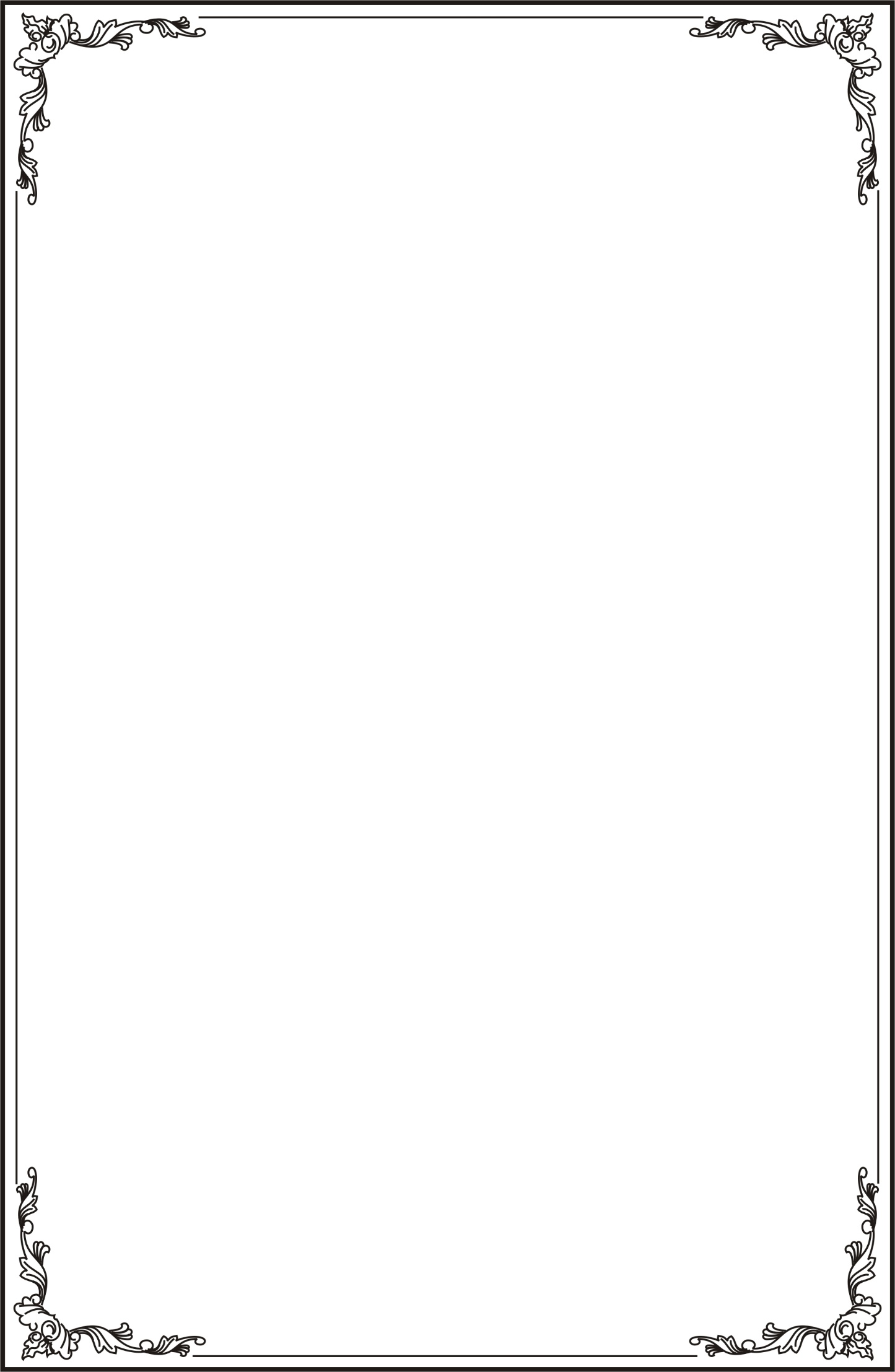
****

TRƯỜNG ĐẠI HỌC NHA TRANG

**KHOA CÔNG NGHỆ THÔNG TIN**

-----🙞🙜🕮🙞🙜-----

****

**BÁO CÁO MÔN HỆ HỖ TRỢ QUYẾT ĐỊNH**

**ĐỀ TÀI:** **XÂY DỰNG HỆ HỖ TRỢ QUYẾT ĐỊNH ĐỂ DỰ ĐOÁN HỒ SƠ CHO VAY**

|  |  |
| --- | --- |
| Họ và tên: | Nguyễn Dương Thanh Thùy |
| Lớp: | 63.HTTT |
| Mã số sinh viên: | 63135651 |
| Giảng viên hướng dẫn: | TS. Nguyễn Khắc Cường |

***Khánh Hòa, tháng 05 năm 2024***

**MỤC LỤC**

[**CHƯƠNG 1: TỔNG QUAN** 1](#_Toc167658114)

