



N H Ó M 3

Ứng dụng Big Data, Machine Learning và Cloud Computing

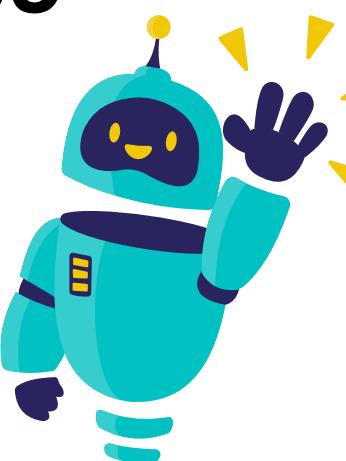
trong phân tích, dự đoán cụm khách
hang dựa trên lịch sử tín dụng, để hỗ
trợ ra quyết định kinh doanh



THÀNH VIÊN NHÓM 3



Đào Mỹ Duyên	K224141654
Trần Nguyên Diễm	K224141653
Nguyễn Ngọc Kim Nga	K224141676
Huỳnh Thị Thanh Nga	K224141675
Nguyễn Thị Ngọc Thanh	K224141692
Trần Ngọc Như Quỳnh	K224141689



BỐI CẢNH VÀ VẤN ĐỀ

- Ngành đang chuyển đổi số mạnh mẽ, ứng dụng **Big Data, Machine Learning, Cloud Computing**.
- Các ngân hàng lớn như **Vietcombank, BIDV** đầu tư khai thác dữ liệu để đáp ứng nhu cầu khách hàng số.
- Dữ liệu lớn giúp hiểu rõ hành vi, rủi ro và nhu cầu tài chính của khách hàng.
- Thách thức: **Dữ liệu khổng lồ, phức tạp**, khó phân tích thủ công, dễ bỏ lỡ cơ hội tối ưu dịch vụ.
- Giải pháp: Ứng dụng **Big Data & ML** để **tự động hóa phân tích, cá nhân hóa dịch vụ** và **nâng cao lợi thế cạnh tranh**.



- **Mục tiêu:** Xây dựng quy trình xử lý dữ liệu lớn và ứng dụng Machine Learning để phân cụm, dự đoán nhóm khách hàng, trực quan hóa kết quả và triển khai qua ứng dụng web thực tế.



- **Lợi ích:**

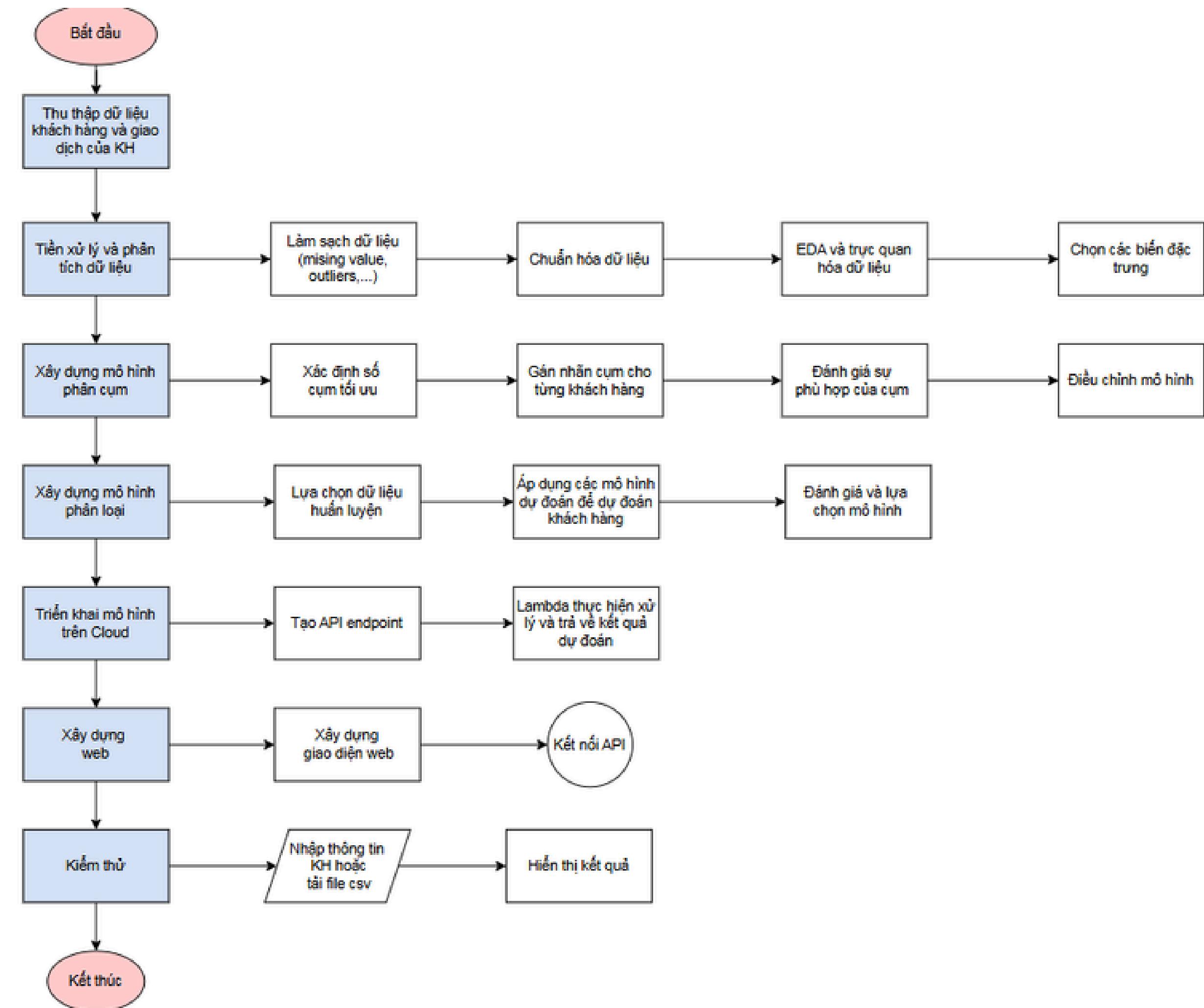
- Hiểu rõ hành vi khách hàng, hỗ trợ ra quyết định chiến lược.
- Cá nhân hóa sản phẩm, tăng hiệu quả marketing, giảm chi phí vận hành.
- Quản lý rủi ro tốt hơn, nhận diện nhóm khách hàng tiềm ẩn nguy cơ cao.

