|  |
| --- |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT**  **---------------------------------------** |
|  |
|  |
| **ĐỒ ÁN KHOA HỌC MÁY TÍNH** |
| **NGÀNH CÔNG NGHỆ THÔNG TIN** |
| **TÊN ĐỀ TÀI:** **Phân tích hành vi khách hàng thông qua dữ liệu Bank-full** |
| **Hà Nội - Năm 2025** |

BỘ GIÁO DỤC VÀ ĐÀO TẠO TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT

**ĐỒ ÁN KHOA HỌC MÁY TÍNH**

**CHUYÊN NGÀNH: KHOA HỌC MÁY TÍNH**

**GIÁO VIÊN HƯỚNG DẪN: Nguyễn Thị Hải Yến**

**SINH VIÊN THỰC HIỆN: Nguyễn Văn Dũng**

|  |  |
| --- | --- |
| BỘ MÔN KHOA HỌC MÁY TÍNH | LỚP: KHMT ỨNG DỤNG DCCTCT66\_07E |

**Hà Nội – Năm 2025**

Mục Lục

[Lời cảm ơn 5](#_Toc191148316)

[Danh mục bảng biểu 5](#_Toc191148317)

[Danh mục hình ảnh 5](#_Toc191148318)

[Mở đầu 5](#_Toc191148319)

[1. Lý do chọn đề tài 5](#_Toc191148320)

[2. Mục tiêu của đề tài 5](#_Toc191148321)

[3. Phạm vi của đề tài 5](#_Toc191148322)

[4. Nội dung nghiên cứu 5](#_Toc191148323)

[5. Bố cục của đề tài 5](#_Toc191148324)

[CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI 5](#_Toc191148325)

[1.1. Lý do chọn đề tài 5](#_Toc191148326)

[1.2. Mục đích của đề tài 5](#_Toc191148327)

[1.3. Nội dung nghiên cứu 5](#_Toc191148328)

[1.4. Mục tiêu nghiên cứu 5](#_Toc191148329)

[1.5. Phạm vi nghiên cứu 5](#_Toc191148330)

[1.6. Cơ sở khoa học và thực tiễn của đề tài 5](#_Toc191148331)

[1.7. Bố cục đồ án 5](#_Toc191148332)

[CHƯƠNG 2: Cơ sở lý thuyết 5](#_Toc191148333)

[2.1. Cơ sở lý luận 5](#_Toc191148334)

[2.1.1. Khái niệm và đặc điểm của hành vi khách hàng trong ngành ngân hàng 5](#_Toc191148335)

[2.1.2. Ứng dụng hệ thống học máy để phân tích hành vi khách hàng thông qua dữ liệu được truyền vào 5](#_Toc191148336)

[2.1.3. Lịch sử và sự phát triển của các mô hình học máy trong phân tích dữ liệu 5](#_Toc191148337)

[2.1.4. Cấu trúc và cách hoạt động của các mô hình phân loại khách hàng trong học máy 5](#_Toc191148338)

[2.2. Các mô hình học máy trong phân tích hành vi khách hàng 5](#_Toc191148339)

[2.2.1. Khái niệm về học máy 5](#_Toc191148340)

[2.2.2. Các mô hình học máy phổ biến 6](#_Toc191148341)

[2.2.3. Ứng dụng mô hình học máy trong phân loại hành vi khách hàng 6](#_Toc191148342)

[2.3. Ngôn ngữ lập trình 6](#_Toc191148343)

[2.3.1. Thư viện Numpy 6](#_Toc191148344)

[2.3.2. Thư viện Pandas 6](#_Toc191148345)

[2.3.3. Thư viện Matplotlib 6](#_Toc191148346)

[2.3.4. Thư viện Scikit-learn (Sklearn) 6](#_Toc191148347)

[CHƯƠNG 3: MÔ TẢ VÀ PHÂN TÍCH HỆ THỐNG BÀI TOÁN PHÂN LOẠI BẰNG HỆ THỐNG HỌC MÁY 6](#_Toc191148348)

[3.1. Bối cảnh nghiên cứu 6](#_Toc191148349)

[3.1.1. Đối tượng nghiên cứu 6](#_Toc191148350)

[3.1.2. Dữ liệu 6](#_Toc191148351)

[3.2. Tiền xử lý dữ liệu 6](#_Toc191148352)

[3.2.1. Cài đặt các thư viện liên quan 6](#_Toc191148353)

[3.2.2. Tải bộ dữ liệu 6](#_Toc191148354)

[3.2.3. Trực quan hóa dữ liệu 6](#_Toc191148355)

[3.2.4. Tăng cường dữ liệu 6](#_Toc191148356)

[3.2.5. Mô hình cây quyết định 6](#_Toc191148357)

[3.2.6. Mô hình phân cụm 6](#_Toc191148358)

[3.2.7. Mô hình phân lớp 6](#_Toc191148359)

[CHƯƠNG 4: LỰA CHỌN MÔ HÌNH CHO BÀI TOÁN PHÂN LOẠI HÀNH VI KHÁCH HÀNG 6](#_Toc191148360)

[4.1 Lựa chọn mô hình 6](#_Toc191148361)

[4.2 Thiết kế hệ thống 6](#_Toc191148362)

[4.3 Kết quả thực nghiệm 6](#_Toc191148363)

[4.4 Phân tích sau khi lựa chọn và huấn luyện mô hình 6](#_Toc191148364)

[KẾT LUẬN 6](#_Toc191148365)

[TÀI LIỆU THAM KHẢO 7](#_Toc191148366)

