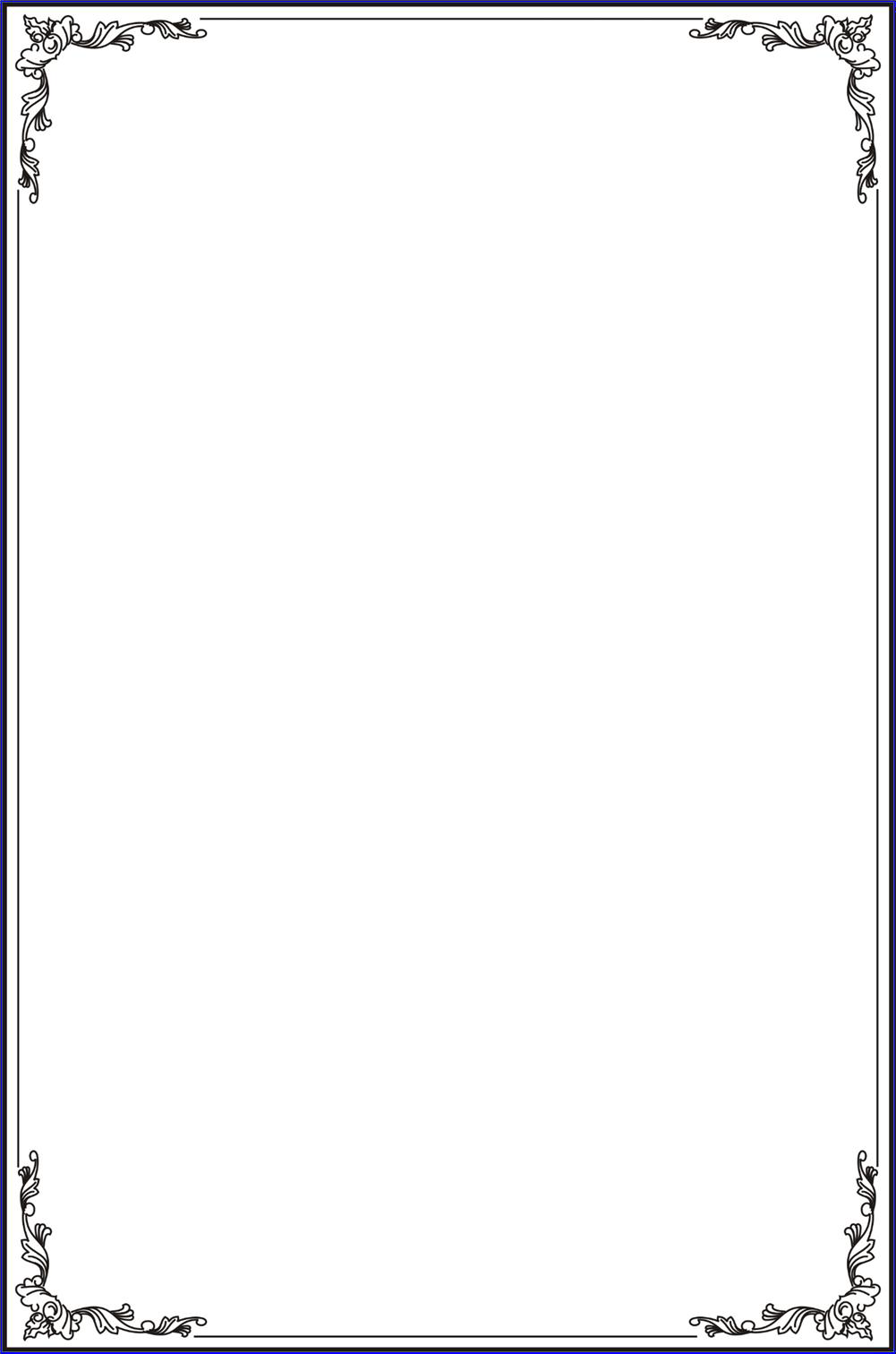
**yTRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHI MINH**



**KHOA CÔNG NGHỆ THÔNG TIN**

---------------------

**BÁO CÁO ĐỒ ÁN**

**ĐỀ TÀI: XÂY DỰNG KHO DỮ LIỆU BÁN HÀNG**

**Môn học: Kho dữ liệu**

**Mã lớp học phần: DAMI330484\_01**

**Giảng viên hướng dẫn: Ths. Nguyễn Văn Thành**

**Nhóm sinh viên thực hiện: Nhóm 6**

**Nguyễn Võ Song Toàn – 20133097**

**Lê Vũ Thế Khôi – 20133057**

**Ngô Văn Khánh – 20133054**

TP Hồ Chí Minh, tháng 05 năm 2023

**MỤC LỤC**

[LỜI CẢM ƠN 1](#_Toc135664486)

[CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI 2](#_Toc135664487)

[1.1. Giới thiệu đề tài 2](#_Toc135664488)

[1.1.1. Lý do chọn đề tài 2](#_Toc135664489)

[1.1.2. Nội dung đề tài 3](#_Toc135664490)

[1.2. Mô tả dữ liệu gốc 4](#_Toc135664491)

[1.3. Xử lý dữ liệu 5](#_Toc135664492)

[1.4. Đổ dữ liệu từ file CSV vào SQL source 7](#_Toc135664493)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 13](#_Toc135664494)

[2.1. Giới thiệu về Khai phá dữ liệu 13](#_Toc135664495)

[2.2. Phương pháp Khai phá dữ liệu 14](#_Toc135664496)

[2.2.1. Các khái niệm cơ bản trong Khai phá dữ liệu 14](#_Toc135664497)

[2.2.1.1. Dữ liệu và các loại hình dữ liệu 14](#_Toc135664498)

[2.2.1.2. Mẫu dữ liệu, thuộc tính, tập dữ liệu 15](#_Toc135664499)

[2.2.1.3. Quá trình tiền xử lý dữ liệu 15](#_Toc135664500)

[2.2.2. Các thuật toán trong Khai phá dữ liệu 16](#_Toc135664501)

[2.2.2.1. Association Rule Mining (Khai phá luật kết hợp) 16](#_Toc135664502)

[2.2.2.2. Microsoft Clustering (Phân cụm) 18](#_Toc135664503)

[2.2.2.3. Microsoft Decision Trees (Cây quyết định) 19](#_Toc135664504)

[CHƯƠNG 3: PHÂN TÍCH DỮ LIỆU (SSAS) 22](#_Toc135664505)

[CHƯƠNG 4: KẾT LUẬN 51](#_Toc135664506)

[4.1. Môi trường làm việc 51](#_Toc135664507)

[4.2. Phân công nhiệm vụ 52](#_Toc135664508)

[4.3. Hướng phát triển 52](#_Toc135664509)

# LỜI CẢM ƠN

Lời mở đầu, nhóm xin gửi lời cảm ơn đến thầy Nguyễn Văn Thành (Giảng viên hướng dẫn môn Kho dữ liệu). Thầy đã cung cấp kiến thức, chỉ bảo và đóng góp những ý kiến quý báu giúp nhóm hoàn thành được đồ án môn học của mình. Nhóm rất mong nhận sự góp ý từ phía thầy nhằm rút ra những kinh nghiệm quý báu và hoàn thiện vốn kiến thức để nhóm có thể tiếp tục hoàn thành những đồ án khác trong tương lai. Xin chân thành cảm ơn Thầy!

Hiện nay, công nghệ thông tin đã được ứng dụng rộng rãi trong tất cả các lĩnh vực của đời sống xã hội đặc biệt là trong lĩnh vực Business Intelligence (BI). Vấn đề cốt lõi trong hệ thống BI là kho dữ liệu (Data Warehouse) và khai phá dữ liệu (Data Mining) vì dữ liệu dùng trong BI là dữ liệu tổng hợp (Nhiều nguồn, nhiều định dạng, phân tán và có tính lịch sử) đó là đặc trưng của kho dữ liệu. Đồng thời việc phân tích dữ liệu trong BI không phải là những phân tích đơn giản (Query, Filtering) mà là những kỹ thuật trong khai phá dữ liệu (Data Mining) dùng để phân loại (classification) phân cụm (Clustering), hay dự đoán (Prediction). Vì vậy BI có mối quan hệ rất chặt chẽ với Data Warehouse và Data Mining.

