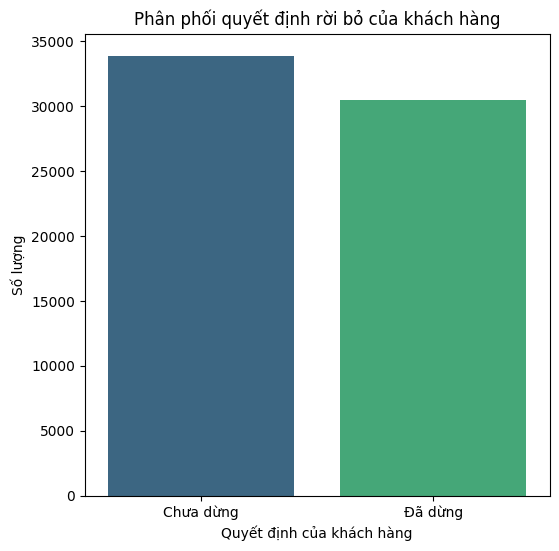
## 1.3 Trực quan hóa dữ liệu trước phân tích dữ liệu

### 1.2.1 Biểu đồ cột

Để cho nhà cung cấp nhìn ra được tổng quan trong việc khách hàng đã đưa ra quyết định có tiếp tục hay ngừng sử dụng dịch vụ của họ hay không như Hình 11.



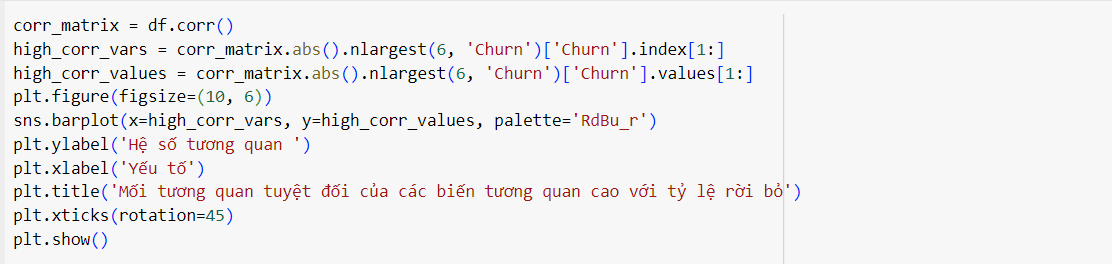
Hình 11: Mã nguồn biểu đồ cột



Hình 12: Kết quả biểu đồ cột

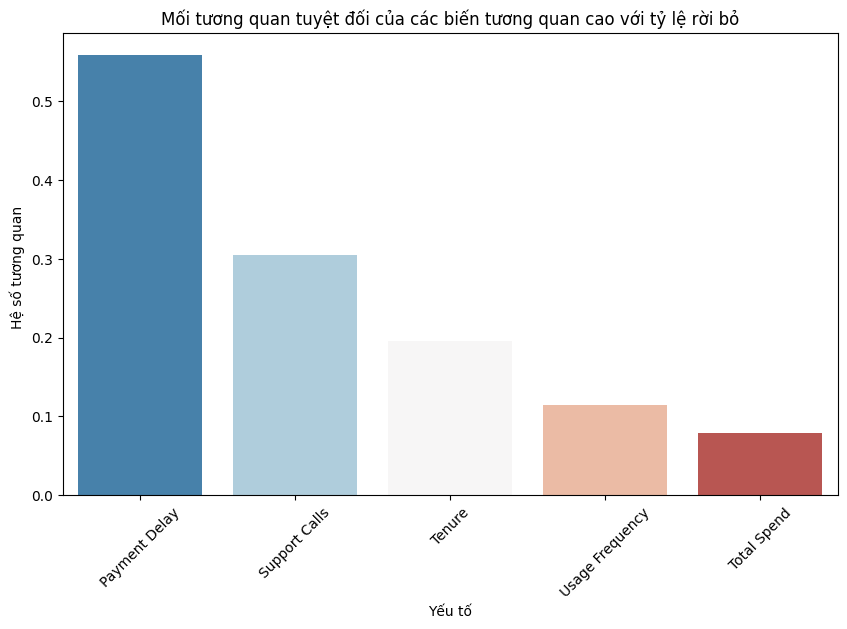
Kết quả như Hình 12 cho thấy việc số lượng khách hàng quyết định tiếp tục sử dụng nhiều hơn số khách hàng đã rời bỏ nhưng không chênh lệch đáng kể, đây cũng là một điều lo ngại khi số người hủy gần bằng người ở lại.

Trong phần này sử dụng biểu đồ cột để so sánh tương quan giữa các yếu tố với quyết định dừng sử dụng dịch vụ của khách hàng.



Hình 13: Biểu đồ cột về hệ số tương quan

Kết quả:

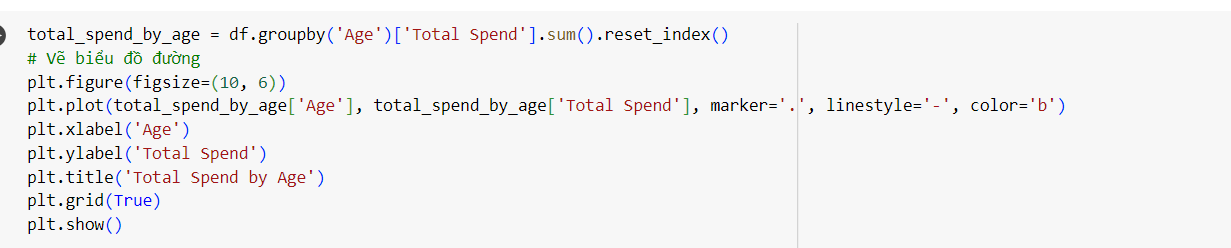


Hình 14: Kết quả biểu đồ cột hệ số tương quan

Nhận được kết quả như , chúng ta thấy được rằng yếu tố “Payment delay” (số ngày trễ hạn thanh toán) có môi tương quan khá cao với tỉ lệ khách hàng rời bỏ nhà cung cấp.

### 1.2.2 Biểu đồ đường

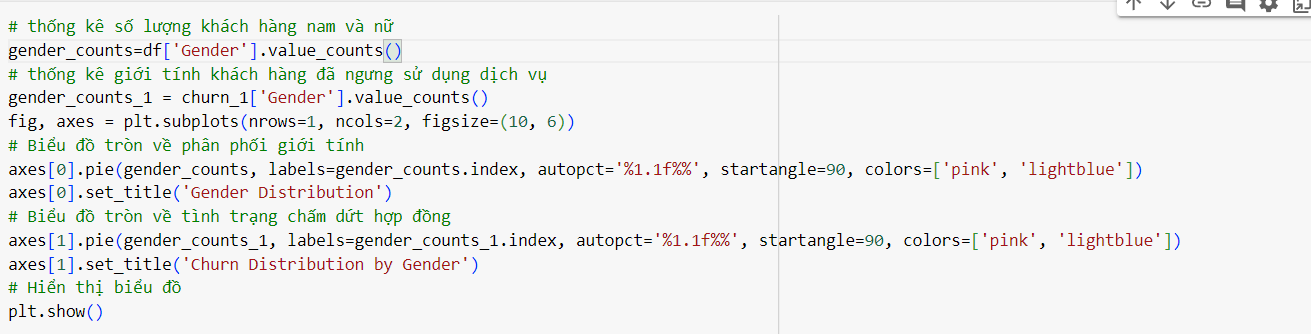
Nhà cung cấp cần nắm được nguồn thu của mình của các nhóm tuổi để có chiến lược phát triển kinh doanh.



Hình 15: Mã nguồn biểu đồ đường

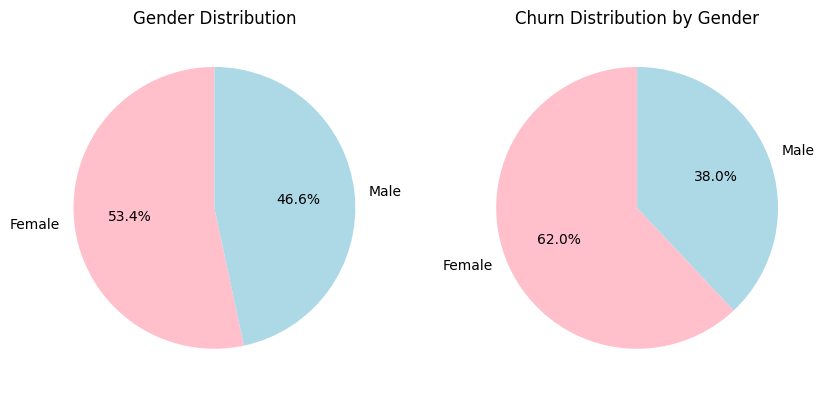
### 1.2.3 Biểu đồ tròn

Thống kê tỉ lệ giới tính phù hợp với việc sử dụng biểu đồ tròn, xem đối tượng nào là khách hàng chiếm trọng số, và đối tượng nào chiếm trọng số trong việc ngừng sử dụng dịch vụ doanh nghiệp.



Hình 16: Mã nguồn biểu đồ tròn

Kết quả:



Hình 17: Kết quả biểu đồ tròn phân phối giới tính

Từ Hình 17 nhận thấy rằng tỉ lệ nam và nữ trong tệp những khách hàng đã đăng kí chênh lệch không nhiều. Ngược lại trong tệp khách hàng quyết định ngừng sử dụng dịch vụ tỉ lệ nữ và nam lệch nhau lớn, đây cũng là một yếu tố quan trọng ảnh hưởng đến mô hình.

### 1.2.4 Biểu đồ phân tán

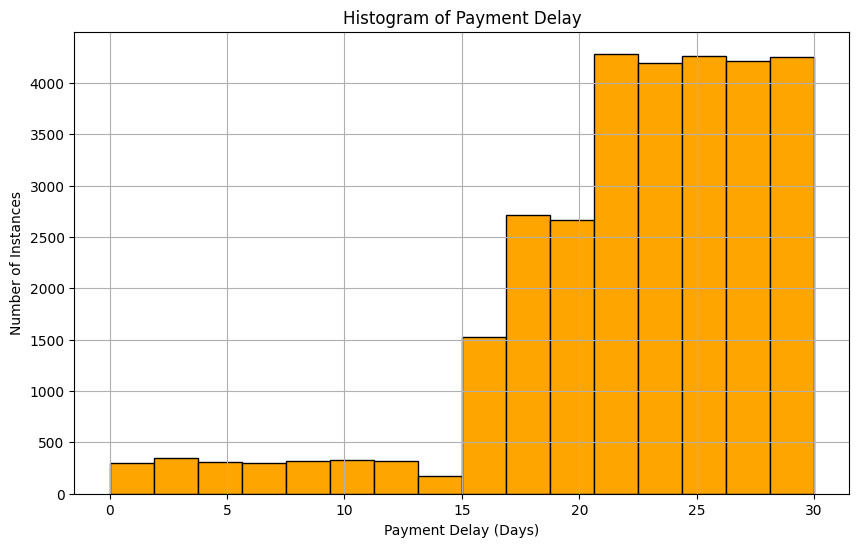
### 1.2.5 Biểu đồ phân phối tần suất

Nhận thấy kết quả hệ số tương quan giữa các yếu tố với quyết định rời bỏ của khách hàng từ Hình 14 thì yếu tố “Payment delay” có tương quan cao nhất. Biểu đồ phân phối tần suất dưới đây biểu thị số ngày trễ hạn dẫn đến việc không tiếp tục sử dụnh.



