|  |
| --- |
| **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

[Hình 1: Biểu đồ thể hiện nghề nghiệp của khách hàng…………………………………………….. 13](#_Toc199357587)

[Hình 2: Biểu đồ thể hiện tình trạng hôn nhân của khách hàng……………………………………. ..13](#_Toc199357588)

[Hình 3: Biểu đồ thể hiện trình độ học vấn của khách hàng……………………………………….. ..13](#_Toc199357589)

[Hình 4: Biểu đồ thể hiện tình trạng nợ xấu của khách hàng……………………………………… ...14](#_Toc199357590)

[Hình 5: Biểu đồ thể hiện tình trạng căn nhà của khách hàng…………………………………….. …14](#_Toc199357591)

[Hình 6: Biểu đồ thể hiện tình trạng khoản vay cá nhân của khách hàng………………………….. ..15](#_Toc199357592)

[Hình 7: Biểu đồ thể hiện phương thức liên hệ với khách hàng……………………………………. ..15](#_Toc199357593)

[Hình 8: Biểu đồ thể hiện khách hàng có đăng ký gửi có kỳ hạn…………………………………... ..15](#_Toc199357594)

[Hình 9: Biểu đồ biểu diễn cột dữ liệu pdays (số ngày kể từ lần liên hệ trước đó) .…………...……..16](#_Toc199357595)

[Hình 10: Biểu đồ biểu diễn cột dữ liệu duration (thời gian của cuộc gọi cuối cùng)……………...... 16](#_Toc199357596)

[Hình 11: Biểu đồ biểu diễn cột dữ liệu age (tuổi của khách hàng) và balance (số dư tài khoản ngân hàng)……………………………………………………………………………………………….. ..17](#_Toc199357597)

[Hình 12: Biểu đồ biểu diễn cột dữ liệu campaign (số lần liên hệ trong chiến dịch hiện tại)……….. 17](#_Toc199357598)

[Hình 13: Biểu đồ biểu diễn cột dữ liệu previous (số lần liên hệ trước đó trong chiến dịch trước)…... 18](#_Toc199357599)

# 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 công nghệ tài chính. 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 đăng ký gửi tiền có kỳ hạn của ngân hàng hay không.

## 3.2. Tiền xử lý dữ liệu

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

Các thư viện bao gồm:

import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

import seaborn as sns

import warnings

warnings.filterwarnings('ignore')

from sklearn.model\_selection import train\_test\_split

from sklearn.tree import DecisionTreeClassifier

from imblearn.over\_sampling import SMOTE

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.

- train\_test\_split: dùng để chia dữ liệu thành 2 (hoặc nhiều) tập: tập huấn luyện (train) và tập kiểm tra (test).

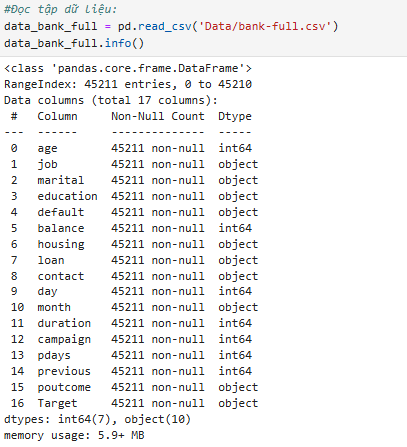
- DecisionTreeClassifier: là một mô hình học có giám sát (supervised learning) dùng để phân loại dữ liệu bằng cách xây dựng một cây quyết định (decision tree).

Cách cài đặt:

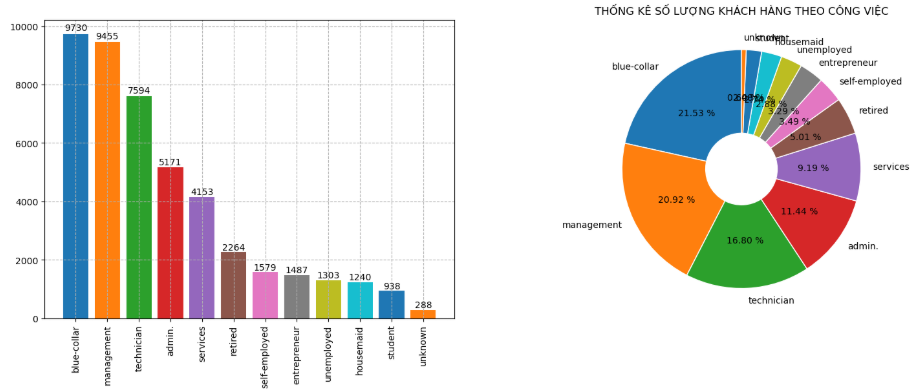
!pip install pandas matplotlib numpy seaborn