Do đó việc ứng dụng mô hình OLAP vào trong lĩnh vực phân tích dữ liệu kinh doanh đang đem lại rất nhiều giá trị cho các doanh nghiệp lớn, giúp doanh nghiệp có được cái nhìn tổng quan về dữ liệu, hỗ trợ ra quyết định, khai phá dữ liệu và giúp định hướng và phát triển doanh nghiệp, vv… Trong thời gian một học kỳ thực hiện đề tài, nhóm chúng em đã vận dụng những kiến thức nền tảng đã tích lũy đồng thời kết hợp với việc học hỏi và nghiên cứu những kiến thức mới vận dụng tối đa những gì đã thu thập được để hoàn thành đề tài đồ án tốt nhất. Tuy nhiên, trong quá trình thực hiện, nhóm chúng em không tránh khỏi những thiếu sót. Chính vì vậy, rất mong nhận được những sự góp ý từ phía Thầy nhằm hoàn thiện những kiến thức mà nhóm chúng em đã học tập và là hành trang để nhóm thực hiện tiếp các đề tài khác trong tương lai.

# CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI

## Giới thiệu đề tài

### Lý do chọn đề tài

Phân tích dữ liệu bán hàng đang mà một hoạt động rất quan trọng của doanh nghiệp, nó ảnh hưởng đến sự tồn tại và phát triển của nhiều doanh nghiệp trên thị trường hiện nay, phân tích dữ liệu kinh doanh có một số vai trò quan trọng như sau:



* Cải thiện sự gắn kết khách hàng

Một trong những thử thách lớn nhất của các nhà bán hàng là làm thế nào để chuyển những khách hàng mua sắm một lần (one-time shoppers) thành những khách hàng trung thành của mình (brand loyalists). Những nhà bán hàng lớn trên thế giới như Amazon đã theo dõi những thói quen tiêu dùng của khách hàng, lịch sử tìm kiếm, thị hiếu, ... Không chỉ những công ty thương mại điện tử eCommerce lớn, mà kể cả những công ty vừa và nhỏ cũng cần sử dụng những phân tích này, kết hợp với phân tích dự báo (predictive analytics) để phát hiện ra những tiềm năng bán hàng, xu hướng bán hàng, các chương trình thu hút sự tham gia của khách hàng hoặc thậm chí là ý tưởng về những sản phẩm mới mà các khách hàng tiềm năng đang cần. Bằng việc kết hợp phân tích bán hàng (retail analytics) vào các mô hình dự báo, chúng ta có thể dự báo được nhu cầu khách hàng và khuyến khích họ quay trở lại.

* Cải thiện quản lý kho hàng và chuỗi cửa hàng

Có nhiều các sản phẩm còn ứ đọng chưa bán được, hoặc chưa đủ các sản phẩm cần thiết có thể dẫn đến các tổn thất lớn cho doanh nghiệp. Sử dụng phân tích dự báo có thể giúp doanh nghiệp giảm chi phí trong quản lý hàng tồn kho, tăng hiệu quả sử dụng các kệ hàng, tập trung nguồn lực vào những khu vực có nhu cầu cao, nắm bắt nhanh các xu hướng bán hàng, tối ưu hóa việc vận chuyển, giảm chi phí tồn kho và cải thiện chuỗi cung ứng, gia tăng lợi nhuận cho doanh nghiệp.

* Cải thiện các quyết định về giá bán

Trong xu hướng cắt giảm chi tiêu của người tiêu dùng và sự gia tăng của các kênh bán hàng, giá bán trở nên một trong những mục tiêu quan trọng mà các nhà quản lý cần cân nhắc; tuy nhiên, việc xác định giá bán phù hợp không phải là một quyết định dễ dàng. Phân tích dữ liệu nâng cao (advanced analytics) sẽ giúp những nhà quản lý có thể có được một mức giá bền vững, đo lường sự khác biệt trong nhu cầu của từng phân khúc khách hàng, xác định các sản phẩm chủ chốt, việc quản lý phân phối các cửa hàng, phân tích hành vi người tiêu dùng qua các kênh phân phối,… Những thông tin này giúp nhà bán hàng có thể đưa ra các chiến lược về giá hợp lý khi đã cân nhắc sự phân hóa về đối tượng khách hàng, kênh bán hàng, đối thủ cạnh tranh, và các mặt hàng,… Theo một báo cao phân tích của Deloitte về “Analytics in Retail” cho thấy rằng các giải pháp về giá bán có thể làm cải thiện tỷ suất hoạt động (margin performance) ngay lập tức lên đến 2%-4% và tăng trưởng doanh thu lên đến 1%-2%.

* Cải thiện các chiến lược Marketing

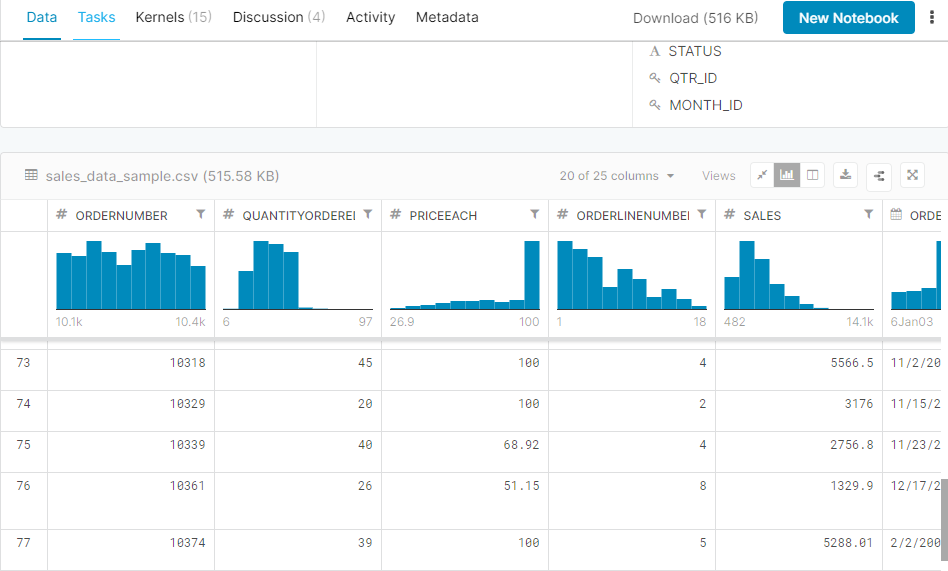
Ngày nay, các nghiên cứu cho thấy chiến lược marketing “cá nhân hóa” (personalized marketing) có tác động mạnh đến hành vi người tiêu dùng. Khi Facebook và Instagram có thể cho chúng ta thấy những quảng cáo liên quan dựa trên những thông tin chúng ta chia sẻ hằng ngày; thì các nhà bán hàng cũng có thể thu thập những thông tin mang tính cá nhân như sở thích, lịch sử tìm kiếm, xu hướng và thói quen tiêu dùng, … Do đó, thay vì tạo ra các chiến lược quảng bá lớn với chi phí khổng lồ nhưng lại có ít tác động trực tiếp, việc ứng dụng phân tích dữ liệu và dự báo có thể “cá nhân hóa” (personalize) các quá trình tiếp thị này. Bằng việc đưa ra các thông tin trực tiếp (direct messages) có nghĩa là chúng ta có thể kiểm soát không chỉ nội dung, mà còn khi nào, bằng cách nào, và tại sao, và cuối cùng, giúp cải thiện ROI và hiệu quả bán hàng cho công ty.

### Nội dung đề tài

Đây là bộ dữ liệu mô tả về dữ liệu bán hàng mẫu được lấy trên trang [www.kaggle.com](http://www.kaggle.com) bao gồm nhiều thông tin về dữ liệu bán hàng, thông tin khách hàng, thông tin về order, doanh thu, giao hàng, … Dữ liệu được dùng để phân khúc, phân tích, phân cụm khách hàng, sản phẩm và dữ liệu được dùng cho phân tích dữ liệu kinh doanh trong ngành bán lẻ.

