BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH

VŨ NGỌC DUY
NGUYỄN NGUYÊN ÁI VÂN
NGUYỄN PHÁT

BÀI BÁO CÁO MÔN HỌC
PHÂN TÍCH DỮ LIỆU

Mã số sinh viên: **2254052018**
2254052087
2254052059

ĐỀ TÀI
"DỰ BÁO QUYẾT ĐỊNH MUA HÀNG HÓA HỮU CƠ CỦA KHÁCH HÀNG"

Giảng viên hướng dẫn: HỒ HƯỚNG THIÊN

TP.HỒ CHÍ MINH, 2024

MỤC LỤC

DANH MỤC HÌNH ẢNH

PHẦN I
GIỚI THIỆU ĐỀ TÀI

1.1 Lý do chọn đề tài

Với nền nông nghiệp nước nhà đang phát triển rất mạnh mẽ hiện nay, phân bón, thuốc bảo vệ thực vật cũng như các loại chất hóa học khác đã hình thành nên một con dao hai lưỡi. Một mặt nó giúp phòng trừ được sâu bệnh, cho năng suất cao, vẻ ngoài bắt mắt thì mặt khác của nó là sự đánh đổi về sức khỏe của người tiêu dùng. Việt Nam hiện nay đang là một trong số những nước có tỉ lệ ung thư liên quan đến việc ăn uống nhiều nhất thế giới. Vì lẽ đó, những sản phẩm mang tính hữu cơ (organic) cũng đang dần được hình thành và phát triển. Đầu năm 2022, với nền nông nghiệp lúa nước sẵn có tại Kiên Giang, dưới tên công ty Đình Kiên đã cho ra mắt các sản phẩm liên quan đến lúa hữu cơ như: gạo, bột gạo, bánh cốm, nếp, rượu.... và một số sản phẩm khác hoàn toàn được trồng trọt dưới mô hình Organic đặt chuẩn của tình nhà. Vì muốn nâng cao khả năng cạnh tranh cũng như phân định được tệp khách hàng cũng như phát triển chiến lược tiếp thị nên chúng em chọn đề tài "Dự báo quyết định mua hàng hữu cơ của khách hàng" để giải quyết các vấn đề nan giải ở trên.

1.2 Giới thiệu

Công ty TNHH Đình Kiên tiền thân là một của hàng phân bón và thuốc trừ sâu với quy mô 3 cơ sở được đặt tại các kênh 7,8,11 xã Thạnh Đông A, huyện Tân Hiệp, tỉnh Kiên Giang. Nay tiếp tục mở rộng với mô hình lúa nước Organic tại tỉnh An Giang - thủ phủ của lúa nước. Với gần 9000 dữ liệu thu thập từ hai tỉnh An Giang và Kiên Giang thông qua khách hàng mua phân bón và thuốc trừ sâu để tiếp cận đến tệp khách hàng tiềm năng thông các nhân tố như giới tính, tuổi tác cũng như các dữ liệu về gia đình.

PHẦN II
PHÂN TÍCH DỮ LIỆU KHÁM PHÁ

2.1 Chuẩn bị và mô tả dữ liệu

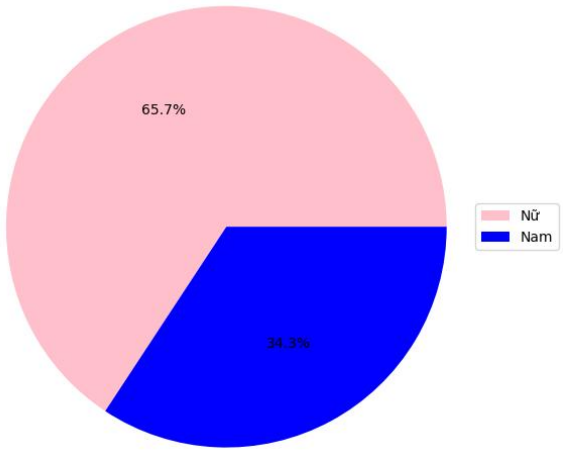
2.1.1 Mô tả thuộc tính

- Gender: Giới tính khách hàng (0:nữ/1:nam).
- Age: Tuổi.
- AffluenceGrade: Mức độ giàu có của khách hàng (Theo bảng phân phối từ 1 cho đến 35, số càng lớn mức độ giàu có càng tăng).
- LoyaltyCardTenure: Số năm mà khách hàng sử dụng các loại thẻ để mua hàng.
- TotalSpend: Tổng chi tiêu của khách hàng.
- OrganicsPurchaseIndicator: Chỉ số quyết định mua hàng của khách hàng (0:Không mua/1:Mua)

