Shape

Description automatically generated

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

**MÔN KHAI PHÁ DỮ LIỆU**

Logo

Description automatically generated-----🙞🙜🕮🙞🙜-----

**BÁO CÁO ĐỀ TÀI CUỐI KỲ**

**⁃⁃⁃⁃⁃⁃⁃⁃⁃⁃⁃**

**DỰ ĐOÁN RỦI RO CHO VAY TÍN DỤNG**

**LỚP HỌC PHẦN: DAMI330484\_23\_2\_03CLC**

**HỌC KỲ II – NĂM HỌC 2023-2024**

**Sinh viên thực hiện:**

**- 21110820: Trương Thị Thùy Dung**

**- 21110175: Nguyễn Văn Hào**

**- 21110155 : Trần Văn Bảo Duy**

**- 21110316 : Bùi Đặng Thùy Thương**

**Giảng viên hướng dẫn : ThS. Trần Trọng Bình**

*Thành phố Hồ Chí Minh, tháng 5 năm 2024*

|  |  |
| --- | --- |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **ĐẠI HỌC SƯ PHẠM KỸ THUẬT TPHCM**  **KHOA CÔNG NGHỆ THÔNG TIN**  **MÔN KHAI PHÁ DỮ LIỆU**  **­­­­­­­­­­­­­­­** | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA**  **VIỆT NAM**  **Độc lập – Tự do – Hạnh phúc** |

-----🙞🙜🕮🙞🙜-----

Logo

Description automatically generated

**BÁO CÁO ĐỀ TÀI CUỐI KỲ**

**⁃⁃⁃⁃⁃⁃⁃⁃⁃⁃⁃**

**DỰ ĐOÁN RỦI RO CHO VAY TÍN DỤNG**

**LỚP HỌC PHẦN:** **DAMI330484\_23\_2\_03CLC**

**HỌC KỲ II – NĂM HỌC 2023-2024**

**Sinh viên thực hiện:**

**- 21110820: Trương Thị Thùy Dung**

**- 21110175: Nguyễn Văn Hào**

**- 21110155 : Trần Văn Bảo Duy**

**- 21110316 : Bùi Đặng Thùy Thương**

**Giảng viên hướng dẫn : ThS. Trần Trọng Bình**

*Thành phố Hồ Chí Minh, tháng 5 năm 2024*

|  |  |
| --- | --- |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **ĐẠI HỌC SƯ PHẠM KỸ THUẬT TPHCM**  **KHOA CÔNG NGHỆ THÔNG TIN**  **MÔN KHAI PHÁ DỮ LIỆU**  **­­­­­­­­­­­­­­­** | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA**  **VIỆT NAM**  **Độc lập – Tự do – Hạnh phúc** |

**DANH SÁCH NHÓM THỰC HIỆN ĐỀ TÀI**

**MÔN NHẬP MÔN DỮ LIỆU LỚN**

**HỌC KỲ II NĂM HỌC 2023-2024**

1. **Giảng viên hướng dẫn:** ThS. Trần Trọng Bình
2. **Mã lớp học**: DAMI330484\_23\_2\_03CLC
3. **Tên đề tài :** Dự đoán rủi ro cho vay tín dụng
4. **Danh sách nhóm viết tiểu luận cuối kỳ:**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **HỌ VÀ TÊN SINH VIÊN** | **Mã số sinh viên** | **Tỉ lệ % tham gia** |
| 01 | Trương Thị Thùy Dung | 21110820 | 100% |
| 02 | Nguyễn Văn Hào | 21110175 | 100% |
| 03 | Trần Văn Bảo Duy | 21110155 | 100% |
| 04 | Bùi Đặng Thùy Thương | 21110316 | 100% |

NHẬN XÉT CỦA GV

Giảng viên chấm điểm

**LỜI CẢM ƠN**

Chúng em xin gửi lời cảm ơn sâu sắc đến Thầy Trần Trọng Bình, Thầy đã tận tình giảng dạy trong suốt quá trình tìm hiểu và học tập môn Khai phá dữ liệu. Trong quá trình nghiên cứu và học, thầy đã nhiệt huyết giảng dạy và những kiến thức mà thầy đã truyền đạt không chỉ là nền tảng cho quá trình nghiên cứu, thực hiện báo cáo cuối kỳ mà còn là tư liệu quý báu để chúng em trong quá trình học tập và tìm hiểu sau này cũng như là trên con đường nghề nghiệp phía trước.

Một lần nữa, em xin bày tỏ lòng biết ơn đến Thầy. Xin kính chúc Thầy luôn dồi dào sức khỏe, đạt được nhiều thành công trong công việc.

*Thành phố Hồ Chí Minh, tháng 5 năm 2024*

|  |  |
| --- | --- |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **ĐẠI HỌC SƯ PHẠM KỸ THUẬT TPHCM**  **KHOA CÔNG NGHỆ THÔNG TIN**  **BỘ MÔN KHAI PHÁ DỮ LIỆU**  **­­­­­­­­­­­­­­­** | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA**  **VIỆT NAM**  **Độc lập – Tự do – Hạnh phúc** |

**KẾ HOẠCH PHÂN CÔNG NHIỆM VỤ**

**THỰC HIỆN ĐỀ TÀI CUỐI KÌ**

**HỌC KỲ II NĂM HỌC 2023-2024**

|  |  |
| --- | --- |
| **Công việc** | **Người thực hiện** |
| 1. Tìm hiểu chung về dataset | Cả nhóm |
| 1. Tiền xử lý dữ liệu | Trương Thị Thuỳ Dung  Trần Văn Bảo Duy |
| 1. Kiểm định giả thuyết | Bùi Đặng Thuỳ Thương |
| 1. Logistic Regression | Trần Văn Bảo Duy |
| 1. KNN | Bùi Đặng Thuỳ Thương |
| 1. Random Forest | Trương Thị Thuỳ Dung  Bùi Đặng Thuỳ Thương |
| 1. AdaBoost | Nguyễn Văn Hào |
| 1. Gradient Boosting | Nguyễn Văn Hào |
| 1. Navie Bayes | Trần Văn Bảo Duy |
| 1. Artificial Neural Network (ANN) | Trương Thị Thuỳ Dung |
| 1. Phân cụm áp dụng Kmeans - clustering | Nguyễn Văn Hào |

**MỤC LỤC**

-----🙞🙜🕮🙞🙜-----

[**DANH MỤC HÌNH ẢNH**](#_Toc167267384)

[**PHẦN MỞ ĐẦU** 1](#_Toc167267385)

[**1.** **Phát biểu về bài toán** 1](#_Toc167267386)

[**2.** **Mục đích thực hiện** 1](#_Toc167267387)

[**3.** **Tổng quan về dữ liệu** 2](#_Toc167267388)

[**4.** **Bố cục đề tài** 2](#_Toc167267389)

[**PHẦN NỘI DUNG** 3](#_Toc167267390)

[**CHƯƠNG 1: DATASET** 3](#_Toc167267391)

[**1.1.** **Tổng quan về tập dữ liệu** 3](#_Toc167267392)

[**1.2.** **Khám phá tổng quát và định hướng phân tích** 3](#_Toc167267393)

[**1.3.** **Xử lý dữ liệu thiếu và dọn dẹp dữ liệu** 5](#_Toc167267394)

[**CHƯƠNG 2. TIỀN XỬ LÝ VÀ CHUẨN HÓA DỮ LIỆU** 7](#_Toc167267395)

[**2.1.** **Các bước phân tích và nhận xét về từng đặc trưng** 7](#_Toc167267396)

[**2.1.2.** **Đặc trưng về sở hữu xe (Car Owners)** 8](#_Toc167267397)

[**2.1.3.** **Đặc trưng tình trạng hôn nhân** 9](#_Toc167267398)

[**2.1.4.** **Đặc trưng nghề nghiệp** 10](#_Toc167267399)

[**2.1.5.** **Tỷ lệ phân bổ của biến Thành phố và Bang** 11](#_Toc167267400)

[**2.1.6.** **Đặc trưng bang người dùng sinh sống** 13](#_Toc167267401)

[**2.2.** **Phân tích mối tương quan giữa các biến** 17](#_Toc167267402)

[**2.3.** **Xử lý ngoại lai** 19](#_Toc167267403)

[**2.3.1.** **Xứ lý ngoại lại của đặc trưng Tuổi** 19](#_Toc167267404)

[**2.3.2.** **Xứ lý ngoại lại của đặc trưng Income** 19](#_Toc167267405)

[**2.4.** **Trực quan cung cấp cái nhìn rõ ràng về sự phân bổ của biến mục tiêu, Risk\_Flag** 20](#_Toc167267406)

[**CHƯƠNG 3: PHÂN TÍCH THỐNG KÊ VÀ KIỂM ĐỊNH GIẢ THUYẾT** 22](#_Toc167267407)

[**3.1.** **Phương pháp thống kê sử dụng** 22](#_Toc167267408)

[**3.2.** **Phân tích và kiểm định giả thuyết** 22](#_Toc167267409)

[**3.3.** **Đánh giá mức độ tin cậy của kết quả** 25](#_Toc167267410)

[**CHƯƠNG 4: MÔ HÌNH HÓA VÀ DỰ ĐOÁN** 27](#_Toc167267411)

[**4.1.** **Chuẩn hóa và chuẩn bị dữ liệu huấn luyện** 27](#_Toc167267412)

[**4.1.1.** **Phân tích thành phần chính (PCA) cho CURRENT\_JOB\_YRS và Experience** 27](#_Toc167267413)

[**4.1.2.** **Chuẩn bị dữ liệu** 29](#_Toc167267414)

[**4.2.** **Mô hình logictis regresion** 30](#_Toc167267415)

[**4.2.1.** **Sơ lược về mô hình** 30](#_Toc167267416)

[**4.2.2.** **Đào tạo và đánh giá mô hình** 31](#_Toc167267417)

[**4.2.3.** **Đánh giá kết quả và giải thích kết quả khai phá** 32](#_Toc167267418)

[**4.3.** **Mô hình KNN** 35](#_Toc167267419)

[**4.3.1.** **Sơ lược về mô hình** 35](#_Toc167267420)

[**4.3.2.** **Đào tạo và đánh giá mô hình** 35](#_Toc167267421)

[**4.3.3.** **Đánh giá kết quả và giải thích kết quả khai phá** 36](#_Toc167267422)

[**4.4.** **Mô hình Random Forest** 37](#_Toc167267423)

[**4.4.1.** **Sơ lược về mô hình** 37](#_Toc167267424)

[**4.4.2.** **Đào tạo và đánh giá mô hình** 38](#_Toc167267425)

[**4.4.3.** **Đánh giá kết quả và giải thích kết quả khai phá** 41](#_Toc167267426)

[**4.5.** **Mô hình AdaBoosting** 43](#_Toc167267427)

[**4.5.1.** **Sơ lược về mô hình** 43](#_Toc167267428)

[**4.5.2.** **Đào tạo và đánh giá mô hình** 44](#_Toc167267429)

[**4.5.3.** **Đánh giá kết quả và giải thích kết quả khai phá** 45](#_Toc167267430)

[**4.6.** **Mô hình GradientBoosting** 46](#_Toc167267431)

[**4.6.1.** **Sơ lược về mô hình** 46](#_Toc167267432)

[**4.6.2.** **Đào tạo và đánh giá mô hình** 46](#_Toc167267433)

[**4.6.3.** **Đánh giá kết quả và giải thích kết quả khai phá** 48](#_Toc167267434)

[**4.7.** **Mô hình Naive Bayes** 50](#_Toc167267435)

[**4.7.1.** **Sơ lược về mô hình** 50](#_Toc167267436)

[**4.7.2.** **Đào tạo và đánh giá mô hình** 51](#_Toc167267437)

[**4.7.3.** **Đánh giá kết quả và giải thích kết quả khai phá** 52](#_Toc167267438)

[**4.8.** **Mô hình ANN** 54](#_Toc167267439)

[**4.8.1.** **Sơ lược về mô hình** 54](#_Toc167267440)

[**4.8.2.** **Đào tạo và đánh giá mô hình** 55](#_Toc167267441)

[**4.8.3.** **Đánh giá kết quả và giải thích kết quả khai phá** 59](#_Toc167267442)

[**4.9.** **Phân cụm sử dụng Kmeans** 61](#_Toc167267443)

[**4.9.1.** **Sơ lược về mô hình** 61](#_Toc167267444)

[**4.9.2.** **Sơ lược về mô hình** 62](#_Toc167267445)

[**4.9.3.** **Ý nghĩa kết quả khai phá** 65](#_Toc167267446)

[**CHƯƠNG 5. KẾT LUẬN VÀ KIẾN NGHỊ** 66](#_Toc167267447)

[**5.1.** **Tóm tắt kết quả nghiên cứu** 66](#_Toc167267448)

[**5.2.** **Kiến nghị dựa trên kết quả** 66](#_Toc167267449)

[**5.3.** **Hướng nghiên cứu và phát triển tiếp theo** 66](#_Toc167267450)

[**PHẦN ĐÁNH GIÁ LÀM VIỆC NHÓM** 68](#_Toc167267451)

[**TÀI LIỆU THAM KHẢO** 73](#_Toc167267452)

# **DANH MỤC HÌNH ẢNH**

[Hình 1. Biểu đồ cột giữa đặc trưng House Ownership và Risk Flag 7](#_Toc167225713)

[Hình 2: Biểu đồ cột giữa đặc trưng Car Ownership và Risk Flag 8](#_Toc167225714)

[Hình 3: Biểu đồ cột giữa đặc trưng Marital Status và Risk Flag 9](#_Toc167225715)

[Hình 4: Biểu đồ cột giữa đặc trưng Profession và Risk Flag 10](#_Toc167225716)

[Hình 5: Biểu đồ cột giữa đặc trưng STATE và Risk Flag 13](#_Toc167225717)

[Hình 6: Biểu đồ tần suất về tuổi tác (Age) 14](#_Toc167225718)

[Hình 7: Biểu đồ tần suất về thu nhập (Income) 16](#_Toc167225719)

[Hình 8: Biểu đồ tương quan giữa các biến 17](#_Toc167225720)

[Hình 9: Biểu đồ hộp về đặc trưng tuổi (Age) 19](#_Toc167225721)

[Hình 10: Biểu đồ hộp của đặc trưng thu nhập (Income) 20](#_Toc167225722)

[Hình 11: Biểu đồ cột và biểu đồ tròn về sự phân bổ của biến mục tiêu Risk\_Flag 20](#_Toc167225723)

[Hình 12: Kết quả kiểm định khả năng vỡ nợ theo sở hữu xe 23](#_Toc167225724)

[Hình 13: Kết quả kiểm định khả năng vỡ nợ theo tình trạng hôn nhân 24](#_Toc167225725)

[Hình 14: Kết quả kiểm định khả năng vỡ nợ theo sở hữu nhà 25](#_Toc167225726)

[Hình 15: DataFrame đã chuẩn hóa cho PCA 28](#_Toc167225727)

[Hình 16: DataFrame sau khi áp dụng PCA 28](#_Toc167225728)

[Hình 17: Nối DataFrame PCA với DataFrame ban đầu 29](#_Toc167225729)

[Hình 18: DataFrame sau khi mã hóa các biến phân loại và loại bỏ cột Id 30](#_Toc167225730)

[Hình 19: Báo cáo các độ đo của Logistic Regression 31](#_Toc167225731)

[Hình 20: Confusion matrix cho dự đoán của Logistic Regression 32](#_Toc167225732)

[Hình 21. Báo cáo các độ đo của mô hình KNN 36](#_Toc167225733)

[Hình 22. Mức độ quan trọng của các thuộc tính trong RandomForest 39](#_Toc167225734)

[Hình 23. Biểu đồ thể hiện mức độ quan trọng của các thuộc tính trong RandomForest 40](#_Toc167225735)

[Hình 24. Báo cáo về các độ đo của mô hình Random Forest 41](#_Toc167225736)

[Hình 25. Confusion matrix cho dự đoán của mô hình Random Forest 42](#_Toc167225737)

[Hình 27: Báo cáo các độ đo của mô hình GradientBoostingClassifier 47](#_Toc167225738)

[Hình 28: Confusion matrix cho dự đoán của GradientBoostingClassifier 48](#_Toc167225739)

[Hình 29: Báo cáo về các độ đo của GradientBoostingClassifier 51](#_Toc167225740)

[Hình 31. Biểu đồ biểu diễn quá trình học của mô hình ANN 59](#_Toc167225741)

# 

# **PHẦN MỞ ĐẦU**

1. **Phát biểu về bài toán**

Học máy - Machine Learning là một lĩnh vực của trí tuệ nhân tạo liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống “học” tự động từ dữ liệu để giải quyết những vấn đề cụ thể.

