**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT**

**KHOA TÀI CHÍNH – NGÂN HÀNG**

**🙠🕮🙢**

**A blue and white circle with text

Description automatically generated with low confidence**

**ĐỒ ÁN CUỐI KỲ - HỌC KỲ 2 (2022 – 2023)**

**MÔN HỌC: PHÂN TÍCH DỮ LIỆU**

**ĐỀ TÀI:**

**PHÂN TÍCH VÀ TRỰC QUAN HÓA DỮ LIỆU TÍN DỤNG KHÁCH HÀNG SỬ DỤNG DỮ LIỆU VAY VỐN**

**Danh sách sinh viên thực hiện:**

|  |  |
| --- | --- |
| K214142052 | Lê Như Ái |
| K214051252 | Nguyễn Thị Tuyết Ngân |
| A picture containing picture frame, rectangle, frame  Description automatically generated K214142065 | Võ Nguyễn Bảo Hân |
| K214142093 | Ngô Thị Minh Trinh |
| K214142067 | Hồ Phạm Hữu |

**Giảng viên hướng dẫn: TS. Nguyễn Thôn Dã**

**TP. Hồ Chí Minh, ngày 29 tháng 5 năm 2023**

**DANH SÁCH THÀNH VIÊN**

|  |  |  |
| --- | --- | --- |
| **Họ và tên** | **MSSV** | **Email** |
| Lê Như Ái | K214142052 | ailn21414@st.uel.edu.vn |
| Nguyễn Thị Tuyết Ngân | K214051252 | nganntt21405@st.uel.edu.vn |
| Võ Nguyễn Bảo Hân | K214142065 | hanvnb21414@st.uel.edu.vn |
| Ngô Thị Minh Trinh | K214142093 | trinhntm21414@st.uel.edu.vn |
| Hồ Phạm Hữu  (NHÓM TRƯỞNG) | K214142067 | huuhp21414@st.uel.edu.vn |

**Lời cảm ơn**

Lời đầu tiên, nhóm chúng em xin chân thành cảm ơn Ban giám hiệu trường Đại học Kinh tế - Luật vì đã tạo điều kiện cho chúng em được tiếp cận với môn “ Phân tích dữ liệu” cũng như về cơ sở vật chất với hệ thống thư viện hiện đại, đa dạng các loại sách, tài liệu thuận lợi cho việc tìm kiếm, nghiên cứu thông tin môn học. Đặc biệt hơn, chúng em xin bày tỏ lòng biết ơn sâu sắc đến giảng viên bộ môn - thầy Nguyễn Thôn Dã đã hỗ trợ và giảng dạy những kiến thức quý báu cho chúng em trong thời gian vừa qua.

Trong quá trình hoàn thành đồ án cuối kỳ “Phân tích và trực quan hóa dữ liệu tín dụng khách hàng sử dụng dữ liệu vay vốn” giảng viên Nguyễn Thôn Dã đã tận tình chỉ dẫn và đưa ra nhiều lời khuyên bổ ích giúp chúng em có thể hoàn thành tốt đồ án của mình. Tuy nhiên, do thời gian và kinh nghiệm thực tiễn còn hạn chế nên chúng em không thể tránh khỏi những sai sót trong quá trình hoàn thành đồ án. Em rất mong có thể nhận được những ý kiến đóng góp, nhận xét của quý thầy, cô để chúng em có thể sổ sung, chỉnh sửa và học hỏi thêm được nhiều kiến thức cho thực tiễn sau này.

Kính chúc thầy có thật nhiều sức khỏe, thành công và hạnh phúc.

Chúng em xin chân thành cảm ơn!

*Nhóm sinh viên thực hiện tiểu luận*

*TP.HCM, ngày 29 tháng 05 năm 2023*

**MỤC LỤC**

[A. Giới thiệu (Introduction) 6](#_Toc136177406)

[B. Các nghiên cứu liên quan (Related work) 8](#_Toc136177407)

[C. Nền tảng lý thuyết (Background) 9](#_Toc136177408)

[**1. Tổng quan về lĩnh vực phân tích tín dụng: 9**](#_Toc136177409)

[**2. Dữ liệu vay vốn. 11**](#_Toc136177410)

[**3. Ngôn ngữ lập trình Python 11**](#_Toc136177411)

[**4. Mô Hình Máy Học 13**](#_Toc136177412)

[**5. Đánh giá hiệu suất mô hình: 14**](#_Toc136177413)

[D. Phương pháp luận nghiên cứu 15](#_Toc136177414)

[**1. Mô hình tổng quan 15**](#_Toc136177415)

[**2. Thu thập dữ liệu 16**](#_Toc136177416)

[**3. Mô tả dữ liệu 17**](#_Toc136177417)

[**4. Tiền xử lý dữ liệu và trích xuất đặc trưng 25**](#_Toc136177418)

[E. Kết quả thử nghiệm và phân tích 27](#_Toc136177419)

[**1. Trực quan hóa dữ liệu dưới dạng biểu đồ và phân tích 27**](#_Toc136177420)

[**2. Kết quả thử nghiệm và phân tích đối với mô hình máy học 36**](#_Toc136177421)

[F. Kết luận 38](#_Toc136177422)

[**1 Ưu điểm: 38**](#_Toc136177423)

[**2. Hạn chế: 39**](#_Toc136177424)

[**3. Phương hướng phát triển 39**](#_Toc136177425)

**DANH MỤC HÌNH ẢNH**

[Hình 1. Mô hình tổng quan 17](#_Toc136179428)

[Hình 2. Các yếu tố từ khảo sát và phỏng vấn 19](#_Toc136179429)

[Hình 3. Sơ đồ các bước tiền xử lý dữ liệu 27](#_Toc136179430)

[Hình 4. Biểu đồ tròn cho Loan\_Period 29](#_Toc136179431)

[Hình 5. Biểu đồ Scatter plot cho 'Loan\_Period' và 'Interest\_Rate' 29](#_Toc136179432)

[Hình 6. Biểu đồ countplot cho biến 'Approved' 30](#_Toc136179433)

[Hình 7. Biểu đồ barplot giữa 'Approved' và 'Monthly\_Income' 30](#_Toc136179434)

[Hình 8. Biểu đồ tròn cho 'Employer\_Category\_1' 31](#_Toc136179435)

[Hình 9. Biểu đồ Barplot giữa 'Loan\_Amount' và 'Employer\_Category' 31](#_Toc136179436)

[Hình 10. Biểu đồ tròn cho 'Employer\_Category\_2' 32](#_Toc136179437)

[Hình 11. Biểu đồ barplot giữa 'Loan\_Amount' và 'Employer\_Category\_2' 32](#_Toc136179438)

[Hình 12. Biểu đồ giữa 'Monthly\_Income' và 'Employer\_Category\_2' 33](#_Toc136179439)

[Hình 13. Biểu đồ tròn cho 'Contacted' 33](#_Toc136179440)

[Hình 14. Biểu đồ giữa 'Monthly\_Income' và 'City\_Category' 34](#_Toc136179441)

[Hình 15. Biểu đồ tròn cho 'Var1' 34](#_Toc136179442)

[Hình 16. Biểu đồ cho 'Employer\_Category\_1' và 'EMI' 35](#_Toc136179443)

[Hình 17. Biểu đồ giữa 'Employer\_Category\_1' và 'Loan\_Amount' 36](#_Toc136179444)

[Hình 18. Biểu đồ giữa 'Employer\_Category\_2' và 'EMI' 36](#_Toc136179445)

[Hình 19. Biểu đồ PairGrid và Heatmap 37](#_Toc136179446)

**Tóm tắt dự án**

Dự án "Phân tích tín dụng khách hàng sử dụng dữ liệu vay vốn" là một dự án sử dụng ngôn ngữ lập trình Python để thực hiện phân tích và đánh giá rủi ro tín dụng của khách hàng dựa trên dữ liệu vay vốn. Mục tiêu của dự án là phân tích các yếu tố liên quan từ dữ liệu từ đó xây dựng một hệ thống thông minh để đánh giá khả năng hoàn trả nợ của khách hàng dựa trên các chỉ số tài chính và thông tin về lịch sử vay vốn.

Dự án bao gồm các bước chính sau:

1. Tiền xử lý dữ liệu: Dữ liệu vay vốn được thu thập từ các nguồn khác nhau và cần được xử lý để loại bỏ dữ liệu thiếu, nhiễu và không cần thiết. Các bước tiền xử lý bao gồm làm sạch dữ liệu, điền giá trị thiếu và chuyển đổi dữ liệu sang định dạng phù hợp cho việc phân tích.

2. Trích xuất đặc trưng: Dữ liệu vay vốn ban đầu có thể chứa nhiều thông tin không cần thiết. Trong bước này, các đặc trưng quan trọng được chọn lọc từ dữ liệu, bao gồm thông tin về thu nhập, lịch sử thanh toán, mục đích vay vốn, tình trạng công việc, v.v. Các đặc trưng này sẽ được sử dụng để xây dựng mô hình dự đoán.

3. Trực quan hóa dữ liệu: Trực quan hóa dữ liệu là quá trình biểu diễn dữ liệu số hoặc thông tin trừu tượng dưới dạng hình ảnh, biểu đồ hoặc đồ họa để giúp hiểu và phân tích dữ liệu một cách dễ dàng và trực quan. Dự án trực quan hóa các dữ liệu đơn biến cho đến đa biến.

4. Xây dựng mô hình: Sử dụng các kỹ thuật học máy và phân tích dữ liệu, một mô hình dự đoán sẽ được xây dựng để đánh giá khả năng hoàn trả nợ của khách hàng. Các mô hình phổ biến như Logistic Regression, Decision Tree, KNeighborsClassifier được áp dụng trong dự án này.

5. Đánh giá mô hình: để đánh giá hiệu suất của mô hình, các phép đo như Accuracy, Precision, Recall, F1-score và kappa score được sử dụng. Điều này giúp đánh giá khả năng dự đoán và độ chính xác của mô hình trong việc phân loại tín dụng.

6. Triển khai và tích hợp: Sau khi mô hình đã được xác định và đánh giá, nó có thể được triển khai vào một ứng dụng hoặc hệ thống để sử dụng thực tế. Ví dụ, một ứng dụng web có thể được xây dựng để cho phép người dùng nhập thông tin cá nhân và nhận được dự đoán về khả năng hoàn trả nợ của họ.

Tóm lại, dự án "Phân tích tín dụng khách hàng sử dụng dữ liệu vay vốn" sử dụng Python để thực hiện phân tích dữ liệu, xây dựng mô hình học máy và đánh giá hiệu suất. Nó mang lại thông tin quan trọng và đáng tin cậy về tín dụng khách hàng, giúp tổ chức tài chính đưa ra quyết định thông minh và quản lý rủi ro hiệu quả trong việc cấp tín dụng.

**Keywords:** Phân tích tín dụng; Dữ liệu vay vốn; Python; Máy học; Đánh giá hiệu suất; Phân loại tín dụng; Big Data; Trực quan hóa dữ liệu.

# A. Giới thiệu (Introduction)

Sự phát triển vượt bậc của Internet và trí tuệ nhân tạo đã mang đến sự thay đổi đáng kể trong cách ngân hàng quản lý, phân tích và xử lý dữ liệu. Trong số những thách thức mà ngân hàng phải đối mặt, việc phân tích tín dụng khách hàng, dự báo khả năng trả nợ và quyết định cho vay luôn là một trong những vấn đề quan trọng. Việc phân tích tín dụng của khách hàng có liên quan mật thiết đến khả năng thanh khoản của ngân hàng, vì sản phẩm và dịch vụ ngân hàng liên quan chặt chẽ đến tài chính của khách hàng. Đặc biệt, với hàng triệu khách hàng (trong thị trường Việt Nam), việc chăm sóc và phát hiện sớm khả năng thanh toán khoản vay của khách hàng chính là một thách thức.

Việc ứng dụng Big Data và trí tuệ nhân tạo (AI) trong ngành ngân hàng là một quá trình phức tạp và đòi hỏi sự cam kết từ các cấp lãnh đạo, và thành công nó mang lại sẽ không thấy được ngay lập tức mà cần phải có thời gian nhất định. Quá trình này bắt đầu từ việc xây dựng chiến lược, thu thập và chuẩn hóa dữ liệu, huấn luyện các mô hình máy học, và sử dụng thông tin thu được từ bộ dữ liệu để đưa ra quyết định. Mặc dù ở Việt Nam, một số ngân hàng đã bắt đầu ứng dụng Big Data và AI vào hoạt động tín dụng, nhưng nhìn chung họ chỉ mới áp dụng một số tính năng chính mà chưa mở rộng hoàn toàn. Để đạt được những lợi ích to lớn từ việc sử dụng Big Data và AI, ngân hàng cần tiếp tục đầu tư và mở rộng áp dụng các tính năng và công nghệ này vào các lĩnh vực khác nhau trong quá trình phân tích tín dụng và quản lý rủi ro. Sự tiến bộ trong việc áp dụng Big Data và AI sẽ giúp ngân hàng tăng cường khả năng ra quyết định thông minh, nâng cao hiệu suất hoạt động và cung cấp dịch vụ tài chính tốt hơn cho khách hàng, việc phân tích tín dụng khách hàng để ra quyết định cho vay cũng không ngoại lệ.

Quay lại thời gian trước đây, để đánh giá tín dụng của khách hàng và quyết định cho vay, các ngân hàng đã sử dụng các phương pháp và quy trình truyền thống. Sau đây là những phương pháp thông thường mà họ đã áp dụng trước khi sự phát triển của trí tuệ nhân tạo (AI) được đưa vào sử dụng: Xem xét lịch sử tín dụng; Đánh giá thu nhập và tài sản; Kiểm tra thông tin cá nhân; Xem xét thông tin công ty; Đánh giá tài chính tổ chức; Kiểm tra thông tin tham chiếu. Các phương pháp truyền thống này dựa trên các quy tắc, kinh nghiệm và đánh giá thủ công từ các nhân viên, chuyên viên ngân hàng. Vậy nên nó tốn khá nhiều thời gian và công sức, độ chính xác không hoàn toàn có thể tin tưởng và phải phụ thuộc nhiều vào kinh nghiệm cũng như cách phân tích của mỗi người. Tuy nhiên, việc sử dụng trí tuệ nhân tạo và các kỹ thuật phân tích dữ liệu mới đã mang lại sự cải tiến đáng kể trong quá trình phân tích tín dụng và đưa ra quyết định cho vay, giúp ngân hàng tăng cường hiệu suất và đảm bảo tính chính xác cao hơn. Có thể thấy, việc áp dụng công nghệ vào lĩnh vực tài chính - ngân hàng nói chung và phân tích tín dụng khách hàng nói riêng là vô cùng cần thiết.

Phân tích tín dụng đóng vai trò vô cùng quan trọng trong quá trình đánh giá độ tin cậy của khách hàng và đưa ra quyết định cho vay thông minh của các ngân hàng và tổ chức tài chính. Với sự gia tăng về lượng dữ liệu cho vay như hiện nay, tiềm năng của việc sử dụng những dữ liệu này là rất lớn. Tuy nhiên, trong thực tế, việc hình dung, khám phá và sử dụng những dữ liệu này thường gặp nhiều khó khăn và phức tạp, đặc biệt đối với những người không có chuyên môn và không có kinh nghiệm với công nghệ và dữ liệu máy tính. Để giải quyết thách thức này, sự áp dụng các kỹ thuật phân tích dữ liệu tiên tiến và trực quan hóa dữ liệu đã trở thành một công cụ hỗ trợ đắc lực trong việc phân tích tín dụng khách hàng. Nhờ vào những kỹ thuật này, người dùng có thể dễ dàng khám phá thông tin quan trọng từ bộ dữ liệu cho vay và hiểu rõ hơn về các yếu tố ảnh hưởng đến khả năng trả nợ của khách hàng cũng như tỷ lệ chấp nhận cho vay của ngân hàng đối với những khách hàng khác nhau. Từ đó, các ngân hàng và tổ chức tài chính có thể nâng cao khả năng đánh giá tín dụng, giảm rủi ro nợ xấu và tối ưu hóa quản lý danh mục cho vay.