!pip install scikit-learn imbalanced-learn

### 3.2.2. Tải bộ dữ liệu

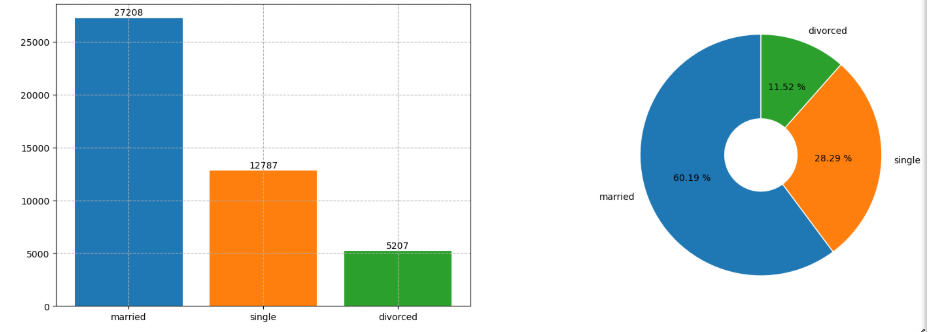


### 3.2.3. Trực quan hóa dữ liệu



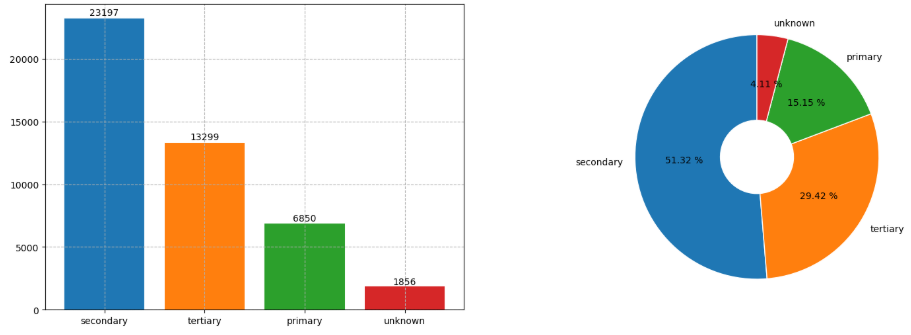
Hình 1: Biểu đồ thể hiện nghề nghiệp của khách hàng

* 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.



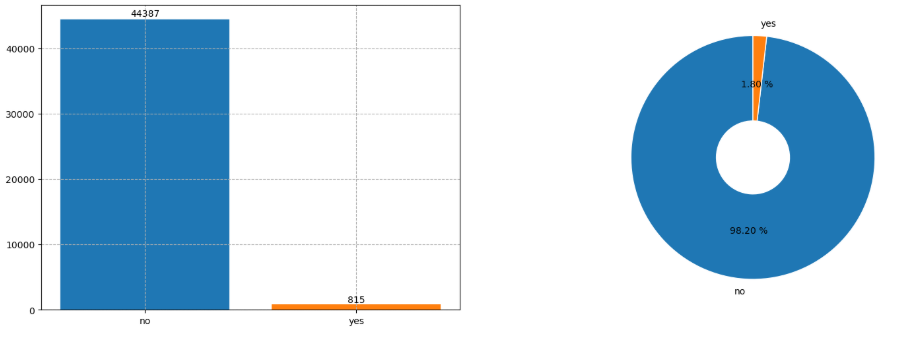
Hình 2: Biểu đồ thể hiện tình trạng hôn nhân của khách hàng

* 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.