2.1.2 Mô tả thống kê

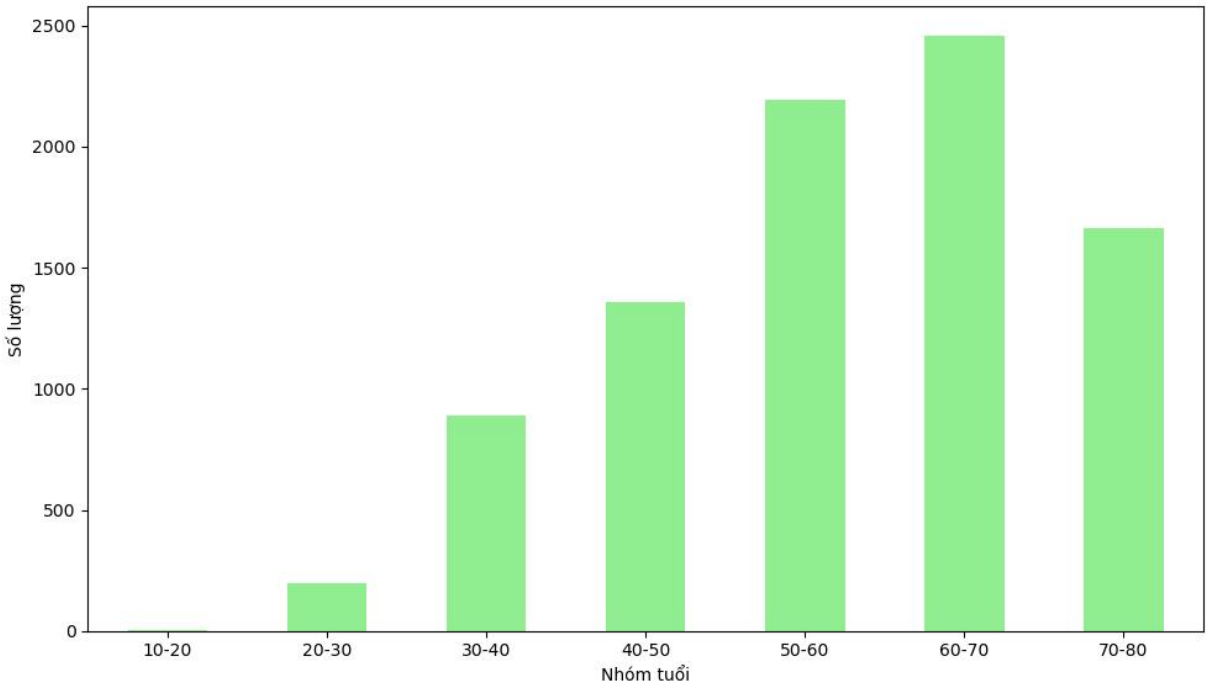
```
df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Gender	8766.0	0.342688	0.474635	0.0	0.0	0.0	1.0	1.0
Age	8766.0	56.912389	13.121142	18.0	48.0	58.0	67.0	79.0
AffluenceGrade	8766.0	8.554985	3.347816	2.0	6.0	8.0	10.0	34.0
LoyaltyCardTenure	8766.0	7.380105	5.217871	2.0	4.0	6.0	9.0	38.0
TotalSpend	8766.0	6398.107347	7636.577461	1.0	1500.0	4884.0	7500.0	110072.0
OrganicsPurchaseIndicator	8766.0	0.179786	0.384031	0.0	0.0	0.0	0.0	1.0



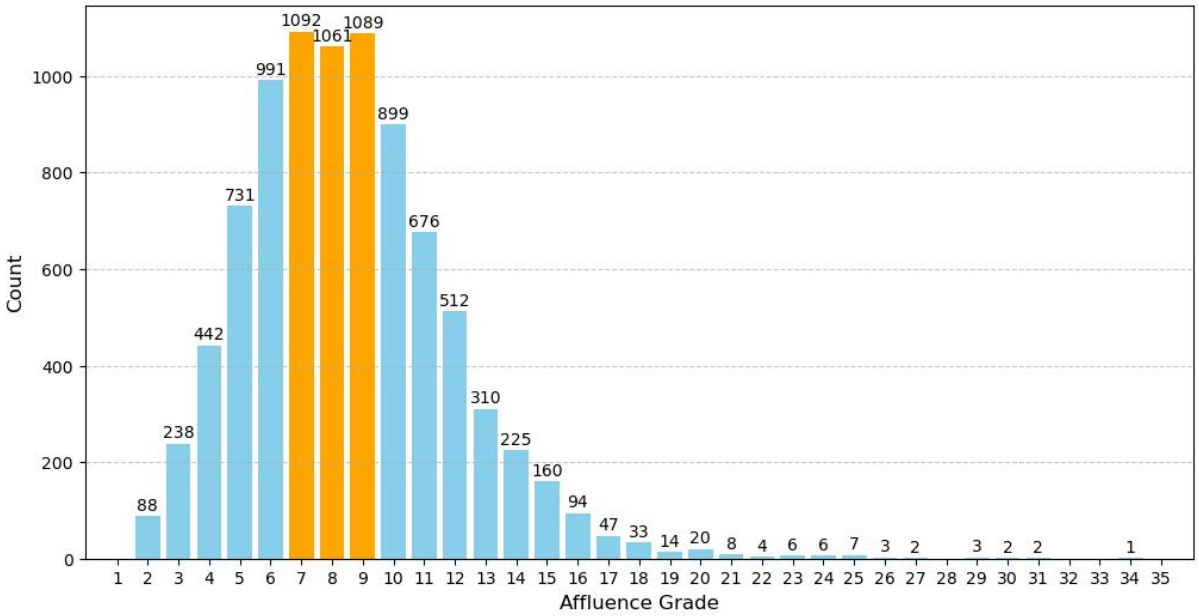
Hình 1. Biểu đồ giới tính tham gia khảo sát

Biểu đồ giới tính cho thấy đa phần người tham gia khảo sát là nữ, chiếm 65,7% tổng số người tham gia khảo sát. Điều này cho thấy, sự quan tâm đến sức khỏe cá nhân và gia đình của những người phụ nữ. Việc phân loại khách hàng của doanh nghiệp nên tập trung vào các "Bà nội trợ" chăm sóc gia đình, chiến lược tiếp thị nên đánh vào tâm lý sức khỏe, không hóa chất, không thuốc bảo vệ thực vật để an toàn cho gia đình và con cái.



Hình 2. Biểu đồ nhóm tuổi tham gia khảo sát

Từ biểu đồ nhóm tuổi tham gia khảo sát cho thấy, nhóm 50-60 và 60-70 có sự quan tâm đến các hàng hóa hữu cơ hơn so với các nhóm trẻ tuổi như 20-30 và 30-40. Qua đó, cho thấy rằng càng lớn tuổi thì việc con người chú trọng vào sức khỏe càng tăng, cao nhất là nhóm 60-70 tuổi. Theo Vinmec International Hospital, độ tuổi có nguy cơ mắc các bệnh về ung thư cao nhất đang vào khoảng 65-74 tuổi với 25%, kế sau đó là 24% trong độ tuổi 55-64. Không những thế mà có dấu hiệu ngày càng trẻ hóa trong độ tuổi từ 25-40 tuổi. Do tâm lý phòng bệnh, mà khi chúng ta càng lơn tuổi thì việc tự bảo vệ bản thân khỏi những bệnh khó chữa, nan y càng quan trong hơn, đặc biệt là trong vấn đề ăn uống.

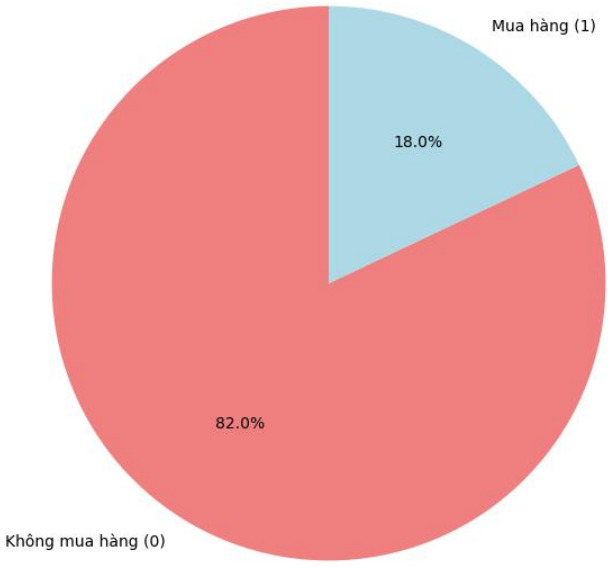


Hình 3. Biểu đồ phản ánh mức độ giàu có của khách hàng tham gia khảo sát

Mức độ giàu có của khách hàng phản ánh lên tài sản mà họ đang nắm giữ, từ 1-5 là những gia đình khó khăn, từ 6-11 là gia đình có tài chính ổn định, 12-20 là gia đình khá giả và trên 20 là những gia đình giàu có. Từ bảng dữ liệu có thể thấy mức tài chính đa phần nằm trong khoảng gia đình có tài chính ổn định, nhiều nhất là khoảng 7-8-9. Cũng nói lên phần nào gia đình tại các vùng nông thôn có mức sống vừa phải, không quá giàu cũng không quá khó khăn. Do đó, mục tiêu của cửa hàng cần nhắm đến nhóm gia đình bình dân và trung lưu.