Datasets này gồm 2824 dòng và có 25 cột thuộc tính

Link Dataset: <https://www.kaggle.com/kyanyoga/sample-sales-data>



## Mô tả dữ liệu gốc

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Ý nghĩa** | **Kiểu dữ liệu** |
| 1 | ORDERNUMBER | Mã Order của hóa đơn | INT |
| 2 | QUANTITYORDERED | Số lượng đã order | INT |
| 3 | PRICEEACH | Giá bán ra | FLOAT |
| 4 | ORDERLINENUMBER | Mã hóa đơn | INT |
| 5 | SALES | Tổng chi phí hóa đơn | FLOAT |
| 6 | ORDERDATE | Ngày order | DATE |
| 7 | STATUS | Trạng thái order | STRING |
| 8 | QTR\_ID | Quý | INT |
| 9 | MONTH\_ID | Tháng | INT |
| 10 | YEAR\_ID | Năm | INT |
| 11 | PRODUCTLINE | Tên sản phẩm | STRING |
| 12 | MSRP | Giá niêm yết | FLOAT |
| 13 | PRODUCTCODE | Mã sản phẩm | STRING |
| 14 | CUSTOMERNAME | Tên khách hàng (công ty) | STRING |
| 15 | PHONE | Số điện thoại khách hàng | STRING |
| 16 | ADDRESSLINE1 | Địa chỉ 1 của khách hàng | STRING |
| 17 | ADDRESSLINE2 | Địa chỉ 2 của khách hàng | STRING |
| 18 | CITY | Tên thành phố | STRING |
| 19 | STATE | Tên tiểu bang | STRING |
| 20 | POSTALCODE | Mã bưu điện | STRING |
| 21 | COUNTRY | Tên đất nước | STRING |
| 22 | TERRITORY | Tên vùng | STRING |
| 23 | CONTACTLASTNAME | Tên liên hệ cuối của khách hàng | STRING |
| 24 | CONTACTFIRSTNAME | Tên liên hệ đầu của khách hàng | STRING |
| 25 | DEALSIZE | Kích cỡ sản phẩm | STRING |

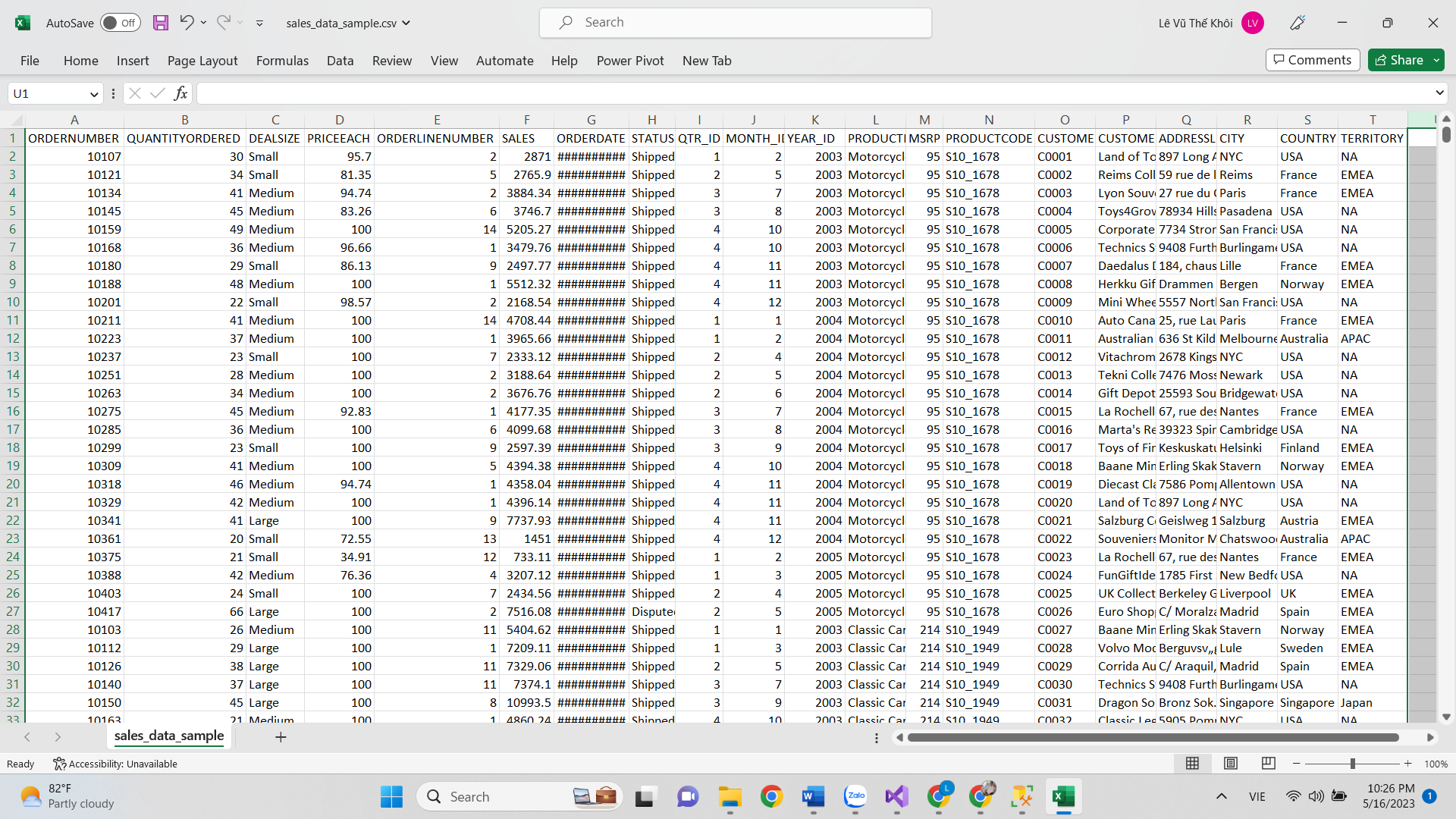
## Xử lý dữ liệu

Xử lý dữ liệu loại bỏ các trường không cần thiết và lọc các trường để tối ưu dữ liệuGraphical user interface, text, application

Description automatically generated

Xóa đi những thuộc tính không cần thiết: CONTACTFIRSTNAME, CONTACTLASTNAME, POSTALCODE, STATE, ADDRESSLINE2.

Tập dữ liệu sau khi xử lý:



## Đổ dữ liệu từ file CSV vào SQL source

Sử dụng ETL bên SSIS và Visual Studio 2019. Các bước thực hiện:

Tạo một database mới trong SQL Server

A screenshot of a computer

Description automatically generated

Sử dụng công cụ SSIS trong Visual Studio 2019 tạo một project mới và tạo kết nối đến cơ sở dữ liệu **SalesDataSource\_Datamining**

A screenshot of a computer

Description automatically generated with medium confidence

Tạo một package mới đặt tên là ExToSource



Sử dụng Data Flow Task, Flat File Source và OLE DB Destination

A screenshot of a computer

Description automatically generated with low confidence

A screenshot of a computer

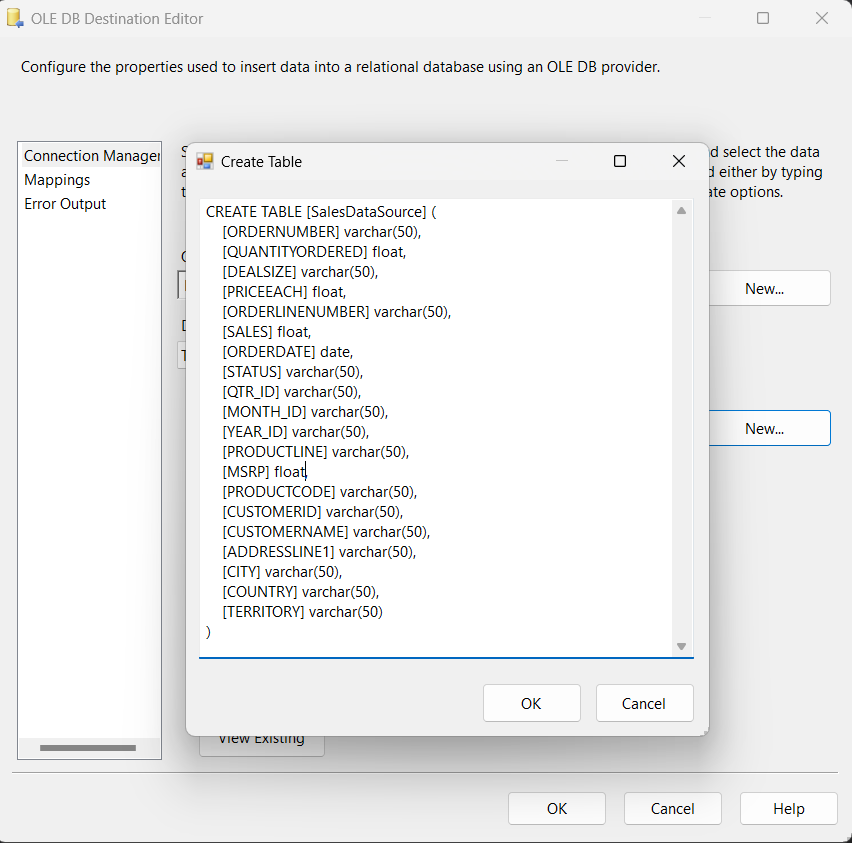
Description automatically generated

Tại CSV\_File chọn đường dẫn đến file data set

A screenshot of a computer

Description automatically generated

Tại SalesDataSource ta sẽ tạo bảng và chỉnh sữa các kiểu dữ liệu cho các thuộc tính



Mapping

A screenshot of a computer

Description automatically generated

Dữ liệu sau khi đổ xong

A screenshot of a computer

Description automatically generated

Tương tự với BankDataSource

A screenshot of a computer

Description automatically generated with medium confidence

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

1. **Giới thiệu về Khai phá dữ liệu**
   1. **Khái niệm và định nghĩa của Khai phá dữ liệu**

Khai phá dữ liệu (Data Mining) là quá trình khám phá và phân tích các mẫu, thông tin tiềm ẩn và tri thức có giá trị từ các tập dữ liệu lớn. Nó nhằm mục đích tìm ra các mối quan hệ, xu hướng và sự tương tác giữa các dữ liệu để giúp đưa ra quyết định thông minh và phát hiện những thông tin quan trọng đằng sau dữ liệu.