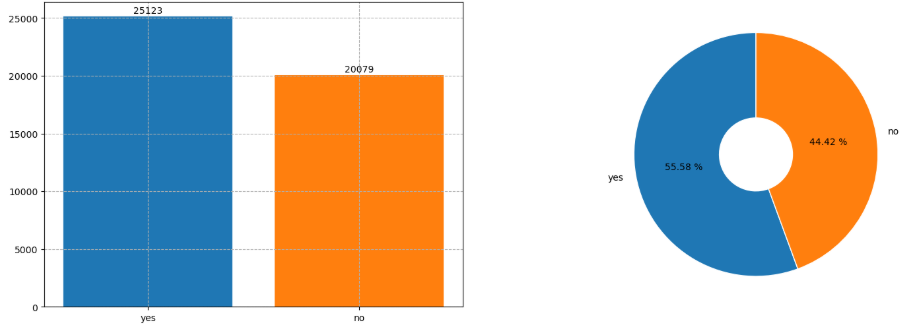
Hình 3: Biểu đồ thể hiện trình độ học vấn của khách hàng

* 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.



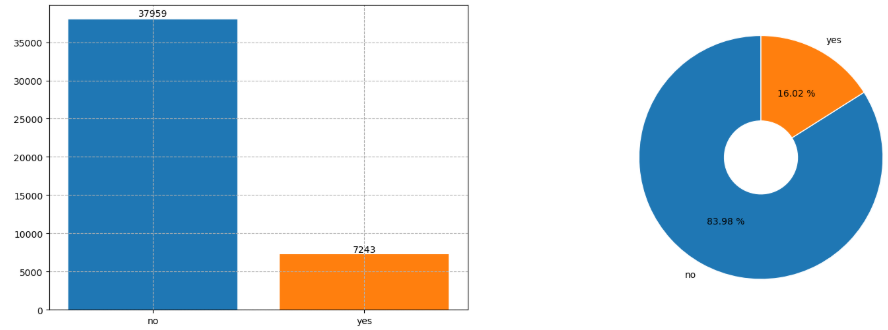
Hình 4: Biểu đồ thể hiện tình trạng nợ xấu của khách hàng

* 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.



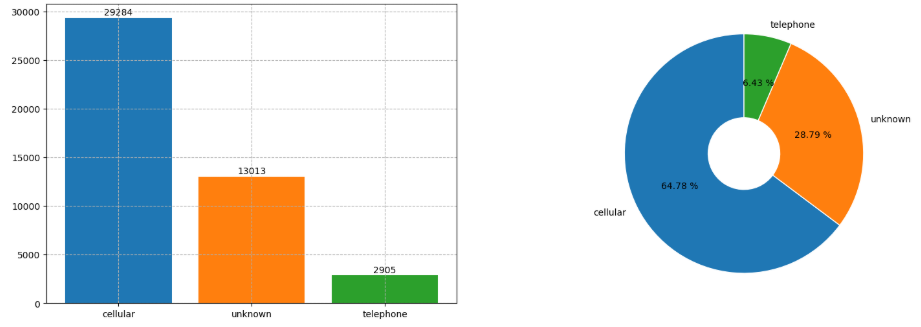
Hình 5: Biểu đồ thể hiện tình trạng căn nhà của khách hàng

* 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.



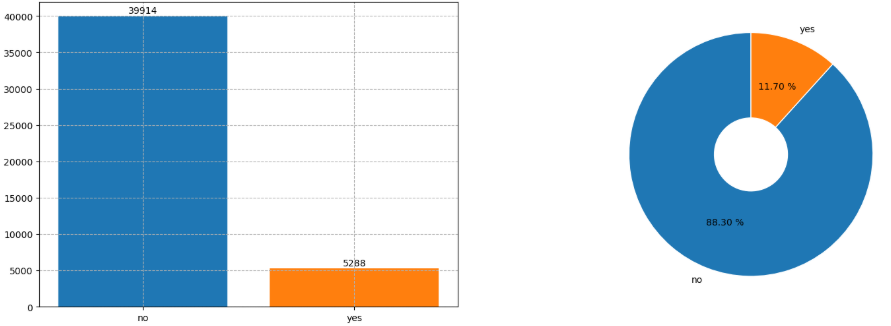
Hình 6: Biểu đồ thể hiện tình trạng khoản vay cá nhân của khách hàng

* 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.



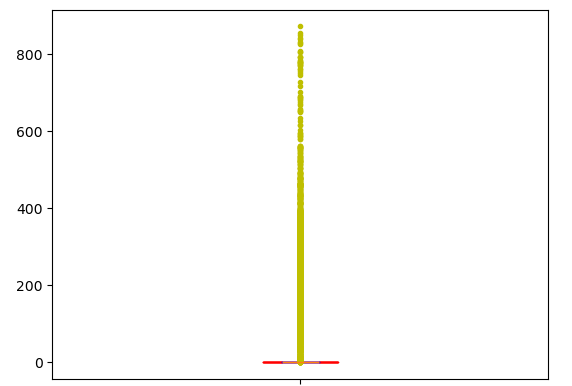
Hình 7: Biểu đồ thể hiện phương thức liên hệ với khách hàng

* 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.