MỤC TIÊU VÀ LỢI ÍCH



Sơ đồ

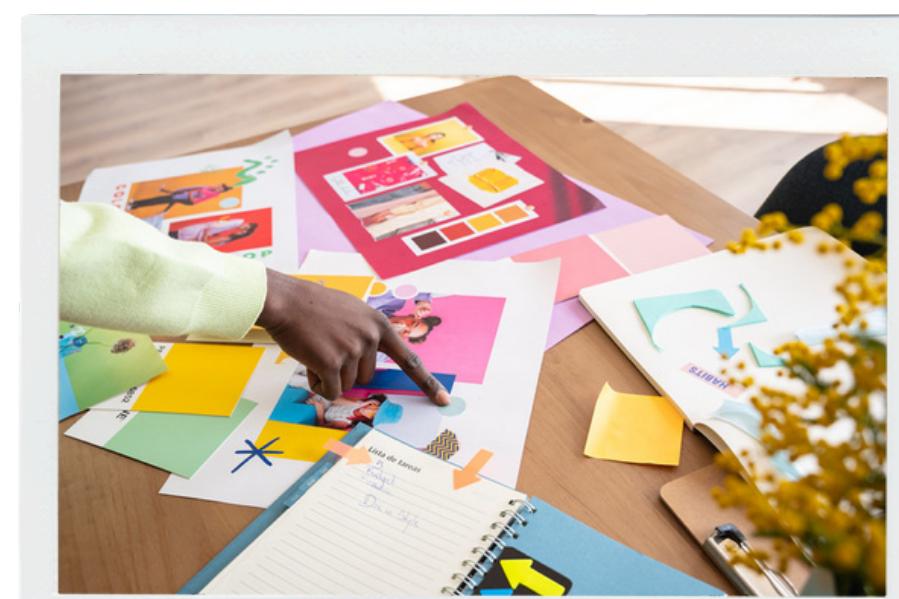
Quy trình phân tích và dự đoán cụm khách hàng ngân hàng



Nguồn dữ liệu

- Transactions: Lịch sử giao dịch của khách hàng (số tiền, loại giao dịch, lỗi, chip/online)
- Users: Thông tin cá nhân, thu nhập, điểm tín dụng, tổng nợ.
- Cards: Thông tin thẻ tín dụng, giới hạn tín dụng, số lượng thẻ, rủi ro dark web

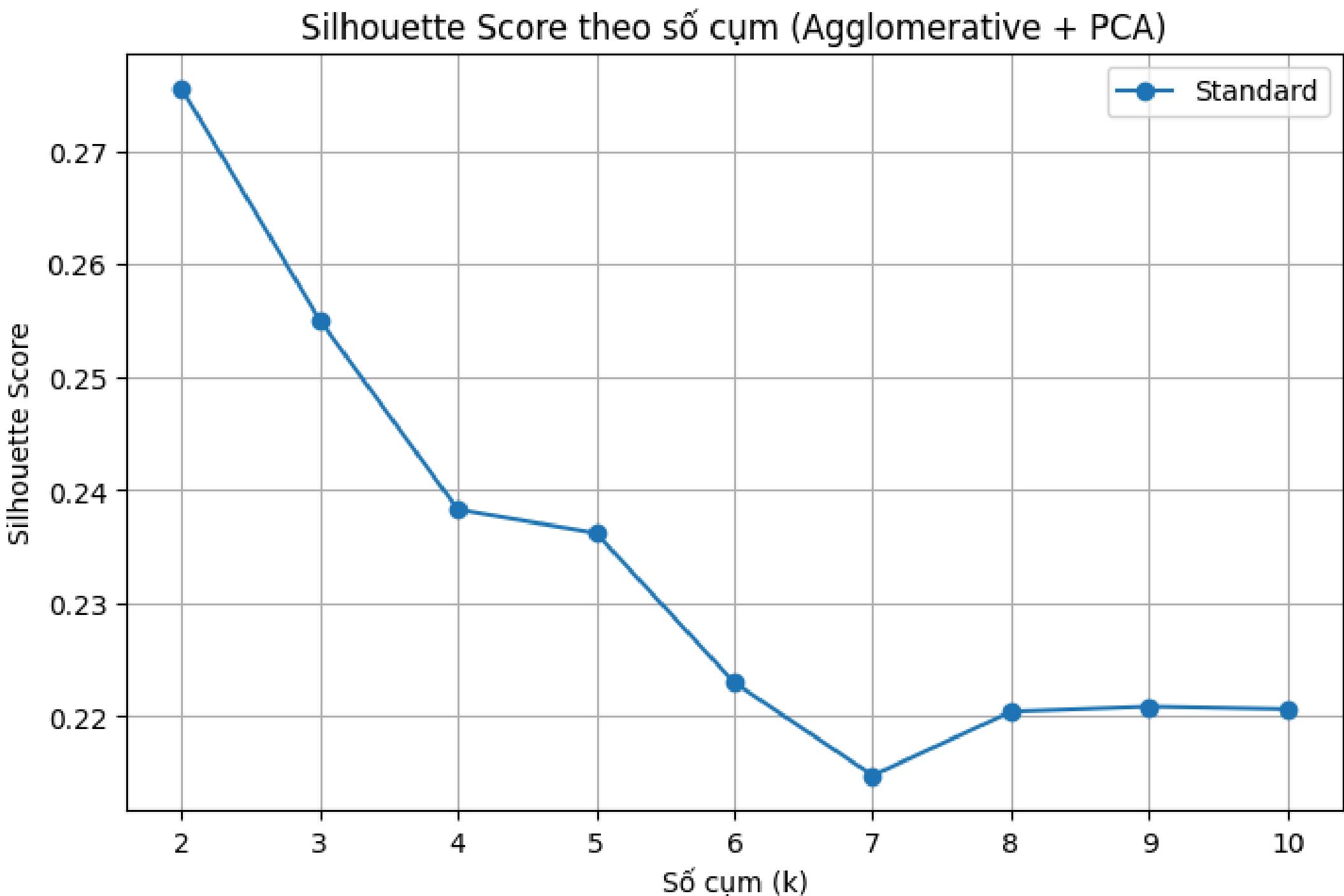
- 1 Google Drive, tải tự động qua thư viện gdown
- 2 Xử lý và tổng hợp dữ liệu
- 3 Phân tích mối tương quan
- 4 Thống kê mô tả, phân loại nhóm tuổi
- 5 Chuẩn bị dữ liệu cho phân cụm