# Lời cảm ơn

# Danh mục bảng biểu

# Danh mục hình ảnh

# Mở đầu

1. Lý do chọn đề tài

Trong thời đại số hóa, dữ liệu đóng vai trò quan trọng trong việc phân tích và dự đoán hành vi khách hàng. Ngành ngân hàng là một trong những lĩnh vực có khối lượng dữ liệu khách hàng lớn, bao gồm thông tin cá nhân, tài chính và lịch sử giao dịch. Việc phân tích hành vi khách hàng thông qua dữ liệu thực tế giúp các tổ chức tài chính đưa ra các quyết định kinh doanh hiệu quả hơn, tối ưu hóa chiến lược marketing và nâng cao trải nghiệm khách hàng. Khách hàng luôn giữ vai trò quan trọng trong thành công của ngân hàng. Với tình hình cạnh tranh gay gắt trong ngành ngân hàng hiện nay, việc hiểu rõ hành vi và nhu cầu tiêu dùng của khách hàng là yếu tố then chốt giúp ngân hàng phát triển và duy trì lợi thế cạnh tranh. Đồng thời, sự phát triển của công nghệ dữ liệu và trí tuệ nhân tạo tạo điều kiện thuận lợi để ngân hàng có thể phân tích và dự đoán chính xác hơn nhu cầu của khách hàng. Ngoài ra, hệ thống ngân hàng đang đối mặt với sự cạnh tranh từ các công ty tài chính công nghệ.

Do đó, việc khai thác dữ liệu khách hàng để cá nhân hóa dịch vụ và tối ưu hóa các chiến dịch tiếp thị là xu hướng tất yếu nhằm tăng cường sự hài lòng và giữ chân khách hàng. Từ những lý do trên, việc lựa chọn đề tài "Phân tích hành vi khách hàng thông qua dữ liệu Bank-full" là cần thiết và có ý nghĩa thực tiễn, góp phần nâng cao hiệu quả hoạt động kinh doanh của ngân hàng.

1. Mục tiêu của đề tài

* Nghiên cứu các yếu tố ảnh hưởng đến hành vi khách hàng dựa trên dữ liệu ngân hàng.
* Xác định các nhóm khách hàng tiềm năng có khả năng tham gia các dịch vụ tài chính.
* Đề xuất các chiến lược marketing dựa trên kết quả phân tích dữ liệu.
* Ứng dụng các phương pháp phân tích dữ liệu để trực quan hóa và diễn giải thông tin.

1. Phạm vi của đề tài

Phạm vi nghiên cứu tập trung vào tập dữ liệu "bank-full.csv", bao gồm thông tin về khách hàng ngân hàng, các chiến dịch marketing và kết quả tiếp cận khách hàng. Đề tài sử dụng phương pháp phân tích dữ liệu, trực quan hóa thông tin và áp dụng các mô hình thống kê để phân tích hành vi khách hàng.

1. Nội dung nghiên cứu

* Thu thập và làm sạch dữ liệu từ tập "bank-full.csv".
* Phân tích thống kê để nhận diện các xu hướng trong hành vi khách hàng.
* Áp dụng mô hình phân tích dữ liệu để xác định các nhóm khách hàng tiềm năng.
* Đưa ra các đề xuất nhằm tối ưu chiến lược tiếp cận khách hàng

1. Bố cục của đề tài

* Bố cục đề tài gồm các chương sau:
* Chương 1: Tổng quan về đề tài
* Chương 2: Phân tích dữ liệu khách hàng
* Chương 3: Kết quả và đề xuất chiến lược
* Chương 4: Kết luận và hướng phát triển

# CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI

## 1.1. Lý do chọn đề tài

Sự phát triển mạnh mẽ của công nghệ dữ liệu và trí tuệ nhân tạo đã giúp các doanh nghiệp, đặc biệt là ngành ngân hàng, khai thác thông tin khách hàng một cách hiệu quả. Việc phân tích dữ liệu khách hàng không chỉ giúp các ngân hàng hiểu rõ nhu cầu, hành vi của khách hàng mà còn hỗ trợ trong việc tối ưu hóa chiến lược kinh doanh, nâng cao trải nghiệm khách hàng và cải thiện chất lượng dịch vụ.

Hiện nay, ngành ngân hàng đang đối mặt với sự cạnh tranh gay gắt từ cả các ngân hàng truyền thống lẫn các công ty fintech. Để giữ vững vị thế, các ngân hàng cần tận dụng dữ liệu khách hàng nhằm dự đoán xu hướng tiêu dùng, phân nhóm khách hàng và đưa ra các chương trình tiếp thị phù hợp.

Phân tích hành vi khách hàng thông qua dữ liệu thực tế là một phương pháp giúp các ngân hàng tối ưu hóa chiến lược tiếp cận khách hàng, cải thiện tỷ lệ chuyển đổi và tăng cường mức độ hài lòng của khách hàng. Do đó, việc nghiên cứu và ứng dụng phân tích dữ liệu vào lĩnh vực ngân hàng có ý nghĩa quan trọng và cần thiết.

## 1.2. Mục đích của đề tài

* Xác định các yếu tố quan trọng ảnh hưởng đến hành vi khách hàng.
* Phân loại khách hàng theo các đặc điểm quan trọng để tối ưu chiến lược tiếp thị.
* Đề xuất các chiến lược kinh doanh phù hợp dựa trên phân tích dữ liệu thực tế.

## 1.3. Nội dung nghiên cứu

Đề tài tập trung vào phân tích dữ liệu khách hàng của ngân hàng thông qua tập dữ liệu "bank-full.csv", với các nội dung chính như:

* Phân tích nhân khẩu học của khách hàng (độ tuổi, nghề nghiệp, tình trạng hôn nhân, v.v.).
* Đánh giá tác động của các chiến dịch marketing đến quyết định của khách hàng.
* Phân loại khách hàng theo các yếu tố tài chính và hành vi tiêu dùng.