Khai phá dữ liệu sử dụng các phương pháp, thuật toán và kỹ thuật từ nhiều lĩnh vực như học máy (machine learning), thống kê, trí tuệ nhân tạo (artificial intelligence) và cơ sở dữ liệu (database) để xác định các mẫu và thông tin có ý nghĩa. Quá trình này bao gồm các bước như tiền xử lý dữ liệu, lựa chọn và áp dụng các thuật toán khai phá dữ liệu, đánh giá và tìm hiểu kết quả, và trình bày thông tin thu được một cách rõ ràng và hữu ích.

* 1. **Lợi ích và ứng dụng của Khai phá dữ liệu trong thực tế**

Khai phá dữ liệu mang lại nhiều lợi ích và có ứng dụng rộng rãi trong nhiều lĩnh vực thực tế. Dưới đây là một số lợi ích và ứng dụng quan trọng của Khai phá dữ liệu:

* **Phát hiện tri thức tiềm ẩn:** Khai phá dữ liệu cho phép tìm ra tri thức, mẫu và thông tin tiềm ẩn trong dữ liệu mà trước đây chưa được biết đến. Điều này giúp cải thiện sự hiểu biết về mối quan hệ giữa các yếu tố và hiệu quả trong quyết định.
* **Dự báo và dự đoán:** Khai phá dữ liệu có thể sử dụng để dự báo và dự đoán các xu hướng và sự kiện trong tương lai dựa trên các mẫu và thông tin hiện có. Điều này giúp đưa ra quyết định chiến lược và kế hoạch tốt hơn trong các lĩnh vực như tiếp thị, tài chính, y tế và nhiều lĩnh vực khác.
* **Phân loại và phân cụm:** Khai phá dữ liệu có thể phân loại các đối tượng vào các nhóm hoặc phân cụm dựa trên các thuộc tính và mẫu tương tự. Điều này giúp tổ chức và phân tích dữ liệu một cách hiệu quả, tìm ra nhóm đích hàng, xác định khách hàng tiềm năng và tối ưu hóa các quy trình kinh doanh.
* **Tối ưu hóa quy trình kinh doanh:** Khai phá dữ liệu giúp cải thiện hiệu suất và tối ưu hóa quy trình kinh doanh bằng cách xác định các yếu tố ảnh hưởng và tìm ra cách cải tiến. Nó có thể giúp tối thiểu hóa lãng phí, tăng cường hiệu quả vận hành, nâng cao chất lượng sản phẩm và dịch vụ, và tăng khả năng cạnh tranh của doanh nghiệp.
* **Phân tích khách hàng và tiếp thị cá nhân hóa:** Khai phá dữ liệu cho phép phân tích thông tin về khách hàng, xu hướng tiêu dùng và hành vi người dùng để tạo ra các chiến lược tiếp thị cá nhân hóa. Điều này giúp tăng cường tương tác với khách hàng, tối ưu hóa chiến dịch quảng cáo, nâng cao trải nghiệm khách hàng và tăng cường khả năng tương tác và chuyển đổi.

Các lợi ích và ứng dụng của Khai phá dữ liệu trong thực tế là vô cùng đa dạng và có thể tùy thuộc vào lĩnh vực cụ thể mà nó được áp dụng. Tuy nhiên, trong mọi trường hợp, Khai phá dữ liệu đóng vai trò quan trọng trong việc phân tích và tìm hiểu dữ liệu để đưa ra quyết định thông minh và tạo ra giá trị cho doanh nghiệp và xã hội.

1. **Phương pháp Khai phá dữ liệu**
2. **Các khái niệm cơ bản trong Khai phá dữ liệu**
   1. **Dữ liệu và các loại hình dữ liệu**

Trong lĩnh vực Khai phá dữ liệu, dữ liệu đóng vai trò quan trọng là nguồn tài nguyên để phân tích và khám phá. Dữ liệu có thể được hiểu là tập hợp các thông tin, sự kiện, số liệu hoặc bất kỳ thông tin nào có thể được thu thập, lưu trữ và xử lý. Có nhiều loại hình dữ liệu khác nhau được sử dụng trong Khai phá dữ liệu, bao gồm:

* **Dữ liệu số (Numerical data):** Đây là loại dữ liệu biểu diễn bằng các giá trị số. Dữ liệu số có thể là dữ liệu liên tục (ví dụ: chiều cao, nhiệt độ) hoặc dữ liệu rời rạc (ví dụ: số lượng, điểm số). Dữ liệu số thường được sử dụng trong các phép tính toán, phân tích thống kê và dự đoán.
* **Dữ liệu văn bản (Text data):** Dữ liệu văn bản là các chuỗi ký tự, bao gồm các văn bản, tài liệu, bài viết, email, tin nhắn và nhiều hình thức khác. Dữ liệu văn bản thường phức tạp và đa dạng, và cần các kỹ thuật xử lý đặc biệt để trích xuất thông tin và tri thức từ dữ liệu này.
* **Dữ liệu hình ảnh (Image data):** Dữ liệu hình ảnh là các dữ liệu biểu diễn dưới dạng hình ảnh hoặc đồ họa. Đây có thể là các hình ảnh tĩnh, bức tranh, biểu đồ hoặc các đối tượng hình ảnh khác. Dữ liệu hình ảnh được sử dụng trong các lĩnh vực như nhận dạng khuôn mặt, phân loại hình ảnh và nhận dạng đối tượng.
* **Dữ liệu âm thanh (Audio data):** Dữ liệu âm thanh là các dữ liệu biểu diễn âm thanh, bao gồm các tệp âm thanh, nhạc, giọng nói và các tín hiệu âm thanh khác. Dữ liệu âm thanh được sử dụng trong các lĩnh vực như nhận dạng giọng nói, xử lý âm thanh và xác định tính chất âm thanh.
* **Dữ liệu chuỗi thời gian (Time series data):** Dữ liệu chuỗi thời gian là các dữ liệu được thu thập liên tục theo thời gian. Đây có thể là dữ liệu về thời tiết, giá cổ phiếu, dữ liệu giao thông và nhiều loại dữ liệu khác. Dữ liệu chuỗi thời gian thường được sử dụng trong dự báo, phân tích xu hướng và phát hiện biểu đồ.

Các loại hình dữ liệu trên đều có đặc điểm riêng và yêu cầu các kỹ thuật và phương pháp xử lý khác nhau trong quá trình Khai phá dữ liệu. Việc hiểu và xử lý chính xác các loại dữ liệu này là rất quan trọng để thu được những thông tin quan trọng và tri thức có giá trị từ dữ liệu.

* 1. **Mẫu dữ liệu, thuộc tính, tập dữ liệu**

Trong quá trình Khai phá dữ liệu, các khái niệm quan trọng cần được hiểu rõ bao gồm mẫu dữ liệu, thuộc tính và tập dữ liệu.