Hình 18: Mã nguồn biểu đồ tần suất

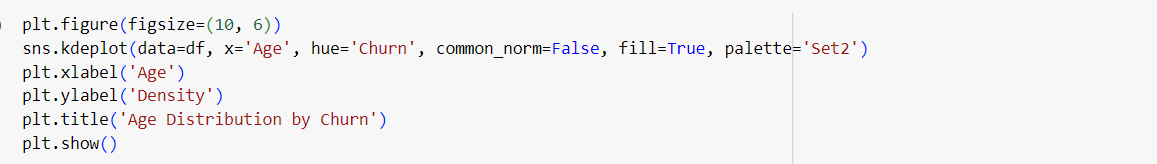
Kết quả:



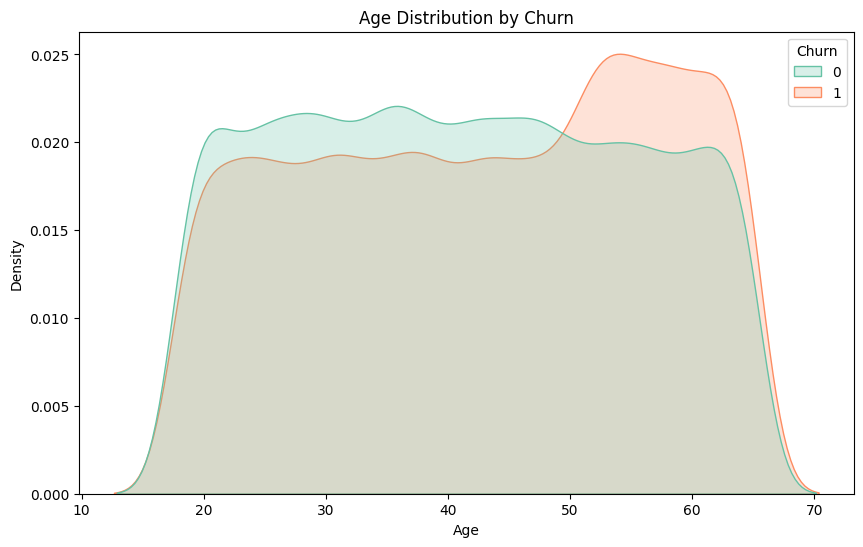
Hình 19: Biểu đồ tần suất

Nhóm người quyết định hủy dịch vụ đại đa số nằm trong khoảng thanh toán trễ lâu từ 20-30 ngày.

### 1.2.6 Biểu đồ mật độ



Hình 20: Mã nguồn biểu đồ mật độ

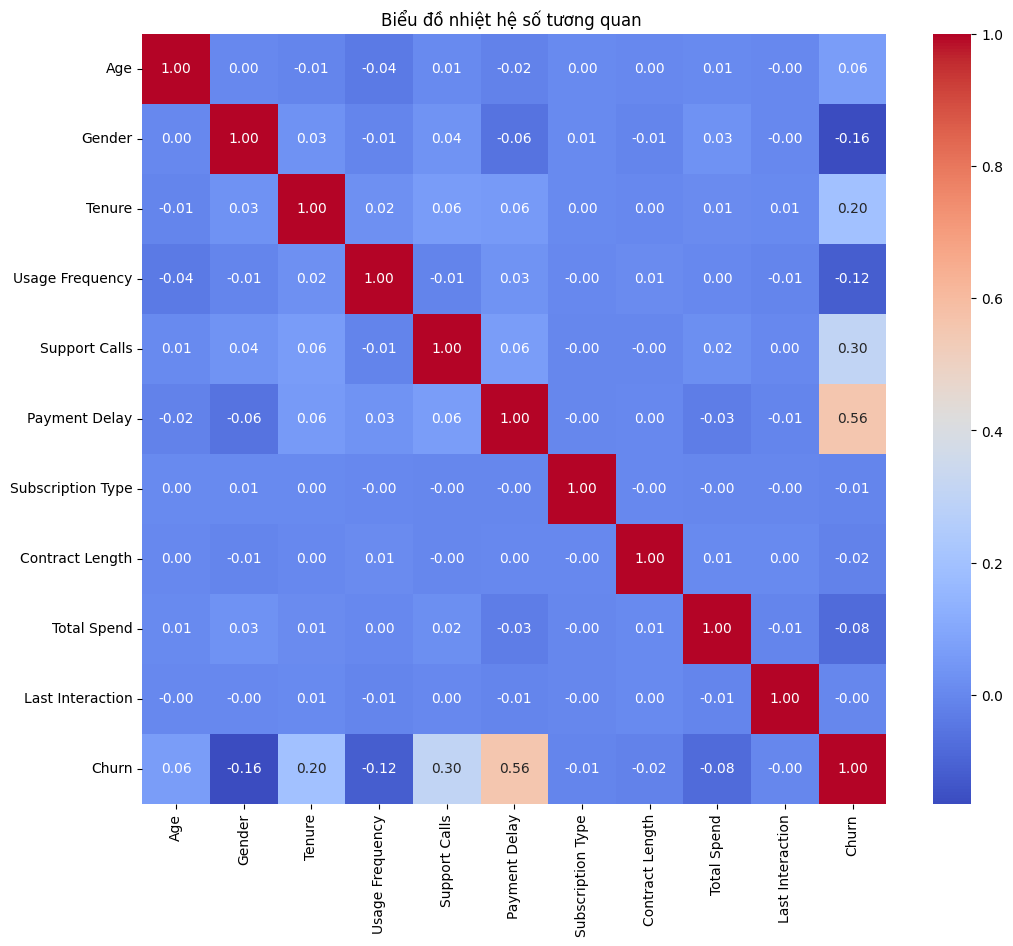


Hình 21: Biểu đồ mật độ

Các khoảng tuổi thuộc từ 20-50 đều có chênh lệch tỉ trọng ổn định, đồng đều qua các khoảng. Đặc biệt khoảng tuổi từ 55-65 độ chênh lệch tỉ trọng giữa việc tiếp tục và ngưng sử dụng lớn hơn nhiều so với các khoảng còn lại.

### 1.2.7 Biểu đồ nhiệt

Biểu đồ nhiệt (Heatmap) trích xuất các mối tương quan của các yếu tố đặc trưng với nhau, hoặc cột mục tiêu, từ đó có thể dựa vào biểu đồ để đưa ra các đặc vào tập huấn luyện và tập kiểm tra.



Hình 22: Biểu đồ nhiệt hệ số tương quan

Kết quả từ Hình 22, ta nhận thấy được các đặc trưng như “Payment Delay”, “Support calls”, “Tenure”, “Usage Frequency” có tương quan với “Churn”.

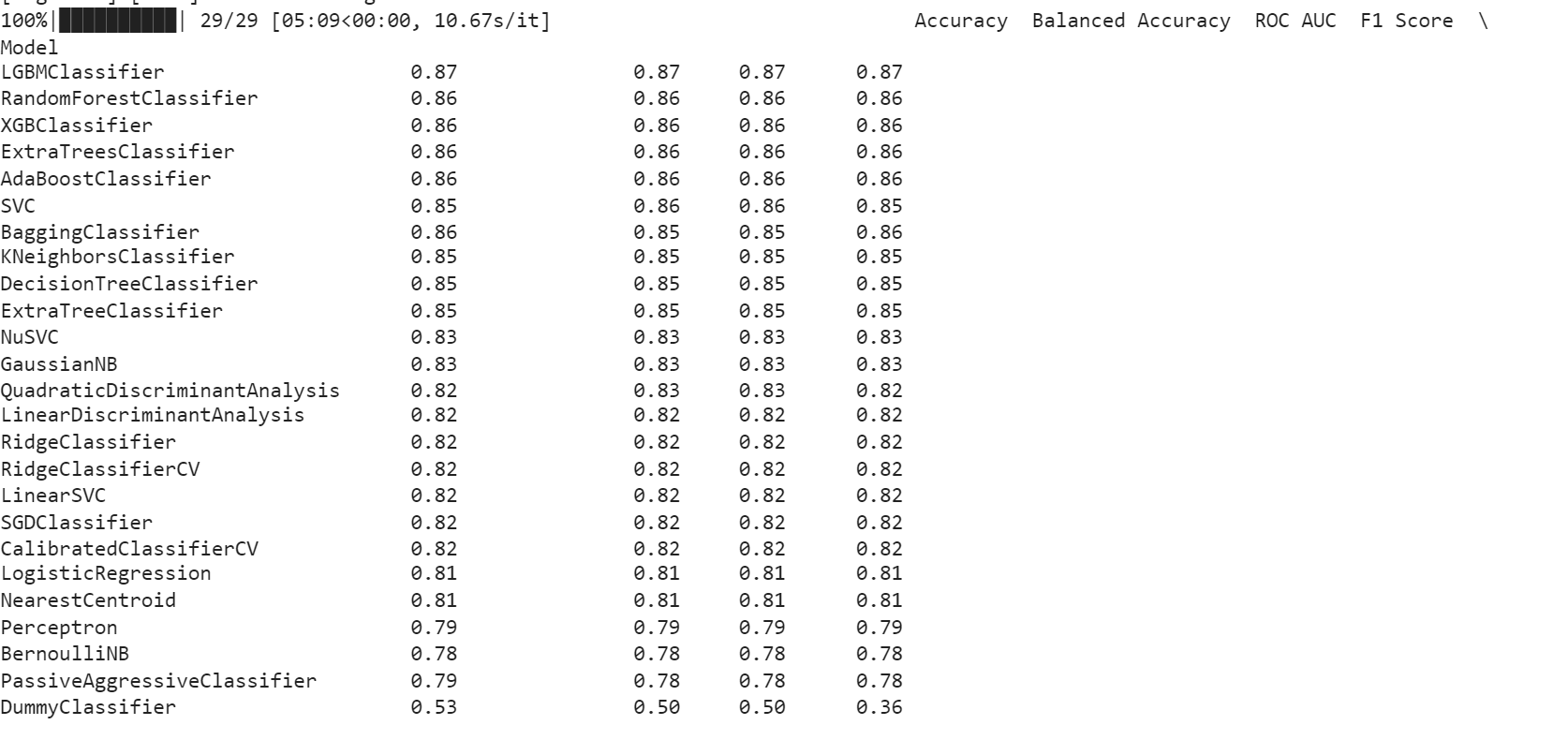
## 1.4 Ứng dụng các mô hình học giám sát vào bài toán

### 1.3.1 Mục đích bài toán

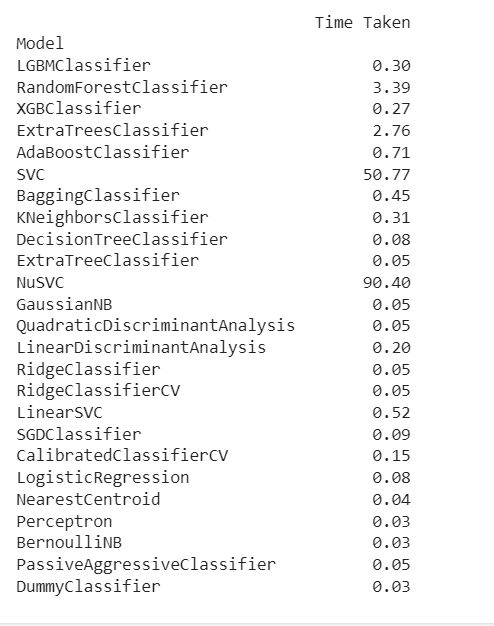
Trong bài toán, chúng em sử dụng mô hình thuật toán trong học có giám sát để ứng dụng dự đoán phân loại khách hàng quyết định ngừng sử dụng dịch vụ của nhà cung cấp hay không.

### 1.3.2 Thư viện Lazy Predict

Lazy Predict là một thư viện Python được sử dụng để tự động đánh giá hiệu suất của nhiều mô hình máy học khác nhau cho một tập dữ liệu cụ thể. Thư viện này tự động chọn và huấn luyện một loạt các mô hình khác nhau mà không cần bạn phải thủ công chỉ định từng mô hình cụ thể. Nó giúp tiết kiệm thời gian trong quá trình đánh giá và so sánh hiệu suất của các mô hình.



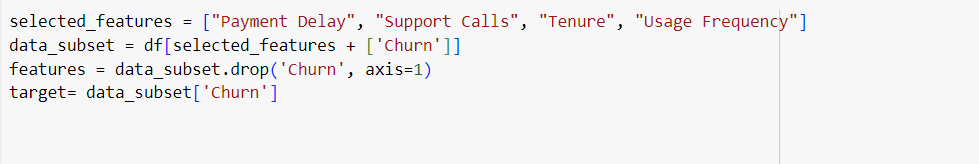
Hình 23: Đánh giá hiệu suất mô hình bằng Lazy Preditct



Hình 24: Thời gian thực hiện từng mô hình

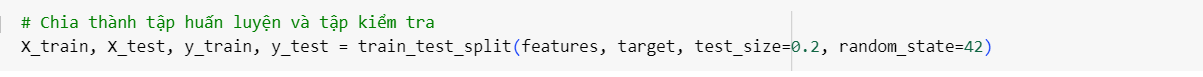
Từ kết quả nhận được 3 mô hình có hiệu suất tốt nhất để thực hiện thủ công từng mô hình: [ LGBMClassifier, RandomForestClassifier, XGBClassifier]

Chọn các đặc trưng có tương quan với mục tiêu để đưa vào mô hình.