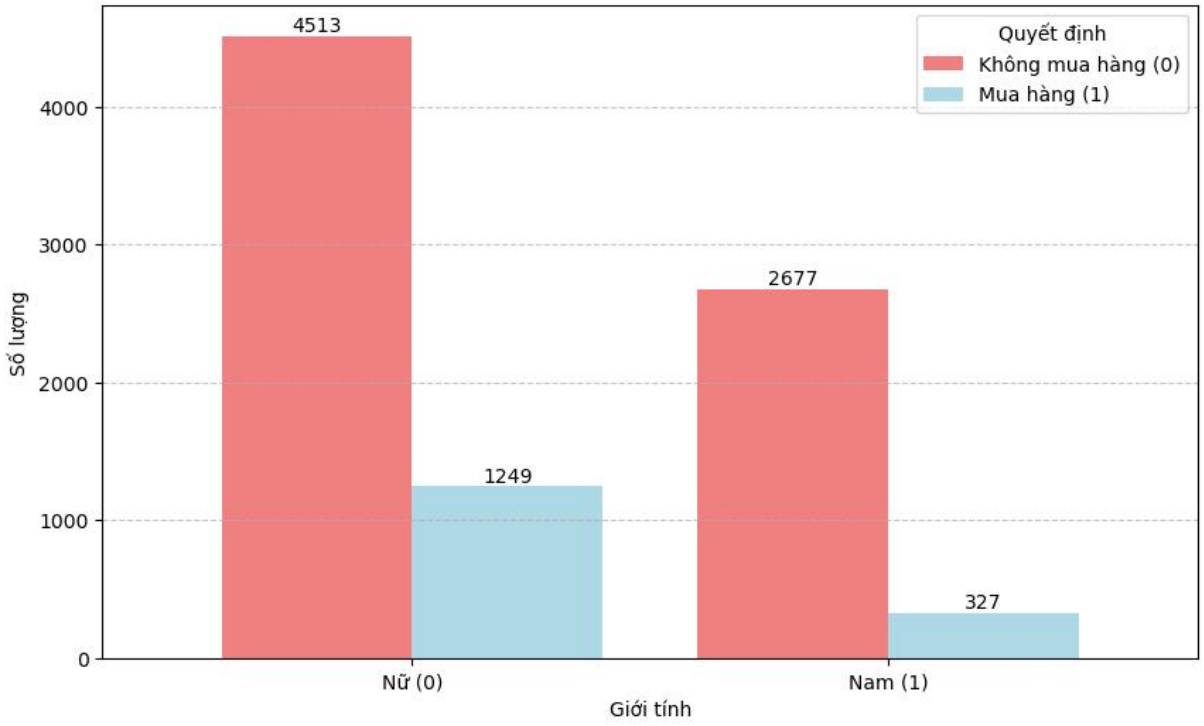
2.2 Phân tích đơn biến

Với gần 9000 người tham gia khảo sát, nhưng chỉ có 18% người khảo sát tích vào ô trống mua hàng hóa hữu cơ, cho thấy hàng hóa hữu cơ vẫn là một ngành hàng sản phẩm mới mẻ với mọi người. Khảo sát chủ yếu ở hai tỉnh An Giang và Kiên Giang, 2 tỉnh không chỉ đi đầu về nền nông nghiệp ở khu vực đồng bằng



Hình 4. Biểu đồ tỉ lệ mua hàng hữu cơ

sông Cửu Long mà còn dẫn đầu về cả nước về sản lượng lúa cũng như các loại trái cây. Với việc đất trồng trọt rộng rãi cũng như kinh nghiệm thâm niên về các loại cây, rau, củ... nên việc mỗi nhà có một mảnh vườn riêng để trồng các loại rau củ cây ăn trái hữu cơ riêng để phục vụ cho gia đình là một điều hiển nhiên. Cũng một phần là quan niệm về hữu cơ trong tiềm thức của người dân về các sản phẩm mang tính hữu cơ, là hữu cơ chỉ cho các loại thực phẩm không dành cho các sản phẩm nên mọi người thường ngần ngại tiếp xúc với các mặt hàng hữu cơ.

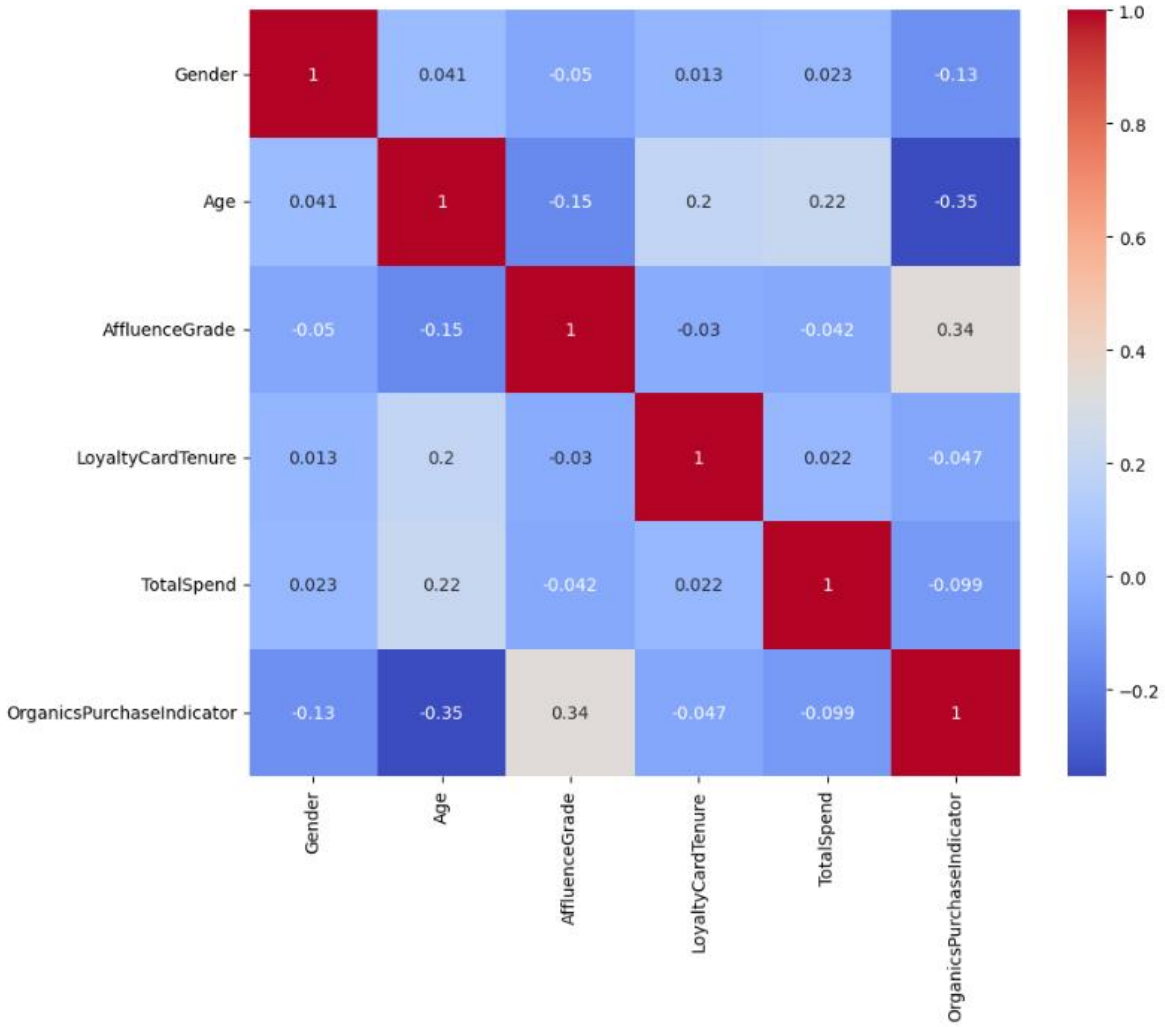


Hình 5. Biểu đồ mua hàng hữu cơ theo giới tính

Số lượng đồng ý mua hàng hóa hữu cơ của nữ giới gấp gần 4 lần so với nam giới. Phụ nữ, đặc biệt là những người làm nội trợ hoặc chăm sóc gia đình, thường quan tâm đến sức khỏe của bản thân và gia đình. Hàng hóa hữu cơ thường được liên kết với chất lượng tốt hơn, ít hóa chất độc hại, và an toàn hơn, nên phụ nữ có xu hướng ưu tiên chúng hơn. Hàng hóa hữu cơ thường được liên kết với lối sống "xanh" hoặc "healthy", những xu hướng này thường phổ biến hơn ở phụ nữ. Họ có thể dễ bị ảnh hưởng bởi các xu hướng xã hội hoặc cộng đồng hơn nam giới.

2.3 Phân tích đa biến

2.3.1 Correlation matrix



Hình 6. Correlation matrix

Nhận xét từng mối quan hệ với OrganicsPurchaseIndicator:

Gender (Giới tính): Hệ số: -0.13. Có mối tương quan yếu ngược chiều giữa giới tính và việc mua hàng hữu cơ. Điều này có thể ngụ ý rằng phụ nữ (giới tính = 0) có xu hướng mua hàng hữu cơ nhiều hơn so với nam giới.