CLUSTERING

AGGLOMERATIVE CLUSTERING

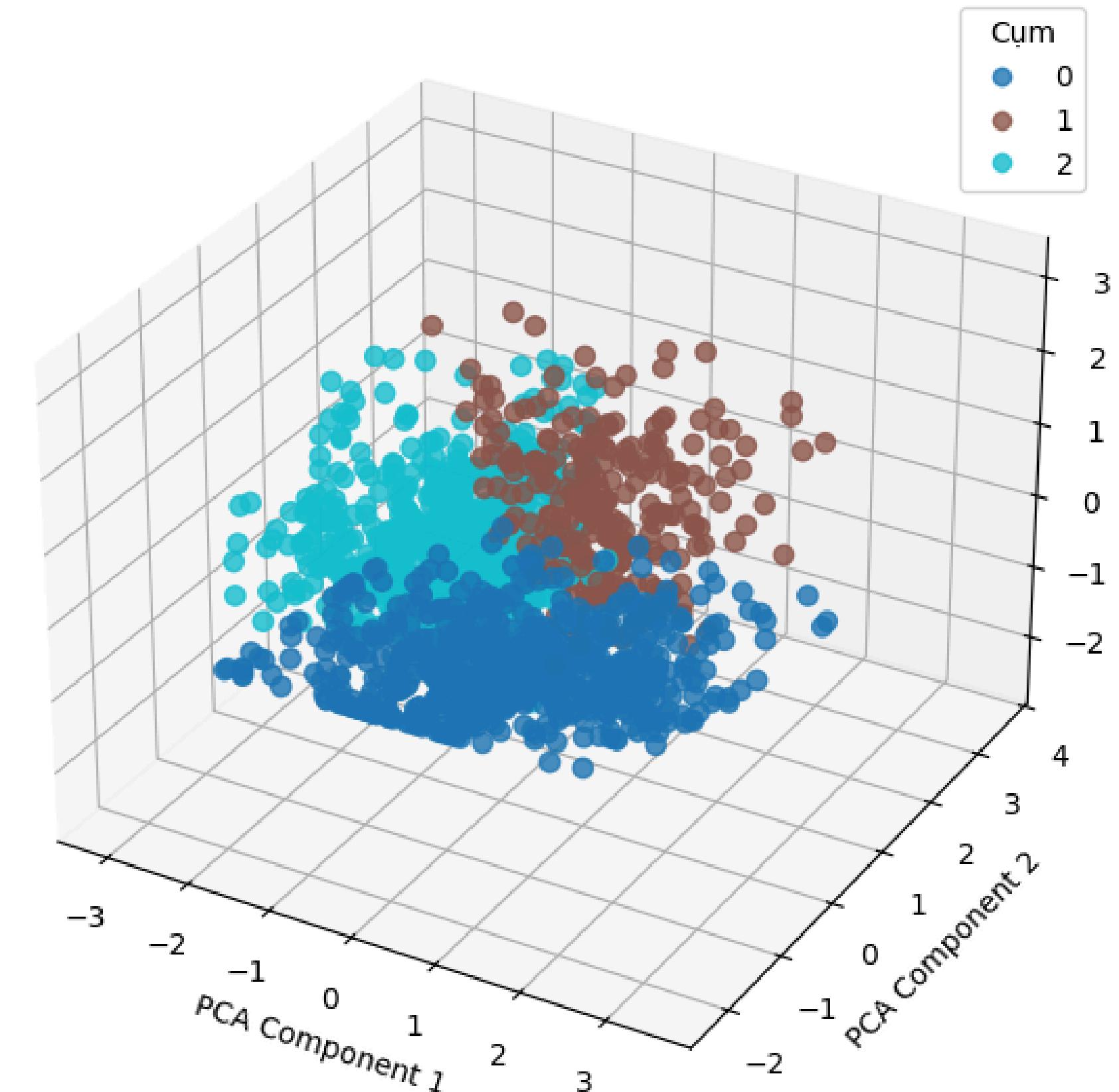
Tìm số cụm tối ưu cho Agglomerative Clustering. Silhouette Score được sử dụng làm chỉ số đánh giá. Dữ liệu được chuẩn hóa và giảm chiều PCA trước khi phân cụm để cải thiện hiệu suất.



AGGLOMERATIVE CLUSTERING

Mô hình Agglomerative Clustering được fit với số cụm tối ưu ($k = 3$). Silhouette score cho kết quả là 0.255, cho thấy chất lượng cụm. Sau đó, chúng ta tính toán đặc trưng trung bình cho mỗi cụm để hiểu các đặc điểm của chúng.