Để có cái nhìn tổng quan hơn về khả năng được cấp vay và khả năng trả nợ của khách hàng, dự án nghiên cứu của chúng tôi đã áp dụng tiến bộ của trí tuệ nhân tạo và học máy vào quá trình phân tích. Các thông tin về hành vi khách hàng được thu thập và lưu trữ bởi ngân hàng hoặc các tổ chức tài chính sẽ được phân tích thông qua công nghệ AI để nhận biết khách hàng có khả năng trả nợ cao hay không, tương ứng với tỷ lệ được cấp vay sẽ là bao nhiêu. Thông qua các phương pháp xử lý và phân tích dữ liệu, thành lập mô hình học máy, biểu đồ được thể hiện dựa trên dữ liệu số, chúng ta sẽ có được cái nhìn rõ hơn và chính xác hơn về bộ dữ liệu. Khả năng dự đoán chính xác khả năng trả nợ của khoản vay là điều cần thiết để giảm thiểu rủi ro tín dụng và đảm bảo hoạt động cho vay bền vững. Qua đó, giúp các chuyên gia tài chính, ngân hàng hoàn thiện chiến lược quản lý khách hàng hiệu quả hơn, đưa ra quyết định về việc cấp vay và giảm thiểu rủi ro cho tổ chức.

Sử dụng các kỹ thuật phân tích dữ liệu và trực quan hóa thông tin giúp người dùng khám phá và tương tác với lượng lớn dữ liệu một cách dễ dàng và dễ hiểu hơn. Mục tiêu chính của trực quan hóa thông tin là biến đổi và trình bày dữ liệu dưới dạng biểu đồ, hình ảnh đại diện, nhằm giúp người dùng hiểu rõ hơn về bộ dữ liệu. Trực quan hóa đóng vai trò quan trọng trong việc cung cấp cái nhìn tổng quan về các bộ dữ liệu, các loại dữ liệu chính, thuộc tính và mối quan hệ giữa chúng. Để thực hiện kỹ thuật trực quan hóa thông tin, cần có một số cấu trúc dữ liệu cụ thể. Từ đó, dự án của chúng tôi là tạo ra một hệ sinh thái khai thác và trực quan hóa dữ liệu, khả năng dự báo có độ chính xác đáng tin cậy. Mục tiêu chính của nghiên cứu này gồm hai phần: (1) dự báo khả năng trả nợ của khách hàng dựa trên dữ liệu khoản vay và (2) hỗ trợ các tổ chức tài chính đưa ra quyết định cho vay sáng suốt.

Kết quả của nghiên cứu này sẽ mang lại những đóng góp có giá trị cho lĩnh vực phân tích tín dụng và cung cấp ý nghĩa thực tiễn cho các tổ chức tài chính. Bằng cách khai thác sức mạnh của phân tích dữ liệu, trực quan hóa và lập mô hình, ngân hàng và các tổ chức tài chính khác có thể nâng cao khả năng đánh giá mức độ tin cậy tín dụng một cách chính xác, giảm thiểu rủi ro vỡ nợ và tối ưu hóa quản lý danh mục cho vay.

# B. Các nghiên cứu liên quan (Related work)

Năm 2017, nhóm nghiên cứu gồm Sivarajah và cộng sự đã thực hiện nghiên cứu phân tích tầm quan trọng của Big Data (BD) và các phương pháp phân tích. Nghiên cứu cho thấy phân tích dữ liệu lớn (BDA) đang ngày càng trở thành một xu hướng thực tiễn mà nhiều tổ chức đang áp dụng với mục đích xây dựng thông tin có giá trị từ BD. Quá trình phân tích, bao gồm việc triển khai và sử dụng các công cụ BDA, được các tổ chức coi là một công cụ để cải thiện hiệu quả hoạt động mặc dù nó có tiềm năng chiến lược, thúc đẩy các nguồn doanh thu mới và đạt được lợi thế cạnh tranh so với các đối thủ kinh doanh. Phân tích được trình bày trong bài báo này đã xác định các nghiên cứu BD có liên quan đã đóng góp cả về mặt khái niệm và thực nghiệm cho việc mở rộng và tích lũy của cải trí tuệ cho BDA trong kỷ luật quản lý tài nguyên tổ chức và công nghệ. [1]

Năm 2021, Doko và cộng sự của anh ấy đã thực hiện bài nghiên cứu đánh giá các mô hình phân tích tín dụng khách hàng dựa trên dữ liệu đăng ký tín dụng của ngân hàng trung ương. Mục tiêu của bài viết này là đánh giá các mô hình học máy khác nhau để tạo ra mô hình chính xác cho việc đánh giá rủi ro tín dụng bằng cách sử dụng dữ liệu từ bộ dữ liệu đăng ký tín dụng thực của Ngân hàng Trung ương Cộng hòa Bắc Macedonia. Trong nghiên cứu này, nhóm chuyên gia đã so sánh năm mô hình học máy để phân loại dữ liệu rủi ro tín dụng, đó là hồi quy logistic, cây quyết định, rừng ngẫu nhiên, máy vectơ hỗ trợ (SVM) và mạng nơ-ron. Dự án đánh giá năm mô hình bằng cách sử dụng các chỉ số học máy khác nhau và đề xuất một mô hình dựa trên dữ liệu đăng ký tín dụng từ ngân hàng trung ương với phương pháp chi tiết có thể dự đoán rủi ro tín dụng dựa trên lịch sử tín dụng của người dân trong nước. Kết quả của chúng tôi cho thấy rằng độ chính xác tốt nhất đạt được bằng cách sử dụng cây quyết định thực hiện trên dữ liệu mất cân bằng có và không chia tỷ lệ, tiếp theo là rừng ngẫu nhiên và hồi quy tuyến tính. [2]

Năm 2022, nhóm chuyên gia gồm Salih Imece, Ugur Gurkan Tosun, Ege Berk Buyukbas, Bilge Koroglu, Sefik Ilkin Serengil đã thực hiện một nghiên so sánh về các phương pháp học máy để phân tích các tác nhân gây nên nợ xấu và giải thích. Họ đã nghiên cứu các thuật toán trên dữ liệu được cung cấp bởi ngân hàng tư nhân ở Thổ Nhĩ Kỳ. Nghiên cứu tập trung dữ liệu lịch sử của khách hàng như: lịch sử hành vi thanh toán của khách hàng, bảng cân đối kế toán, các khoản thanh toán bằng thẻ tín dụng trước đây, rủi ro và giới hạn số tiền khách hàng ở các ngân hàng khác từ cơ quan báo cáo người tiêu dùng, lĩnh vực kinh doanh của khách hàng. Cụ thể: (i) thực hiện phân tích để lọc các thông tin không mang giá trị cần thiết để loại trừ khỏi phân tích bằng cách tính toán các giá trị còn thiếu; (ii) sau đó, họ đã nghiên cứu về mối tương quan của các biến thông tin; (iii) thay thế các biến có liên quan còn thiếu với các giá trị trung bình được tính toán dựa trên cột. (iv) cuối cùng, áp dụng bình thường hóa cho rủi ro và giới hạn số tiền được báo cáo bởi cơ quan báo cáo người tiêu dùng. Tóm lại, các giai đoạn của dự án máy học của nhóm chuyên gia này cụ thể là thu thập dữ liệu, tiền xử lý, mô hình hóa, đánh giá và lập luận của các mô hình được đào tạo theo cách dễ hiểu của con người. [3]

Năm 2022, một nhóm chuyên gia gồm Allaymoun và cộng sự của mình đã công bố một nghiên cứu về Big Data và trực quan hóa dữ liệu giúp ngân hàng ra quyết định. Nghiên cứu này xem xét các tài liệu hiện có về các ứng dụng Big Data và trực quan hóa dữ liệu trong ngân hàng bằng cách sử dụng phương pháp phân tích thư mục. Các bài báo được chọn từ 2012 đến 2020 được phân tích dựa trên tỷ lệ trích dẫn, kết quả nghiên cứu và triển khai. Mặc dù các ứng dụng Big Data có tốc độ phát triển nhanh chóng trong lĩnh vực ngân hàng, nhưng số lượng tài liệu nghiên cứu trong lĩnh vực này vẫn còn hạn chế. Đánh giá xác định một số chủ đề bao gồm đầu tư, lợi nhuận, cạnh tranh, phân tích rủi ro tín dụng, khả năng trả nợ, tội phạm ngân hàng và fintech. Báo cáo đã nhấn mạnh tầm quan trọng và việc sử dụng Big Data và ứng dụng trực quan hóa dữ liệu trong lĩnh vực ngân hàng và thảo luận về các cơ hội nghiên cứu trong tương lai về phân tích dữ liệu cho ngành ngân hàng. [4]

Năm 2022, Machado và cộng sự đã công bố bài báo nghiên cứu liên quan đến việc đánh giá rủi ro tín dụng của khách hàng thương mại bằng thuật toán học máy. Nghiên cứu đã chỉ ra rằng với lượng lớn dữ liệu khách hàng có sẵn cho các công ty tài chính, việc sử dụng các phương pháp thống kê truyền thống (ví dụ: hồi quy) để dự đoán điểm tín dụng của khách hàng có thể không mang lại hiệu suất dự đoán tốt nhất. Các thuật toán học máy đã được khám phá trong tài liệu chấm điểm tín dụng để tăng khả năng dự đoán. Trong bài báo này, họ đã dự đoán điểm tín dụng của khách hàng thương mại bằng thuật toán học máy kết hợp kết hợp các phương pháp học máy không giám sát và giám sát. Nghiên cứu cho thấy rằng các mô hình kết hợp vượt trội so với các mô hình riêng lẻ của chúng trong việc dự đoán điểm tín dụng của khách hàng thương mại. Hơn nữa, trong khi tài liệu hiện có bỏ qua điểm tín dụng trong quá khứ, nghiên cứu cho thấy rằng hiệu suất dự đoán của các mô hình kết hợp sẽ cao hơn khi bao gồm các thông tin trong quá khứ này. [5]

# C. Nền tảng lý thuyết (Background)

## 1. Tổng quan về lĩnh vực phân tích tín dụng:

1. *Phân tích tính dụng là gì?*

   Quá trình chọn lọc, nghiên cứu, đánh giá của các chuyên viên tín dụng để hiểu rõ về dữ liệu người vay, cơ cấu và mục đích vay cũng như người thanh khoản cho vay của khách hàng. Quá trình này được định nghĩa là phân tích dữ liệu.

   Phân tích tín dụng có thể được giải thích một cách đơn giản là công việc thu thập, rà sót và xử lý thông tin có khoa học cùng với các kỹ thuật, công cụ phân tích. Mục tiêu là đưa ra những nhận định về mức độ rủi ro hay tin cậy có ý nghĩa kinh tế đối với doanh nghiệp và cá nhân, các giải pháp mới cho dự án kinh doanh, sản xuất từ đó đưa ra các quyết định cấp tín dụng của các tổ chức tín dụng.

   Một hệ thống tính điểm dành riêng được tích hợp cho bên đánh giá được bao gồm với tác dụng là giảm thiểu thiệt hại, nợ xấu và tối đa hóa lợi ích. Kết quả cuối cùng từ việc phân tích sẽ quyết định việc được cấp tín dụng, cho vay hay không đối với đối tượng.

   Đối tượng của phân tích tín dụng là các cá nhân tổ chức có nhu cầu tín dụng bao gồm:

* Khách hàng là doanh nghiệp
* Khách hàng là cá nhân

1. *Mục đích của phân tích tín dụng:*

* ***Khắc phục tình trạng không chính xác của thông tin:*** việc không hạn chế được các thông tin sai lệch có thể dẫn đến những đánh giá không chính xác về đối tượng khách hàng dẫn đến những quyết định không được đảm bảo.
* ***Đánh giá chính xác nhu cầu tín dụng của khách hàng:***Sau khi phân tích tín dụng được diễn ra, các mục đích vay vốn, nhu cầu vay vốn được làm rõ từ đó  là cơ sở để tổ chức tín dụng đưa ra các sản phẩm tín dụng phù hợp với đối tượng khách hàng.
* ***Nhận định được các rủi ro:*** Phân tích tín dụng giúp khách hàng hiểu về khách hàng, đánh giá được lợi ích, khó khăn, ưu điểm, nhược điểm của khách hàng. Và dựa vào đó đánh giá, đo lường mức độ rủi ro của các khoản vay và giúp ích cho quá trình quyết định.

1. *Nội dung của phân tích tín dụng:*

* ***Thu thập và xử lý thông tin:*** Đây là bước đầu của công cuộc phân tích nhưng là một bước quan trọng quyết định đến kết quả của các bước tiếp theo. Công việc này sẽ được thực hiện thông qua các nguồn như: dữ liệu lưu trữ, hồ sơ đề nghị cấp tín dụng của khách hàng, các nguồn bên ngoài. Các thông tin được thu thập sẽ được phân tích, đối chiếu, so sánh, đánh giá với các các nguồn trong quá khứ và hiện tại để đưa ra các thông tin chính xác.
* ***Phân tích khách hàng:*** Tùy vào loại đối tượng khách hàng để thực hiện phân tích:
* *Đánh giá tài sản:* Đối với khách hàng là doanh nghiệp thì đánh giá dựa trên bảng cân đối kế toán. Đối với khách hàng là hộ kinh doanh/cá nhân sẽ đánh giá thông qua tình hình kinh doanh, tài sản cá nhân, các khoản thu nhập chính và phụ.
* *Đánh giá các khoản nợ:* các khoản nợ của khách hàng chắc chắn được quan tâm như nợ ngân hàng khác, nợ cũ, nợ nhà cung cấp, nợ lương nhân viên và các khoản nợ tồn tại khác.
* *Phân tích chi phí, doanh thu:* thường là với khách hàng là doanh nghiệp.
* *Phân tích luồng tiền:* dự tính các dòng tiền để xác định dòng tiền trả nợ.
* *Phân tích tài chính:* đánh giá tình hình khách hàng dựa trên các tỷ lệ về tài chính.
* ***Thẩm định phương án, dự án vay vốn:***đánh giá khả năng, tiềm năng của dự án kinh doanh có mục đích vay vốn, khách hàng có trả được nợ hay không, có nguồn trả nợ thực tế không hay phụ thuộc và dự án, phương án kinh doanh… Đây là công việc quan trọng cho quyết định của tổ chức tín dụng.
* ***Phân tích tài sản đảm bảo:*** Tập trung vào việc đánh giá khả năng thanh lý của các tài sản đảm bảo khi cần. Hướng đến khía cạnh pháp lý của tài sản, khả năng thanh lý tài sản đó, giá thịt trường và khả năng kiểm soát.

## 2. Dữ liệu vay vốn.

Trong dự án này dữ liệu vay vốn đóng vai trò chính, đây là dữ liệu được sử dụng cho quá trình phân tích và đưa ra quyết định cấp tín dụng. Dữ liệu vay vốn bao gồm các thông tin về khách hàng, như thu nhập, lịch sử tín dụng, công việc hiện tại, dữ liệu về tài sản, và các thông tin liên quan cần thiết cho quá trình vay vốn.

Dữ liệu vay vốn được lấy từ các nguồn khác nhau đa phần là từ các nguồn có cơ sở, tin cậy như hồ sơ khách hàng, giao dịch tài chính bên cạnh đó là các nguồn dữ liệu đa dạng khác bên ngoài. Các biến số trong dữ liệu được chọn lọc,xác định để phù hợp cho quá trình phân tích tín dụng, tiếp theo là các hoạt động thu thập các dữ liệu liên quan để phản ảnh các yếu tố được xác định trước đó trong quá trình, phân tích, đánh giá khả năng trả nợ và rủi ro tín dụng.