Và hiện nay với lượng dữ liệu khổng lồ của người dùng Internet cùng với sự bùng nổ của cách mạng phần cứng càng làm cho tốc độ máy tính được cải thiện thì Học máy đang trở thành xu hướng và ngày càng len lỏi sâu vào cuộc sống của chúng ta. Machine learning là một thành phần quan trọng của lĩnh vực khoa học dữ liệu đang phát triển. Thông qua việc sử dụng các phương pháp thống kê, các thuật toán được đào tạo để đưa ra các phân loại hoặc dự đoán và khám phá những thông tin chi tiết từ chính các dự án khai thác dữ liệu.

Thông qua các thông tin chi tiết có được để thúc đẩy việc đưa ra quyết định đối với các ứng dụng và doanh nghiệp, tác động mạnh đến các chỉ số tăng trưởng. Khi dữ liệu lớn tiếp tục nhu cầu mở rộng và phát triển đòi hỏi nhu cầu tuyển dụng các nhà khoa học dữ liệu sẽ tăng lên. Họ sẽ được yêu cầu giúp xác định các câu hỏi kinh doanh có liên quan nhất và dữ liệu để trả lời chúng.

Bài toán của machine learning thường được chia làm hai loại là dự đoán (prediction) và phân loại (classification).

1. **Mục đích thực hiện**

Thông qua phân tích dữ liệu, đưa ra giả thuyết và thiết kế các mô hình để giải quyết các vấn đề thực tế. Từ đó áp dụng các thuật toán học máy như hồi quy, phân loại và gom cụm, từ đó hiểu rõ hơn về cách mỗi phương pháp hoạt động và khi nào nên áp dụng chúng. Đồng thời, qua các thao tác cũng mang mục đích nắm vững quy trình làm việc trong dự án học máy, từ việc thu thập dữ liệu đến triển khai và đánh giá mô hình.

1. **Tổng quan về dữ liệu**

Dữ liệu là yếu tố quan trọng nhất trong bất kỳ dự án Machine Learning nào. Trong quá trình này, dữ liệu không chỉ là nguồn thông tin mà còn là "nhiên liệu" cung cấp cho mô hình Machine Learning.

Việc thành công của một dự án Machine Learning phụ thuộc phần lớn vào chất lượng và độ đa dạng của dữ liệu. Trước khi xây dựng mô hình, việc hiểu và tiền xử lý dữ liệu là bước quan trọng, bao gồm làm sạch dữ liệu, khám phá tính chất của nó, chuẩn hóa và mã hóa các biến, xử lý dữ liệu bị thiếu, và phân chia dữ liệu thành các tập huấn luyện, kiểm tra và validation.

Quản lý dữ liệu cũng là một phần không thể thiếu, đảm bảo rằng dữ liệu được bảo mật và toàn vẹn trong suốt quá trình. Tóm lại, dữ liệu là yếu tố quyết định thành công của một dự án Machine Learning, và việc làm việc hiệu quả với dữ liệu là chìa khóa để xây dựng mô hình hiệu suất cao.

1. **Bố cục đề tài**

# **PHẦN NỘI DUNG**

# **CHƯƠNG 1: DATASET**

* 1. **Tổng quan về tập dữ liệu**

Cho vay tài chính là một trong những dịch vụ phổ biến nhất được ngân hàng cung cấp, cho dù đó là vay mua nhà, vay mua ô tô, vay học vấn và v.v. Nhưng đồng thời, có những rủi ro liên quan đến việc cho vay tiền đến khách hàng. Ngân hàng có thể nào tìm ra nguy cơ liên quan đến khách hàng nào đã nộp đơn vay không?

Một tổ chức về cho vay muốn dự đoán ai có thể là những người không thực hiện nghĩa vụ thanh toán cho sản phẩm vay tiêu dùng. Họ có dữ liệu về hành vi lịch sử của khách hàng dựa trên những gì họ đã quan sát được. Do đó, khi họ thu hút khách hàng mới, họ muốn dự đoán ai là người có nguy cơ hơn và ai là người không có nguy cơ.

Link tập dữ liệu:

|  |
| --- |
| [**https://www.kaggle.com/datasets/subhamjain/loan-prediction-based-on-customer-behavior**](https://www.kaggle.com/datasets/subhamjain/loan-prediction-based-on-customer-behavior) |

* 1. **Khám phá tổng quát và định hướng phân tích**

***Các nhãn data trong tập dữ liệu:***

* ID: Mã số duy nhất của người dùng.
* Income: Thu nhập của người dùng.
* Age: Tuổi của người dùng.
* Experience: Kinh nghiệm chuyên môn của người dùng tính bằng năm.
* Profession: Nghề nghiệp của người dùng.
* Married/Single: Tình trạng hôn nhân của người dùng, có kết hôn hay chưa.
* House\_Ownership: Tình trạng sở hữu nhà, có sở hữu nhà, thuê nhà hay không sở hữu nhà.
* Car\_Ownership: Người dùng có sở hữu xe hơi hay không.
* STATE: Bang hoặc vùng lãnh thổ cư trú của người dùng.
* CITY: Thành phố hoặc khu vực cư trú của người dùng.
* CURRENT\_JOB\_YRS: Số năm kinh nghiệm làm việc hiện tại của người dùng.
* CURRENT\_HOUSE\_YRS: Số năm cư trú tại địa chỉ hiện tại của người dùng.
* Risk\_Flag: Biến mục tiêu, chỉ liệu người dùng đã gặp rủi ro về việc vỡ nợ hay không.

***Định hướng phân tích:***

**Phân tích phân phối biến mục tiêu (Risk\_Flag)**: Có thể xem xét phân phối của biến mục tiêu để hiểu tỷ lệ người dùng đã vỡ nợ so với tỷ lệ không vỡ nợ trong dữ liệu.

**Phân tích tương quan:** Có thể tiến hành phân tích tương quan giữa các biến đầu vào (Income, Age, Experience, Profession, Marital\_Status, House\_Ownership, Car\_Ownership, STATE, CITY, CURRENT\_JOB\_YRS, CURRENT\_HOUSE\_YRS) và biến mục tiêu (Risk\_Flag) để xác định mối quan hệ giữa chúng.

**Khám phá dữ liệu bất thường**: Kiểm tra dữ liệu để phát hiện các giá trị bất thường, dữ liệu còn thiếu, hoặc các điểm ngoại lai có thể ảnh hưởng đến kết quả phân tích.

**Mô hình hóa dự đoán**: Dựa trên dữ liệu lịch sử và các biến đầu vào, tổ chức có thể xây dựng các mô hình dự đoán để phân loại khách hàng thành hai nhóm: người có nguy cơ vỡ nợ và người không có nguy cơ vỡ nợ. Các mô hình có thể bao gồm học máy phân loại như Random Forest, Logistic Regression, hoặc Support Vector Machines.

**Đánh giá mô hình**: Tổ chức cần đánh giá hiệu suất của các mô hình dự đoán bằng cách sử dụng các phép đo như độ chính xác, độ nhạy, độ đặc hiệu và F1-score để đảm bảo rằng mô hình có khả năng dự đoán chính xác và hiệu quả.

**Tối ưu hóa chiến lược cho vay**: Dựa trên kết quả phân tích và dự đoán từ mô hình, tổ chức có thể tối ưu hóa chiến lược cho vay của mình bằng cách tập trung vào việc cung cấp các sản phẩm vay cho nhóm người không có nguy cơ vỡ nợ và áp đặt các biện pháp hạn chế đối với nhóm có nguy cơ cao hơn.

* 1. **Xử lý dữ liệu thiếu và dọn dẹp dữ liệu**

|  |
| --- |
| df.info() |

|  |
| --- |
| <class 'pandas.core.frame.DataFrame'>  RangeIndex: 252000 entries, 0 to 251999  Data columns (total 13 columns):  # Column Non-Null Count Dtype  --- ------ -------------- -----  0 Id 252000 non-null int64  1 Income 252000 non-null int64  2 Age 252000 non-null int64  3 Experience 252000 non-null int64  4 Married/Single 252000 non-null object  5 House\_Ownership 252000 non-null object  6 Car\_Ownership 252000 non-null object  7 Profession 252000 non-null object  8 CITY 252000 non-null object  9 STATE 252000 non-null object  10 CURRENT\_JOB\_YRS 252000 non-null int64  11 CURRENT\_HOUSE\_YRS 252000 non-null int64  12 Risk\_Flag 252000 non-null int64  dtypes: int64(7), object(6)  memory usage: 25.0+ MB |

DataFrame này có 252.000 dòng, chia thành 13 cột. Mỗi cột được liệt kê cùng với thông tin về số lượng giá trị không rỗng

|  |
| --- |
| numerical\_summary = df.describe() # For numerical variables  numerical\_columns\_count = numerical\_summary.shape[1]  print(f'There are {numerical\_columns\_count} numeric columns in the dataset') |

|  |
| --- |
| categorical\_summary = df.describe(include=['O']).T  categorical\_columns\_count = categorical\_summary.shape[0]  print(f'There are {categorical\_columns\_count} categorical columns in the dataset') |

Kiểm tra tổng các cột số trong dữ liệu và kiểm tra các loại dữ liệu trong tập dữ liệu. Kết quả thu được 7 cột dữ liệu số và 6 loại cột trong tập dữ liệu

|  |
| --- |
| There are 7 numeric columns in the dataset  There are 6 categorical columns in the dataset |

Tiếp theo kiểm tra các giá trị NULL trogn tập dữu liệu và loại bỏ nó

|  |
| --- |
| df.isnull().sum() |

|  |
| --- |
| Id 0  Income 0  Age 0  Experience 0  Married/Single 0  House\_Ownership 0  Car\_Ownership 0  Profession 0  CITY 0  STATE 0  CURRENT\_JOB\_YRS 0  CURRENT\_HOUSE\_YRS 0  Risk\_Flag 0  dtype: int64 |

Không có dữ liệu rỗng nào trong tập dữ liệu này

# **CHƯƠNG 2. TIỀN XỬ LÝ VÀ CHUẨN HÓA DỮ LIỆU**

* 1. **Các bước phân tích và nhận xét về từng đặc trưng**
     1. **Đặc trưng về sở hữu nhà** A graph showing a house ownership vs a flag

        Description automatically generated

Hình 1. Biểu đồ cột giữa đặc trưng House Ownership và Risk Flag

Phân bố sở hữu nhà: Phần lớn các cá nhân trong tập dữ liệu là người thuê nhà ('người thuê'), đông hơn đáng kể những người sở hữu nhà của họ ('sở hữu') hoặc không có quyền sở hữu ổn định ('norent\_noown').

Phân bố rủi ro: Trên tất cả các loại quyền sở hữu nhà, số lượng cá nhân không có rủi ro (Risk Flag = 0) cao hơn số người được gắn cờ là có rủi ro (Risk Flag = 1).

So sánh giữa các rủi ro: Tỷ lệ cá nhân bị gắn cờ rủi ro trong danh mục cho thuê cao hơn so với tỷ lệ trong danh mục sở hữu, cho thấy rằng người thuê nhà có thể gặp phải các yếu tố dẫn đến xếp hạng rủi ro cao hơn. Điều này có thể là do sự bất ổn về tài chính, tình trạng kinh tế thấp hơn hoặc điều kiện sống kém an toàn hơn liên quan đến tài sản thuê.

* + 1. **Đặc trưng về sở hữu xe (Car Owners)**

A graph showing a car ownership vs car ownership

Description automatically generated

Hình 2: Biểu đồ cột giữa đặc trưng Car Ownership và Risk Flag

Phân phối quyền sở hữu ô tô: Số lượng người không sở hữu ô tô ('no') lớn hơn so với những người có ('yes'). Điều này cho thấy rằng việc sở hữu ô tô không phổ biến trong nhóm dân số được trình bày trong tập dữ liệu.

Phân bố rủi ro: Trong số cả chủ sở hữu ô tô và người không sở hữu ô tô, phần lớn được phân loại là không có rủi ro (Risk Flag = 0). Điều này cho thấy rằng dân số nói chung, bất kể sở hữu ô tô, có xu hướng có mức độ rủi ro thấp.

Những người sở hữu ô tô dường như có tỷ lệ cảnh báo rủi ro thấp hơn so với những người không sở hữu ô tô. Điều này cho thấy rằng việc sở hữu ô tô có thể tương quan với mức độ rủi ro thấp hơn, điều này có thể là do các yếu tố kinh tế xã hội khác nhau như sự ổn định tài chính cao hơn của những người sở hữu ô tô.

