

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM
KHOA CÔNG NGHỆ THÔNG TIN**



HCMUTE

ĐỒ ÁN CUỐI KỲ

PHÂN TÍCH HÀNH VI KHÁCH HÀNG ĐỂ DỰ ĐOÁN VIỆC CHO VAY

Môn: Khai phá dữ liệu

Thực hiện: Nhóm 5. Thứ 4, tiết 7, 10.

GVHD: Nguyễn Văn Thành

Thành phố Hồ Chí Minh, tháng 05 năm 2023

DANH SÁCH NHÓM LÀM ĐỒ ÁN CUỐI KỲ
MÔN KHAI PHÁ DỮ LIỆU
HỌC KỲ II NĂM HỌC 2022-2023

- 1. Mã lớp môn học:** DAMI330484_22_2_01 (Thứ 4, tiết 7, 10)
- 2. Giảng viên hướng dẫn:** Nguyễn Văn Thành
- 3. Tên đề tài:** Phân tích hành vi khách hàng để dự đoán việc cho vay
- 4. Danh sách nhóm viết tiểu luận cuối kỳ:**

STT	Họ tên sinh viên	Mã số sinh viên	Tỷ lệ tham gia %	Kí tên
1	Nguyễn Trị Quốc	20133084	100%	
2	Nguyễn Ý	20133117	100%	
3	Nguyễn Thanh Tùng	20133111	100%	
4	Đinh Quang Thắng	20133089	100%	

- Tỷ lệ % = 100%
- Trưởng nhóm: Nguyễn Trị Quốc

Nhận xét của giáo viên:

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Tháng 05 năm 2023
Giáo viên chấm điểm

LỜI CẢM ƠN

Lời đầu tiên, nhóm em xin được gửi lời cảm ơn đặc biệt đến Thầy Nguyễn Văn Thành - Giảng viên phụ trách môn Khai phá dữ liệu – trường đại học Sư Phạm Kỹ Thuật Tp.Hồ Chí Minh.

Trong thời gian nhóm em làm đồ án, tụi em đã nhận được nhiều sự giúp đỡ từ thầy. Thầy đã cung cấp đầy đủ kiến thức, chỉ bảo và đóng góp những ý kiến quý báu giúp tụi em có thể hoàn thành được đồ án môn học của mình một cách tốt nhất.

Xuất phát từ mục đích học tập, tìm hiểu sâu hơn các kiến thức về tương tác dữ liệu, cũng như tìm hiểu kỹ về quy trình nghiệp vụ của lên ý tưởng, xây dựng kho dữ liệu, thống kê dữ liệu. Nhóm chúng em đã thực hiện đồ án “Phân tích hành vi khách hàng để dự đoán việc cho vay”. Trong quá trình thực hiện đồ án, dựa trên kiến thức được Thầy cung cấp qua các buổi học lý thuyết cũng như thực hành trên lớp, kết hợp với việc tự tìm hiểu những công cụ và kiến thức mới, nhóm đã cố gắng thực hiện đồ án một cách tốt nhất. Tuy nhiên, đồ án còn chưa được hoàn thiện và có nhiều sai sót.

Nhóm rất mong nhận được sự góp ý từ thầy nhằm rút ra những kinh nghiệm quý báu và hoàn thiện vốn kiến thức để nhóm có thể hoàn thành những đồ án, dự án khác trong tương lai.

Nhóm chúng em xin chân thành cảm ơn quý thầy!

LỜI MỞ ĐẦU

Trong thời đại số hóa hiện nay, việc nắm bắt và phân tích hành vi khách hàng trở thành yếu tố quan trọng trong việc đưa ra quyết định kinh doanh thông minh. Đặc biệt trong lĩnh vực tài chính, dự đoán khả năng vay vốn của khách hàng dựa trên hành vi và thông tin cá nhân có thể giúp ngân hàng và tổ chức tín dụng đánh giá rủi ro và quyết định cho vay một cách hiệu quả hơn.

Mục tiêu của đồ án này là áp dụng các kỹ thuật khai thác dữ liệu (Data Mining) để phân tích tập dữ liệu "Loan Prediction Based on Customer Behavior" và xây dựng mô hình dự đoán khả năng vay vốn của khách hàng dựa trên hành vi và các yếu tố tài chính liên quan. Bằng cách tìm hiểu và áp dụng các thuật toán và phương pháp phân tích dữ liệu, chúng tôi hy vọng đồ án này sẽ mang lại những thông tin giá trị và cái nhìn sâu sắc về khách hàng và quá trình vay vốn.

Trong bài tiểu luận này, nhóm chúng em sẽ tiến hành phân tích dữ liệu, tiền xử lý dữ liệu, xây dựng mô hình dự đoán và đánh giá mô hình để nghiên cứu về việc dự đoán khả năng vay vốn dựa trên hành vi khách hàng. Qua việc áp dụng các kỹ thuật phân tích dữ liệu, chúng em hy vọng có thể khám phá các mô hình và mối quan hệ giữa các biến để xây dựng một mô hình dự đoán chính xác để các tổ chức tín dụng và ngân hàng đưa ra các quyết định cho vay thông minh và hiệu quả.

MỤC LỤC

LỜI CẢM ƠN	3
LỜI MỞ ĐẦU	4
Chương 1: Lý do chọn Dataset và giới thiệu tổng quan Dataset	7
1.1 Lý do hình thành dự án.....	7
1.1.1 Vấn đề nhận thấy	7
1.1.2 Giải pháp	7
1.1.3 Mục tiêu và ý nghĩa của dự án.....	8
1.2 Giới thiệu tổng quan dataset.....	9
1.2.1 Nguồn dữ liệu sử dụng	9
1.2.1.1 Giới thiệu nơi cấp dữ liệu.....	9
1.2.1.2 Hướng dẫn tải dataset thực hiện trong đồ án.....	9
1.2.2 Mô tả chi tiết dữ liệu.....	9
1.2.2.1 Thông số dataset	9
1.2.2.2 Dữ liệu sau khi trích xuất.....	10
1.2.2.3 Mô tả chi tiết các thuộc tính trong dataset	10
1.2.3 Giới thiệu các công cụ được sử dụng trong đồ án.....	11
1.2.3.1 Tổng quan về Visual Studio 2019 Community	11
1.2.3.2 Tổng quan về SQL Server 2019 Developer.....	13
1.2.3.3 Tổng quan về Visual Studio Code	14
1.2.3.4 Giới thiệu về ngôn ngữ lập trình python3	15
Chương 2: Cơ sở lý thuyết.....	15
2.1 Kỹ thuật Naive Bayes	15
2.2 Kỹ thuật Decision Tree.....	16
2.3 Kỹ thuật K-Means	17
2.4 Kỹ thuật Apriori.....	18
Chương 3: Phân tích khám phá dữ liệu (EDA).....	20
3.1 Biểu đồ cột để phân tích thống kê cột Risk Flag	20
3.2 Biểu đồ cột để phân tích thống kê cột Profession.....	21
3.3 Biểu đồ cột để phân tích thống kê cột Current Job Year	21
3.4 Biểu đồ cột để phân tích thống kê cột Experience	22
3.5 Biểu đồ hộp để phân tích thống kê cột Income	22
3.6 Biểu đồ hộp để phân tích thống kê cột Age	23
3.7 Biểu đồ tròn để phân tích thống kê cột House Ownership.....	23
3.8 Biểu đồ tròn để phân tích thống kê cột Car Ownership.....	23
3.9 Biểu đồ tròn để phân tích thống kê cột Marital Status.....	24
3.10 Biểu đồ tròn để phân tích thống kê cột State	24
Chương 4: Tiền xử lý dữ liệu và xây dựng mô hình	25

4.1 Tiền xử lý dữ liệu	25
4.2 Xây dựng mô hình	25
4.2.1 Xây dựng cơ sở dữ liệu bằng công cụ SSIS	25
4.2.1.1 Quá trình tạo mới project SSIS	25
4.2.1.2 Quá trình đổ dữ liệu từ excel vào database	27
4.2.1.2.1 Quá trình tạo Flat File Connection Manager	27
4.2.1.2.2 Quá trình tạo OLEDB Connection Manager	29
4.2.1.2.3 Quá trình tạo SSIS Package.....	32
4.2.1.2.4 Quá trình tạo Control Flow	33
4.2.1.2.4.1 Excute SQL Task	33
4.2.1.2.4.2 Data Flow Task	34
4.2.1.2.5 Kết quả chạy SSIS Package	37
4.2.2 Xây dựng mô hình khai phá dữ liệu bằng công cụ SSAS	38
Chương 5: Giải quyết nghiệp vụ	45
5.1 Phát biểu nghiệp vụ	45
5.2 Giải quyết nghiệp vụ bằng các kỹ thuật	46
5.2.1 Kỹ thuật phân tích phân loại (Classification Analysis)	46
5.2.1.1 Kỹ thuật Naive Bayes	46
5.2.1.2 Kỹ thuật Decision Tree.....	48
5.2.2 Kỹ thuật phân tích theo cụm (Clustering Analysis)	50
5.2.2.1 Kỹ thuật K-Means	50
5.2.3 Kỹ thuật kết hợp (Association Analysis).....	52
5.2.3.1 Kỹ thuật Apriori	52
Chương 6: Đánh giá mô hình	53
6.1 Bảng so sánh các thuật toán.....	53
Chương 7: Kết luận	54
7.1 Kết quả đạt được	54
7.2 Những hạn chế	54
7.3 Bảng phân công nhiệm vụ trong nhóm	54
7.4 Tài liệu tham khảo.....	55

Chương 1: Lý do chọn Dataset và giới thiệu tổng quan Dataset

1.1 Lý do hình thành dự án

1.1.1 Vấn đề nhận thấy

Trong thời đại số hóa hiện nay, việc nắm bắt và phân tích hành vi khách hàng trở thành yếu tố quan trọng trong việc đưa ra quyết định kinh doanh thông minh. Đặc biệt trong lĩnh vực tài chính, dự đoán khả năng vay vốn của khách hàng dựa trên hành vi và thông tin cá nhân có thể giúp ngân hàng và tổ chức tín dụng đánh giá rủi ro và quyết định cho vay một cách hiệu quả hơn.

Một trong những vấn đề quan trọng của việc cho vay là đánh giá và quản lý rủi ro tín dụng. Ngân hàng cần đảm bảo rằng khách hàng mà họ cho vay có khả năng trả nợ và tuân thủ các điều khoản hợp đồng. Điều này đòi hỏi phải đánh giá chính xác khả năng thanh toán của khách hàng dựa trên các yếu tố như thu nhập, lịch sử tín dụng và hành vi thanh toán trước đó.

1.1.2 Giải pháp

Trong việc giải quyết các vấn đề và thách thức liên quan đến việc cho vay của các ngân hàng, một trong những giải pháp quan trọng là áp dụng Data Mining. Data Mining là quá trình khám phá và phân tích các mẫu, thông tin và kiến thức hữu ích từ dữ liệu để hỗ trợ quyết định và dự đoán.

Phân tích hồ sơ tín dụng, chúng ta có thể phân tích hồ sơ tín dụng của khách hàng để đánh giá khả năng trả nợ và rủi ro tín dụng. Chúng ta có thể xây dựng các mô hình dự đoán dựa trên các yếu tố như thu nhập, lịch sử tín dụng, số lần vay trước đó, v.v. Điều này giúp ngân hàng đưa ra quyết định cho vay thông minh và chính xác hơn.

Từ đó ta có thể xây dựng mô hình và áp dụng các thuật toán như Decision Trees (Cây quyết định), K-mean và các loại thuật toán khác để có thể đánh giá hành vi khách hàng để dự đoán khả năng trả nợ và quyết định cho vay.

1.1.3 Mục tiêu và ý nghĩa của dự án

Mục tiêu và ý nghĩa của dự án "Phân tích hành vi khách hàng để dự đoán việc cho vay" là tạo ra một mô hình dự đoán khả năng vay vốn dựa trên hành vi khách hàng. Dự án này mang lại nhiều ý nghĩa quan trọng cho các ngân hàng và các tổ chức tài chính.

Mục tiêu chính của dự án là xây dựng một mô hình dự đoán khả năng trả nợ của khách hàng. Điều này giúp ngân hàng đánh giá rủi ro tín dụng và quyết định xem liệu khách hàng có nên được cho vay hoặc không. Dự án giúp cải thiện quy trình xét duyệt vay vốn, đảm bảo tính chính xác và khách quan trong việc đưa ra quyết định cho vay.

Dự án cung cấp cơ sở dữ liệu và các phương pháp phân tích để tối ưu hóa quy trình cho vay của ngân hàng. Bằng cách sử dụng các thuật toán Data Mining, dự án có thể phân tích và xác định các yếu tố quan trọng trong quyết định cho vay và cung cấp thông tin giúp ngân hàng tối ưu hóa chính sách và quy trình xét duyệt vay.

Dự án đóng góp vào việc tăng tính minh bạch và công bằng trong quy trình cho vay. Thay vì dựa vào quyết định chủ quan của nhân viên ngân hàng, mô hình dự đoán dựa trên dữ liệu và thuật toán giúp đưa ra quyết định công bằng và khách quan. Điều này giúp đảm bảo rằng các khách hàng được đánh giá dựa trên năng lực tài chính thực sự và giảm thiểu sự thiên vị.

Dự án giúp giảm thiểu rủi ro tín dụng của ngân hàng bằng cách đánh giá và dự đoán khả năng trả nợ của khách hàng. Bằng cách phân tích dữ liệu về hành vi và lịch sử tín dụng, ngân hàng có thể xác định các yếu tố có liên quan đến khả năng trả nợ và rủi ro. Điều này giúp ngân hàng đưa ra quyết định cho vay thông minh hơn, giảm thiểu rủi ro tín dụng và đảm bảo tính bền vững của hoạt động cho vay.

1.2 Giới thiệu tổng quan dataset

1.2.1 Nguồn dữ liệu sử dụng

Tập dữ liệu này được cung cấp công khai bởi SUBHAM SURANA trên kaggle.com và được chấp nhận sử dụng miễn phí cho mục đích nghiên cứu và phân tích.

1.2.1.1 Giới thiệu nơi cấp dữ liệu

Tập dữ liệu được sử dụng trong dự án này được lấy từ Kaggle - một trang web cung cấp các tập dữ liệu mở và dữ liệu khoa học để các nhà nghiên cứu, kỹ sư dữ liệu và các chuyên gia có thể sử dụng.

1.2.1.2 Hướng dẫn tải dataset thực hiện trong đồ án

Link tải dataset: <https://www.kaggle.com/datasets/subhamjain/loan-prediction-based-on-customer-behavior/download?datasetVersionNumber=1>

Nguồn dữ liệu này được thu thập từ các hoạt động giao dịch và hành vi khách hàng trong lĩnh vực cho vay. Nó bao gồm một loạt các thông tin về khách hàng như tuổi, giới tính, thu nhập, nghề nghiệp, nơi ở, lịch sử thanh toán, v.v.

1.2.2 Mô tả chi tiết dữ liệu

Tập dữ liệu "Loan Prediction Based on Customer Behavior" chứa thông tin về các khách hàng và hành vi tài chính liên quan đến việc vay vốn. Tập dữ liệu có tổng cộng 13 cột và 252.000 hàng.

1.2.2.1 Thông số dataset

Dữ liệu gồm có 3 dataset nhưng ta chỉ cần 1 dataset để phân tích:

Training Data.csv có: 252.000 (dòng) * 13(cột), với mỗi dòng là thông tin cá nhân của khách hàng.

1.2.2.2 Dữ liệu sau khi trích xuất

Id	Income	Age	Experience	Married/Single	House_Ownership	Car_Ownership	Profession
1	1303834	23	3	single	rented	no	Mechanical_engineer
2	7574516	40	10	single	rented	no	Software_Developer
3	3991815	66	4	married	rented	no	Technical_writer
4	6256451	41	2	single	rented	yes	Software_Developer
5	5768871	47	11	single	rented	no	Civil_servant
6	6915937	64	0	single	rented	no	Civil_servant
7	3954973	58	14	married	rented	no	Librarian
8	1706172	33	2	single	rented	no	Economist
9	7566849	24	17	single	rented	yes	Flight_attendant
10	8964846	23	12	single	rented	no	Architect
11	4634680	78	7	single	rented	no	Flight_attendant
12	6623263	22	4	single	rented	no	Designer
13	9120988	28	9	single	rented	no	Physician
14	8043880	57	12	single	rented	no	Financial_Analyst
15	9420838	48	6	single	rented	no	Technical_writer
16	5694236	39	2	married	rented	yes	Economist
17	7315840	71	8	married	rented	no	Air_traffic_controller
18	3666346	56	12	single	rented	no	Politician
19	2241112	28	8	single	rented	no	Police_officer
20	5431918	40	1	single	rented	no	Artist
21	9225468	54	14	single	rented	no	Surveyor
22	6506739	50	4	single	rented	no	Politician
23	9157379	72	13	single	rented	yes	Design_Engineer

CITY	STATE	CURRENT_JOB_YRS	CURRENT_HOUSE_YRS	Risk_Flag
Rewa	Madhya_Pradesh	3	13	0
Parbhani	Maharashtra	9	13	0
Alappuzha	Kerala	4	10	0
Bhubaneswar	Odisha	2	12	1
Tiruchirappalli[10]	Tamil_Nadu	3	14	1
Jalgaon	Maharashtra	0	12	0
Tiruppur	Tamil_Nadu	8	12	0
Jamnagar	Gujarat	2	14	0
Kota[6]	Rajasthan	11	11	0
Karimnagar	Telangana	5	13	0
Hajipur[31]	Bihar	7	12	0
Adoni	Andhra_Pradesh	4	14	0
Erode[17]	Tamil_Nadu	9	12	0
Kollam	Kerala	8	10	0
Madurai	Tamil_Nadu	6	10	1
Anantapuram[24]	Andhra_Pradesh	2	10	0
Kamarhati	West_Bengal	8	14	0
Bhusawal	Maharashtra	12	11	1
Sirsa	Haryana	6	14	0
Amaravati	Andhra_Pradesh	1	14	0
Secunderabad	Telangana	8	10	0
Ahmedabad	Gujarat	4	11	0
Ajmer	Rajasthan	9	10	0

1.2.2.3 Mô tả chi tiết các thuộc tính trong dataset

Tên thuộc tính	Mô tả
Income	Thu nhập của khách hàng

Age	Tuổi của khách hàng
experience	Số năm kinh nghiệm trong lĩnh vực của khách hàng
profession	Lĩnh vực của khách hàng
married	Tình trạng hôn nhân của khách hàng
house_ownership	Sở hữu nhà hay thuê nhà hay không có
car_ownership	Có sở hữu xe hơi hay không
risk_flag	Có lịch sử trễ hạn trả nợ (1: Yes, 0: No)
current_job_years	Số năm kinh nghiệm trong công việc hiện tại
current_house_years	Số năm ở nơi cư trú hiện tại
city	Thành phố cư trú của khách hàng
State	Bang cư trú của khách hàng

1.2.3 Giới thiệu các công cụ được sử dụng trong đồ án

- Visual Studio 2019 Community: Tích hợp các công nghệ
 - SQL Server Integration Services (SSIS)
 - SQL Server Analysis Services (SSAS)
- SQL Server 2019 Developer
- Ngôn ngữ lập trình: Python3
- IDE sử dụng: Visual Studio Code

1.2.3.1 Tổng quan về Visual Studio 2019 Community

Visual Studio 2019 Community là một phiên bản miễn phí của môi trường phát triển tích hợp (Integrated Development Environment - IDE) Visual Studio 2019. Nó cung cấp một loạt các công cụ và tài nguyên để phát triển ứng dụng trên nhiều nền tảng, bao gồm ứng dụng di động, desktop và web.