Phân cụm khách hàng (Agglomerative + PCA 3D) – $k=3$
Silhouette=0.255

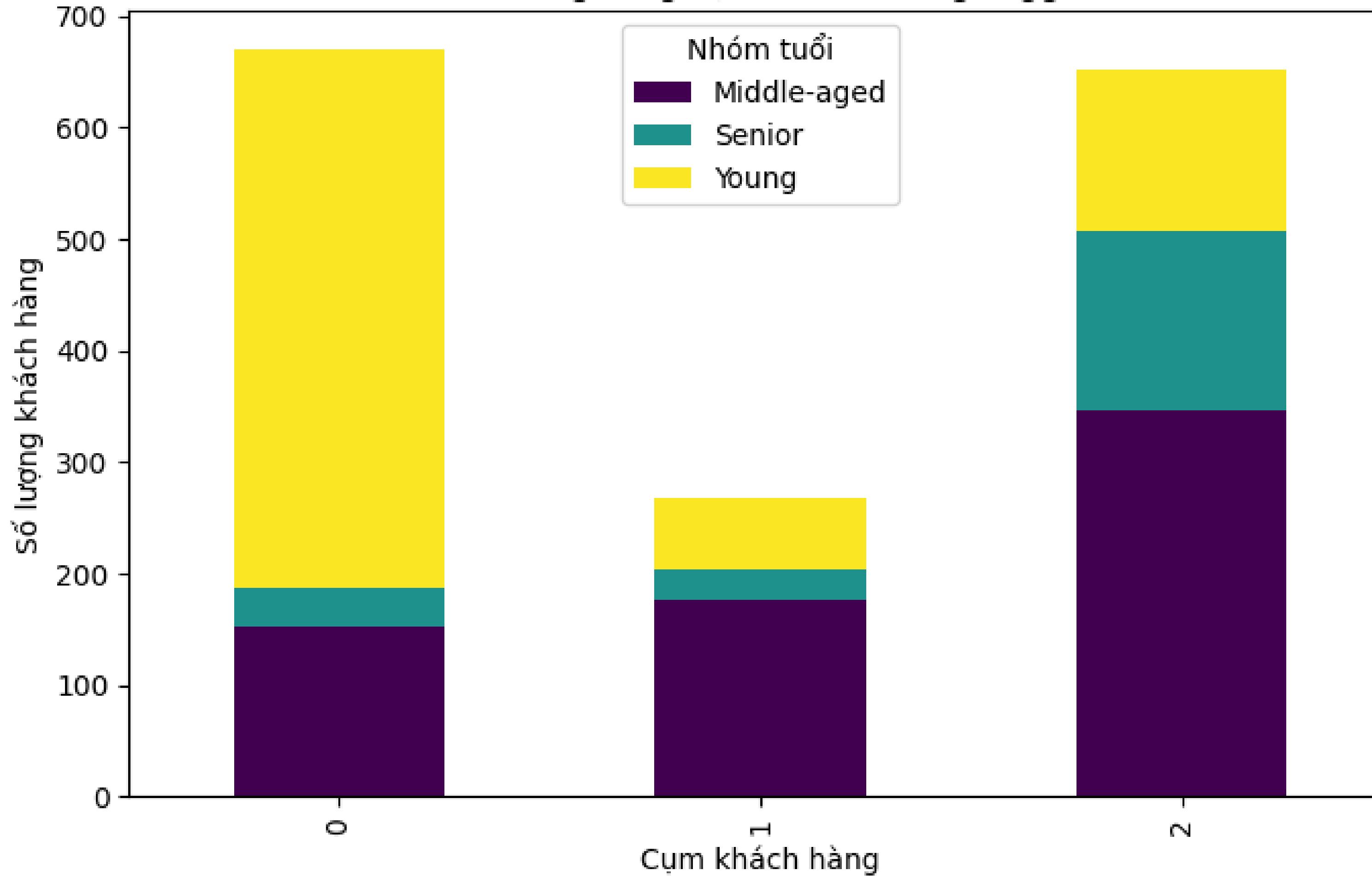


PHÂN TÍCH CỤM CON AGGLOMERATIVE



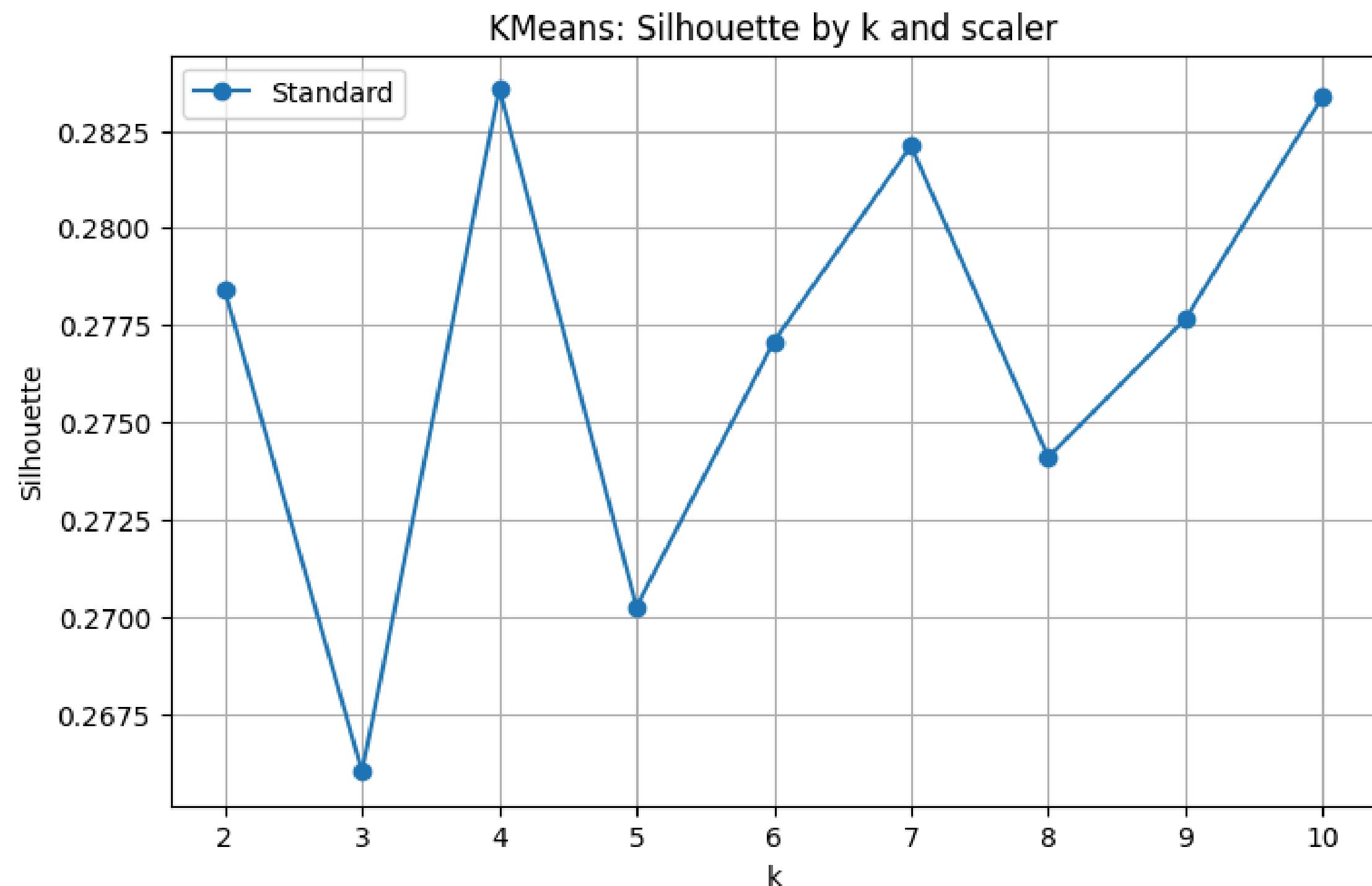
Agglomerative Clustering được kết hợp với thông tin tuổi để tạo 9 cụm con (3 cụm chính × 3 nhóm tuổi). Điều này cung cấp những hiểu biết chi tiết hơn về các phân khúc khách hàng. Phân phối tuổi trong mỗi cụm được trực quan hóa dưới dạng biểu đồ cột xếp chồng.