* **Mẫu dữ liệu (Data Instance):** Mẫu dữ liệu là một quan sát hoặc một ví dụ cụ thể trong tập dữ liệu. Nó đại diện cho một bộ các giá trị hoặc thuộc tính được đo hay quan sát tại một thời điểm cụ thể. Ví dụ, trong một tập dữ liệu khách hàng, mỗi khách hàng có thể được xem như một mẫu dữ liệu.
* **Thuộc tính (Attribute):** Thuộc tính là đặc trưng hoặc thông tin mà ta quan tâm thu thập và lưu trữ trong mỗi mẫu dữ liệu. Thuộc tính có thể là các giá trị số, nhị phân, văn bản, hình ảnh hoặc bất kỳ kiểu dữ liệu nào khác. Ví dụ, trong tập dữ liệu khách hàng, các thuộc tính có thể bao gồm tuổi, giới tính, địa chỉ, số điện thoại, và hành vi mua hàng.
* **Tập dữ liệu (Data Set):** Tập dữ liệu là một tập hợp các mẫu dữ liệu được tổ chức và lưu trữ cùng nhau. Nó bao gồm tất cả các mẫu dữ liệu và các thuộc tính tương ứng của chúng. Tập dữ liệu có thể bao gồm hàng ngàn hoặc hàng triệu mẫu dữ liệu, tùy thuộc vào quy mô của bài toán và nguồn dữ liệu sẵn có.

Trong quá trình Khai phá dữ liệu, chúng ta cần xác định và xử lý các thuộc tính của mỗi mẫu dữ liệu trong tập dữ liệu. Điều này bao gồm việc xác định kiểu dữ liệu, đặc tính và giá trị của thuộc tính. Sau đó, tập dữ liệu sẽ được sử dụng để áp dụng các thuật toán Khai phá dữ liệu như Association Rule Mining, Microsoft Clustering và Microsoft Decision Trees để khám phá thông tin ẩn trong dữ liệu và đưa ra các phân tích, dự đoán hoặc quyết định có ích trong thực tế.

* 1. **Quá trình tiền xử lý dữ liệu**

Quá trình tiền xử lý dữ liệu là giai đoạn quan trọng trong quá trình Khai phá dữ liệu. Nó bao gồm các bước chuẩn bị dữ liệu trước khi áp dụng các thuật toán khai phá. Mục tiêu của quá trình này là làm sạch, chuẩn hóa và chuyển đổi dữ liệu để đảm bảo tính toàn vẹn, chính xác và phù hợp với yêu cầu của thuật toán khai phá.

Các bước quan trọng trong quá trình tiền xử lý dữ liệu bao gồm:

* Loại bỏ dữ liệu lặp: Kiểm tra và loại bỏ các bản ghi trùng lặp trong tập dữ liệu. Dữ liệu lặp có thể gây nhiễu và ảnh hưởng đến kết quả khai phá.
* Xử lý dữ liệu thiếu: Kiểm tra và xử lý các giá trị thiếu trong tập dữ liệu. Dữ liệu thiếu có thể làm giảm chất lượng và độ tin cậy của quá trình khai phá.
* Chuẩn hóa dữ liệu: Thực hiện chuẩn hóa dữ liệu để đảm bảo rằng các thuộc tính có cùng phạm vi hoặc phân phối tương tự. Các phương pháp chuẩn hóa có thể bao gồm chuẩn hóa Min-Max, chuẩn hóa Z-score và chuẩn hóa decimal scaling.
* Biến đổi dữ liệu: Thực hiện biến đổi dữ liệu để chuyển đổi dữ liệu theo một phân phối khác hoặc theo một đơn vị khác. Biến đổi dữ liệu có thể bao gồm logarit, căn bậc hai, chuẩn hóa mũ, v.v.
* Rời rạc hóa dữ liệu: Biến đổi các thuộc tính số thành dạng rời rạc để xử lý dữ liệu dạng danh mục hoặc để áp dụng các thuật toán khai phá phù hợp.
* Giảm số chiều dữ liệu: Áp dụng các kỹ thuật giảm số chiều dữ liệu như PCA (Principal Component Analysis) để giảm số lượng thuộc tính trong tập dữ liệu, từ đó giảm chi phí tính toán và cải thiện hiệu suất khai phá.

Quá trình tiền xử lý dữ liệu đảm bảo tính chuẩn xác và phù hợp của dữ liệu trước khi áp dụng các thuật toán khai phá. Nó giúp tăng độ tin cậy và chất lượng của kết quả khai phá, từ đó cung cấp thông tin hữu ích và hỗ trợ quyết định trong thực tế.

1. **Các thuật toán trong Khai phá dữ liệu**
   1. **Association Rule Mining (Khai phá luật kết hợp)**

Association Rule Mining là một trong những phương pháp phổ biến trong lĩnh vực khai phá dữ liệu. Nó được sử dụng để tìm ra các quy tắc kết hợp (association rules) giữa các mục tiêu trong tập dữ liệu. Cụ thể, phương pháp này tìm kiếm các mẫu kết hợp giữa các thuộc tính hoặc các mục tiêu xảy ra cùng nhau trong dữ liệu.

Chi tiết phần "Microsoft Clustering" bao gồm các khái niệm và quy trình như sau:

1. **Khái niệm:**
   * Tập dữ liệu: Tập dữ liệu chứa các bản ghi về các sự kiện hoặc giao dịch.
   * Mục tiêu (Itemset): Tập hợp các mục hoặc thuộc tính được xem xét trong tập dữ liệu.
   * Hạng mục (Item): Một phần tử đơn lẻ trong tập mục tiêu.
   * Luật kết hợp (Association rule): Quy tắc mô tả mối quan hệ giữa các mục tiêu dựa trên sự xuất hiện chung trong tập dữ liệu.
   * Hệ số hỗ trợ (Support): Tỉ lệ xuất hiện của một mục tiêu hoặc một luật kết hợp trong tập dữ liệu.
   * Hệ số độ tin cậy (Confidence): Tỉ lệ mà một luật kết hợp được xác định theo đúng trong các trường hợp khi một mục tiêu xảy ra.
2. **Quy trình Association Rule Mining:**
   * Bước 1: Tạo tập ứng viên (Candidate itemset): Tìm tất cả các tập hợp con có thể có từ tập mục tiêu ban đầu.
   * Bước 2: Đo hỗ trợ và độ tin cậy: Đo hỗ trợ và độ tin cậy cho mỗi luật kết hợp trong tập ứng viên.
   * Bước 3: Lựa chọn luật kết hợp: Chọn các luật kết hợp dựa trên các ngưỡng hỗ trợ và độ tin cậy đã định trước.
   * Bước 4: Đánh giá và tinh chỉnh: Đánh giá và tinh chỉnh các luật kết hợp để tìm ra những luật quan trọng và hữu ích.
   * Bước 5: Đưa ra kết quả: Trình bày kết quả của quá trình khai phá luật kết hợp, bao gồm các luật kết hợp có hỗ trợ và độ tin cậy đạt ngưỡng.

**Lợi ích và ứng dụng của Association Rule Mining:**

* Tìm kiếm quy tắc kết hợp giữa các mục tiêu trong dữ liệu giúp hiểu rõ hơn về mối quan hệ giữa chúng.
* Cung cấp thông tin hữu ích trong lĩnh vực tiếp thị, dự báo, tư vấn và quyết định kinh doanh.
* Giúp phân tích sự ảnh hưởng và tương tác giữa các mục tiêu trong dữ liệu.
* Hỗ trợ việc đưa ra các quyết định dựa trên những quy tắc và xu hướng xuất hiện trong dữ liệu.

Tóm lại, phương pháp Association Rule Mining đóng vai trò quan trọng trong việc khai phá dữ liệu bằng cách tìm kiếm và phân tích các quy tắc kết hợp giữa các mục tiêu trong tập dữ liệu. Nó mang lại lợi ích và ứng dụng quan trọng trong việc hiểu và đưa ra quyết định dựa trên thông tin từ dữ liệu.

* 1. **Microsoft Clustering (Phân cụm)**

Microsoft Clustering là một trong những phương pháp phân cụm (clustering) phổ biến trong khai phá dữ liệu. Phương pháp này giúp tổ chức và nhóm các đối tượng dữ liệu vào các nhóm dựa trên sự tương đồng giữa chúng. Microsoft Clustering sử dụng các thuật toán phân cụm để xác định cấu trúc và mối quan hệ trong dữ liệu.