* + 1. **Đặc trưng tình trạng hôn nhân**

**A graph of a person with a blue and orange bar

Description automatically generated with medium confidence**

Hình 3: Biểu đồ cột giữa đặc trưng Marital Status và Risk Flag

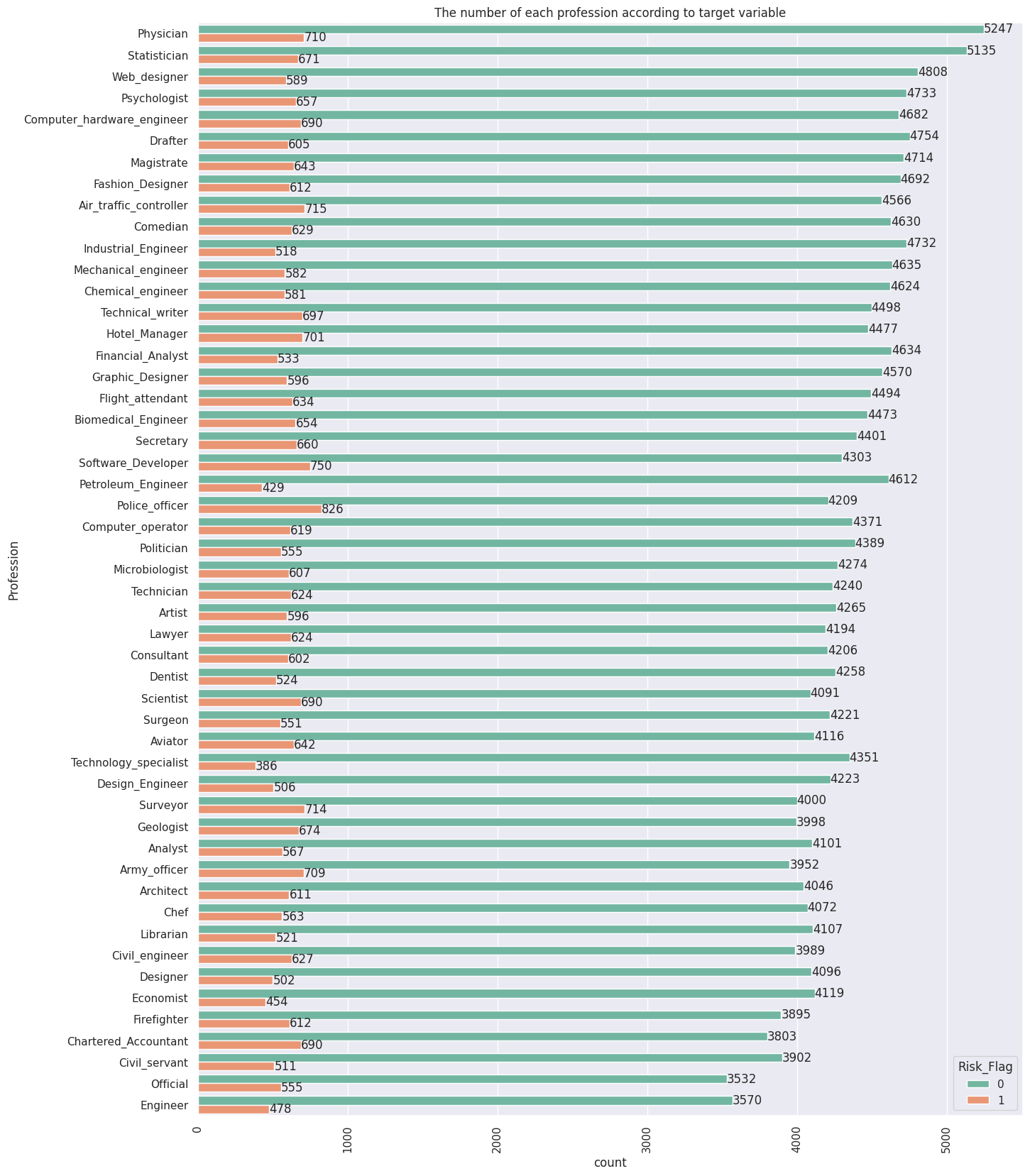
Phân bố tình trạng hôn nhân: Có số lượng cá nhân độc thân cao hơn so với những người đã kết hôn trong tập dữ liệu. Điều này có thể gợi ý một nhóm nhân khẩu học trẻ hơn hoặc một nhóm dân số có xu hướng kết hôn muộn hơn hoặc ít hơn.

Phân bố rủi ro: Tỷ lệ cá nhân có cờ rủi ro là 1 (nguy cơ cao hơn) ở những người độc thân tương đối cao hơn so với những người đã kết hôn. Điều này có thể chỉ ra rằng việc độc thân có liên quan đến nguy cơ rủi ro tăng lên một chút, có thể do các yếu tố như sự ổn định tài chính kém hơn hoặc trách nhiệm xã hội khác biệt so với những người đã kết hôn.

Tổng số người độc thân không có rủi ro cao hơn đáng kể so với số người đã kết hôn, phản ánh số lượng người độc thân cơ bản lớn hơn. Tuy nhiên, tỷ lệ rủi ro và không rủi ro trong nhóm độc thân cũng cao hơn, cho thấy các yếu tố rủi ro cụ thể hoặc khả năng dễ bị tổn thương liên quan đến việc độc thân.

Các cá nhân đã kết hôn có tỷ lệ rủi ro tương đối thấp hơn, có thể do thu nhập tổng hợp, trách nhiệm chung hoặc các yếu tố kinh tế xã hội khác góp phần ổn định và đánh giá rủi ro thấp hơn.

* + 1. **Đặc trưng nghề nghiệp**



Hình 4: Biểu đồ cột giữa đặc trưng Profession và Risk Flag

***Phân bố nghề nghiệp:***

* Bộ dữ liệu bao gồm nhiều ngành nghề khác nhau, với sự khác biệt đáng kể về số lượng giữa các ngành nghề khác nhau.
* Một số ngành nghề như 'Bác sĩ', 'Nhà thống kê' và 'Nhà thiết kế web' phổ biến hơn trong tập dữ liệu.

***Phân bố rủi ro:***

* Trong hầu hết các ngành nghề, số cá nhân không có rủi ro (Risk Flag = 0) đông hơn đáng kể những người có rủi ro (Risk Flag = 1).
* Một số ngành nghề nhất định như 'Bác sĩ', 'Nhà thiết kế web' và 'Nhà tâm lý học' có số lượng cá nhân không gặp rủi ro đặc biệt cao, cho thấy mức độ rủi ro thấp hơn liên quan đến những nghề nghiệp này.

Hình ảnh trực quan nêu bật các ngành nghề có tỷ lệ rủi ro và không rủi ro khác nhau. Ví dụ: các ngành nghề như 'Kỹ sư dầu khí' và 'Nhà phát triển phần mềm' cho thấy số lượng cá nhân gặp rủi ro cao hơn so với nhiều nghề khác, cho thấy rằng các yếu tố ngành hoặc đặc điểm công việc cụ thể có thể ảnh hưởng đến việc đánh giá rủi ro.

Mặt khác, các ngành nghề như 'Nhà thống kê' và 'Máy tính, kỹ sư phần cứng' cho thấy sự phân bổ rủi ro tương đối cân bằng nhưng vẫn có mức độ chi phối rủi ro thấp hơn.

* + 1. **Tỷ lệ phân bổ của biến Thành phố và Bang**
* ***Địa điểm có rủi ro cao:***

Bhubaneswar, Odisha có tỷ lệ rủi ro cao nhất khoảng 32,62%. Điều này có thể chỉ ra các yếu tố địa phương hoặc điều kiện kinh tế góp phần làm tăng mức độ rủi ro của người dân.

Gwalior, Madhya Pradesh và Bettiah, Bihar cũng có tỷ lệ rủi ro cao đáng kể, lần lượt là 27,27% và 26,70%. Những khu vực này có thể có những thách thức kinh tế xã hội cụ thể góp phần làm tăng mức độ rủi ro cao hơn này.

* ***Vị trí có rủi ro thấp hơn:***

Ở phía bên kia của quang phổ, các thành phố như Gandhinagar, Gujarat và Dehradun, Uttarakhand có tỷ lệ rủi ro thấp hơn đáng kể, lần lượt khoảng 2,61% và 2,63%. Điều này cho thấy các điều kiện kinh tế hoặc nhân khẩu học ổn định hơn có thể giảm thiểu các yếu tố rủi ro.

* ***Những biến đổi về mặt địa lý:***

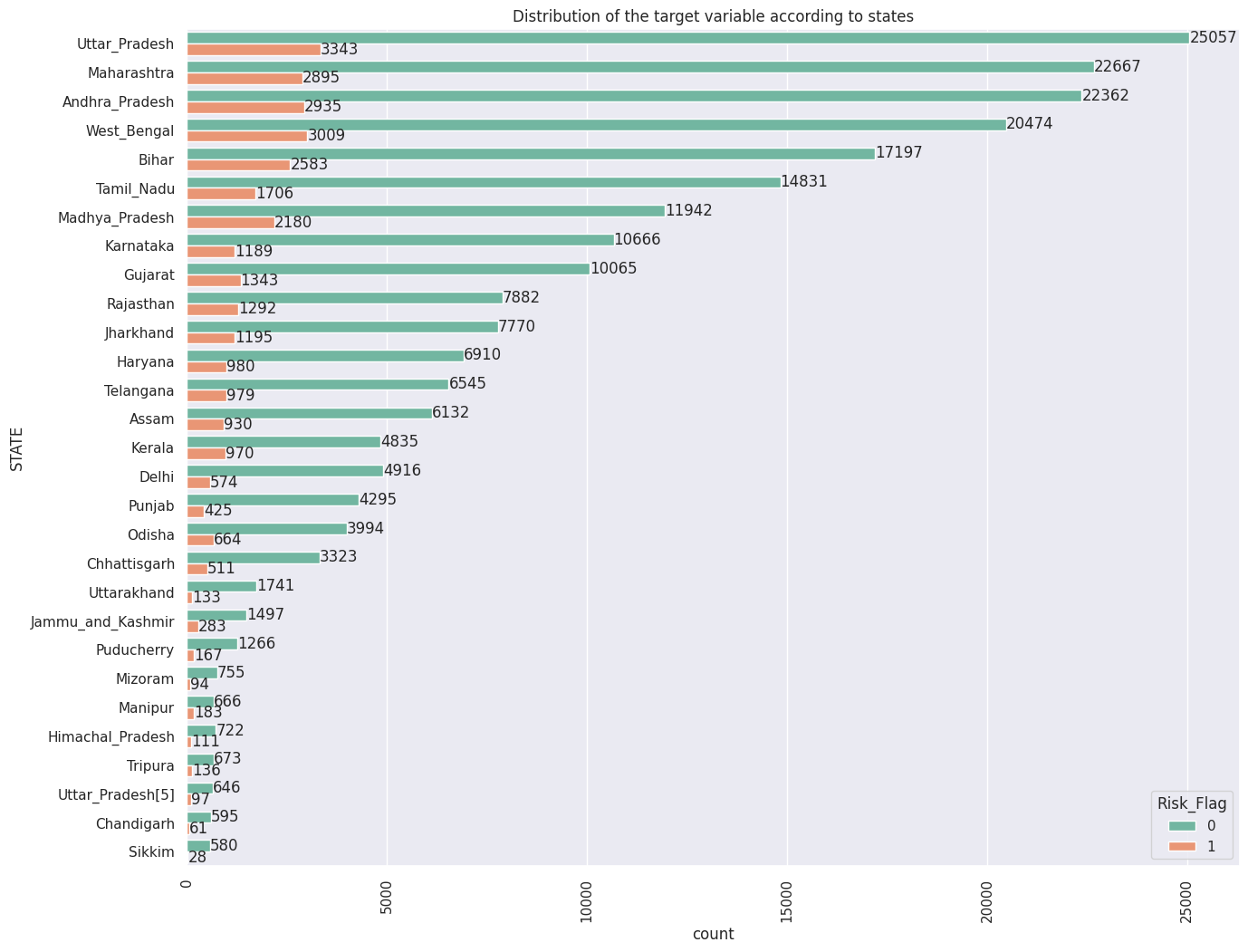
Có sự khác biệt đáng kể về tỷ lệ rủi ro giữa các tiểu bang và thành phố khác nhau. Sự đa dạng này có thể phản ánh các điều kiện kinh tế, cơ hội việc làm hoặc chính sách khu vực khác nhau ảnh hưởng đến sự ổn định tài chính và rủi ro.

* ***Ý nghĩa đối với chính sách và kinh doanh:***

Các khu vực có rủi ro cao có thể yêu cầu giáo dục tài chính có mục tiêu, các chương trình phát triển kinh tế hoặc các sản phẩm cho vay và bảo hiểm chuyên biệt để giải quyết các yếu tố rủi ro cơ bản.

Ngược lại, các khu vực có rủi ro thấp hơn có thể là địa điểm hấp dẫn đối với các doanh nghiệp do có cơ sở người tiêu dùng ổn định hơn.

* + 1. **Đặc trưng bang người dùng sinh sống**

****

Hình 5: Biểu đồ cột giữa đặc trưng STATE và Risk Flag

* ***Phân phối theo tiểu bang:***

Uttar Pradesh, Maharashtra và Andhra Pradesh có số lượng cá thể cao nhất trong tập dữ liệu, cho thấy các bang này có số lượng đại diện lớn hơn trong dữ liệu.

Các bang nhỏ hơn như Sikkim và Chandigarh có số lượng thấp nhất, phản ánh dân số nhỏ hơn hoặc số lượng đại diện ít hơn trong bộ dữ liệu này.

* ***Phân phối cờ rủi ro:***

Ở hầu hết mọi tiểu bang, số người không có rủi ro (Cờ rủi ro = 0) đông hơn đáng kể số người có rủi ro (Cờ rủi ro = 1). Điều này cho thấy xu hướng chung là rủi ro thấp ở hầu hết các bang.

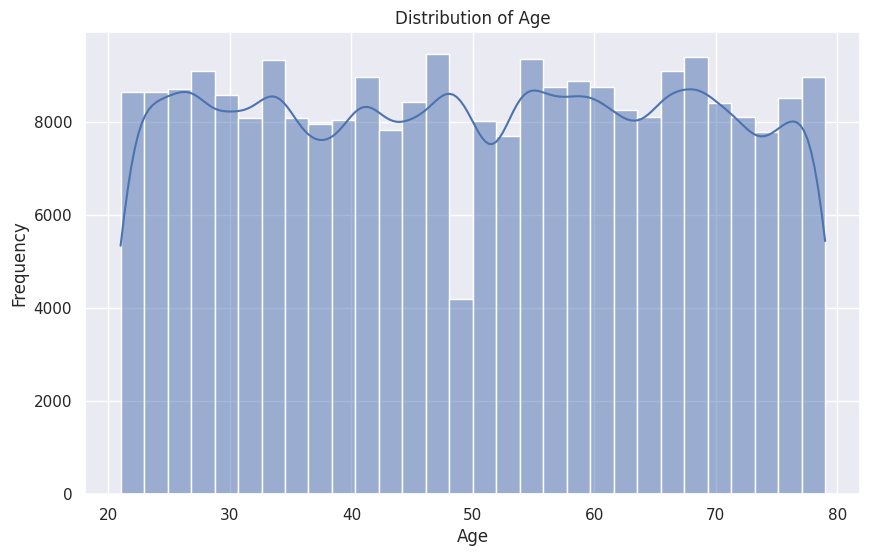
Các bang có số lượng cá nhân gắn cờ rủi ro cao nhất bao gồm Uttar Pradesh, Maharashtra, Andhra Pradesh, Tây Bengal và Bihar. Điều này có thể chỉ ra các yếu tố khu vực cụ thể góp phần tạo nên mức độ rủi ro cao hơn ở những khu vực này.

* ***Phân tích so sánh:***

Tỷ lệ cờ rủi ro thay đổi tùy theo tiểu bang. Ví dụ, Bihar và Odisha có tỷ lệ cá nhân bị gắn cờ rủi ro tương đối cao hơn so với các bang như Kerala và Delhi. Điều này có thể là do các điều kiện kinh tế xã hội khác nhau, cơ hội việc làm hoặc các chính sách khu vực ảnh hưởng đến sự ổn định tài chính.

Các bang như Kerala, Punjab và Haryana có tỷ lệ rủi ro tương đối thấp hơn, cho thấy điều kiện kinh tế ổn định hơn hoặc các yếu tố giảm thiểu rủi ro tốt hơn.

* + 1. ***Đặc trưng về tuổi tác***



Hình 6: Biểu đồ tần suất về tuổi tác (Age)

* ***Phân bố theo độ tuổi:***

Sự phân bổ độ tuổi dường như tương đối đồng đều giữa các nhóm tuổi khác nhau, cho thấy sự thể hiện độ tuổi đa dạng trong tập dữ liệu.

Có một khoảng cách đáng chú ý ở độ tuổi 50, cho thấy có thể có sự bất thường về dữ liệu hoặc sự thiếu hiểu biết về các cá nhân trong nhóm tuổi này.

* ***Thời kỳ đỉnh cao:***

Một số nhóm tuổi có tần suất cao hơn một chút, biểu thị các nhóm cá nhân ở những độ tuổi này. Ví dụ: độ tuổi khoảng 30, 40 và 60 dường như có số lượng cao hơn.

* ***KDE (Ước tính mật độ hạt nhân):***

Dòng KDE cung cấp ước tính mịn về phân bổ độ tuổi. Nó khẳng định sự phân bố tương đối đồng đều về độ tuổi với một số biến động nhỏ.