Dữ liệu vay vốn được thu thập sau đó tổ chức và lưu trữ dưới dạng bảng dữ liệu,  có thể được định dạng CSV, Excel hoặc các cơ sở dữ liệu quan hệ. Đối với dự án đang tiến hành nghiên cứu, bộ dữ liệu với định dạng CSV được sử dụng để làm nguồn dữ liệu thông tin để tiến hành phân tích tín dụng khách hàng.

Sau khi bộ dữ liệu vay vốn được xác định cho dự án, các bước tiếp theo sẽ tiến hành các xử lý trên bộ dữ liệu nhờ vào các thư viện trong Python, dữ liệu được tiền xử lý và khám phá. Các phép biến đổi, chọn lọc dữ liệu, xử lý các sai sót của dữ liệu, xử lý giá trị thiếu, trích xuất đặc trưng từ đó chuẩn bị cho các bước phân tích tiếp theo.

Dữ liệu vay vốn sau khi đã được tiền xử lý sẽ được sử dụng trong việc xây dựng và huấn luyện các mô hình học máy. Các mô hình như Decision Tree, K-Nearest Neighbors, và Logistic Regression có thể được áp dụng để phân loại khách hàng, dự đoán khả năng trả nợ và đưa ra quyết định cấp tín dụng.

Dữ liệu vay vốn đóng vai trò quan trọng của dữ liệu trong dự án "Phân tích tín dụng khách hàng sử dụng dữ liệu vay vốn". Dữ liệu vay vốn được thu thập, tiền xử lý và sử dụng để xây dựng các mô hình học máy để đánh giá khả năng trả nợ và đưa ra quyết định cấp tín dụng.

## 3. Ngôn ngữ lập trình Python

Ngôn ngữ Python là ngôn ngữ chính được sử dụng cho dự án. Python là một ngôn ngữ rất phổ biến với nhiều ứng dụng đa dạng cho nhiều lĩnh vực khác nhau.

* + 1. *Tổng quan về ngôn ngữ Python:*

Python là ngôn ngữ lập trình bậc cao, được ứng dụng trong rất nhiều lĩnh vực khác nhau như xây dựng trang web, phần mềm,phân tích dữ liệu, tự động hóa. Python có thể sử dụng để tạo nhiều chương trình khác nhau, không giới hạn cho một lĩnh vực, vấn đề cụ thể nào. Chính vì sự đa năng cùng với sự thuận tiện, dễ sử dụng của ngôn ngữ này đối với người dùng, Python ngày nay được biết đến như là một trong những ngôn ngữ phổ biến nhất hiện tại.

1. *Ứng dụng của Python:*

* Phân tích dữ liệu và học máy:

Khoa học dữ liệu dùng Python như là một trong những yếu tố chính, ngôn ngữ này hỗ trợ các phép tính toán thống kê phức tạp, trực quan hóa các dữ liệu, xây dựng mô hình học máy, phân tích, khám phá dữ liệu và các công việc khác liên quan đến quá trình nghiên cứu, đánh giá dữ liệu.

Python có thể tạo ra đa dạng các dạng trực quan hóa dữ liệu khác nhau, từ các dạng bảng biểu,biểu đồ cơ bản chuyên dùng trong thống kê, phân tích cho đến các biểu đồ 3D. Python cũng các thư viện thực hiện các chương trình phân tích dữ liệu và học máy nhanh hơn và hiệu quả hơn.

* Một số ứng dụng nổi bậc khác của Python không được ứng dụng trong dự án như là: Phát triển Web, Tự động hóa và phát triển phần mềm.

1. *Ứng dụng của Python đối với dự án:*

Python là một ngôn ngữ lập trình phổ biến và mạnh mẽ được sử dụng rộng rãi trong lĩnh vực phân tích dữ liệu và machine learning. Sử dụng Python trong dự án phân tích tín dụng mang lại nhiều lợi ích:

* ***Thư viện phân tích dữ liệu:*** Python có các thư viện phổ biến như Pandas, NumPy và Scikit-learn cung cấp các công cụ và chức năng mạnh mẽ cho việc phân tích và xử lý dữ liệu. Pandas được sử dụng để đọc dữ liệu từ tệp CSV (như final2.csv), tiền xử lý dữ liệu, và thực hiện các thao tác trên DataFrame như lọc, nhóm, và tính toán thống kê. NumPy cung cấp các hàm số học và thao tác trên mảng nhiều chiều, hữu ích cho việc tính toán số liệu trong phân tích tín dụng. Scikit-learn cung cấp các thuật toán machine learning và công cụ đánh giá hiệu suất mô hình.
* ***Xây dựng và huấn luyện mô hình:*** Python hỗ trợ việc xây dựng và huấn luyện mô hình machine learning thông qua các thư viện như Scikit-learn và TensorFlow. Scikit-learn cung cấp các thuật toán phân loại và hồi quy, bao gồm cả Decision Tree, K-Nearest Neighbors và Logistic Regression, cho việc phân tích tín dụng. Các thuật toán này có thể được áp dụng trên dữ liệu vay vốn để tạo ra mô hình dự đoán khả năng trả nợ của khách hàng. TensorFlow là một thư viện phổ biến cho việc xây dựng và huấn luyện các mô hình deep learning, có thể được sử dụng trong các trường hợp phức tạp hơn.
* ***Trực quang hóa kết quả****:* Python cung cấp các thư viện như Matplotlib và Seaborn để tạo đồ thị và biểu đồ trực quan hóa dữ liệu phân tích. Điều này giúp hiển thị các kết quả phân tích tín dụng một cách trực quan và dễ hiểu, giúp người dùng dễ dàng tìm hiểu và đưa ra quyết định.

Sử dụng Python trong dự án phân tích tín dụng khách hàng giúp tận dụng các thư viện và công cụ mạnh mẽ để thực hiện các công việc phân tích dữ liệu, xây dựng và huấn luyện mô hình, và trực quan hóa kết quả. Python là một ngôn ngữ linh hoạt và dễ sử dụng, phù hợp cho việc phát triển dự án phân tích tín dụng.

## 4. Mô Hình Máy Học

Mô hình học máy là một công cụ phân tích dữ liệu được xây dựng dựa trên các thuật toán và mô hình toán học, cho phép máy tính tự động học từ dữ liệu và đưa ra dự đoán hoặc quyết định dựa trên những gì đã học được. Trong các dự án phân tích tín dụng mô hình máy học thể được dùng để:

* ***Đánh giá rủi ro:*** Mô hình học máy có thể phân loại khách hàng thành các nhóm rủi ro khác nhau dựa trên dữ liệu tín dụng. Bằng cách phân loại khách hàng thành các nhóm rủi ro, ngân hàng hoặc tổ chức tài chính có thể đánh giá rủi ro và áp dụng các biện pháp phòng ngừa phù hợp.
* ***Dự đoán khả năng trả nợ:*** Mô hình học máy có thể được huấn luyện trên dữ liệu lịch sử về khách hàng và kết quả trả nợ để dự đoán khả năng trả nợ của khách hàng mới. Điều này giúp ngân hàng hoặc tổ chức tài chính đưa ra quyết định cấp tín dụng dựa trên khả năng dự đoán của mô hình.
* ***Phát hiện gian lận:*** Mô hình học máy có thể phân loại các hoạt động gian lận trong việc vay vốn, như vay không trả hoặc sử dụng thông tin giả mạo. Bằng cách sử dụng mô hình để phát hiện các hành vi gian lận, ngân hàng hoặc tổ chức tài chính có thể ngăn chặn và cảnh báo kịp thời.
* ***Quản lý rủi ro:*** Mô hình học máy có thể được sử dụng để theo dõi và phân tích tình trạng vay nợ của khách hàng, từ đó đưa ra các biện pháp phòng ngừa và quản lý rủi ro một cách chặt chẽ. Bằng cách theo dõi và đánh giá tình trạng vay nợ, ngân hàng hoặc tổ chức tài chính có thể đưa ra các quyết định liên quan đến tín dụng và quản lý rủi ro một cách hiệu quả.

Trong dự án phân tích tín dụng, Python cung cấp các thư viện và framework mạnh mẽ để xây dựng và huấn luyện các mô hình học máy. Các thư viện phổ biến như Scikit-learn, TensorFlow và Keras cho phép tiến hành các tác vụ như tiền xử lý dữ liệu, xây dựng mô hình, huấn luyện và đánh giá mô hình. Sử dụng Python và các thư viện này, dự án phân tích tín dụng có thể triển khai các mô hình học máy để tăng cường quản lý rủi ro và đưa ra quyết định cấp tín dụng một cách chính xác và hiệu quả.

## 5. Đánh giá hiệu suất mô hình:

Khi xây dựng dự án "Phân tích tín dụng khách hàng sử dụng dữ liệu vay vốn" và sau khi xây dựng mô hình học máy, việc đánh giá hiệu suất và độ chính xác của mô hình là một bước quan trọng để đo lường khả năng dự đoán và độ tin cậy của mô hình đối với việc phân loại dữ liệu tín dụng.

Có một số phép đo thông thường được sử dụng trong đánh giá hiệu suất mô hình, bao gồm:

* **Accuracy (Độ chính xác):** Accuracy là tỷ lệ phần trăm giữa số lượng dự đoán chính xác và tổng số lượng mẫu dữ liệu. Đây là một phép đo đơn giản và dễ hiểu, nhưng nó chỉ phù hợp khi dữ liệu được cân bằng, tức là tỷ lệ giữa các lớp dữ liệu là tương đương.
* **Precision (Độ chính xác dương tính):** Precision đo lường tỷ lệ giữa số lượng dự đoán dương tính chính xác và tổng số lượng dự đoán dương tính. Đây là một phép đo quan trọng để đánh giá khả năng của mô hình trong việc đưa ra các dự đoán chính xác cho nhãn dương tính (ví dụ: khách hàng được phân loại là khách hàng tín dụng tốt).
* **Recall (Độ chính xác nhận diện):** Recall đo lường tỷ lệ giữa số lượng dự đoán dương tính chính xác và tổng số lượng mẫu thực tế dương tính. Đây là một phép đo quan trọng để đánh giá khả năng của mô hình trong việc nhận diện các mẫu thực tế dương tính. Recall có ý nghĩa khi chúng ta quan tâm đến việc không bỏ sót các trường hợp tích cực (ví dụ: không bỏ sót khách hàng tín dụng tốt).
* **F1-score:** F1-score là một phép đo kết hợp giữa Precision và Recall. Nó tính toán trung bình điều hòa của hai giá trị này. F1-score thường được sử dụng khi chúng ta muốn cân bằng giữa Precision và Recall. F1-score cung cấp một phép đo tổng quan về hiệu suất của mô hình, đồng thời cân nhắc cả Precision và Recall.

Ngoài ra một thang đo khác là phép đo kappa score, hay còn được gọi là Cohen's kappa, là một phép đo được sử dụng để đánh giá mức độ đồng nhất (agreement) giữa các người đánh giá trong việc phân loại dữ liệu. Nó được sử dụng phổ biến trong các bài toán phân loại và đánh giá hiệu suất mô hình, đặc biệt khi có sự không cân bằng giữa các lớp dữ liệu.

Kappa score tính toán độ đồng nhất bằng cách so sánh số liệu thực tế và số liệu dự đoán của các người đánh giá. Nó điều chỉnh độ đồng nhất ngẫu nhiên (chance agreement) giữa các người đánh giá, loại bỏ sự ảnh hưởng của sự tình cờ và đưa ra một đánh giá chính xác hơn về mức độ đồng nhất.

Giá trị kappa score thường nằm trong khoảng từ -1 đến 1. Các trường hợp sau có ý nghĩa như sau:

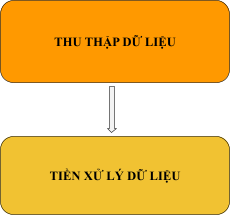
* Kappa score = 1: Hoàn toàn đồng nhất, các người đánh giá hoàn toàn đồng ý.
* Kappa score = 0: Đồng nhất ngẫu nhiên, không có sự đồng ý vượt qua sự tình cờ.
* Kappa score < 0: Đồng nhất kém hơn mức ngẫu nhiên, sự không đồng ý vượt qua sự tình cờ.

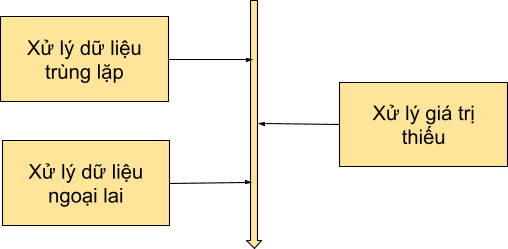
Trong dự án "Phân tích tín dụng khách hàng sử dụng dữ liệu vay vốn", việc sử dụng kappa score có thể giúp đánh giá mức độ đồng nhất giữa các nhân viên đánh giá tín dụng hoặc đánh giá mức độ đồng nhất giữa dữ liệu thực tế và dữ liệu dự đoán từ mô hình học máy. Điều này giúp đảm bảo tính nhất quán và tin cậy trong quá trình phân loại khách hàng và đưa ra quyết định về cấp tín dụng.

Việc đánh giá hiệu suất mô hình thông qua các phép đo như Accuracy, Precision, Recall và F1-score, Kappa score  là một phần quan trọng trong dự án phân tích tín dụng khách hàng. Điều này giúp đảm bảo rằng mô hình có khả năng dự đoán chính xác và đáng tin cậy trong việc phân loại khách hàng và đưa ra quyết định về cấp tín dụng.

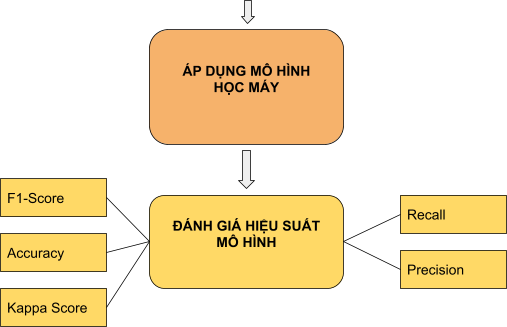
# D. Phương pháp luận nghiên cứu

## 1. Mô hình tổng quan

****

****

****

****

# Hình . Mô hình tổng quan

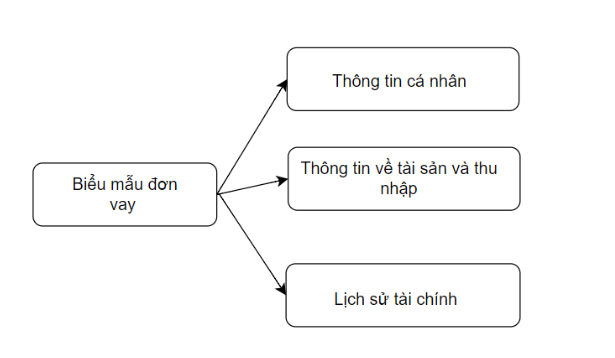
***Diễn giải mô hình tổng quan:***

1. **Thu thập dữ liệu:** Dự án bắt đầu bằng việc thu thập dữ liệu liên quan đến tín dụng khách hàng. Các thông tin này có thể bao gồm lịch sử vay vốn, thu nhập, tuổi tác, hồ sơ tín dụng, số lần trễ hẹn thanh toán, số lần vay không trả và các yếu tố khác có thể ảnh hưởng đến khả năng và rủi ro về tín dụng.
2. **Tiền xử lý dữ liệu:** Tiến hành làm sạch dữ liệu với các bước chính là xử lý dữ liệu bị thiếu, xử lý dữ liệu ngoại lai, xử lý dữ liệu trùng lặp
3. **Trực quan hóa dữ liệu:** Trong giai đoạn này, các phương pháp và công cụ phân tích dữ liệu được áp dụng để hiểu và khám phá các mẫu, xu hướng và thông tin quan trọng trong dữ liệu về tín dụng khách hàng. Các phương pháp phân tích có thể bao gồm phân tích mô tả, phân tích biểu đồ, phân tích đa biến, và các kỹ thuật khác để khám phá mối quan hệ giữa các biến và đặc điểm quan trọng của khách hàng.
4. **Áp dụng mô hình học máy:** Tiếp theo, dự án sử dụng các mô hình học máy để dự đoán và phân loại tín dụng khách hàng. Các mô hình Decision Tree, Logistic Regression, Kneighbors Classifie được áp dụng.
5. **Đánh giá hiệu suất mô hình:** Sau khi xây dựng mô hình, quá trình đánh giá hiệu suất được thực hiện để đo lường khả năng dự đoán và độ chính xác của mô hình. Các phép đo như Accuracy, Precision, Recall, F1-score và Kappa score được sử dụng để đánh giá mức độ chính xác và đáng tin cậy của mô hình trong việc phân loại tín dụng.