Chi tiết phần "Microsoft Clustering" bao gồm các khái niệm và quy trình như sau:

1. **Khái niệm:**
   * Phân cụm (Clustering): Quá trình tổ chức các đối tượng dữ liệu vào các nhóm (cụm) dựa trên sự tương đồng giữa chúng.
   * Đối tượng (Object): Một đơn vị dữ liệu cần được phân cụm.
   * Tương đồng (Similarity): Đánh giá mức độ giống nhau giữa các đối tượng dữ liệu.
   * Nhóm (Cluster): Tập hợp các đối tượng dữ liệu có tính chất tương tự nằm trong cùng một cụm.
   * Thuật toán phân cụm (Clustering algorithm): Các phương pháp tính toán và quy tắc được sử dụng để phân cụm dữ liệu.
2. **Quy trình Microsoft Clustering:**
   * Bước 1: Chuẩn bị dữ liệu: Tiền xử lý dữ liệu bằng cách loại bỏ nhiễu, xử lý giá trị thiếu và chuẩn hóa dữ liệu nếu cần.
   * Bước 2: Chọn thuật toán: Lựa chọn và cấu hình thuật toán phân cụm phù hợp dựa trên yêu cầu và tính chất của dữ liệu.
   * Bước 3: Xác định số lượng cụm: Xác định số lượng cụm mong muốn hoặc sử dụng các phương pháp định lượng để tìm ra số lượng cụm tối ưu.
   * Bước 4: Thực hiện phân cụm: Áp dụng thuật toán phân cụm để phân loại các đối tượng dữ liệu vào các nhóm tương ứng.
   * Bước 5: Đánh giá và tinh chỉnh: Đánh giá chất lượng của phân cụm và điều chỉnh các tham số nếu cần thiết.
   * Bước 6: Đưa ra kết quả: Trình bày kết quả của quá trình phân cụm, bao gồm sự phân chia của các đối tượng vào từng cụm và tính chất của mỗi cụm.

**Lợi ích và ứng dụng của Microsoft Clustering:**

* Nhận biết và tìm hiểu các nhóm trong dữ liệu mà không cần biết trước về cấu trúc của chúng.
* Phân tích đặc điểm và tính chất của từng nhóm, giúp hiểu rõ hơn về sự tương đồng và khác biệt giữa các đối tượng dữ liệu.
* Áp dụng trong lĩnh vực phân loại khách hàng, phân đoạn thị trường, phân tích hành vi người dùng, phân tích hệ thống gợi ý, và phân cụm dữ liệu lớn.
* Giúp tạo ra các chiến lược tiếp thị, quản lý khách hàng và ra quyết định dựa trên hiểu biết sâu sắc về sự tương đồng và nhóm hóa của dữ liệu.

Tóm lại, Microsoft Clustering là phương pháp phân cụm dữ liệu mạnh mẽ và linh hoạt trong khai phá dữ liệu. Nó giúp tổ chức và nhóm các đối tượng dữ liệu vào các nhóm tương tự nhau, đồng thời cung cấp lợi ích và ứng dụng quan trọng trong việc hiểu, phân tích và ra quyết định dựa trên thông tin từ dữ liệu.

* 1. **Microsoft Decision Trees (Cây quyết định)**

Microsoft Decision Trees là một trong những phương pháp quyết định (decision) phổ biến trong khai phá dữ liệu. Phương pháp này sử dụng cấu trúc cây quyết định để đưa ra các quyết định và dự đoán dựa trên việc phân tích các quan hệ và luật logic trong dữ liệu.

Chi tiết phần "Microsoft Decision Trees" bao gồm các khái niệm và quy trình như sau:

1. **Khái niệm:**
   * Cây quyết định (Decision Tree): Mô hình biểu diễn các quyết định và dự đoán dưới dạng cấu trúc cây, trong đó các nút đại diện cho các quyết định và các nhánh đại diện cho các kết quả dự đoán hoặc quyết định tiếp theo.
   * Quyết định (Decision): Các hành động hoặc kết quả được đưa ra dựa trên các luật logic hoặc quan hệ giữa các thuộc tính trong dữ liệu.
   * Quy tắc (Rule): Các điều kiện và hành động được xác định để đưa ra quyết định dựa trên dữ liệu.
   * Đặc tính (Attribute): Các thuộc tính hoặc đặc trưng của dữ liệu được sử dụng để đưa ra quyết định hoặc dự đoán.
   * Cắt nhánh (Split): Quá trình phân tách cây quyết định thành các nhánh con dựa trên thuộc tính và giá trị của chúng.
2. **Quy trình Microsoft Decision Trees:**
   * Bước 1: Chuẩn bị dữ liệu: Tiền xử lý dữ liệu bằng cách loại bỏ nhiễu, xử lý giá trị thiếu và chuẩn hóa dữ liệu nếu cần.
   * Bước 2: Chọn thuật toán: Lựa chọn và cấu hình thuật toán cây quyết định phù hợp dựa trên yêu cầu và tính chất của dữ liệu.
   * Bước 3: Xây dựng cây quyết định: Áp dụng thuật toán để xây dựng cây quyết định bằng cách chia các nút và cắt nhánh dựa trên các thuộc tính và giá trị của chúng.
   * Bước 4: Đánh giá và tinh chỉnh: Đánh giá chất lượng của cây quyết định và điều chỉnh các tham số nếu cần thiết.
   * Bước 5: Dự đoán và đánh giá: Sử dụng cây quyết định để dự đoán kết quả và đánh giá độ chính xác của mô hình.

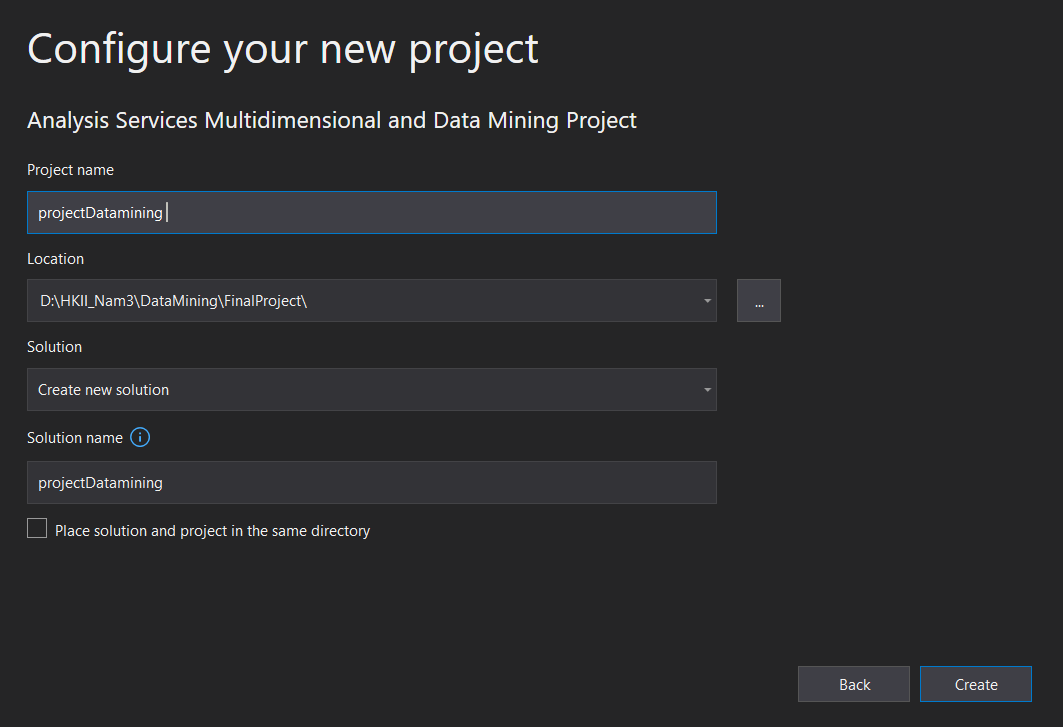
Lợi ích và ứng dụng của Microsoft Decision Trees:

* Dễ hiểu và diễn giải: Cây quyết định có cấu trúc rõ ràng và dễ hiểu, giúp người dùng dễ dàng diễn giải kết quả và quyết định của mô hình.
* Hỗ trợ ra quyết định: Mô hình cây quyết định cung cấp quyết định và dự đoán dựa trên các luật logic và quan hệ trong dữ liệu, hỗ trợ trong việc ra quyết định và lập kế hoạch.
* Khám phá thông tin: Phân tích cây quyết định giúp hiểu rõ hơn về mối quan hệ giữa các thuộc tính và quyết định, khám phá thông tin ẩn trong dữ liệu.
* Ứng dụng rộng rãi: Microsoft Decision Trees được áp dụng trong nhiều lĩnh vực như phân loại dữ liệu, dự báo, phân tích hành vi khách hàng, và hệ thống gợi ý.