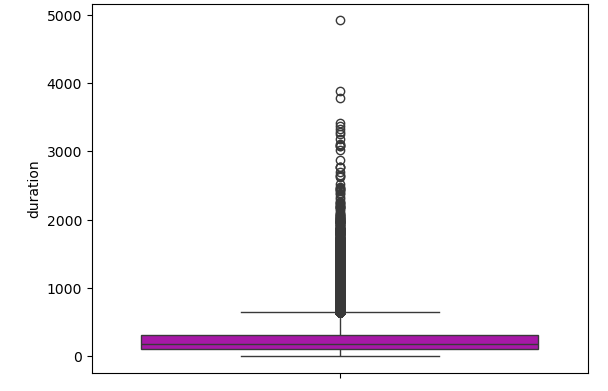
Hình 8: Biểu đồ thể hiện khách hàng có đăng ký gửi có kỳ hạn

* 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.



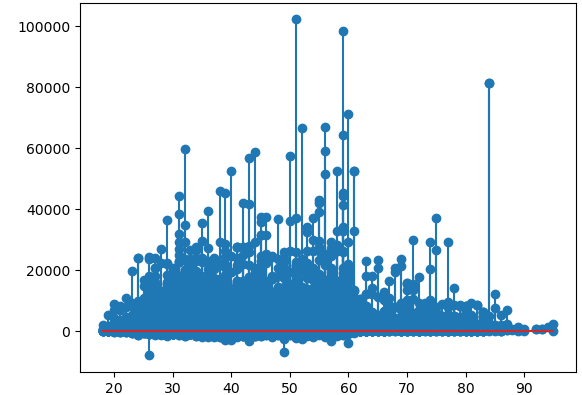
Hình 9: Biểu đồ biểu diễn cột dữ liệu pdays (số ngày kể từ lần liên hệ trước đó)

* Cột pdays có phân phối rất lệch, nhiều dữ liệu ngoại lai, và rất nhiều giá trị mặc định (999) 🡪 Nếu đưa vào mô hình học máy sẽ gây ảnh hưởng lớn đến hiệu suất và độ chính xác 🡪 Loại bỏ cột dữ liệu này.



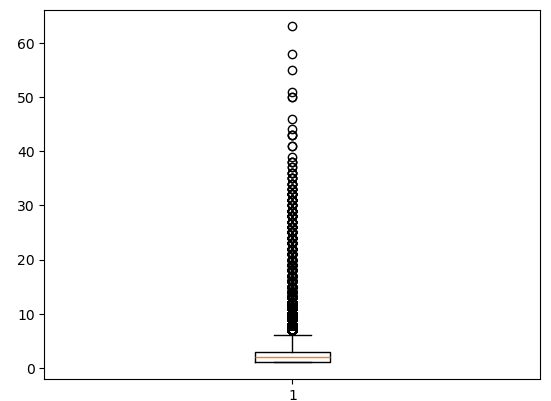
Hình 10: Biểu đồ biểu diễn cột dữ liệu duration (thời gian của cuộc gọi cuối cùng)

* Cột duration có phân bố lệch mạnh và có 3 điểm dữ liệu ngoại lai 🡪 Nên xử lý các dữ liệu ngoại lai để tránh gây ảnh hưởng đến mô hình học máy.



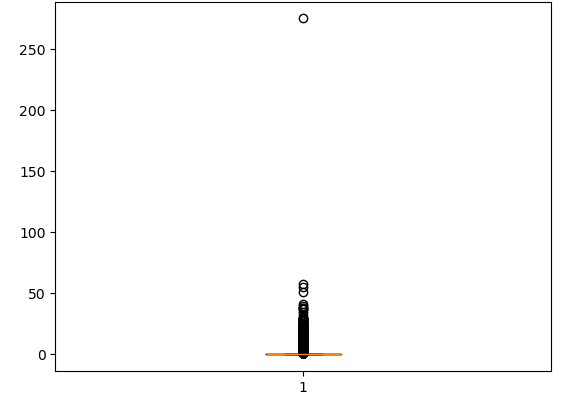
Hình 11: Biểu đồ biểu diễn cột dữ liệu age (tuổi của khách hàng) và balance (số dư tài khoản ngân hàng)