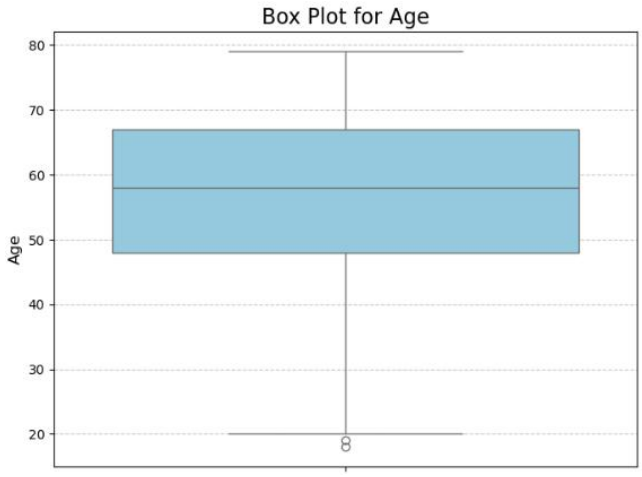
Age (Tuổi): Hệ số: -0.35. Tương quan ngược chiều vừa phải. Người trẻ có xu hướng mua hàng hữu cơ nhiều hơn so với người lớn tuổi.

AffluenceGrade (Mức độ giàu có) : Hệ số: 0.34. Tương quan cùng chiều vừa phải. Người có mức độ giàu có cao hơn có xu hướng mua hàng hữu cơ nhiều hơn.

LoyaltyCardTenure (Thời gian sử dụng thẻ) :Hệ số: -0.047. Mối tương quan yếu gần như không đáng kể giữa thời gian sử dụng thẻ khách hàng thân thiết và quyết định mua hàng hữu cơ.

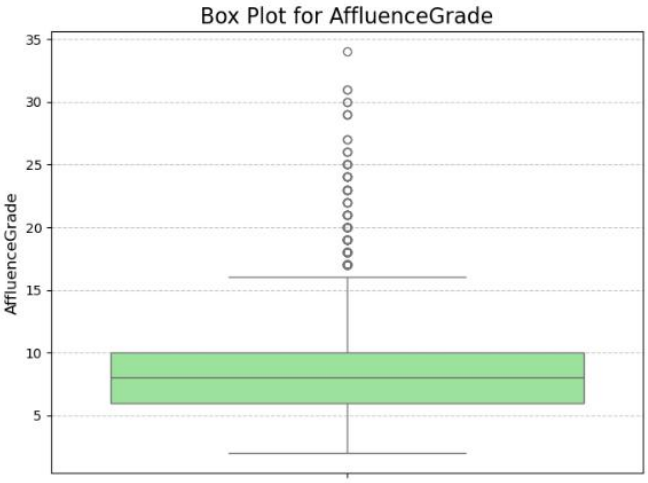
TotalSpend (Tổng chi tiêu) : Hệ số: -0.099. Tương quan yếu ngược chiều, không rõ ràng giữa tổng chi tiêu và việc mua hàng hữu cơ. Người chi tiêu ít không hẳn có xu hướng mua hàng hữu cơ nhiều hơn.

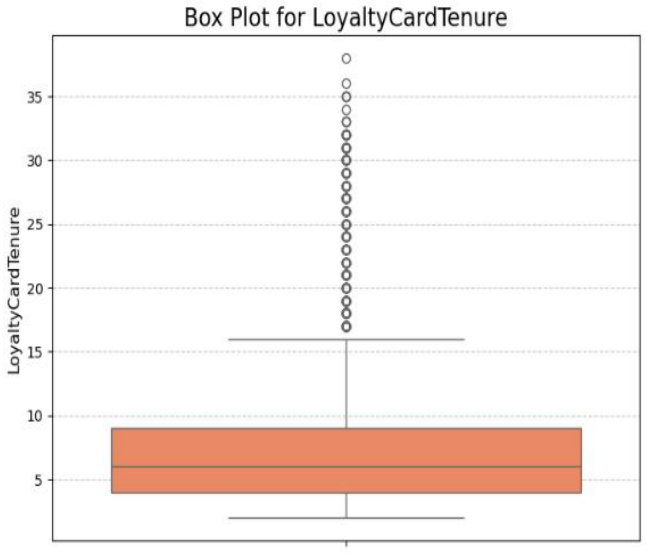
2.3.2 Box Plot



Phân bố độ tuổi khá rộng, tuy nhiên nhóm khách hàng chính tập trung trong độ tuổi trung niên. Điều này gợi ý rằng các chiến lược tiếp thị nên hướng tới nhóm đối tượng từ 50 đến 70 tuổi để tối ưu hóa hiệu quả.

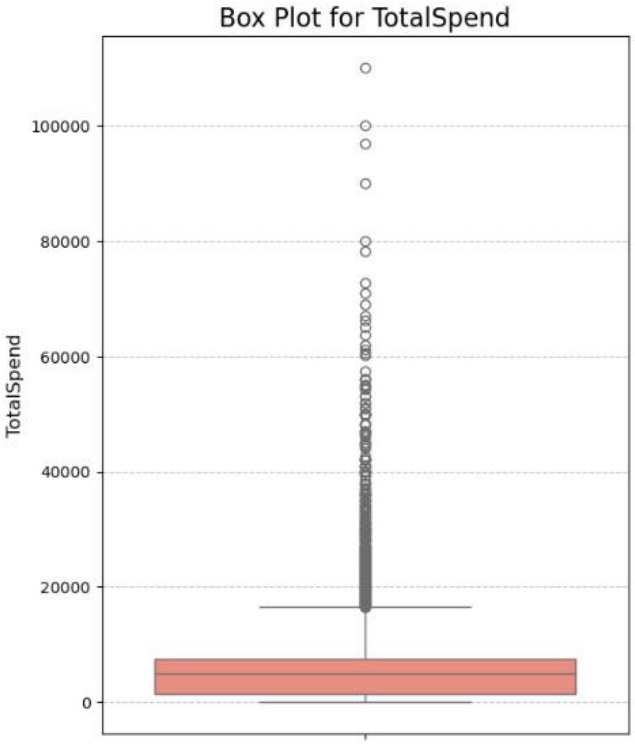
Phần lớn khách hàng có mức độ giàu có ở mức trung bình. Tuy nhiên, có một số lượng đáng kể các khách hàng có mức độ giàu có cao hơn. Điều này cho thấy cơ sở khách hàng của bạn khá đa dạng về mức thu nhập.





Phần lớn khách hàng đã sử dụng thẻ thành viên trong một khoảng thời gian tương đối ngắn. Tuy nhiên, có một số lượng đáng kể các khách hàng đã sử dụng thẻ thành viên trong thời gian dài. Điều này cho thấy bạn có một lượng khách hàng trung thành nhất định, đồng thời cũng có một lượng khách hàng mới.

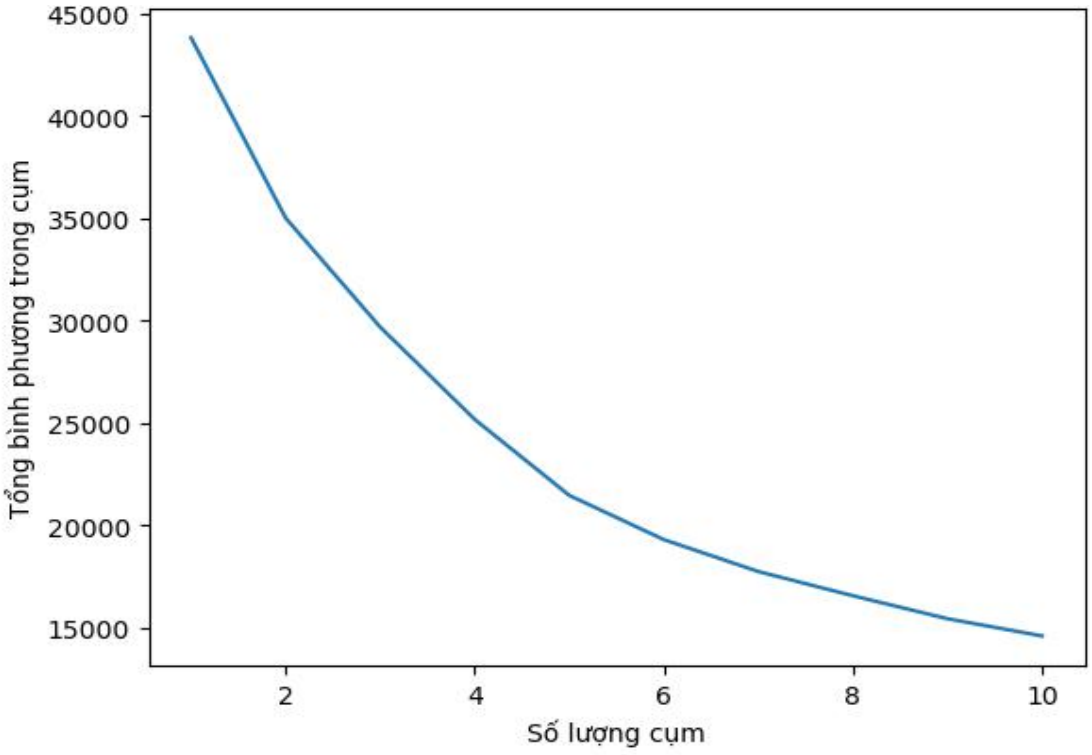
Phần lớn khách hàng có tổng chi tiêu tương đối thấp. Tuy nhiên, có một số khách hàng có tổng chi tiêu rất cao, tạo thành các giá trị ngoại lệ.Điều này cho thấy có sự chênh lệch lớn về mức độ chi tiêu giữa các khách hàng.



\

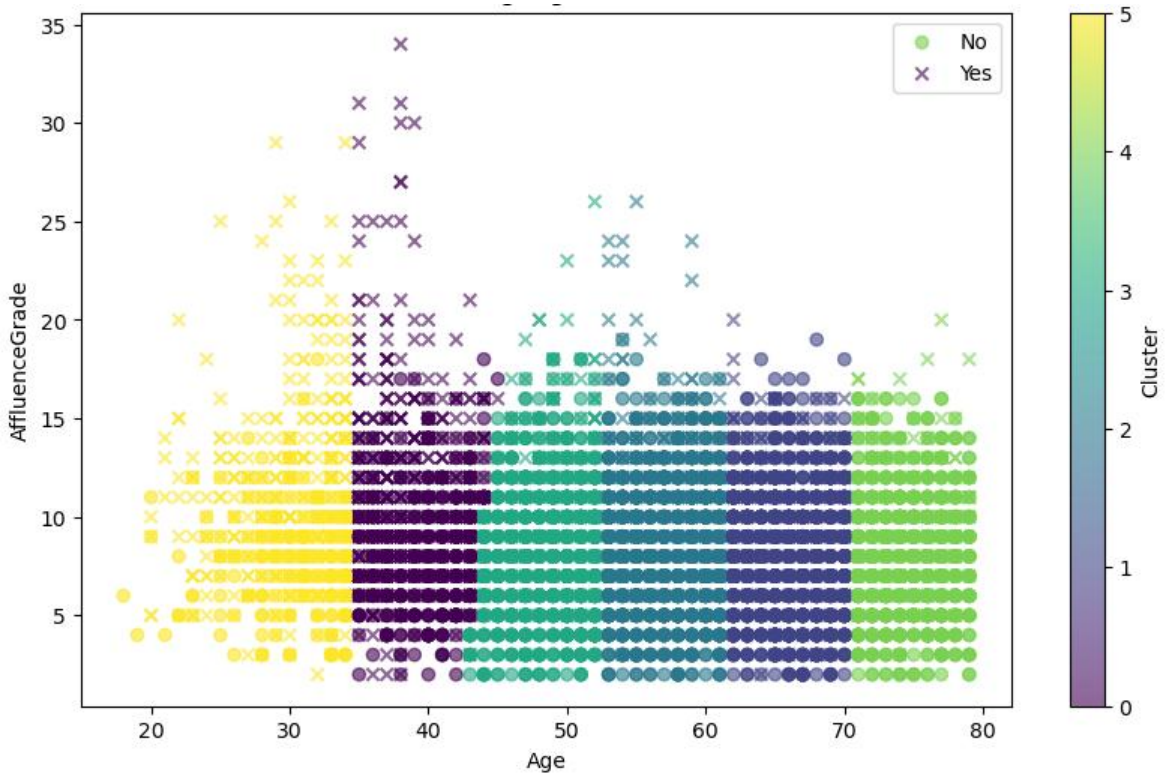
PHẦN III
PHÂN CỤM

3.1 Elbow



Hình 7. Elbow

Biểu đồ Elbow là công cụ quan trọng để xác định số lượng cụm tối ưu trong phân cụm K-means. Trên biểu đồ, trục x thể hiện số lượng cụm, trong khi trục y biểu diễn Tổng bình phương trong cụm (WCSS) - một đánh giá về độ nhóm của dữ liệu. 'Elbow' trên biểu đồ là điểm mà thêm cụm không cải thiện đáng kể WCSS.Nhìn chung, *số cụm tối ưu cho dữ liệu là 6*, được xác định dựa trên phương pháp Elbow.