Dưới đây là một số điểm nổi bật về Visual Studio 2019 Community:

- Tích hợp đa nền tảng: Visual Studio 2019 Community hỗ trợ phát triển ứng dụng trên nhiều nền tảng, bao gồm Windows, Android, iOS và web. Điều này cho phép bạn xây dựng ứng dụng đa nền tảng và tái sử dụng mã nguồn dễ dàng.
- Ngôn ngữ lập trình đa dạng: Visual Studio 2019 Community hỗ trợ nhiều ngôn ngữ lập trình, bao gồm C#, Visual Basic, C++, Python, JavaScript và nhiều ngôn ngữ khác. Điều này cho phép bạn lựa chọn ngôn ngữ phù hợp với dự án của mình.
- Công cụ và trình biên dịch mạnh mẽ: Visual Studio 2019 Community cung cấp nhiều công cụ và trình biên dịch tiên tiến để hỗ trợ quá trình phát triển. Bạn có thể sử dụng trình biên dịch thông minh IntelliSense, gỡ lỗi mạnh mẽ, kiểm tra mã, quản lý phiên bản và nhiều tính năng khác để tăng năng suất và chất lượng phát triển.
- Tích hợp công cụ và dịch vụ của Microsoft: Visual Studio 2019 Community tích hợp với các công cụ và dịch vụ của Microsoft như Azure, SQL Server, Office và nền tảng điện toán đám mây Microsoft Azure. Điều này giúp bạn dễ dàng tích hợp ứng dụng của mình với các dịch vụ và nền tảng của Microsoft.
- Cộng đồng lớn và hỗ trợ: Visual Studio 2019 Community có một cộng đồng sôi nổi và hỗ trợ rộng rãi từ cộng đồng phát triển phần mềm. Bạn có thể tìm kiếm thông tin, hỏi đáp và chia sẻ kiến thức với cộng đồng để giúp giải quyết các vấn đề trong quá trình phát triển.

Visual Studio 2019 Community là một lựa chọn phổ biến cho các nhà phát triển ứng dụng, sinh viên và nhóm phát triển nhỏ, cho phép họ phát triển các ứng dụng đa nền tảng một cách dễ dàng và hiệu quả.

1.2.3.2 Tổng quan về SQL Server 2019 Developer

SQL Server 2019 Developer là một phiên bản của hệ quản trị cơ sở dữ liệu SQL Server 2019 do Microsoft phát triển. Nó là phiên bản dành riêng cho các nhà phát triển phần mềm và cung cấp nhiều tính năng và công cụ mạnh mẽ để phát triển, thử nghiệm và triển khai ứng dụng dựa trên cơ sở dữ liệu SQL Server.

Dưới đây là một số điểm nổi bật về SQL Server 2019 Developer:

- **Quản lý cơ sở dữ liệu:** SQL Server 2019 Developer cung cấp các tính năng quản lý cơ sở dữ liệu như tạo, cập nhật và xóa cơ sở dữ liệu, bảng, quyền truy cập và các đối tượng khác. Bạn có thể sử dụng Transact-SQL để thao tác với cơ sở dữ liệu.
- **Tính năng cao cấp:** SQL Server 2019 Developer bao gồm các tính năng cao cấp như Replication, Always On Availability Groups, Transparent Data Encryption và Dynamic Data Masking. Điều này cho phép bạn xây dựng các ứng dụng có tính bảo mật, khả năng chịu lỗi và khả năng mở rộng cao.
- **Hỗ trợ đa nền tảng:** SQL Server 2019 Developer hỗ trợ triển khai cơ sở dữ liệu trên nhiều nền tảng, bao gồm Windows, Linux và Docker. Điều này cho phép bạn triển khai ứng dụng dựa trên SQL Server trên nền tảng mà bạn chọn.
- **Tích hợp với các công cụ phát triển:** SQL Server 2019 Developer tích hợp tốt với các công cụ phát triển phổ biến như Visual Studio và Azure DevOps. Bạn có thể sử dụng công cụ này để phát triển, gỡ lỗi và triển khai ứng dụng liên quan đến cơ sở dữ liệu SQL Server.
- **Cộng đồng hỗ trợ và tài liệu:** SQL Server 2019 Developer có sự hỗ trợ từ cộng đồng phát triển phần mềm và Microsoft. Bạn có thể tìm kiếm thông tin, tham gia diễn đàn và sử dụng các tài liệu hướng dẫn để giải quyết các vấn đề và tối ưu hóa sử dụng SQL Server.

SQL Server 2019 Developer là một công cụ mạnh mẽ và linh hoạt để phát triển và quản lý cơ sở dữ liệu SQL Server. Nó cung cấp nhiều tính năng và tùy chọn linh hoạt để phù hợp với nhu cầu phát triển ứng dụng của bạn.

1.2.3.3 Tổng quan về Visual Studio Code

Visual Studio Code là một trình biên tập mã nguồn mở và miễn phí, được phát triển bởi Microsoft. Đây là một trong những trình biên tập mã nguồn phổ biến nhất hiện nay, được sử dụng rộng rãi bởi các nhà phát triển phần mềm, các chuyên gia IT, các nhà khoa học dữ liệu và nhiều người khác. Visual Studio Code hỗ trợ đa ngôn ngữ, bao gồm C++, C#, Python, JavaScript, HTML, CSS và nhiều ngôn ngữ lập trình khác.

Visual Studio Code có nhiều tính năng hữu ích cho các nhà phát triển, bao gồm:

- **IntelliSense:** đây là tính năng tự động hoàn thành mã, giúp giảm thiểu thời gian gõ code.
- **Debugging:** tính năng này giúp các nhà phát triển dễ dàng tìm ra lỗi trong mã nguồn và sửa chúng.
- **Extensions:** Visual Studio Code hỗ trợ nhiều extensions, cho phép người dùng thêm tính năng và chức năng vào trình biên tập.
- **Version Control:** Visual Studio Code tích hợp với các hệ thống quản lý phiên bản phổ biến như Git, SVN và Mercurial.
- **Terminal:** tính năng này cho phép người dùng truy cập vào dòng lệnh trực tiếp từ trình biên tập.

Visual Studio Code là một công cụ đáng tin cậy cho các nhà phát triển phần mềm và là một phần quan trọng của quy trình phát triển phần mềm hiện đại.

1.2.3.4 Giới thiệu về ngôn ngữ lập trình python3

Python là một ngôn ngữ lập trình đa mục đích, được phát triển bởi Guido van Rossum vào năm 1991. Python có cú pháp đơn giản, dễ đọc và dễ hiểu, và được sử dụng rộng rãi trong nhiều lĩnh vực, bao gồm khoa học dữ liệu, máy học, phát triển web, đồ họa, tự động hóa và nhiều ứng dụng khác.

Python hỗ trợ nhiều phong cách lập trình, bao gồm lập trình hướng đối tượng, lập trình hàm, lập trình thủ tục và lập trình lập trình logic. Nó cũng cung cấp nhiều thư viện và framework mạnh mẽ để giải quyết các vấn đề phức tạp, bao gồm NumPy, Pandas, Matplotlib, Scikit-learn, TensorFlow và Django.

Python được sử dụng phổ biến trong khoa học dữ liệu và máy học do tính linh hoạt và sức mạnh của nó trong việc Xử lý dữ liệu và phân tích. Với Python, người dùng có thể dễ dàng thực hiện các nhiệm vụ phức tạp như thu thập dữ liệu, chuẩn bị dữ liệu, phân tích dữ liệu và đưa ra dự đoán.

Python là một ngôn ngữ lập trình được sử dụng rộng rãi và phát triển nhanh chóng. Vì vậy, nó có một cộng đồng lớn, phong phú và nhiều tài nguyên cho những người mới bắt đầu và những người đã có kinh nghiệm.

Chương 2: Cơ sở lý thuyết

2.1 Kỹ thuật Naive Bayes

Naive Bayes là một phương pháp phân loại trong machine learning dựa trên giả định độc lập giữa các đặc trưng của dữ liệu. Nó dựa trên định lý Bayes để tính xác suất của một mẫu thuộc về một lớp nhất định dựa trên các đặc trưng của mẫu đó.

Giả định độc lập trong Naive Bayes giả định rằng các đặc trưng của một mẫu là độc lập với nhau khi đã biết lớp của mẫu đó. Mặc dù giả định này không luôn đúng trong thực tế, nhưng Naive Bayes vẫn cho kết quả tốt và được sử dụng rộng rãi trong các bài toán phân loại.

Giải thuật Naive Bayes có thể được áp dụng cho các loại dữ liệu khác nhau, bao gồm dữ liệu rời rạc (discrete) và dữ liệu liên tục (continuous). Có ba biến thể chính của Naive Bayes: Naive Bayes đa thức (Multinomial Naive Bayes) cho dữ liệu rời rạc đếm số lần xuất hiện, Naive Bayes đa tín hiệu (Gaussian Naive Bayes) cho dữ liệu liên tục tuân theo phân phối Gaussian, và Naive Bayes đa biến thức (Bernoulli Naive Bayes) cho dữ liệu nhị phân.

Mặc dù giải thuật Naive Bayes có nhược điểm trong việc giả định độc lập giữa các đặc trưng, nhưng nó thường cho hiệu suất tốt và tốn ít thời gian huấn luyện và dự đoán so với các giải thuật phức tạp hơn.

2.2 Kỹ thuật Decision Tree

Giải thuật Decision Tree (cây quyết định) là một thuật toán học máy được sử dụng cho các bài toán phân loại và dự đoán. Nó tạo ra một cấu trúc cây quyết định dựa trên các quyết định và phân nhánh dựa trên các đặc trưng của dữ liệu.

Cách hoạt động của giải thuật Decision Tree như sau:

1. Ban đầu, cây quyết định được xây dựng từ một tập dữ liệu huấn luyện. Quá trình xây dựng bắt đầu bằng việc chọn đặc trưng quan trọng nhất để phân chia tập dữ liệu thành các nhánh.
2. Đặc trưng được chọn sẽ tạo thành một nút quyết định trên cây. Dữ liệu được chia thành các nhánh con tương ứng với các giá trị của đặc trưng này.
3. Quá trình phân chia được tiếp tục trên các nhánh con, bằng cách chọn đặc trưng tiếp theo tốt nhất để tiếp tục phân chia.
4. Quá trình xây dựng cây tiếp tục cho đến khi một điều kiện dừng được đáp ứng, ví dụ như khi tất cả các điểm dữ liệu trong một nhánh đều thuộc cùng một lớp hoặc khi đạt đến một độ sâu tối đa đã được định nghĩa trước.

5. Khi cây quyết định hoàn thành, nó có thể được sử dụng để phân loại các điểm dữ liệu mới bằng cách đi từ gốc đến lá của cây dựa trên các quyết định đã học được.

Các đặc điểm của giải thuật Decision Tree bao gồm:

- Dễ hiểu và diễn giải: Cây quyết định tạo ra một biểu đồ cây dễ hiểu và diễn giải được, giúp người dùng hiểu cách quyết định được thực hiện.
- Xử lý dữ liệu số và rời rạc: Decision Tree có thể xử lý cả dữ liệu số và rời rạc mà không cần quá nhiều tiền xử lý dữ liệu.
- Dễ bị overfitting: Decision Tree có xu hướng tạo ra cây phức tạp và dễ bị overfitting trên dữ liệu huấn luyện. Các biện pháp như cắt tỉa (pruning) có thể được áp dụng để giảm overfitting.
- Nhạy cảm với nhiễu: Decision Tree có thể bị ảnh hưởng bởi nhiễu trong dữ liệu, đặc biệt khi sử dụng các độ đo không tốt cho việc chọn đặc trưng.

Giải thuật Decision Tree được sử dụng rộng rãi trong nhiều lĩnh vực, bao gồm phân loại văn bản, dự đoán chuỗi thời gian, phát hiện gian lận, và hỗ trợ ra quyết định trong quản lý và kinh doanh.

2.3 Kỹ thuật K-Means

Giải thuật K-means là một thuật toán phân cụm (clustering) trong lĩnh vực học máy và khai phá dữ liệu. Nó được sử dụng để phân nhóm các điểm dữ liệu thành các cụm (cluster) dựa trên đặc trưng của chúng.

Cách hoạt động của giải thuật K-means như sau:

1. Chọn số lượng cụm K muốn phân chia.
2. Chọn K điểm dữ liệu ngẫu nhiên làm các trung tâm ban đầu của các cụm.
3. Gán mỗi điểm dữ liệu vào cụm có trung tâm gần nhất (sử dụng khoảng cách Euclid).

4. Tính toán lại trung tâm của mỗi cụm bằng cách lấy trung bình của tất cả các điểm dữ liệu thuộc cụm đó.
5. Lặp lại các bước 3 và 4 cho đến khi không có sự thay đổi nào trong việc gán các điểm dữ liệu vào các cụm.
6. Kết quả cuối cùng là các cụm dữ liệu đã được phân chia.

Các đặc điểm của giải thuật K-means bao gồm:

- Phương pháp phân cụm dựa trên không gian: K-means phân cụm dựa trên khoảng cách Euclidean giữa các điểm dữ liệu. Nó cố gắng tối thiểu hóa tổng bình phương khoảng cách giữa các điểm dữ liệu và điểm trung tâm của cụm tương ứng.
- Phụ thuộc vào số lượng cụm: Kết quả của K-means phụ thuộc vào số lượng cụm K được xác định trước. Một số phương pháp có thể được sử dụng để chọn số lượng cụm tối ưu, chẳng hạn như phân tích độ biến thiên (elbow method) hoặc phương pháp Silhouette.
- Nhạy cảm với vị trí ban đầu: Kết quả của K-means có thể khác nhau dựa trên vị trí ban đầu của các điểm trung tâm. Do đó, quá trình chạy K-means có thể được lặp lại nhiều lần với các vị trí ban đầu khác nhau để kiểm tra tính ổn định và đạt được kết quả tốt nhất.

Giải thuật K-means được sử dụng rộng rãi trong nhiều ứng dụng, bao gồm phân loại ảnh, phân nhóm khách hàng, nhận dạng ngôn ngữ, và nhiều lĩnh vực khác trong khai phá dữ liệu và học máy.

2.4 Kỹ thuật Apriori

Giải thuật Apriori là một thuật toán được sử dụng trong khai phá dữ liệu để tìm các luật kết hợp (association rules) giữa các mục (items) trong tập dữ liệu. Đặc biệt, nó tìm kiếm các luật kết hợp có sự xuất hiện chung giữa các mục với một ngưỡng minh họa (min-support) và một ngưỡng minh họa cho luật (min-confidence).

Cách hoạt động của giải thuật Apriori như sau:

1. Đầu tiên, giải thuật xác định tất cả các mục đơn lẻ có mặt trong tập dữ liệu và tính toán sự xuất hiện (support) của chúng, tức là tần suất xuất hiện của mỗi mục trong tập dữ liệu.
2. Tiếp theo, giải thuật tạo ra các tập ứng viên (candidate sets) của các luật kết hợp. Ban đầu, các tập ứng viên chỉ chứa các mục đơn lẻ. Sau đó, giải thuật kiểm tra sự xuất hiện (support) của các tập ứng viên này trong tập dữ liệu.
3. Sau khi có các tập ứng viên và sự xuất hiện của chúng, các tập ứng viên có sự xuất hiện lớn hơn hoặc bằng ngưỡng min-support được lựa chọn để trở thành tập mục phổ biến (frequent itemsets).
4. Tiếp theo, giải thuật tạo ra các luật kết hợp từ các tập mục phổ biến đã tìm thấy. Đối với mỗi tập mục phổ biến, giải thuật tạo ra tất cả các tập con của nó và kiểm tra sự xuất hiện của các tập con này trong tập dữ liệu.
5. Các luật kết hợp được xác định dựa trên sự xuất hiện (support) của các tập con và các tập mục phổ biến. Các luật có sự xuất hiện lớn hơn hoặc bằng ngưỡng min-confidence được chọn làm luật kết hợp cuối cùng.
6. Kết quả cuối cùng là các luật kết hợp được tìm thấy trong tập dữ liệu.

Các đặc điểm của giải thuật Apriori bao gồm:

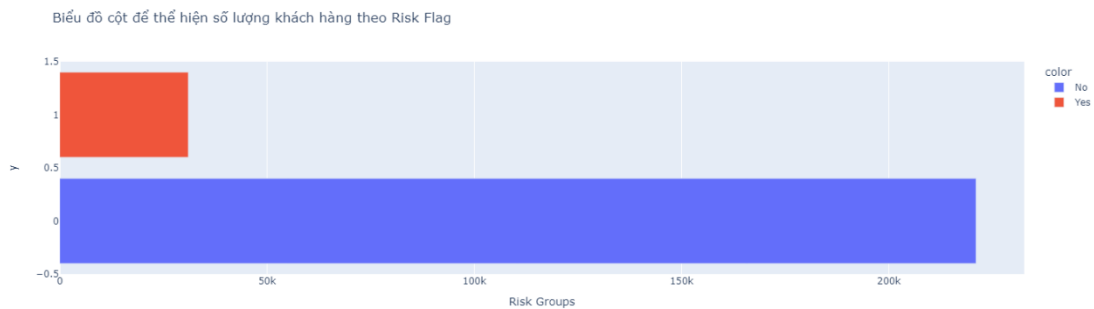
- Dễ triển khai: Giải thuật Apriori dễ triển khai và có hiệu suất tốt đối với tập dữ liệu nhỏ và vừa.
- Đòi hỏi không gian lưu trữ lớn: Apriori yêu cầu không gian lưu trữ lớn để lưu trữ và xử lý các tập mục phổ biến và các luật kết hợp.
- Nhạy cảm với số lượng mục: Khi số lượng mục tăng lên, giải thuật Apriori trở nên không hiệu quả do số lượng tập ứng viên tăng một cách kết quả.
- Phụ thuộc vào ngưỡng min-support và min-confidence: Kết quả của giải thuật Apriori phụ thuộc vào việc chọn các ngưỡng min-support và min-confidence phù hợp.

Giải thuật Apriori được sử dụng rộng rãi trong việc phân tích hành vi người dùng, gợi ý sản phẩm, phân loại các mẫu trong dữ liệu bán hàng và nhiều ứng dụng khác trong lĩnh vực khai phá dữ liệu.

Chương 3: Phân tích khám phá dữ liệu (EDA)

Trực quan hóa dữ liệu là một phần quan trọng của Khai thác dữ liệu. Nó giúp chúng ta có được hình ảnh thông tin chi tiết về tập dữ liệu, chẳng hạn như một số tính năng hoặc mẫu nổi bật trong tập dữ liệu, có thể giúp chúng ta chọn các thuật toán phù hợp để áp dụng.

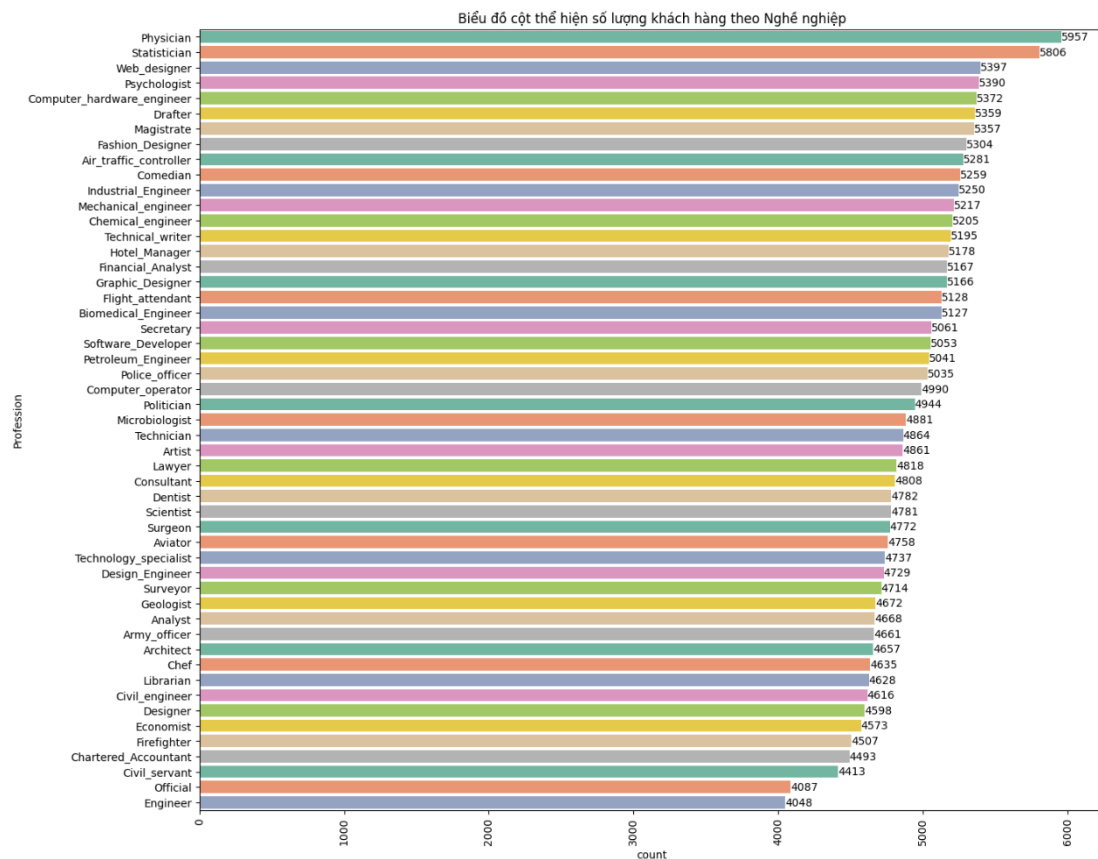
3.1 Biểu đồ cột để phân tích thống kê cột Risk Flag



Kết luận:

- Ta có thể thấy được tập dữ liệu khá mất cân bằng khi tỉ lệ giữa những khách hàng không có lịch sử trễ hạn trả nợ chiếm đa số trong tập dữ liệu.
- Điều này có thể dẫn đến việc kết quả có thể không chính xác khi chúng ta chạy các thuật toán dự đoán trên tập dữ liệu mất cân bằng.