Hình 25: Chọn các đặc trưng đưa vào mô hình

Dữ liệu được chia ngẫu nhiên thành tập huấn luyện chiếm 80% kích thước mẫu và tập kiểm thử chiếm 20% kích thước mẫu.



Hình 26: Chia ngẫu nhiễn tập huấn luyện và tập kiểm tra

### 1.3.2 Mô hình Light Gradient Boosting Machine (LGBM)

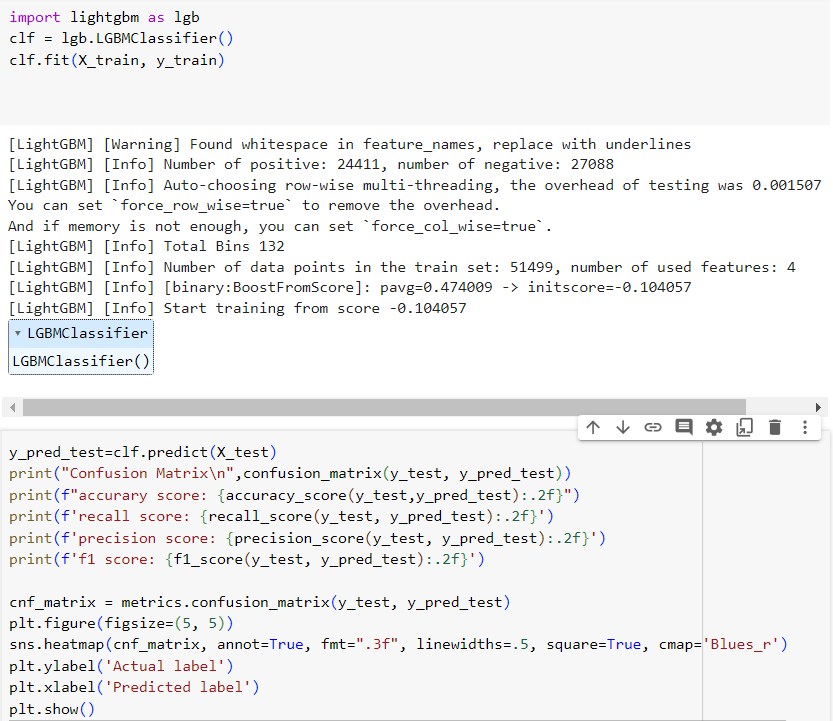
1. Giới thiệu mô hình LGBM

LGBM (Light Gradient Boosting Machine) là một framework gradient boosting sử dụng các thuật toán học cây quyết định. Nó được thiết kế để phân phối và hiệu quả với những ưu điểm sau:

* **Tốc độ đào tạo nhanh và hiệu quả cao**
* **Sử dụng bộ nhớ ít hơn**
* **Độ chính xác cao**
* **Hỗ trợ học song song và GPU,...**

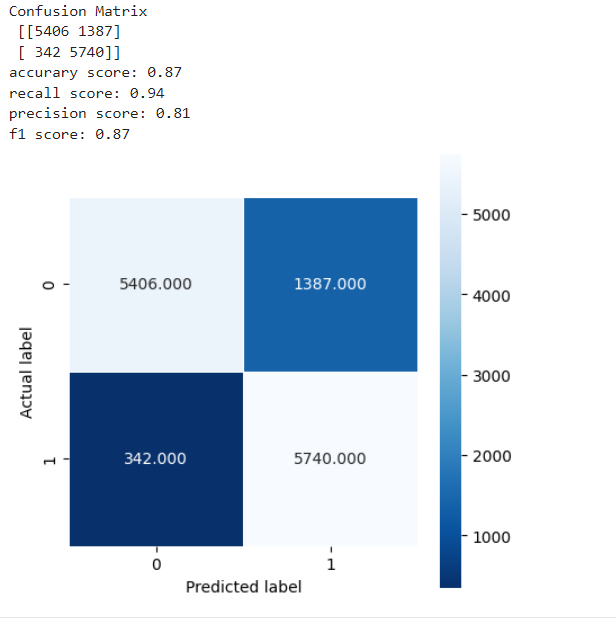
**2. Thực thi mô hình**

**Từ tập dữ liệu huấn luyện và kiểm thử đã được chia ở trên tiến hành huấn luyện và đưa ra kết quả mô hình LGBM như** Hình 27**.**



Hình 27: Mã nguồn huấn luyện mô hình LGBM

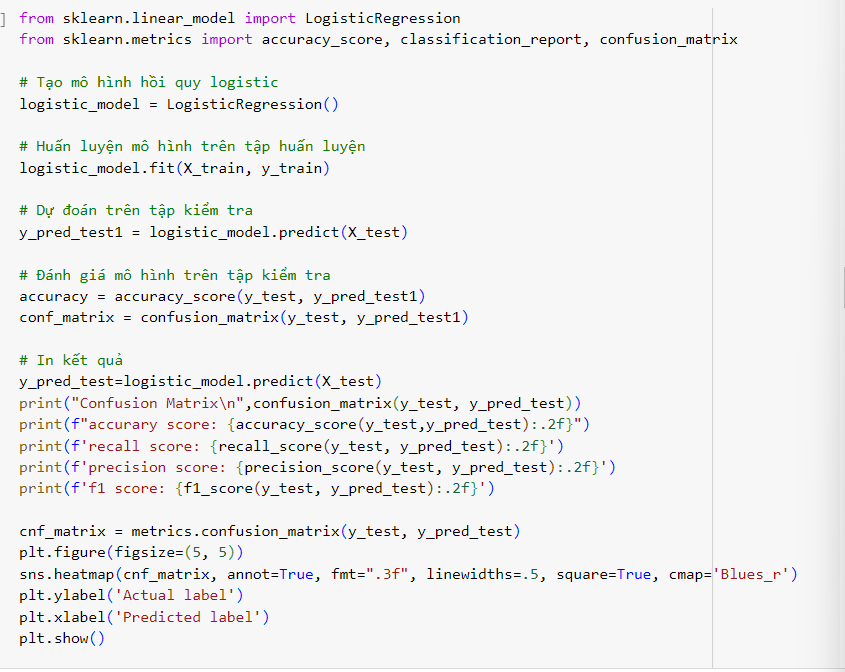
Kết quả về độ đo Accurary, Recall, precision, F1 và biểu đồ ma trận của mô hình LGBM như Hình 28.



Hình 28: Kết quả sau khi huấn luyện mô hình LGBM

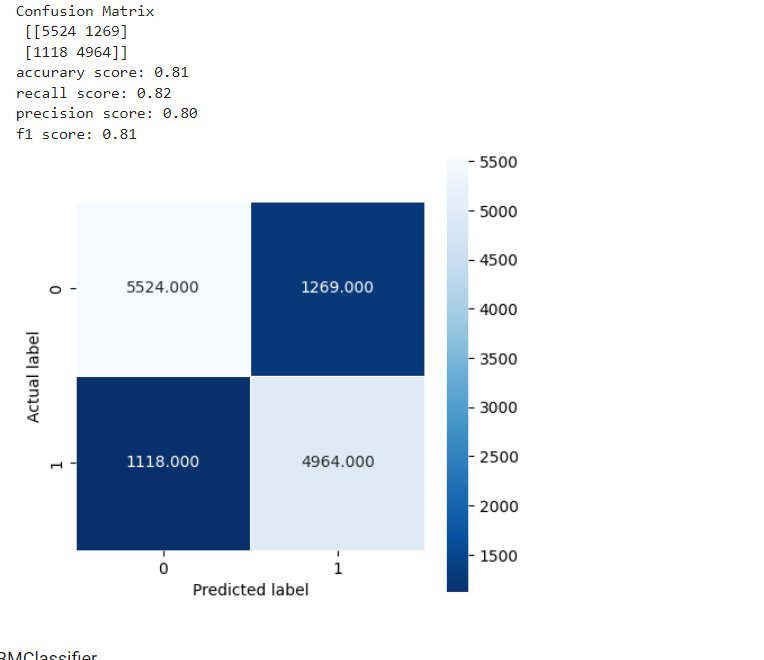
### 1.3.3 Mô hình hồi quy Logistic

Tạo mô hình và huấn luyện mô hình từ tập huấn luyện và kiểm thử đã tạo từ trước như Hình 29.

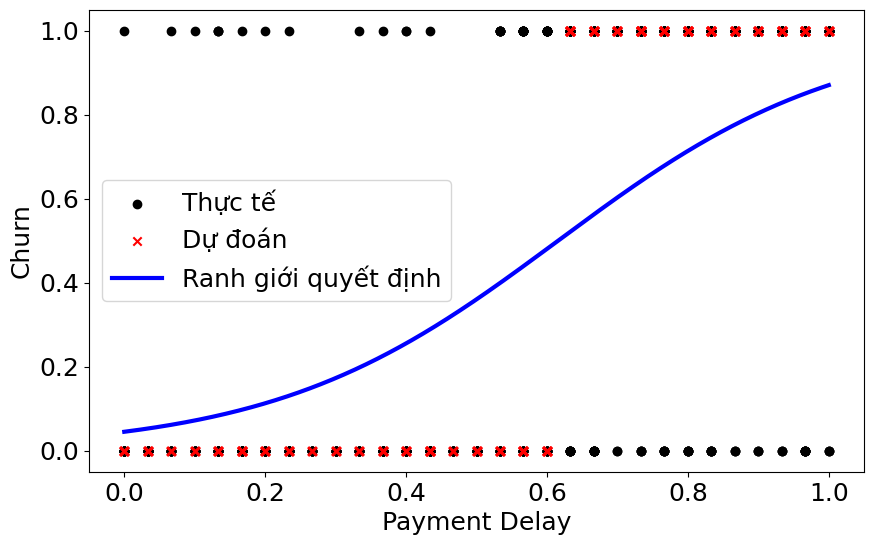


Hình 29: Mã nguồn huấn luyện mô hình hồi quy Logistic

Kết quả về độ đo Accurary, Recall, precision, F1 và biểu đồ ma trận của mô hình hồi quy Logistic như Hình 30.



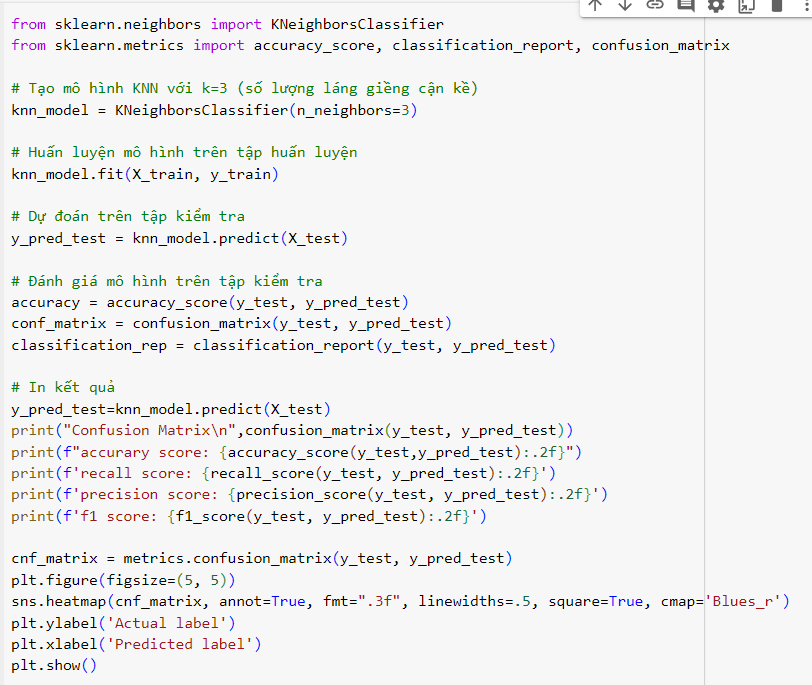
Hình 30: Kết quả sau khi huấn luyện mô hình hồi quy Logistic



Hình 31: Ranh giới quyết định mô hình Logistic

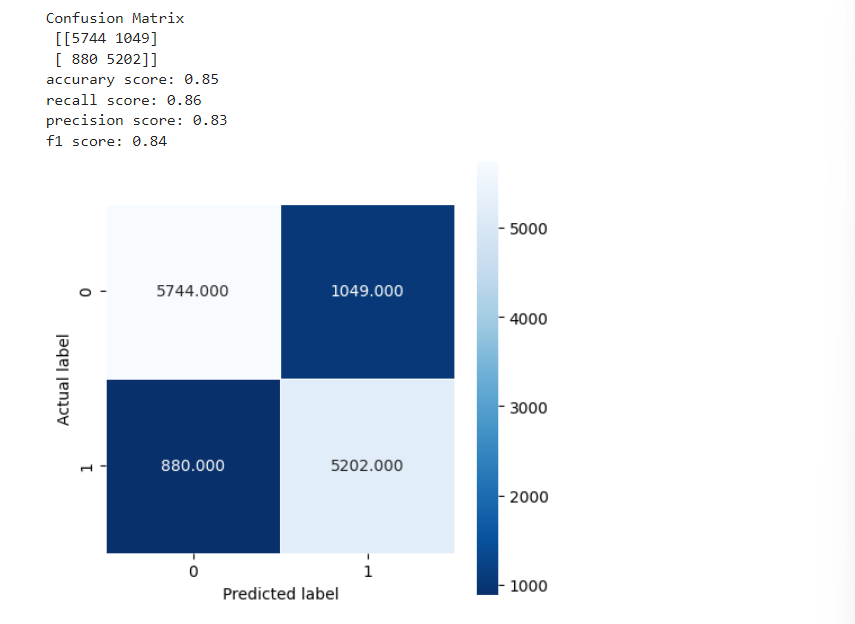
### 1.3.4 Mô hình K-Nearest Neighbors

Tạo mô hình và huấn luyện mô hình từ tập huấn luyện và kiểm thử đã tạo từ trước như **Error! Reference source not found.**.