Tóm lại, Microsoft Decision Trees là một phương pháp quyết định quan trọng trong khai phá dữ liệu. Nó sử dụng cấu trúc cây quyết định để đưa ra các quyết định và dự đoán dựa trên quan hệ và luật logic trong dữ liệu. Microsoft Decision Trees cung cấp lợi ích và ứng dụng quan trọng trong việc hiểu, diễn giải, và ra quyết định dựa trên thông tin từ dữ liệu.

# CHƯƠNG 3: PHÂN TÍCH DỮ LIỆU (SSAS)

Tạo một project mới

****

Tạo một data source và đặt tên là **Sales Data Source Datamining.ds**

A screenshot of a computer

Description automatically generated with medium confidence

Tạo kết nối đến Server name: **LEVUTHEKHOI**

Kết nối đến cơ sở dữ liệu: **SalesDataSource\_Datamining**

A screenshot of a computer

Description automatically generated

Tạo một Data Source View

A screenshot of a computer

Description automatically generated with medium confidence

Sử dụng bảng SalesDataSource

A screenshot of a computer

Description automatically generated

Đặt tên là **Sales Data Source Datamining**

A screenshot of a computer

Description automatically generated with medium confidence

nhấn **Finish**

A screenshot of a computer

Description automatically generated

**2.1 Giải thuật Microsoft Clustering**

Tạo một Mining Structures và chọn giải thuật Clustering

**A screenshot of a computer

Description automatically generated**

Chọn bảng Sales Data Source

A screenshot of a computer

Description automatically generated with medium confidence

Chọn khóa chính, các thuộc tính đầu vào và click Next

A screenshot of a computer

Description automatically generated with medium confidence

Chọn % test khoảng 20%

A screenshot of a test

Description automatically generated with medium confidence

Đặt tên là **Sales Data Clustering.dmm**

A picture containing text, screenshot, font

Description automatically generated

click **Next**

A screenshot of a computer

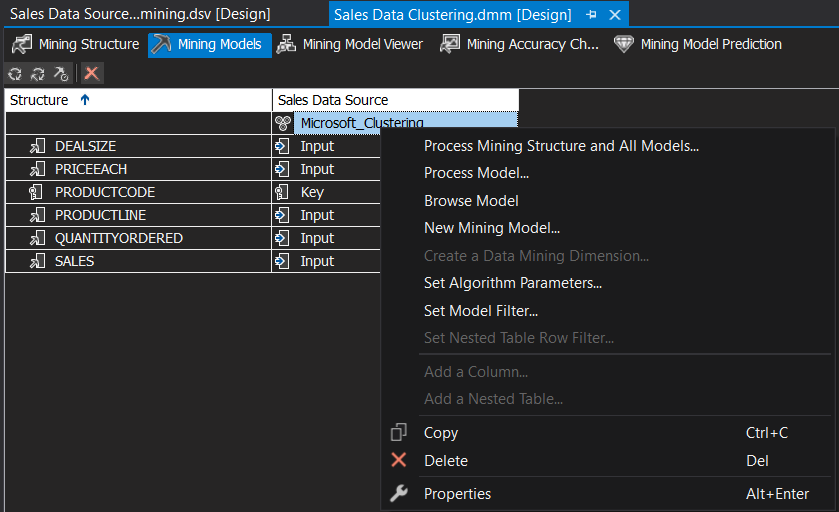
Description automatically generated

Click chuột phải vào project và chọn Deploy

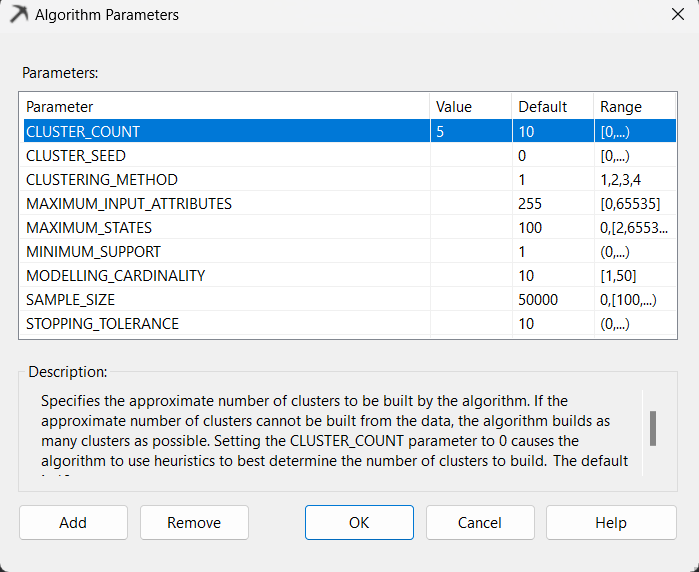
A screenshot of a computer

Description automatically generated

Chọn số cụm mà ta muốn: tại Mining Models chọn Set Algorithm Parameters



Số cụm mà nhóm chọn là 5 cụm



Click **Mining Model Viewer**

Mức độ đậm nhạt của các cụm sẽ phản ánh số dòng dữ liệu trong các cụm nhiều hay ít

Với Model view như hình thì đây là số cụm của dòng dữ liệu. Tổng là 88 dòng trong đó cụm 1 là màu đậm nhất tương ứng với 33 dòng, kế tiếp là cụm 2 với 20 dòng dữ liệu, ít hơn 1 dòng dữ liệu là cụm 3, tiếp sau đó là cụm 4 với 14 dòng dữ liệu và cuối cùng là cụm 5 cụm nhạt nhất với 2 dòng dữ liệu.



A screenshot of a computer

Description automatically generated

Click vào **Cluster Profiles** (dùng để xem rõ đặc điểm của các cụm)

A screenshot of a computer

Description automatically generated with medium confidence

- DEALSIZE (Size của sản phẩm: S, M, L): Cụm 1, cụm 4 và cụm 5 chiếm tỉ lệ là size M cụm 2 và 3 là size L

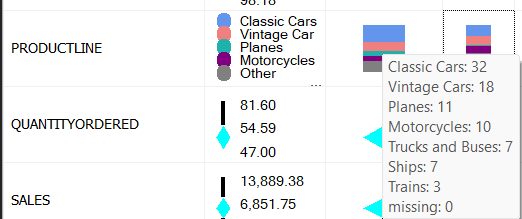
- PRICEEACH (Giá bán ra): Thấp nhất là 98.18$ và cao nhất là 100.00$, trung bình là 99.92$ +/- 0.58 là độ lệch chuẩn.

A picture containing text, screenshot, font, line

Description automatically generated

+ Ở 4 cụm đầu tiên giá bán ra của tất cả sản phẩm đều đạt 100.00$ tuy nhiên ở cụm 5 giá bán ra chỉ đạt giá trị trung bình khoảng 96.71 +/- 2.08

- PRODUCLINE(Loại sản phẩm): Dựa vào số liệu ta có thể thấy Classic Cars chiếm số lượng nhiều nhất với 32 loại và số lượng ít nhất là Trains chỉ có 3 loại



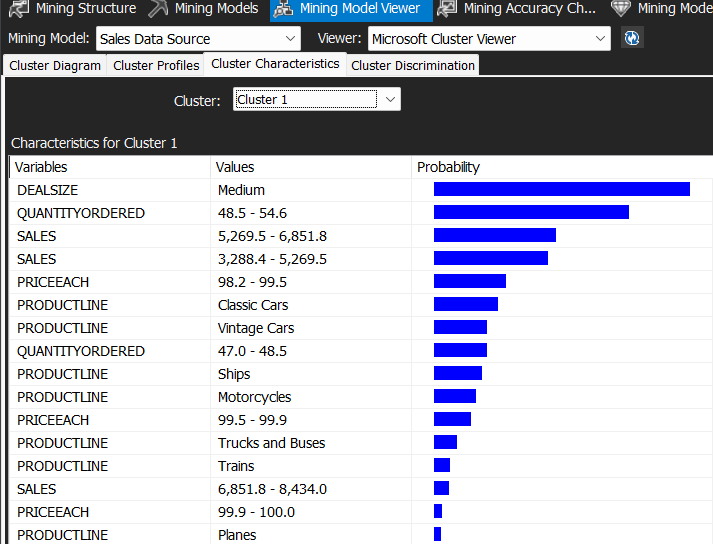
- QUANTITYORDERED (Số lượng đặt hàng): Với tổng giá trị trung bình của số lượng đặt hàng là 54.59 +/- 9.00. Số lượng cao nhất là ở cụm 3 và thấp nhất là cụm 1



- SALES: Giá trị trung bình của doanh số là 6,851.75+/- 2,345.88 tr. Doanh số thấp nhất là ở cụm 5 và doanh số cao nhất thuộc về cụm 2