[**1.1** **LÝ DO CHỌN ĐỀ TÀI** 1](#_Toc167658115)

[**1.2** **MỤC TIÊU ĐỀ TÀI** 1](#_Toc167658116)

[**1.3** **PHƯƠNG PHÁP NGHIÊN CỨU** 2](#_Toc167658117)

[**CHƯƠNG 2: XỬ LÝ DỮ LIỆU** 3](#_Toc167658118)

[**2.1** **XÁC ĐỊNH BÀI TOÁN** 3](#_Toc167658119)

[**2.2** **THU THẬP VÀ TỔNG HỢP DỮ LIỆU** 3](#_Toc167658120)

[**2.3** **TIỀN XỬ LÝ DỮ LIỆU** 4](#_Toc167658121)

[**2.3.1** **Đọc dữ liệu** 4](#_Toc167658122)

[**2.3.2** **Xử lý các cột để phù hợp với mô hình học máy** 4](#_Toc167658123)

[**CHƯƠNG 3: TRIỂN KHAI MÔ HÌNH** 11](#_Toc167658124)

[**3.1** **XÂY DỰNG MÔ HÌNH** 11](#_Toc167658125)

[**3.2** **KIỂM ĐỊNH MÔ HÌNH** 11](#_Toc167658126)

[**CHƯƠNG 4: ĐÁNH GIÁ** 12](#_Toc167658127)

[**4.1** **CHỌN MÔ HÌNH PHÙ HỢP** 12](#_Toc167658128)

[**4.2** **THỰC HIỆN MÔ HÌNH** 13](#_Toc167658129)

[**CHƯƠNG 5: THẢO LUẬN** 14](#_Toc167658130)

[**5.1** **ƯU ĐIỂM** 14](#_Toc167658131)

[**5.2** **HẠN CHẾ** 14](#_Toc167658132)

[**5.3** **HƯỚNG PHÁT TRIỂN** 14](#_Toc167658133)

[**CHƯƠNG 6: KẾT LUẬN** 16](#_Toc167658134)

# 

# **CHƯƠNG 1: TỔNG QUAN**

* 1. **LÝ DO CHỌN ĐỀ TÀI**

Vay tiền đóng vai trò then chốt trong việc tạo điều kiện giao dịch tài chính và cho phép cá nhân và doanh nghiệp thực hiện ước mơ của họ [4]. Cho dù là để mua nhà, mở rộng kinh doanh, hay chi trả các chi phí không mong đợi, vay tiền là một phần không thể thiếu của sự phát triển kinh tế và tiến bộ cá nhân [9]. Tuy nhiên, quá trình phê duyệt vay tiền liên quan đến việc ra quyết định phức tạp, trong đó các tổ chức tài chính đánh giá nhiều yếu tố để xác định khả năng trả nợ của người nộp đơn. Khi nhu cầu vay tiền tiếp tục tăng, nhu cầu về cơ chế phê duyệt vay tiền hiệu quả và hiệu quả trở nên ngày càng quan trọng [2].

Theo truyền thống, quá trình phê duyệt vay tiền thường được đặc trưng bởi các phương pháp đánh giá thủ công dựa nhiều vào dữ liệu lịch sử và các tiêu chí cứng nhắc [11]. Những phương pháp truyền thống này thường gặp khó khăn trong việc thích nghi với cảnh quan tài chính đang thay đổi, dẫn đến hiệu suất không cao và quyết định không tối ưu. Nhược điểm của những phương pháp này trở nên đặc biệt rõ ràng khi đối mặt với điều kiện kinh tế phức tạp và động địa, dẫn đến sự trễ trệ, không chính xác, và đôi khi là bỏ qua những người nộp đơn có khả năng trả nợ tiềm năng [8].

Nhờ tiến bộ khoa học kỹ thuật, học máy đã mang lại một sự thay đổi mạnh mẽ trong lĩnh vực phê duyệt vay tiền. Các nghiên cứu khác nhau đã khám phá ứng dụng của các thuật toán học máy để tăng cường độ chính xác và hiệu quả của quá trình ra quyết định [7]. Mặc dù đã có những bước tiến đáng kể trong hướng này, nhưng vẫn tồn tại một khoảng cách đáng kể trong việc hiểu và giải thích cơ sở quyết định của những mô hình phức tạp này. Sự thiếu minh bạch đặt ra một thách thức lớn khi các bên liên quan, bao gồm người nộp đơn và các cơ quan quy định, đang tìm kiếm một hiểu biết rõ ràng hơn về các yếu tố ảnh hưởng đến kết quả phê duyệt vay tiền [8]. Sự thiếu minh bạch này làm suy yếu niềm tin vào quá trình cho vay và gây trở ngại cho việc xác định và giảm thiểu các độ chệch hoặc lỗi tiềm ẩn trong quá trình ra quyết định [1]. Do đó, có một nhu cầu cấp bách cho sự minh bạch và trách nhiệm lớn hơn trong việc phát triển và triển khai các mô hình để đảm bảo sự công bằng và đáng tin cậy trong các thực hành cho vay.