Hình 32: Mã nguồn huấn luyện mô hình KNN

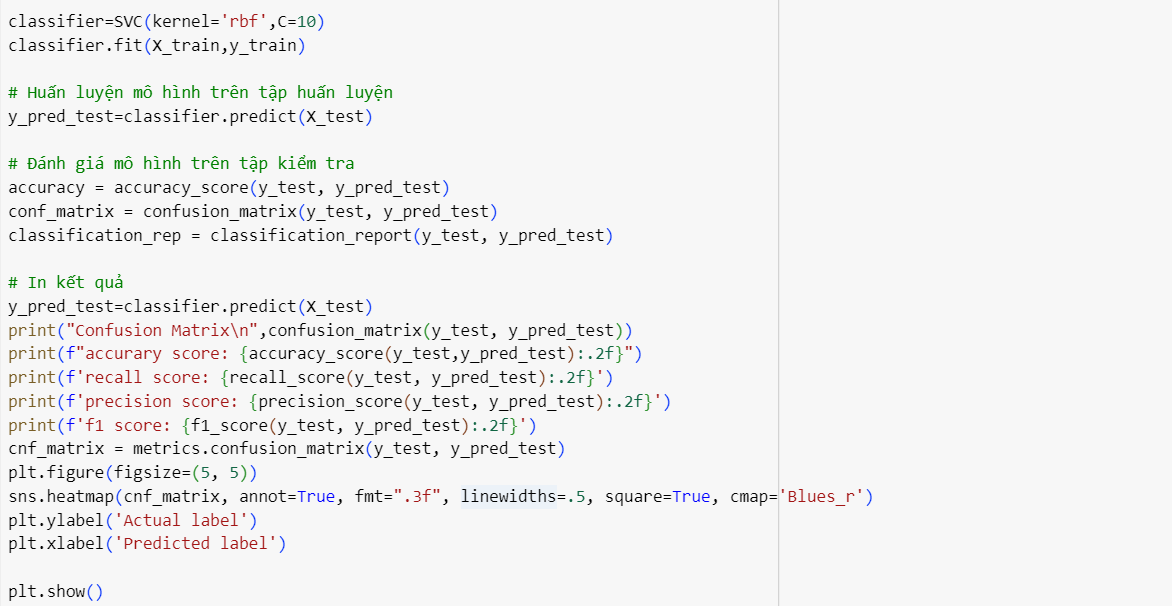
Kết quả về độ đo Accurary, Recall, precision, F1 và biểu đồ ma trận của mô hình KNN như Hình 33.



Hình 33: Kết quả huấn luyện mô hình KNN

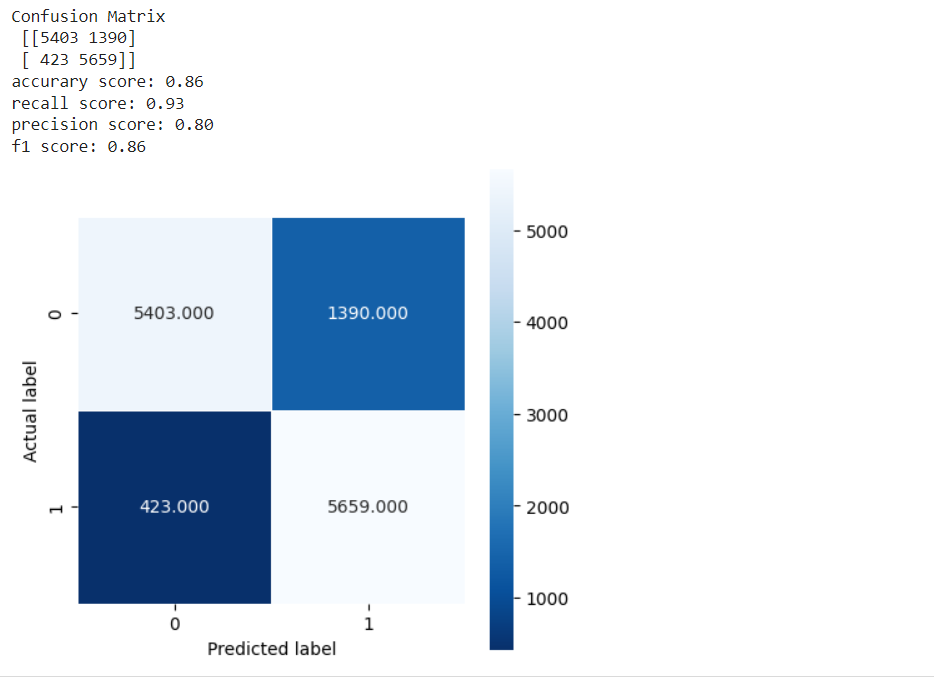
### 1.3.5 Mô hình Supoort Vecto Machine

Tạo mô hình và huấn luyện mô hình từ tập huấn luyện và kiểm thử đã tạo từ trước như Hình 34.

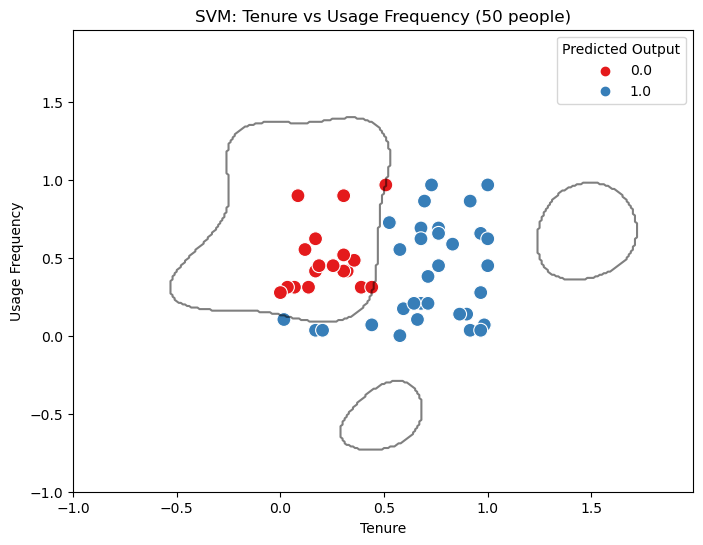


Hình 34: Mã nguồn huấn luyện mô hình SVM

Kết quả về độ đo Accurary, Recall, precision, F1 và biểu đồ ma trận của mô hình KNN như Hình 35.



Hình 35: Kết quả huấn luyện mô hình SVM



Hình 36: Trực quan phân loại mô hình SVM

## 1.5 Thảo luận, phân tích, đánh giá và kết luận về kết quả nhận được sau khi tích dữ liệu

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accurary | Recall Score | Precision | F1 Score | Elapsed time (s) |
| LGBM | 0.87 | 0.94 | 0.81 | 0.87 | 0.08 |
| SVM | 0.86 | 0.93 | 0.80 | 0.86 | 30.87 |
| KNN | 0.85 | 0.86 | 0.83 | 0.84 | 0.29 |
| Logistic | 0.81 | 0.82 | 0.80 | 0.81 | 0.003 |

1. Mô hình LGBM:

* Đạt hiệu suất tốt và độ nhớ lại (Recall) cao ( tỷ lệ bỏ sót các điểm thực sự là Positive là thấp).
* Precision khá cao (Khả năng dự đoán các điểm Positive thực sự là Positive là khá cao) .
* Thời gian thực hiện rất nhanh, chỉ mất 0.08 giây.

2. Mô hình SVM:

* SVM với kernel RBF cũng cho kết quả tốt với độ nhớ lại cao.
* Precision giảm so với LGBM, nhưng độ chính xác vẫn khá cao.
* Thời gian thực hiện lớn hơn nhiều so với LGBM.

3. Mô hình KNN:

* Kết quả độ chính xác và độ nhớ lại ổn định, không quá cao nhưng cũng không quá thấp.
* Precision và F1 Score đều khá tốt.
* Thời gian thực hiện cũng khá nhanh.

4. Mô hình hồi quy Logistic:

* Logistic Regression cung cấp kết quả khá tốt, nhưng thấp hơn so với các mô hình trước.
* Độ nhớ lại và độ chính xác đều ổn định, nhưng không cao bằng LGBM và SVM.
* Thời gian thực hiện rất nhanh, chỉ mất 0.003 giây.

Tổng kết:

* LGBM và SVM (kernel RBF) đều có hiệu suất tốt, với SVM có độ nhớ lại cao hơn nhưng cần thời gian tính toán lâu hơn.
* KNN cũng đạt được kết quả tốt, với sự cân bằng giữa độ chính xác và độ nhớ lại.
* Logistic cung cấp kết quả không bằng các mô hình khác nhưng thời gian thực hiện rất nhanh.

# 

# PHẦN 2: HỌC KHÔNG GIÁM SÁT (UNSUPERVISED LEARNING)

## 2.1 Giới thiệu mục đích bài toán

### 2.1.1 Giới thiệu dữ liệu

Tập dữ liệu sử dụng trong phần học không giám sảt này liên quan đến chi tiêu các mặt hàng. Dưới đây

\* ID: Định danh duy nhất của khách hàng

\* Year\_Birth: Năm sinh của khách hàng

\* Education: Trình độ học vấn của khách hàng

\* Marital\_Status: Tình trạng hôn nhân của khách hàng

\* Income: Thu nhập hàng năm của hộ gia đình khách hàng

\* Kidhome: Số lượng trẻ em trong hộ gia đình của khách hàng

\* Teenhome: Số lượng thanh thiếu niên trong hộ gia đình của khách hàng

\* Dt\_Customer: Ngày đăng ký của khách hàng với công ty

\* Recency: Số ngày kể từ lần mua cuối cùng của khách hàng

\* Complain: 1 nếu khách hàng phàn nàn trong vòng 2 năm qua, 0 nếu ngược lại

\* MntWines: Số tiền chi tiêu cho rượu trong 2 năm qua

\* MntFruits: Số tiền chi tiêu cho trái cây trong 2 năm qua

\* MntMeatProducts: Số tiền chi tiêu cho thịt trong 2 năm qua

\* MntFishProducts: Số tiền chi tiêu cho cá trong 2 năm qua

\* MntSweetProducts: Số tiền chi tiêu cho đồ ngọt trong 2 năm qua

\* MntGoldProds: Số tiền chi tiêu cho vàng trong 2 năm qua

\* NumDealsPurchases: Số lượng mua hàng với giảm giá

\* AcceptedCmp1: 1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch 1, 0 nếu ngược lại

\* AcceptedCmp2: 1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch 2, 0 nếu ngược lại

\* AcceptedCmp3: 1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch 3, 0 nếu ngược lại

\* AcceptedCmp4: 1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch 4, 0 nếu ngược lại

\* AcceptedCmp5: 1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch 5, 0 nếu ngược lại

\* Response: 1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch cuối cùng, 0 nếu ngược lại

\* NumWebPurchases: Số lượng mua hàng thông qua trang web của công ty

\* NumCatalogPurchases: Số lượng mua hàng sử dụng catalog

\* NumStorePurchases: Số lượng mua hàng trực tiếp tại cửa hàng

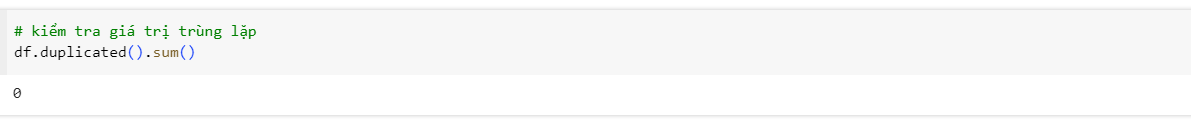
\* NumWebVisitsMonth: Số lượng lượt truy cập trang web của công ty trong tháng qua

## 2.2 Tiền xử lí dữ liệu

Dữ liệu ban đầu gồm loại dữ liệu số và dữ liệu phân loại không nhất quán, để phục vụ cho việc trực quan hóa dữ liệu và thực hiện bài toán phân cụm việc tiền xử lí dữ liệu đóng vai trò rất quan trọng.