Để xem rõ hơn về các cụm ta có thể click vào **Cluster Characteristics** và chọn cụm muốn xem tại **Cluster**



Nếu muốn so sánh giữa 2 cụm với nhau thì ta chọn **Cluster Discrimination** và chọn 2 cụm muốn so sánh. Ở đây ta chọn so sánh hai cụm 1 và 3

A screenshot of a computer

Description automatically generated with medium confidence

Cuối cùng ta có thể dự đoán sản phẩn đó thuộc cụm nào bằng cách click vào nút **Mining Model Pecdiction** và cung cấp những thông tin và nhấn **Result**

A screenshot of a computer

Description automatically generated

Kết quả cho ra thì sản phẩm này thuộc cụm 5

A screenshot of a computer

Description automatically generated with medium confidence

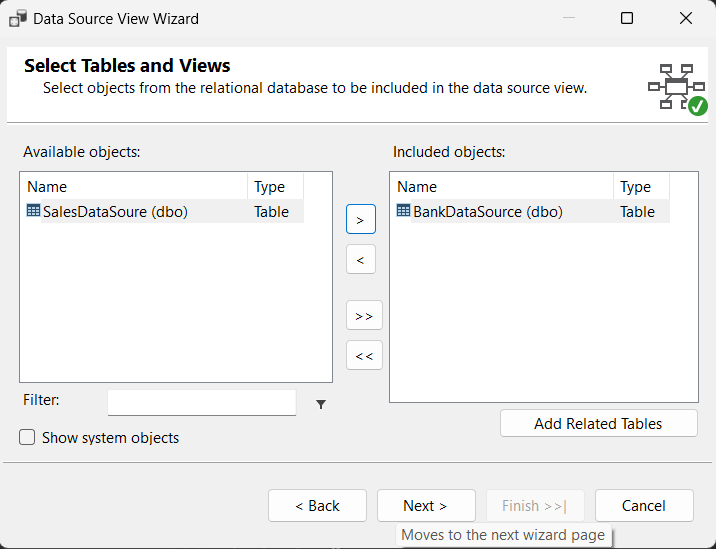
**2.2 Giải thuật Decesion Tree**

Tạo một Data Source View

A screenshot of a computer

Description automatically generated with medium confidence

Sử dụng bảng BankDataSource



Đặt tên là **Bank Data Source Datamining**

A screenshot of a computer

Description automatically generated with low confidence

nhấn **Finish**

A screenshot of a computer

Description automatically generated

Tạo một Mining Structures và chọn giải thuật Clustering

A screenshot of a computer

Description automatically generated

Chọn bảng Bank Data Source

A screenshot of a computer

Description automatically generated with medium confidence

Chọn khóa chính (id), các thuộc tính đầu vào (tất cả), biến dự đoán (có cho vay hay không) và click Next

A screenshot of a computer

Description automatically generated

Đặt tên là **Sales Data Clustering.dmm**

A picture containing text, font, screenshot

Description automatically generated

click **Next**

A screenshot of a computer

Description automatically generated

Click chuột phải vào project và chọn Deploy

A screenshot of a computer

Description automatically generated

Mức độ ảnh hưởng của từng thuộc tính đối với mô hình

A screenshot of a computer

Description automatically generated

Đọc mô hình: Click chuột phải và chọn **Show Legend**

**A computer screen shot of a computer

Description automatically generated with low confidence**

Bảng chọn **Mining Legend** sẽ xuất hiện và theo như mô hình dự báo thì nếu người này không có con, đã có gia đình, không có tài sản thế chấp và có tài khoảng tiết kiệm thì khả năng ngân hàng sẽ cho vay là 89.23% và khả năng không cho vay là 10.77%

**A screenshot of a computer

Description automatically generated with medium confidence**

Đọc một biến khác

A picture containing text, screenshot, software, multimedia software

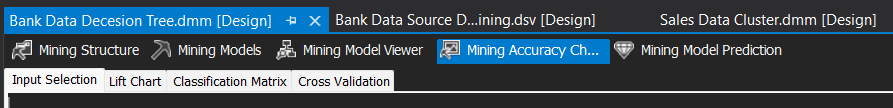
Description automatically generated

Tương tự bảng chọn Mining Lenged sẽ xuất hiện, mô hình này sẽ dự đoán kết quả của người này là đồng ý cho vay với điều kiện người này có 1 con và thu nhập hàng năm lớn hơn 16160.868$ với khả năng là 97.79% và khả năng không cho vay là 2.81%.

A screenshot of a computer

Description automatically generated with medium confidence

Để xem độ chính xác của mô hình thì ta chọn vào **Mining Accurary Chat** và ở đây ta có thể xem bằng 3 cách: **Lift Chart**, **Classification Matrix** và **Cross Validation**



**Lift Chart** dự đoán được Score của mô hình là 89%

A screenshot of a computer

Description automatically generated with medium confidence

**Classification Matrix** sẽ cho ra một ma trận như hình

A screenshot of a computer

Description automatically generated

Để tính được độ chính xác bằng phương pháp này ta sẽ làm như sau:

Accuracy = (B2+C3) / SUM(B2:C3) = 86.67%

Giá trị trên trên đường chéo B2, C3 là giá trị đúng, 2 giá trị còn lại là giá trị sai.

A screenshot of a computer

Description automatically generated with medium confidence

Để dự đoán một người có thể vay tiền được hay không thì ta chọn vào **Mining Model Prediction**

A screenshot of a computer

Description automatically generated

Sau khi thiết lập xong các thông số thì ta chọn Result

A screenshot of a computer

Description automatically generated

Kết quả sẽ là đồng ý cho vay với xác suất 93%

A screenshot of a computer

Description automatically generated with medium confidence

**2.3 Giải thuật Association Rule**

Tạo một Mining Structures và chọn giải thuật **Association Rule**

A screenshot of a computer

Description automatically generated

Chọn bảng Bank Data Source

A screenshot of a computer

Description automatically generated with medium confidence

Chọn key là id, input là tất cả các thuộc tính đầu vào và các biến dự đoán là: Current Account, Lend, Mortgage, Savings Account.

A screenshot of a computer

Description automatically generated

Đặt tên là Bank Data AR và nhấn Finish

A screenshot of a computer

Description automatically generated with medium confidence

Chuột phải vào project và chọn **Process** để Deploy

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

Sau khi Deploy xong, để xem được kết quả thì ta click vào **Mining Model Viewer** để xem kết quả và click vào tab **Rules** để xem một số luật mà ta tìm được

+ Ví dụ: Ở luật đầu tiên thì xác xuất là 100% khi mà lương hàng năm >=46431.5375, có xe thì kết quả dự đoán sẽ là có tài khoảng tiết kiệm

A screenshot of a computer

Description automatically generated

**Dependency Network** chính là mạng phụ thuộc, ở đây thì nó sẽ mô tả mức độ phụ thuộc giữa các Item trong luật kết hợp theo mức độ mạnh yếu.

Trong ảnh chính là độ mạnh nhất của luật kết hợp, tức là nếu như lương của người này lớn hơn hoặc bằng con số 46431… thì đồng ý cho họ vay.

A screenshot of a computer

Description automatically generated with medium confidence

Để dự báo dự đoán thì ta click vào **Mining Model Prediction** và chọn **Singleton Query**

A screenshot of a computer

Description automatically generated

Chọn mô hình và chọn các biến dự đoán

A screenshot of a computer

Description automatically generated

Kết quả sau khi nhấn Result thì người này có tài khoản tiết kiệm, có tài khoản hiện tại và có thể cho vay.

A screenshot of a computer

Description automatically generated

# CHƯƠNG 4: KẾT LUẬN

## Môi trường làm việc

* Hệ điều hành: Window 11.
* Hệ quản trị cơ sở dữ liệu: SQL Server 2019 (Enterprise Edition).
* Môi trường thực hiện thao tác: Microsoft Visual Studio 2019 (Professional).

## Phân công nhiệm vụ

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Công việc** | **Decision Tree** | **Association Rule** | **Microsoft Clustering** | **Trực quan hoá trên Python** | **Trực quan hóa trên Power BI** | **Tiến độ** |
| Lê Vũ  Thế Khôi | x | x | x |  | x | 100% |
| Ngô Văn Khánh |  |  | x | x |  | 100% |
| Nguyễn Võ Song Toàn | x |  |  |  | x | 100% |

## Hướng phát triển

* Thực hiện thao tác Data Mining trên dữ liệu. Từ đó rút ra những kiến thức ẩn trong dữ liệu.