Đường KDE cũng làm nổi bật mức giảm ở độ tuổi 50, phù hợp với khoảng trống quan sát được trong các thanh biểu đồ.

* ***Quan sát chung:***

Sự phân bổ tổng thể cho thấy rằng tập dữ liệu không nghiêng nhiều về một nhóm tuổi cụ thể, khiến nó trở thành một đại diện tốt để phân tích xu hướng ở các độ tuổi khác nhau.

Các khoảng cách ở độ tuổi 50 có thể cần điều tra thêm để hiểu liệu đó có phải là do vấn đề thu thập dữ liệu hay các yếu tố cơ bản khác hay không.

* + 1. **Đặc trưng về thu nhập**

A graph of a distribution of income

Description automatically generated

Hình 7: Biểu đồ tần suất về thu nhập (Income)

* ***Phân phối thu nhập:***

Sự phân bổ thu nhập dường như khá đồng đều giữa các mức thu nhập khác nhau, cho thấy rằng tập dữ liệu bao gồm nhiều mức thu nhập khác nhau.

Không có đỉnh hoặc đáy đáng kể trong phân bổ, cho thấy dữ liệu thu nhập được trải đều và không bị lệch nhiều về bất kỳ khung thu nhập cụ thể nào.

* ***KDE (Ước tính mật độ hạt nhân):***

Đường KDE cung cấp ước tính mịn về phân phối thu nhập. Nó xác nhận sự phân phối tương đối đồng đều với những biến động nhỏ.

Đường KDE cũng nêu bật sự gia tăng nhẹ về mật độ ở cả đầu dưới và đầu cao hơn của phổ thu nhập, điều này cho thấy sự tập trung nhỏ của các cá nhân có thu nhập rất thấp hoặc rất cao.

* ***Quan sát chung:***

Sự phân bổ tổng thể cho thấy rằng tập dữ liệu đại diện cho một nhóm cá nhân đa dạng với mức thu nhập khác nhau, khiến nó phù hợp cho các phân tích yêu cầu nhiều dữ liệu thu nhập.

Sự phân bố đồng đều cũng chỉ ra rằng tập dữ liệu không có độ lệch thu nhập đáng kể, điều này có lợi cho việc phân tích và lập mô hình khách quan.

* 1. **Phân tích mối tương quan giữa các biến**

A screenshot of a computer

Description automatically generated

Hình 8: Biểu đồ tương quan giữa các biến

* ***Quan sát chung:***

Hầu hết các biến thể hiện mối tương quan rất thấp hoặc không có mối tương quan với nhau, được biểu thị bằng các giá trị gần bằng 0. Điều này cho thấy các đặc điểm tương đối độc lập với nhau.

* ***Tương quan mạnh mẽ:***

Mối tương quan mạnh nhất được quan sát là giữa Kinh nghiệm và CURRENT\_JOB\_YRS (0,65). Điều này có ý nghĩa vì số năm kinh nghiệm của một cá nhân có thể liên quan chặt chẽ đến số năm họ đã làm trong công việc hiện tại.

Một mối tương quan tương đối cao khác được thấy giữa CURRENT\_JOB\_YRS và Tuổi (0,20), điều này được mong đợi vì những người lớn tuổi có thể có nhiều năm kinh nghiệm làm việc hơn.

* ***Tương quan Risk Flag:***

Risk\_Flag cho thấy mối tương quan rất thấp với tất cả các tính năng khác, trong đó cao nhất là mối tương quan yếu 0,03 với Id. Điều này cho thấy rằng các đặc điểm trong tập dữ liệu không ảnh hưởng mạnh đến cờ rủi ro hoặc các tương tác phức tạp hơn có thể ảnh hưởng đến rủi ro không được nắm bắt bằng tuyến tính đơn giản. các mối tương quan.

* ***Không có mối tương quan tiêu cực đáng chú ý:***

Không có mối tương quan tiêu cực đáng kể nào được quan sát trong tập dữ liệu, chỉ ra rằng các tính năng không có mối quan hệ nghịch đảo với nhau một cách có ý nghĩa.

* ***Độc lập của các biến phân loại:***

Các biến phân loại như Married/Single, House\_Ownership, Car\_Ownership, Profession, CITY, and STATE cho thấy mối tương quan rất thấp với các biến khác, cho thấy tính độc lập của chúng.

* 1. **Xử lý ngoại lai**
     1. **Xứ lý ngoại lại của đặc trưng Tuổi**

**A blue rectangular object with gray lines

Description automatically generated**

Hình 9: Biểu đồ hộp về đặc trưng tuổi (Age)

Không có điểm dữ liệu nào được biểu diễn như ngoại lệ trên biểu đồ, điều này cho thấy không có giá trị tuổi nào quá cao hoặc quá thấp so với phần còn lại của nhóm dữ liệu.

* + 1. **Xứ lý ngoại lại của đặc trưng Income**

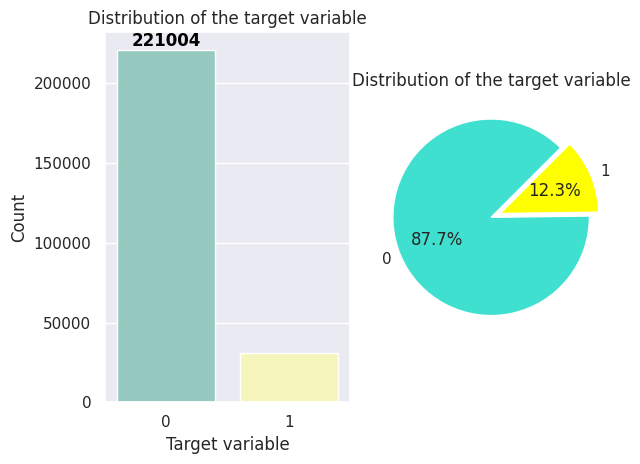
**A graph showing a blue rectangle

Description automatically generated**

Hình 10: Biểu đồ hộp của đặc trưng thu nhập (Income)

Không có điểm ngoại lai nào được hiển thị, dữ liệu có thể rất đồng đều và không có giá trị nào quá cao hoặc quá thấp so với phần còn lại của tập dữ liệu

* 1. **Trực quan cung cấp cái nhìn rõ ràng** **về sự phân bổ của biến mục tiêu, Risk\_Flag**

****

Hình 11: Biểu đồ cột và biểu đồ tròn về sự phân bổ của biến mục tiêu Risk\_Flag

* ***Tóm tắt về trực quan hóa dữ liệu***
* Lớp 0 chiếm 88% tập dữ liệu, trong khi lớp 1 chỉ chiếm 12%. Các lớp bị sai lệch rất nhiều, chúng ta cần giải quyết vấn đề này
* Không có ngoại lệ trong tập dữ liệu. Nhưng chúng ta cần chuẩn hóa Age và Income
* Mối tương quan chặt chẽ giữa Experience và CURRENT\_JOB\_YRS. Có thể bỏ một cột trong quá trình lựa chọn tính năng hoặc sử dụng Phân tích thành phần chính (PCA)
* Married/Single House\_Ownership Car\_Ownership có thể được mã hóa nhị phân hoặc dùng mã hóa one-hot
* Có thể tìm mối liên hệ giữa biến mục tiêu và biến phân loại bằng phép kiểm Chi bình phương

# **CHƯƠNG 3: PHÂN TÍCH THỐNG KÊ VÀ KIỂM ĐỊNH GIẢ THUYẾT**

* 1. **Phương pháp thống kê sử dụng**

***Phương pháp chi bình phương (Chi-Square)***

Kiểm định chi bình phương (còn gọi là kiểm định chi bình phương hoặc kiểm định χ 2 ) là một kiểm định giả thuyết thống kê được sử dụng trong phân tích các bảng dự phòng khi cỡ mẫu lớn. Nói một cách đơn giản hơn, thử nghiệm này chủ yếu được sử dụng để kiểm tra xem hai biến phân loại (hai chiều của bảng dự phòng) có độc lập trong việc ảnh hưởng đến thống kê thử nghiệm hay không ( các giá trị trong bảng ).

***Công thức:***

***Trong đó:***

* O là tần số quan sát, tần số thực sự thu được từ mẫu ngẫu nhiên
* E là tần số kỳ vọng, tần số dự đoán khi giả định hai biến độc lập nhau
* *Có hai giả thuyết kiểm định Chi-Square:*
* H0: hai biến thống kê độc lập với nhau
* H1: hai biến thống kê phụ thuộc với nhau
  1. **Phân tích và kiểm định giả thuyết**

***Giả thuyết kiểm định khả năng vỡ nợ theo sở hữu xe***

Mục đích: Kiểm định sự khác biệt về khả năng vỡ nợ giữa những người sở hữu xe và những người không sở hữu xe, giúp hiểu rõ hơn về ảnh hưởng của việc có xe đến khả năng vỡ nợ.

Nhận thức vấn đề: Nhận định ban đầu về giả thuyết là người không sở hữu xe có rủi ro vỡ nợ cao hơn. Chi-Square test (kiểm định chi bình phương) được áp dụng để kiểm tra giả thuyết này, có hay không sự khác biệt về khả năng vỡ nợ giữa hai nhóm sở hữu xe và không sở hữu xe.

+ Giả thuyết Không (H0): Phản ánh giả định rằng không có sự khác biệt về khả năng vỡ nợ giữa hai nhóm, hoặc khả năng vỡ nợ của nhóm không sở hữu xe không cao hơn nhóm sở hữu xe (có khoảng tin cậy là 95% và mức ý nghĩa là 5%).

+ Giả thuyết Nghịch (H1): Có sự khác biệt đáng kể về khả năng vỡ nợ giữa hai nhóm người.

Giả định rằng các phương sai của hai nhóm là không đồng nhất (điều này thường phù hợp hơn khi làm việc với dữ liệu thực tế) để thực hiện việc kiểm định dựa trên các giá trị của hai nhóm mẫu. Ta sẽ đi sâu vào phân tích thống kê đã được tiến hành để so sánh khả năng vỡ nợ giữa người sở hữu xe và không sở hữu xe, sử dụng kiểm định chi bình phương:

A screenshot of a computer

Description automatically generated

Hình 12: Kết quả kiểm định khả năng vỡ nợ theo sở hữu xe

Nhận xét: Kết quả của kiểm định Chi-Square cho thấy sự chênh lệch đáng kể, với p-value < mức ý nghĩ, bác bỏ H0 cho thấy sự khác biệt về sở hữu xe giữa hai nhóm là có ý nghĩa thống kê.

***Giả thuyết kiểm định khả năng vỡ nợ theo trình trạng hôn nhân***

Mục đích: Kiểm định sự khác biệt về khả năng vỡ nợ giữa những người độc thân và đã kết hôn, giúp hiểu rõ hơn về ảnh hưởng của tình trạng hôn nhân đến khả năng vỡ nợ.

Nhận thức vấn đề: Nhận định ban đầu về giả thuyết là người độc thân có rủi ro vỡ nợ cao hơn. Chi-Square test (kiểm định chi bình phương) được áp dụng để kiểm tra giả thuyết này, có hay không sự khác biệt về khả năng vỡ nợ giữa hai nhóm độc thân và đã kết hôn.

+ Giả thuyết Không (H0): Phản ánh giả định rằng không có sự khác biệt về khả năng vỡ nợ giữa hai nhóm, hoặc khả năng vỡ nợ của nhóm độc thân không cao hơn nhóm đã kết hôn (có khoảng tin cậy là 95% và mức ý nghĩa là 5%).

+ Giả thuyết Nghịch (H1): Có sự khác biệt đáng kể về khả năng vỡ nợ giữa hai nhóm người.

Giả định rằng các phương sai của hai nhóm là không đồng nhất (điều này thường phù hợp hơn khi làm việc với dữ liệu thực tế) để thực hiện việc kiểm định dựa trên các giá trị của hai nhóm mẫu. Ta sẽ đi sâu vào phân tích thống kê đã được tiến hành để so sánh khả năng vỡ nợ giữa người độc thân và đã kết hôn, sử dụng kiểm định chi bình phương:

A screenshot of a computer

Description automatically generated

Hình 13: Kết quả kiểm định khả năng vỡ nợ theo tình trạng hôn nhân

Nhận xét: Kết quả của kiểm định Chi-Square cho thấy sự chênh lệch đáng kể, với p-value < mức ý nghĩ, bác bỏ H0 cho thấy sự khác biệt về tình trạng hôn nhân giữa hai nhóm là có ý nghĩa thống kê.

***Giả thuyết kiểm định khả năng vỡ nợ theo sở hữu nhà***

Mục đích: Kiểm định sự khác biệt về khả năng vỡ nợ giữa những người sở hữu nhà, người thuê nhà,người không có nhà và không thuê nhà, giúp hiểu rõ hơn về ảnh hưởng của việc sở hữu nhà đến khả năng vỡ nợ.

Nhận thức vấn đề: Nhận định ban đầu về giả thuyết là người thuê nhà có rủi ro vỡ nợ cao hơn. Chi-Square test (kiểm định chi bình phương) được áp dụng để kiểm tra giả thuyết này, có hay không sự khác biệt về khả năng vỡ nợ giữa các nhóm người thuê nhà, người sở hữu nhà và người không thuê không sở hữu.

+ Giả thuyết Không (H0): Phản ánh giả định rằng không có sự khác biệt về khả năng vỡ nợ giữa các nhóm, hoặc khả năng vỡ nợ của nhóm thuê nhà không cao hơn nhóm còn lại (có khoảng tin cậy là 95% và mức ý nghĩa là 5%).

+ Giả thuyết Nghịch (H1): Có sự khác biệt đáng kể về khả năng vỡ nợ giữa các nhóm người.

Giả định rằng các phương sai của hai nhóm là không đồng nhất (điều này thường phù hợp hơn khi làm việc với dữ liệu thực tế) để thực hiện việc kiểm định dựa trên các giá trị của hai nhóm mẫu. Ta sẽ đi sâu vào phân tích thống kê đã được tiến hành để so sánh khả năng vỡ nợ giữa những người sở hữu nhà, người thuê nhà,người không có nhà và không thuê nhà, sử dụng kiểm định chi bình phương:

A screenshot of a computer

Description automatically generated

Hình 14: Kết quả kiểm định khả năng vỡ nợ theo sở hữu nhà

Nhận xét: Kết quả của kiểm định Chi-Square cho thấy sự chênh lệch đáng kể, với p-value < mức ý nghĩ, bác bỏ H0 cho thấy sự khác biệt về việc sở hữu nhà giữa các nhóm là có ý nghĩa thống kê.

* 1. **Đánh giá mức độ tin cậy của kết quả**

Việc kiểm định thống kê đã cung cấp bằng chứng rõ ràng về sự khác biệt trong mối quan hệ giữa các nhóm người:

Nhóm người theo sở hữu nhà: có vẻ như những người không thuê nhà hoặc không sở hữu và thuê có tỷ lệ vỡ nợ cao hơn so với những người sở hữu nhà. Điều này có thể phản ánh khả năng tài chính ổn định hơn của những người sở hữu nhà

Nhóm người theo sở hữu xe: có thể thấy rằng người không sở hữu xe có tỷ lệ vỡ nợ cao hơn người sở hữu xe.

Nhóm người theo tình trạng hôn nhân: số liệu cho thấy có sự khác biệt rõ khi người độc thân có tỷ lệ vỡ nợ cao hơn người đã kết hôn.

* Các kết quả có độ tin cậy cao. Giả thuyết kiểm định cho thấy rằng các yếu tố như sở hữu nhà, tình trạng hôn nhân, và sở hữu xe có thể có ảnh hưởng đáng kể đến khả năng vỡ nợ của một cá nhân. Việc phân tích sâu hơn các yếu tố này có thể giúp các tổ chức tài chính đưa ra quyết định vay vốn một cách chính xác hơn và giảm thiểu rủi ro vỡ nợ.