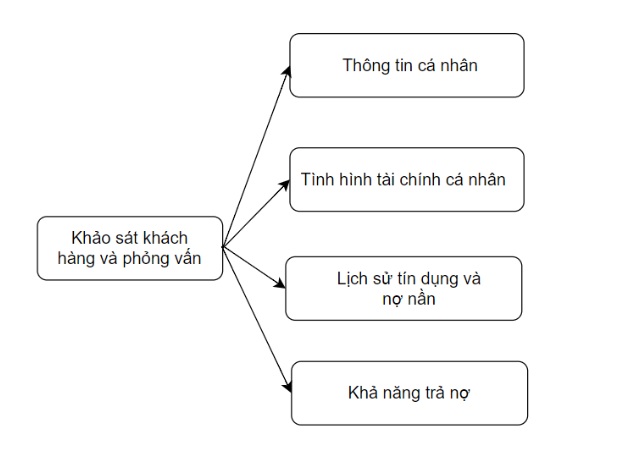
## 2. Thu thập dữ liệu

Việc thu thập dữ dữ liệu có thể từ một hoặc nhiều nguồn, ở đây chúng em thu thập được 69718 dữ liệu ID khách hàng vay vốn trong ngân hàng hay còn gọi là ( mã số định danh cho từng khách hàng).

Với mong muốn dữ liệu thu thập là tốt nhất nên chúng em đã thu thập thông tin khách hàng  qua nội dung của biểu mẫu đơn vay như: thông tin cá nhân ( sau đó được mã hóa dưới dạng ID), thu nhập, tài sản và mục đích vay vốn,....bên cạnh đó khảo sát và phỏng vấn khách hàng với các câu hỏi để có thể thu thập dữ liệu một cách tốt nhất được mô tả ở Hình 1 và Hình 2.

****

**Hình 1. Các yếu tố từ biểu mẫu đơn vay**

****

# Hình . Các yếu tố từ khảo sát và phỏng vấn

## 3. Mô tả dữ liệu

Bộ dữ liệu bao gồm 69718 hàng tương đương 69718 số định danh của  từng khách hàng và có 20 cột tương đương 20 thuộc tính như sau:

 Thuộc tính                                        Kiểu                                 Miêu tả

ID                                                      Phân loại             Mã số định danh cho mỗi khách hàng

Gender                                              Phân loại             Male ( nam), Female ( nữ)

DOB                                                  Phân loại             Ngày tháng năm sinh của khách hàng

Lead\_Creation \_Date                        Phân loại             Ngày tạo thông tin về khách hàng

City\_Code                                         Phân loại             Mã số định danh cho thành phố

City\_Category                                   Phân loại             Phân loại thành phố ( A, B hoặc C)

Employer\_Code                                Phân loại             Mã số định danh cho nhà tuyển dụng

Employer\_Category1                        Phân loại             Phân loại nhà tuyển dụng theo mục 1

Customer\_Existing\_Primary\_           Phân loại             Mã ngân hàng mà khách hàng đang sử dụng

\_Bank\_Code

Primary\_Bank\_Type                         Phân loại             Loại ngân hàng chính mà khách hàng sử dụng

Contacted                                         Phân loại             Trạng thái liên hệ khách hàng

Source                                              Phân loại              Nguồn thông tin về khách hàng

Source\_Category                             Phân loại              Phân loại nguồn thông tin về khách hàng

Existing\_EMI                                    Số                        Số tiền trả trước hàng tháng ( nếu có)

Loan\_Amount                                  Số                        Số tiền vay

Employer\_Category2                       Số                        Phân loại nhà tuyển dụng theo mục 2

Monthly\_Income                              Số                        Thu nhập hàng tháng của khách hàng

Loan\_Period                                    Số                        Thời hạn vay ( số tháng )

Interest\_Rate                                  Số                         Lãi suất vay

EMI                                                 Số                         Khoản trả hàng tháng

Var1                                                Số                         Biến số 1

Approved                                        Phân loại              Trạng thái phê duyệt

Chi tiết hơn ở tập dữ liệu, chúng ta có thể thống kê sơ bộ về dữ liệu. Dưới đây là một số thông tin sơ bộ về bộ dữ liệu

* + Số lượng mẫu ( dòng ) : 69718 dòng
  + Số lượng thuộc tính ( cột ) : 22 cột

**1. ID**

- Số lượng: 69718 dòng

- Mỗi ID có một giá trị đại diện cho một khách hàng

- Đây là trường dữ liệu để xác định khách hàng

**2. Gender**

* Số lượng nam giới là: 39953 ( người ). Tỷ lệ nam giới chiếm khoảng 57.31% trong tổng số khách hàng.
* Số lượng nữ giới là 29765 ( người). Tỷ lệ nữ giới chiếm khoảng 42.69% trong tổng số khách hàng.
* Vì vậy ta có thể thấy rằng số lượng nam giới trong bộ dữ liệu cao hơn nữ giới. Từ sự chênh lệch này có thể gây ra ảnh hưởng đến phân tích và đưa ra quyết định, vì nếu dữ liệu không cân bằng về mặt giới tính tính, các mô hình kết quả có thể bị ảnh hưởng theo chiều của nhóm dữ liệu lớn hơn.
* Để đảm bảo công bằng và đúng đắn trong phân tích dữ liệu, việc xử lý dữ liệu không cân bằng giới tính có thể được áp dụng, ví dụ như sử dụng kỹ thuật undersampling hoặc oversampling để tạo ra một bộ dữ liệu cân bằng hơn giữa nam giới và nữ giới.

1. **DOB**

* Có tổng cộng 6381 giá trị duy nhất trong cột DOB. Điều này cho thấy có sự đa dạng trong ngày sinh.
* Giá trị  “31/8/1987” xuất hiện nhiều nhất với 49 lần xuất hiện. Bên cạnh đó các giá trị khác như “31/08/1990”, “31/08/1988” và “15/07/1991” cũng xuất hiện khá nhiều lần.
* Giá trị trung bình là “ 15/17/1989” cho thấy có sự trung lập về phân bố ngày sinh trong dữ liệu.
* Độ lệch chuẩn là 5645 ngày, cho thấy sự đa dạng và biến động lớn về ngày sinh.

**4. Lead\_Creation\_Date**

* Tổng số lượng dòng là 43570. cho thấy bảng dữ liệu này chứa thông tin về nhiều khách hàng khác nhau
* Có 56 giá trị suy nhất trong cột “Lead\_Creation\_Date” cho thấy sự đa dạng về ngày tạo hồ sơ vay vốn
* Ngày “22/09/2016” xuất hiện nhiều nhất trong cột, với tần suất là 1629 lần. Điều này cho thấy ngày này có một sự tập trung đáng kể trong quá trình tạo hồ sơ vay vốn
* Giá trị đầu tiên trong cột là “13/07/2016”, còn giá trị cuối cùng là “30/09/2016”. Điều này cho thấy dữ liệu trong cột được thu thập trong khoảng thời gian từ tháng 7 đến tháng 9 năm 2016

1. **City\_Code**

* Cột City\_Code chứa thông tin về mã thành phố của khách hàng
* Có 678 giá trị duy nhất trong cột, cho thấy có sự đa dạng về mã thành phố hoặc vùng
* Mã thành phố/vùng mang giá trị “C10001” xuất hiện nhiều nhất với tần suất là 10009 lần
* Không có thông tin về các giá trị thiếu (null) trong cột “City\_Code”

Thông tin này có thể hữu ích để xem xét sự phân bố và phân loại các khách hàng theo mã thành phố trong trong việc đưa ra quyết định liên quan đến vấn đề vay vốn của khách hàng

1. **City\_Category**

* Số lượng dữ liệu có 68904 mẫu
* Số lượng giá trị duy nhất có 3 giá trị duy nhất trong cột “City\_Category”
* Giá trị xuất hiện nhiều nhất là : Giá trị “A” xuất hiện nhiều nhất trong cột, với tần suất là 49890 lần
* Tần suất của giá trị “A” xuất hiện: 49890 lần trong dữ liệu
* Dựa vào kết quả này, ta có thể kết luận rằng “A” là giá trị phổ biến nhất trong cột “City\_Category” với hơn 72% số lượng mẫu

1. **Employer\_Code**

* Tổng cộng có 65700 dòng dữ liệu trong cột “Employer\_Code” tức là có 657000 khách hàng khác nhau trong bộ dữ liệu này
* Có 36617 giá trị duy nhất trong cột tức là có 36617 nhà tuyển dụng khác nhau mà khách hàng của chúng ta làm việc
* Giá trị “COM0000002” xuất hiện nhiều nhất trong cột này với tần suất là 457 lần. Điều này cho thấy nhà tuyển dụng đại diện bởi mà “COM0000002” có sự phổ biến cao trong bộ dữ liệu và có khả năng là một những tuyển dụng lớn hoặc phổ biến trong ngành

Từ kết quả trên, ta có thể nhận thấy sự đa dạng về nhà tuyển dụng trong bộ dữ liệu này, nhưng cũng có một số nhà tuyển dụng chiếm ưu thế và xuất hiện nhiều hơn các nhà tuyển dụng khác. Thông tin này có thể giúp ta hiểu hơn về phân phối và tầm quan trọng của các nhà tuyển dụng trong quá trình đánh giá khả năng vay vốn của khách hàng

1. **Emloyer\_Category1**

* Tổng cộng có 65700 quan sát trong cột “Employer\_Category1”
* Có 3 giá trị duy nhất trong cột gồm “A”, “B”, “C”.
* Giá trị “A” xuất hiện nhiều nhất, với tần suất là 33338. Điều này cho thấy giá trị cho thấy giá trị “A” chiếm đa số trong dữ liệu và là giá trị phổ biến nhất
* Giá trị “B” có tần suất 16458, tức là xuất hiện khoảng 25% trong dữ liệu
* Giá trị “C” có tần suất là 15904, tức là xuất hiện khoảng 24% trong dữ liệu

Từ các kết quả trên, ta có thể suy luận rằng trong cột “Employer\_Category1”, giá trị “A” chiếm phần lớn và phổ biến nhất trong khi giá trị “B” và giá trị “C” xuất hiện với tần suất tương đối gần nhau

Điều này có thể cho thấy mức độ phân bố và đa dạng trong các nhóm của cột “Employer\_Category1”

1. **Customer\_Existing\_Primary\_Bank\_Code**

* Có tổng cộng 60327 dòng dữ liệu trong cột này tức là có 60327 khách hàng được ghi lại thông tin về ngân hàng chính hiện có của họ
* Có 57 giá trị duy nhất trong cột, cho thấy có 57 ngân hàng khác nhau mà khách hàng đang sở hữu
* Giá trị phổ biến trong cột là “B001”, xuất hiện 14200 lần. Điều này cho thấy có một ngân hàng phổ biến mà nhiều khách hàng đang sử dụng làm ngân hàng chính của họ
* Các giá trị còn lại xuất hiện với tần suất thấp hơn, cho ta thấy sự đa dạng về ngân hàng mà khách hàng đang sở hữu. Mỗi giá trị đại diện cho một ngân hàng khác nhau

Thông qua thống kê này, ta có cái nhìn tổng quan về sự phân bố và tần suất xuất hiện của các ngân hàng khác nhau mà khách hàng đang sở hữu. Điều này có thể cung cấp thông tin quan trọng về quan hệ tài chính của khách hàng với các ngân hàng, sự đa dạng trong lựa chọn ngân hàng chính, và có thể tạo cơ sở để phân tích sâu hơn về khách hàng và hành vi tài chính của họ

1. **Primary\_Bank\_Type**

* Trong bộ dữ liệu có 60327 mẫu. Có hai giá trị P và G duy nhất trong cột này
* Giá trị “P” là giá trị xuất hiện nhiều nhất với tần suất là 39622 lần. Điều này cho thấy hầu hết các khách hàng có loại ngân hàng chính được đánh dấu là “P”
* Giá trị “G” ít xuất hiện hơn, với tần suất xuất hiện là 20705 lần.

1. **Contacted**

* Số lượng mẫu có tổng cộng 69718 mẫu dữ liệu trong cột “Contacted”. Điều này cho thấy tổng số khách hàng đã có thông tin liên hệ là 69718
* Cột chỉ có hai giá trị “Y” và “N”. Điều này cho thấy chỉ có hai trạng thái liên hệ được ghi nhận trong tập dữ liệu
* Giá trị xuất hiện nhiều nhất là “Y” ( đã liên hệ) là giá trị xuất hiện nhiều nhất trong cột “Contacted”. Điều này cho thấy tỷ lệ khách hàng là khá cao
* Số lần xuất hiện của giá trị “Y” xuất hiện 45279 lần trong cột “Contacted”. Điều này cho thấy hơn 45279 khách hàng đã được liên hệ

Với thông tin này ta có thể thấy rằng, trong tập dữ liệu tỷ lệ khách hàng đã được liên hệ là khá cao, chiếm khoảng 65% tổng số khách hàng. Điều này có thể gợi ý rằng chiến dịch liên hệ khách hàng đã diễn ra một cách hiệu quả và đạt được mục tiêu liên hệ đến hầu hết khách hàng trong tập dữ liệu.

1. **Source**

* Cột “Source trong bảng dữ liệu có 29 giá trị duy nhất với tổng số quan sát là 69718. Giá trị “S122” xuất hiện nhiều nhất với tần suất là 30942 chiếm khoản 44.39% tổng số quan sát
* Các giá trị khác cũng có mức độ xuất hiện khá đa dạng với “S133” chiếm 14.01%, “S143” chiếm 6.45%, “S127” chiếm 5.29% và “S159” chiếm 4.48%. Các giá trị còn lại có tỷ lệ xuất hiện thấp hơn 4% cho mỗi giá trị
* Kết quả này cho rằng giá trị “S122” là nguồn dữ liệu phổ biến nhất trong cột “Source”, có sự ưu tiên cao hơn so với các giá trị khác. Điều này có thể gợi ý rằng nguồn dữ liệu “S122” có thể có mức độ quan trọng cao hơn trong phân tích dữ liệu hoặc các quyết định kinh doanh liên quan

1. **Source\_Category**

* Cột “Source\_Category có 7 giá trị khác nhau, được đại diện bằng các chữ cái A, B,C,D,E, F và G. Giá trị “B” xuất hiện nhiều nhất trong cột với tần suất là 29816.
* Thông qua cột trên, ta có thể suy ra rằng các thông tin liên quan đến nguồn tài nguyên mà khách hàng đã sử dụng để dễ tiếp cận dịch vụ hoặc sản phẩm. Mỗi giá trị của “Source\_Category”  có thể đại diện cho một hình thức tiếp cận khác nhau, chẳng hạn như quảng cáo trực tuyến (A), quảng cáo trên kênh truyền hình (B), tiếp thị trực tiếp ( C ), đối tác liên kết (D) và nhiều hình thức khác.
* Thông qua việc thống kê mô tả, chúng ta có cái nhìn tổng quan về sự phân bố của các giá trị trong cột “Source\_Category”. Điều này có thể giúp chúng ta hiểu rõ hơn về nguồn tài nguyên được sử dụng phổ biến nhất và tối ưu hóa các hoạt động tiếp thị hoặc quảng cáo để tăng cường khả năng tiếp cận và tương tác với khách hàng.