🡪 Từ đó đưa ra các giải pháp giúp chiến dịch tiếp thị dựa trên hành vi của khách hàng hiệu quả hơn.

## 1.4. Mục tiêu nghiên cứu

* Hiểu rõ hơn về hành vi khách hàng dựa trên dữ liệu thực tế.
* Phát hiện các mô hình hành vi tiêu biểu của khách hàng.
* Đưa ra các dự đoán về khả năng khách hàng tham gia vào các dịch vụ tài chính.

## 1.5. Phạm vi nghiên cứu

Phạm vi nghiên cứu bao gồm:

* Dữ liệu: Sử dụng tập dữ liệu "bank-full.csv" với hơn 45.000 khách hàng.
* Thời gian: Phân tích dựa trên dữ liệu trong một khoảng thời gian nhất định.
* Phương pháp: Sử dụng phân tích thống kê, trực quan hóa dữ liệu và mô hình dự đoán.

## 1.6. Cơ sở khoa học và thực tiễn của đề tài

* Cơ sở khoa học: Dựa trên các phương pháp phân tích dữ liệu và mô hình dự đoán hành vi khách hàng.
* Cơ sở thực tiễn: Áp dụng vào ngành ngân hàng để tối ưu hóa chiến lược kinh doanh và marketing.

## 1.7. Bố cục đồ án

Đề tài gồm 4 chương chính:

* Chương 1: Tổng quan về đề tài.
* Chương 2: Phân tích dữ liệu khách hàng.
* Chương 3: Kết quả phân tích và đề xuất chiến lược.
* Chương 4: Kết luận và hướng phát triển.

# CHƯƠNG 2: Cơ sở lý thuyết

## 2.1. Cơ sở lý luận

### 2.1.1. Khái niệm và đặc điểm của hành vi khách hàng trong ngành ngân hàng

Hành vi khách hàng trong ngành ngân hàng là tập hợp các phản ứng, thái độ, và hành động của khách hàng đối với sản phẩm và dịch vụ ngân hàng.

Điều này bao gồm:

* Việc khách hàng lựa chọn sản phẩm
* Tần suất giao dịch
* Mức độ phản hồi với các chiến dịch tiếp thị và sự trung thành với thương hiệu.

Đặc điểm nổi bật của hành vi khách hàng trong lĩnh vực này là tính đa dạng, thường xuyên thay đổi và chịu ảnh hưởng bởi nhiều yếu tố như kinh tế, tâm lý, xã hội và công nghệ.

🡪 Việc phân tích hành vi khách hàng giúp ngân hàng nhận diện nhu cầu tiềm ẩn, nâng cao khả năng giữ chân khách hàng và cá nhân hóa dịch vụ.

### 2.1.2. Ứng dụng hệ thống học máy để phân tích hành vi khách hàng thông qua dữ liệu được truyền vào

* Hệ thống học máy (machine learning) là một tập hợp các thuật toán cho phép máy tính học từ dữ liệu đầu vào để phát hiện các mẫu (patterns) và đưa ra dự đoán mà không cần lập trình rõ ràng từng bước.
* Trong lĩnh vực ngân hàng, học máy có thể được ứng dụng để phân loại khách hàng theo mức độ tiềm năng, dự đoán khả năng họ tham gia vào các dịch vụ như tiền gửi, vay vốn, thẻ tín dụng,...

🡪 Thông qua việc huấn luyện mô hình trên các tập dữ liệu lịch sử, ngân hàng có thể rút ra những đặc điểm hành vi tiêu biểu, giúp tối ưu hóa chiến lược tiếp thị và nâng cao hiệu quả hoạt động kinh doanh.

### 2.1.3. Lịch sử và sự phát triển của các mô hình học máy trong phân tích dữ liệu

* Học máy bắt đầu được nghiên cứu từ những năm 1950 với các mô hình đơn giản như Perceptron.
* Đến thập niên 1990, các thuật toán như cây quyết định, K-Nearest Neighbors (KNN), và mạng Bayes đã được ứng dụng rộng rãi trong phân tích dữ liệu. Sự phát triển của phần cứng và dữ liệu lớn (big data) đã thúc đẩy mạnh mẽ việc áp dụng học máy trong thập kỷ gần đây.
* Ngày nay, các mô hình tiên tiến như Random Forest, Gradient Boosting, và mạng nơ-ron nhân tạo (Neural Networks) không chỉ được ứng dụng trong công nghệ mà còn trong các lĩnh vực như ngân hàng, y tế, giáo dục,... nhằm nâng cao khả năng phân tích, dự đoán và ra quyết định tự động.

### 2.1.4. Cấu trúc và cách hoạt động của các mô hình phân loại khách hàng trong học máy

Các mô hình phân loại khách hàng thường hoạt động theo quy trình sau: tiền xử lý dữ liệu, trích xuất đặc trưng, huấn luyện mô hình, đánh giá và dự đoán. Mỗi mô hình có cấu trúc và nguyên lý hoạt động riêng:

* Decision Tree (Cây quyết định): Mô hình phân chia dữ liệu thành các nhánh dựa trên điều kiện logic để đi đến kết quả phân loại.
* Random Forest: Tập hợp nhiều cây quyết định hoạt động song song, giúp giảm thiểu quá khứ học lệch và nâng cao độ chính xác.
* Logistic Regression: Dựa trên xác suất tuyến tính, thường được sử dụng để phân loại nhị phân (có/không).
* K-Nearest Neighbors (KNN): Phân loại mẫu mới dựa trên đa số nhãn của các điểm dữ liệu gần nhất.
* Support Vector Machine (SVM): Tìm siêu mặt phẳng tối ưu phân tách các nhóm dữ liệu khác nhau.