# **CHƯƠNG 4: MÔ HÌNH HÓA VÀ DỰ ĐOÁN**

* 1. **Chuẩn hóa và chuẩn bị dữ liệu huấn luyện**
     1. **Phân tích thành phần chính (PCA) cho CURRENT\_JOB\_YRS và Experience**

PCA (Principal Component Analysis) là một kỹ thuật giảm số chiều phổ biến trong học máy và phân tích dữ liệu. PCA được sử dụng để giảm số lượng các biến số trong dữ liệu xuống thành một tập hợp các biến số nhỏ hơn mà vẫn giữ lại được phần lớn thông tin của dữ liệu gốc. Mục tiêu chính của PCA là giảm số chiều dữ liệu, loại bỏ nhiễu và giúp việc trực quan hóa dữ liệu dễ dàng hơn.

|  |
| --- |
| features = ["CURRENT\_JOB\_YRS","Experience"]  df\_for\_pca = df[features]  scaled\_df\_for\_pca = (df\_for\_pca - df\_for\_pca.mean(axis=0))/df\_for\_pca.std()  scaled\_df\_for\_pca |

A screenshot of a calculator

Description automatically generated

Hình 15: DataFrame đã chuẩn hóa cho PCA

Ảnh có chứa văn bản, Phông chữ, ảnh chụp màn hình, màu trắng

Mô tả được tạo tự động

Hình 16: DataFrame sau khi áp dụng PCA

|  |
| --- |
| df1 = pd.concat([df,df\_pca],axis=1)  df1.head() |

Ảnh có chứa văn bản, Phông chữ, hàng, số

Mô tả được tạo tự động

Hình 17: Nối DataFrame PCA với DataFrame ban đầu

* + 1. **Chuẩn bị dữ liệu**
       1. **Mã hóa nhãn cho các biến phân loại**

Các mô hình học máy thường yêu cầu đầu vào là các giá trị số và không thể làm việc trực tiếp với các giá trị phân loại dạng chữ hoặc ký hiệu, do đó ta cần đưa về dạng số.

|  |
| --- |
| features = ['Married/Single','Car\_Ownership','Profession','CITY','STATE', 'House\_Ownership']  label\_encoder = LabelEncoder()  for col in features:      df1[col] = label\_encoder.fit\_transform(df1[col]) |

* + - 1. **Loại bỏ cột dư thừa**

|  |
| --- |
| df2 = df1.drop(["Id", "CURRENT\_JOB\_YRS", "Experience"],axis=1)  df2 |

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động

Hình 18: DataFrame sau khi mã hóa các biến phân loại và loại bỏ cột dư thừa

* + 1. **Chia dữ liệu**
       1. **Loại bỏ thuộc tính đích và chia tập dữ liệu**

|  |
| --- |
| X = df2.drop(['Risk\_Flag'],axis=1)  y = df2.Risk\_Flag  X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size = 0.3, random\_state = 42, stratify=y) |

* + - 1. **Xử lý mất cân bằng dữ liệu**

Khi dữ liệu huấn luyện bị mất cân bằng, tức là số lượng các mẫu trong các lớp không đều nhau, mô hình học máy có thể gặp nhiều vấn đề dẫn đến hiệu suất kém. Kỹ thuật SMOTE (Synthetic Minority Over-sampling Technique) là một phương pháp tạo ra các mẫu tổng hợp mới từ các mẫu hiện có của lớp thiểu số, giúp nâng cao hiệu suất và độ tin cậy của mô hình trong các ứng dụng thực tế.

|  |
| --- |
| sm = SMOTE(random\_state = 1000)  X\_res, y\_res = sm.fit\_resample(X\_train, y\_train) |

* 1. **Mô hình logictis regresion**
     1. **Sơ lược về mô hình**

Logistic Regression là thuật toán học máy được sử dụng cho bài toán phân loại nhị phân (chỉ có 2 lớp). Khác với hồi quy tuyến tính, Logistic Regression xây dựng một hàm sigmoid để ước tính xác suất của một mẫu dữ liệu thuộc về lớp nào. Thuật toán này có ưu điểm là dễ hiểu, dễ triển khai và hiệu quả với dữ liệu tuyến tính. Tuy nhiên, Logistic Regression chỉ xử lý tốt dữ liệu tuyến tính và cần chuẩn hóa dữ liệu đầu vào trước khi sử dụng.

* + 1. **Đào tạo và đánh giá mô hình**

|  |
| --- |
| from sklearn.linear\_model import LogisticRegression  from sklearn.preprocessing import StandardScaler  # Chuẩn hóa dữ liệu  scaler = StandardScaler()  X\_train\_scaled = scaler.fit\_transform(X\_res)  X\_test\_scaled = scaler.transform(X\_test)  model = LogisticRegression(max\_iter = 500000)  model.fit(X\_train\_scaled, y\_res)  y\_pred = model.predict(X\_test\_scaled)  accuracy\_logis = model.score(X\_test\_scaled, y\_test)  accuracy\_logis  print(classification\_report(y\_test,y\_pred)) |

Ảnh có chứa văn bản, ảnh chụp màn hình, biên lai, Phông chữ

Mô tả được tạo tự động

Hình 19: Báo cáo các độ đo của Logistic Regression

* + 1. **Đánh giá kết quả và giải thích kết quả khai phá**

|  |
| --- |
| from sklearn.metrics import ConfusionMatrixDisplay  import matplotlib.pyplot as plt  %matplotlib inline  con\_max = ConfusionMatrixDisplay.from\_predictions(y\_test, y\_pred)  plt.show() |

Ảnh có chứa văn bản, ảnh chụp màn hình, hàng, biểu đồ

Mô tả được tạo tự động

Hình 20: Confusion matrix cho dự đoán của Logistic Regression

* **Lớp 0 (Không vỡ nợ)**:
  + **Precision (0.89)**: Mô hình có độ chính xác cao khi dự đoán khách hàng không vỡ nợ, với 89% dự đoán là chính xác.
  + **Recall (0.51)**: Tuy nhiên, chỉ có 51% khách hàng thực sự không vỡ nợ được mô hình nhận diện chính xác.
  + **F1-score (0.65)**: Điểm F1 cho thấy mô hình có hiệu suất tốt nhưng vẫn còn không gian để cải thiện, đặc biệt là về mặt recall.
* **Lớp 1 (Vỡ nợ)**:
  + **Precision thấp (0.13)**: Chỉ có 13% trong số những dự đoán vỡ nợ của mô hình là chính xác, cho thấy nhiều báo động giả.
  + **Recall (0.55)**: Mô hình có khả năng nhận diện 55% khách hàng thực sự vỡ nợ.
  + **F1-score (0.22)**: Điểm F1 thấp cho thấy mô hình không hiệu quả trong việc dự đoán lớp vỡ nợ, chủ yếu do precision rất thấp.
* **Accuracy tổng thể (0.51)**: Mô hình chính xác 51% trên toàn bộ dữ liệu kiểm tra.

***Nhận xét chung:***

* Mô hình này có vẻ tốt hơn trong việc nhận diện khách hàng không vỡ nợ so với việc nhận diện khách hàng vỡ nợ.
* Precision thấp cho lớp vỡ nợ có thể dẫn đến việc từ chối tín dụng cho nhiều khách hàng tiềm năng không vỡ nợ.
* Recall tương đối thấp cho cả hai lớp cho thấy mô hình có thể bỏ qua một số trường hợp thực sự vỡ nợ hoặc không vỡ nợ.
* **Ưu điểm:**
* Hồi quy logistic dễ thực hiện, diễn giải hơn và đào tạo rất hiệu quả.
* Hồi quy logistic không đưa ra giả định nào về sự phân bố của các lớp trong không gian đặc trưng.
* Có thể dễ dàng mở rộng sang nhiều lớp (hồi quy đa thức) và chế độ xem xác suất tự nhiên của các dự đoán lớp.
* Không chỉ cung cấp thước đo mức độ phù hợp của một yếu tố dự đoán (kích thước hệ số) mà còn cả hướng liên kết của nó (tích cực hoặc tiêu cực).
* Rất nhanh trong việc phân loại các bản ghi chưa biết.
* Độ chính xác tốt cho nhiều tập dữ liệu đơn giản và nó hoạt động tốt khi tập dữ liệu có thể phân tách tuyến tính.
* Có thể giải thích các hệ số của mô hình như là các chỉ số về tầm quan trọng của đặc điểm.
* Hồi quy logistic ít có xu hướng khớp quá mức nhưng nó có thể khớp quá mức trong các tập dữ liệu nhiều chiều. Người ta có thể xem xét các kỹ thuật Chính quy hóa (L1 và L2) để tránh overfitting trong các tình huống này.
* **Nhược điểm:**
* Nếu số lượng quan sát nhỏ hơn số lượng tính năng thì không nên sử dụng Hồi quy logistic, nếu không, có thể dẫn đến tình trạng overfitting.
* Hạn chế chính của hồi quy logistic là giả định về tính tuyến tính giữa biến phụ thuộc và biến độc lập.
* Các vấn đề phi tuyến tính không thể được giải quyết bằng hồi quy logistic vì nó có bề mặt quyết định tuyến tính. Dữ liệu có thể phân tách tuyến tính hiếm khi được tìm thấy trong các tình huống thực tế.
* Hồi quy logistic yêu cầu mức độ đa cộng tuyến trung bình hoặc không có giữa các biến độc lập.
* Thật khó để có được các mối quan hệ phức tạp bằng cách sử dụng hồi quy logistic. Các thuật toán mạnh mẽ và nhỏ gọn hơn như Mạng thần kinh có thể dễ dàng vượt trội hơn thuật toán này.
* Trong hồi quy tuyến tính, các biến độc lập và phụ thuộc có liên quan tuyến tính. Nhưng Hồi quy logistic cần các biến độc lập có liên quan tuyến tính với tỷ lệ logarit (log(p/(1-p)).
* **Ý nghĩa của khai phá:**
  + Độ chính xác tổng thể thấp: Mô hình Logistic đạt độ chính xác tổng thể 51%, cho thấy mô hình phân loại đúng khá ít các trường hợp.
  + Mô hình này có vẻ tốt hơn trong việc nhận diện khách hàng không vỡ nợ so với việc nhận diện khách hàng vỡ nợ.
  + Precision thấp cho lớp vỡ nợ có thể dẫn đến việc từ chối tín dụng cho nhiều khách hàng tiềm năng không vỡ nợ.
  + Recall tương đối thấp cho cả hai lớp cho thấy mô hình có thể bỏ qua một số trường hợp thực sự vỡ nợ hoặc không vỡ nợ.
  1. **Mô hình KNN**
     1. **Sơ lược về mô hình**

KNN (viết tắt của K-Nearest Neighbors, hay thuật toán k láng giềng gần nhất) là thuật toán học máy đơn giản và linh hoạt, có thể được sử dụng cho cả bài toán phân loại và hồi quy. KNN hoạt động bằng cách dự đoán dựa trên các điểm gần nhất (k điểm) trong tập huấn luyện, sau đó dự đoán nhãn (phân loại) hoặc giá trị trung bình (hồi quy) cho mẫu dữ liệu mới dựa trên nhãn hoặc giá trị trung bình của k láng giềng gần nhất. Ưu điểm của KNN là dễ hiểu, không cần giả định về phân bố dữ liệu và hiệu quả với nhiều dạng dữ liệu. Tuy nhiên, KNN cũng có một số nhược điểm như: nhạy cảm với nhiễu trong dữ liệu, cần chọn giá trị k phù hợp và tốn thời gian tính toán khi xử lý tập dữ liệu lớn.

***Ưu điểm:***

* Đơn giản và dễ dàng triển khai: KNN có cách hoạt động đơn giản, dễ hiểu và dễ triển khai.
* Độ phức tạp tính toán thấp: Việc dự đoán kết quả cho dữ liệu mới rất đơn giản và không đòi hỏi nhiều tính toán phức tạp.

***Nhược điểm:***

* Phụ thuộc vào giá trị K: Lựa chọn giá trị K quan trọng, và nếu K quá nhỏ, mô hình dễ bị ảnh hưởng bởi nhiễu dữ liệu.
* Tốn thời gian tính toán: Việc tính toán khoảng cách với tất cả các đối tượng trong tập dữ liệu khiến KNN cần nhiều thời gian để thực hiện.
  + 1. **Đào tạo và đánh giá mô hình**

|  |
| --- |
| from sklearn.neighbors import KNeighborsClassifier  param\_grid = {      'n\_neighbors': [3, 5, 7, 9],      'p': [1, 2],      'metric': ['minkowski', 'euclidean']  }  knn = KNeighborsClassifier()  grid\_search = GridSearchCV(estimator=knn, param\_grid=param\_grid, cv=5, verbose=3, scoring='accuracy')  grid\_search.fit(X\_res, y\_res)  best\_params = grid\_search.best\_params\_  best\_score = grid\_search.best\_score\_  best\_params, best\_score  knn\_model = KNeighborsClassifier(n\_neighbors=9,metric='minkowski',p=1)  knn\_model.fit(X\_res, y\_res)  y\_pred = knn\_model.predict(X\_test)  accuracy\_knn = knn\_model.score(X\_test, y\_test)  accuracy\_knn |

* + 1. **Đánh giá kết quả và giải thích kết quả khai phá**

|  |
| --- |
| print(classification\_report(y\_test,y\_pred)) |

Ảnh có chứa văn bản, ảnh chụp màn hình, biên lai, Phông chữ

Mô tả được tạo tự động

Hình 21. Báo cáo các độ đo của mô hình KNN

Nhận xét:

* Độ chính xác tổng thể cao: Mô hình KNN đạt độ chính xác tổng thể 87%, cho thấy mô hình phân loại đúng phần lớn các trường hợp.
* Precision cho lớp 1 thấp: Precision của lớp 1 (các trường hợp có rủi ro) là 0.46, nghĩa là chỉ 46% trong số các dự đoán dương tính là đúng. Điều này cho thấy mô hình có nhiều dự đoán dương tính sai.
* Recall cho lớp 1 tốt hơn: Recall của lớp 1 là 0.60, nghĩa là mô hình tìm được 60% các trường hợp có rủi ro thực sự. Điều này tốt hơn so với mô hình AdaBoost trước đó.
* F1-Score của lớp 1: F1-Score của lớp 1 là 0.52, cho thấy mô hình KNN có sự cân bằng giữa Precision và Recall tốt hơn so với AdaBoost.
* Macro và Weighted Averages: Các chỉ số trung bình (macro và weighted) cho thấy hiệu suất tổng thể của mô hình là ổn định và đáng tin cậy, đặc biệt với weighted avg do sự mất cân bằng giữa hai lớp.
  1. **Mô hình Random Forest**
     1. **Sơ lược về mô hình**

Random Forest là một phương pháp ensemble học máy mạnh mẽ dựa trên việc kết hợp nhiều cây quyết định độc lập. Mỗi cây được xây dựng trên một tập con ngẫu nhiên của dữ liệu, và các cây được huấn luyện để giảm thiểu phương sai và ngăn ngừa overfitting. Khi dự đoán, Random Forest lấy trung bình (trong trường hợp hồi quy) hoặc bỏ phiếu đa số (trong trường hợp phân loại) từ các cây riêng lẻ. Điều này giúp Random Forest có khả năng dự đoán mạnh mẽ và ổn định, nhưng mô hình có thể trở nên phức tạp và khó giải thích

***Ưu điểm***

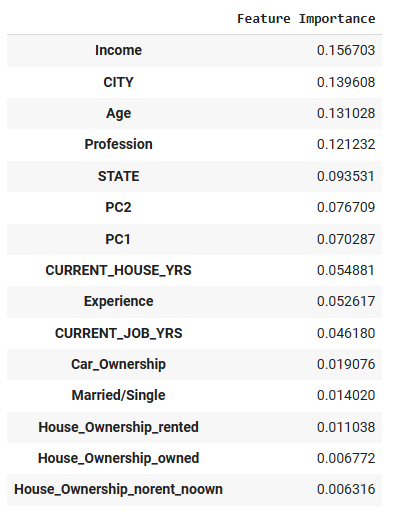
* Độ chính xác cao: Random Forest sử dụng nhiều cây quyết định, mỗi cây được huấn luyện trên một tập dữ liệu con khác nhau. Bằng cách kết hợp các dự đoán của các cây, nó giảm biến thiên và tạo ra dự đoán chính xác hơn.
* Ổn định và đáng tin cậy: Random Forest thường ít bị ảnh hưởng bởi các nhiễu loạn trong dữ liệu và có khả năng tránh overfitting tốt hơn các mô hình đơn giản hơn.