* 1. **MỤC TIÊU ĐỀ TÀI**

Nghiên cứu này tập trung vào việc nâng cao quá trình ra quyết định về việc phê duyệt vay tiền trong bối cảnh kinh tế số thông qua áp dụng một phương pháp học máy có khả năng giải thích. Bằng cách so sánh và lựa chọn từ 5 kỹ thuật học máy khác nhau như Logistic Regression, Support Vector Machines, Decision Tree, Random Forest và XGBoost, nghiên cứu nhấn mạnh vào tiềm năng của phương pháp này trong việc tạo ra một hướng tiếp cận mới mẻ cho các tổ chức tài chính. Mục tiêu của nghiên cứu là đảm bảo rằng quá trình ra quyết định về việc phê duyệt vay tiền được thực hiện một cách hiệu quả, đáng tin cậy và minh bạch hơn trong bối cảnh kinh tế số ngày nay.

Ngoài ra, nghiên cứu cũng nhấn mạnh tầm quan trọng của khả năng giải thích trong việc triển khai các giải pháp học máy, đặc biệt là trong các lĩnh vực có ảnh hưởng kinh tế và xã hội lớn. Sự minh bạch trong quá trình ra quyết định không chỉ giúp tăng cường niềm tin từ phía các bên liên quan mà còn giúp phát hiện và giảm thiểu các sai sót và thiên vị có thể xuất hiện trong quá trình ra quyết định. Điều này làm tăng sự công bằng và đáng tin cậy của hệ thống quyết định và góp phần thúc đẩy sự phát triển bền vững của nền kinh tế và xã hội.

* 1. **PHƯƠNG PHÁP NGHIÊN CỨU**

Bài báo cáo sẽ chi tiết hóa cách thức tiếp cận vấn đề nghiên cứu, bao gồm các bước thực hiện và công cụ được sử dụng để thu thập dữ liệu, phân tích dữ liệu và đánh giá kết quả. Cụ thể:

* Thu thập dữ liệu: Mô tả về nguồn dữ liệu được sử dụng trong báo cáo này.
* Phương pháp phân tích: Đề cập đến các phương pháp phân tích dữ liệu được áp dụng trong nghiên cứu, bao gồm cả phương pháp học máy và các kỹ thuật thống kê. Gồm 5 kỹ thuật như đã đề cập ở phần mục tiêu đề tài.
* Đánh giá kết quả: Mô tả cách tiếp cận để đánh giá hiệu suất và tính hiệu quả của phương pháp được sử dụng gồm: Precision, Recall và F1-Score.
* Tinh chỉnh mô hình: Dùng RandomizedSearchCV để thay đổi các tham số đầu vào trong mô hình học máy cho ra kết quả tối ưu.
* Các bước thực hiện: Trình bày chi tiết về các bước cụ thể mà nghiên cứu đã thực hiện để đạt được kết quả và rút ra các kết luận.

Thông qua việc trình bày một cách chi tiết và logic về phương pháp nghiên cứu, độc giả sẽ có cái nhìn rõ ràng và đầy đủ về cách thức nghiên cứu đã được thực hiện và đạt được kết quả như thế nào.

# **CHƯƠNG 2: XỬ LÝ DỮ LIỆU**

* 1. **XÁC ĐỊNH BÀI TOÁN**

Bài toán về quá trình phê duyệt vay tiền là một thách thức phức tạp trong ngành tài chính hiện đại, đặc biệt là khi đối mặt với sự biến động của cảnh quan kinh tế và yêu cầu ngày càng cao về hiệu quả và minh bạch. Vấn đề cơ bản là làm thế nào để tổ chức tài chính có thể đánh giá khả năng trả nợ của người nộp đơn một cách chính xác và hiệu quả nhất. Đồng thời, cần xác định làm thế nào để tăng cường minh bạch và trách nhiệm trong quá trình phê duyệt vay tiền, từ việc xây dựng các mô hình đến triển khai chúng trong thực tế. Điều này không chỉ giúp tăng cường niềm tin của các bên liên quan vào quá trình cho vay, mà còn giúp giảm thiểu độ chệch và lỗi tiềm ẩn, đồng thời đảm bảo sự công bằng và đáng tin cậy trong các thực hành tài chính.