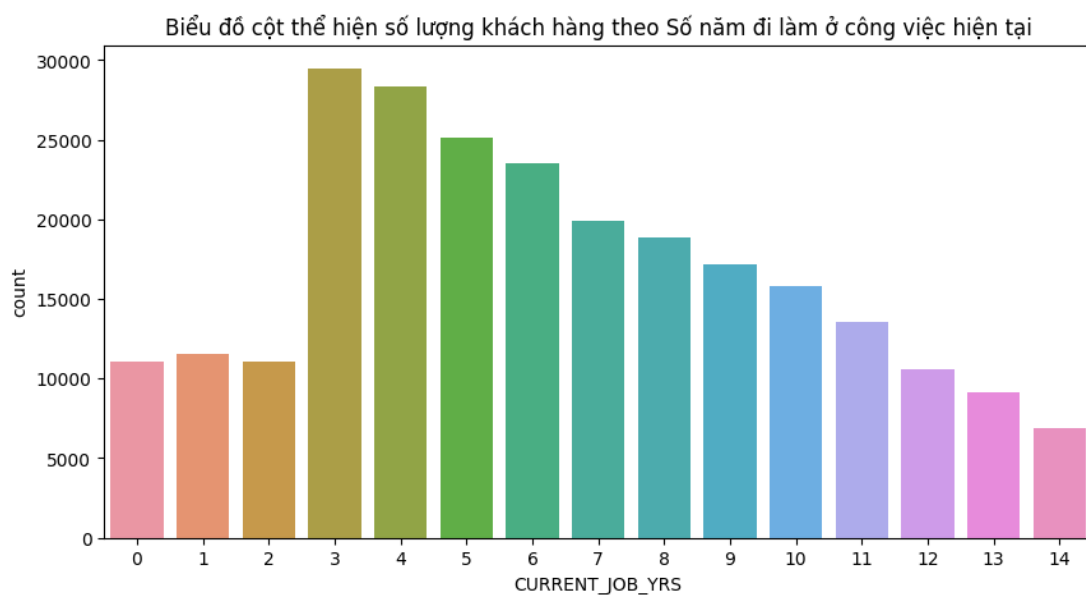
3.2 Biểu đồ cột để phân tích thống kê cột Profession



Kết luận:

- Nghề "Bác sĩ" là nghề phổ biến nhất, theo sát là nghề "Nhà thống kê".
- Đáng ngạc nhiên là Nghề ít được chọn nhất là Nghề "Kỹ sư".

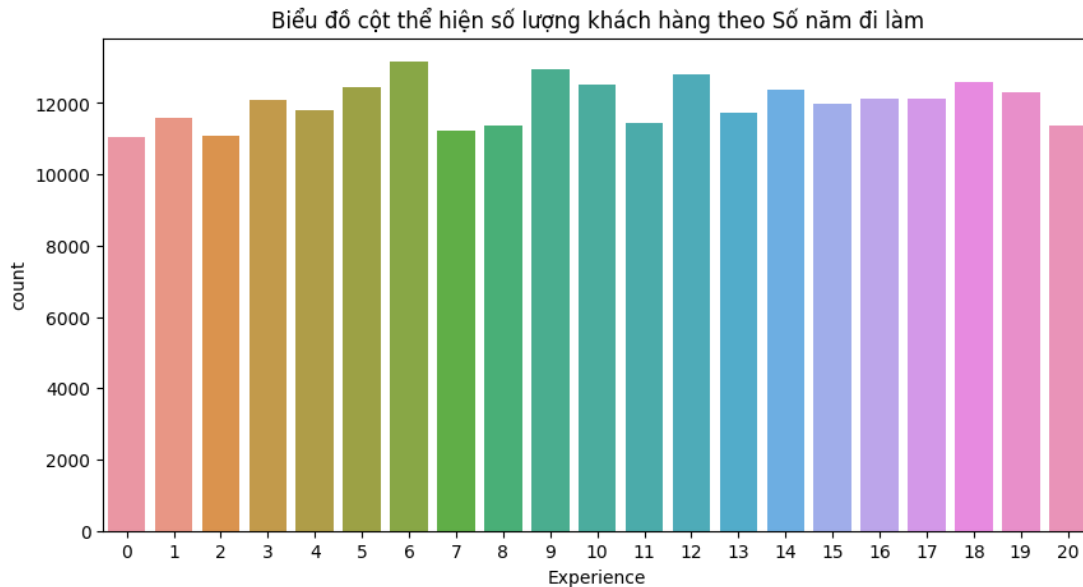
3.3 Biểu đồ cột để phân tích thống kê cột Current Job Year



Kết luận:

- Ta có thể thấy được tập khách hàng tập trung nhiều nhất là những người hiện đang làm việc từ 3 năm và giảm dần đến năm thứ 14.

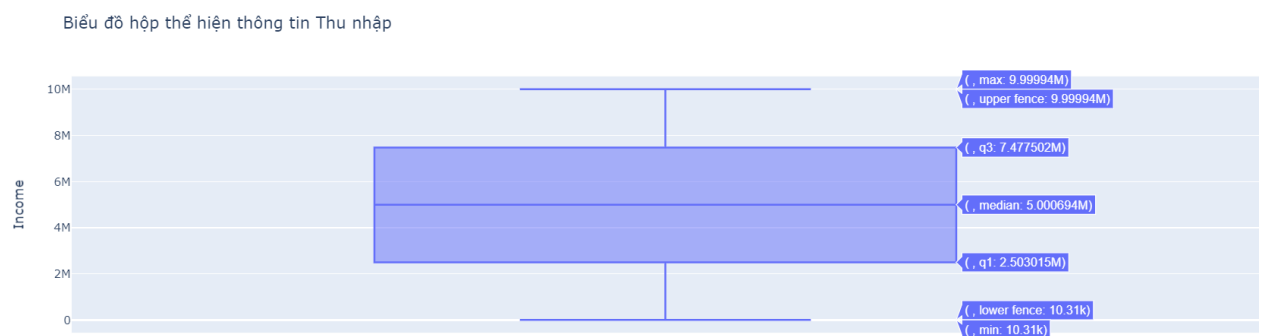
3.4 Biểu đồ cột để phân tích thống kê cột Experience



Kết luận:

- Ta có thể thấy được tập dữ liệu trải dài khá đồng đều, không tập trung rõ rệt ở bao nhiêu năm.
- Chúng ta dù đi làm nhiều hay ít ai cũng sẽ vay ngân hàng.

3.5 Biểu đồ hộp để phân tích thống kê cột Income

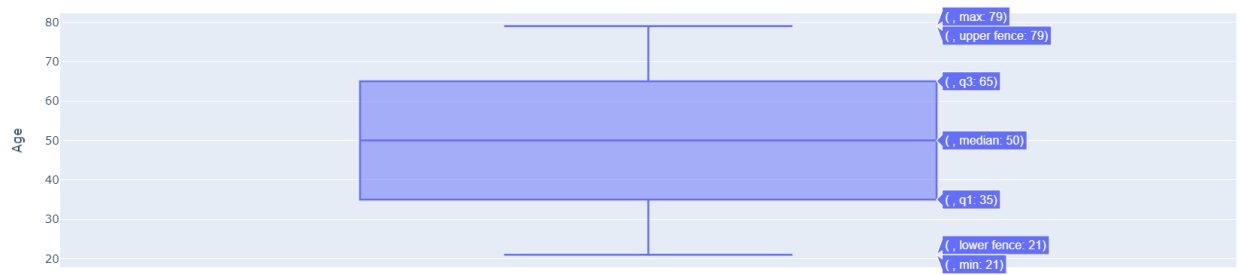


Kết luận:

- Thu nhập trung bình 5 triệu, thu nhập cao nhất: 9.99 triệu, thu nhập thấp nhất: 10,3 ngàn
- Không có outlier.

3.6 Biểu đồ hộp để phân tích thống kê cột Age

Biểu đồ hộp thể hiện thông tin Độ tuổi

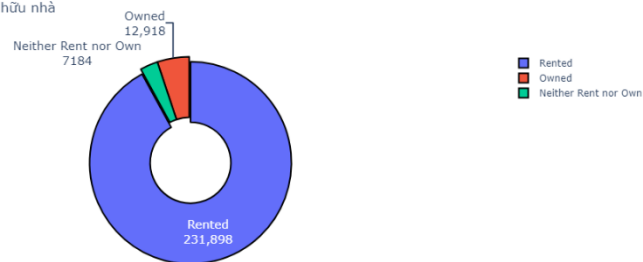


Kết luận:

- Tuổi trung bình là khoảng 50.
- Tuổi tối đa: 79, Tuổi tối thiểu: 21.

3.7 Biểu đồ tròn để phân tích thống kê cột House Ownership

Biểu đồ tròn thể hiện số lượng khách hàng theo Sở hữu nhà



Kết luận:

- Khoảng 92% toàn bộ khách hàng thuê Nhà, chiếm khoảng 232 nghìn người.
- Thuê một ngôi nhà dường như là một lựa chọn rõ ràng hơn là sở hữu một ngôi nhà, chiếm gần 5,13% toàn bộ khách hàng.
- Chỉ 2,85% tổng khách hàng không sở hữu Nhà cũng như không thuê.

3.8 Biểu đồ tròn để phân tích thống kê cột Car Ownership

Biểu đồ tròn thể hiện số lượng khách hàng theo Sở xe hơi



Kết luận:

- Khoảng 69,8% khách hàng không sở hữu Ô tô, chiếm khoảng 176 nghìn người.
- Ngược lại, khoảng 30,2% khách hàng sở hữu một chiếc Ô tô.

3.9 Biểu đồ tròn để phân tích thống kê cột Marital Status

Biểu đồ tròn thể hiện số lượng khách hàng theo Tình trạng hôn nhân



Kết luận:

- Khoảng 89,8% khách hàng là "Độc thân", chiếm khoảng 226 nghìn người.
- Ngược lại, chỉ có 10,2% khách hàng là "Đã kết hôn".

3.10 Biểu đồ tròn để phân tích thống kê cột State

Biểu đồ tròn thể hiện số lượng khách hàng theo Bang cư trú



Kết luận:

- Bang chiếm tỷ lệ khách hàng tối đa trong bộ dữ liệu này là "Uttar Pradesh", tiếp theo là "Maharashtra" và "Andhra Pradesh".
- "Sikkim" có ít khách hàng nhất trong tập dữ liệu này, chỉ chiếm khoảng 0,24% toàn bộ tập dữ liệu.

Chương 4: Tiền xử lý dữ liệu và xây dựng mô hình

4.1 Tiền xử lý dữ liệu

a) Đọc dữ liệu

```
# Đọc dữ liệu
df = pd.read_csv('./Training Data.csv')
```

b) Thay đổi thuộc tính

```
# Thay đổi tên thuộc tính
df.columns = df.columns.str.lower()
df.rename(columns={"married/single": "married_single"}, inplace=True)

# Loại bỏ thuộc tính 'city'
df = df.drop('city', axis = 1)

# Chuyển kiểu dữ liệu của thuộc tính risk_flag từ 1, 0 thành 'Yes', 'No'
df["risk_flag"] = df["risk_flag"].map({0: 'No', 1: 'Yes'})
```

c) Cân bằng tập dữ liệu

```
df = df.sample(frac=1)

risk_data = df.loc[df["risk_flag"] == 'Yes']
not_risk_data = df.loc[df["risk_flag"] == 'No'][:30996]

normal_distributed_data = pd.concat([risk_data, not_risk_data])

loan = normal_distributed_data.sample(frac=1, random_state=42)
```

c) Lưu tập dữ liệu

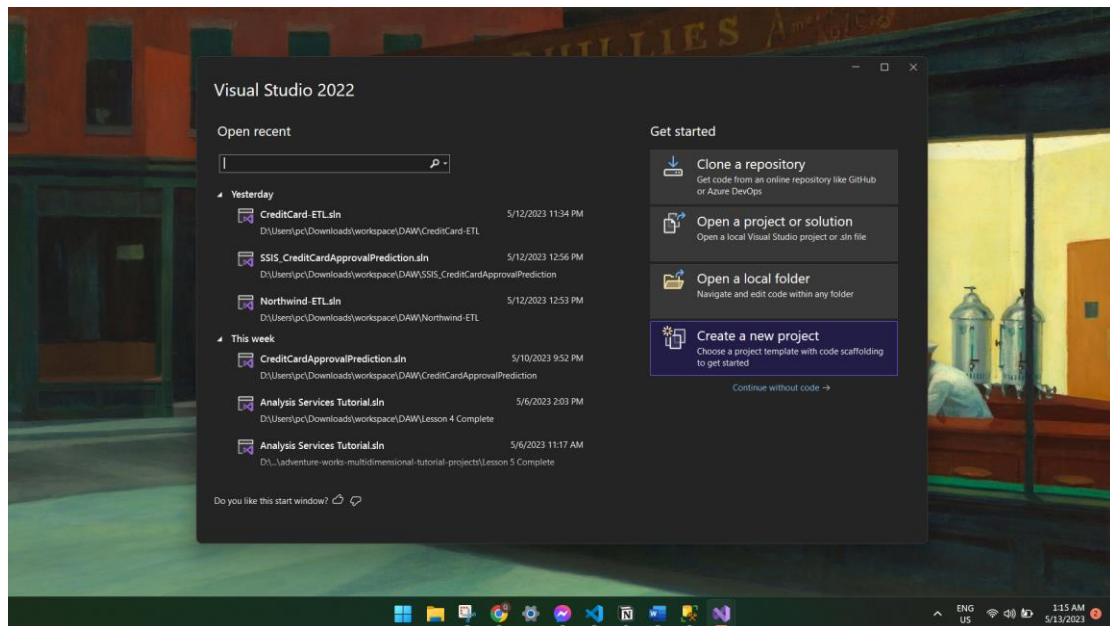
```
# Lưu dữ liệu
loan.to_csv('data.csv', index=False)
```

4.2 Xây dựng mô hình

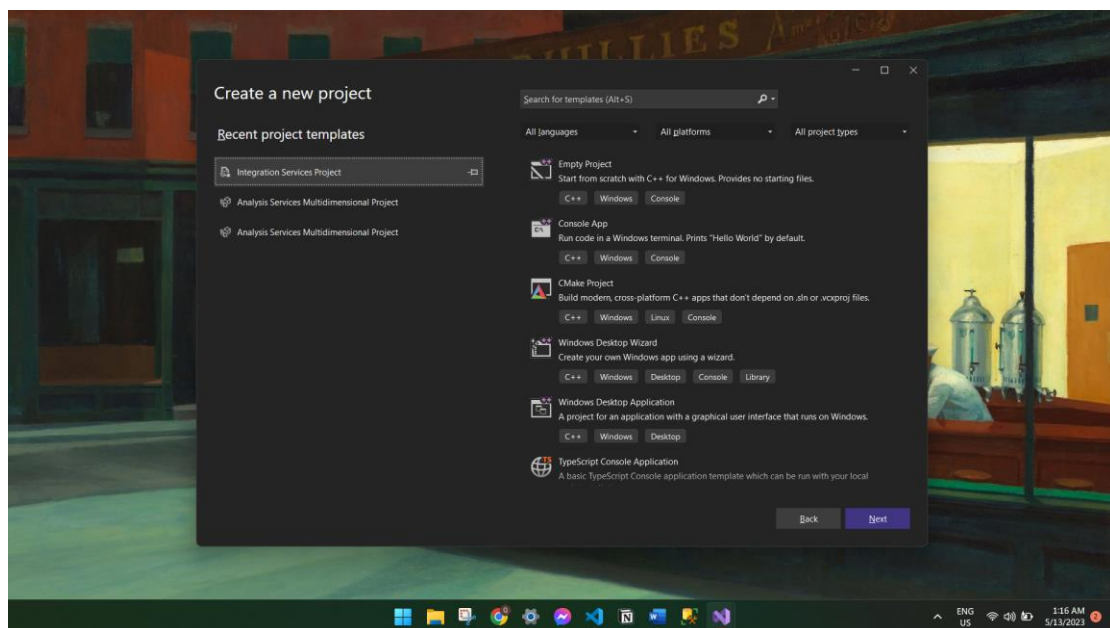
4.2.1 Xây dựng cơ sở dữ liệu bằng công cụ SSIS

4.2.1.1 Quá trình tạo mới project SSIS

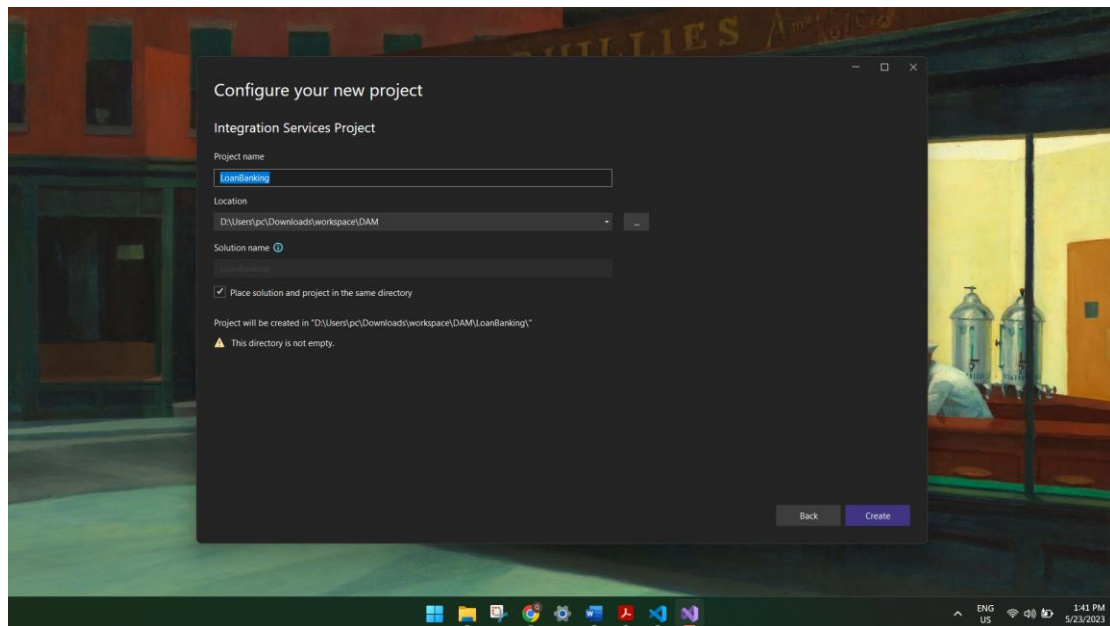
Mở Visual Studio 2019 -> Chọn Create New Project.