* Mặc dù dữ liệu balance của mỗi khách hàng không đồng đều và có dữ liệu âm nhưng không thể nào loại bỏ những dữ liệu đó được vì có những người có thể nợ ngân hàng và cũng có những người có thu nhập cao 🡪 Nên phân dữ liệu này thành các nhóm để mô hình học máy dễ dàng xử lý.
* Tuổi từ 30 đến 60 có mật độ dày đặc hơn 🡪 Đây là nhóm khách hàng chính rơi vào độ tuổi lao động.



Hình 12: Biểu đồ biểu diễn cột dữ liệu campaign (số lần liên hệ trong chiến dịch hiện tại)

* Cột campaign có phân bố lệch mạnh và có 5 điểm dữ liệu ngoại lai 🡪 Nên xử lý các dữ liệu ngoại lai để tránh gây ảnh hưởng đến mô hình học máy.



Hình 13: Biểu đồ biểu diễn cột dữ liệu previous (số lần liên hệ trước đó trong chiến dịch trước)

* Cột dữ liệu có rất nhiều giá trị thấp và một số ít giá trị rất cao.
* Cột dữ liệu này cũng không ảnh hưởng quá nhiều đến mô hình học máy 🡪 Có thể loại bỏ được.

### 3.2.4. Tăng cường dữ liệu

Tăng cường dữ liệu là quá trình tạo ra dữ liệu mới từ dữ liệu hiện có bằng cách áp dụng các biến đổi, thay đổi hoặc tổng hợp một cách hợp lý mà không làm thay đổi bản chất hoặc nhãn của dữ liệu.

Mục tiêu chính là:

- Tăng số lượng dữ liệu huấn luyện.

- Cải thiện khả năng tổng quát hóa (generalization) của mô hình.

- Giảm hiện tượng quá khớp (overfitting).

### 3.2.5. Mô hình cây quyết định

- Thuật toán Cây quyết định (Decision Tree) là một thuật toán học có giám sát, được sử dụng cho cả bài toán phân loại và hồi quy.

- Cây quyết định hoạt động bằng cách xây dựng một cấu trúc dạng cây, trong đó mỗi nút đại diện cho một đặc trưng (feature), mỗi nhánh đại diện cho một giá trị của đặc trưng đó, và mỗi lá (leaf) đại diện cho một nhãn hoặc giá trị đầu ra.

- Cây quyết định rất phổ biến trong các ứng dụng thực tiễn, đặc biệt khi yêu cầu mô hình dễ giải thích và có thể xử lý cả dữ liệu phân loại và hồi quy.

### 3.2.6. Mô hình phân cụm

Phân cụm dữ liệu (clustering) là một phương pháp xử lý thông tin quan trọng và phổ biến, nó nhằm khám phá mối liên hệ giữa các mẫu dữ liệu bằng cách tổ chức chúng thành các cụm tương tự.

Phân cụm dữ liệu nhằm mục đích chính là khai phá cấu trúc của mẫu dữ liệu để thành lập các nhóm dữ liệu từ tập dữ liệu lớn, theo đó, cho phép người ta đi sâu vào phân tích và nghiên cứu cho từng cụm dữ liệu này nhằm khám phá và tìm kiếm các thông tin tiềm ẩn, hữu ích phục vụ cho ra quyết định.

Phân cụm dữ liệu là một kỹ thuật quan trọng trong lĩnh vực khai phá dữ liệu và học máy, thuộc nhóm học không giám sát (unsupervised learning). Mục tiêu của phân cụm dữ liệu đó là chia các đối tượng thành các cụm thuần nhất và phân biệt với nhau, tức là thành các nhóm đối tượng thỏa mãn 2 điều kiện sau:

- Độ tương tự của các đối tượng trong mỗi nhóm cao nhất có thể (tiêu chuẩn liên kết chặt).

- Các đối tượng trong các nhóm khác nhau phân biệt nhất có thể (tiêu chuẩn tách rời).

Các thuật toán phân cụm:

- K-Means.

- Hierarchical Clustering.

- DBSCAN.

- Mean Shift.