* 1. **THU THẬP VÀ TỔNG HỢP DỮ LIỆU**

Tập dữ liệu được sử dụng trong nghiên cứu này được lấy từ bài nghiên cứu của Sergio [10]. Tập dữ liệu bao gồm 8 đặc điểm cần thiết liên quan đến đơn xin vay tiền, với biến mục tiêu là phân loại nhị phân của trạng thái duyệt vay (đã duyệt hoặc bị từ chối), và không có dữ liệu thiếu hay trùng lặp. Loại dữ liệu và mô tả cho mỗi đặc điểm được mô tả trong Bảng 1.

|  |  |  |
| --- | --- | --- |
| **Tên cột** | **Kiểu dữ liệu** | **Mô tả đặc trưng tên cột** |
| age | Numeric | Tuổi tác |
| job | Categorical | Loại công việc |
| marital | Categorical | Tình trạng hôn nhân |
| education | Categorical | Trình độ học vấn |
| default | Categorical | Có tín dụng mặc định hay không |
| balance | Numeric | Số dư trung bình hàng tháng ($) |
| housing | Categorical | Có vay mua nhà không? |
| loan | Categorical | Cột được gán nhãn nên cho vay hay không |

***Bảng 1. Mô tả dữ liệu***

Một giai đoạn tiền xử lý kỹ lưỡng đảm bảo tập dữ liệu được chuẩn bị tốt cho việc huấn luyện mô hình. Điều này bao gồm một cuộc kiểm tra kỹ lưỡng về các giá trị thiếu sót trong bất kỳ đặc điểm nào, và nếu nhận diện, các chiến lược phù hợp như điền giá trị hoặc loại bỏ được áp dụng để xử lý dữ liệu bị thiếu sót [3]. Các đặc điểm phân loại trải qua mã hóa one-hot để chuyển đổi chúng thành một định dạng phù hợp cho các thuật toán học máy. Ngoài ra, tập dữ liệu được chia thành các tập huấn luyện và kiểm tra, với 80% được phân bổ cho huấn luyện và 20% cho kiểm tra. Việc phân chia chiến lược này đảm bảo mô hình được huấn luyện trên một phần đáng kể của dữ liệu trong khi vẫn giữ một tập dữ liệu riêng và độc lập để đánh giá hiệu suất của nó một cách không thiên vị [5].

* 1. **TIỀN XỬ LÝ DỮ LIỆU**
     1. **Đọc dữ liệu**

Trước hết cần import các thư viện cần thiết cho quá trình xử lý dữ liệu vào. Dùng câu lệnh “pip install (tên thư viện)” trong Terminal của Python để tải thư viện về máy.

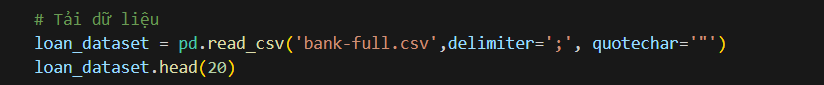
A screen shot of a computer

Description automatically generated

***Hình 1. Đoạn mã sử dụng các thư viện trong Python***

Sau đó, tác giả dùng thư viện pandas để đọc dữ liệu.

A black screen with white text

Description automatically generated

***Hình 2. Hiển thị dữ liệu đã đọc***

Có thể thấy rằng tập dữ liệu này bao gồm 8 đặc trưng, trong đó chỉ có 2 cột là thuộc kiểu giá trị số, còn lại các cột khác bắt buộc phải được tinh chỉnh về dạng số để có thể sử dụng trong mô hình máy học. Việc tinh chỉnh dữ liệu này đòi hỏi một quy trình tiền xử lý cẩn thận và chi tiết nhằm đảm bảo chất lượng và tính toàn vẹn của dữ liệu trước khi đưa vào huấn luyện mô hình.

* + 1. **Xử lý các cột để phù hợp với mô hình học máy**
       1. **Chỉnh các cột có giá trị nhị phân để xử lý**

Các cột “default”, “housing” và “loan” là những cột có giá trị nhị phân, với các giá trị “no” và “yes”. Để sử dụng những cột này trong mô hình, cần phải chuyển đổi chúng thành các giá trị số, cụ thể là chuyển đổi “no” thành giá trị 0 và “yes” thành giá trị 1.

A black screen with white text

Description automatically generated**A computer code with colorful text

Description automatically generated**

***Hình 3. Dữ liệu các cột có giá trị nhị phân sau khi được chuyển đổi***

Việc chuyển đổi các cột nhị phân này không chỉ giúp cho mô hình máy học xử lý dễ dàng hơn mà còn giúp tăng cường khả năng diễn giải kết quả. Khi các giá trị nhị phân được chuyển đổi thành số, các thuật toán máy học có thể trực tiếp sử dụng chúng trong các tính toán mà không cần thêm các bước xử lý phức tạp.

* + - 1. **Chuyển đổi cột education, marital, job**

***Cột education:***

***A graph of blue and orange bars

Description automatically generated***

***Hình 4. Biểu đồ countplot phân loại theo trình độ học vấn***

Dựa vào biểu đồ, tác giả xóa cột mang giá trị “unknown” vì chiếm tỷ trọng nhỏ, và không có định danh cụ thể. Sau đó dữ liệu sẽ được mã hóa, giá trị của cột “education” sẽ chuyển từ các chuỗi ký tự biểu thị trình độ học vấn sang các giá trị số tương ứng. Điều này giúp cho các mô hình máy học có thể xử lý dữ liệu hiệu quả hơn và khai thác được các mối quan hệ tiềm ẩn giữa trình độ học vấn và khả năng trả nợ của người vay.