Tùy theo mục tiêu và đặc điểm dữ liệu, các mô hình sẽ được lựa chọn để tối ưu hóa kết quả phân tích.

## 2.2. Các mô hình học máy trong phân tích hành vi khách hàng

Một số mô hình học máy phổ biến được sử dụng trong phân tích hành vi khách hàng bao gồm:

* Logistic Regression: Dễ triển khai và giải thích, phù hợp với bài toán phân loại nhị phân như “có/không tham gia”.
* Decision Tree và Random Forest: Dễ trực quan hóa và hiệu quả cao với dữ liệu phức tạp
* Gradient Boosting Machines (GBM): Cho độ chính xác cao nhờ kết hợp nhiều mô hình yếu.
* K-Nearest Neighbors (KNN): Hiệu quả khi dữ liệu không quá lớn và dễ hiểu.
* Neural Networks: Mô hình phức tạp, hiệu quả cao trong việc phát hiện các mối quan hệ phi tuyến tính trong dữ liệu lớn.

Việc lựa chọn mô hình phù hợp phụ thuộc vào đặc điểm của tập dữ liệu và mục tiêu phân tích. Trong đề tài này, các mô hình sẽ được áp dụng và so sánh để chọn ra mô hình tối ưu nhằm phân loại và dự đoán hành vi khách hàng.

### 2.2.1. Khái niệm về học máy

Học máy (Machine Learning - ML) là một nhánh của trí tuệ nhân tạo (AI) cho phép hệ thống tự động học hỏi và cải thiện từ dữ liệu mà không cần lập trình rõ ràng. ML hoạt động dựa trên việc xây dựng các mô hình toán học để phân tích dữ liệu và đưa ra dự đoán hoặc quyết định. Các mô hình này có thể học từ dữ liệu đầu vào, xác định mẫu và điều chỉnh để nâng cao độ chính xác theo thời gian.

### 2.2.2. Các mô hình học máy phổ biến

Các mô hình học máy có thể được chia thành bốn loại chính:

1. Học có giám sát (Supervised Learning): Dữ liệu huấn luyện bao gồm cả đầu vào và đầu ra mong muốn, giúp mô hình học cách ánh xạ từ đầu vào đến đầu ra. Các thuật toán phổ biến:

* Hồi quy tuyến tính (Linear Regression)
* Hồi quy logistic (Logistic Regression)
* Cây quyết định (Decision Tree)
* Random Forest
* Mạng nơ-ron nhân tạo (Neural Networks)

1. Học không giám sát (Unsupervised Learning): Mô hình học từ dữ liệu không có nhãn, phát hiện mẫu hoặc cấu trúc ẩn trong dữ liệu. Các thuật toán phổ biến:

* Phân cụm K-Means
* Phân tích thành phần chính (PCA)
* Mạng tự mã hóa (Autoencoders)

1. Học bán giám sát (Semi-supervised Learning): Kết hợp giữa học có giám sát và học không giám sát, sử dụng một lượng nhỏ dữ liệu có nhãn và một lượng lớn dữ liệu không có nhãn để cải thiện hiệu suất mô hình. Các thuật toán phổ biến:

* Mô hình bán giám sát dựa trên đồ thị (Graph-based Semi-supervised Learning)
* Mô hình tự huấn luyện (Self-training)
* Mạng nơ-ron bán giám sát (Semi-supervised Neural Networks)

1. Học tăng cường (Reinforcement Learning): Mô hình học thông qua thử nghiệm và phản hồi từ môi trường để tối ưu hóa hành vi. Các thuật toán phổ biến:

* Q-learning
* Deep Q-Network (DQN)
* Chính sách sâu (Deep Policy Networks)

### 2.2.3. Ứng dụng mô hình học máy trong phân loại hành vi khách hàng

## 2.3. Ngôn ngữ lập trình:

Học máy thường sử dụng các ngôn ngữ lập trình như Python, R, Java, nhưng Python là ngôn ngữ phổ biến nhất nhờ sự đa dạng của các thư viện hỗ trợ. Dưới đây là một số thư viện quan trọng trong Python:

### 2.3.1. Thư viện Numpy

NumPy là thư viện cơ bản hỗ trợ tính toán khoa học trong Python, cung cấp các cấu trúc dữ liệu mạnh mẽ như mảng đa chiều (ndarray) và các phép toán ma trận tối ưu. Nó giúp xử lý dữ liệu nhanh chóng và hiệu quả trong các ứng dụng học máy.

### 2.3.2. Thư viện Pandas

Pandas cung cấp các cấu trúc dữ liệu linh hoạt như DataFrame và Series để xử lý dữ liệu có cấu trúc. Nó hỗ trợ đọc/ghi dữ liệu từ nhiều nguồn khác nhau (CSV, Excel, SQL) và cung cấp các công cụ mạnh mẽ để thao tác dữ liệu, lọc, nhóm và tổng hợp dữ liệu.

### 2.3.3. Thư viện Matplotlib

Matplotlib là thư viện giúp tạo biểu đồ và trực quan hóa dữ liệu. Nó hỗ trợ vẽ biểu đồ đường, cột, tán xạ, histogram, v.v. để phân tích và trình bày dữ liệu một cách trực quan.