Chọn **Intergration Services Project**.



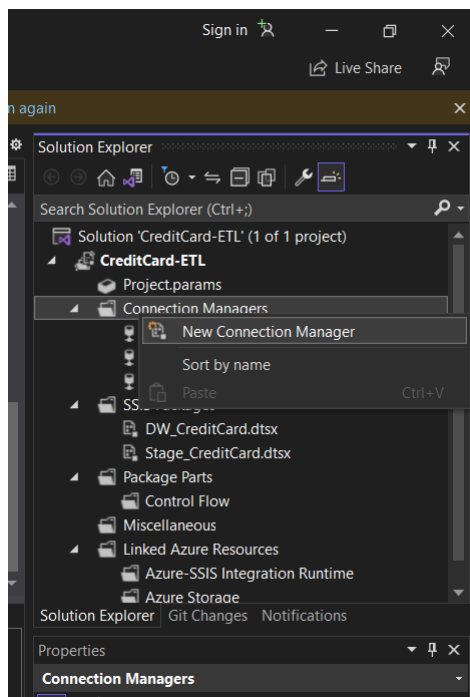
Đặt tên project **CreditCard-ETL** -> Ấn nút **Create**.



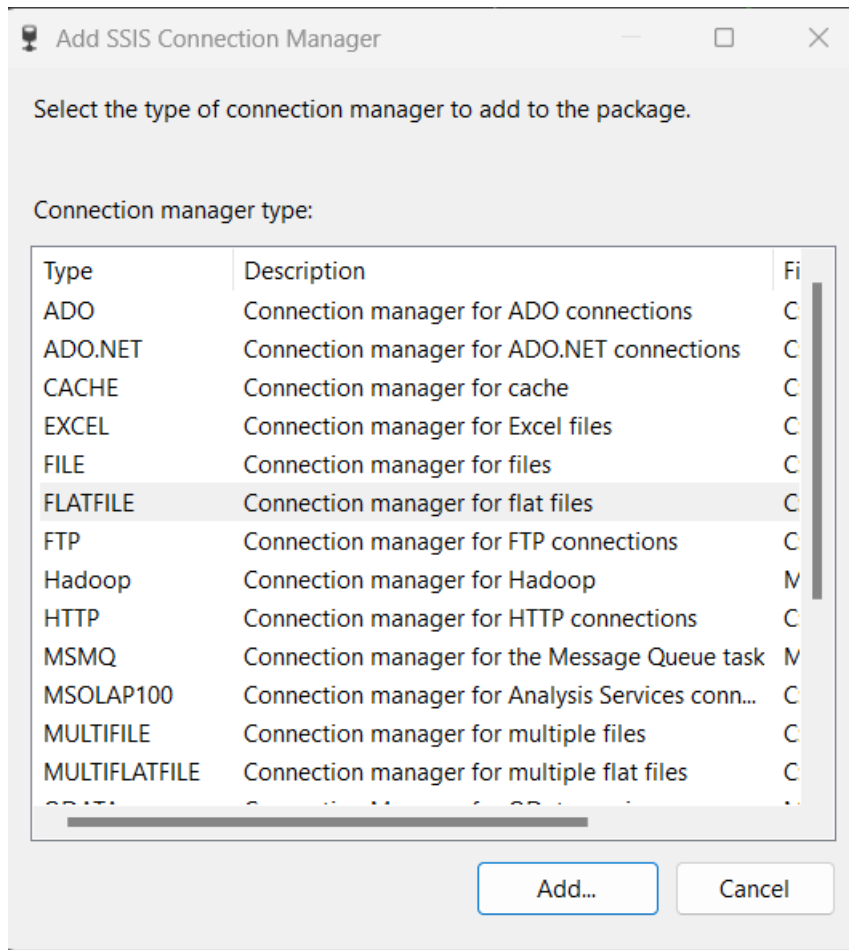
4.2.1.2 Quá trình đổ dữ liệu từ excel vào database

4.2.1.2.1 Quá trình tạo Flat File Connection Manager

Chuột phải vào **Connection Managers** -> **New Connection Managers**.



Chọn **FLATFILE** -> Ấn nút **ADD**.



Nhập đường dẫn đến **File name** -> Click chọn **Columns** -> Ấn nút **OK**.

Flat File Connection Manager Editor

Connection manager name: Flat File Connection Manager 1

Description:

General
Columns
Advanced
Preview

Select a file and specify the file properties and the file format.

File name: C:\Downloads\DH\DAM\DoAn\data.csv Browse...

Locale: English (United States) ☐ Unicode

Code page: 1252 (ANSI - Latin I)

Format: Delimited

Text qualifier: <none>

Header row delimiter: {CR}{LF}

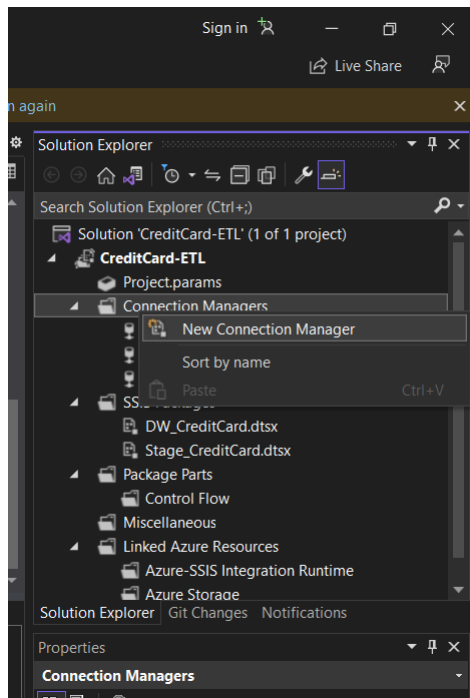
Header rows to skip: 0

☒ Column names in the first data row

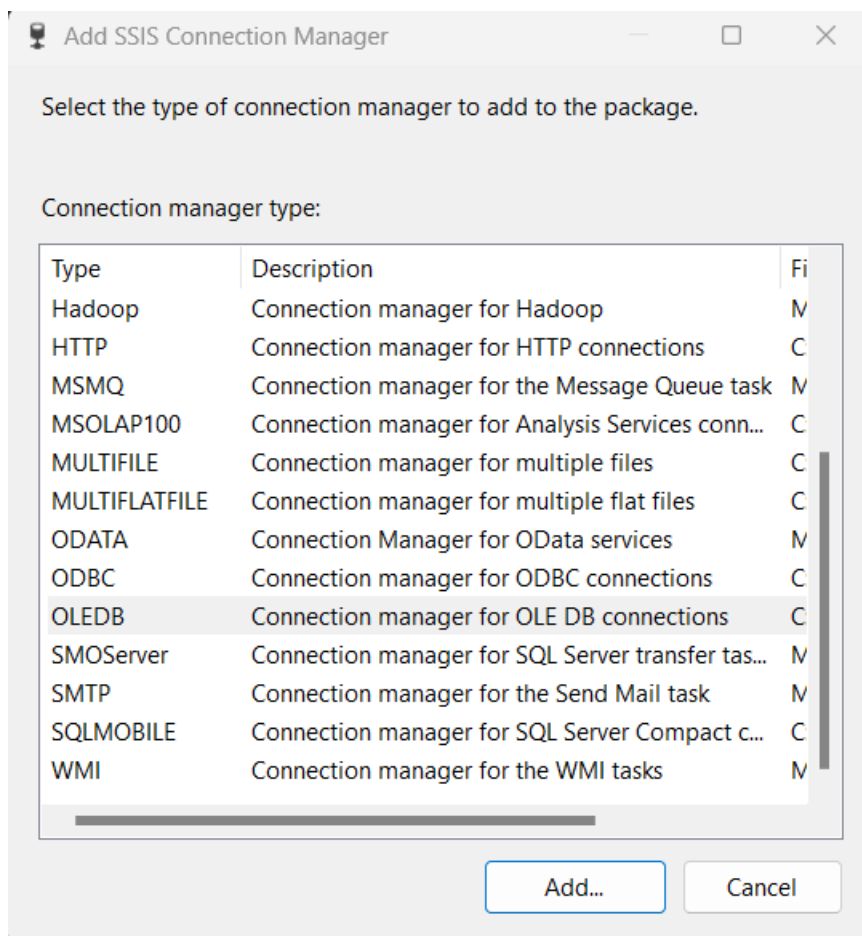
OK Cancel Help

4.2.1.2.2 Quá trình tạo OLEDB Connection Manager

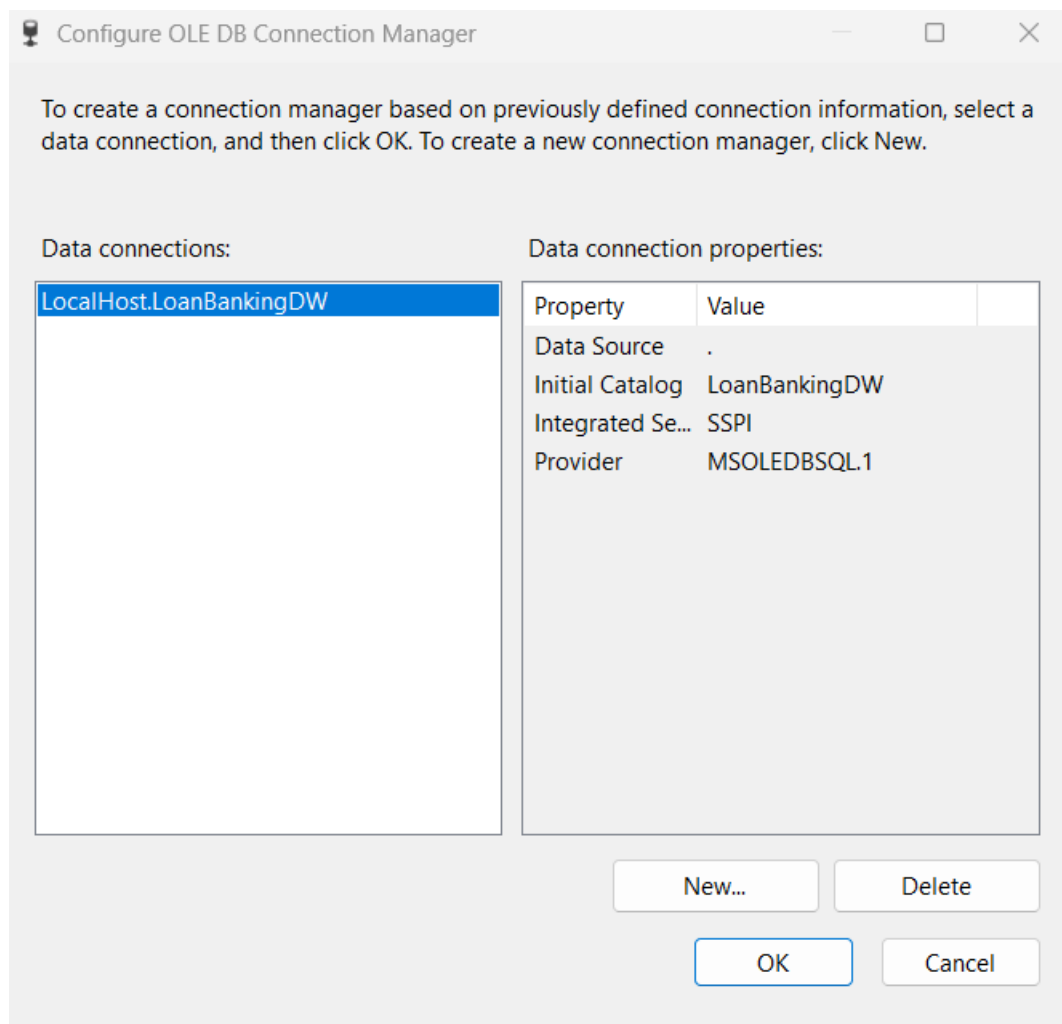
Chuột phải vào **Connection Managers** -> **New Connection Managers**.



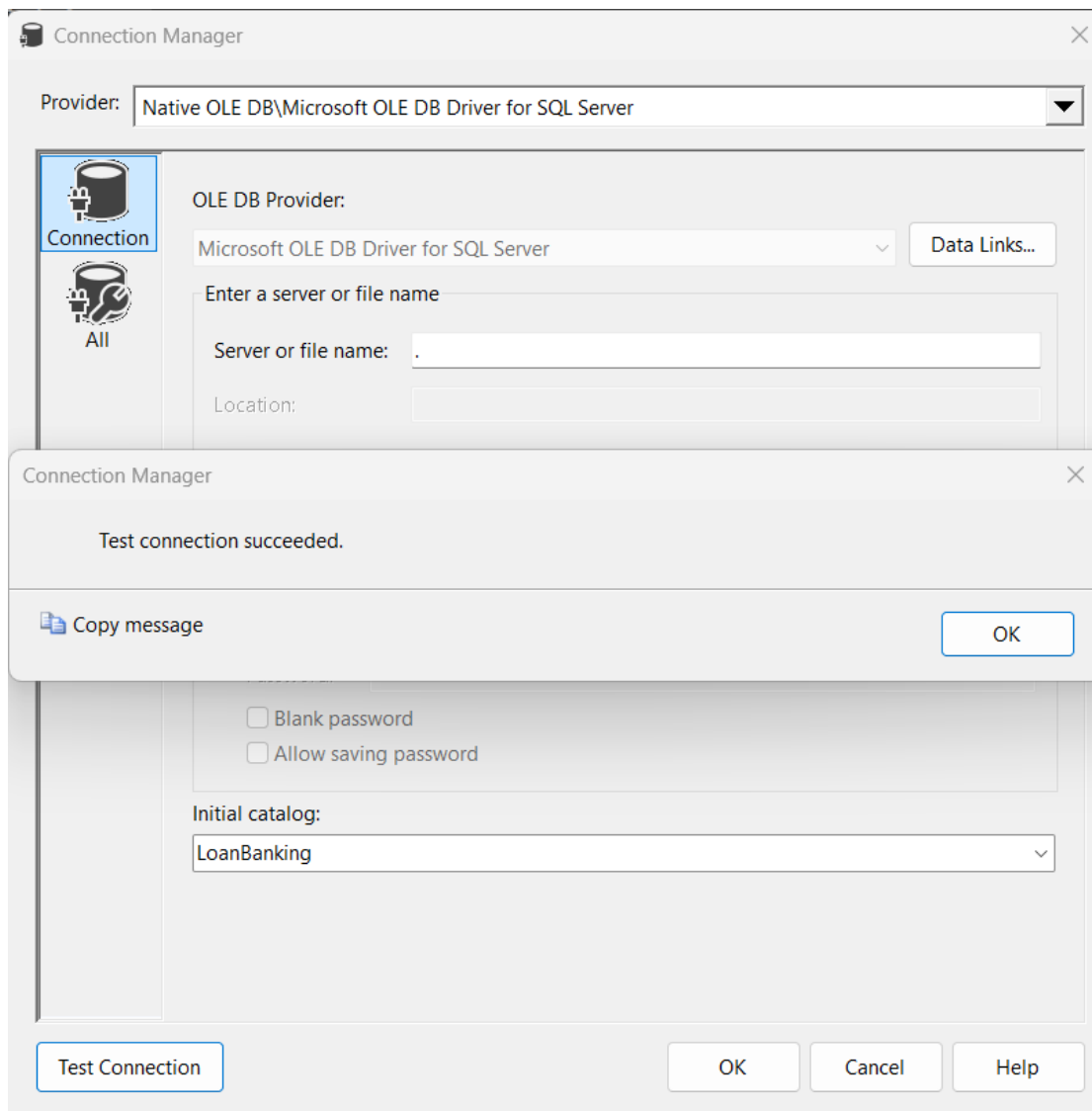
Chọn **OleDb** -> Ấn nút **ADD**.



Chọn **New**.

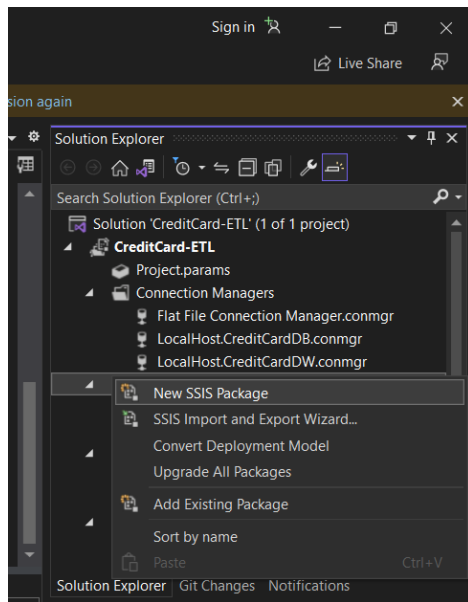


Provider chọn **Native OLE DB\Microsoft OLE DB Driver for SQL**
Server -> **Server or file name** gõ '.' để mặc định là localhost -> Tạo
LoanBanking database bên SQL Server -> **Initial Catalog** chọn
CreditCardDB -> Ấn **Test Connection**.

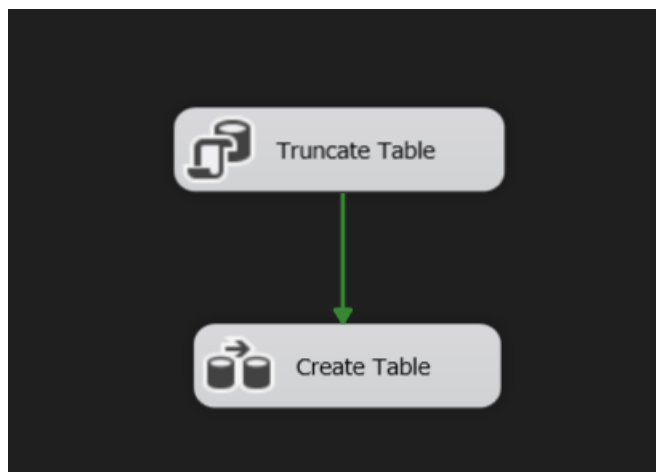


4.2.1.2.3 Quá trình tạo SSIS Package

Chọn **SSIS Package** -> **New SSIS Package** -> Đổi tên package thành **Stage_LoanBaking.dtsx**.