Cột "education" có ý nghĩa quan trọng vì nó có thể ảnh hưởng đến quyết định phê duyệt vay tiền. Trình độ học vấn cao thường liên quan đến các yếu tố như thu nhập cao hơn, ổn định công việc tốt hơn, và do đó, rủi ro thấp hơn trong việc hoàn trả khoản vay.

***A black screen with white text

Description automatically generatedA screen shot of a computer code

Description automatically generated***

***Hình 5. Cột education sau khi được mã hóa***

***Cột marital:***

***A graph of a married couple

Description automatically generated with medium confidence***

***Hình 6. Biểu đồ countplot phân loại tình trạng hôn nhân***

Tương tự như cột “education”, việc mã hóa dữ liệu cho cột “marital” là một bước quan trọng trong quá trình tiền xử lý dữ liệu để chuẩn bị cho mô hình máy học. Cột “marital” biểu thị tình trạng hôn nhân của người nộp đơn, với các giá trị phân loại như “single” (độc thân), “married” (đã kết hôn), và “divorced” (ly hôn). Thông tin này có thể rất hữu ích trong việc đánh giá khả năng trả nợ của người nộp đơn, vì tình trạng hôn nhân có thể ảnh hưởng đến thu nhập, chi tiêu và trách nhiệm tài chính của họ.

A screen shot of a computer program

Description automatically generated

***Hình 7. Cột marital sau khi được mã hóa***

***Cột job:***

***A graph of a number of people

Description automatically generatedHình 8. Biểu đồ countplot phân loại công việc***

Tương tự như cột “education” và “marital”, việc mã hóa cột “job” cũng là một bước quan trọng trong quá trình tiền xử lý dữ liệu để chuẩn bị cho mô hình máy học. Cột “job” thường biểu thị ngành nghề hoặc công việc của người nộp đơn vay, với các giá trị phân loại như “admin.” (hành chính), “technician” (kỹ thuật viên), “services” (dịch vụ), và các ngành nghề khác. Thông tin về nghề nghiệp có thể cung cấp cái nhìn sâu sắc về thu nhập, tính ổn định tài chính và khả năng trả nợ của người nộp đơn. Bằng cách loại bỏ các giá trị “unknown” và mã hóa các giá trị còn lại sẽ đảm bảo rằng dữ liệu được chuẩn bị tốt nhất cho quá trình huấn luyện mô hình.

A black screen with white text

Description automatically generatedA computer code on a black background

Description automatically generated

***Hình 9. Cột job sau khi được mã hóa***

* + - 1. **Chuyển đổi cột age**

Vì cột age có độ trải dữ liệu từ 18 tuổi đến 95 tuổi nên tác giả sẽ tiến hành chia nhóm tuổi. Việc chia nhóm tuổi là một phần quan trọng trong việc tiền xử lý dữ liệu khi chuẩn bị cho mô hình máy học, giúp tạo ra các nhóm có mật độ dữ liệu cân đối và giảm thiểu sự biến động giữa các nhóm tuổi.

Công thức được sử dụng để tính toán số lượng nhóm (k) dựa trên số lượng quan sát (n) trong cột “age”. Công thức này dựa trên ý tưởng về sự đồng đều về mật độ dữ liệu giữa các nhóm khi chia tỷ lệ theo căn bậc ba của tổng số quan sát.

A screenshot of a computer

Description automatically generated

***Hình 10. Cột age sau khi được mã hóa***

* + - 1. **Chuẩn hóa các giá trị dùng MinMaxScaler**

A screenshot of a computer screen

Description automatically generatedMinMaxScaler sẽ đảm bảo phạm vị đồng nhất, chuyển đổi dữ liệu sao cho tất cả các đặc trưng nằm trong phạm vi [0, 1]. Điều này giúp đồng nhất phạm vi của tất cả các đặc trưng, loại bỏ sự không cân đối trong đầu vào và giúp mô hình học máy hội tụ nhanh hơn.

A screen shot of a computer code

Description automatically generated

***Hình 11. Dữ liệu sau khi được chuẩn hóa MinMaxScaler***

# **CHƯƠNG 3: TRIỂN KHAI MÔ HÌNH**

* 1. **XÂY DỰNG MÔ HÌNH**

Trước khi xây dựng mô hình, tác giả tiến hành chia dữ liệu thành 2 bảng phục vụ cho quá trình học máy. Một bảng gán giá trị X sẽ gồm các cột không bao gồm cột “loan”, bảng được gán cho Y sẽ chỉ gồm cột nhãn gán “loan”.