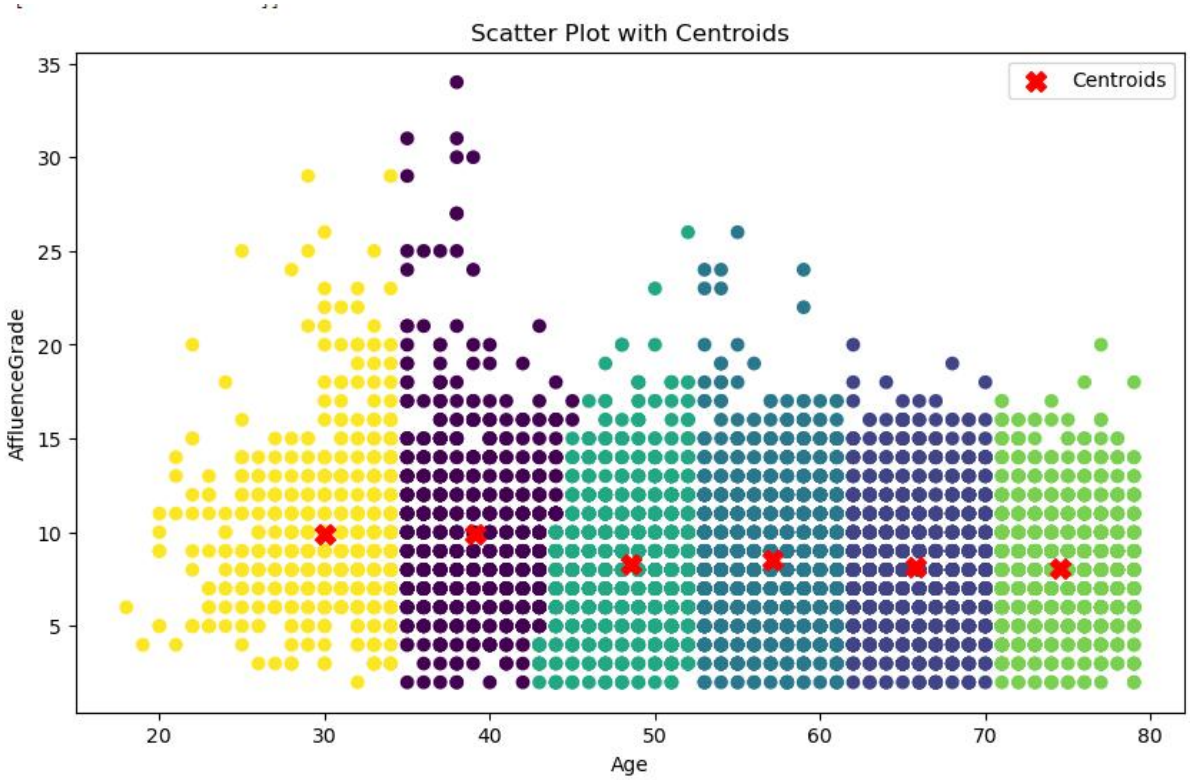
Hình 8. Biểu đồ phân tán trực quan mối quan hệ giữa Age và AffluenceGrade

Biểu đồ phân tán trực quan hóa mối quan hệ giữa Age và AffluenceGrade, hai tính năng quan trọng trong tập dữ liệu. Mỗi điểm trên biểu đồ đại diện cho một cá nhân trong tập dữ liệu và màu của điểm biểu thị cụm mà cá nhân đó thuộc về dựa trên phân tích phân cụm K-means:

- Age: Tính năng này được thể hiện trên trục x. Các cá nhân có mức độ AffluenceGrade khác nhau.
- AffluenceGrade: Tính năng này được thể hiện trên trục y. Các cá nhân có nhiều giá trị Age khác nhau.

Clusters: Các màu khác nhau trên biểu đồ đại diện cho các cụm khác nhau. Dùng thuật toán phân cụm K-mean đã nhóm các cá nhân thành các cụm riêng biệt dựa trên Age và AffluenceGrade.

3.2 Tâm cụm



Hình 9. Biểu đồ phân tán có tâm mối quan hệ giữa Age và AffluenceGrade

Vị trí tâm

[[39.11 9.89]

[65.80 8.17]

[57.10 8.50]

[48.56 8.29]

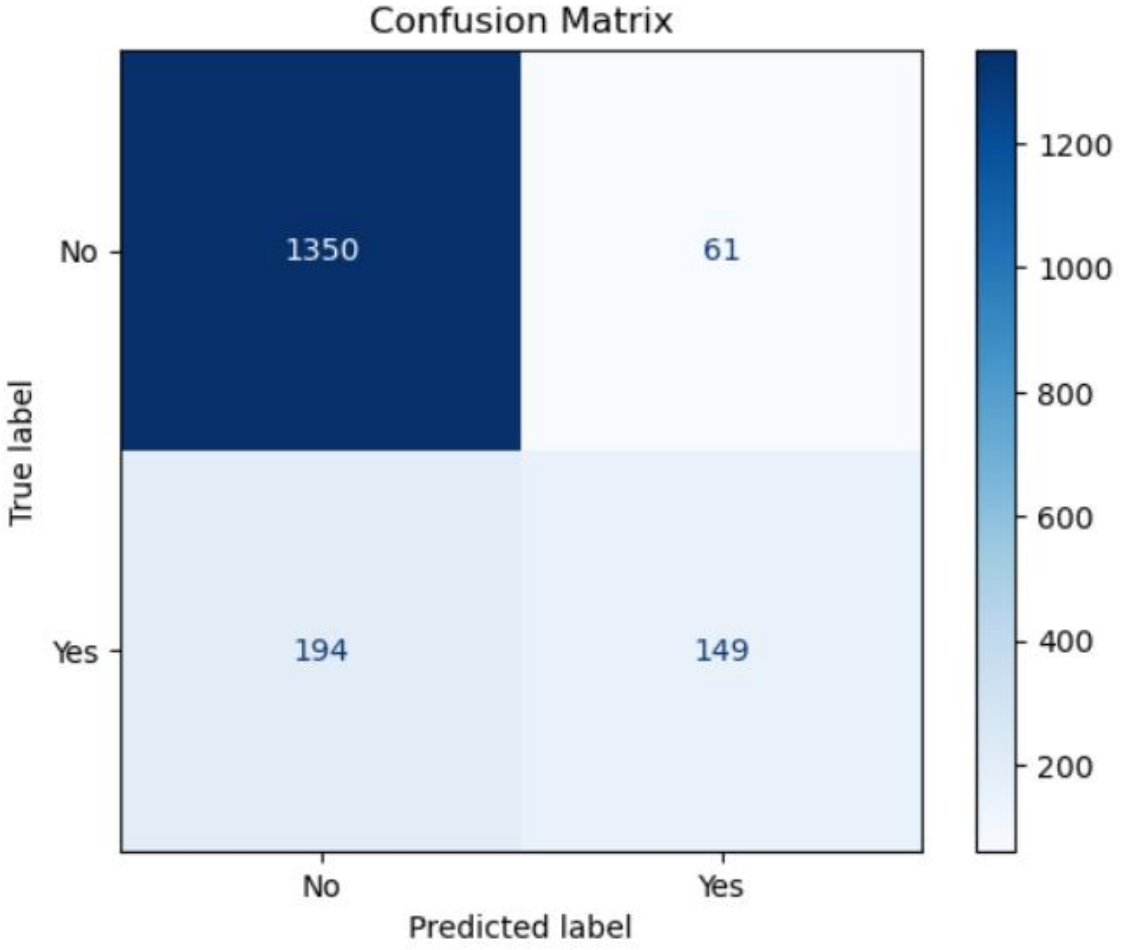
[74.55 8.06]

[30.01 9.87]]

Các điểm dữ liệu không tạo thành một đường thẳng rõ ràng, điều này cho thấy không có mối quan hệ tuyến tính trực tiếp giữa tuổi và mức độ giàu có.Tâm của các cụm có thể giúp ta xác định đặc điểm của từng phân khúc khách hàng. Ví dụ, một cụm có tâm ở vùng tuổi trẻ và mức độ giàu có cao có thể đại diện cho nhóm khách hàng trẻ, thành đạt.

PHẦN IV
THUẬT TOÁN PHÂN LOẠI

4.1 Confusion Matrix



Hình 10. Confusion Matrix

Biểu đồ trên là một Confusion Matrix, được sử dụng để đánh giá hiệu suất của mô hình phân loại trong việc dự đoán quyết định mua hàng của khách hàng. Confusion Matrix này bao gồm 4 ô, mỗi ô đại diện cho một kết quả dự đoán. Số 0 trên cả hai trục đại diện cho kết quả dự đoán quyết định không mua hàng, trong khi số 1 trên cả hai trục đại diện cho kết quả dự đoán quyết định mua hàng.

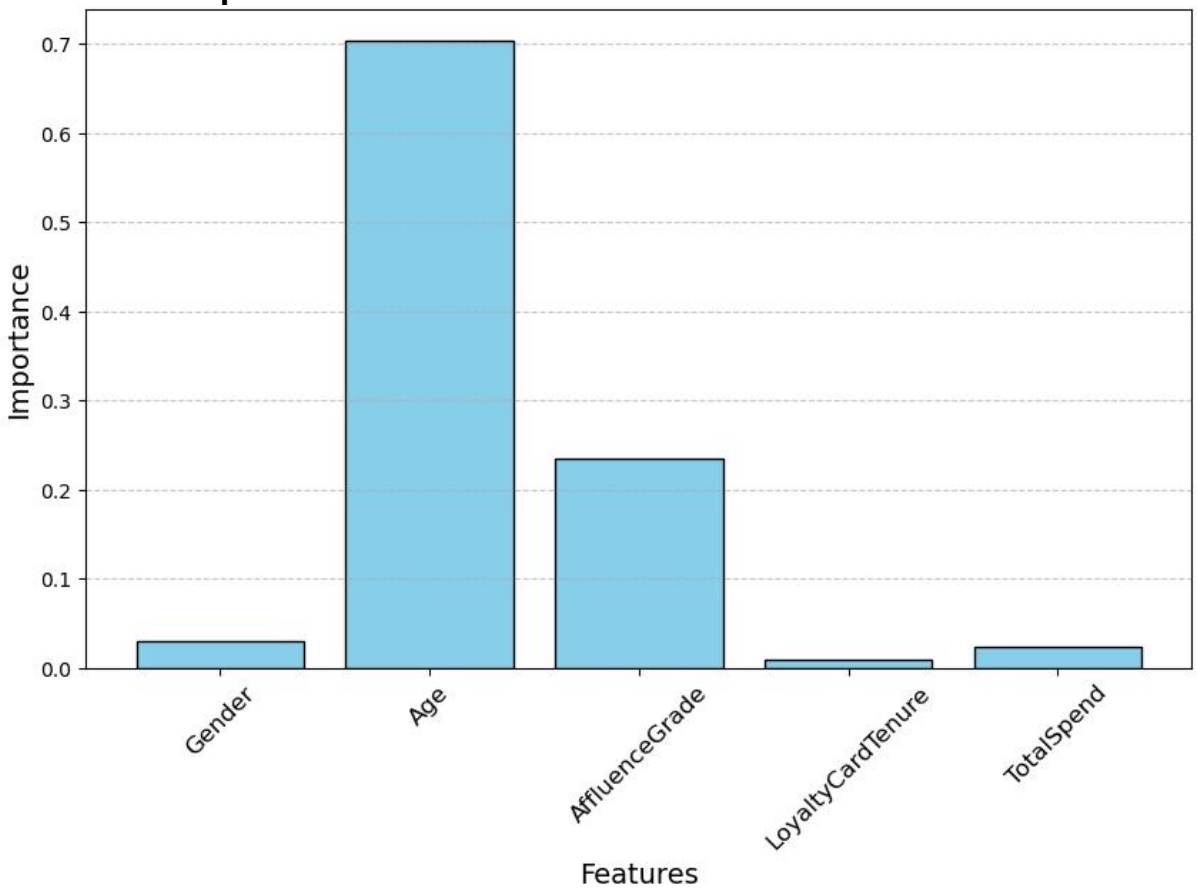
True Positive (TP): Có 149 mẫu thuộc lớp "Yes" và được dự đoán đúng là "Yes".

True Negative (TN): Có 1350 mẫu thuộc lớp "No" và được dự đoán đúng là "No".