Phân bố nhóm tuổi trong từng cụm khách hàng (Agglomerative, k=3)





K-MEANS CLUSTERING



THỬ NGHIỆM THAM SỐ

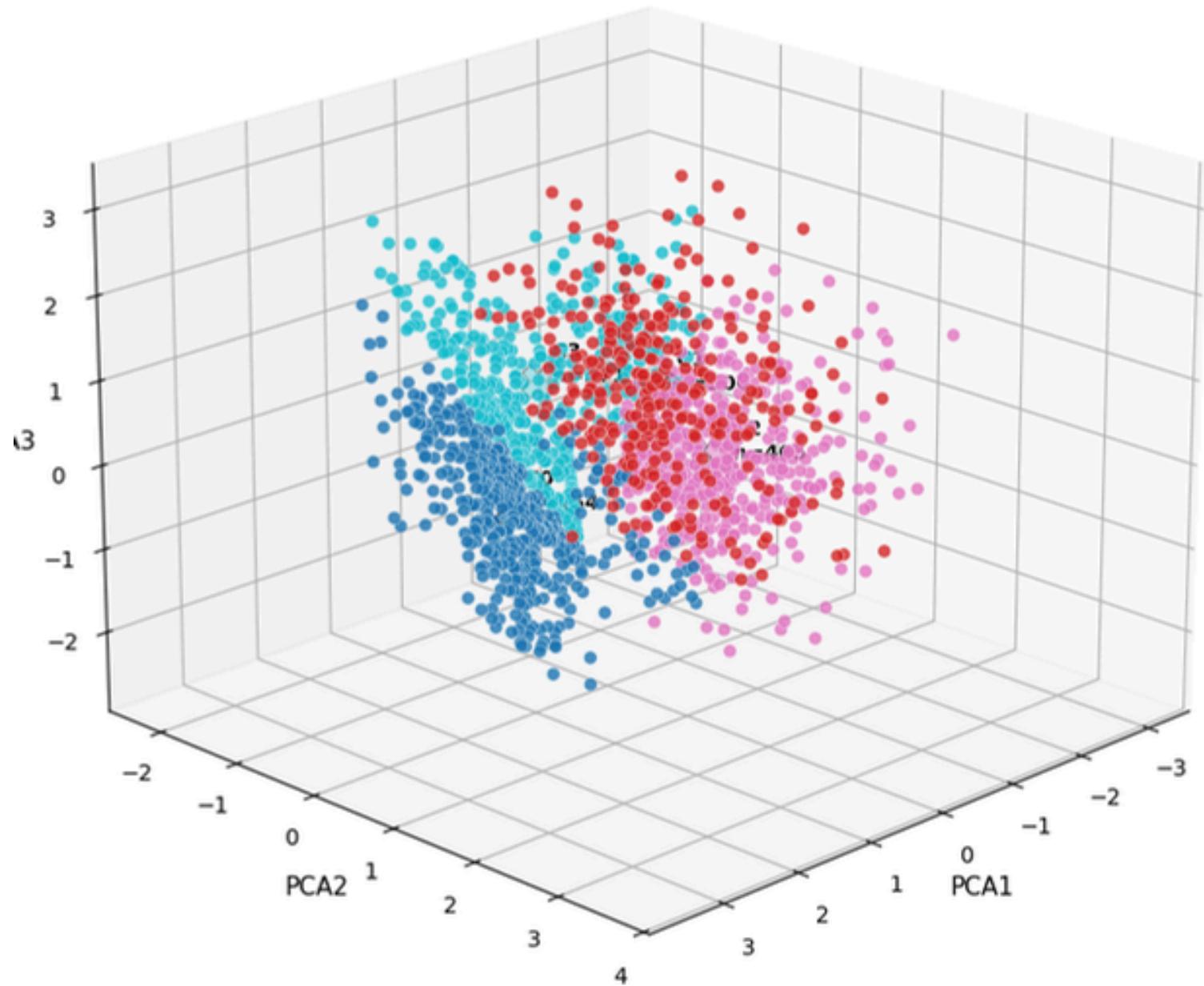
K-Means cũng trải qua tìm kiếm lướt để xác định số cụm tối ưu. Kết quả cho thấy k=4 là tối ưu nhất với Silhouette Score = 0.2836. Điều này cao hơn so với Agglomerative Clustering (0.255), cho thấy K-Means tạo ra các cụm tách biệt hơn.