**A computer screen with colorful text

Description automatically generated**

***Hình 10. Chia dữ liệu thành 2 bảng***

Tiếp tục tiến hành xây dựng mô hình dựa trên 2 thư viện “scikit-learn” và “xgboost” để áp dụng thuật toán Logistic Regression, Support Vector Machines (SVM), Decision Tree, Random Forest, XGBoost.

* 1. **KIỂM ĐỊNH MÔ HÌNH**

Phương pháp kiểm định mô hình tác giả sử dụng là phương pháp tách tập dữ liệu thành tập huấn luyện và tập kiểm tra kết hợp với kỹ thuật cross-validation để đánh giá hiệu suất của mô hình. Dưới đây là mô tả chi tiết về phương pháp kiểm định mô hình:

* Dữ liệu được chia thành hai phần: một phần dành cho việc huấn luyện mô hình (tập huấn luyện) và một phần dành cho việc kiểm tra hiệu suất của mô hình (tập kiểm tra). Tỉ lệ phân chia được thiết lập là 80% cho tập huấn luyện và 20% cho tập kiểm tra.
* Đánh giá hiệu suất trên tập kiểm tra: Sau khi mô hình được huấn luyện, nó được sử dụng để dự đoán nhãn cho dữ liệu trong tập kiểm tra (X\_test). Các chỉ số đánh giá như accuracy, precision, recall và F1-score được tính toán dựa trên so sánh giữa nhãn dự đoán và nhãn thực tế trên tập kiểm tra [6].
* Cross-validation: Để đánh giá hiệu suất của mô hình một cách chính xác hơn và tránh overfitting, kỹ thuật cross-validation được sử dụng. Bài báo cáo sử dụng StratifiedKFold với 5 folds để chia dữ liệu thành 5 phần bằng nhau, mỗi lần 4 phần được sử dụng để huấn luyện mô hình và phần còn lại được sử dụng để đánh giá mô hình. Quá trình này được lặp lại 5 lần (tương ứng với số fold) và đánh giá được thực hiện trên các fold khác nhau. Kết quả cuối cùng là sự trung bình và độ lệch chuẩn của các chỉ số đánh giá từ các fold.

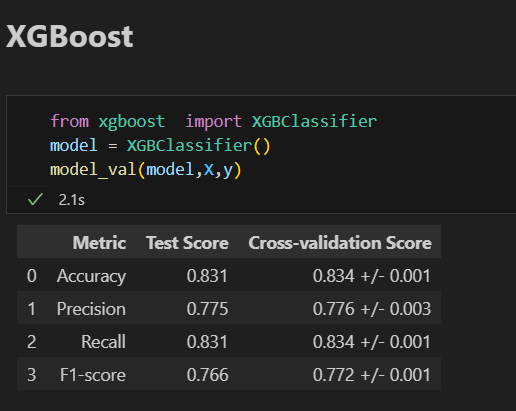
# **CHƯƠNG 4: ĐÁNH GIÁ**

* 1. **CHỌN MÔ HÌNH PHÙ HỢP**

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer program

Description automatically generatedSau khi chạy một hàm đánh giá đối với 5 mô hình học máy đã đề cập. Tác giả nhận thấy lựa chọn mô hình XGBoost là có độ chính xác cũng như chỉ số precision cao nhất. Lựa chọn mô hình XGBoost không chỉ dựa trên kết quả thử nghiệm ban đầu mà còn dựa trên quá trình kiểm tra kỹ lưỡng và so sánh với các mô hình khác. Kết quả này cung cấp một cơ sở vững chắc cho việc sử dụng XGBoost trong quy trình phê duyệt khoản vay, đảm bảo rằng các quyết định được thực hiện một cách chính xác, đáng tin cậy và hiệu quả hơn trong môi trường kinh tế số hiện nay.

***Hình 11. Lựa chọn mô hình hiệu quả nhất***

Sau khi lựa chọn mô hình XGBoost, tác giả tiếp tục sử dụng kỹ thuật điều chỉnh siêu tham số (Hyperparameter Tuning) để tối ưu hóa hiệu suất của mô hình. Quá trình này được thực hiện bằng cách sử dụng Randomized Search, một phương pháp tìm kiếm ngẫu nhiên trong không gian siêu tham số để tìm ra các giá trị tốt nhất. Kết quả của quá trình tìm kiếm cho thấy mô hình đạt được độ chính xác (accuracy) là 83% và chỉ số precision là 78,1%.

* 1. **THỰC HIỆN MÔ HÌNH**

Để tạo ra một công cụ tiện lợi và thân thiện với người dùng cho việc phê duyệt hồ sơ vay tiền, tác giả đã xây dựng một giao diện đơn giản bằng thư viện Tkinter của Python. Tkinter là một thư viện chuẩn cho GUI (Giao diện Người dùng Đồ họa) trong Python, cho phép tạo ra các ứng dụng có giao diện người dùng một cách dễ dàng và hiệu quả.