False Positive (FP): Có 61 mẫu thuộc lớp "No" nhưng bị dự đoán sai là "Yes".

False Negative (FN): Có 194 mẫu thuộc lớp "Yes" nhưng bị dự đoán sai là "No".

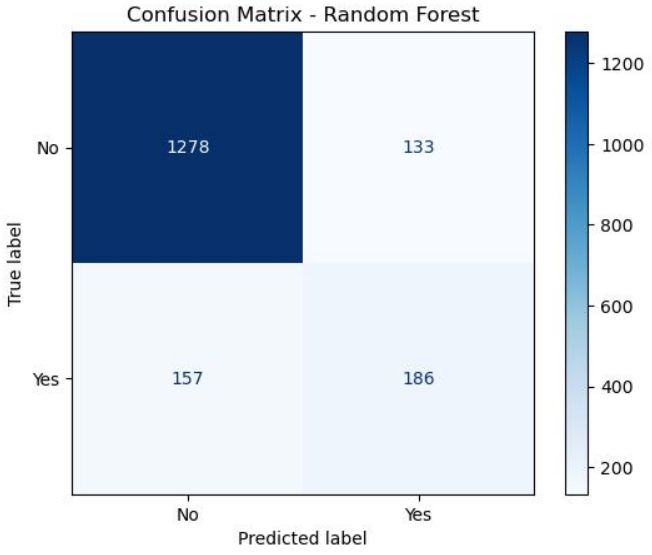
4.2 Features Importance



Hình 11. Biểu đồ thuộc tính quan trọng

PHẦN V
CÁC MÔ HÌNH

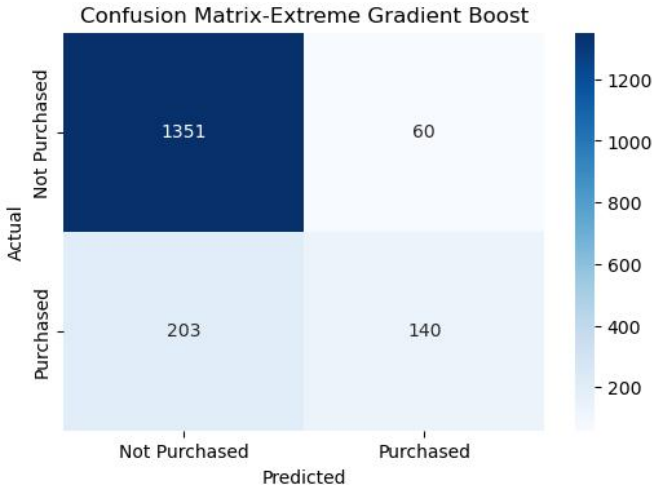
5.1 Mô hình Random Forest



Báo cáo phân loại - Random Forest:

	precision	recall	f1-score	support
No	0.89	0.91	0.90	1411
Yes	0.58	0.54	0.56	343
accuracy			0.83	1754
macro avg	0.74	0.72	0.73	1754
weighted avg	0.83	0.83	0.83	1754

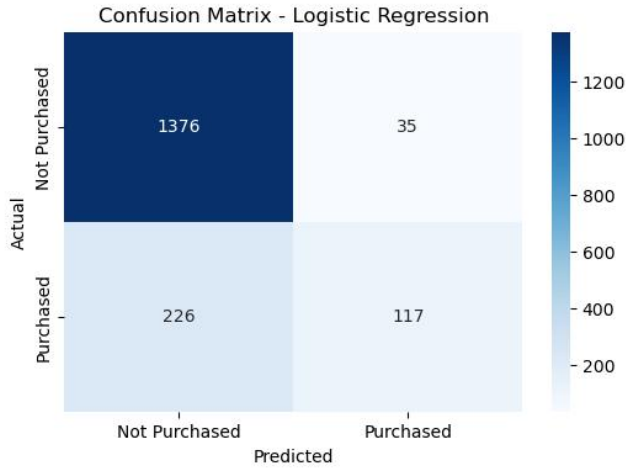
5.2 Mô hình Extreme Gradient Boost



Báo cáo phân loại Extreme Gradient Boost:

	precision	recall	f1-score	support
0	0.87	0.96	0.91	1411
1	0.70	0.41	0.52	343
accuracy			0.85	1754
macro avg	0.78	0.68	0.71	1754
weighted avg	0.84	0.85	0.83	1754

5.3 Mô hình Logistic Regression



Báo cáo phân loại Logistic Regression:

	precision	recall	f1-score	support
0	0.86	0.98	0.91	1411
1	0.77	0.34	0.47	343
accuracy			0.85	1754
macro avg	0.81	0.66	0.69	1754
weighted avg	0.84	0.85	0.83	1754

5.4 Đánh giá mô hình

Các mô hình có mức độ chính xác như sau:

Mô hình Random Forest (RF): 83%

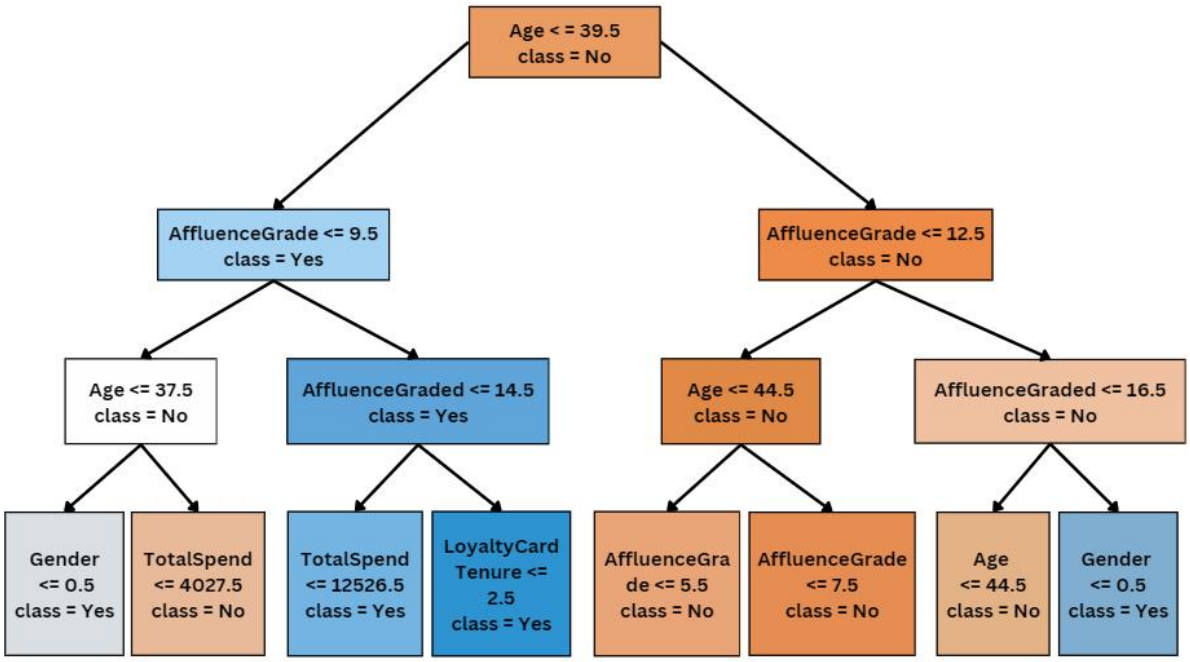
Mô hình Extreme Gradient Boost (EGB): 85%

Mô hình Logistic Regression (LR): 85%

Tuy mô hình EGB và LR có mức độ chính xác cao hơn mô hình RF, nhưng mức RECALL của YES tại 2 mô hình quá thấp so với mô hình RF nên sử dụng mô hình RF để làm mô hình mẫu cho cây quyết định.

PHẦN VI
CÂY QUYẾT ĐỊNH

6.1 Cây quyết định



Hình 12. Trực quan hóa cây quyết định

6.1.1 Độ chính xác cây quyết định

Để biết độ chính xác của cây quyết định, ta có thể sử dụng phương pháp xác thực chéo (cross-validation). Xác thực chéo là một kỹ thuật đánh giá hiệu suất của mô hình học máy trên dữ liệu mới.

Có nhiều phương pháp xác thực chéo khác nhau, nhưng một phương pháp phổ biến là xác thực chéo k-fold. Trong xác thực chéo k-fold, dữ liệu được chia thành k phần bằng nhau. Sau đó, mô hình được huấn luyện trên k-1 phần và thử nghiệm trên phần còn lại. Quá trình này được lặp lại k lần, mỗi lần sử dụng một phần khác nhau làm tập thử nghiệm.