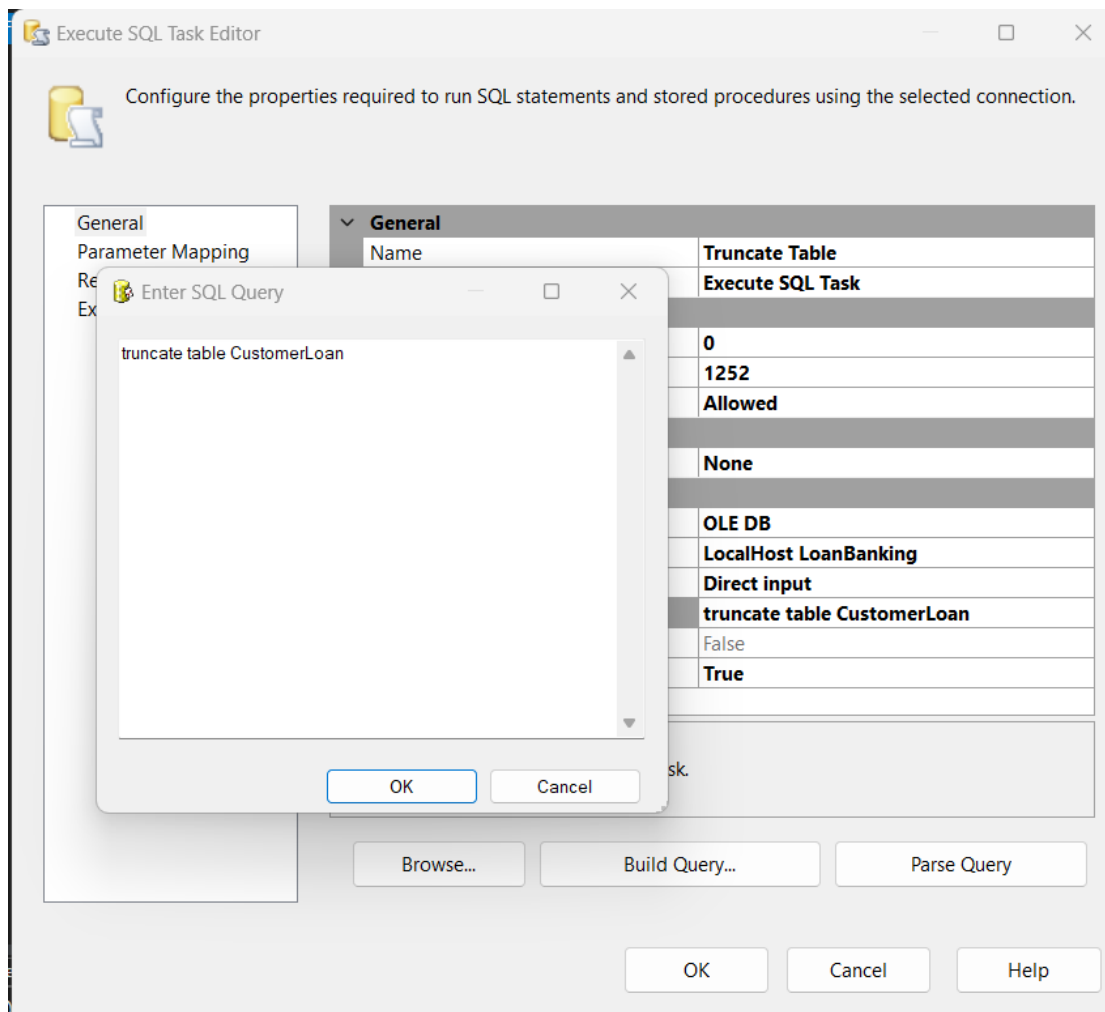


4.2.1.2.4 Quá trình tạo Control Flow



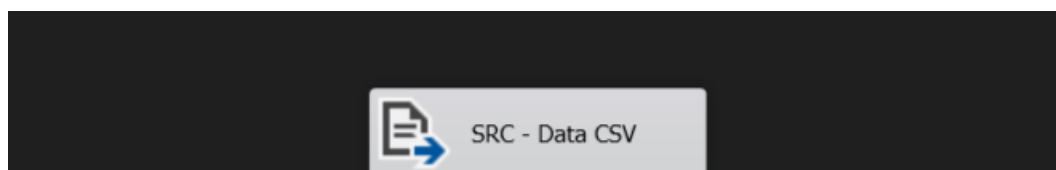
4.2.1.2.4.1 Excute SQL Task

Làm sạch các bảng trong database

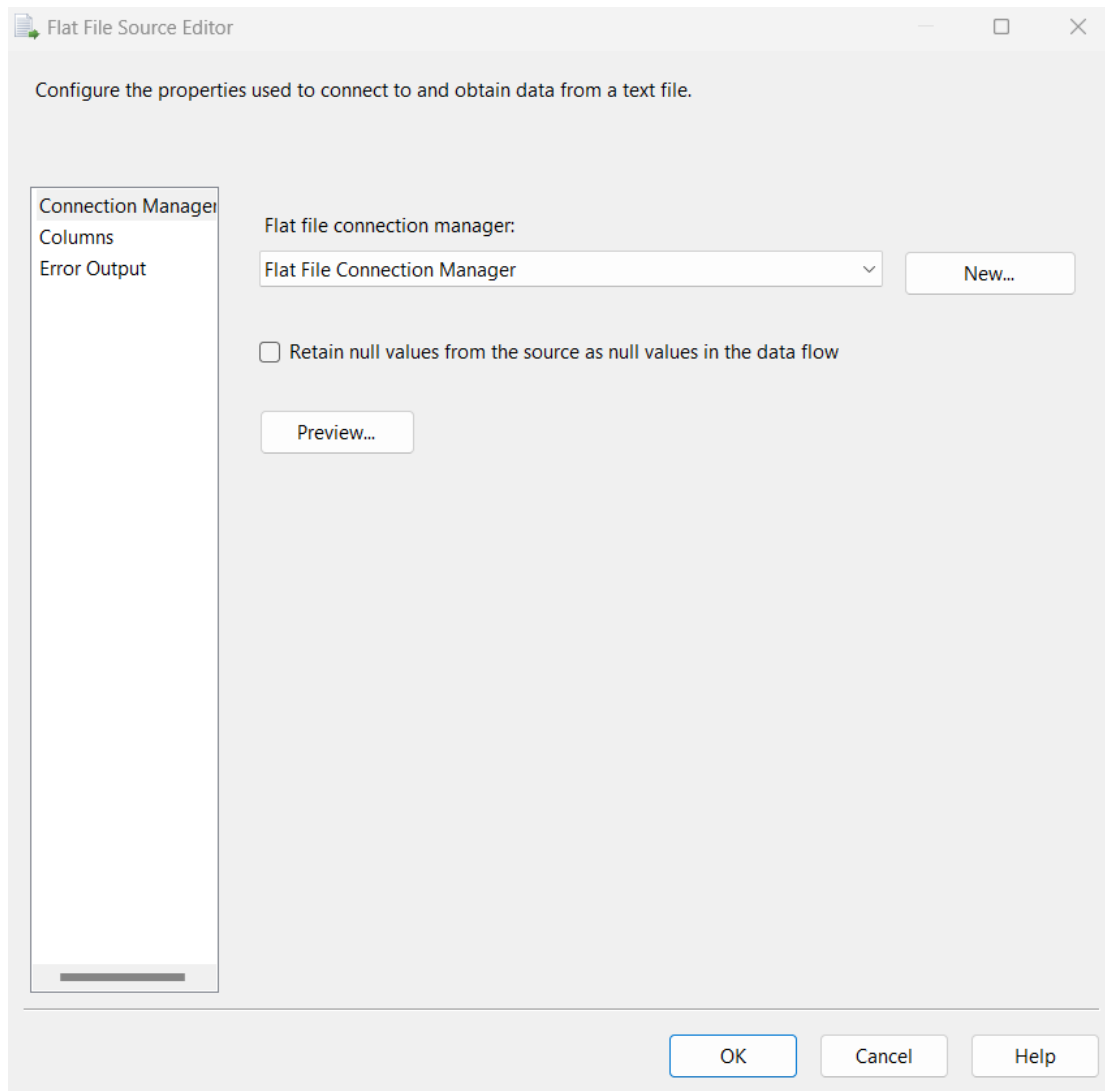


4.2.1.2.4.2 Data Flow Task

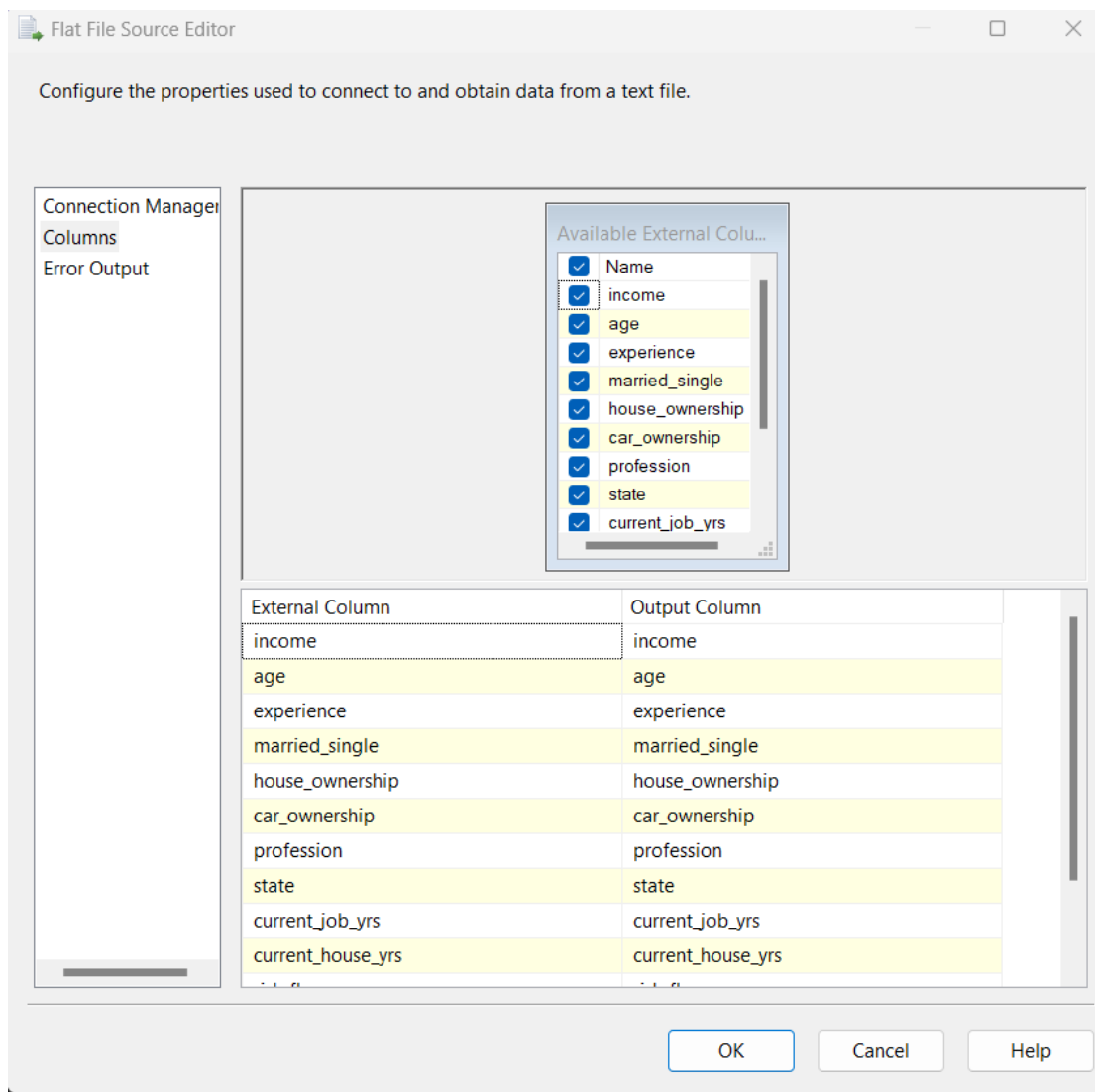
Tạo 1 **Source Assistant** tên **SRC – Data CSV**.



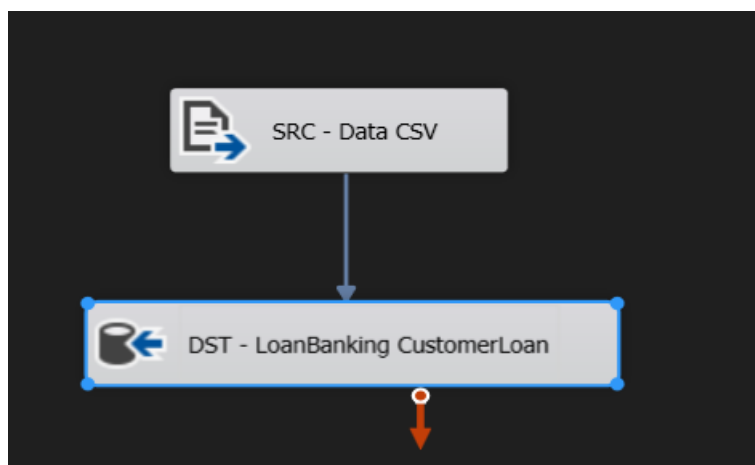
Nhấp đúp vào để cấu hình -> Chọn **Flat File Connection Manager**.



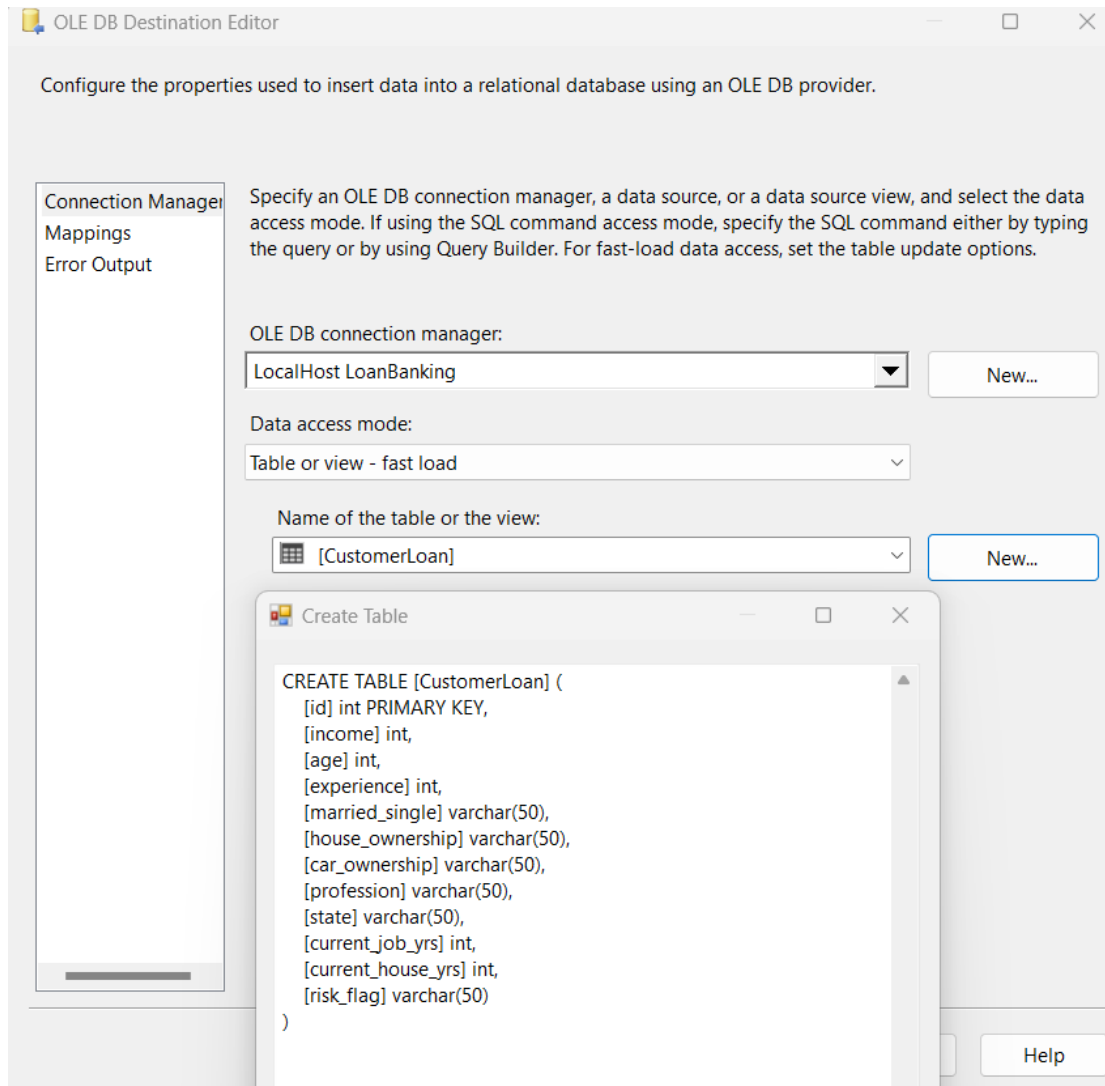
Chọn thuộc tính phù hợp -> Nhấn **OK**.



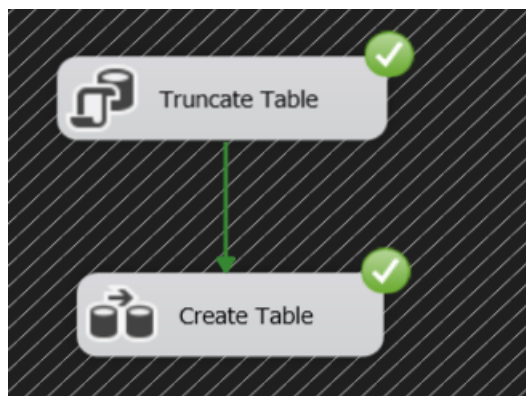
Tạo **Destination Assistant** tên **DST - LoanBanking CustomerLoan**.



Nhấn đúp để cấu hình -> Chọn kết nối đến **LoanBanking** -> Ấn **New...**
-> Nhập câu lệnh SQL phù hợp -> Nhấn **OK**.