***Nhược điểm***

* Khó hiểu và khó giải thích: Do sự kết hợp của nhiều cây quyết định, việc giải thích kết quả của Random Forest có thể phức tạp.
* Tốn thời gian và tài nguyên: Huấn luyện nhiều cây quyết định đòi hỏi nhiều thời gian và tài nguyên tính toán.
  + 1. **Đào tạo và đánh giá mô hình**

Tạo mô hình RandomForestClassifier và huấn luyện, sau đó xuất ra mức độ quan trọng của thuộc tính

|  |
| --- |
| from sklearn.ensemble import RandomForestClassifier  gr\_rf\_model = RandomForestClassifier(class\_weight="balanced", random\_state=101, max\_depth=12, n\_estimators= 200, min\_samples\_split = 3)  gr\_rf\_model.fit(X\_train,y\_train)  rf\_feature\_imp = pd.DataFrame(index = X.columns, data = gr\_rf\_model.feature\_importances\_,                                columns = ["Feature Importance"]).sort\_values("Feature Importance", ascending = False)  rf\_feature\_imp |



Hình 22. Mức độ quan trọng của các thuộc tính trong RandomForest

|  |
| --- |
| ax = sns.barplot(x=rf\_feature\_imp["Feature Importance"], y=rf\_feature\_imp.index)  plt.ylabel("Feature")  plt.title("Feature Importance for Random Forest")  plt.show() |

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, biểu đồ

Mô tả được tạo tự động

Hình 23. Biểu đồ thể hiện mức độ quan trọng của các thuộc tính trong RandomForest

Từ biểu đồ trên, chọn các thuộc tính có mức quan trọng cao và thực hiện huấn luyện lại

|  |
| --- |
| X\_new = X[["Income", "CITY", "Age", "Profession", "PC1", "PC2", "STATE"]]  X\_train\_new, X\_test\_new, y\_train\_new, y\_test\_new = train\_test\_split(X\_new, y, test\_size = 0.3, stratify=y, random\_state = 101)  rf\_model = RandomForestClassifier(class\_weight="balanced", random\_state=101)  rf\_model.fit(X\_train\_new,y\_train\_new)  y\_pred\_new = rf\_model.predict(X\_test\_new)  accuracy\_rf = rf\_model.score(X\_test\_new, y\_test\_new)  accuracy\_rf |

* + 1. **Đánh giá kết quả và giải thích kết quả khai phá**

|  |
| --- |
| print(classification\_report(y\_test,y\_pred)) |

Ảnh có chứa văn bản, ảnh chụp màn hình, biên lai, Phông chữ

Mô tả được tạo tự động

Hình 24. Báo cáo về các độ đo của mô hình Random Forest

|  |
| --- |
| plt.figure(figsize=(10,6))  y\_pred\_new = rf\_model.predict(X\_test\_new)  cf\_matrix = confusion\_matrix(y\_test\_new, y\_pred\_new)  group\_names = ['True Negatives (TN)', 'False Positives (FP)\n(Type I Error)',                 'False Negatives (FN)\n(Type II Error)', 'True Positives (TP)']  group\_counts = ['{0:0.0f}'.format(value) for value in cf\_matrix.flatten()]  group\_percentages = ['{0:.2%}'.format(value) for value in cf\_matrix.flatten()/np.sum(cf\_matrix)]  labels = [f'{v1}\n{v2}\n{v3}' for v1, v2, v3 in zip(group\_names, group\_counts, group\_percentages)]  labels = np.asarray(labels).reshape(2, 2)  ax = sns.heatmap(cf\_matrix, annot=labels, fmt="", annot\_kws={'size': 13}, cmap='viridis')  ax.set(xlabel='Predicted Class', ylabel = 'Actual Class') |

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Mô tả được tạo tự động

Hình 25. Confusion matrix cho dự đoán của mô hình Random Forest

Nhận xét

**- True Negatives (TN):**

- Số lượng: 60606

- Tỷ lệ: 80.17%

- Đây là số lượng mẫu thuộc lớp 0 (âm tính) mà mô hình dự đoán chính xác là âm tính. Con số này khá lớn, cho thấy mô hình có khả năng phát hiện các trường hợp âm tính khá tốt.

**- False Positives (FP) (Type I Error):**

- Số lượng: 5695

- Tỷ lệ: 7.53%

- Đây là số lượng mẫu thuộc lớp 0 nhưng mô hình lại dự đoán là lớp 1 (dương tính). Tỷ lệ này không quá cao, nhưng vẫn cần được cải thiện để giảm thiểu các dự đoán sai lệch.

**- False Negatives (FN) (Type II Error):**

- Số lượng: 2177

- Tỷ lệ: 2.89%

- Đây là số lượng mẫu thuộc lớp 1 (dương tính) nhưng mô hình lại dự đoán là lớp 0. Tỷ lệ này khá thấp, cho thấy mô hình có khả năng phát hiện các trường hợp dương tính tốt, nhưng vẫn còn một số trường hợp bị bỏ sót.

**- True Positives (TP):**

- Số lượng: 7122

- Tỷ lệ: 9.41%

- Đây là số lượng mẫu thuộc lớp 1 mà mô hình dự đoán chính xác là dương tính. Tỷ lệ này thể hiện khả năng của mô hình trong việc nhận diện các trường hợp dương tính.

**Tổng quan:**

- Mô hình Random Forest này có độ chính xác tương đối cao trong việc nhận diện các trường hợp âm tính (TN) và dương tính (TP).

- Số lượng FP và FN không quá cao, nhưng vẫn cần được chú ý để cải thiện hơn nữa.

- Để cải thiện mô hình, nên xem xét việc cân bằng lại dữ liệu (data balancing).

* 1. **Mô hình AdaBoosting**
     1. **Sơ lược về mô hình**

AdaBoost (Adaptive Boosting) là một thuật toán boosting phổ biến trong học máy, hoạt động bằng cách kết hợp nhiều mô hình yếu (weak learners) để tạo ra một mô hình mạnh (strong learner). AdaBoost huấn luyện từng mô hình yếu tuần tự, và mỗi mô hình tập trung vào những mẫu mà các mô hình trước đó phân loại sai. Trọng số của các mẫu khó phân loại sẽ tăng lên, giúp mô hình cuối cùng có hiệu suất cao hơn. AdaBoost thường sử dụng cây quyết định đơn giản như các mô hình yếu và nổi bật ở khả năng tăng độ chính xác mà không dễ bị overfitting. Tuy nhiên, AdaBoosting cũng có thể overfitting nếu không chọn tham số phù hợp.

***Ưu điểm***

* Tăng cường các mẫu khó phân loại: AdaBoosting tập trung vào các mẫu khó phân loại bằng cách tăng trọng số của chúng trong các lần huấn luyện kế tiếp, giúp cải thiện hiệu suất mô hình trên các mẫu này.
* Khả năng kết hợp nhiều mô hình đơn giản: AdaBoosting kết hợp nhiều mô hình yếu (weak learners) để tạo ra một mô hình mạnh hơn, giúp nâng cao độ chính xác và khả năng tổng quát hóa.

***Nhược điểm***

* Nhạy cảm với nhiễu và ngoại lệ: AdaBoost có thể bị ảnh hưởng bởi nhiễu và ngoại lệ trong dữ liệu, dẫn đến việc phân loại sai.
* Dễ bị overfitting: Trong một số trường hợp, AdaBoost có thể dẫn đến overfitting trên tập dữ liệu huấn luyện.
  + 1. **Đào tạo và đánh giá mô hình**

|  |
| --- |
| from sklearn.ensemble import AdaBoostClassifier  param\_grid = {      'n\_estimators': [50, 100, 150],      'learning\_rate': [0.01, 0.1, 1]  }  grid\_search = GridSearchCV(estimator=AdaBoostClassifier(random\_state=15000), param\_grid=param\_grid, cv=5, scoring='accuracy', verbose=2, n\_jobs=-1)  grid\_search.fit(X\_train\_scaled, y\_res)  best\_params = grid\_search.best\_params\_  best\_score = grid\_search.best\_score\_  print(f"Best Parameters: {best\_params}")  print(f"Best Cross-Validation Score: {best\_score}")  best\_model = grid\_search.best\_estimator\_  y\_pred\_best = best\_model.predict(X\_test\_scaled)  accuracy\_ada = accuracy\_score(y\_test, y\_pred\_best)  precision\_best = precision\_score(y\_test, y\_pred\_best)  recall\_best = recall\_score(y\_test, y\_pred\_best)  f1\_best = f1\_score(y\_test, y\_pred\_best)  roc\_auc\_best = roc\_auc\_score(y\_test, y\_pred\_best)  print(f"Test Set Accuracy: {accuracy\_ada}")  print(f"Test Set Precision: {precision\_best}")  print(f"Test Set Recall: {recall\_best}")  print(f"Test Set F1-score: {f1\_best}")  print(f"Test Set ROC-AUC: {roc\_auc\_best}")  print("Confusion Matrix for Best Model:")  print(confusion\_matrix(y\_test, y\_pred\_best))  print("\nClassification Report for Best Model:")  print(classification\_report(y\_test, y\_pred\_best)) |

* + 1. **Đánh giá kết quả và giải thích kết quả khai phá**

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, tài liệu

Mô tả được tạo tự động

Hình 26. Báo cáo các độ đo của AdaBoostClassifier

**Nhận xét:**

**- Độ chính xác (Accuracy):** Mô hình đạt độ chính xác tổng thể là 56%, thấp hơn so với kỳ vọng cho một mô hình dự đoán rủi ro.

**- Precision và Recall (lớp 1):** Precision cho lớp 1 (các trường hợp có rủi ro) là 0.15, nghĩa là mô hình chỉ đúng 15% khi dự đoán là có rủi ro. Recall là 0.53, nghĩa là mô hình tìm được 53% các trường hợp có rủi ro thực sự.

**- F1-Score (lớp 1):** F1-Score cho lớp 1 là 0.23, cho thấy sự cân bằng giữa Precision và Recall không tốt.

**- ROC-AUC:** Điểm ROC-AUC là 0.547, cho thấy mô hình chỉ hơi tốt hơn một chút so với dự đoán ngẫu nhiên.

* 1. **Mô hình GradientBoosting**
     1. **Sơ lược về mô hình**

Gradient Boosting là một phương pháp ensemble khác cũng dựa trên boosting, nhưng sử dụng gradient descent để tối ưu hóa mô hình. Gradient Boosting xây dựng các mô hình yếu tuần tự, mỗi mô hình mới cố gắng sửa lỗi của mô hình trước đó bằng cách giảm gradient của hàm mất mát. Điều này giúp mô hình học tập theo từng bước nhỏ và liên tục cải thiện hiệu suất. Gradient Boosting nổi bật với khả năng xử lý tốt các dữ liệu phi tuyến và độ chính xác cao, nhưng đòi hỏi thời gian huấn luyện lâu và có nguy cơ overfitting nếu không được điều chỉnh cẩn thận.

* + 1. **Đào tạo và đánh giá mô hình**

|  |
| --- |
| from sklearn.model\_selection import GridSearchCV  from sklearn.ensemble import GradientBoostingClassifier  param\_grid = {      'n\_estimators': [50, 100],      'learning\_rate': [0.01, 0.1],      'max\_depth': [3, 5]  }  model = GradientBoostingClassifier()  grid\_search = GridSearchCV(model, param\_grid, cv=3, verbose=2)  grid\_search.fit(X\_train\_scaled, y\_res)  best\_params = grid\_search.best\_params\_  best\_model = GradientBoostingClassifier(\*\*best\_params)  best\_model.fit(X\_train\_scaled, y\_res)  accuracy\_gra = best\_model.score(X\_test\_scaled, y\_test)  accuracy\_gra  y\_pred = best\_model.predict(X\_test)  print(classification\_report(y\_test,y\_pred)) |

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động

Hình 27: Báo cáo các độ đo của mô hình GradientBoostingClassifier

* + 1. **Đánh giá kết quả và giải thích kết quả khai phá**

|  |
| --- |
| con\_max = ConfusionMatrixDisplay.from\_predictions(y\_test, y\_pred)  plt.show() |

Ảnh có chứa văn bản, ảnh chụp màn hình, hàng, biểu đồ

Mô tả được tạo tự động

Hình 28: Confusion matrix cho dự đoán của GradientBoostingClassifier

* **Ưu điểm**
* Có khả năng thực hiện tốt cả bài toán phân loại và hồi quy nên chúng ta có thể coi nó như một thuật toán linh hoạt.
* Các thuật toán tăng cường độ dốc với cây quyết định có thể xử lý các giá trị còn thiếu trong dữ liệu bằng cách sử dụng các phân tách thay thế trong quá trình xây dựng cây.
* Việc tăng cường độ dốc tương đối mạnh mẽ đối với các ngoại lệ, vì ảnh hưởng của các điểm dữ liệu riêng lẻ được giảm thiểu thông qua phương pháp tổng hợp.
* **Nhược điểm**
* Mặc dù Gradient Boosting mạnh mẽ đối với các ngoại lệ, nhưng nó cũng tốn kém về mặt tính toán, đặc biệt là khi xử lý các tập dữ liệu lớn hoặc cây sâu. Việc đào tạo mô hình có thể cần nhiều thời gian và nguồn lực hơn so với các thuật toán đơn giản hơn.
* Trong trường hợp các mô hình phức tạp hoặc tốc độ học cao, việc giảm độ dốc có thể dễ xảy ra vấn đề quá khớp.
* Giảm dần độ dốc thường được coi là một mô hình hộp đen hoặc ít giải thích hơn. Việc hiểu mối quan hệ chính xác giữa các yếu tố dự đoán và kết quả có thể là một thách thức do cấu trúc tổng thể phức tạp.
* **Ý nghĩa khai phá**

Lớp 0 (Không vỡ nợ):

* + Precision cao (0.94): Khi mô hình dự đoán một khách hàng không vỡ nợ, nó chính xác đến 94%.
  + Recall thấp hơn (0.70): Có 70% của tổng số khách hàng thực sự không vỡ nợ được mô hình dự đoán chính xác.
  + F1-score (0.80): Điểm số này cho thấy một sự cân bằng giữa precision và recall, nhưng vẫn có không gian để cải thiện, đặc biệt là về mặt recall.