K-MEANS CLUSTERING

KMeans clusters on PCA (k=4) — Silhouette=0.284

- Cluster 0 (n=454)
- Cluster 1 (n=300)
- Cluster 2 (n=465)
- Cluster 3 (n=371)
- ✖ Centroid



MÔ HÌNH CUỐI CÙNG

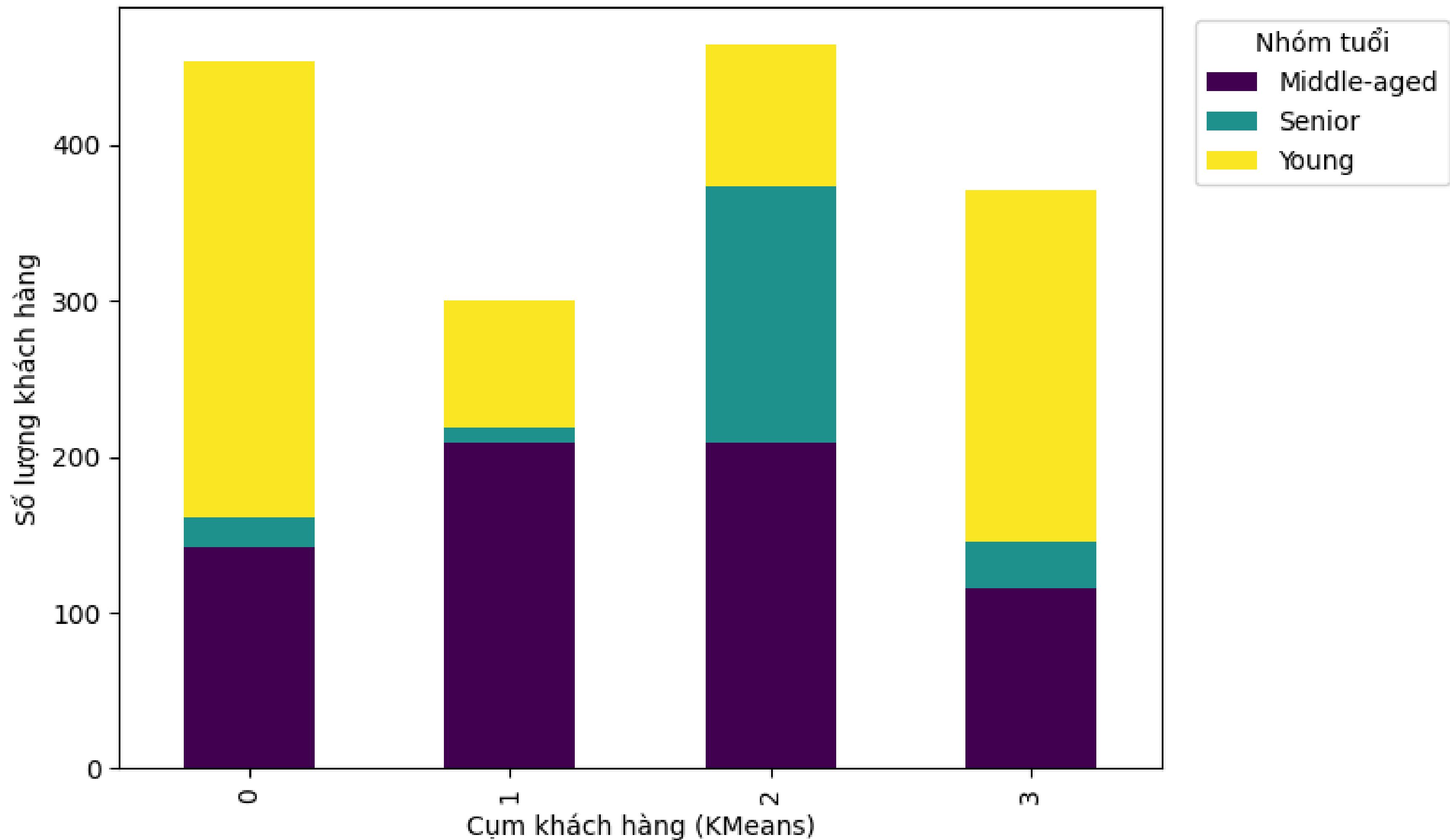
Mô hình K-Means cuối cùng được fit với $k = 4$. Phân phối khách hàng qua các cụm là khá cân bằng: Cụm 0 (454), Cụm 1 (300), Cụm 2 (465), Cụm 3 (371). Mỗi cụm có đặc trưng riêng về nơ, chi tiêu trung bình và điểm tín dụng.

PHÂN TÍCH CỤM CON K-MEANS



K-Means cũng được phân tích theo nhóm tuổi, tạo ra 12 cụm con (4 cụm chính \times 3 nhóm tuổi). Bảng tóm tắt cho thấy sự phân phối khác nhau của tuổi trong các cụm: Cụm 0 chủ yếu là trẻ (293), Cụm 1 là trung niên (209), Cụm 2 là người cao tuổi (164). Điều này cho thấy K-Means đã nắm bắt được các mẫu hành vi liên quan đến tuổi tác.

Phân bố nhóm tuổi trong từng cụm khách hàng (KMeans)



SO SÁNH HAI
PHƯƠNG PHÁP
PHÂN CỤM

AGGLOMERATIVE
&
KMEANS

Phương pháp phân cụm	Số cụm	Mô tả cụm (Debt – Spending – Credit)	Nhận xét / Điểm mạnh	Ứng dụng cho classification
Agglomerative	3	<p>Cluster 0: Nợ thấp (13–68k), chi tiêu rất thấp, credit ~710 → nhóm an toàn, ít rủi ro</p> <p>Cluster 1: Nợ cao (93–100k), chi tiêu ~44–48, credit thấp 650–680 → rủi ro cao</p> <p>Cluster 2: Nợ TB (12–45k), chi tiêu ~41–42, credit ~735 → nhóm cân bằng</p>	<ul style="list-style-type: none"> + Phân tách rõ ràng nợ cao vs nợ thấp + Dễ hiểu, đơn giản → Không tách được nhóm “VIP” hay chi tiết hơn 	Có thể dùng cho phân loại cơ bản, nhưng khó tách các nhóm đặc thù
KMeans	4	<p>Cluster 0: Nợ TB (20–74k), credit thấp 660–670, chi tiêu nhỏ → nhóm nợ vừa, tín dụng kém</p> <p>Cluster 1: Nợ rất cao (95–110k), credit khá 705, chi tiêu cao → nợ cao, rủi ro trung bình</p> <p>Cluster 2: Nợ thấp (11–32k), credit tốt 720+, chi tiêu TB → an toàn, low risk</p> <p>Cluster 3: Nợ TB (21–58k), credit cao 770–790, chi tiêu thấp → VIP, tín dụng cao</p>	<ul style="list-style-type: none"> + Tách được nhóm VIP và nhóm rủi ro + Credit score trải rộng → dễ gán nhãn “low/medium/high/premium” + Phân lớp đa chiều hơn 	Phù hợp cho classification/predict → Dễ gán nhãn cho khách hàng mới dựa trên debt/spending/credit