```
Cross-Validation Scores: [0.87457241 0.86537365 0.86423274 0.8682259  0.85567598]
Độ chính xác trung bình: 0.8656161354927633
```

Độ chính xác của mô hình được tính bằng cách tính trung bình độ chính xác của mô hình trên k lần lặp lại. Trong trường hợp này, độ chính xác của cây quyết định là

86,56%, điều này có nghĩa là cây quyết định có thể dự đoán đúng khả năng mua hàng hóa hữu cơ ở 86,56% khách hàng.

6.1.2 Các yếu tố ảnh hưởng chính

AffluenceGrade: Được chọn làm thuộc tính phân nhánh đầu tiên, chứng minh rằng độ giàu có là yếu tố quan trọng nhất. Điều này hợp lý, vì khách hàng giàu có thường có khả năng chi tiêu cao hơn.

Age: Khách hàng trẻ tuổi có khả năng mua hàng cao hơn, đặc biệt khi họ nằm trong phân khúc có AffluenceGrade thấp.

LoyaltyCardTenure: Thời gian sử dụng thẻ khách hàng thân thiết cũng đóng vai trò quan trọng, thể hiện mức độ trung thành ảnh hưởng đến quyết định mua sắm.

6.1.3 Áp dụng thực tiễn

6.1.3.1 Khách hàng tiềm năng (class = Yes)

Những đặc điểm nổi bật của nhóm khách hàng có khả năng mua hàng:

a) AffluenceGrade ≤ 9.5 và Age ≤ 37.5: Khách hàng trẻ tuổi, có mức giàu có trung bình.

b) Tổng chi tiêu ≤ 22306.5, hoặc LoyaltyCardTenure ≤ 8.5. AffluenceGrade > 9.5 và TotalSpend > 15644.0: Mức chi tiêu cao, dù không phải là nhóm giàu nhất.

c) LoyaltyCardTenure > 7.5 và Age ≤ 37.5: Khách hàng trung thành và trẻ tuổi.

d) AffluenceGrade > 16.5, nhưng LoyaltyCardTenure > 7.5: Khách hàng giàu có và trung thành.

6.1.3.2 Khách hàng ít tiềm năng (class = No)

Những đặc điểm thường gặp của nhóm không mua hàng:

a) AffluenceGrade > 12.5 và Age ≤ 39.5: Tuổi trẻ nhưng thuộc nhóm khách hàng giàu có.

b) TotalSpend ≤ 7350.0: Khách hàng có mức chi tiêu rất thấp.

c) AffluenceGrade ≤ 5.5 hoặc Age > 44.5: Khách hàng thuộc nhóm ít giàu có hoặc cao tuổi.

6.2 Dự đoán quyết định

Từ thiết lập cây quyết định để dự đoán khả năng mua hàng của khách hàng, sử

dụng nguồn dữ liệu thực tế về gia đình của sinh viên như sau:

- Nữ, 50 tuổi, tài chính 16, năm sử dụng thẻ 3, tổng chi tiêu 16800