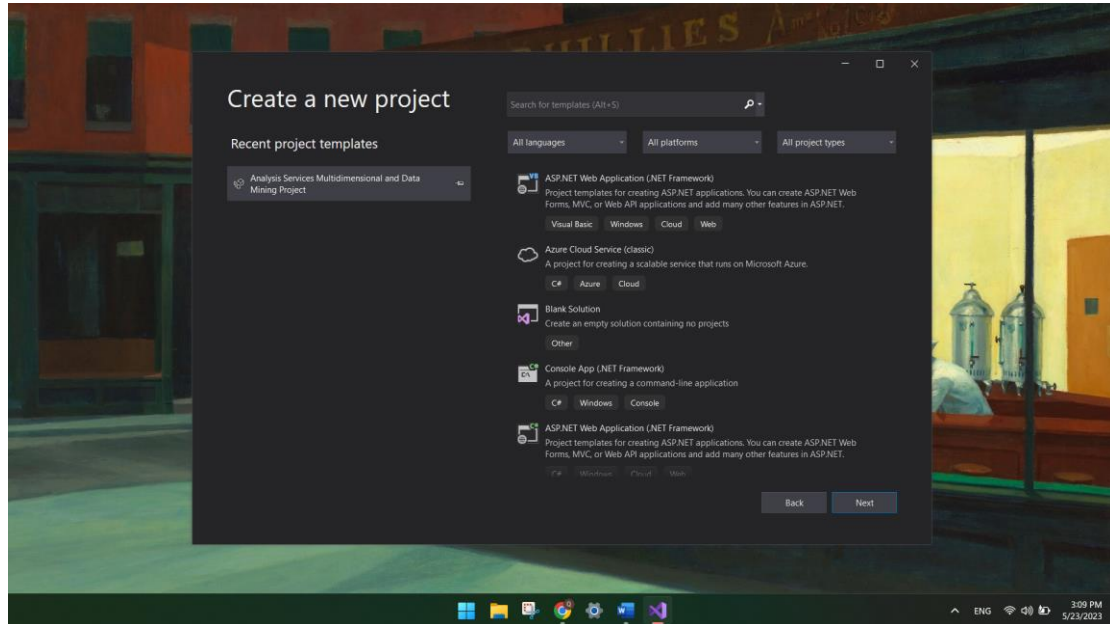


4.2.1.2.5 Kết quả chạy SSIS Package

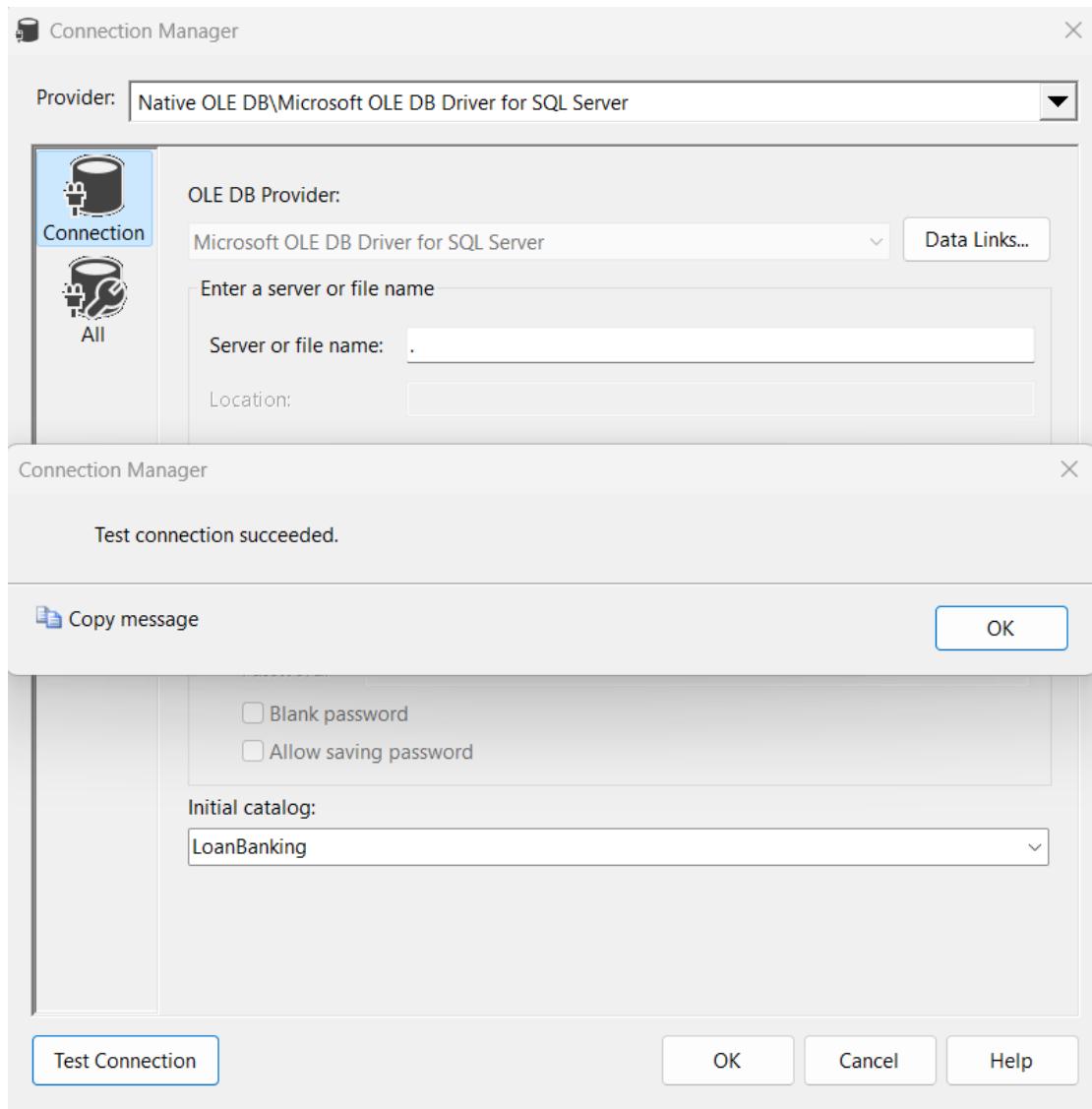


4.2.2 Xây dựng mô hình khai phá dữ liệu bằng công cụ SSAS

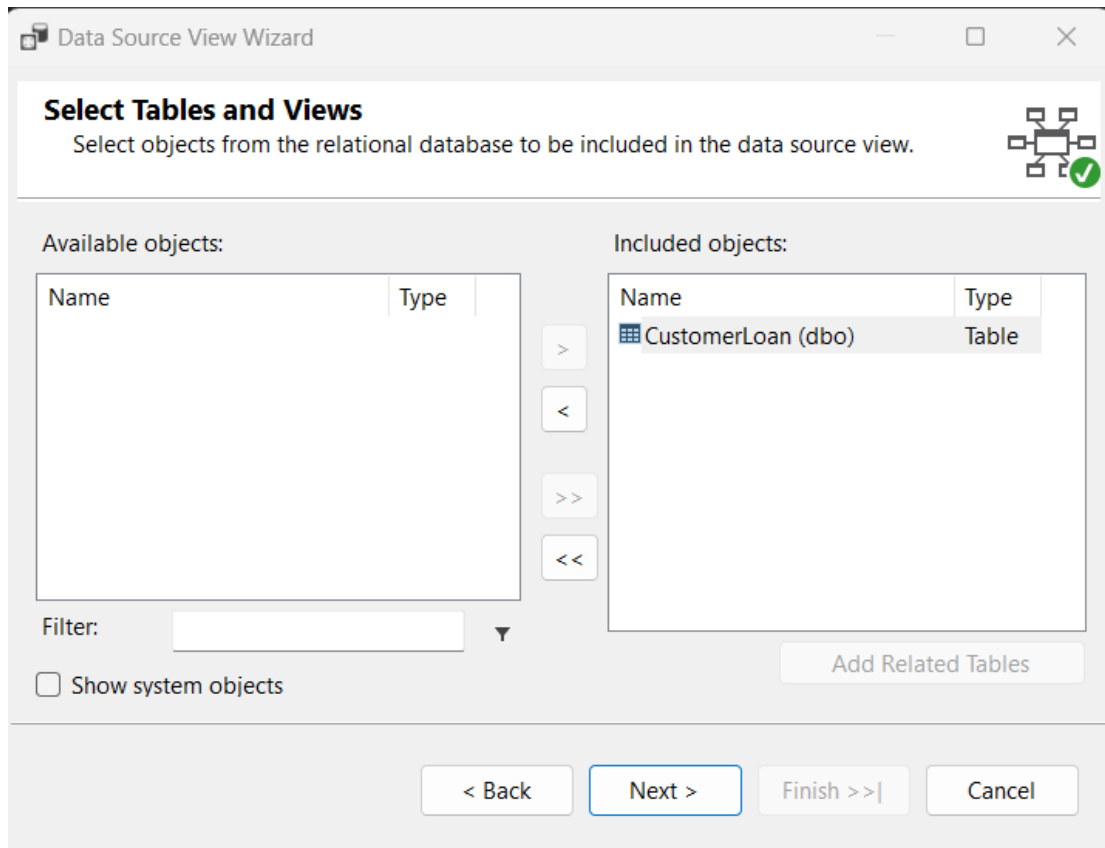
Bước 1: Mở Visual Studio 2019 -> Chọn Create a new project -> Chọn Analysis Services Multidimensional and Data Mining Project.



Bước 2: Tạo Data Source

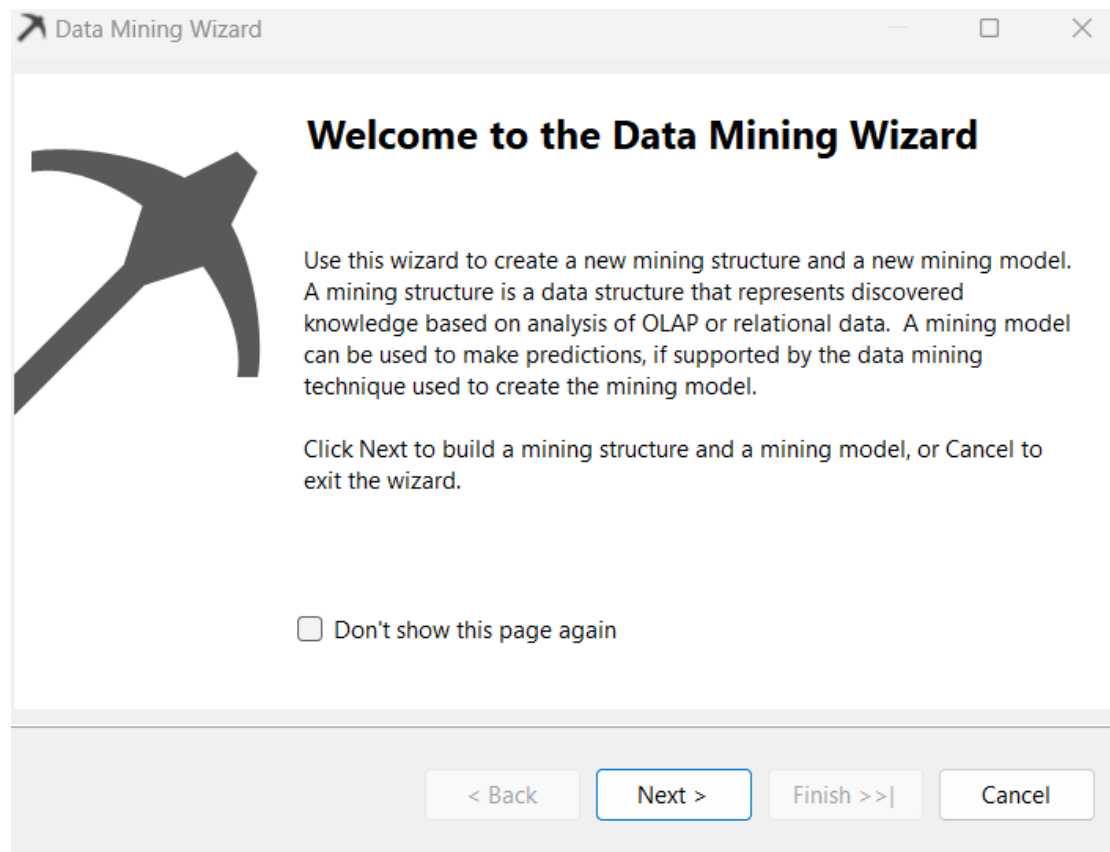


Bước 3: Tạo Data Source View

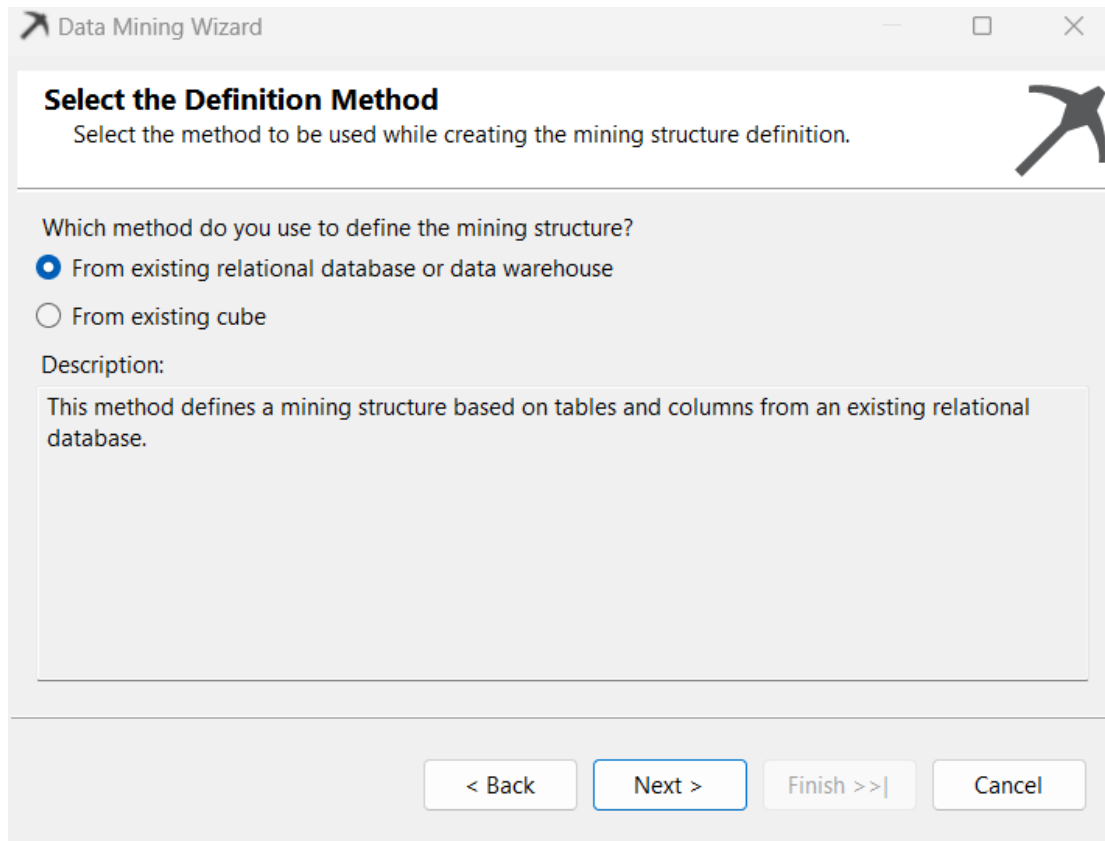


Bước 4: Tạo một Mining model structure

Bấm phải chuột vào **Mining Structures** -> Chọn **New Data Mining Structure** -> Chọn **Next**.



Chọn **From existing relational database or data warehouse** -> Chọn **Next**.



Select the Definition Method
Select the method to be used while creating the mining structure definition.

Which method do you use to define the mining structure?

☒ From existing relational database or data warehouse

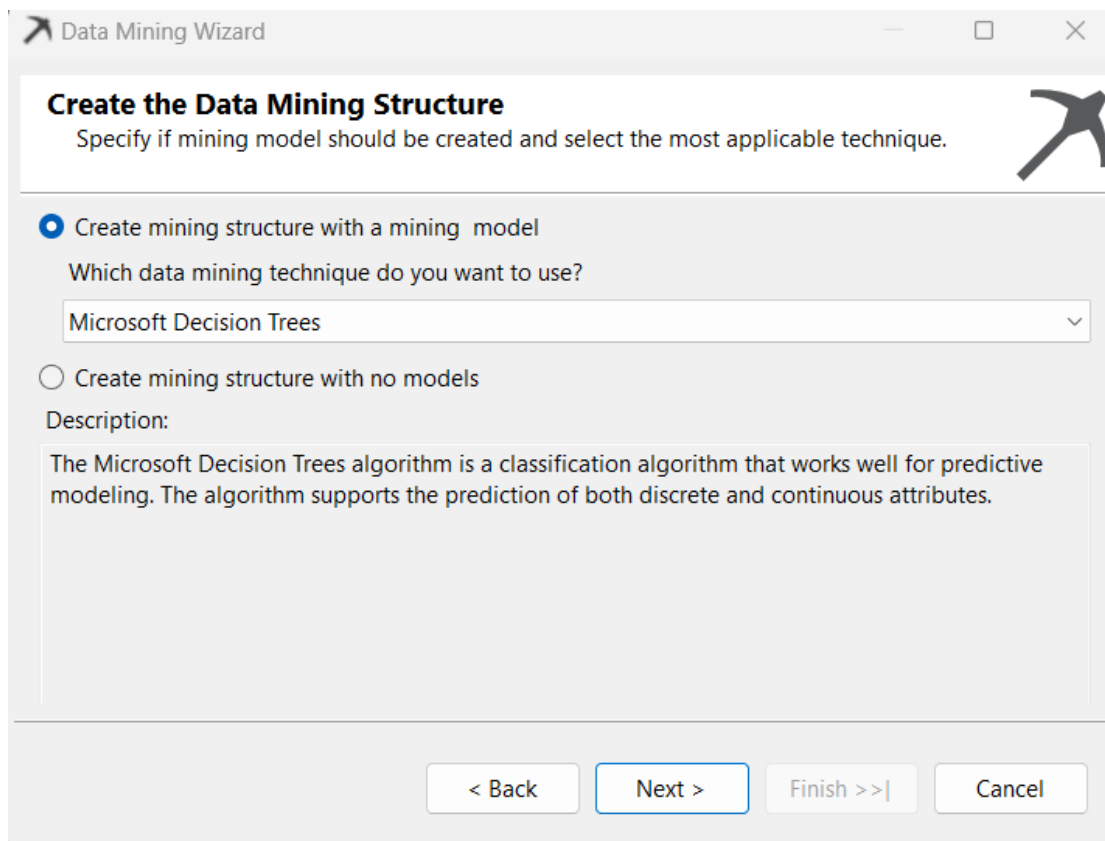
☐ From existing cube

Description:

This method defines a mining structure based on tables and columns from an existing relational database.

< Back Next > Finish >>| Cancel

Chọn **Create mining structure with a mining model** -> Chọn **Microsoft Decision Trees** -> Chọn **Next** -> Chọn **Next**.



Create the Data Mining Structure
Specify if mining model should be created and select the most applicable technique.

☒ Create mining structure with a mining model

Which data mining technique do you want to use?

Microsoft Decision Trees

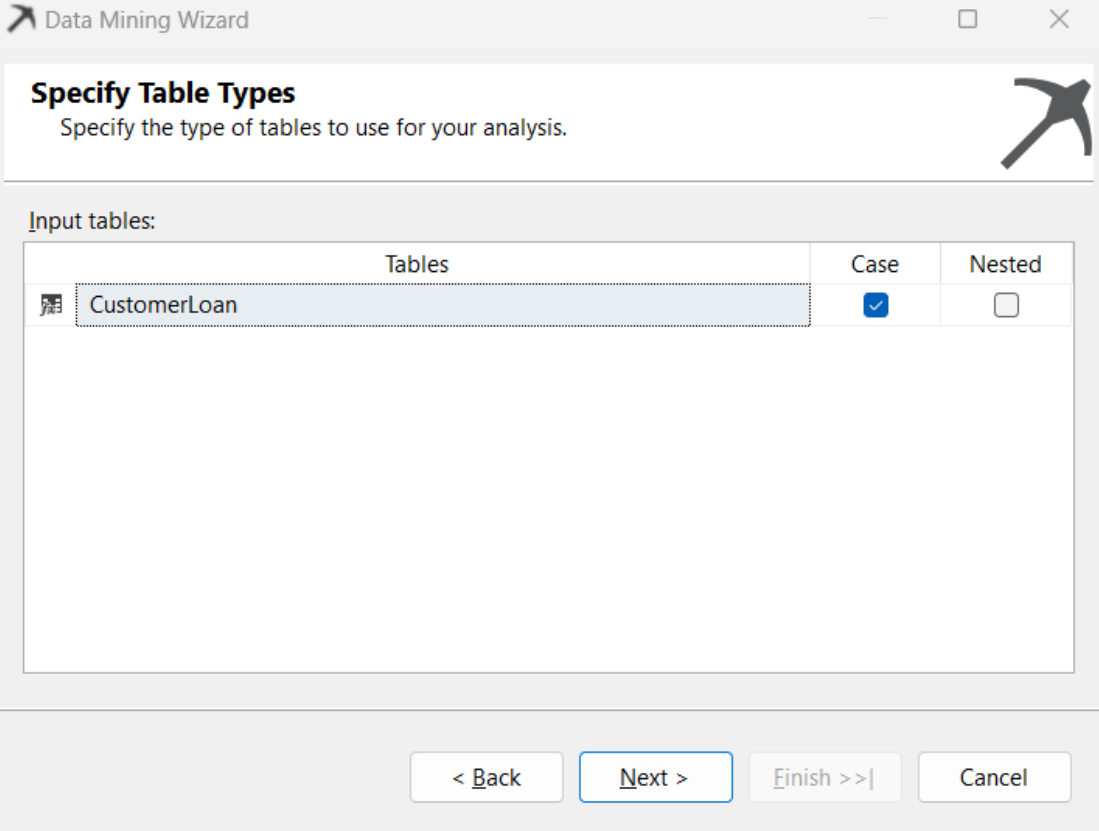
☐ Create mining structure with no models

Description:

The Microsoft Decision Trees algorithm is a classification algorithm that works well for predictive modeling. The algorithm supports the prediction of both discrete and continuous attributes.

< Back Next > Finish >>| Cancel

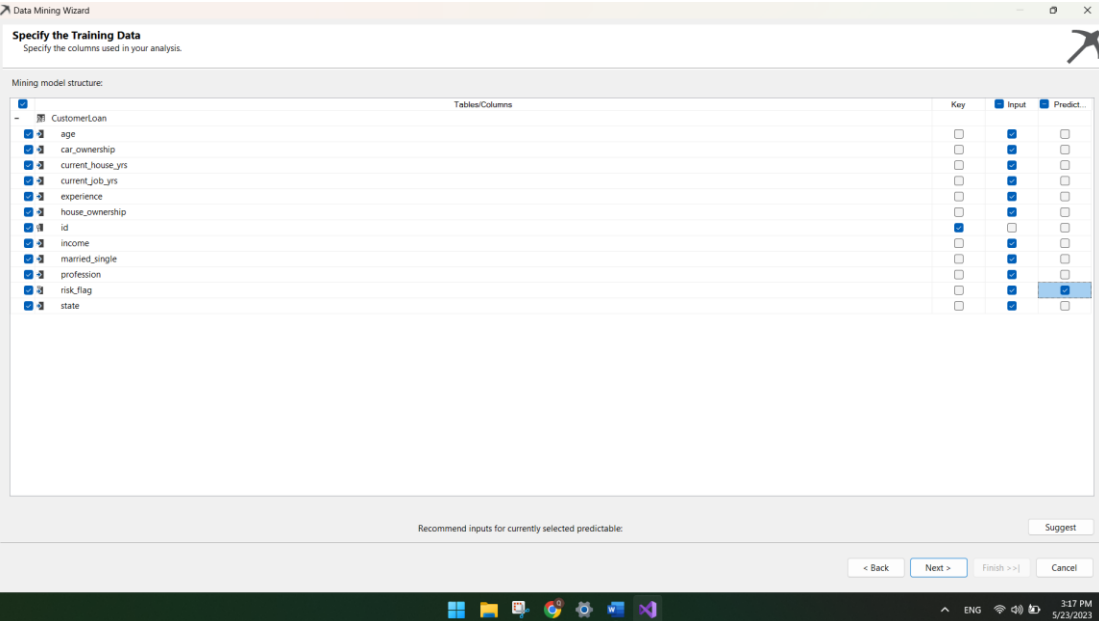
Chọn **Case** -> Chọn **Next**.