**A screenshot of a computer

Description automatically generated**Giao diện này bao gồm nhiều ô nhập liệu để người dùng có thể nhập vào các thông tin cần thiết cho việc đánh giá hồ sơ vay tiền. Các thông tin này bao gồm tuổi, loại công việc, tình trạng hôn nhân, trình độ học vấn, tình trạng tín dụng, số dư tài khoản trung bình hàng tháng cũng như tình trạng vay mua nhà.

Khi người dùng đã nhập đầy đủ thông tin vào các ô nhập liệu và nhấn nút “Dự đoán”, hệ thống sẽ sử dụng mô hình học máy XGBoost đã được huấn luyện và tối ưu hóa để đưa ra dự đoán về việc hồ sơ có được phê duyệt vay tiền hay không. Kết quả dự đoán sẽ được hiển thị trên giao diện người dùng. Nếu hồ sơ được phê duyệt, hệ thống sẽ hiển thị thông báo “Được duyệt”, ngược lại là “Không được duyệt”.

***Hình 12. Giao diện phần mềm hồ sơ cho vay***

# **CHƯƠNG 5: THẢO LUẬN**

* 1. **ƯU ĐIỂM**

Nghiên cứu này đã mang lại nhiều ưu điểm quan trọng trong việc nâng cao quy trình phê duyệt khoản vay bằng cách sử dụng mô hình học máy. Đầu tiên, việc áp dụng XGBoost cùng với các kỹ thuật như Random Search để điều chỉnh siêu tham số đã giúp mô hình đạt được độ chính xác cao (83%) cũng như chỉ số precision cao (78,1%), đảm bảo rằng quá trình ra quyết định về việc phê duyệt vay tiền được thực hiện một cách hiệu quả và đáng tin cậy hơn. Điều này chứng tỏ XGBoost có khả năng dự đoán chính xác và ổn định hơn so với 4 kỹ thuật học máy khác nhau bao gồm Logistic Regression, Support Vector Machines, Decision Tree, Random Forest. Sự cải tiến dùng mô hình học máy này giúp các tổ chức tài chính có thể xử lý các yêu cầu vay tiền một cách nhanh chóng và chính xác, đồng thời giảm thiểu rủi ro từ các quyết định sai lầm.

* 1. **HẠN CHẾ**

Mặc dù phương pháp XGBoost cho thấy nhiều ưu điểm nổi bật, nhưng nghiên cứu cũng gặp phải một số hạn chế. Một trong những hạn chế chính là việc mô hình có thể bị ảnh hưởng bởi sự không cân bằng dữ liệu, khi mà số lượng các trường hợp được phê duyệt vay tiền và không được phê duyệt không tương đồng nhau. Điều này có thể dẫn đến việc mô hình thiên vị và không phản ánh chính xác thực tế. Bên cạnh đó, quá trình huấn luyện mô hình XGBoost yêu cầu thời gian và tài nguyên tính toán đáng kể, điều này có thể là một trở ngại đối với các tổ chức có hạn chế về công nghệ và tài chính. Ngoài ra, việc điều chỉnh siêu tham số (Hyperparameter tuning) mặc dù giúp cải thiện hiệu suất nhưng lại khá phức tạp và tốn thời gian.

* 1. **HƯỚNG PHÁT TRIỂN**

Để khắc phục các hạn chế hiện tại và tiếp tục nâng cao hiệu suất của mô hình, nghiên cứu trong tương lai có thể tập trung vào một số hướng phát triển sau:

* Cân bằng dữ liệu: Sử dụng các kỹ thuật như oversampling hoặc undersampling để xử lý sự không cân bằng trong dữ liệu, nhằm đảm bảo mô hình học được từ cả hai nhóm dữ liệu (phê duyệt và không phê duyệt) một cách cân bằng.
* Cải tiến thuật toán: Thử nghiệm và áp dụng các thuật toán mới hoặc các biến thể của XGBoost, chẳng hạn như LightGBM hoặc CatBoost, để xem xét khả năng cải thiện hiệu suất và giảm thời gian huấn luyện.
* Học sâu (Deep Learning): Áp dụng các mô hình học sâu để khai thác tiềm năng của các mạng neuron trong việc ra quyết định phê duyệt vay tiền, đồng thời so sánh hiệu suất với các mô hình học máy hiện tại.
* Tích hợp thêm dữ liệu: Mở rộng tập dữ liệu bằng cách tích hợp thêm các nguồn dữ liệu khác, chẳng hạn như lịch sử tín dụng, hành vi giao dịch và các thông tin tài chính khác để tăng cường khả năng dự đoán của mô hình.
* Giải thích và minh bạch hóa: Sử dụng các phương pháp như LIME hoặc SHAP để giải thích rõ ràng hơn các quyết định của mô hình, giúp các tổ chức tài chính tin tưởng hơn vào các quyết định dựa trên mô hình học máy.