```
Nhập giới tính (0: nữ, 1: nam):  0
Nhập độ tuổi:  50
Nhập mức độ giàu có (AffluenceGrade):  16
Nhập số năm sử dụng thẻ khách hàng trung thành (LoyaltyCardTenure):  3
Nhập tổng chi tiêu (TotalSpend):  16800
Người này không có khả năng mua hàng (No).
```

- Nam, 65 tuổi, tài chính 21, năm sử dụng thẻ 1, tổng chi tiêu 22430

```
Nhập giới tính (0: nữ, 1: nam):  1
Nhập độ tuổi:  65
Nhập mức độ giàu có (AffluenceGrade):  21
Nhập số năm sử dụng thẻ khách hàng trung thành (LoyaltyCardTenure):  1
Nhập tổng chi tiêu (TotalSpend):  22430
Người này có khả năng mua hàng (Yes).
```

- Nữ, 59 tuổi, tài chính 10, năm sử dụng thẻ 7, tổng chi tiêu 14790

```
Nhập giới tính (0: nữ, 1: nam):  1
Nhập độ tuổi:  59
Nhập mức độ giàu có (AffluenceGrade):  10
Nhập số năm sử dụng thẻ khách hàng trung thành (LoyaltyCardTenure):  7
Nhập tổng chi tiêu (TotalSpend):  14790
Người này không có khả năng mua hàng (No).
```

PHẦN VII
TỔNG KẾT

Phần lớn người tham gia khảo sát là nữ giới, chủ yếu thuộc nhóm tuổi trung niên (50-70 tuổi). Đây là nhóm quan tâm nhiều nhất đến sức khỏe và có xu hướng mua hàng hữu cơ. Đa phần khách hàng có mức sống ổn định đến khá giả, phù hợp với chiến lược tiếp thị tập trung vào nhóm trung lưu và cao cấp.

Các yếu tố như độ giàu có (AffluenceGrade), tuổi (Age), và thời gian sử dụng thẻ khách hàng thân thiết (LoyaltyCardTenure) có ảnh hưởng đáng kể đến quyết định mua hàng hữu cơ. Mức độ tương quan giữa các biến với việc mua hàng hữu cơ được đánh giá, trong đó độ giàu có và tuổi là những yếu tố quan trọng.

Mô hình Random Forest được chọn làm mô hình chính với độ chính xác 83%, cân bằng tốt giữa khả năng phân loại và mức độ hồi tưởng (Recall). Cây quyết định giúp xác định rõ các đặc điểm khách hàng tiềm năng và đưa ra các khuyến nghị chiến lược. Doanh nghiệp nên tập trung tiếp thị vào nhóm khách hàng trung niên và giới trẻ với thông điệp nhấn mạnh sức khỏe, chất lượng sản phẩm, và lợi ích của hàng hóa hữu cơ. Phát triển các chương trình thẻ khách hàng thân thiết để tăng sự trung thành và cải thiện trải nghiệm mua sắm. Tỷ lệ khách hàng mua hàng hữu cơ còn thấp, điều này cho thấy cần đẩy mạnh giáo dục người tiêu dùng về lợi ích của sản phẩm hữu cơ. Tiếp tục mở rộng phạm vi khảo sát để có cái nhìn toàn diện hơn về thị trường.

Kết quả phân tích đã cung cấp thông tin giá trị cho doanh nghiệp trong việc định hình chiến lược tiếp thị và phát triển sản phẩm. Đồng thời, nó mở ra cơ hội cải thiện khả năng cạnh tranh trong ngành hàng hữu cơ.

BẢNG PHÂN CÔNG

Họ và Tên	MSSV	Nhiệm vụ	Mức độ hoàn thành
Vũ Ngọc Duy	2254052018	Thu thập dữ liệu, làm sạch. Nội dung phần II, IV, VI. Thiết kế PDF.	100%
Nguyễn Nguyên Ái Vân	2254052087	Thu thập dữ liệu. Nội dung phần I, II. Thiết kế Slide	100%
Nguyễn Phát	2254052059	Thu thập dữ liệu. Nội dung phần III, V. Thiết kế Slide	100%

TÀI LIỆU THAM KHẢO

1. Cây quyết định - https://trituenhantao.io/kien-thuc/decision-tree/

2. Bài báo cáo phân tích dữ liệu khóa K21 - https://colab.research.google.com/drive/1ZD_XXjgFLkesDO_QN2tJWv6uvWQDn03F#scrollTo=nwl279hyVi32&uniqifier=2

3. Bài báo cáo khai phá dữ liệu khóa K21 - https://github.com/TDNhatCuong/Analyze_Heart_Attack_Risk/blob/main/Code_Analyze_Condition%26Health_Of_People_At_Risk_Of_Heart_Attack.ipynb

4. Tìm hiểu về cây quyết định (Kênh Youtube) - https://www.youtube.com/c/M%C3%ACAIblog