1. **Existing\_EMI**

Cột “Existing\_EMI” đại diện cho số tiền trả góp hàng tháng hiện có của khách hàng trước khi xin vay. Dựa trên kết quả thống kê mô tả, ta có thể diễn giải như sau:

* Số lượng mẫu: 69667 mẫu trong cột “Existing\_EMI”
* Giá trị trung bình khoảng 360.95. Điều này cho biết trung bình mỗi khách hàng có số tiền trả góp hàng tháng khoảng 360.95
* Độ lệch chuẩn là khoảng 2288.45. Điều này cho biết có sự biến động lớn trong số tiền trả góp hàng tháng của khách hàng, với một phạm vị rộng từ 0 đến 545436.5
* Giá trị nhỏ nhất là 0. Điều này cho biết có một số khách hàng không có số tiền trả góp hàng tháng
* Phân vị 25%: 25% các giá trị trong cột có giá trị nhỏ hơn hoặc bằng 0
* Phân vị 50% các giá trị có giá trị nhỏ hơn hoặc bằng 0. Điều này cho biết có một số lượng lớn khách hàng không có số tiền trả góp hàng tháng
* Phân vị 75% các giá trị trong cột có giá trị nhỏ hơn hoặc bằng 350. Điều này cho thấy có một số khách hàng có số tiền trả góp hàng tháng nhỏ hơn hoặc bằng 350
* Giá trị lớn nhất: 545436.5. Điều này cho biết có một số khách hàng có số tiền trả góp hàng tháng đáng kể

Tổng quan, trong cột “Existing\_EMI” có sự biến động lớn với phần lớn khách hàng, tạo nên sự chênh lệch và độ lệch chuẩn cao trong dữ liệu.

1. **Loan\_Amount**

* Số lượng mẫu: 42008. Điều này cho biết có 42008 mẫu dữ liệu trong cột “Loan\_Amount”
* Giá trị trung bình: 39429.299181 đây là giá trị trung bình của các khoản vay trong cột. Cho biết giá trị trung bình của các khoản vay là 39429,3
* Độ lệch chuẩn: 30726.406835 dùng để đo lường mức độ biến động của các khoản vay trong cột.
* Giá trị nhỏ nhất: 5000. Đây là giá trị nhỏ nhất trong cột “Loan\_Amount”, cho thấy khoản vay nhỏ nhất là 5000
* Phân vị 25% cho thấy rằng 255 số lượng mẫu trong cột có giá trị nhỏ hơn hoặc bằng 20000
* Phân vị 50% là giá trị giữa, chia dữ liệu thành hai thành phần bằng nhau. Trong trường hợp này, 50% số lượng mẫu có giá trị trong cột nhỏ hơn hoặc bằng 3000
* Phân vị 75% cho thấy rằng 75% số lượng mẫu trong cột có giá trị nhỏ hơn hoặc bằng 50000
* Giá trị lớn nhất: 300000 cho thấy khoản vay lớn nhất là 300000

1. **Employer\_Category2**

* Cột “Employer\_Category2” có 65420 dòng không bị thiếu dữ liệu, tức là có 65420 giá trị khác nhau
* Giá trị trung bình của cột là 3.720208, cho biết giá trị trung bình gần với giá trị Quartile 3, tức là phần lớn giá trị trong cột tập trung vào giá trị 4
* Độ lệch chuẩn là 0.807347, cho thấy sự phân tán của các giá trị xung quanh giá trị trung bình không quá lớn
* Giá trị nhỏ nhất là 1 và giá trị lớn nhất là 4 tức là các giá trị trong cột nằm trong khoảng 1 đến 4

1. **Monthly\_Income**

* Số lượng mẫu: 69718 giá trị hợp lệ trong cột
* Giá trị trung bình: 5622.13 đây là giá trị trung bình của thu nhập hàng tháng của khách hàng
* Độ lệch chuẩn: 174760.8 biểu thị mức độ biến động của dữ liệu. Độ lệch chuẩn lớn cho thấy các giá trị trong cột có sự phân tán rộng
* Giá trị nhỏ nhất: 0 đại diện cho giá trị thu  nhập hàng tháng nhỏ nhất trong dữ liệu
* Phân vị 25%: 1650 , chỉ có 25% giá trị thu nhập hàng tháng nhỏ hơn hoặc bằng 1650
* Trung vị 50%: 2500, đây là giá trị nằm ở giữa danh sách sau khi được sắp xếp theo thứ tự tăng dần. Nửa số lượng giá trị thu nhập hàng tháng nhỏ hơn hoặc bằng 2500 và nửa còn lại lớn hơn hoặc bằng 2500
* Phân vị 75% : 4000, chỉ có 25% giá trị thu nhập hàng tháng lớn hơn 4000
* Giá trị tối đa: 38383840, đại diện cho giá trị thu nhập hàng tháng lớn nhất trong dữ liệu

1. **Loan\_Period**

* Số lượng mẫu: có tổng cộng 42008 mẫu dữ liệu được ghi nhận trong cột
* Giá trị trung bình của thời gian vay vốn là 3.89 tháng. Cho biết sự trung bình hóa thời gian vay của các khoản vay trong dữ liệu
* Độ lệch chuẩn là: 1.17 thể hiện mức độ phân tán của thời gian vay vốn trong dữ liệu. Độ lệch chuẩn cao cho thấy sự biến động lớn giữa các khoản vay về thời gian vay
* Giá trị nhỏ nhất: thời gian vay vốn nhỏ nhất là 1 tháng, cho thấy có các khoản vay ngắn hạn trong vốn dữ liệu
* Phân vị 25% : chỉ ra rằng 25% các khoản vay có thời gian vay ít hơn hoặc bằng 3 tháng

Phân vị 50% cho thấy 50% các khoản vay có thời gian vay ít hơn hoặc bằng 4 tháng

* Phân vị 75% chỉ ra rằng 75% các khoản vay có thời gian vay ít hơn hoặc bằng 5 tháng
* Giá trị lớn nhất: thời gian vay vốn lớn nhất là 6 tháng, cho thấy có các khoản vay dài hạn trong dữ liệu

Từ thông tin trên, ta có thể thấy rằng trong tập dữ liệu này, thời gian vay vốn chủ yếu tập trung trong khoảng từ 3 đến 5 tháng, với trung bình là 3.89 tháng. Tuy nhiên cũng có một phần nhỏ các khoản vay ngắn ( dưới 3 tháng) và dài hạn ( từ 5 đến 6 tháng).

1. **Interest\_Rate**

* Số lượng 22279 khoản vay có thông tin về lãi suất
* Giá trị trung bình: 19.214192 cho biết giá trị trung bình của lãi suất cho các khoản vay
* Độ lệch chuẩn: 5.84 cho thấy mức độ biến động của lãi suất trong tập dữ liệu. Một độ lệch chuẩn lớn hơn thể hiện mức độ biến động cao hơn của dữ liệu
* Giá trị tối thiểu: 11.99 đại diện cho lãi suất thấp trong tập dữ liệu
* Phân vị 25% : 15.25 cho thấy 25% các khoản vay có lãi suất thấp hơn giá trị này
* Giá trị trung vị : 18.00 đây là giá trị trung vị của lãi suất. nửa trên các khoản vay có lãi suất cao hơn và nửa dưới có lãi suất thấp hơn giá trị này
* Giá trị 75% : 20.00 chỉ có 25% các khoản vay có lãi suất cao hơn giá trị này
* Giá trị tối đa: 37.00 đại diện cho lãi suất cao nhất trong tập dữ liệu

Kết quả này cho thấy rằng phân bố lãi suất trong tập dữ liệu có độ biến động từ 11.99% đến 37.00%. Giá trị trung bình của lãi suất là khoảng 19.21% và phần lớn các khoản vay có lãi suất nằm trong khoảng từ 15.25% đến 20.00%.

1. **EMI**

* Số lượng: 22279 đây là số lượng mẫu có giá trị không bị thiếu trong cột EMI
* Giá trị trung bình 1101.42 đây là giá trị trung bình của số tiền trả góp hàng tháng, nó biểu thị mức trung bình của các khoản vay
* Độ lệch chuẩn: 752.62 đây là một đánh giá về sự biến động của dữ liệu. Giá trị này cho thấy mức độ dao động của số tiền trả góp hàng tháng xung quanh giá trị trung bình
* Giá trị nhỏ nhất: 118 thể hiện số tiền trả góp hàng tháng nhỏ nhất
* Giá trị lớn nhất: 13556 thể hiện số tiền trả góp hàng tháng lớn nhất trong tập dữ liệu

Từ các thông số trên, chúng ta có thể nhận thấy rằng cột “EMI” có sự biến động đáng kể trong dữ liệu với mức trung bình là 1101.42 và độ lệch chuẩn là 752.62. Khoảng giá trị từ 25% đến 75% tập trung chủ yếu trong khoảng từ 649 đến 1295 nhưng giá trị EMI có thể dao động từ 118 đến 13556.

1. **Var1**

* Số lượng mẫu có tổng cộng 69718 mẫu dữ liệu trong cột Var1.
* Giá trị trung bình là 3.948583. Điều này cho thấy giá trị trung bình của các biến số trong phân loại cột.
* Độ lệch chuẩn: là một đánh giá về mức độ phân tán trong cột, trong trường hợp này độ lệch chuẩn của cột “Var1” khoảng 3.819217, cho thấy sự biến đổi lớn của các giá trị.
* Giá trị nhỏ nhất là 0 và giá trị lớn nhất là 10 điều này cho thấy cột “Var1” chứa các giá trị từ 0 đến 10.

1. **Approved**

* Số lượng mẫu trong cột là 60718.
* Tỷ lệ phần trăm không được chấp nhận là 98.53%.
* Tỷ lệ phần trăm được chấp nhận là 1.46%.

## 4. Tiền xử lý dữ liệu và trích xuất đặc trưng

**4.1.**  **Tiền xử lý dữ liệu**

****

# Hình . Sơ đồ các bước tiền xử lý dữ liệu

Sau khi đã thu thập và mô tả dữ liệu, chúng ta tiến hành quá trình tiền xử lý dữ liệu. Vì được lấy nguồn từ thực tế, bộ dữ liệu này có một số dữ liệu không hoàn chỉnh, không nhất quán và dữ liệu nhiễu. Chính vì vậy, bước tiền xử lý dữ liệu là vô cùng cần thiết để mang đến những thông tin hữu ích cho bộ dữ liệu. Tiền xử lý dữ liệu bao gồm các kỹ thuật biến đổi dữ liệu từ nguyên bản thành một phiên bản dễ hiểu hơn và phù hợp hơn với mục đích sử dụng. Hiện nay, có rất nhiều phương pháp tiền xử lý dữ liệu, một số phương pháp tiền xử lý dữ liệu mà nhóm chúng tôi sử dụng cho bộ dữ liệu này là: xử lý dòng trùng lặp, xử lý dữ liệu bị thiếu và xử lý dữ liệu ngoại lai (outlier).

*Làm sạch dữ liệu*: quá trình này bao gồm việc điền vào các giá trị thiếu và xử lý dòng trùng lặp. Chúng tôi hoàn toàn loại bỏ dòng trùng lặp, thay thế giá trị bị thiếu bằng dữ liệu xuất hiện nhiều nhất đối với đặc trưng hạng mục, và giá trị xấp xỉ trung bình đối với đặc trưng dạng số. (1)

*Xử lý dữ liệu ngoại lai:* đôi khi, tập dữ liệu sẽ chứa các giá trị lớn không khớp với mẫu tổng thể, chúng không thể đưa vào bất kỳ danh mục hoặc nhóm nào vì sẽ làm sai lệch bộ dữ liệu. Loại dữ liệu này được gọi là các giá trị ngoại lai. Nhóm chúng tôi sử dụng Z-score để giải quyết vấn đề này. Z-score được sử dụng phổ biến trong việc phát hiện các ngoại lai bằng các đo lường độ chênh lệch và định rõ xem một giá trị có xu hướng rời xa so với phân phối dữ liệu chung hay không. (2)

*Biểu diễn đặc trưng dưới dạng vector:* để có thể tích hợp với các thuật toán máy học, giảm kích thước dữ liệu và tính toán hiệu quả hơn, phương pháp này là đóng vai trò vô cùng quan trọng. Nhóm chúng tôi sử dụng biến giả (dummy variable) để đưa các giá trị hạng mục thành dạng 0 và 1. (3)

**4.2. Trích xuất đặc trưng**

Trích xuất đặc trưng trong phân tích dữ liệu là quá trình chuyển đổi dữ liệu từ dạng ban đầu thành các đặc trưng có ý nghĩa và hữu ích để phân tích. Mục đích của việc trích xuất đặc trưng là tìm ra những thuộc tính quan trọng và đáng chú ý trong dữ liệu để sử dụng cho mô hình hoặc phân tích sau này.

*Chọn các đặc trưng cần thiết cho máy học:* lựa chọn đặc trưng là quá trình giảm số lượng biến đầu vào khi xây dựng một mô hình dự đoán. Việc giảm số lượng đặc trưng đầu vào nhằm giảm chi phí tính toán khi mô hình hóa và, trong một số trường hợp, cải thiện hiệu suất của mô hình. Ở đây, nhóm chúng tôi đã chọn 26 danh mục cần thiết cho quá trình máy học. (4)

**4.3. Chuẩn hóa dữ liệu và xử lý tình trạng mất cân bằng lớp**

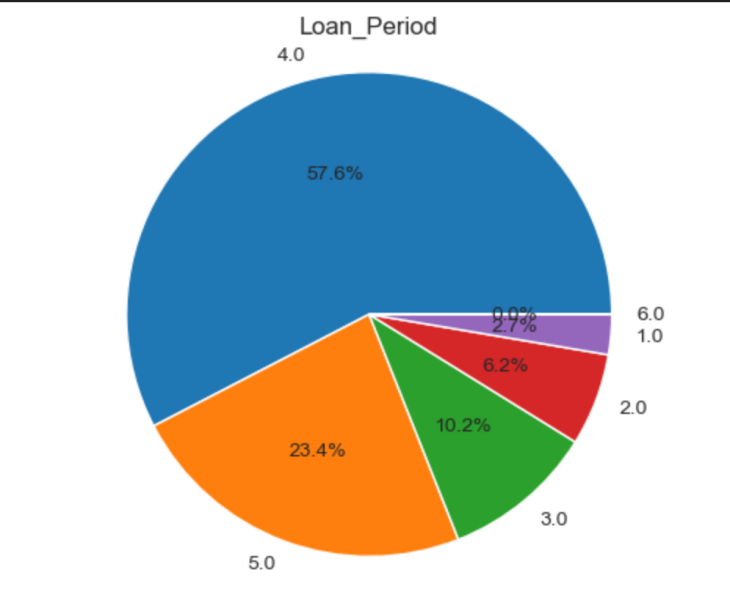
*Chuẩn hóa (Normalization):* là một kỹ thuật được sử dụng để đảm bảo rằng tất cả các dữ liệu trong cơ sở dữ liệu có một phạm vi tương tự.(5) Ở đây, nhóm chúng tôi sử dụng MinMaxScaler để chuẩn hóa dữ liệu nhằm đảm bảo rằng các đặc trưng trong các tập dữ liệu có cùng phạm vi hoặc phân phối chung.