Lớp 1 (Vỡ nợ):

* + Precision thấp (0.24): Chỉ có 24% trong số những dự đoán vỡ nợ của mô hình là chính xác.
  + Recall cao hơn (0.69): Mô hình có khả năng nhận diện 69% của những trường hợp thực sự vỡ nợ.
  + F1-score thấp (0.36): Điểm số này cho thấy mô hình không hiệu quả lắm trong việc dự đoán lớp vỡ nợ, chủ yếu do precision thấp.

Accuracy tổng thể (0.70): Mô hình chính xác 70% trên toàn bộ dữ liệu kiểm tra.

Nhận xét chung:

* + Mô hình có vẻ tốt trong việc nhận diện khách hàng không vỡ nợ nhưng không hiệu quả trong việc nhận diện khách hàng vỡ nợ.
  + Precision thấp cho lớp 1 có thể dẫn đến việc từ chối tín dụng cho nhiều khách hàng tiềm năng không vỡ nợ.
  1. **Mô hình Naive Bayes**
     1. **Sơ lược về mô hình**

Naive Bayes là một thuật toán học máy phổ biến được sử dụng cho các bài toán phân loại. Thuật toán này dựa trên nguyên tắc của định lý Bayes, và giả định "ngây thơ" (naive) rằng các đặc trưng là độc lập với nhau. Nguyên lý hoạt động của Naive Bayes là tính toán xác suất của mỗi lớp dựa trên các đặc trưng. Mặc dù giả định "ngây thơ" thường không đúng trong thực tế, nhưng nó giúp giảm độ phức tạp của mô hình và tăng hiệu suất tính toán. Naive Bayes có các biến thể như Gaussian Naive Bayes, Multinomial Naive Bayes và Bernoulli Naive Bayes, phù hợp với các loại dữ liệu khác nhau như dữ liệu số, văn bản, hoặc dữ liệu nhị phân. Mặc dù đơn giản và dễ hiểu, Naive Bayes có thể không hiệu quả nếu các đặc trưng không độc lập với nhau, và có thể không hoạt động tốt trên dữ liệu có phân phối không chuẩn. Tuy nhiên, với dữ liệu lớn và có số chiều lớn, Naive Bayes thường cho kết quả tốt.

* + 1. **Đào tạo và đánh giá mô hình**

|  |
| --- |
| from sklearn.naive\_bayes import GaussianNB  from sklearn.preprocessing import StandardScaler  from sklearn.pipeline import Pipeline  from sklearn.feature\_selection import SelectKBest, f\_classif  pipeline = Pipeline([      ('scaler', StandardScaler()),      ('feature\_selection', SelectKBest(f\_classif)),      ('classifier', GaussianNB(var\_smoothing=1e-8))  ])  # Huấn luyện mô hình với pipeline  pipeline.fit(X\_res, y\_res)  # Dự đoán trên tập kiểm tra và tính toán độ chính xác  y\_pred = pipeline.predict(X\_test)  accuracy\_naive = pipeline.score(X\_test, y\_test)  print("Accuracy:", accuracy\_naive)  print(classification\_report(y\_test,y\_pred)) |

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, biên lai

Mô tả được tạo tự động

Hình 29: Báo cáo về các độ đo của GradientBoostingClassifier

* + 1. **Đánh giá kết quả và giải thích kết quả khai phá**

|  |
| --- |
| con\_max = ConfusionMatrixDisplay.from\_predictions(y\_test, y\_pred)  plt.show() |

Ảnh có chứa ảnh chụp màn hình, văn bản, hàng, biểu đồ

Mô tả được tạo tự động

Hình 30: Confusion matrix cho dự đoán của Naive Bayes

* **Ưu điểm**
* Thuật toán này hoạt động nhanh chóng và có thể tiết kiệm rất nhiều thời gian.
* Naive Bayes phù hợp để giải các bài toán dự đoán đa lớp.
* Nếu giả định về tính độc lập của các tính năng là đúng thì nó có thể hoạt động tốt hơn các mô hình khác và yêu cầu ít dữ liệu huấn luyện hơn nhiều.
* Naive Bayes phù hợp hơn với các biến đầu vào phân loại hơn là các biến số.
* **Nhược điểm**
* Naive Bayes cho rằng tất cả các yếu tố dự đoán (hoặc đặc điểm) đều độc lập, hiếm khi xảy ra trong đời thực. Điều này hạn chế khả năng áp dụng thuật toán này trong các trường hợp sử dụng trong thế giới thực.
* Thuật toán này phải đối mặt với 'vấn đề tần số bằng 0', trong đó nó gán xác suất bằng 0 cho một biến phân loại có danh mục trong tập dữ liệu thử nghiệm không có sẵn trong tập dữ liệu huấn luyện. Sẽ tốt nhất nếu bạn sử dụng kỹ thuật làm mịn để khắc phục vấn đề này.
* Ước tính của nó có thể sai trong một số trường hợp, vì vậy bạn không nên quá coi trọng kết quả xác suất của nó.
* **Ý nghĩa khai phá**

Lớp 0 (Không vỡ nợ):

* + Precision cao (0.89): Khi mô hình dự đoán một khách hàng không vỡ nợ chính xác đến 89%
  + Recall thấp (0.46): Có 46% của tổng số khách hàng thực sự không vỡ nợ được mô hình dự đoán chính xác.
  + F1-score cao hơn (0.6): Điểm số này cho thấy mô hình dự đoán chính xác 60% lớp không vỡ nợ, chủ yếu do recall thấp.

Lớp 1 (Vỡ nợ):

* + Precision thấp (0.13): Chỉ có 13% trong số những dự đoán vỡ nợ của mô hình là chính xác.
  + Recall cao hơn (0.59): Mô hình có khả năng nhận diện 59% của những trường hợp thực sự vỡ nợ.
  + F1-score thấp (0.22): Điểm số này cho thấy mô hình không hiệu quả lắm trong việc dự đoán lớp vỡ nợ, chủ yếu do precision rất thấp.

Accuracy tổng thể (0.47): Mô hình chính xác 47% trên toàn bộ dữ liệu kiểm tra.

Nhận xét chung:

* + Mô hình này có vẻ không cân bằng giữa hai lớp. Mặc dù có precision cao cho lớp không vỡ nợ, nhưng recall thấp cho thấy mô hình bỏ qua nhiều trường hợp thực sự không vỡ nợ.
  + Đối với lớp vỡ nợ, mô hình có recall cao nhưng precision rất thấp, điều này có nghĩa là mô hình dự đoán quá nhiều trường hợp là vỡ nợ khi chúng không phải, dẫn đến nhiều báo động giả.
  + Độ chính xác tổng thể thấp cho thấy mô hình này có thể không phù hợp để đưa ra quyết định về việc cấp tín dụng.
  1. **Mô hình ANN**
     1. **Sơ lược về mô hình**

Artificial Neural Networks (ANN) là một mô hình học máy phỏng theo cấu trúc và chức năng của mạng nơron sinh học, bao gồm các tầng nơron liên kết với nhau. Mỗi nơron nhận đầu vào, thực hiện một phép tính và truyền kết quả đến các nơron khác ở tầng tiếp theo. ANN có khả năng học các mô hình phức tạp và phi tuyến từ dữ liệu, được sử dụng rộng rãi trong nhiều lĩnh vực như nhận dạng hình ảnh, xử lý ngôn ngữ tự nhiên và dự đoán chuỗi thời gian. Tuy nhiên, ANN đòi hỏi lượng dữ liệu lớn để huấn luyện, cần nhiều tài nguyên tính toán và dễ bị overfitting nếu không được điều chỉnh đúng cách.

***Ưu điểm***

* Xử lý dữ liệu chuỗi hoặc có tính thời gian: ANN có khả năng xử lý các dữ liệu chuỗi hoặc có tính thời gian, bằng cách sử dụng bộ nhớ ngắn hạn để lưu trữ trạng thái trước đó.
* Khả năng nhận dạng các mẫu dữ liệu và đưa ra quyết định: ANN có thể học hỏi từ dữ liệu đầu vào và thực hiện nhiều nhiệm vụ phức tạp như nhận dạng mẫu, phân loại và dự đoán.

***Nhược điểm***

* Khó huấn luyện: Huấn luyện mô hình ANN đòi hỏi kiến thức chuyên sâu và kỹ thuật phức tạp.
* Dễ bị vấn đề biến mất hoặc bùng nổ gradient: Trong quá trình huấn luyện, mô hình có thể gặp vấn đề về gradient biến mất (vanishing gradient) hoặc gradient bùng nổ (exploding gradient).
  + 1. **Đào tạo và đánh giá mô hình**

|  |
| --- |
| from sklearn.preprocessing import OneHotEncoder, StandardScaler, QuantileTransformer  import tensorflow as tf  from tensorflow import keras  from tensorflow.keras import callbacks  import joblib  from sklearn.compose import ColumnTransformer  from tensorflow.keras.models import Sequential  from tensorflow.keras.layers import Dense  from tensorflow.keras.callbacks import EarlyStopping  scaler = StandardScaler()  X\_train\_scaled = scaler.fit\_transform(X\_res)  X\_test\_scaled = scaler.transform(X\_test)  ann\_model = tf.keras.models.Sequential([      tf.keras.layers.Dense(256, activation='elu', kernel\_initializer='he\_normal', input\_shape=(X\_train\_Scaled.shape[1],)),      tf.keras.layers.Dense(128, activation='elu', kernel\_initializer='he\_normal'),      tf.keras.layers.BatchNormalization(),      tf.keras.layers.Dense(64, activation='elu', kernel\_initializer='he\_normal'),      tf.keras.layers.Dense(32, activation='elu', kernel\_initializer='he\_normal'),      tf.keras.layers.BatchNormalization(),      tf.keras.layers.Dense(8, activation='elu', kernel\_initializer='he\_normal'),      tf.keras.layers.BatchNormalization(),      tf.keras.layers.Dense(1, activation='sigmoid')  ])  ann\_model.summary() |

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, tài liệu

Mô tả được tạo tự động

|  |
| --- |
| # Compile the model  ann\_model.compile(optimizer='Nadam',                loss='binary\_crossentropy',                metrics=['accuracy'])  import time  n\_epochs = 23  new\_training = 1  time\_chay = 0  if new\_training:      model\_saver = keras.callbacks.ModelCheckpoint('/content/drive/MyDrive/Colab Notebooks/LoanPrediction\_Model/ann\_model.h5',monitor='val\_accuracy', save\_best\_only=True)      early\_stopper = keras.callbacks.EarlyStopping(monitor='val\_loss', patience=5)      performance\_sched = keras.callbacks.ReduceLROnPlateau(monitor='val\_loss', patience=2, factor=0.2)      #with tf.device('/gpu:0'):      start\_time = time.time()      history = ann\_model.fit(X\_train\_Scaled, y\_train, epochs=n\_epochs,                                batch\_size=32,                                validation\_split=0.1),                                callbacks=[model\_saver, early\_stopper, performance\_sched], verbose=1)      end\_time = time.time()      time\_chay = end\_time - start\_time      print("Thời gian chạy: ", time\_chay)      history = history.history      joblib.dump(history,'/content/drive/MyDrive/Colab Notebooks/LoanPrediction\_Model/history\_ann\_model')  else:      try:          ann\_model = keras.models.load\_model('/content/drive/MyDrive/Colab Notebooks/LoanPrediction\_Model/ann\_model.h5')          history = joblib.load('/content/drive/MyDrive/Colab Notebooks/LoanPrediction\_Model/history\_ann\_model')      except Exception as e:          history = joblib.load('/content/drive/MyDrive/Colab Notebooks/LoanPrediction\_Model/history\_ann\_model')  history.keys()  print("Train\_Accuracy:", history['accuracy'][-1])  print("Val\_Accuracy:", history['val\_accuracy'][-1])  #endregion |

Ảnh có chứa văn bản, Phông chữ, ảnh chụp màn hình, số

Mô tả được tạo tự động

|  |
| --- |
| # %% Plot learning curves  import pandas as pd  pd.DataFrame({'Training acc':history['accuracy'],'Validation acc':history['val\_accuracy']}).plot(figsize=(8, 5))  plt.grid(True)  plt.xlim(0, n\_epochs)  plt.ylim(0, 1)  plt.xlabel(f'epoch, time: {time\_chay}')  plt.xticks(np.arange(0,n\_epochs+1,5))  plt.savefig("/content/drive/MyDrive/Colab Notebooks/LoanPrediction\_Model/ann\_model")  plt.title("Accuracy")  plt.show() |

Ảnh có chứa văn bản, ảnh chụp màn hình, hàng, Sơ đồ

Mô tả được tạo tự động

Hình 31. Biểu đồ biểu diễn quá trình học của mô hình ANN

* + 1. **Đánh giá kết quả và giải thích kết quả khai phá**

|  |
| --- |
| y\_pred = (ann\_model.predict(X\_test\_scaled) > 0.5).astype("int32")  #y\_pred = model.predict\_classes(X\_test)  accuracy\_ann = accuracy\_score(y\_test, y\_pred)  print(confusion\_matrix(y\_test, y\_pred))  print(classification\_report(y\_test, y\_pred)) |

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động

Hình 32. Báo cáo các độ đo của ANN

**Hiệu suất Mô hình:**

* Training Accuracy: 0. 8889
* Validation Accuracy: 0.9315
* Test Accuracy: 0.87

**Quá trình Huấn luyện:**

* Loss và Accuracy cải thiện qua các epoch:
  + Ban đầu, loss giảm và accuracy tăng đáng kể.
  + Từ epoch thứ 13 trở đi, các chỉ số không cải thiện đáng kể và có dấu hiệu dao động nhẹ, cho thấy mô hình có thể đã đạt tới mức bão hòa.

**Đánh giá:**

* Training và Validation Accuracy đều cao: Điều này cho thấy mô hình có khả năng học tốt từ dữ liệu huấn luyện và tổng quát tốt trên dữ liệu kiểm tra.
* Không có dấu hiệu Overfitting rõ ràng: Accuracy trên tập huấn luyện và kiểm tra không chênh lệch quá lớn, cho thấy mô hình không bị overfit.
* Kết quả đánh giá các độ đo khá tốt, mô hình có thể được sử dụng để dự đoán.

**Tổng kết độ chính xác của các mô hình:**

|  |
| --- |
| data = {      'Model': ['Logistic Regression', 'KNN', 'Random Forest', 'AdaBoost', 'Gradient Boosting', 'Naive Bayes', 'ANN'],      'Accuracy': [accuracy\_logis, accuracy\_knn, accuracy\_rf, accuracy\_ada, accuracy\_gra, accuracy\_naive, accuracy\_ann]  }  df = pd.DataFrame(data)  df |

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động

Hình 33. DataFrame biểu diễn độ chính xác của từng mô hình

* 1. **Phân cụm sử dụng Kmeans**
     1. **Sơ lược về mô hình**

Phân cụm K-means là thuật toán học máy không giám sát phổ biến, phân nhóm dữ liệu thành các cụm có liên quan. Thuật toán lặp lại hai bước: gán mỗi điểm dữ liệu cho cụm gần nhất dựa trên khoảng cách đến trung tâm cụm, và cập nhật vị trí trung tâm cụm cho đến khi hội tụ

***Ưu điểm:***

* Đơn giản và dễ hiểu: Thuật toán K-means dễ thực hiện và cài đặt. Nó phân chia dữ liệu thành các cụm dựa trên khoảng cách giữa các điểm dữ liệu và trung tâm của cụm.
* Phân cụm tự động: K-means tự động xác định số lượng cụm (K) mà không cần thông tin trước.