ĐÁNH GIÁ & ĐỀ XUẤT

Cluster	Tên cụm mô tả (Độ tuổi - Chi tiêu - Nợ)	Đặc trưng nổi bật	Insight chính	Gợi ý hành động (Action)	Giá trị kinh doanh	Rủi ro / Thách thức
0_Young	Trẻ - Chi tiêu rất thấp - Nợ cao	Tổng nợ cao (~74.7k), chi tiêu thấp (~0.6), tín dụng TB (~671)	Có tiềm năng tài chính nhưng chưa kích hoạt hành vi tiêu dùng	Tạo gói ưu đãi "chi tiêu đầu tiên", cashback nhỏ	Tiềm năng mở rộng doanh thu từ giới trẻ	Dễ rời bỏ nếu không thấy lợi ích trực tiếp
0_Middle-aged	Trung niên - Chi tiêu thấp - Nợ TB	Nợ TB (~68.9k), chi tiêu vừa (~7.5), tín dụng ~668	Ôn định, chi tiêu đều nhưng không cao	Duy trì ưu đãi định kỳ, tăng cường chăm sóc	Doanh thu ổn định	Khó mở rộng thêm chi tiêu
0_Senior	Lớn tuổi - Chi tiêu thấp - Nợ thấp	Nợ thấp (~20k), chi tiêu thấp (~2.4), tín dụng ~662	Trung thành, rủi ro thấp	Giới thiệu sản phẩm tiết kiệm, bảo hiểm	Ôn định, ít rủi ro	Tăng trưởng doanh thu hạn chế
1_Young	Trẻ - Chi tiêu cao - Nợ cao	Nợ cao (~97.5k), chi tiêu lớn (~50.3), tín dụng tốt (~706)	Khách trẻ năng động, thích chi tiêu, rủi ro thấp	Tập trung giữ chân qua chương trình tích điểm, cashback	Doanh thu lớn từ nhóm chi tiêu cao	Dễ mất nếu không có ưu đãi liên tục
1_Middle-aged	Trung niên - Chi tiêu cao - Nợ cao	Nợ cao (~95.7k), chi tiêu cao (~48.8), tín dụng tốt (~703)	Nhóm ôn định, chi tiêu mạnh	Ưu tiên chương trình khách hàng thân thiết	Tăng trưởng doanh thu ổn định	Có thể giảm chi tiêu nếu kinh tế biến động
1_Senior	Lớn tuổi - Chi tiêu cao - Nợ cao	Nợ cao (~109.6k), chi tiêu cao (~48.7), tín dụng tốt (~705)	Nhóm đáng tin cậy, có sức chi tiêu lớn	Giới thiệu thẻ hạng cao, ưu đãi cao cấp	Nguồn doanh thu lớn, rủi ro thấp	Dễ bị đối thủ cạnh tranh thu hút
2_Young	Trẻ - Chi tiêu TB - Nợ TB	Nợ TB (~32.9k), chi tiêu vừa (~41.6), tín dụng cao (~724)	Quản lý tài chính tốt, ít rủi ro	Đề xuất sản phẩm đầu tư an toàn, tiết kiệm sinh lời	Ôn định, an toàn, ít ngẫu	Biên lợi nhuận thấp
2_Middle-aged	Trung niên - Chi tiêu TB - Nợ TB	Nợ TB (~27.9k), chi tiêu ~43.7, tín dụng ~719	Khách hàng ổn định, trung thành	Bán chéo sản phẩm tài chính	Lợi nhuận bền vững	Chi tiêu khó tăng mạnh
2_Senior	Lớn tuổi - Chi tiêu TB - Nợ thấp	Nợ thấp (~11.3k), chi tiêu ~42.3, tín dụng ~713	Hành vi tài chính bền vững	Gợi ý gói bảo hiểm, tiết kiệm lãi cao	Doanh thu ổn định, rủi ro thấp	Ít tiềm năng mở rộng tiêu dùng
3_Young	Trẻ - Chi tiêu thấp - Nợ TB	Nợ TB (~58k), chi tiêu thấp (~3.6), tín dụng cao (~773)	Có năng lực tài chính tốt nhưng ít chi tiêu	Khuyến khích chi tiêu qua gói ưu đãi, đầu tư linh hoạt	Tiềm năng doanh thu cao nếu kích hoạt	Khó thuyết phục hành vi chi tiêu
3_Middle-aged	Trung niên - Chi tiêu TB - Nợ TB	Nợ TB (~53k), chi tiêu ~17.3, tín dụng rất cao (~789)	Rất đáng tin cậy, thận trọng tài chính	Giới thiệu sản phẩm đầu tư an toàn, sinh lời ổn định	Ít rủi ro, năng cao uy tín danh mục	Hành vi bảo thủ, khó kích thích chi tiêu
3_Senior	Lớn tuổi - Chi tiêu thấp - Nợ thấp	Nợ thấp (~21.6k), chi tiêu ~12.8, tín dụng cao (~783)	Khách hàng trung thành, tín dụng tốt	Ưu đãi nhẹ, tri ân khách hàng trung niên	Gửi vong hình ảnh thương hiệu uy tín	Khó mở rộng doanh thu từ chi tiêu

KẾT LUẬN

K-Means được lựa chọn làm mô hình phân cụm chính vì nó cung cấp Silhouette score cao hơn và phân biệt các phân khúc khách hàng tốt hơn, đặc biệt là nhóm VIP (Cụm 3).

===== So sánh các mô hình Classification (KMeans) =====

Mô hình	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.973	0.974	0.973	0.972
SVM	0.958	0.961	0.958	0.958
Random Forest	0.954	0.956	0.954	0.954
XGBoost	0.954	0.955	0.954	0.954
Gradient Boosting	0.948	0.950	0.948	0.947
Decision Tree	0.929	0.932	0.929	0.928

Logistic Regression vượt trội với độ chính xác cao nhất (0.973) và F1-score (0.972). Điều này cho thấy rằng mối quan hệ giữa các đặc trưng và các cụm con là khá tuyến tính.