*Tình trạng mất cân bằng lớp (class imbalance):* đây là vấn đề thường gặp trong dữ liệu. Nhóm chúng tôi sử dụng phương pháp SMOTE (Synthetic Minority over-sampling TEchnique) tạo ra dữ liệu bổ sung từ lớp thiểu số, nhằm khắc phục tình trạng thiếu dữ liệu. Phương pháp này dựa trên việc tạo ra các ví dụ trên các đường nối giữa một điểm và K điểm láng giềng gần nhất của nó. (6)

# E. Kết quả thử nghiệm và phân tích

## 1.1. Trực quan hóa dữ liệu dưới dạng biểu đồ và phân tích

**1.1.1. Phân tích đơn biến**

****

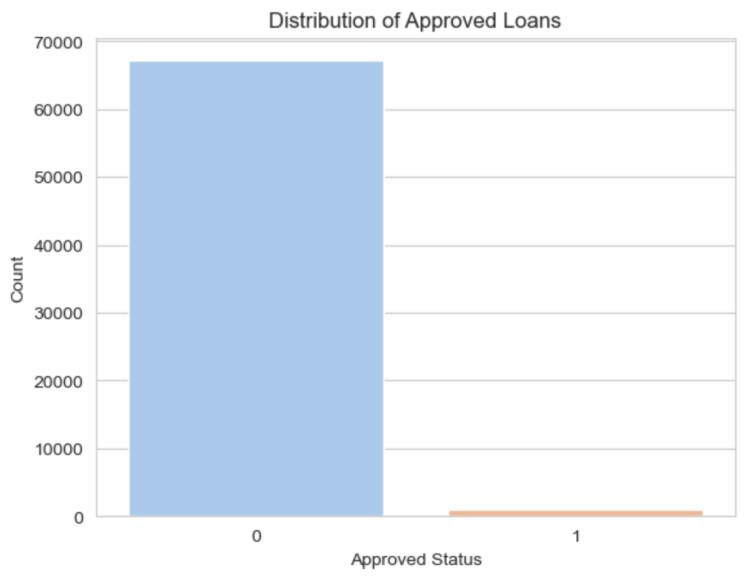
# Hình . Biểu đồ tròn cho Loan\_Period

57% tổng số khách hàng của ngân hàng có xu hướng trả khoản vay trong vòng 4 năm và chỉ có một khách hàng vay trong vòng 6 năm



# Hình . Biểu đồ Scatter plot cho 'Loan\_Period' và 'Interest\_Rate'

Biểu đồ scatter plot này sẽ cho ta cái nhìn tổng quan về mối quan hệ giữa 'Loan\_Period' và 'Interest\_Rate', và cũng thể hiện sự phân bố của giá trị 'Loan\_Amount' thông qua kích thước của điểm trên biểu đồ.



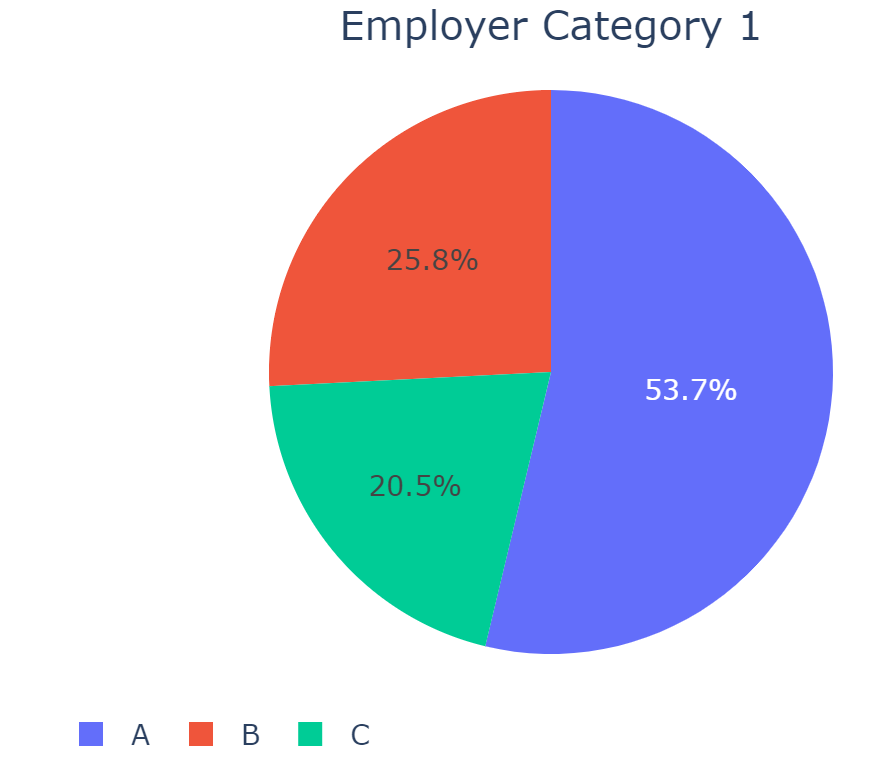
# Hình . Biểu đồ countplot cho biến 'Approved'

Đoạn mã trên sử dụng thư viện Seaborn để vẽ biểu đồ countplot, mô tả phân phối của biến 'Approved' trong dataframe với 0 là số người bị từ chối vay tín dụng, 1 là số người được đồng ý cho vay tín dụng.



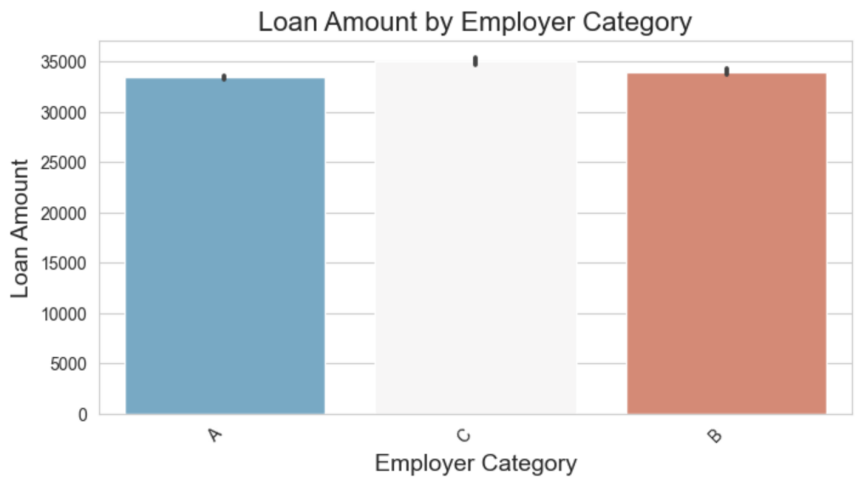
# Hình . Biểu đồ barplot giữa 'Approved' và 'Monthly\_Income'

Biểu đồ barplot bên thể hiện sự khác biệt về thu nhập hàng tháng trung bình giữa các trường hợp được chấp nhận và bị từ chối. Chúng ta phát hiện ra rằng mức độ phê duyệt của khoản vay (Approved) cao hơn đối với những khách hàng có thu nhập hàng tháng (Monthly\_Income) cao hơn, điều này có thể cho thấy mức độ quan trọng của thu nhập hàng tháng đối với việc được phê duyệt khoản vay.



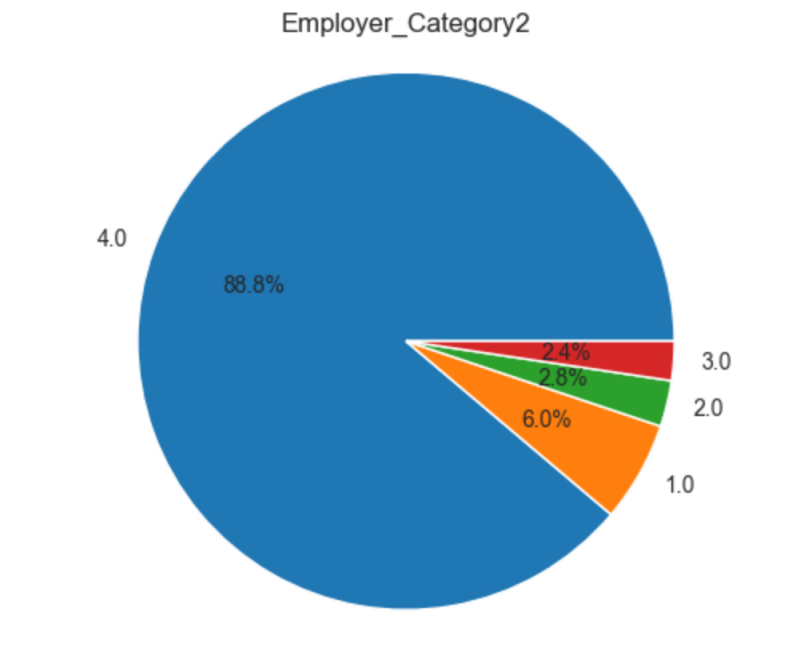
# Hình . Biểu đồ tròn cho 'Employer\_Category\_1'

Ta thấy: Category1 (A) là phổ biến nhất. Category1 (C) là ít phổ biến nhất.



# Hình . Biểu đồ Barplot giữa 'Loan\_Amount' và 'Employer\_Category'

Đoạn mã trên sử dụng thư viện Seaborn để vẽ biểu đồ barplot, mô tả mức độ vay (Loan Amount) theo danh mục nhà tuyển dụng (Employer Category) trong dataframe. Kết quả: Loại C là loại khách hàng có khoản vay lớn nhất, trong khi cả loại A và B đều giống nhau



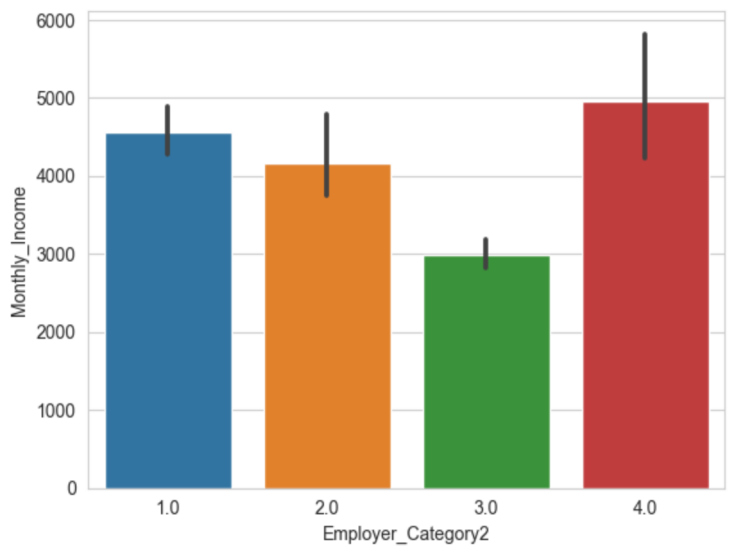
# Hình . Biểu đồ tròn cho 'Employer\_Category\_2'

Ta thấy: Danh mục khách hàng (4.0) là phổ biến nhất. Danh mục khách hàng (3.0) là ít phổ biến nhất.



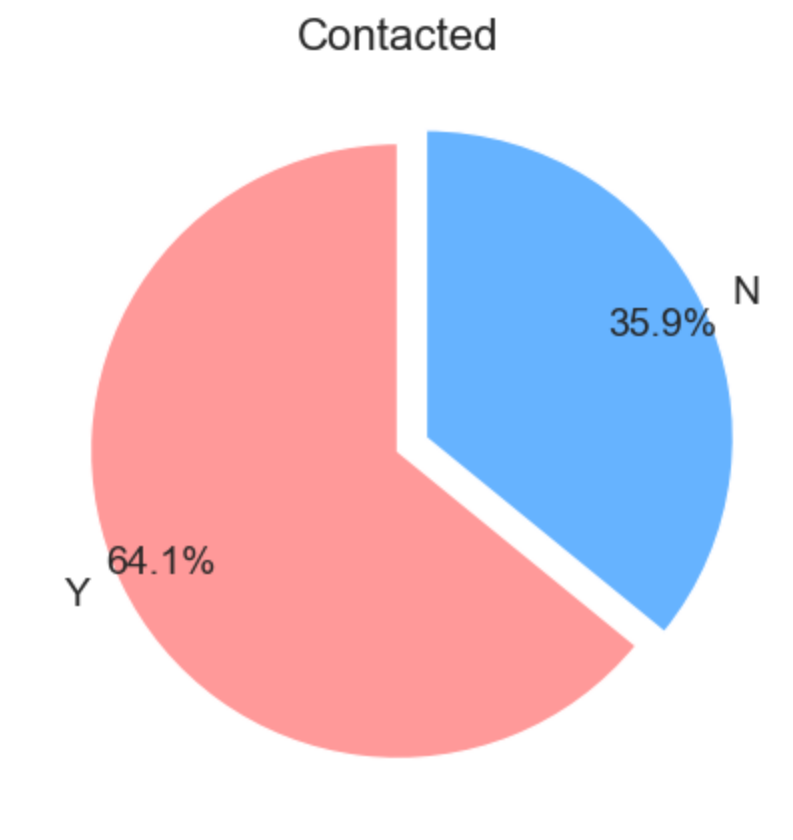
# Hình . Biểu đồ barplot giữa 'Loan\_Amount' và 'Employer\_Category\_2'

Mã trên sử dụng thư viện Seaborn để vẽ biểu đồ cột dọc (barplot) mô tả mức độ vay trung bình (Loan Amount) theo danh mục nhà tuyển dụng (Employer Category 2) trong dataframe. Kết quả: Danh mục (1.0) là danh mục khách hàng có khoản vay lớn nhất, trong khi danh mục (3.0) có khoản vay nhỏ nhất.