***Nhược điểm:***

* Phụ thuộc vào số lượng cụm K: Hiệu suất của K-means phụ thuộc vào việc chọn số lượng cụm K. Nếu chọn sai K, kết quả phân cụm có thể không chính xác.
* Nhạy cảm với các giá trị ngoại lai (outliers): K-means dễ bị ảnh hưởng bởi các điểm dữ liệu ngoại lai, gây sai sót trong việc phân cụm.
* Chi phí tính toán: Thuật toán K-means thực hiện vòng lặp tính toán khoảng cách giữa các điểm dữ liệu và trung tâm cụm. Với số cụm K lớn và dữ liệu phân cụm lớn, chi phí tính toán có thể tăng.
  + 1. **Sơ lược về mô hình**

Tìm số cụm thích hợp bằng cách đánh giá inertia

|  |
| --- |
| from sklearn.cluster import KMeans  # Fitting K-means and evaluate inertia for different values of k.  num\_clusters = [i for i in range(2,11)]  def kmeans\_inertia(num\_clusters, x\_values):      inertia = []      for num in num\_clusters:          kms = KMeans(n\_clusters = num, random\_state = 123)          kms.fit(x\_values)          inertia.append(kms.inertia\_)      return inertia  inertia = kmeans\_inertia(num\_clusters, X\_scaled)  inertia |

**Ảnh có chứa văn bản, Sơ đồ, hàng, ảnh chụp màn hình

Mô tả được tạo tự động**

Hình 32. Giá trị inertia cho từng cụm

Như chúng ta thấy, inertia giảm đáng kể khi số lượng cụm tăng từ 2 lên 6. Tuy nhiên, sau đó, sự khác biệt là ít hơn rất nhiều. Biểu đồ dường như mô tả một khuỷu tay ở sáu cụm, nhưng không có phương pháp rõ ràng để xác nhận rằng mô hình sáu cụm là tối ưu. Do đó, silhouette scores nên được kiểm tra.

|  |
| --- |
| from sklearn.metrics import silhouette\_score  num\_clusters = [i for i in range(2,11)]  def kmeans\_sil(num\_clusters,x\_values):      sil\_score = []      for num in num\_clusters:          kms = KMeans(n\_clusters = num, random\_state = 123)          kms.fit(x\_values)          sil\_score.append(silhouette\_score(x\_values,kms.labels\_))      return sil\_score  sil\_score = kmeans\_sil(num\_clusters,X\_scaled)  sil\_score |

**Ảnh có chứa biểu đồ, Sơ đồ, hàng, văn bản

Mô tả được tạo tự động**

Hình 33. Điểm silhoutte tương ứng với từng số cụm

Biểu đồ chỉ ra rằng điểm Silhouette giảm đáng kể từ 2 xuống 3 cụm, sau đó tăng dần khi dữ liệu được phân chia thành sáu cụm. Vì vậy, 6 cụm mang lại điểm Silhouette tương đối tốt. Do đó, chúng ta chọn 6 là số cụm tối ưu để xây dựng mô hình KMeans.

|  |
| --- |
| kmeans6 = KMeans(n\_clusters = 6, random\_state = 123)  kmeans6.fit(X\_scaled)  df2['Cluster'] = kmeans6.labels\_  df2  df2.groupby(['Cluster','Risk\_Flag']).size() |

**Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động**

Hình 34. Số lượng bình thường và rủi ro tương ứng với từng cụm

* + 1. **Ý nghĩa kết quả khai phá**
* Số lượng không rủi ro (Risk\_Flag = 0) luôn cao hơn đáng kể so với số lượng có rủi ro (Risk\_Flag = 1) trong mỗi cụm.
* Cụm 2 và 5 có số lượng không rủi ro cao nhất, lần lượt là 49596 và 52397, điều này cho thấy hai cụm này có thể có các yếu tố mạnh mẽ hơn trong việc ngăn chặn rủi ro vỡ nợ.
* Cụm 3 có số lượng không rủi ro thấp nhất là 16858, nhưng cũng có tỷ lệ rủi ro thấp (1737 so với 16858), có thể là do cụm này nhỏ hơn các cụm khác.
* Tỷ lệ rủi ro vỡ nợ (Risk\_Flag = 1) tăng dần từ cụm 0 đến 5, với cụm 5 có số lượng rủi ro cao nhất là 8738.

# **CHƯƠNG 5. KẾT LUẬN VÀ KIẾN NGHỊ**

* 1. **Tóm tắt kết quả nghiên cứu**

Dữ liệu: Sử dụng dữ liệu lịch sử về hành vi của khách hàng để xây dựng mô hình dự đoán.

Mô hình học máy: Áp dụng các mô hình học máy như Random Forest, Logistic Regression, AdaBoosting, Navie Bayes, ANN để dự đoán khả năng vỡ nợ của khách hàng mới. Áp dụng kỹ thuật phân cụm dùng mô hình Kmeans để phân cụm khách hàng thành các nhóm và đánh giá độ rủi ro dựa trên tỷ lệ rủi ro từng nhóm

Đánh giá mô hình: Sử dụng các phương pháp đánh giá như precision, recall, F1-score và ROC-AUC để đánh giá hiệu suất của mô hình trên tập dữ liệu kiểm tra.

Kết quả: Hiểu được và phân tích được sự phù hợp và không phù hợp của các loại mô hình khác nhau, tối ưu chỉ số accuracy đến mức tối đa

* 1. **Kiến nghị dựa trên kết quả**

Tích hợp mô hình vào quy trình quyết định: tích hợp mô hình dự đoán vào quy trình xác định rủi ro và xác nhận khách hàng mới. Điều này giúp tổ chức nhanh chóng xác định khách hàng có rủi ro cao hơn và đưa ra biện pháp phòng tránh kịp thời.

Cập nhật và tinh chỉnh định kỳ: đề xuất cập nhật định kỳ mô hình dự đoán dựa trên dữ liệu mới và tinh chỉnh mô hình để duy trì hiệu suất cao.

* 1. **Hướng nghiên cứu và phát triển tiếp theo**

Nâng cao hiệu suất mô hình: tiếp tục nghiên cứu và thử nghiệm các phương pháp mới để cải thiện hiệu suất của mô hình, bao gồm việc sử dụng các thuật toán học máy tiên tiến và tối ưu hóa siêu tham số.

Mở rộng dữ liệu: nghiên cứu việc tích hợp thêm các thông tin mới vào dữ liệu, như dữ liệu tín dụng từ các tổ chức bên ngoài để cải thiện dự đoán về khả năng vỡ nợ.

Nghiên cứu về giải pháp phòng tránh: tìm hiểu và phát triển các giải pháp phòng tránh để giảm thiểu rủi ro về vỡ nợ từ nhóm khách hàng có khả năng cao.

# **PHẦN ĐÁNH GIÁ LÀM VIỆC NHÓM**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **BẢNG ĐÁNH GIÁ QUÁ TRÌNH LÀM ĐỒ ÁN CỦA NHÓM**  **Họ tên: Trương Thị Thùy Dung**  **Vị trí công việc: Nhóm trưởng**  **Các đánh giá theo các tiêu bên dưới bằng cách chấm điểm từ 1 tới 10 tương ứng với các mức đánh giá từ thấp nhất đến cao nhất được liệt kê chi tiết** | | | | | |
| **Tiêu chí** | **Tiêu chí thấp nhất / cao nhất** | **Số điểm đánh giá** | | | |
| **Dung** | **Duy** | **Hào** | **Thương** |
| * 1. Mục tiêu | * Không rõ ràng, cụ thể, thiếu cam kết * Rõ ràng, cụ thể, có tính cam kết đầy đủ với tất cả thành viên trong nhóm | 10 | 10 | 10 | 10 |
| * 1. Sự cởi mở | * Các thành viên đề phòng, thận trọng trong thảo luận nhóm và làm việc * Các thành viên thoải mái bày tỏ suy nghĩ, cảm xúc và ý tưởng trong thảo luận và làm việc | 10 | 10 | 10 | 10 |
| * 1. Tin cậy | * Các thành viên nghi ngờ, thiếu tin cậy nhau * Các thành viên tin tưởng lẫn nhau | 10 | 10 | 10 | 10 |
| * 1. Giải quyết mâu thuẫn | * Các thành viên bị động hoặc tránh để xảy ra mâu thuẫn, khác biệt trong khi làm việc * Các thành viên chủ động đề cập đến các mâu thuẫn, khác biệt để tìm phương hướng giải quyết tốt nhất | 10 | 10 | 10 | 10 |
| * 1. Sự phối hợp | * Các thành viên bị động đề nghị phối hợp hoặc nhận trợ giúp * Các thành viên chủ động và cảm thấy thoải mái khi phối hợp làm việc | 10 | 10 | 10 | 10 |
| * 1. Sự tham gia | * Quá trình thảo luận, làm việc bị chi phối bởi một vài thành viên * Tất cả thành viên đều tham gia chủ động, tích cực vào thảo luận và làm việc | 10 | 10 | 10 | 10 |
| * 1. Đưa ra quyết định | * Quá trình đưa ra quyết định của nhóm bị chi phối bởi một vài thành viên * Tất cả thành viên đều được tham gia thực chất vào quá trình đưa ra quyết định | 10 | 10 | 10 | 10 |
| * 1. Sự linh hoạt | * Bó buộc, cứng nhắc tuân thủ các quy định, nguyên tắc * Chủ động, linh hoạt điều chỉnh quy tắc làm việc để ứng phó với các tình huống phát sinh | 10 | 10 | 10 | 10 |
| * 1. Sử dụng nguồn nhân lực | * Khả năng, kiến thức của từng cá nhân không được sử dụng tốt * Khả năng, kiến thức của từng cá nhân được sử dụng tối đa | 10 | 10 | 10 | 10 |
| * 1. Tinh thần trách nhiệm | * Không thực hiện đầy đủ nhiệm vụ, thường xuyên trễ hạn và không cam kết với công việc. * Có trách nhiệm cao, hoàn thành nhiệm vụ đúng hạn và luôn cam kết với công việc của mình. | 10 | 10 | 10 | 10 |
| ***Tổng điểm*** | | ***100*** | ***100*** | ***100*** | ***100*** |

***Đánh giá chung:***

* Động lực làm việc của cá nhân khi làm việc nhóm: các thành viên đều có động lực làm việc rất cao, luôn cam kết và nỗ lực hết mình trong quá trình thực hiện dự án.
* Động lực làm việc chung của nhóm: nhóm có động lực làm việc rất cao, tinh thần đồng đội và sự hợp tác chặt chẽ giúp đạt được hiệu quả công việc tối ưu.
* Tính hiệu quả của quy trình làm việc hiện tại: quy trình làm việc của nhóm rất hiệu quả, với sự phân công nhiệm vụ rõ ràng và sử dụng tối đa khả năng của từng thành viên.

Với đánh giá tối đa cho tất cả các yếu tố, nhóm đã thể hiện một thái độ làm việc tích cực, chủ động, hợp tác và sáng tạo, tạo nên một môi trường làm việc hiệu quả và năng động. Điều này sẽ giúp nhóm đạt được kết quả tốt nhất trong việc thực hiện đồ án công nghệ thông tin về training các mô hình học máy dự đoán

|  |  |
| --- | --- |
| **BẢNG ĐÁNH GIÁ ƯU NHƯỢC CỦA QUÁ TRÌNH LÀM VIỆC NHÓM** | |
| **Ưu điểm** | **Nhược điểm** |
| 1. ***Điểm ảnh hưởng từ quá trình làm việc của cả nhóm và chia cho các thành viên*** | |
| Các thành viên nhóm đều nỗ lực làm việc vì kết quả cả cả nhóm cũng như chính bản thân mình. Phân công công việc từ đó cũng rõ ràng và công bằng | Đóng góp của mỗi cá nhân không được phản ánh trong việc cho điểm Sinh viên kém có thể được hưởng lợi từ công việc của những sinh viên chăm chỉ Sinh viên giỏi có thể bị kìm lại bởi sinh viên kém Không tạo động lực cho sinh viên |
| 1. ***Có sự hợp tác giữa các thành viên, tận dụng kiến thức của nhau*** | |
| Thành viên yếu hơn có cơ hội học hỏi và nhận sự giúp đỡ từ những người có kinh nghiệm hơn.  Khi có sự hỗ trợ từ nhóm, áp lực công việc sẽ được chia sẻ, giúp mọi người cảm thấy an toàn hơn. | Sự chênh lệch về đóng góp và nỗ lực giữa các thành viên có thể dẫn đến mâu thuẫn, xung đột trong nhóm.  Những người có năng lực tốt có thể bị ảnh hưởng bởi những người làm việc kém, không phát huy hết khả năng của mình |
| 1. ***Thái độ cảu từng thành viên*** | |
| Làm việc nhóm giúp các thành viên gắn kết và hiểu nhau hơn, tạo ra một môi trường làm việc tích cực | Nếu có thành viên không chịu trách nhiệm hoặc làm việc không hiệu quả, cả nhóm sẽ bị ảnh hưởng. |
| 1. ***Tổ chức các buổi họp định kỳ*** | |
| Các buổi họp định kỳ giúp nhóm nắm bắt được tiến độ của dự án, xác định những phần công việc đã hoàn thành và những phần còn đang thực hiện.  Các buổi họp định kỳ tạo điều kiện cho các thành viên trao đổi thông tin, chia sẻ ý kiến và thảo luận về các phương án tốt nhất để thực hiện công việc. | Việc tổ chức các buổi họp định kỳ có thể tốn nhiều thời gian, nhất là khi nhóm phải thảo luận về nhiều vấn đề phức tạp.  Các buổi họp có thể làm gián đoạn công việc của các thành viên, đặc biệt nếu họ đang làm việc hiệu quả và không muốn bị phân tâm. |

***Giải pháp để tối ưu hóa:***

* Phân công nhiệm vụ cụ thể cho từng thành viên để đảm bảo mọi người biết rõ vai trò của mình.
* Xây dựng kế hoạch làm việc chi tiết từ đầu đến cuối để tránh lãng phí thời gian.
* Thường xuyên tổ chức buổi họp để cập nhật tiến độ và giải quyết vấn đề kịp thời.
* Khuyến khích sự giao tiếp và chia sẻ ý kiến giữa các thành viên để tăng cường tinh thần nhóm.
* Nhận biết và giải quyết mâu thuẫn ngay khi chúng phát sinh để tránh tác động tiêu cực đến quá trình làm việc.
* Sử dụng công nghệ và công cụ hỗ trợ như Google Drive, GitHub để tối ưu hóa quá trình làm việc.
* Luôn sẵn lòng nhận phản hồi và đề xuất cải tiến để nâng cao hiệu suất làm việc của nhóm.

# **TÀI LIỆU THAM KHẢO**

1. ***Loan Default Prediction Based on Customer Behavior Using Xgboost And SMOTE ENN* - Mehul Wankhede**(Truy cập ngày 10/05/2024)

Link: <https://medium.com/@mswankhede7/loan-prediction-based-on-customer-behavior-using-xgboost-and-smote-enn-983b35380af8>

1. ***Loan Prediction Based on Customer Behavior* -** **SUBHAM SURANA**(Truy cập ngày 12/05/2024)

Link: <https://www.kaggle.com/datasets/subhamjain/loan-prediction-based-on-customer-behavior/data>

1. ***Chi-squared test* -** **From Wikipedia, the free encyclopedia** (Truy cập ngày 09/05/2024)

Link: <https://en.wikipedia.org/wiki/Chi-squared_test>

1. ***Google Colab* -** **Google**

Link: <https://colab.research.google.com/>

1. ***Naive Bayes Classifier : Advantages and Disadvantages*** (Truy cập ngày 12/05/2024)

Link: https://machinelearninginterview.com/topics/machine-learning/naive-bayes-classifier-advantages-and-disadvantages/

1. **Advantages and Disadvantages of Logistic Regression** (Truy cập ngày 12/05/2024)

Link: https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/