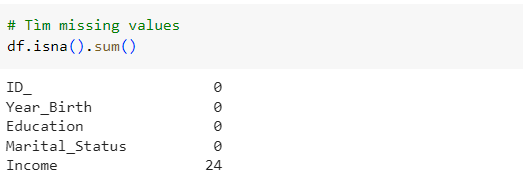
### 2.2.1 Làm sạch dữ liệu

1. Kiểm tra giá trị trùng lặp

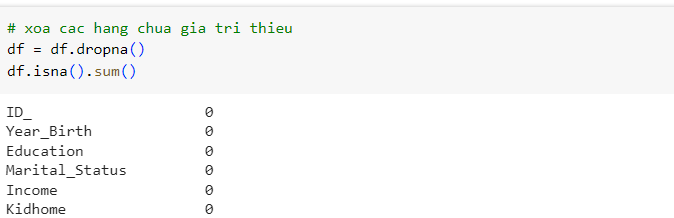


Hình 37: Kiểm tra giá trị trùng lặp

2. Xử lí giá trị thiếu



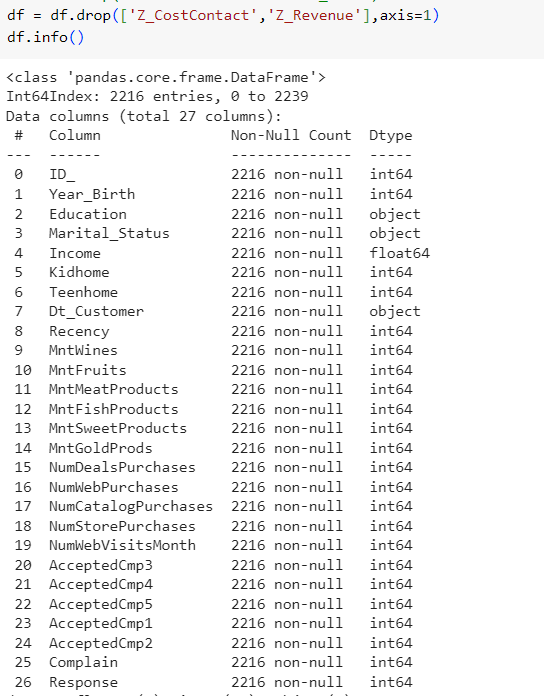
Hình 38: Kiểm tra giá trị thiếu



Hình 39: Xử lí giá trị thiếu

3. Loại bỏ cột không cần thiết

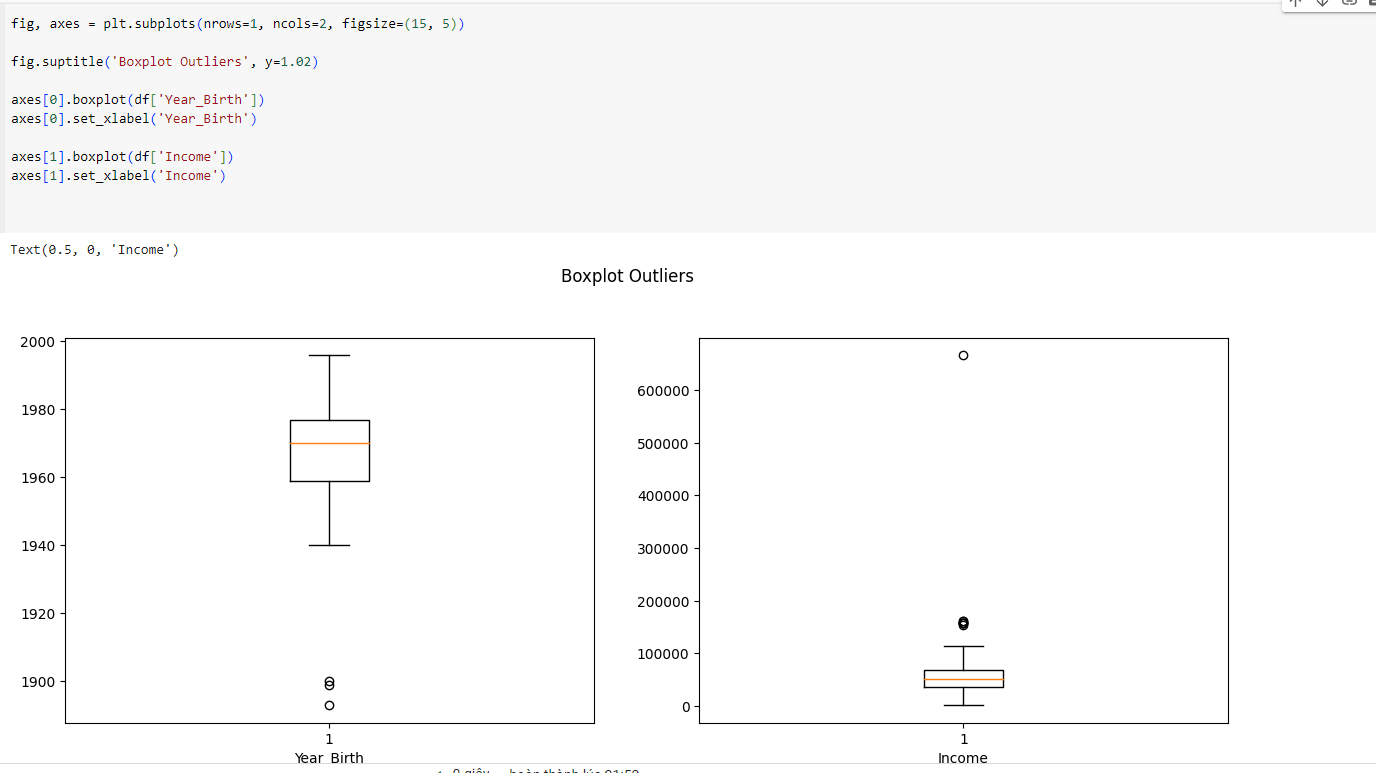
Vì cột “Z\_CostContact” và “Z\_Revenue” có 1 giá trị duy nhất, có thể loại bỏ hai cột này đi.



Hình 40: Loại bỏ cột không cần thiết

1. Xử lí giá trị nhiễu

Xử lí giá trị nhiễu bằng cách sử dụng biểu đồ hộp, từ đó có thể nhận ra giá trị ngoại lai xóa khỏi dữ liệu



Hình 41: Xử lí giá trị nhiễu

### 2.2.2 Tích hợp dữ liệu

Dựa vào tính năng các cột để có thể tích hợp các cột lại với nhau thành 1 cột:



Hình 42: Tích hợp dữ liệu

### 2.2.3 Chuyển đổi dữ liệu

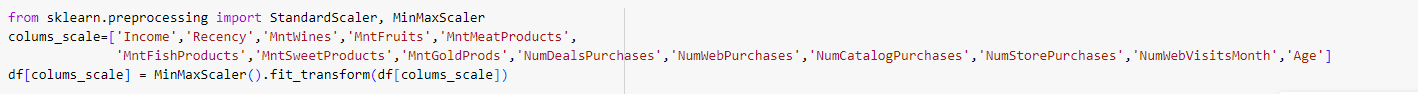
1. Chuyển hóa dữ liệu

Chuyển hóa dữ liệu dạng phân loại sang dạng số như dưới đây: 

Hình 43: Chuyển hóa dữ liệu

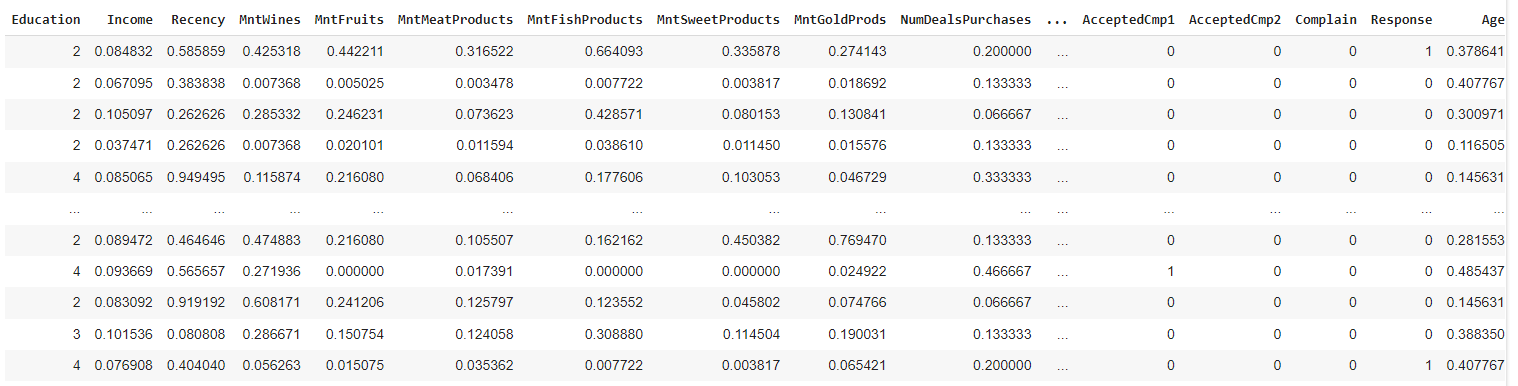
2. Chuẩn hóa dữ liệu

Chuẩn hóa dữ liệu là một bước quan trọng trong quá trình tiền xử lý dữ liệu, nhằm chuyển các đặc trưng định lượng (number) về cùng một thang đo chung, giúp cho việc biểu diễn dữ liệu dễ dàng và các mô hình phân tích nhất là các mô hình học máy (machine learning) hoạt động hiệu quả hơn.



Hình 44: Chuẩn hóa dữ liệu

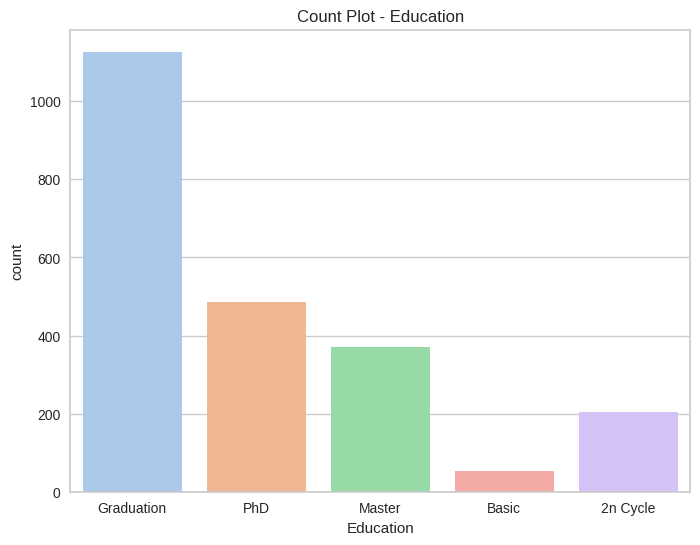
Sau khi kết thúc quá trình tiền xử lí dữ liệu, kết quả như dưới đây



Hình 45: Kết quả tiền xử lí dữ liệu

## 2.3 Trực quan hóa dữ liệu

### 2.3.1 Biểu đồ cột



Hình 46: Biểu đồ cột tình trạng học vấn

Từ kết quả biểu đồ cho thấy trình độ học vấn đạt bằng tốt nghiệp đại học nhiều hơn bất kì trình độ học vấn nào, điều này cũng dễ hiểu khi các công việc hiện tại có bằng cấp vẫn là yếu tố đáng chú ý.