### 3.2.7. Mô hình phân lớp

Phân lớp dữ liệu (classification) là một kỹ thuật cốt lõi trong học có giám sát (supervised learning). Mục tiêu của phương pháp này là xây dựng một mô hình dự đoán nhãn (hoặc lớp) của các đối tượng mới dựa trên một tập dữ liệu đã biết trước nhãn. Nói cách khác, mô hình phân lớp học từ các ví dụ đã được gán nhãn để có thể gán đúng nhãn cho các dữ liệu chưa biết.

Mục đích: dự đoán những nhãn phân lớp cho các bộ dữ liệu /mẫu mới

* Đầu vào: Một tập các mẫu dữ liệu huấn luyện, với một nhãn phân lớp cho mỗi mẫu dữ liệu.
* Đầu ra: Mô hı̀nh (bộ phân lớp) dựa trên tập huấn luyện và nhãn cho mẫu mớ đưa vào.

Phân lớp dữ liệu chia làm 2 loại:

* Phân lớp nhị phân (Binary Classification)
* Phân lớp đã lớp (Multi-class Classification)

Các mô hình phân lớp bao gồm:

* Decision Tree (Cây quyết định).
* Naive Bayes.
* K-Nearest Neighbors (K-NN).
* Logistic Regression.
* Support Vector Machine (SVM).

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

## Lựa chọn mô hình

* Dạng bài toán:
* Học có giám sát (Supervised Learning) Tập dữ liệu bao gồm nhiều thuộc tính đầu vào (independent features) và một thuộc tính đầu ra (Target).
* Thuộc tính Target đã có nhãn rõ ràng với hai giá trị: yes (đăng ký) và no (không đăng ký).

🡪 Đây là bài toán học có giám sát vì mô hình được học từ dữ liệu có nhãn.

* Dạng nhãn đầu ra:
* Biến mục tiêu (Target) là biến rời rạc, mang 2 giá trị phân biệt 🡪 Đây là bài toán phân lớp (classification).
* Do chỉ có 2 lớp (yes/no) 🡪 Đây là bài toán phân lớp nhị phân (binary classification).
* Đặc điểm dữ liệu:
* Tập dữ liệu sử dụng bao gồm các thuộc tính hỗn hợp:
* Định lượng (numerical): age, balance, duration, campaign.
* Định tính (categorical): job, marital, education, default, housing, loan, contact.
* Một số biến dạng yes/no, một số biến dạng danh mục nhiều giá trị (job, education…).

🡪 Do vậy, mô hình cần có khả năng xử lý tốt cả hai loại dữ liệu này mà không yêu cầu biến đổi phức tạp.

* Lý do lựa chọn mô hình Cây quyết định (Decision Tree):
* Phù hợp với bài toán phân lớp nhị phân.
* Xử lý tốt cả dữ liệu định tính và định lượng mà có thể không cần chuẩn hóa hoặc biến đổi phức tạp.
* Có khả năng trực quan hóa mô hình dễ dàng dưới dạng cây, thuận tiện cho việc phân tích hành vi khách hàng và giải thích kết quả.
* Dễ dàng xác định các thuộc tính quan trọng.
* Tốc độ huấn luyện nhanh, dễ triển khai và kiểm thử trên nhiều tập dữ liệu.

## 4.2 Thiết kế hệ thống

Hệ thống được thiết kế gồm 7 bước xây dựng mô hình chính:

- Xác định bài toán và Thu thập dữ liệu.

* Bài toán phân tích hành vi khách hàng.
* Tên Dataset: bank-full.csv
* File dữ liệu chứa thông 45211 khách hàng của ngân hàng
* Mỗi dòng ứng với thông tin của 1 khách hàng, bao gồm 17 thuộc tính:
* age: Tuổi của khách hàng.
* job: Nghề nghiệp của khách hàng.
* marital: Tình trạng hôn nhân.
* education: Trình độ học vấn.
* default: Khách hàng có nợ xấu hay không? (yes/no).
* balance: Số dư tài khoản ngân hàng.
* housing: Khách hàng có khoản vay thế chấp nhà không? (yes/no).
* loan: Khách hàng có khoản vay cá nhân không? (yes/no).
* contact: Phương thức liên hệ với khách hàng (yes/no).
* day: Ngày trong tháng mà khách hàng được liên hệ lần gần nhất.
* month: Tháng mà khách hàng được liên hệ lần gần nhất.
* duration: Thời gian của cuộc gọi cuối cùng (tính bằng giây).
* campaign: Số lần liên hệ trong chiến dịch hiện tại.
* pdays: Số ngày kể từ lần liên hệ trước đó.
* previous: Số lần liên hệ trước đó trong chiến dịch trước.
* poutcome: Kết quả của chiến dịch tiếp thị trước đó (success = thành công, failure = thất bại, unknown = không xác định).
* Target: Mục tiêu khách hàng có đăng ký gửi tiền có kỳ hạn của ngân hàng không? (yes/no).