### 2.3.4. Thư viện Scikit-learn (Sklearn)

Scikit-learn là thư viện mạnh mẽ nhất trong Python cho học máy, cung cấp các công cụ để:

* Tiền xử lý dữ liệu (chuẩn hóa, mã hóa nhãn, chọn đặc trưng)
* Xây dựng mô hình (hồi quy, phân loại, phân cụm)
* Đánh giá mô hình (cross-validation, metric đánh giá)
* Tinh chỉnh mô hình (GridSearchCV, RandomizedSearchCV)

Scikit-learn giúp đơn giản hóa việc triển khai các mô hình học máy và là công cụ không thể thiếu cho các nhà khoa học dữ liệu.

# CHƯƠNG 3: MÔ TẢ VÀ PHÂN TÍCH HỆ THỐNG BÀI TOÁN PHÂN LOẠI BẰNG HỆ THỐNG HỌC MÁY

## 3.1. Bối cảnh nghiên cứu

### 3.1.1. Đối tượng nghiên cứu

Đối tượng nghiên cứu trong đề tài này là hành vi tiêu dùng của các khách hàng trong việc tham gia vào chiến dịch gửi tiền kỳ hạn tại ngân hàng. Dựa trên thông tin cá nhân, tài chính và hành vi tương tác với các chiến dịch tiếp thị trước đó, mô hình học máy sẽ giúp phân loại khả năng tham gia của khách hàng.

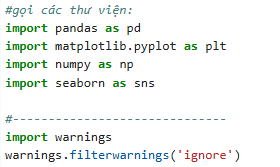
### 3.1.2. Dữ liệu

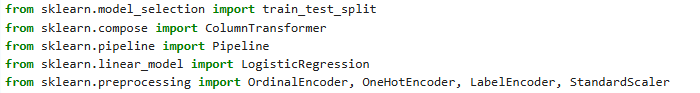
Tập dữ liệu sử dụng là "bank-full.csv", chứa thông tin từ chiến dịch marketing của một ngân hàng tại Bồ Đào Nha. Dữ liệu bao gồm các biến định tính (job, marital, education, contact...) và định lượng (age, balance, duration, campaign, pdays, previous...), cùng với biến Target (y) thể hiện khách hàng có tham gia gửi tiền hay không.

## 3.2. Tiền xử lý dữ liệu

### 3.2.1. Cài đặt các thư viện liên quan

Các thư viện bao gồm:





Trong đó:

- pandas: Thư viện mạnh mẽ dùng để thao tác và phân tích dữ liệu có cấu trúc (dữ liệu bảng). Cho phép dễ dàng đọc file CSV, truy xuất, lọc và xử lý dữ liệu.

- numpy: Thư viện hỗ trợ tính toán số học hiệu suất cao, đặc biệt hữu ích với các mảng nhiều chiều (arrays) và xử lý dữ liệu số.

- matplotlib.pyplot và seaborn: Hai thư viện vẽ biểu đồ phổ biến. matplotlib cung cấp khả năng tạo các biểu đồ tùy chỉnh, trong khi seaborn hỗ trợ trực quan hóa dữ liệu thống kê với giao diện đẹp và dễ sử dụng.

- warnings.filterwarnings('ignore') giúp loại bỏ các cảnh báo không cần thiết để tránh gây nhiễu khi hiển thị kết quả.

- sklearn.model\_selection.train\_test\_split: Hàm chia dữ liệu thành tập huấn luyện và tập kiểm tra, rất cần thiết để đánh giá hiệu suất của mô hình.

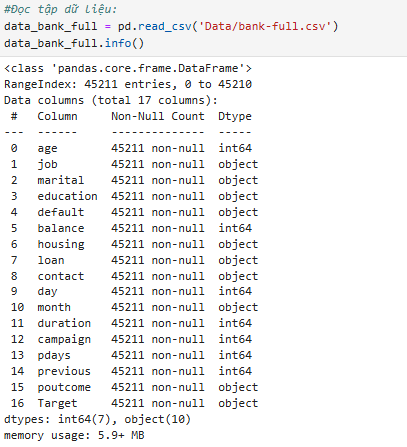
- sklearn.compose.ColumnTransformer: Cho phép xử lý dữ liệu cột theo từng nhóm riêng biệt (ví dụ: mã hóa dữ liệu phân loại và chuẩn hóa dữ liệu số cùng lúc).

- sklearn.pipeline.Pipeline: Dùng để xây dựng chuỗi các bước xử lý dữ liệu và huấn luyện mô hình một cách mạch lạc, tiện lợi và tự động.

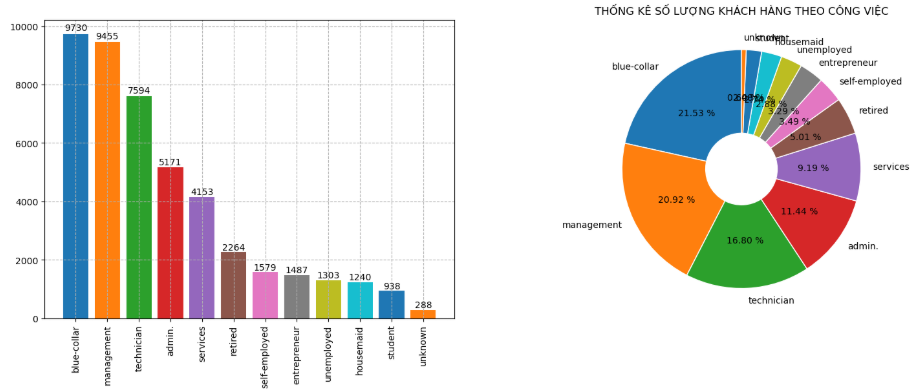
- sklearn.linear\_model.LogisticRegression: Mô hình hồi quy logistic — một mô hình phân loại nhị phân đơn giản và phổ biến trong học máy:

* OrdinalEncoder: Mã hóa dữ liệu phân loại theo thứ tự (ordinal).
* OneHotEncoder: Mã hóa dữ liệu phân loại không theo thứ tự (nominal) bằng cách tạo các cột nhị phân.
* LabelEncoder: Mã hóa nhãn đầu ra (target) dưới dạng số nguyên.
* StandardScaler: Chuẩn hóa dữ liệu đầu vào (số thực) để có giá trị trung bình 0 và độ lệch chuẩn 1.