### 2.3.2 Biểu đồ đường



Hình 47: Biểu đồ đường dao động ngày cuối cùng mua với thu nhập

Kết quả biểu đồ đường cho thấy giao động thu nhập của các người có số ngày kể từ lần mua cuối cùng là không đồng đều.

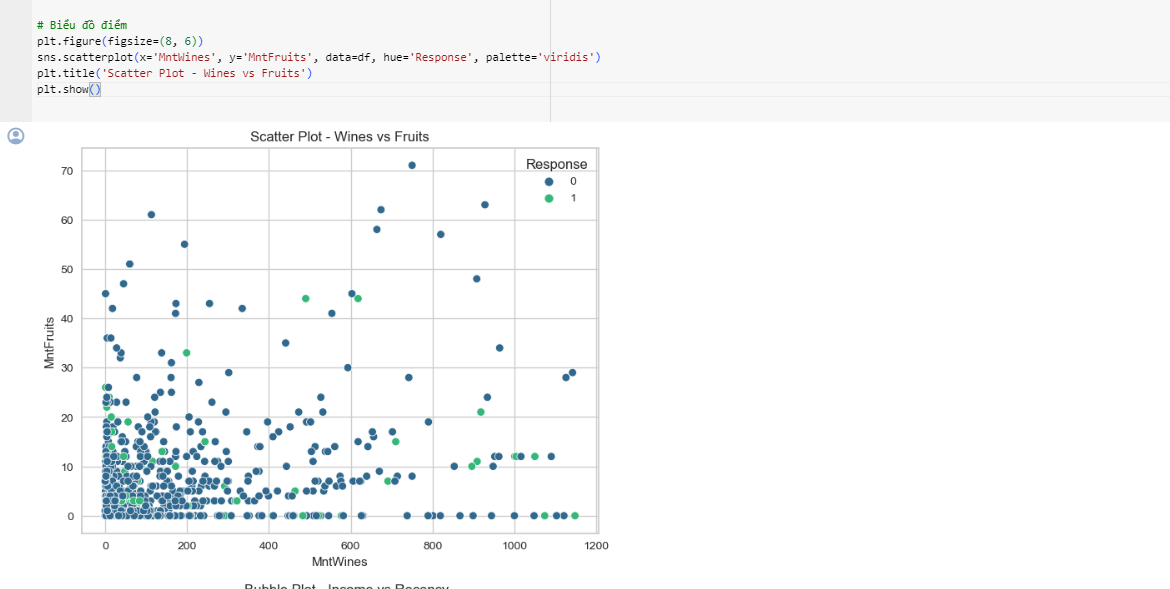
### 2.3.3 Biểu đồ tròn



Hình 48: Biểu đồ tròn phân phối số lượng trẻ trong gia đình

Từ biểu đồ tròn phân phối số lượng trẻ em có trong gia đình, cho thấy số lượng trẻ em không có trong gia đình chiếm gần 58%, trong khi 2 yếu tố còn lại chỉ chiếm hơn 42%. Từ đó cho thấy đây cũng là một yếu tố để phân cụm chi tiêu khách hàng.

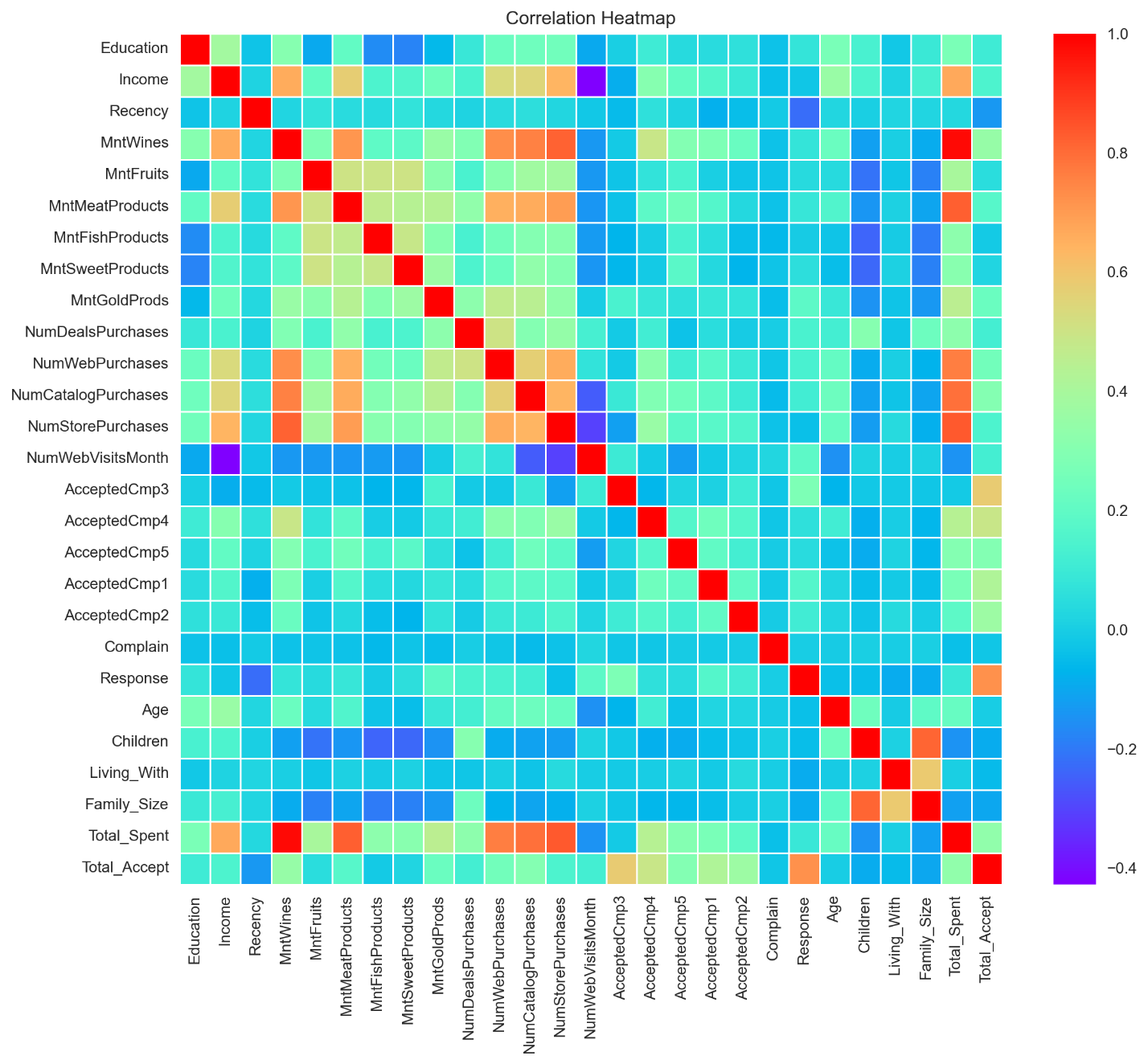
### 2.3.4 Biểu đồ phân tán



Hình 49: Biểu đồ phân tán tiền mua rượu và tiền mua trái cây

Biểu đồ phân tán tiền mua rượu và trái cái cây của mỗi người và được nhận ưu đãi cho chiến dịch lần cuối hay không.

### 2.3.5 Biểu đồ nhiệt



Hình 50: Biểu đồ nhiệt

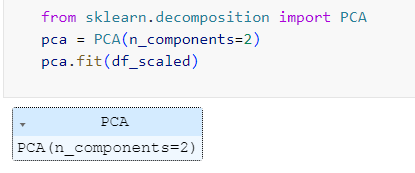
Biểu đồ nhiệt thể hiện tổng quát hệ số tương quan các cột dữ liệu trong datasets với nhau.

## 2.4 Ứng dụng các mô hình vào bài toán

### 2.4.1 Thuật toán PCA (Principal component analysis)

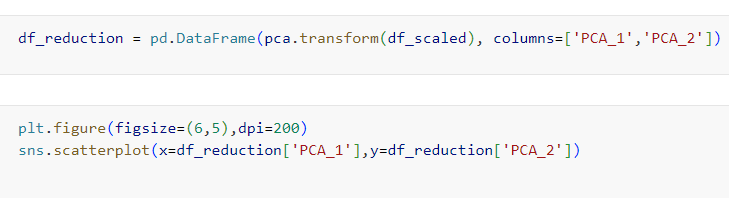
PCA là viết tắt của "Principal Component Analysis," một phương pháp trong thống kê và máy học được sử dụng để giảm số chiều của dữ liệu trong không gian đặc trưng. Mục tiêu của PCA là giảm sự phức tạp của dữ liệu bằng cách chuyển đổi các biến tương quan thành các biến không tương quan mới, gọi là các thành phần chính (principal components). Các thành phần chính được sắp xế theo độ giảm dần của phương sai, với ý nghĩa là thành phần đầu tiên giữ lại nhiều thông tin nhất về biến đổi của dữ liệu.

Ở bộ dữ liệu này, chúng tôi giảm chiều PCA về 2 chiều:

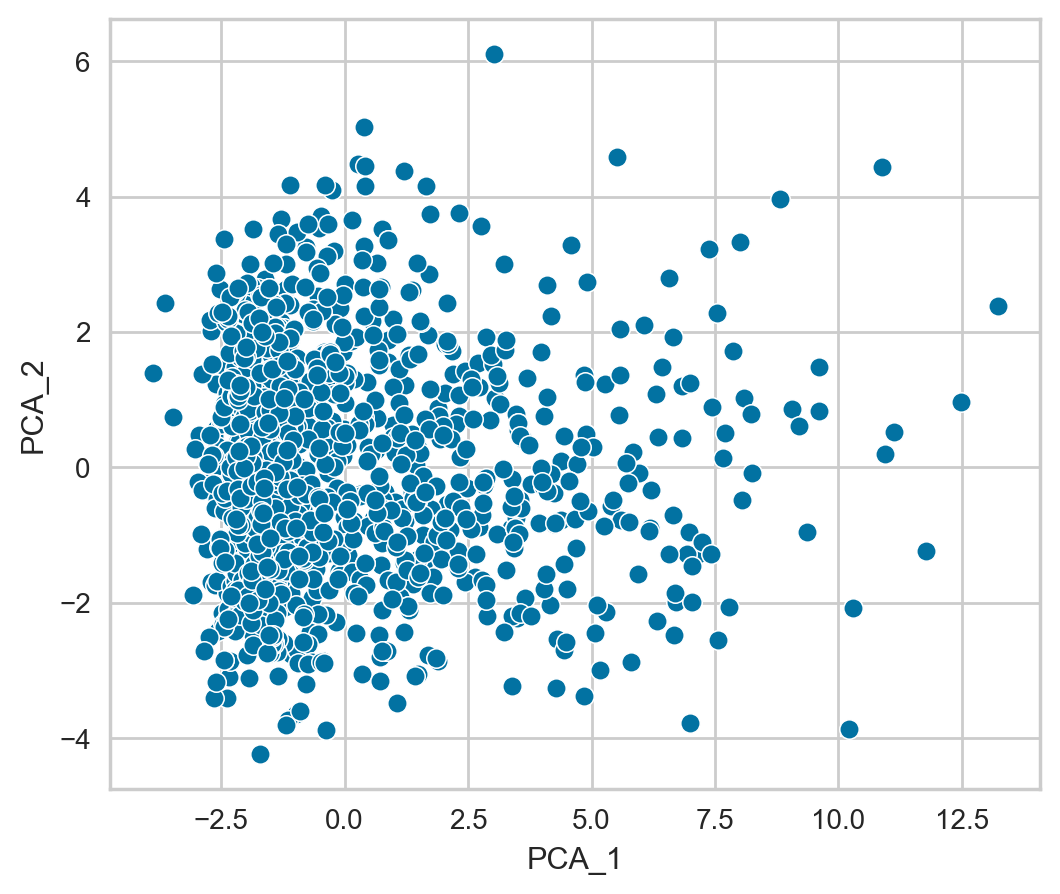


Hình 51: Triển khai thuật toán PCA

Trực quan hóa bằng Biểu đồ điểm:



Hình 52: Mã nguồn biểu đồ điểm sau khi giảm chiều



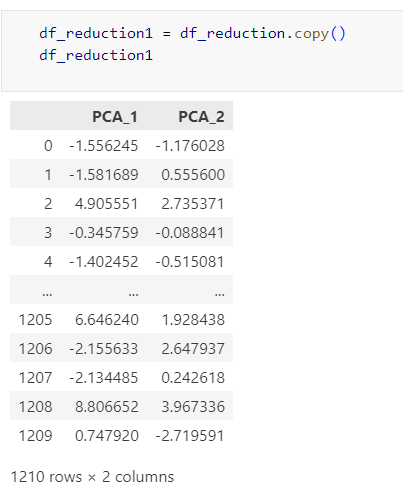
Hình 53: Biểu đồ phân tán sau khi áp dụng PCA

Việc giảm chiều này sẽ hỗ trợ trong việc trực quan hóa cho những mô hình học không giám sát dưới đây.