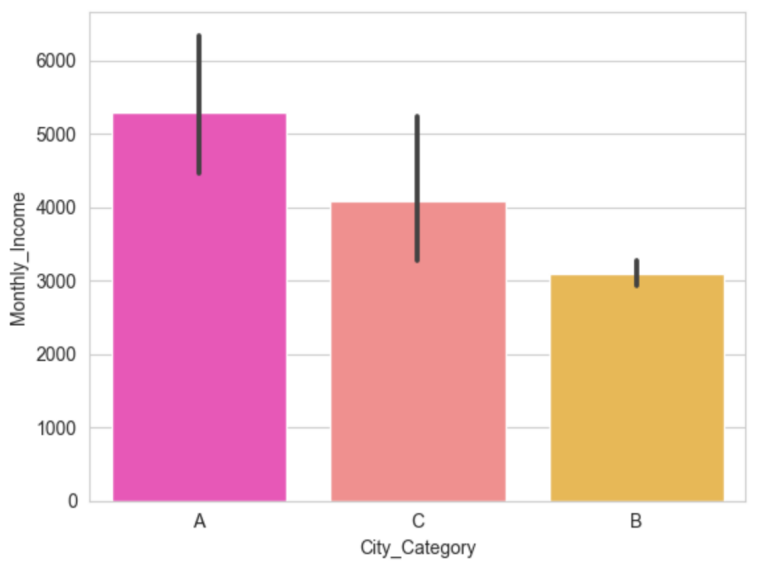
# Hình . Biểu đồ giữa 'Monthly\_Income' và 'Employer\_Category\_2'

Đoạn mã trên tính giá trị trung bình của thu nhập hàng tháng (Monthly\_Income) theo danh mục nhà tuyển dụng (Employer\_Category2) trong DataFrame 'df' bằng cách sử dụng phương thức groupby và mean.Kết quả: Nhóm khách hàng thứ 3 có thu nhập bình quân tháng thấp nhất và giá trị khoản vay bình quân thấp nhất



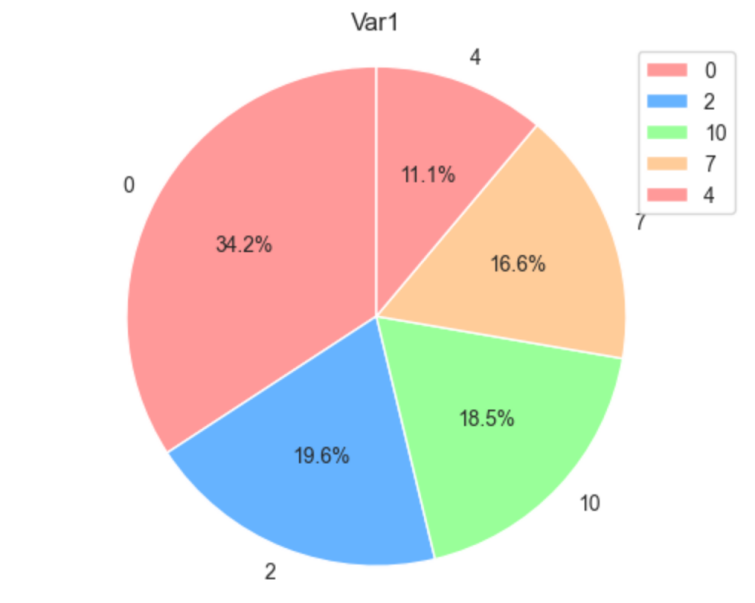
# Hình . Biểu đồ tròn cho 'Contacted'

Kết quả: có khoảng 64% khách hàng đã được liên hệ cùng với đó là hơn 35% khách hàng chưa được liên hệ.



# Hình . Biểu đồ giữa 'Monthly\_Income' và 'City\_Category'

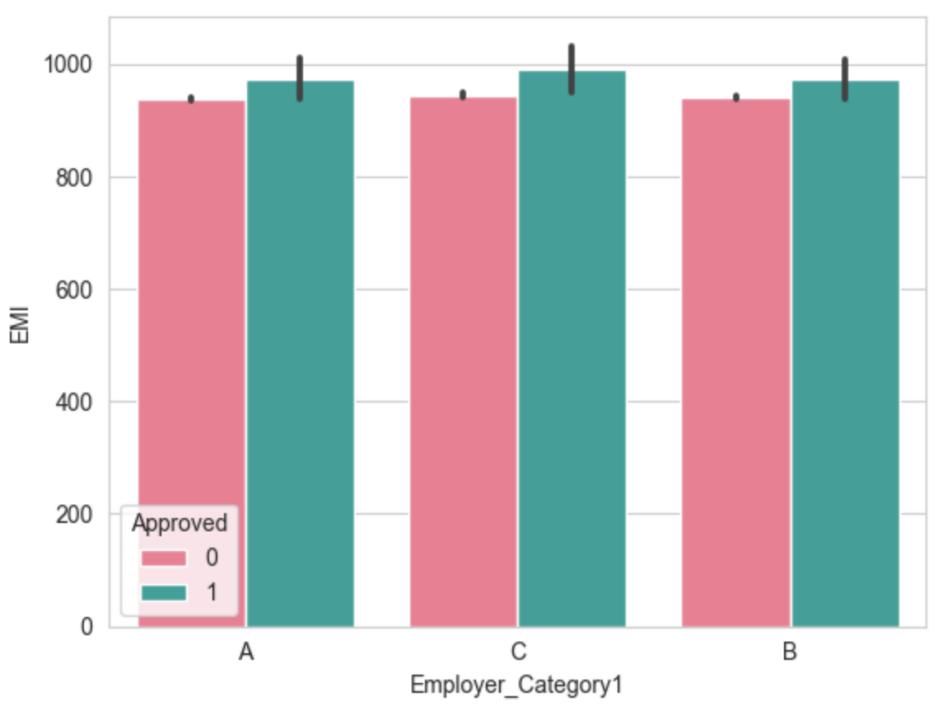
Mã trên tính giá trị trung bình của cột "Monthly\_Income" theo từng giá trị trong cột "City\_Category" của DataFrame 'df'. Kết quả: Khách hàng sống ở thành phố A có thu nhập hàng tháng cao nhất.



# Hình . Biểu đồ tròn cho 'Var1'

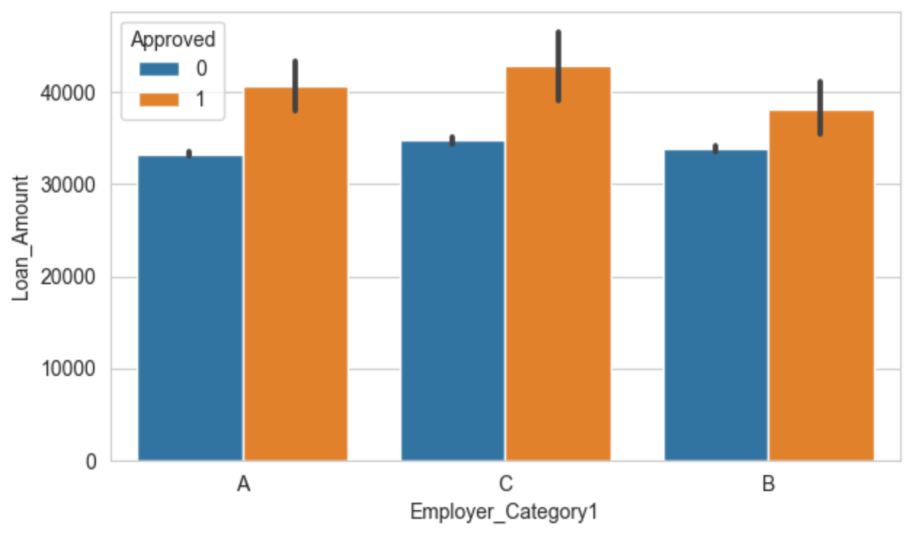
Kết quả: Số lượng khách hàng phân bố ở nhóm 0 và 1 chiếm tỉ lệ cao nhất trong cột Var1

**1.1.2. Phân tích lưỡng biến**

****

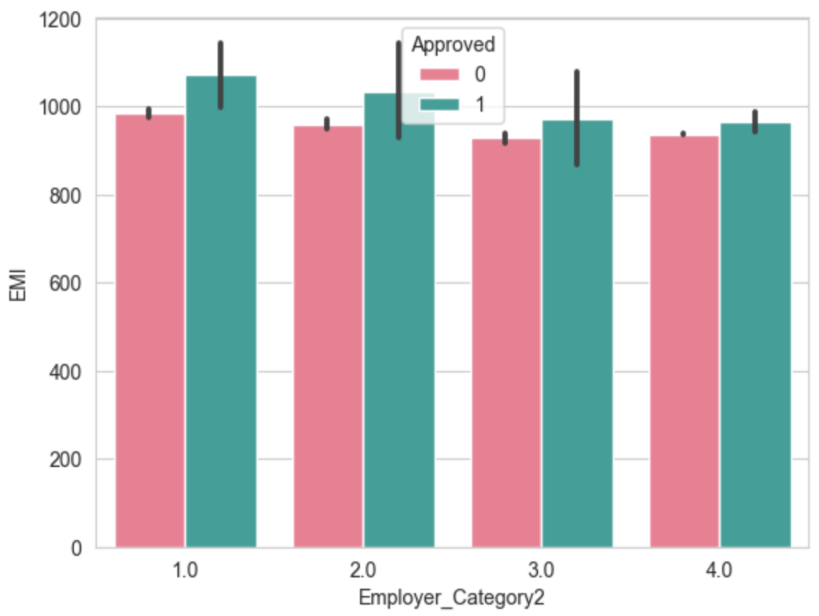
# Hình . Biểu đồ cho 'Employer\_Category\_1' và 'EMI'

Mã trên sử dụng hàm sns.barplot để tạo biểu đồ cột (bar plot) với trục x là "Employer\_Category1", trục y là "EMI" và dữ liệu được lấy từ DataFrame 'df'. Kết quả: Hầu hết người sử dụng lao động Loại 1 trả tiền Đóng EMI trong trường hợp khoản vay bị từ chối. Nhà tuyển dụng Loại 1 (C) là cao nhất trong EMI trong trường hợp khoản vay được chấp nhận.



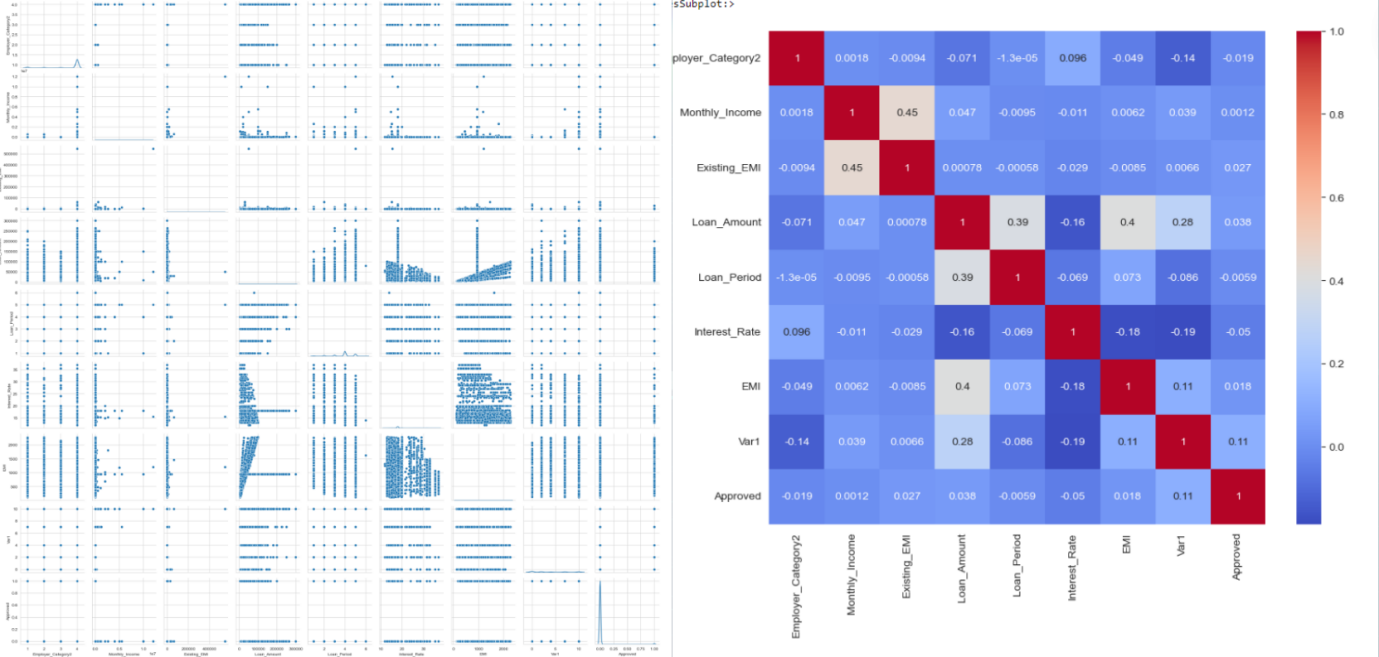
# Hình . Biểu đồ giữa 'Employer\_Category\_1' và 'Loan\_Amount'

Kết quả: (C) là loại có số tiền vay được chấp thuận cao nhất (B) là loại có số tiền vay được chấp thuận thấp nhất.



# Hình . Biểu đồ giữa 'Employer\_Category\_2' và 'EMI'

Kết quả: Nhóm khách hàng thứ nhất có giá trị phí bảo hiểm bình quân hàng tháng cao nhất trong trường hợp từ chối và chấp nhận. Điều này là do danh mục đầu tiên là cao nhất trong các giá trị cho vay trung bình.



# Hình . Biểu đồ PairGrid và Heatmap

Hai biểu đồ PairGrid và Heatmap cho thấy rằng có mối tương quan tích cực mạnh mẽ giữa EMI và Số tiền cho vay.

**1.1.3. Trả lời các câu hỏi phân tích**

Thông qua việc phân tích bộ dữ liệu, lập và quan sát các biểu đồ, mô hình trực quan hóa dữ liệu, ta có thể trả lời các câu hỏi sau:

*Giới tính của người nộp đơn thường xuyên nhất là gì?*

→ Hơn một nửa khách hàng của ngân hàng là nam giới. Số lượng khách hàng nam nhiều hơn nữ. Lượng khách hàng nam của ngân hàng vay nhiều hơn khách hàng nữ.

*Thu nhập trung bình hàng tháng của mỗi giới tính là bao nhiêu?*

→ Nữ giới: 2957.057255; Nam giới: 6318.155893 Thu nhập trung bình hàng tháng của nam giới cao hơn và do đó họ dễ chấp nhận các khoản vay hơn

*Số tiền cho vay có ảnh hưởng đến EMI không?*→ Có mối tương quan thuận giữa Số tiền cho vay và EMI.

*Giá trị khoản vay lớn nhất được thực hiện bởi khách hàng ở mọi thành phố là bao nhiêu?*→ Là 300000.0

*Tổng thu nhập hàng tháng của khách hàng từ mỗi thành phố là bao nhiêu?*→ Thành phố A: 261542007.4 Thành phố B: 22141113.4 Thành phố C: 47146955.0

*Loại ngân hàng chính thường xuyên nhất là gì?*→ Loại ngân hàng P.  
*Danh mục nguồn cho vay được chấp nhận và từ chối nhiều nhất?*→ Danh mục nguồn (B) bị từ chối nhiều nhất đối với Khoản vay.

*Giới tính nào được chấp nhận nhất?*→ Khách hàng nam nhận được sự chấp nhận cao hơn.

1.2. Kết quả thử nghiệm và phân tích đối với mô hình máy họcĐể thu hẹp và cải thiện hiệu suất cho bộ dữ liệu, nhóm chúng tôi chỉ sử dụng dữ liệu của Primary\_Bank\_Type loại P và 26 hạng mục được chọn ở trên cho mô hình máy học.

**1.2.1. Kết quả thử nghiệm**

Trong nghiên cứu này, chúng tôi đã áp dụng ba mô hình phân loại trên bộ dữ liệu của chúng tôi và so sánh các yếu tố kết quả. Dưới đây là bảng kết quả cho các mô hình máy học:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Model Name | Kappa Score | F1 Score | Accuracy | Recall | Precision |
| 1 | Decision Tree Classifer | 0.638075 | 0.787879 | 0.817941 | 0.676221 | 0.943704 |
| 2 | Kneighbors Classifier | 0.391224 | 0.608145 | 0.695612 | 0.472399 | 0.853362 |
| 3 | Logistic Regression | 0.480113 | 0.742718 | 0.734050 | 0.767741 | 0.719274 |

**1.2.2. Phân tích**

**I. Decision Tree Classifier**

Dựa trên bảng kết quả, chúng ta có một mô hình phân loại có tên là "Decision Tree Classifier". Dưới đây là các giá trị đánh giá hiệu suất của mô hình:

- Kappa Score: Kappa Score đạt giá trị 0.638075, cho thấy mức độ đồng thuận của mô hình hơi cao hơn so với đồng thuận ngẫu nhiên.

- F1 Score: F1 Score đạt giá trị 0.787879, biểu thị mức độ kết hợp của độ chính xác và độ phủ của mô hình.

- Accuracy: Accuracy (Độ chính xác) của mô hình đạt giá trị 0.817941, biểu thị tỷ lệ dự đoán chính xác của mô hình trên tập dữ liệu kiểm tra.