- Chuẩn bị dữ liệu.

* Quan sát dữ liệu.
* Kiểm tra và xử lý dữ liệu khuyết thiếu, trùng lặp và ngoại lai.
* Trích xuất biến độc lập (X), biến phụ thuộc (y)
* Phân tách tập dữ liệu huấn luyện (Train set) – Kiểm thử (Test set)

- Lựa chọn mô hình phù hợp.

* Dựa trên cột dữ liệu Target (biến phụ thuộc) có 2 biến yes/no là biến rời rạc 🡪 bài toán phân lớp nhị phân.
* Giữa phổ biến của bài toán phân lớp là KNN và Decision Tree thì bài toán Decision Tree xử lý dữ liệu nhanh hơn, dễ trực quan hóa và xử lý dữ liệu phức tạp tốt hơn.

- Huấn luyện mô hình.

- Đánh giá mô hình.

- Nâng cao độ chính xác của mô hình.

- Dự đoán với mô hình xây dựng được.

## 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 88,72% trên tập test và khoảng 89,4% trên tập train. Mô hình thuộc mức Best fit. 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.

Mặc dù mô hình chỉ đạt 88,72% nhưng cũng có thể từ dữ liệu cung cấp huấn để huấn luyện mô hình 🡪 Giúp cho ngân hàng có thể xác định các khách hàng cần chăm sóc đặc biệt cũng như có thể liên hệ để tiếp thị khách hàng mở tài khoản hay gửi tiền có kỳ cũng như các dịch vụ khác trong các chiến dịch sắp tới hiệu quả hơn.

## 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 hoàn thiện toàn bộ quy trình phân tích hành vi khách hàng trong ngành ngân hàng bằng các mô hình học máy. Từ giai đoạn tiền xử lý dữ liệu, khai thác đặc trưng, cho đến lựa chọn và huấn luyện mô hình, tất cả đều được triển khai theo một quy trình khoa học và có hệ thống. Việc tiền xử lý dữ liệu bao gồm làm sạch, loại bỏ dữ liệu ngoại lai, xử lý các giá trị thiếu và chuyển đổi dữ liệu dạng phân loại thành dạng số, giúp đảm bảo chất lượng dữ liệu đầu vào cho mô hình học máy.

Trong quá trình lựa chọn mô hình, Decision Tree (cây quyết định) được chọn là mô hình chính do khả năng diễn giải rõ ràng, dễ hiểu, phù hợp với các tổ chức tài chính cần sự minh bạch trong quyết định. Cây quyết định đã cho kết quả tốt với độ chính xác cao, dễ dàng nhận diện các yếu tố ảnh hưởng mạnh đến hành vi khách hàng như tình trạng nghề nghiệp, độ tuổi, thu nhập, thời gian liên hệ, v.v.

* Các kết quả thu được đã mang lại một số giá trị thực tiễn nổi bật:
* Xác định được các yếu tố then chốt ảnh hưởng đến quyết định của khách hàng, từ đó giúp ngân hàng tối ưu hóa chiến lược tiếp thị.
* Phân nhóm và phân loại khách hàng chính xác theo khả năng tham gia các sản phẩm của 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.
* Xây dựng hệ thống phân tích thời gian thực (real-time analytics) để hỗ trợ các quyết định tức thời trong hoạt động marketing hoặc chăm sóc khách hàng.

# TÀI LIỆU THAM KHẢO

1. Vũ Hữu Tiệp, Machine Learning cơ bản, 2018.
2. Đặng Văn Nam, Slide bài giảng Kỹ nghệ tri thức và học máy, Đại học Mỏ -Địa chất, 2025.
3. Nguyễn Thị Phương Bắc, Slide bài giảng Khai phá dữ liệu, Đại học Mỏ - Địa chất, 2024.
4. Aurélien Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd edition, O'Reilly Media,2019.