### 2.4.2 Mô hình Kmeans

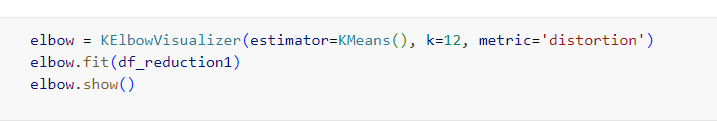
K-Means Clustering là một thuật toán trong machine learning và data mining được sử dụng để phân loại một tập dữ liệu thành các nhóm (clusters) dựa trên các đặc trưng của chúng. Phương pháp này thuộc về loại các thuật toán unsupervised learning, nghĩa là không yêu cầu thông tin nhãn trước đó về từng điểm dữ liệu.

Tạo một biến data mới cho việc phân cụm K-Means:

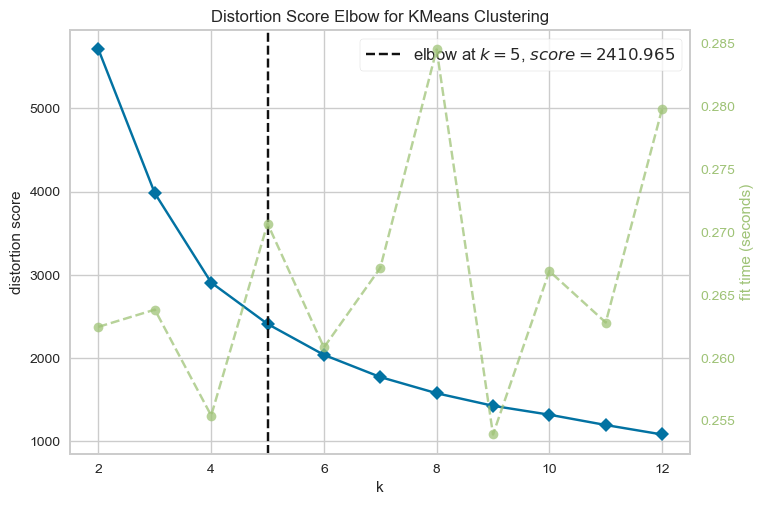


Hình 54: Chuẩn bị dữ liệu cho K-Means

Vẽ đường Elbow để xác định số K hợp lí:



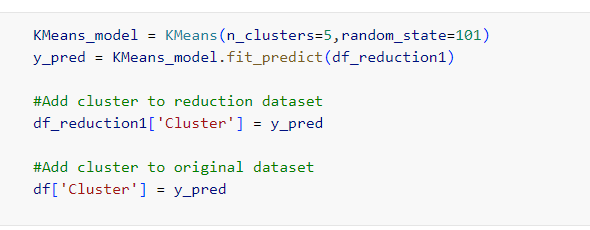
Hình 55: Mã nguồn tìm hệ số k phù hợp cho model K-Means



Hình 56: Kết quả tìm hệ số K cho mô hình K-means

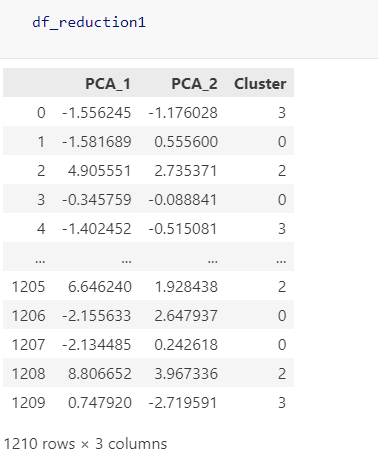
Từ kết quả ta thấy số K hợp lí là 5

Khởi tạo mô hình K-Means như dưới đây.



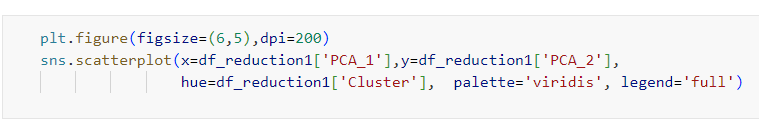
Hình 57: Mã nguồn huấn luyện mô hình K-Means

Kết quả sau thực hiện mô hình K-means phân cụm như dưới đây.

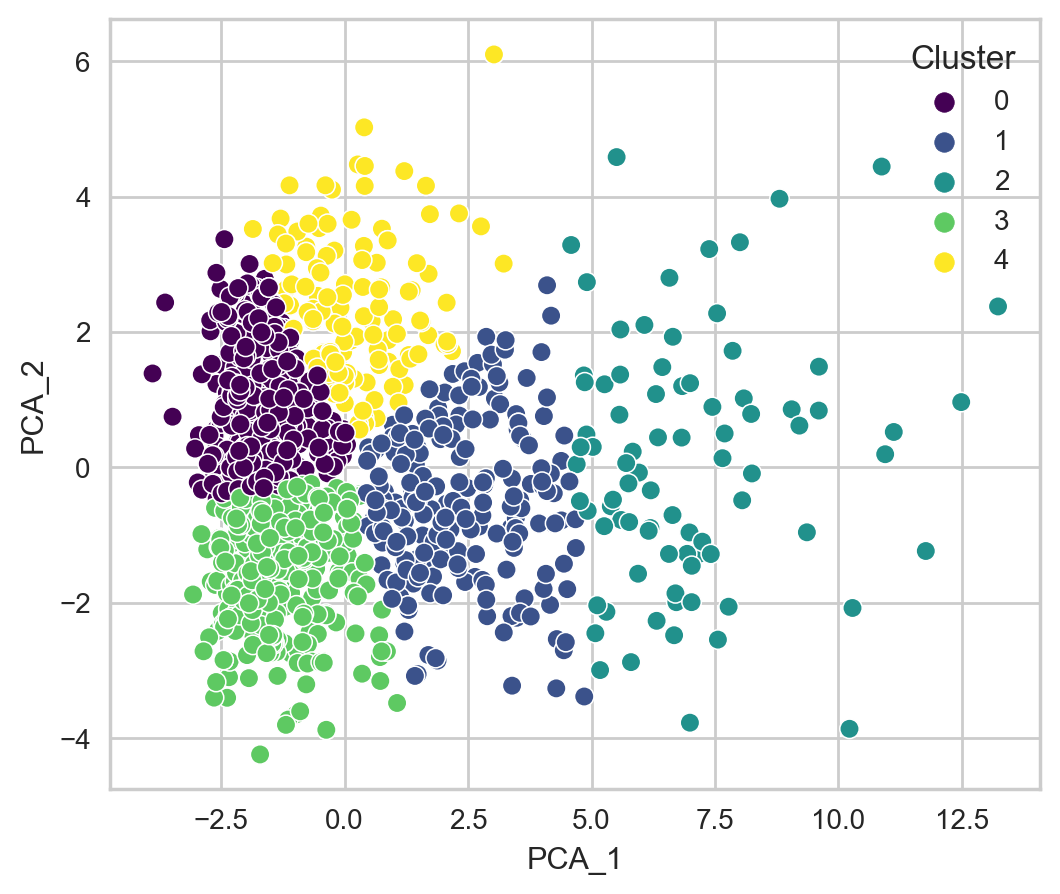


Hình 58: Kết quả phân cụm bằng K-means

Trực quan hóa dữ liệu sau khi phân cụm bằng K-Means như dưới đây.



Hình 59: Mã nguồn trực quan hóa phân cụm sau khi dùng K-means



Hình 60: Biểu đồ phân cụm bằng K-means

### 2.4.3 Mô hình DBSCAN

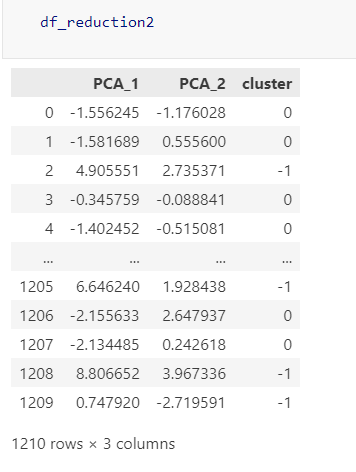
DBSCAN (Density-Based Spatial Clustering of Applications with Noise) là một thuật toán phân cụm (clustering) dựa trên mật độ trong không gian đặc trưng. Điều này có nghĩa là DBSCAN tập trung vào việc phát hiện các vùng có mật độ cao của điểm dữ liệu và tự động phân loại chúng thành các cụm (clusters). DBSCAN cũng có khả năng xác định các điểm nhiễu, tức là các điểm không thuộc vào bất kỳ cluster nào.

Khởi tạo mô hình DBSCAN như dưới đây:



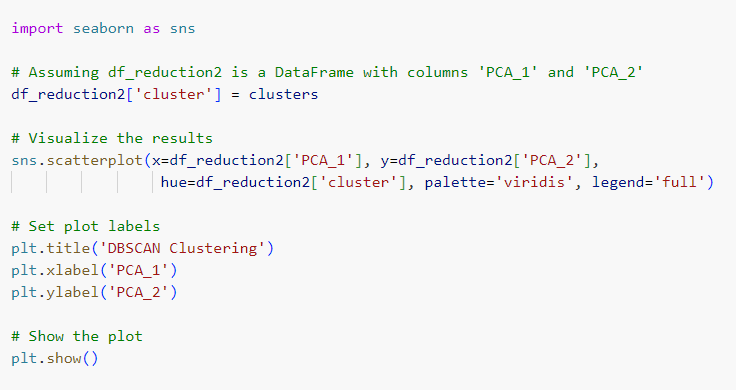
Hình 61: Khởi tạo mô hình phân cụm bằng DBSCAN

Kết quả sau khi phân cụm bằng Dbscan như hình ảnh dưới đây:



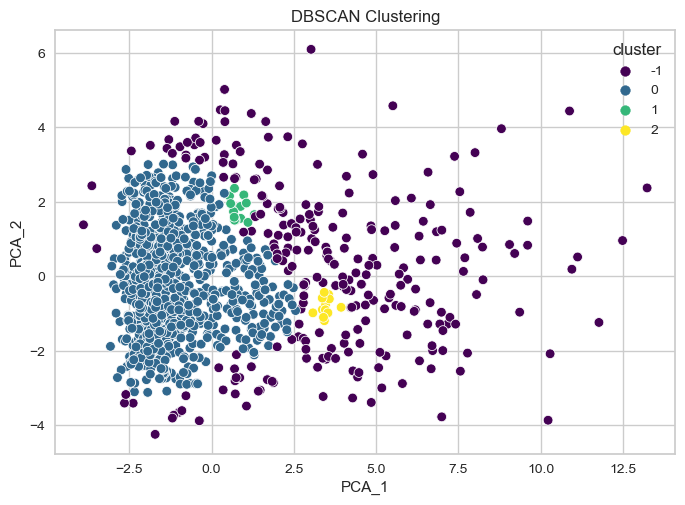
Hình 62: Kết quả phân cụm bằng DBSCAN

Trực quan hóa kết quả phân cụm bằng DBScan như hình dưới đây:



Hình 63: Mã nguồn trực quan hóa phân cụm bằng DBScan

Dưới đây là kết quả phân cụm bằng DBScan bằng biểu đồ trực quan



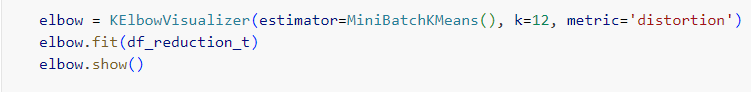
Hình 64: Kết quả trực quan hóa phân cụm bằng DBScan

### 2.4.4 Mô hình MiniBatchKMeans

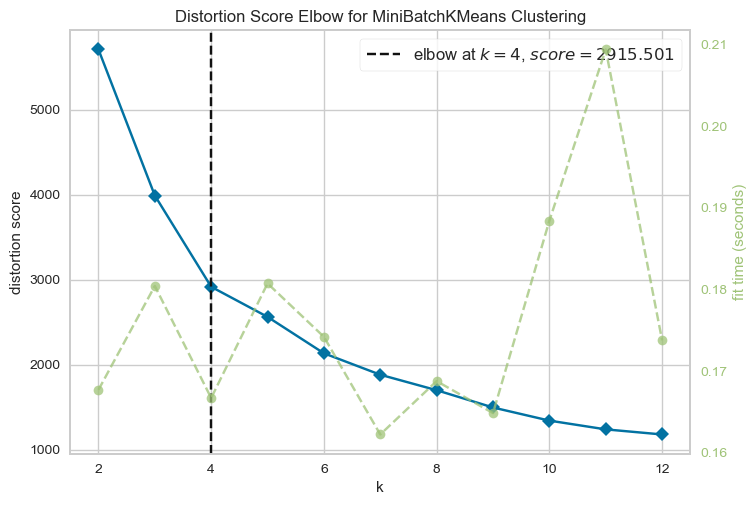
MiniBatchKMeans là một biến thể của thuật toán K-Means Clustering được thiết kế để xử lý dữ liệu lớn một cách hiệu quả hơn. Trong thuật toán K-Means truyền thống, toàn bộ tập dữ liệu được sử dụng để cập nhật centroids và phân loại các điểm vào các nhóm. Điều này có thể trở nên không hiệu quả khi làm việc với dữ liệu lớn.