### 3.2.2. Tải bộ dữ liệu

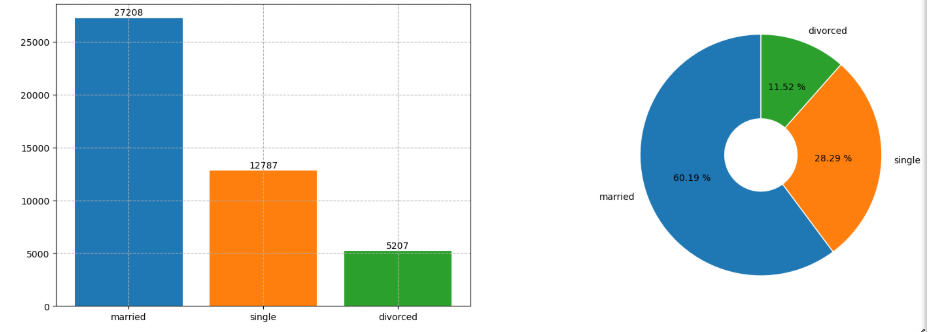


### 3.2.3. Trực quan hóa dữ liệu



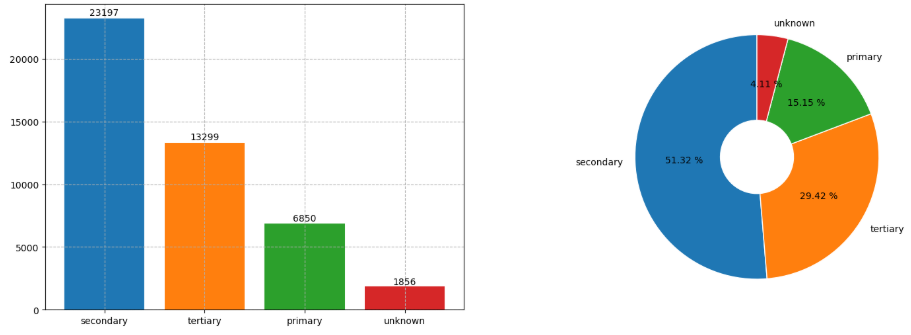
Hình 1: Biểu đồ thể hiện nghề nghiệp của khách hàng

* Ba nhóm chính chiếm gần 60% tổng số khách hàng: blue-collar, management, technician.
* Các nhóm như student, housemaid, unemployed, unknown có số lượng khách hàng rất nhỏ.
* Điều này có thể phản ánh thị trường mục tiêu chính là những người đang làm việc ổn định trong các ngành kỹ thuật, quản lý và lao động phổ thông.



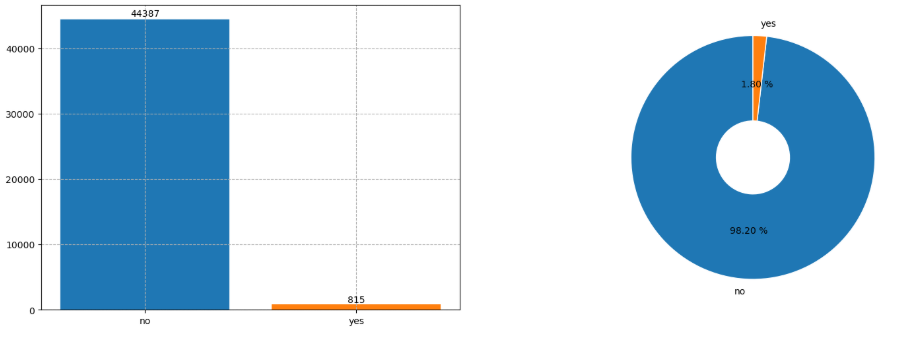
Hình 2: Biểu đồ thể hiện tình trạng hôn nhân của khách hàng

* Khách hàng đã kết hôn chiếm tỷ lệ lớn nhất (hơn 60%), cho thấy nhóm này là đối tượng chính trong cơ sở dữ liệu.
* Người độc thân chiếm tỷ lệ đáng kể (~28%), là nhóm tiềm năng nếu doanh nghiệp muốn mở rộng thị trường.
* Người đã ly hôn có tỷ lệ thấp, chỉ hơn 11%, có thể là nhóm ít được nhắm đến trong các chiến lược tiếp cận.