Xem trước các dự đoán

Hầu hết các dự đoán là chính xác, và các lỗi thường xảy ra với các cụm con tương tự (ví dụ 0_Young và 0_Middle-aged có đặc trưng tương tự nhau).

	age	total_debt	avg_amount	credit_score	Thực tế	Dự đoán	Correct
0	46	81752.0	0.000000	615	0_Middle-aged	0_Middle-aged	True
1	39	36199.0	48.909705	763	2_Young	2_Young	True
2	31	103896.0	0.000000	651	0_Young	0_Young	True
3	58	65836.0	0.000000	696	0_Middle-aged	0_Middle-aged	True
5	38	97252.0	0.000000	697	0_Young	0_Young	True
7	39	114299.0	0.000000	629	0_Young	0_Young	True
8	39	66737.0	0.000000	685	0_Young	0_Young	True
9	64	107130.0	77.995284	728	1_Middle-aged	1_Middle-aged	True
10	21	76915.0	0.000000	607	0_Young	0_Young	True
11	39	72318.0	28.910111	701	1_Young	1_Young	True



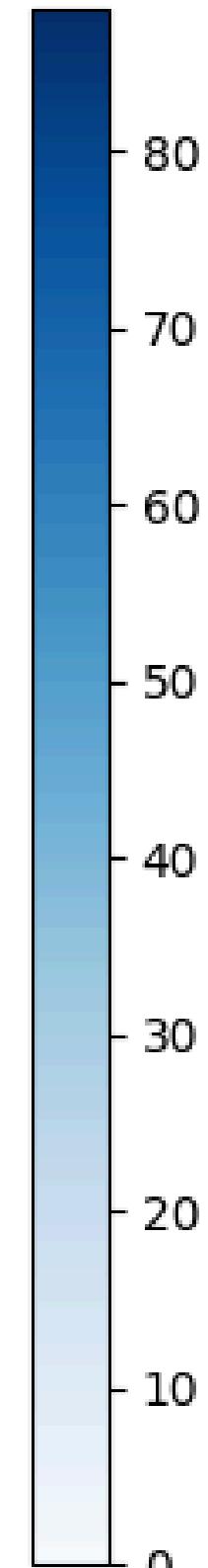
CONFUSION MATRIX

- Ma trận nhầm lẫn cho thấy hầu hết các mẫu nằm trên đường chéo chính (dự đoán đúng).

=> Export mô hình tốt nhất lên S3

Confusion Matrix - Logistic Regression

True label	Predicted label											
	0_Middle-aged	0_Senior	0_Young	1_Middle-aged	1_Senior	1_Young	2_Middle-aged	2_Senior	2_Young	3_Middle-aged	3_Senior	3_Young
0_Middle-aged	41	0	0	2	0	0	0	0	0	0	0	0
0_Senior	0	5	0	0	0	0	0	1	0	0	0	0
0_Young	0	0	88	0	0	0	0	0	0	0	0	0
1_Middle-aged	0	0	0	63	0	0	0	0	0	0	0	0
1_Senior	0	0	0	1	2	0	0	0	0	0	0	0
1_Young	0	0	0	0	0	23	0	0	1	0	0	0
2_Middle-aged	1	0	0	2	0	0	60	0	0	0	0	0
2_Senior	0	0	0	0	0	0	0	49	0	0	0	0
2_Young	0	0	0	0	0	0	0	0	27	0	0	0
3_Middle-aged	1	0	0	2	0	0	1	0	0	31	0	0
3_Senior	0	0	0	0	0	0	0	1	0	0	8	0
3_Young	0	0	0	0	0	0	0	0	0	0	0	67



XÂY DỰNG API DỰ ĐOÁN

Mục tiêu của giai đoạn này là cho phép người dùng gửi dữ liệu đầu vào và nhận lại kết quả phân cụm dự đoán thông qua một endpoint công khai

1

Thiết lập
môi trường
và phân
quyền cho
Lambda qua
AWS IAM

2

Tạo và cấu
hình hàm
Lambda

3

Triển khai
mã xử lý mô
hình

4

Xây dựng API
Gateway

5

Kiểm thử
API dự đoán
bằng
Postman



N H Ó M 3

Demo website

BIG DATA



ƯU ĐIỂM

- Kết hợp hiệu quả **Big Data, Machine Learning, Cloud Computing** và **Web Application**.
- **Dự đoán nhanh cụm khách hàng mới** mà không cần phân cụm lại toàn bộ dữ liệu → tiết kiệm thời gian và chi phí.
- Ứng dụng web **trực quan, dễ sử dụng**, hỗ trợ hai chế độ:
 - Dự đoán **từng khách hàng** (nhập trực tiếp).
 - Dự đoán **hàng loạt** (tải tệp CSV).
- **Đáp ứng linh hoạt** nhu cầu quản lý và phân tích dữ liệu trong ngân hàng.
- Có **tính ứng dụng cao và khả năng mở rộng**, phù hợp thực tiễn ngành tài chính – ngân hàng.



NHƯỢC ĐIỂM



- **Đặc trưng đầu vào còn ít**, chủ yếu gồm tuổi, thu nhập, điểm tín dụng, mức nợ → chưa phản ánh đầy đủ hành vi tài chính.
- **Dữ liệu đã qua xử lý hoặc tổng hợp**, không phải dữ liệu gốc theo từng giao dịch → giảm độ chi tiết của phân tích.
- **Giới hạn về nguồn dữ liệu thực tế**, chỉ sử dụng dữ liệu mô phỏng do không có quyền truy cập dữ liệu ngân hàng thật.
- **Thời gian triển khai ngắn**, chưa mở rộng sang **Deep Learning** hoặc **phân tích thời gian thực**.
- Tuy nhiên, dự án vẫn **đạt được mục tiêu đề ra**, chứng minh tính **khả thi** của **ML** và **Big Data** trong phân nhóm khách hàng tài chính.



THANK YOU

Presented By:
NHÓM 3

Date:
07/11/2025

Drive tổng hợp:

https://drive.google.com/drive/folders/14g4A6joWieCgntmVanwCaFKh8-3YYtdj?usp=drive_link