- Recall: Recall (Độ phủ) của mô hình đạt giá trị 0.676221, biểu thị tỷ lệ các trường hợp dự đoán đúng với lớp dương (positive) so với tất cả các trường hợp thực tế lớp dương.

- Precision: Precision (Độ chính xác) của mô hình đạt giá trị 0.943704, biểu thị tỷ lệ các trường hợp dự đoán đúng với lớp dương so với tất cả các trường hợp được dự đoán là lớp dương.

Kết quả này cho thấy mô hình Decision Tree Classifier có một mức độ đồng thuận tương đối và đạt hiệu suất tốt với độ chính xác cao và độ chính xác dự đoán lớp dương khá cao. Tuy nhiên, độ phủ của mô hình có thể cải thiện để đảm bảo rằng nó dự đoán đúng nhiều hơn các trường hợp thực tế của lớp dương.

**II. Kneighbors Classifier**

Dựa trên bảng kết quả, chúng ta có một mô hình phân loại có tên là "KNeighbors Classifier". Dưới đây là các giá trị đánh giá hiệu suất của mô hình:

- Kappa Score: Kappa Score đạt giá trị 0.391224, cho thấy mức độ đồng thuận của mô hình thấp hơn so với đồng thuận ngẫu nhiên.

- F1 Score: F1 Score đạt giá trị 0.608145, biểu thị mức độ kết hợp của độ chính xác và độ phủ của mô hình.

- Accuracy: Accuracy (Độ chính xác) của mô hình đạt giá trị 0.695612, biểu thị tỷ lệ dự đoán chính xác của mô hình trên tập dữ liệu kiểm tra.

- Recall: Recall (Độ phủ) của mô hình đạt giá trị 0.472399, biểu thị tỷ lệ các trường hợp dự đoán đúng với lớp dương (positive) so với tất cả các trường hợp thực tế lớp dương.

- Precision: Precision (Độ chính xác) của mô hình đạt giá trị 0.853362, biểu thị tỷ lệ các trường hợp dự đoán đúng với lớp dương so với tất cả các trường hợp được dự đoán là lớp dương.

Kết quả này cho thấy mô hình KNeighbors Classifier có một mức độ đồng thuận thấp và đạt hiệu suất tương đối thấp. Độ chính xác và độ chính xác dự đoán lớp dương của mô hình có thể cao, tuy nhiên, độ phủ và kappa score thấp, cho thấy mô hình không dự đoán chính xác các trường hợp của lớp dương. Điều này có thể chỉ ra rằng mô hình cần được cải thiện để đạt được hiệu suất tốt hơn trên tập dữ liệu kiểm tra.

**III. Logistic Regression**

Dựa trên bảng kết quả, chúng ta có một mô hình Logistic Regression. Dưới đây là các giá trị đánh giá hiệu suất của mô hình:

- Kappa Score: Kappa Score đạt giá trị 0.480113, cho thấy mức độ đồng thuận của mô hình có một mức độ trung bình.

- F1 Score: F1 Score đạt giá trị 0.742718, biểu thị mức độ kết hợp của độ chính xác và độ phủ của mô hình.

- Accuracy: Accuracy (Độ chính xác) của mô hình đạt giá trị 0.734050, biểu thị tỷ lệ dự đoán chính xác của mô hình trên tập dữ liệu kiểm tra.

- Recall: Recall (Độ phủ) của mô hình đạt giá trị 0.767741, biểu thị tỷ lệ các trường hợp dự đoán đúng với lớp dương (positive) so với tất cả các trường hợp thực tế lớp dương.

- Precision: Precision (Độ chính xác) của mô hình đạt giá trị 0.719274, biểu thị tỷ lệ các trường hợp dự đoán đúng với lớp dương so với tất cả các trường hợp được dự đoán là lớp dương.

Kết quả này cho thấy mô hình Logistic Regression có một mức độ đồng thuận trung bình và hiệu suất tương đối. Độ chính xác và độ phủ của mô hình đạt giá trị tương đối cao. Có sự cân đối giữa precision và recall, điều này cho thấy mô hình có khả năng dự đoán tốt cả hai lớp dương và lớp âm. Tuy nhiên, để đạt hiệu suất tốt hơn, có thể cần xem xét cải thiện mô hình thông qua điều chỉnh các siêu tham số hoặc sử dụng các phương pháp khác để tăng cường đặc trưng.

# F. Kết luận

## 1. Ưu điểm*:*

 Đánh giá được rủi ro tín dụng: Phân tích tín dụng khách hàng giúp ngân hàng hoặc các tổ chức tài chính đánh giá được rủi ro khi cho khách hàng vay tiền. Việc phân tích phân tích sẽ giúp giảm thiểu rủi ro mất tiền và tăng khả năng thu hồi nợ.

Tăng hiệu quả vận hành: Phân tích tín dụng cung cấp thông tin chi tiết về khách hàng giúp tăng cường hiệu quả vận hành của tổ chức tài chính. Bằng cách sử dụng các thuật toán và mô hình phân tích dữ liệu, ngân hàng và các tổ chức tài chính có thể tự động hóa quá trình vay vốn, giảm thời gian xử lý và tăng khả năng chính xác trong việc đưa ra quyết định.

Phát hiện gian lận tín dụng: các ngân hàng hoặc tổ chức tín dụng có thể phát hiện các hành vi gian lận trong việc vay vốn. Phân tích các mô hình từ dữ liệu có thể ngăn chặn và cảnh báo các hoạt động gian lận như vay không trả hoặc sử dụng thông tin giả mạo.

Tăng cường quản lý rủi ro: Giúp quản lý rủi ro một cách hiệu quả hơn bằng cách theo dõi và phân tích tình trạng vay nợ của khách hàng nhằm đưa ra các biện pháp phòng ngừa và quản lý rủi ro một cách chặt chẽ nhằm đảm bảo sự ổn định và bền vững của hoạt động tài chính.

Đối với dự án phân tích tín dụng khách hàng sử dụng dữ liệu vay vốn nhóm đã sử dụng ba mô hình bao gồm Decision Tree Classifier, Kneighbors Classifier và Logistic Regression để có thể đánh giá kết quả một cách chính xác nhất:

+ Đa dạng hóa được phương pháp đánh giá và kiểm tra, đưa ra cái nhìn tổng quan, đa chiều về dự báo tín dụng.

+ Tăng được độ chính xác và tin cậy.

+ Xem xét được các yếu tố quan trọng như Decision Tree Classifier cho thấy được các quyết định và quy tắc quan trọng trong việc phân loại tín dụng. Kneighbors Classifier và Logistic Regression cung cấp về thông tin trọng số và các biến đầu vào lên kết quả dự báo.

+ Khả năng đánh giá hiệu suất: so sánh và đánh giá hiệu suất giữa ba mô hình giúp nhóm có cái nhìn tổng quan và lựa chọn được mô hình phù hợp với mục tiêu đề ra ban đầu

## 2. Hạn chế:

Các kết quả mà mô hình đưa ra có sự chênh lệch khác nhau vì vậy cần phải có sự xem xét và kiểm tra sâu hơn để hiểu được nguyên nhân vì sao lại có sự chênh lệch như vậy.

Mô hình Decision Tree Classifier, Kneighbors Classifier có tính diễn giải kém hơn so với một số mô hình thuật toán khác.

Việc lựa chọn và tinh chỉnh mô hình đòi hỏi các kiến thức và chuyên môn kinh nghiệm cao vì thế nếu không có sự lựa chọn đúng sẽ rất khó cho việc tối ưu hiệu quả mô hình.

Rủi ro của việc dự báo: mặc dù có thể đưa ra dự báo về khả năng trả nợ của khách hàng nhưng vẫn không thể đảm bảo được tính chính xác tuyệt đối. Việc này có thể xảy ra trong quá trình dự báo và đánh giá hiệu quả.

Mô hình phân tích dữ liệu có thể không linh hoạt trong các tình huống đặc biệt hoặc không bình thường. Khi đánh giá cho các khách hàng có tình huống đặc biệt các mô hình trên sẽ không đảm bảo đáp ứng được đầy đủ dẫn đến các quyết định không chính xác.

## 3. Phương hướng phát triển

1. **Mở rộng và cải thiện mô hình**

* Áp dụng và so sánh các mô hình phân tích dữ liệu khác nhau như Random Forest, Support Vector Machines (SVM), Neural Networks, Gradient Boosting và Ensemble Methods. Mỗi mô hình có ưu điểm và hạn chế riêng, và việc sử dụng nhiều mô hình khác nhau có thể cung cấp cái nhìn đa chiều về cải thiện tính chính xác của dự báo
* Tinh chỉnh các siêu tham số của mô hình để tối ưu hóa hiệu suất dự báo. Sử dụng các kỹ thuật như Grid Search hoặc Random Search để tìm kiếm các giá trị tối ưu cho các tham số của mô hình

1. **Xử lý dữ liệu không cân bằng**

* Đối với dữ liệu có sự mất cân bằng mạnh giữa các giá trị 0 và 1 như trong cột “Approved”, áp dụng các phương pháp xử lý dữ liệu không cân bằng như Oversampling, Undersampling, hoặc Synthetic Minority Over-sampling Technique (SMOTE). Các phương pháp này giúp tạo ra một tập dữ liệu cân bằng hơn và đảm bảo rằng mô hình không bị thiên vị về một lớp

1. **Đánh giá mô hình sâu hơn**

* Ngoài việc đánh giá tỷ lệ chính xác của mô hình, cần xem xét các độ đo khác nhau như Recall ( Độ nhạy), Precision ( Độ chính xác), F1-Score để có cái nhìn toàn diện về hiệu suất của mô hình. Điều này giúp đánh giá khả năng dự báo của mô hình trong việc phân loại khách hàng là “Approved” hoặc “Not Approved” một cách chính xác và cân nhắc hơn
* Vẽ đồ thị ROC và tính toán AUC ( Area Under the Curve) để đánh giá khả năng phân loại mô hình. Đồ thị ROC cung cấp thông tin về tỷ lệ dương tính và tỷ lệ âm tính, và AUc đo lường khả năng phân loại của mô hình trên toàn bộ các ngưỡng
* Áp dụng Cross-validation để đánh giá hiệu suất của mô hình trên nhiều tập dữ liệu kiểm tra khác nhau và đảm bảo rằng tính đáng tin cậy của kết quả đánh giá

1. **Tăng cường diễn giải**

* Sử dụng các phương pháp giải thích mô hình như LIME ( Local Interpretable Model-agnostic Explanations ) hoặc SHAP (SHapley Additive exPlanations) để hiểu rõ hơn về quá trình ra quyết định của các mô hình phức tạp như Random Forest, Neutral Networks. Các phương pháp này giúp chúng ta giải thích tầng lớp đặc trưng quan trọng và quyết định quan trọng trong mô hình

1. **Đánh giá và tối ưu hiệu suất**

* Theo dõi và đánh giá hiệu suất của mô hình theo thời gian. Thực hiện kiểm tra định kỳ và cập nhật mô hình khi có dữ liệu mới hoặc có thay đổi trong môi trường kinh doanh.
* Tối ưu hóa mô hình bằng cách thay đổi các tham số và kiểm tra hiệu suất. Tính toán các độ đo hiệu suất khác nhau trên tập dữ liệu kiểm tra và tinh chỉnh mô hình để đạt được hiệu suất tốt nhất

1. **Tăng cường tính diễn giải của mô hình**

* Nếu tính diễn giải là yếu tố quan trọng, có thể xem xét sử dụng mô hình có tính diễn giải cao hơn Logistic Regression hoặc Rule-based Models để đưa ra các quyết định có thể giải thích được.

1. **Sử dụng dữ liệu mở rộng**

* Nếu dữ liệu hiện có không đủ để đạt được hiệu suất tốt, có thể xem xét sử dụng các nguồn dữ liệu bên ngoài hoặc dữ liệu mở rộng để cung cấp thông tin và cải thiện khả năng dự báo.

Tổng quan, phát triển dự án phân tích tín dụng khách hàng trong vay vốn đòi hỏi việc thử nghiệm và đánh giá nhiều mô hình khác nhau, cải thiện tính chính xác và diễn giải của mô hình, và liên tục theo dõi và tối ưu hiệu suất. Ngoài ra, cần lưu ý rằng dự báo tín dụng có thể mang theo một số rủi ro và không đảm bảo tính chính xác tuyệt đối, do đó cần áp dụng các biện pháp kiểm soát rủi ro và xem xét nhiều yếu tố khác nhau để đưa ra quyết định về việc cấp vay tín dụng

**Tài liệu tham khảo:**

[1] Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of business research*, *70*, 263-286.

[2] Doko, F., Kalajdziski, S., & Mishkovski, I. (2021). Credit risk model based on central bank credit registry data. *Journal of Risk and Financial Management*, *14*(3), 138.

[3] Serengil, S. I., Imece, S., Tosun, U. G., Buyukbas, E. B., & Koroglu, B. (2022). A Comparative Study of Machine Learning Approaches for Non Performing Loan Prediction with Explainability. *International Journal of Machine Learning and Computing*, *12*(5).

[4] Allaymoun, M. H., Qaradh, S., Salman, M., & Hasan, M. (2022). Big Data Analysis and Data Visualization to Help Make a Decision-Islamic Banks Case Study. In *Explore Business, Technology Opportunities and Challenges‎ After the Covid-19 Pandemic* (pp. 54-63). Cham: Springer International Publishing.

[5] Machado, M. R., & Karray, S. (2022). Assessing credit risk of commercial customers using hybrid machine learning algorithms. *Expert Systems with Applications*, *200*, 116889.

S. Ahmed, R. Paul and A. S. M. L. Hoque, "Knowledge discovery from academic data using Association Rule Mining," 2014 17th International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2014, pp. 314-319, doi: 10.1109/ICCITechn.2014.7073107.

V. Aggarwal, V. Gupta, P. Singh, K. Sharma and N. Sharma, "Detection of Spatial Outlier by Using Improved Z-Score Test," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019, pp. 788-790, doi: 10.1109/ICOEI.2019.8862582.

Daniel B. Suits (1957) Use of Dummy Variables in Regression Equations, Tạp chí của Hiệp hội Thống kê Hoa Kỳ, 52:280, 548-551, DOI: [10.1080/01621459.1957.10501412](https://doi.org/10.1080/01621459.1957.10501412)

Deepa, B., & Ramesh, K. (2022). Epileptic seizure detection using deep learning through min max scaler normalization. International Journal of Health Sciences, 6(S1), 10981–10996. <https://doi.org/10.53730/ijhs.v6nS1.7801>

Dina Elreedy, Amir F. Atiya (2019). A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. Information Sciences, 505, 32-64.[https://doi.org/10.1016/j.ins.2019.07.070](https://doi.org/10.1016/j.ins.2019.07.070?fbclid=IwAR1LMypDQxrhv9qTLTeHsI6NfqpC71OrG-HKs2NoUMPfoB9AVG1eDdWc9bI)

Jason Brownlee (2019). Feature Selection with numerical and categorical data. <https://elearn.daffodilvarsity.edu.bd/pluginfile.php/1225702/mod_label/intro/Feature%20Selection%20with%20numerical%20and%20categorical%20data.pdf>

Aggarwal, V., Gupta, V., Singh, P., Sharma, K., & Sharma, N. (2019). Detection of spatial outlier by using improved Z-score test. Paper presented at the 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI).

Brownlee, J. (2019). How to choose a feature selection method for machine learning. Machine Learning Mastery, 10.

Suits, D. B. (1957). Use of dummy variables in regression equations. Journal of the American Statistical Association, 52(280), 548-551.

 Analytics Vidhya. (2021). 5 Techniques to Handle Imbalanced Data for a Classification Problem.https://www.analyticsvidhya.com/blog/2021/06/5-techniques-to-handle-imbalanced-data-for-a-classification-problem/