The screenshot shows the 'Specify Table Types' window of the Data Mining Wizard. The title bar reads 'Data Mining Wizard'. The main heading is 'Specify Table Types' with the instruction 'Specify the type of tables to use for your analysis.' Below this, the 'Input tables:' section contains a table with three columns: 'Tables', 'Case', and 'Nested'. The 'CustomerLoan' table is selected in the 'Tables' column. In the 'Case' column, the checkbox is checked, while in the 'Nested' column, it is unchecked. At the bottom, there are four buttons: '< Back', 'Next >', 'Finish >>|', and 'Cancel'.

Tables	Case	Nested
CustomerLoan	<input checked="" type="checkbox"/>	<input type="checkbox"/>

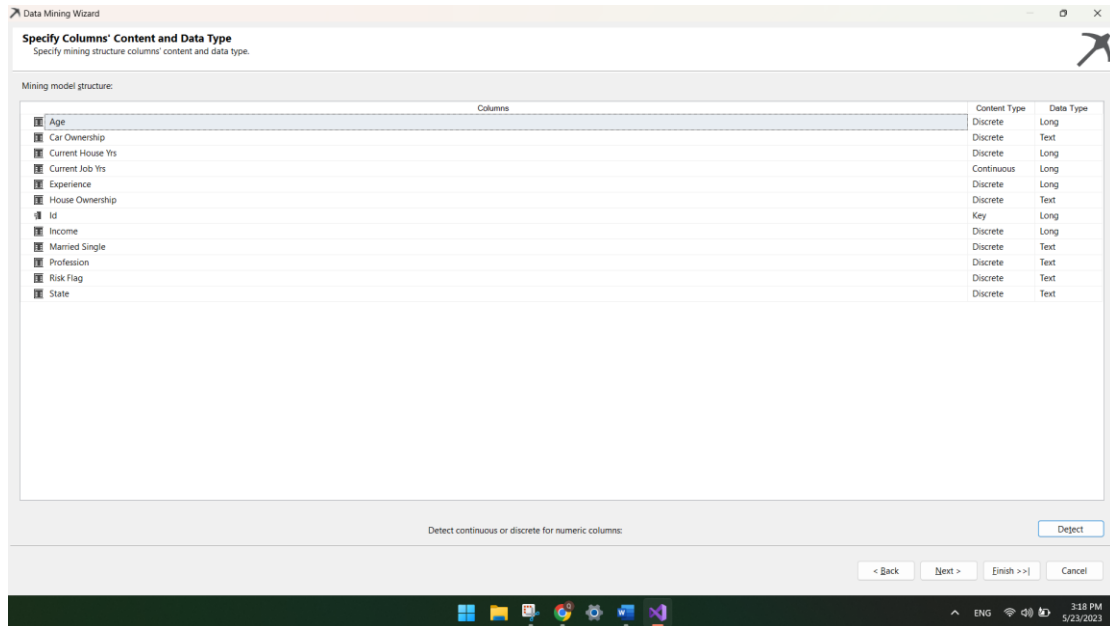
Chọn thuộc tính cho trường **Key**, **Input** và **Predict** -> Chọn **Next**.



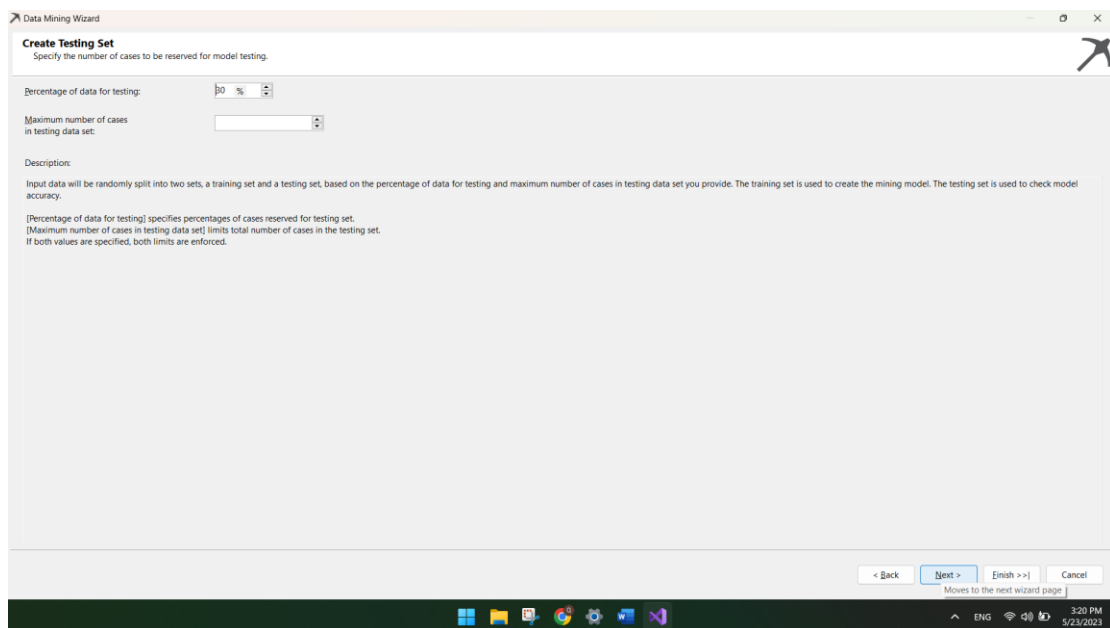
The screenshot shows the 'Specify the Training Data' window of the Data Mining Wizard. The title bar reads 'Data Mining Wizard'. The main heading is 'Specify the Training Data' with the instruction 'Specify the columns used in your analysis.' Below this, the 'Mining model structure:' section contains a table with four columns: 'Tables/Columns', 'Key', 'Input', and 'Predict...'. The 'CustomerLoan' table is selected in the 'Tables/Columns' column. In the 'Key' column, the checkbox is checked. In the 'Input' column, the checkbox is checked. In the 'Predict...' column, the checkbox is checked. At the bottom, there is a 'Recommend inputs for currently selected predictable:' section with a 'Suggest' button. Below this, there are four buttons: '< Back', 'Next >', 'Finish >>|', and 'Cancel'.

Tables/Columns	Key	Input	Predict...
CustomerLoan	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

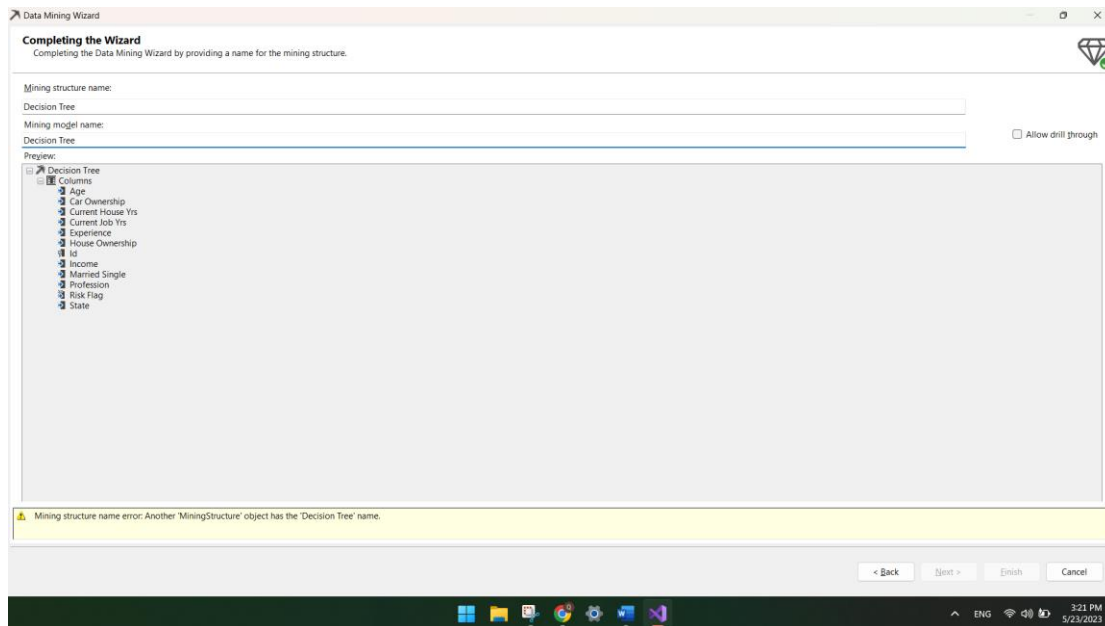
Chọn **Detect** để tự động xác định nội dung và kiểu dữ liệu các thuộc tính được sử dụng cho mô hình khai phá -> Chọn **Next**.



Chia tập dữ liệu theo tỷ lệ 70% cho **training** và 30% cho **testing** -> Chọn **Next**.



Đặt tên cho **Mining Structure** và **Mining Model** -> Chọn **Finish**.



Bước 5: Làm tương tự từ Bước 4 cho các Mining Model khác.

Chương 5: Giải quyết nghiệp vụ

5.1 Phát biểu nghiệp vụ

Câu hỏi nghiệp vụ cho kỹ thuật phân loại:

- Làm thế nào để dự đoán khả năng trả nợ của một khách hàng dựa trên các thông tin như tuổi, thu nhập, nghề nghiệp và lịch sử mắc nợ xấu?
- Có thể phân loại khách hàng thành hai nhóm: nhóm khách hàng có khả năng không mắc nợ xấu và nhóm khách hàng có khả năng mắc nợ xấu?

Câu hỏi nghiệp vụ cho kỹ thuật phân cụm:

- Có thể phân nhóm khách hàng thành các nhóm dựa trên hành vi và thông tin khách hàng?
- Có thể xác định các đặc trưng chung của các nhóm khách hàng dựa trên việc họ mắc nợ xấu hay không?

Câu hỏi nghiệp vụ cho kỹ thuật kết hợp:

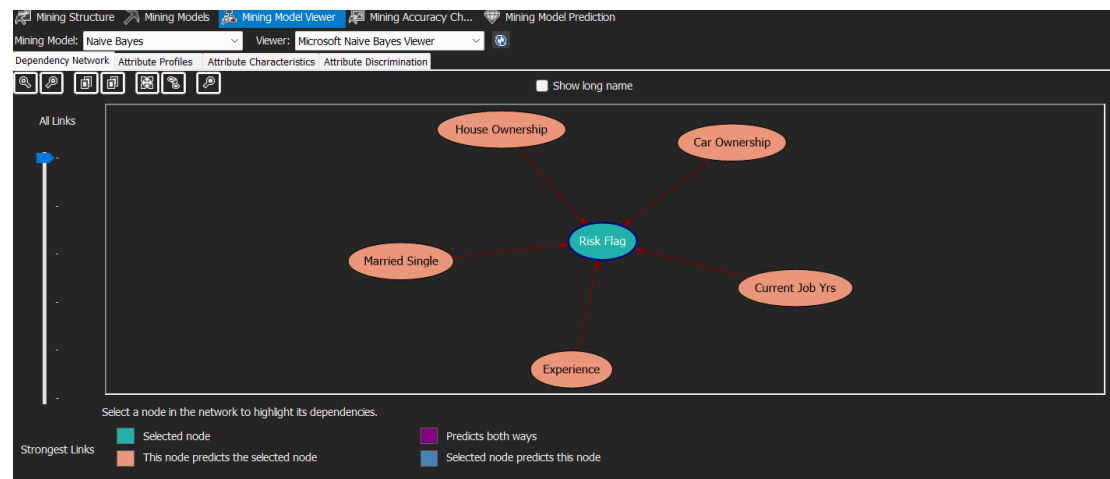
- Có thể tìm ra các quy tắc kết hợp giữa các yếu tố như thông tin của khách hàng để dự đoán khả năng khách hàng có thể thanh toán khoản vay đúng hạn hay không?
- Có thể tìm ra các quy tắc kết hợp giữa các yếu tố như thông tin của khách hàng để đưa ra các khuyến nghị về cho khách hàng vay?

5.2 Giải quyết nghiệp vụ bằng các kỹ thuật

5.2.1 Kỹ thuật phân tích phân loại (Classification Analysis)

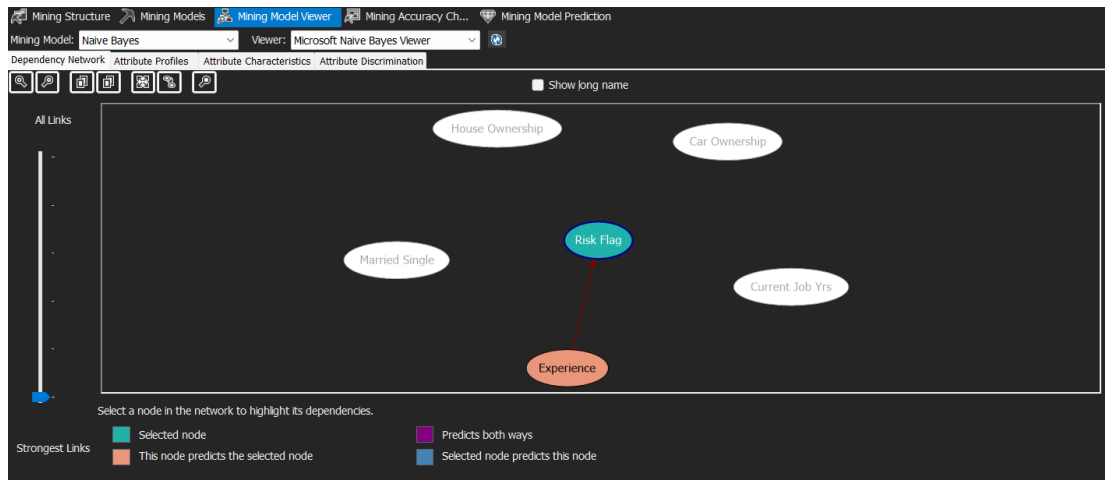
5.2.1.1 Kỹ thuật Naive Bayes

Sau khi xây dựng mô hình khai phá dữ liệu được giới thiệu trong phần 4.2 chúng ta có được biểu đồ các thuộc tính có mối tương quan cao đến thuộc tính được dự đoán

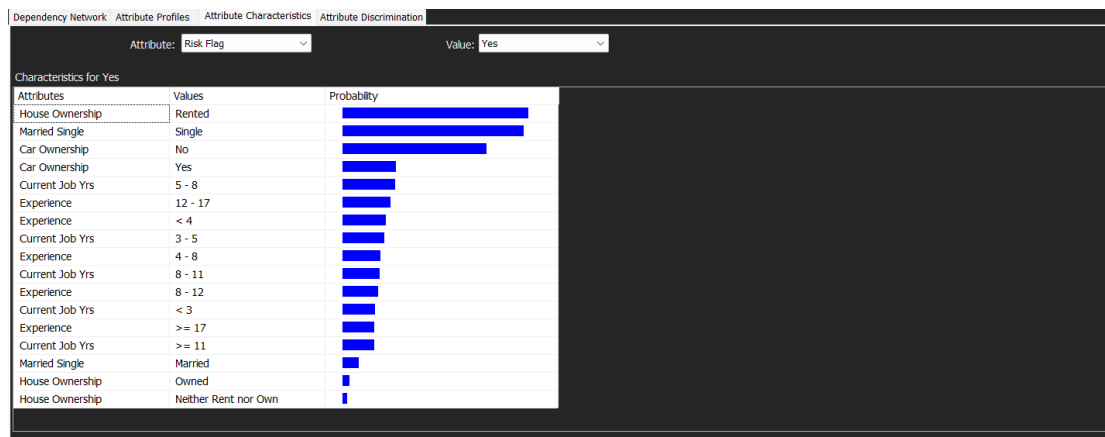


Phân tích kết quả:

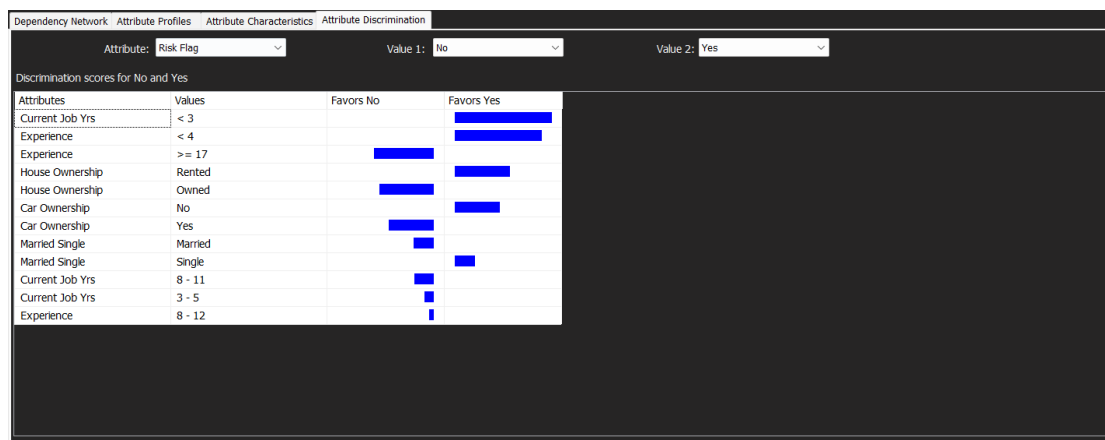
Dựa vào biểu đồ sau ta có thấy thuộc tính Experience có mối liên kết mạnh nhất đến kết quả của thuộc tính dự đoán.



Dựa vào biểu đồ sau ta có thấy những khách hàng có loại nhà là thuê có khả năng cao mắc nợ xấu.



Dựa vào biểu đồ sau ta có thấy những khách hàng có số năm đi làm nhỏ hơn 3 có khả năng cao mắc nợ xấu. Mặt khác, những khách hàng có kinh nghiệm đi làm từ 17 năm trở lên có khả năng cao không mắc nợ xấu.



Kết luận:

Dựa trên những gì ta phân tích, ta có thể đưa ra những kết luận sau:

- Tính không ổn định nhà cửa: Việc ở nhà thuê có thể cho thấy khách hàng không sở hữu tài sản như nhà đất, điều này có thể gây khó khăn trong việc đảm bảo ổn định tài chính. Việc không có tài sản có thể tạo ra rủi ro tài chính cao hơn và làm tăng khả năng gặp khó khăn trong việc trả nợ.
- Tính không ổn định công việc: Việc ở nhà thuê có thể cho thấy khách hàng không có sự ổn định trong công việc. Điều này có thể làm tăng nguy cơ mất việc hoặc thay đổi thu nhập không đáng tin cậy, làm giảm khả năng trả nợ đúng hạn.
- Ổn định công việc: Nhiều năm kinh nghiệm làm việc cho thấy khách hàng đã có sự ổn định trong công việc của mình. Điều này có thể cho thấy khả năng ổn định thu nhập, đảm bảo khả năng thanh toán các khoản vay một cách đều đặn.
- Kỹ năng quản lý tài chính: Khi làm việc trong nhiều năm, khách hàng có thể đã phát triển kỹ năng quản lý tài chính cá nhân tốt hơn. Điều này có thể bao gồm việc có kế hoạch tài chính, khả năng tiết kiệm và cẩn thận trong việc quản lý và trả nợ.