MiniBatchKMeans giả định rằng chỉ một phần nhỏ (mini-batch) của dữ liệu được sử dụng để cập nhật centroids ở mỗi bước lặp. Điều này giúp giảm độ phức tạp tính toán và làm tăng tốc độ hội tụ của thuật toán. Bằng cách này, MiniBatchKMeans có thể xử lý các tập dữ liệu lớn mà vẫn giữ được tính chính xác tương đương với K-Means truyền thống.

Vì là 1 biến thể của KMeans Clustering, MiniBatchKMeans cũng cần xác định số k bằng Elbow.

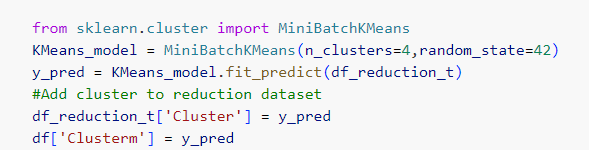


Hình 65: Mã nguồn tìm hệ số K cho MiniBactch Kmeans



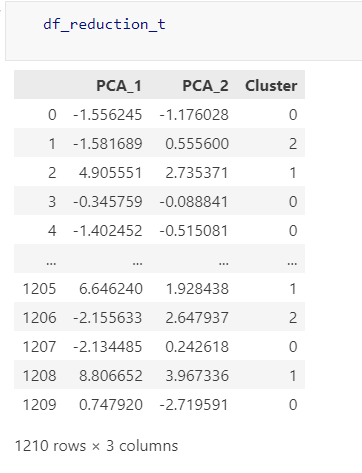
Hình 66: Kết quả tìm hệ số K cho MinniBatch Kmeans

Sau khi xác định tham số K tốt nhất cho mình thì tiến hành xây dựng mô hình MiniBatchKMeans:



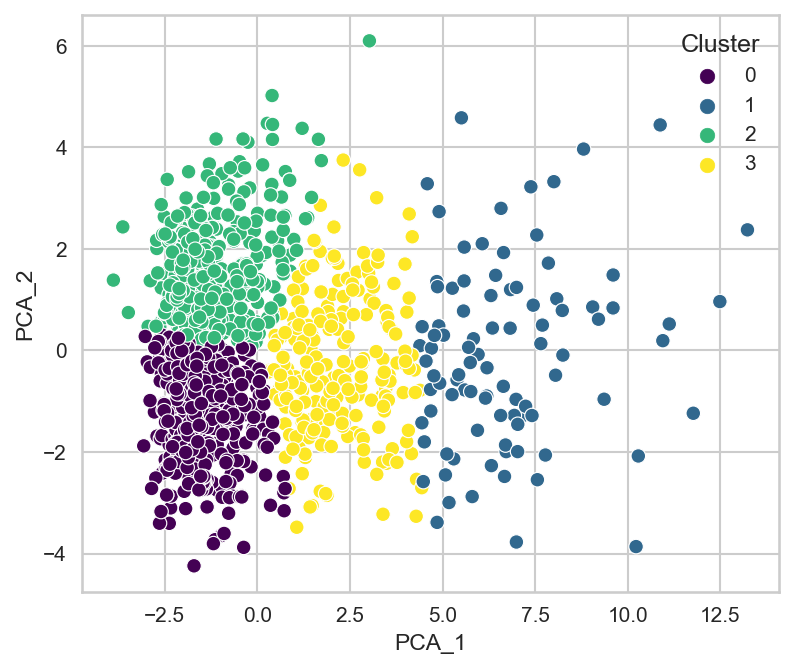
Hình 67: Mã nguồn mô hình MiniBatchKmeans

Kết quả sau phân cụm bằng mô hình MiniBatch Kmeans, như dưới đây



Hình 68: Kết quả phân cụm bằng MiniBatch Kmeans

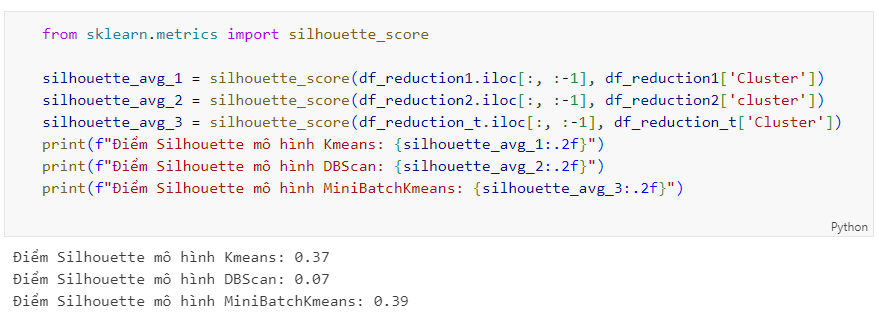
Trực quan hóa kết quả phân cụm bằng MiniBatchKMeans như dưới đây.



Hình 69: Kết quả trực quan hóa bằng MiniBatchKmeans

## **2.5** Thảo luận, phân tích, đánh giá và kết luận về kết quả nhận được sau khi tích dữ liệu

Từ ba mô hình đã được biểu thị ở trên, ta tính điểm silhouette\_score để đánh giá độ hiệu quả của mô hình.



Hình 70: Mã nguồn và kết quả kiểm tra của các mô hình

Dựa trên điểm Silhouette, mô hình K-Means (0.37) và MiniBatchKMeans (0.39) đều cho thấy chất lượng phân cụm tốt hơn so với mô hình DBScan (0.07). Điểm Silhouette là một đánh giá về mức độ đồng nhất và tách biệt giữa các cụm, và điểm cao hơn thường chỉ ra rằng phân cụm của mô hình đó là tốt.

Vì vậy, nếu mục tiêu là phân loại dữ liệu thành các nhóm một cách chất lượng và hiệu quả, có thể ưa thích sử dụng mô hình K-Means hoặc MiniBatchKMeans thay vì DBScan trong trường hợp này.

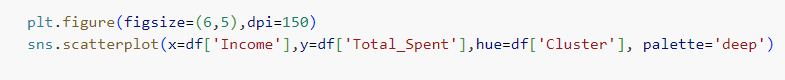
K-Means là một phương pháp phân cụm đơn giản nhưng mạnh mẽ, thích hợp cho dữ liệu có cấu trúc hình cụm tròn và có kích thước tương đồng. Nó tốt cho việc phân chia dữ liệu thành các nhóm với đồng nhất cao.

MiniBatchKMeans là một biến thể của K-Means thích hợp cho dữ liệu lớn và có thể cung cấp sự đồng nhất tương đối tốt (0.39) trong trường hợp của bạn.

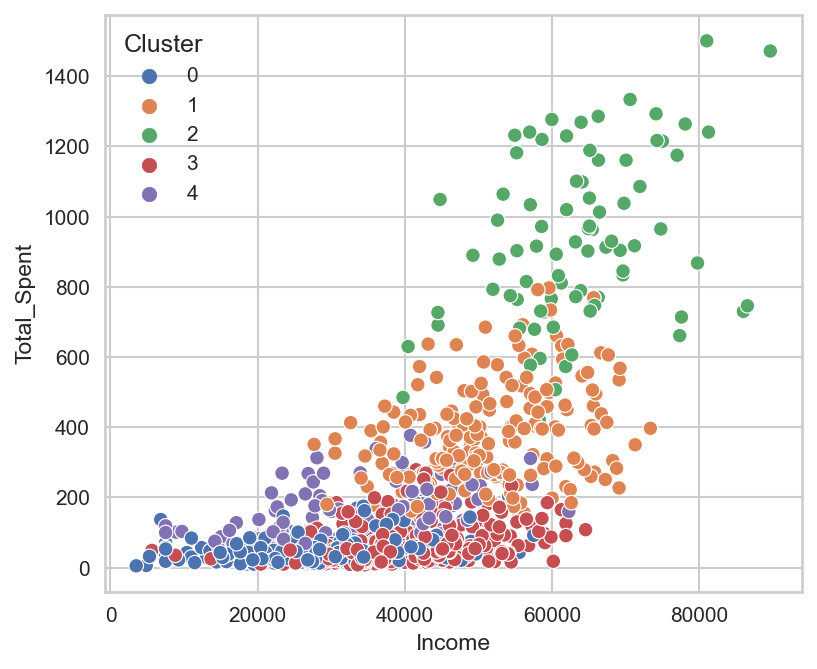
Trong khi đó, DBScan (0.07) có thể không phù hợp với dữ liệu hoặc yêu cầu các điều kiện về mật độ và hình dạng cụm khác nhau để hoạt động tốt. DBSCan đã không đạt được mức độ đồng nhất cao như K-Means.

Do đó, dựa trên đánh giá Silhouette, khuyến nghị sử dụng K-Means hoặc MiniBatchKMeans để phân tích dữ liệu và tạo các cụm trong dữ liệu này.

Qua biểu đồ nhiệt được vẽ ở trên, chúng tôi chọn ra 2 cột có độ tương quan khá cao là 'Income' ,'Total\_Spent' , sử dụng nó để chiếu và phân cụm về 5 cụm có Tổng thu nhập và tổng chi tiêu khác nhau.



Hình 71: Mã nguồn trực quan phân cụm dựa trên tổng thu và chi



Hình 72: Trực quan phân cụm dựa vào 2 tổng thu và tổng chi

Từ biểu đồ trên, chúng tôi phân tích cụm dựa trên thu nhập và tổng chi tiêu như sau:

\* Cluster 0 : Thu nhập thấp & Chi tiêu thấp.

\* Cluster 1 : Thu nhập trung bình & Chi tiêu trung bình đến cao.

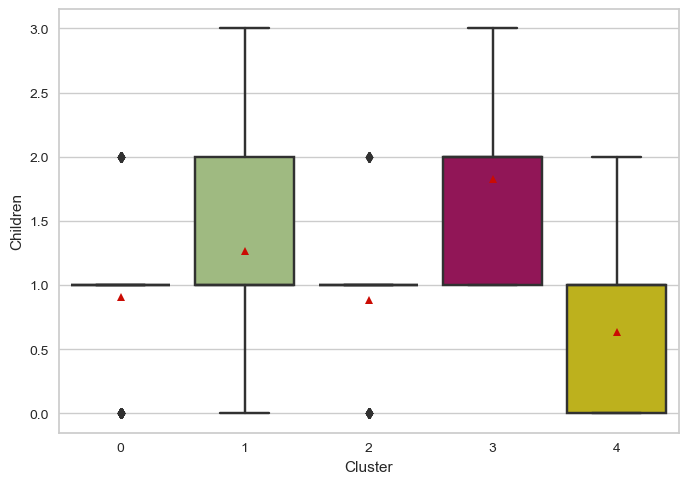
\* Cluster 2 : Thu nhập Cao & Chi tiêu cao.

\* Cluster 3 : Thu nhập trung bình đến cao & Chi tiêu thấp.

\* Cluster 4 : Thu nhập thấp & Chi tiêu thấp đến trung bình.

Tương tự với biểu đồ boxplot để cho thấy số lượng con cái của từng cụm:





Hình 73: Biểu đồ hộp về số lượng trẻ em của mỗi cụm

Giải thích:

\* Cluster 0 : Đa số có 1 con cái

\* Cluster 1 : Đa số có 1-2 con cái

\* Cluster 2 : Đa số có 1 con cái

\* Cluster 3 : Đa số có 1-2 con cái

\* Cluster 4 : Đa số có 0-1 con cái

# TÀI LIỆU THAM KHẢO