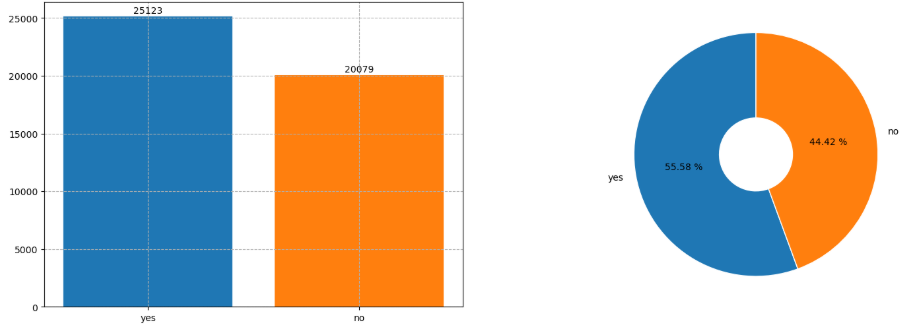
Hình 3: Biểu đồ thể hiện trình độ học vấn của khách hàng

* Phần lớn khách hàng có trình độ học vấn trung học phổ thông (hơn 50%). Đây là nhóm chiếm ưu thế rõ rệt trong dữ liệu.
* Khách hàng có trình độ đại học/cao đẳng chiếm gần 30%, cũng là nhóm quan trọng.
* Tỷ lệ người học tiểu học tuy thấp hơn nhưng vẫn đáng kể (15.15%), cho thấy sự đa dạng trong trình độ học vấn của khách hàng.
* Dữ liệu bị thiếu (unknown) tương đối ít, chỉ hơn 4%, nên không ảnh hưởng nhiều đến phân tích tổng thể.



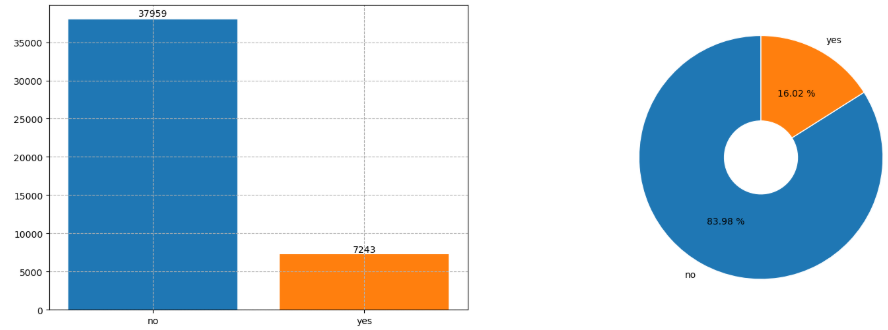
Hình 4: Biểu đồ thể hiện tình trạng nợ xấu của khách hàng

* Tỷ lệ khách hàng có nợ xấu rất thấp, chỉ khoảng 1.8%, cho thấy:
* Ngân hàng/tổ chức tín dụng đang quản lý rủi ro tín dụng tốt.
* Khách hàng trong tập dữ liệu có độ tin cậy tài chính cao.



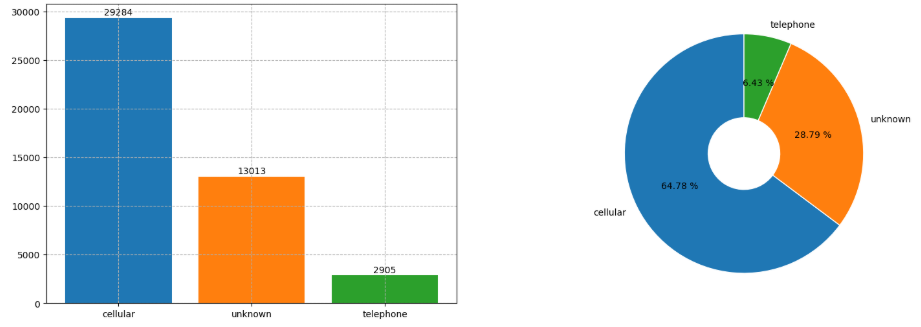
Hình 5: Biểu đồ thể hiện tình trạng căn nhà của khách hàng

* Hơn một nửa số khách hàng (55.58%) có vay thế chấp nhà, cho thấy:
* Đây là loại sản phẩm tín dụng phổ biến và được nhiều khách hàng sử dụng.
* Có thể phản ánh mức độ ổn định tài chính nhất định, vì vay thế chấp thường yêu cầu khả năng trả nợ tốt.
* Tỷ lệ không vay (44.42%) vẫn chiếm phần đáng kể, là nhóm tiềm năng để ngân hàng mở rộng các chương trình vay mua nhà.



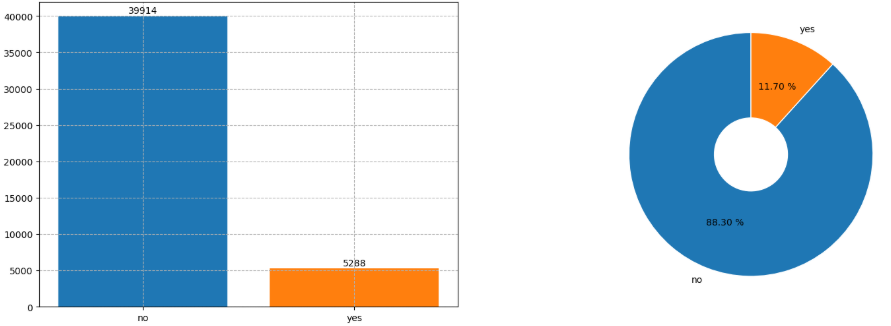
Hình 6: Biểu đồ thể hiện tình trạng khoản vay cá nhân của khách hàng

🡪Tỷ lệ khách hàng có khoản vay cá nhân khá thấp, chỉ chiếm 16.02%, trong khi phần lớn khách hàng (gần 84%) không có khoản vay này.



Hình 7: Biểu đồ thể hiện phương thức liên hệ với khách hàng

* Cellular (di động) là phương thức liên hệ phổ biến nhất với gần 65% khách hàng. Điều này phù hợp với xu hướng hiện đại khi khách hàng ưu tiên liên lạc qua điện thoại di động.
* Tỷ lệ "unknown" khá cao (gần 29%), cho thấy một khoảng trống dữ liệu đáng chú ý.
* Telephone (điện thoại bàn) chỉ chiếm 6.43%, cho thấy phương thức này đang giảm vai trò, có thể do khách hàng chuyển sang dùng di động nhiều hơn.