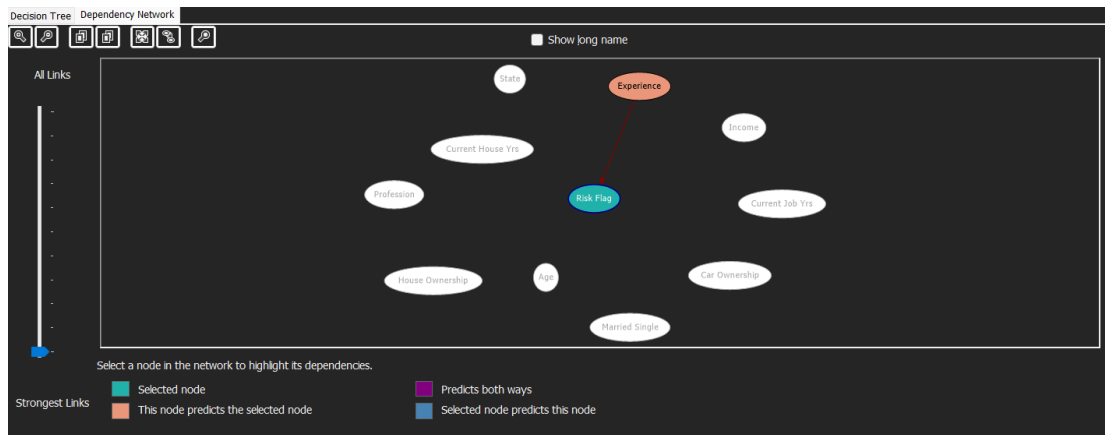
5.2.1.2 Kỹ thuật Decision Tree

Sau khi xây dựng mô hình khai phá dữ liệu được giới thiệu trong phần 4.2 chúng ta có được mô hình cây quyết định như sau



Phân tích kết quả:

Dựa vào biểu đồ sau ta có thấy thuộc tính Experience có mối liên kết mạnh nhất đến kết quả của thuộc tính dự đoán.



Kết luận:

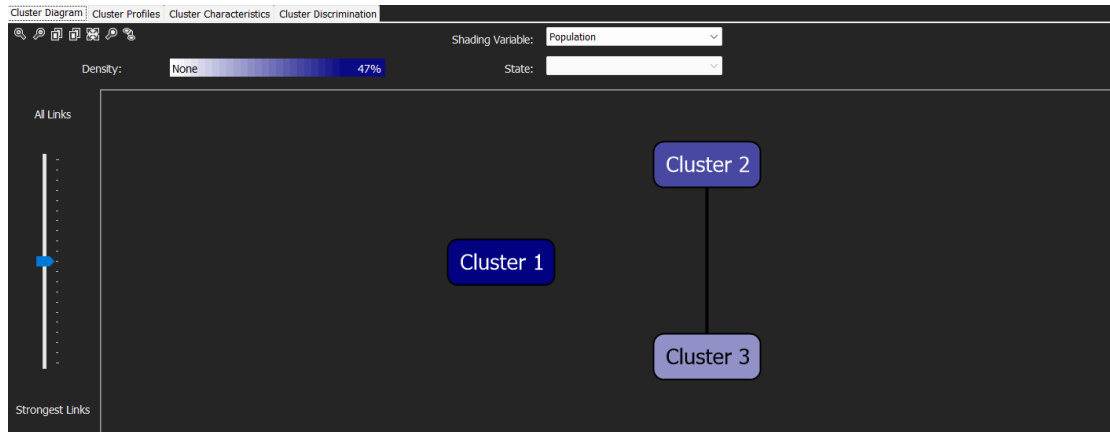
Dựa trên những gì ta phân tích, ta có thể đưa ra những kết luận sau:

- Kinh nghiệm đi làm ảnh hưởng đến khả năng quản lý tài chính: Mỗi tương quan lớn giữa kinh nghiệm đi làm và khả năng mắc nợ xấu cho thấy rằng khách hàng có kinh nghiệm đi làm lâu dài có khả năng quản lý tài chính tốt hơn và ít gặp khó khăn trong việc trả nợ. Kinh nghiệm đi làm có thể giúp khách hàng tích lũy kiến thức và kỹ năng quản lý tài chính, hiểu rõ hơn về các khía cạnh tài chính cá nhân và đưa ra quyết định tài chính thông minh.
- Tương quan giữa kinh nghiệm đi làm và sự ổn định công việc: Kinh nghiệm đi làm cũng thể hiện mức độ ổn định công việc của khách hàng. Khách hàng có kinh nghiệm đi làm lâu dài thường có khả năng duy trì công việc ổn định và thu nhập đáng tin cậy. Điều này có thể làm giảm khả năng gặp khó khăn tài chính và mắc nợ xấu.
- Tương quan với lịch sử tín dụng: Kinh nghiệm đi làm có thể ảnh hưởng đến lịch sử tín dụng của khách hàng. Khách hàng có kinh nghiệm đi làm lâu dài có thể có một lịch sử tín dụng tốt hơn, ví dụ như thời gian trả nợ đúng hạn và không có nợ chồng chất. Điều này có thể làm giảm khả năng mắc nợ xấu và tăng khả năng được đánh giá tích cực bởi tổ chức tín dụng.

5.2.2 Kỹ thuật phân tích theo cụm (Clustering Analysis)

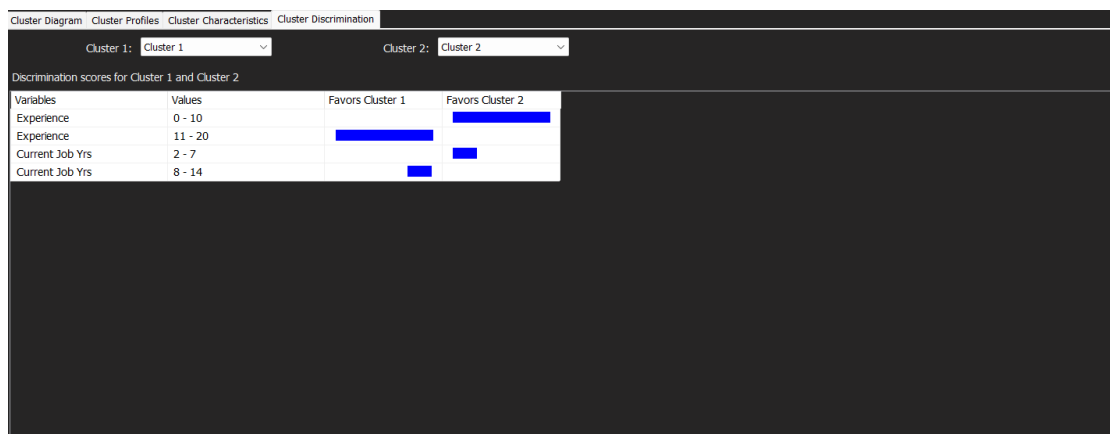
5.2.2.1 Kỹ thuật K-Means

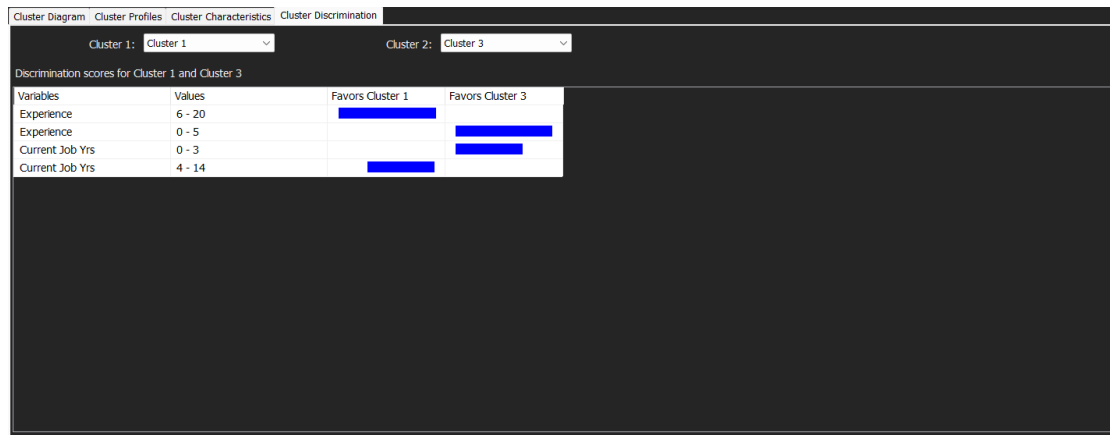
Sau khi xây dựng mô hình khai phá dữ liệu được giới thiệu trong phần 4.2 chúng ta có được mô hình các cụm như sau



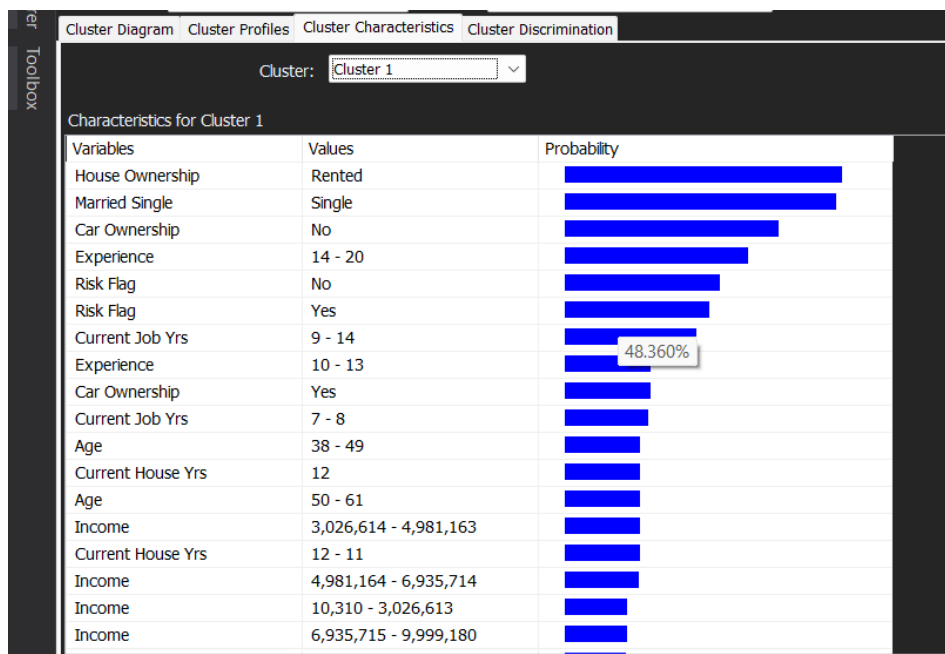
Phân tích kết quả:

Dựa vào biểu đồ so sánh giữa cụm 1, cụm 2 và cụm 3 ta có thấy cluster 1 là những người có nhiều kinh nghiệm trong công việc từ 10 năm trở lên. Mặt khác, cụm 2 là những người có kinh nghiệm trong công việc từ 5 năm trở lên và cuối cùng cụm 3 là những người ít kinh nghiệm trong công việc.





Ngoài ra ta còn có thể thấy khả năng mắc nợ xấu của cụm 1 chỉ có 48% thấp nhất trong 3 cụm (cụm 2: 49%, cụm 3: 55%)



Kết luận:

Dựa trên những gì ta phân tích, ta có thể đưa ra những kết luận sau:

- Mối quan hệ giữa kinh nghiệm làm việc và khả năng mắc nợ xấu: Có xu hướng cho thấy khách hàng có nhiều năm kinh nghiệm đi làm (cụm 1) có khả năng mắc nợ xấu thấp hơn so với khách hàng có ít năm kinh nghiệm đi làm (cụm 2 và cụm 3). Điều này có thể cho thấy mối quan hệ giữa kinh nghiệm làm việc và khả năng quản lý tài chính, đồng thời cũng có thể chỉ ra rằng khách hàng có kinh nghiệm làm việc lâu dài có khả năng duy trì và thanh toán nợ tốt hơn.

- Khách hàng ở cụm 1 có khả năng mắc nợ xấu thấp nhất: Thông tin cho thấy khách hàng ở cụm 1 có khả năng mắc nợ xấu thấp nhất trong cả ba cụm. Điều này có thể cho thấy rằng khách hàng có nhiều năm kinh nghiệm làm việc từ 10 năm đến hơn 20 năm có khả năng quản lý tài chính tốt hơn và ít gặp khó khăn trong việc trả nợ.

5.2.3 Kỹ thuật kết hợp (Association Analysis)

5.2.3.1 Kỹ thuật Apriori

Sau khi xây dựng mô hình khai phá dữ liệu được giới thiệu trong phần 4.2 chúng ta có được các luật được thể hiện trong hình được sắp xếp dựa trên mức độ quan trọng và xác suất.

Pr...	Importance	Rule
0.645	0.115	Current Job Yrs < 3, Age < 32 -> Risk Flag = Yes
0.615	0.095	Experience < 4, Age < 32 -> Risk Flag = Yes
0.604	0.085	Experience >= 17, Income = 2076631 - 4116905 -> Risk Flag = No
0.603	0.085	House Ownership = Owned, Married Single = Single -> Risk Flag = No
0.601	0.085	Experience >= 17, Car Ownership = Yes -> Risk Flag = No
0.601	0.083	Experience >= 17, Age >= 67 -> Risk Flag = No
0.600	0.083	House Ownership = Owned -> Risk Flag = No
0.596	0.080	Income = 4116905 - 5968991, Car Ownership = Yes -> Risk Flag = No
0.594	0.078	Current Job Yrs >= 11, Car Ownership = Yes -> Risk Flag = No
0.591	0.075	Experience >= 17, Current Job Yrs = 3 - 5 -> Risk Flag = No
0.588	0.073	Current Job Yrs < 3, Current House Yrs < 11 -> Risk Flag = Yes
0.586	0.074	State = Madhya_Pradesh, House Ownership = Rented -> Risk Flag = Yes
0.585	0.070	Current Job Yrs < 3, Income < 8031925 -> Risk Flag = Yes
0.584	0.070	Current Job Yrs < 3, Income < 2076631 -> Risk Flag = Yes
0.581	0.068	Experience >= 17, Current House Yrs = 12 - 13 -> Risk Flag = No
0.581	0.074	Current Job Yrs < 3, Car Ownership = No -> Risk Flag = Yes
0.579	0.066	Current House Yrs < 11, Experience < 4 -> Risk Flag = Yes

Phân tích kết quả:

Dựa vào biểu đồ trên ta có thấy những khách hàng có số năm đi làm hiện tại bé hơn 3 và độ tuổi từ 32 trở xuống có khả năng cao mắc nợ xấu. Mặt khác, những khách hàng có kinh nghiệm làm việc từ 17 năm trở lên và thu nhập trong khoảng 2 triệu đến 4 triệu có khả năng cao không mắc nợ xấu.

Kết luận:

Dựa trên những gì ta phân tích, ta có thể đưa ra những kết luận sau:

- Thiếu kinh nghiệm quản lý tài chính: Khách hàng ít kinh nghiệm nghề nghiệp có thể thiếu kinh nghiệm và kiến thức về quản lý tài chính cá nhân. Điều này có thể làm giảm khả năng đánh giá và ứng phó với các rủi ro tài chính, dẫn đến khả năng cao mắc nợ xấu.

- Thu nhập thấp và ổn định kém: Khách hàng mới vào nghề thường có thu nhập thấp và không ổn định. Sự thiếu ổn định thu nhập có thể tạo ra khó khăn trong việc thanh toán các khoản nợ và làm tăng khả năng mắc nợ xấu.

Chương 6: Đánh giá mô hình

6.1 Bảng so sánh các thuật toán

Mô hình	Điểm Precision (SSAS)
Naive Bayes	0.58
Decision Tree	0.63
K-Mean	0.53
Apriori	0.56

Từ bảng trên ta có thấy được mô hình Decision Tree cho điểm Precision cao nhất, có thể nói rằng thuật toán Decision Tree có khả năng tốt trong việc dự đoán và phân loại khách hàng dựa trên hành vi và thông tin khách hàng.

Điểm Precision cao cho thấy Decision Tree có tỷ lệ dự đoán chính xác các trường hợp True Positive là khá cao. Điều này có nghĩa là thuật toán Decision Tree có khả năng phân loại các khách hàng có khả năng hoặc không có khả năng mắc nợ xấu với mức chính xác cao hơn so với các thuật toán phân loại khác trong bảng so sánh.

Tuy nhiên, để đánh giá toàn diện hiệu suất của thuật toán Decision Tree trên tập dữ liệu cụ thể này, cần xem xét và so sánh các chỉ số đánh giá khác như Recall, F1-score và Accuracy. Ngoài ra, cần kiểm tra các yếu tố khác như cân bằng giữa các lớp dữ liệu, xử lý giá trị thiếu và các thao tác tiền xử lý dữ liệu để đảm bảo tính khách quan và đáng tin cậy của kết quả đánh giá.

Do đó, dựa trên thông tin từ bảng so sánh thuật toán, Decision Tree có điểm Precision cao nhất cho thấy tiềm năng và hiệu quả của thuật toán trong việc phân loại khách hàng. Tuy nhiên, việc đánh giá và chọn thuật toán phù hợp còn phụ thuộc vào mục tiêu và yêu cầu cụ thể của dự án.

Chương 7: Kết luận

7.1 Kết quả đạt được

Trong kỳ học vừa qua , nhóm đã tìm hiểu và vận dụng kiến thức về xây dựng mô hình, áp dụng các kỹ thuật khai phá dữ liệu và đạt được các kết quả như sau:

- Nắm rõ các khái niệm cơ bản về kỹ thuật khai phá dữ liệu và tính chất của các kỹ thuật.
- Nắm vững kiến thức và có thể vận dụng, xây dựng một mô hình hoàn chỉnh dùng để khai thác dữ liệu.
- Trang bị kiến thức về các công cụ SSIS, SSAS.
- Xây dựng được mô hình khai phá dữ liệu cho riêng mình.
- Hiểu và áp dụng được các loại kỹ thuật khai phá từ cơ bản đến nâng cao.
- Đặt câu hỏi nghiệp vụ và giải quyết chúng.

7.2 Những hạn chế

Do thời gian hạn ngắn cộng với khối lượng công việc nhiều nên trong quá trình thực hiện đồ án nhóm còn gặp phải một số vấn đề:

- Chưa áp dụng được nhiều loại kỹ thuật khác nhau.
- Quá trình tạo mô hình dữ liệu còn quá sơ sài, đơn giản.
- Phân tích các mô hình chưa đủ sâu.
- Chưa áp dụng được nhiều công cụ khác ngoài SSAS để áp dụng khai phá dữ liệu.

7.3 Bảng phân công nhiệm vụ trong nhóm

Công việc	Người thực hiện			
	Nguyễn Trí Quốc	Nguyễn Ý	Nguyễn Thanh Tùng	Đinh Quang Thắng
Chọn và tìm hiểu tập dữ liệu	25%	25%	25%	25%

Nghiên cứu tính cần thiết để xây dựng kho dữ liệu	25%	25%	25%	25%
Khảo sát nghiên cứu, phân tích báo cáo nghiệp vụ	25%	25%	25%	25%
Xây dựng mô hình khai phá	25%	25%	25%	25%
Áp dụng các kỹ thuật phân loại để khai phá dữ liệu	35%	15%	35%	15%
Áp dụng các kỹ thuật phân cụm để khai phá dữ liệu	20%	30%	20%	30%
Áp dụng các kỹ thuật kết hợp để khai phá dữ liệu	50%	10%	30%	10%
Đánh giá mô hình	20%	40%	20%	40%
Viết báo cáo	50%		50%	

7.4 Tài liệu tham khảo

- **Công cụ SSAS:** <https://learn.microsoft.com/en-us/analysis-services/data-mining/data-mining-ssas?view=asallproducts-allversions>
- **Công cụ SSIS:** <https://learn.microsoft.com/vi-vn/sql/integration-services/sql-server-integration-services?view=sql-server-2017>
- **Kỹ thuật khai phá dữ liệu:** Jiawei Han, Micheline Kamber, Jian Pei (2011): *Data Mining Concepts and Techniques 3rd Edition*