Nhờ vào những hướng phát triển này, mô hình phê duyệt vay tiền có thể trở nên mạnh mẽ và đáng tin cậy hơn, góp phần hỗ trợ các tổ chức tài chính ra quyết định một cách chính xác và minh bạch hơn trong bối cảnh kinh tế số đang ngày càng phát triển.

# **CHƯƠNG 6: KẾT LUẬN**

Trong bài báo cáo này, tác giả đã tạo và đánh giá các mô hình học máy cho khả năng chấp nhận vay. Để hiểu tập dữ liệu và nắm bắt quy trình phê duyệt vay, bài viết bắt đầu bằng việc thực hiện phân tích dữ liệu khám phá. Sau đó, thực hiện huấn luyện và đánh giá một số mô hình phân loại, bao gồm Logistic Regression, Support Vector Machines, Decision Tree, Random Forest và XGBoost. Dựa trên độ chính xác làm tiêu chí đánh giá hiệu suất của các mô hình này, tác giả phát hiện ra rằng mô hình phân loại XGBoost vượt trội hơn các mô hình khác và có độ chính xác cao nhất là 83% trên tập kiểm tra. Do đó, có thể kết luận rằng mô hình XGBoost là hiệu quả trong dự báo việc phê duyệt vay dựa trên các đặc điểm được cung cấp. Phương pháp được trình bày trong nghiên cứu này cung cấp một cơ cấu vững chắc và dễ hiểu có thể đóng vai trò quan trọng trong việc định hình tương lai của việc ra quyết định tài chính. Nó tạo ra một bức tranh tài chính hỗn hợp, hiệu quả và minh bạch hơn, phù hợp với các yêu cầu phát triển của nền kinh tế số đang diễn ra. Các mô hình đã đạt được kết quả khích lệ, nhưng vẫn còn tiềm năng để phát triển và tiến hành nghiên cứu bổ sung.

**TÀI LIỆU THAM KHẢO**

[1]. Brotcke, L. (2022), “Time to Assess Bias in Machine Learning Models for Credit Decisions”, *Journal of Risk and Financial Management*, tập 15(4), 165.

https://doi.org/10.3390/jrfm15040165

[2]. Dansana, D., Patro, S. G. K., Mishra, B. K., Prasad, V., Razak, A., & Wodajo, A. W. (2023), “Analyzing the impact of loan features on bank loan prediction using the Random Forest algorithm”, *Engineering Reports*.

[3]. Hui, S. H., Khai, W. K., XinYing, C., & Wai, P. W. (2023), “Prediction of customer churn for ABC Multistate Bank using machine learning algorithms”, *Malaysian Journal of Computing (MJoC)*, tập 8(2), 1602-1619.

[4]. Kariv, D., & Coleman, S. (2015), “Toward a theory of financial bricolage: the impact of small loans on new businesses”, *Journal of Small Business and Enterprise Development*, tập 22(2), 196-224.

[5]. Noviandy, T. R., Idroes, G. M., Maulana, A., Hardi, I., Ringga, E. S., & Idroes, R. (2023), “Credit Card Fraud Detection for Contemporary Financial Management Using XGBoostDriven Machine Learning and Data Augmentation Techniques”, *Indatu Journal of Management and Accounting*, tập 1(1), 29-35.

[6]. Noviandy, T. R., Maulana, A., Idroes, G. M., Irvanizam, I., Subianto, M., & Idroes, R. (2023), “QSAR-Based Stacked Ensemble Classifier for Hepatitis C NS5B Inhibitor Prediction”, *2023 2nd International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE),* 220–225.

[7]. Orji, U. E., Ugwuishiwu, C. H., Nguemaleu, J. C. N., & Ugwuanyi, P. N. (2022), “Machine Learning Models for Predicting Bank Loan Eligibility”, *2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development*

*(NIGERCON)*, 1-5.

[8]. Purificato, E., Lorenzo, F., Fallucchi, F., & De Luca, E. W. (2023), “The Use of Responsible Artificial Intelligence Techniques in the Context of Loan Approval Processes”, *International Journal of Human–Computer Interaction*, tập 39(7), 1543-1562.

[9]. Saiti, D., & Trenovski, B. (2022). “The impact of loans and interest rates on economic growth in the Republic of North Macedonia”, *Knowledge-International Journal*, tập 50(1), 15-20.

[10]. S. Moro, R. Laureano and P. Cortez (2011), “Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology”, *Proceedings of the European Simulation and Modelling Conference - ESM'2011*, 117-121.

[11]. Tchakoute Tchuigoua, H. (2018), “Which types of microfinance institutions decentralize the loan approval process?”, *The Quarterly Review of Economics and Finance*, tập 67, 237-244.