Hình 1: Biểu đồ thể hiện nghề nghiệp của khách hàng

Tập dữ liệu bị mất cân bằng trung bình (Moderate):

- Chỉ 11.70% khách hàng có hành vi tích cực (đăng ký), trong khi phần lớn (gần 9/10) không đăng ký.

- Điều này rất quan trọng trong bài toán phân loại, vì mô hình có thể nghiêng về việc dự đoán "no" nếu không xử lý đúng cách.

### 3.2.4. Tăng cường dữ liệu

Tăng cường dữ liệu là quá trình xử lý để cải thiện chất lượng dữ liệu đầu vào cho mô hình học máy. Điều này có thể bao gồm: thay thế hoặc loại bỏ giá trị ngoại lai, chuẩn hóa dữ liệu, xử lý dữ liệu bị thiếu và tạo ra các đặc trưng mới để mô hình học tốt hơn.

### 3.2.5. Mô hình cây quyết định

Mô hình cây quyết định là thuật toán phân loại dựa trên việc chia nhỏ dữ liệu theo các thuộc tính có ảnh hưởng nhất, theo cấu trúc dạng cây. Mỗi nút trong cây biểu diễn một thuộc tính và mỗi nhánh là một giá trị của thuộc tính đó, kết quả ở lá cây là phân loại đầu ra.

### 3.2.6. Mô hình phân cụm

Mô hình phân cụm là phương pháp phân chia các điểm dữ liệu thành các nhóm sao cho các điểm trong cùng một nhóm có đặc điểm tương đồng. Mô hình này được sử dụng để khám phá các mẫu hành vi khách hàng chưa biết trước trong dữ liệu.

### 3.2.7. Mô hình phân lớp

Mô hình phân lớp là quá trình xây dựng thuật toán để gán nhãn cho dữ liệu đầu vào thuộc vào một trong nhiều lớp đầu ra có sẵn. Các mô hình phân lớp bao gồm:

* Hồi quy logistic K-Nearest Neighbors (KNN)
* Random Forest Gradient Boosting
* Hiệu quả của mô hình được đánh giá bằng các chỉ số như độ chính xác (accuracy), độ nhạy (recall), độ đặc hiệu (precision) và F1-score.

# CHƯƠNG 4: LỰA CHỌN MÔ HÌNH CHO BÀI TOÁN PHÂN LOẠI HÀNH VI KHÁCH HÀNG

## 4.1 Lựa chọn mô hình

Dựa trên đặc điểm của dữ liệu và mục tiêu phân loại, mô hình cây quyết định (Decision Tree) được lựa chọn vì tính dễ hiểu, dễ trực quan hóa và hiệu quả trong việc xử lý các thuộc tính phân loại rời rạc. Mô hình này cũng phù hợp với việc phân tích hành vi khách hàng trong bối cảnh dữ liệu ngân hàng đa dạng về loại thuộc tính.

## 4.2 Thiết kế hệ thống

Hệ thống được thiết kế gồm các bước chính:

- Tiền xử lý dữ liệu: bao gồm làm sạch dữ liệu, mã hóa nhãn và chuẩn hóa. Chia dữ liệu thành tập huấn luyện và tập kiểm tra.

- Xây dựng mô hình cây quyết định bằng thư viện scikit-learn.

- Đánh giá mô hình bằng các chỉ số như độ chính xác.

## 4.3 Kết quả thực nghiệm

Mô hình cây quyết định cho kết quả khá tốt trên tập dữ liệu kiểm tra, với độ chính xác trung bình đạt khoảng 88%. Mô hình cũng cho phép nhận diện rõ các yếu tố quan trọng ảnh hưởng đến hành vi khách hàng như: thời lượng cuộc gọi, tình trạng vay vốn, trình độ học vấn và nghề nghiệp.

## 4.4 Phân tích sau khi lựa chọn và huấn luyện mô hình

Kết quả cho thấy mô hình cây quyết định có khả năng giải thích tốt các quyết định phân loại. Các nhánh của cây cung cấp cái nhìn rõ ràng về cách các đặc trưng ảnh hưởng đến quyết định cuối cùng.

# KẾT LUẬN

1. Kết quả đạt được

Đề tài đã thực hiện toàn bộ quy trình phân tích hành vi khách hàng bằng mô hình học máy. Từ khâu tiền xử lý, lựa chọn mô hình cho đến đánh giá kết quả đều được triển khai chặt chẽ. Mô hình cây quyết định đã cho kết quả khả quan với độ chính xác cao và khả năng diễn giải tốt.

Kết quả không chỉ giúp phân loại hành vi khách hàng một cách hiệu quả mà còn làm nền tảng để triển khai các chiến lược kinh doanh, tiếp thị và chăm sóc khách hàng tối ưu hơn trong ngành ngân hàng

1. Hướng nghiên cứu của đề tài

Trong tương lai, đề tài có thể được mở rộng theo các hướng:

* Áp dụng thêm mô hình nâng cao như Random Forest để cải thiện độ chính xác.
* Kết hợp nhiều nguồn dữ liệu hơn như hành vi trực tuyến, dữ liệu từ mạng xã hội.
* Xây dựng hệ thống phân tích thời gian thực để hỗ trợ ra quyết định nhanh chóng.
* Phân tích chuyên sâu từng nhóm khách hàng để cá nhân hóa sản phẩm ngân hàng.

# TÀI LIỆU THAM KHẢO