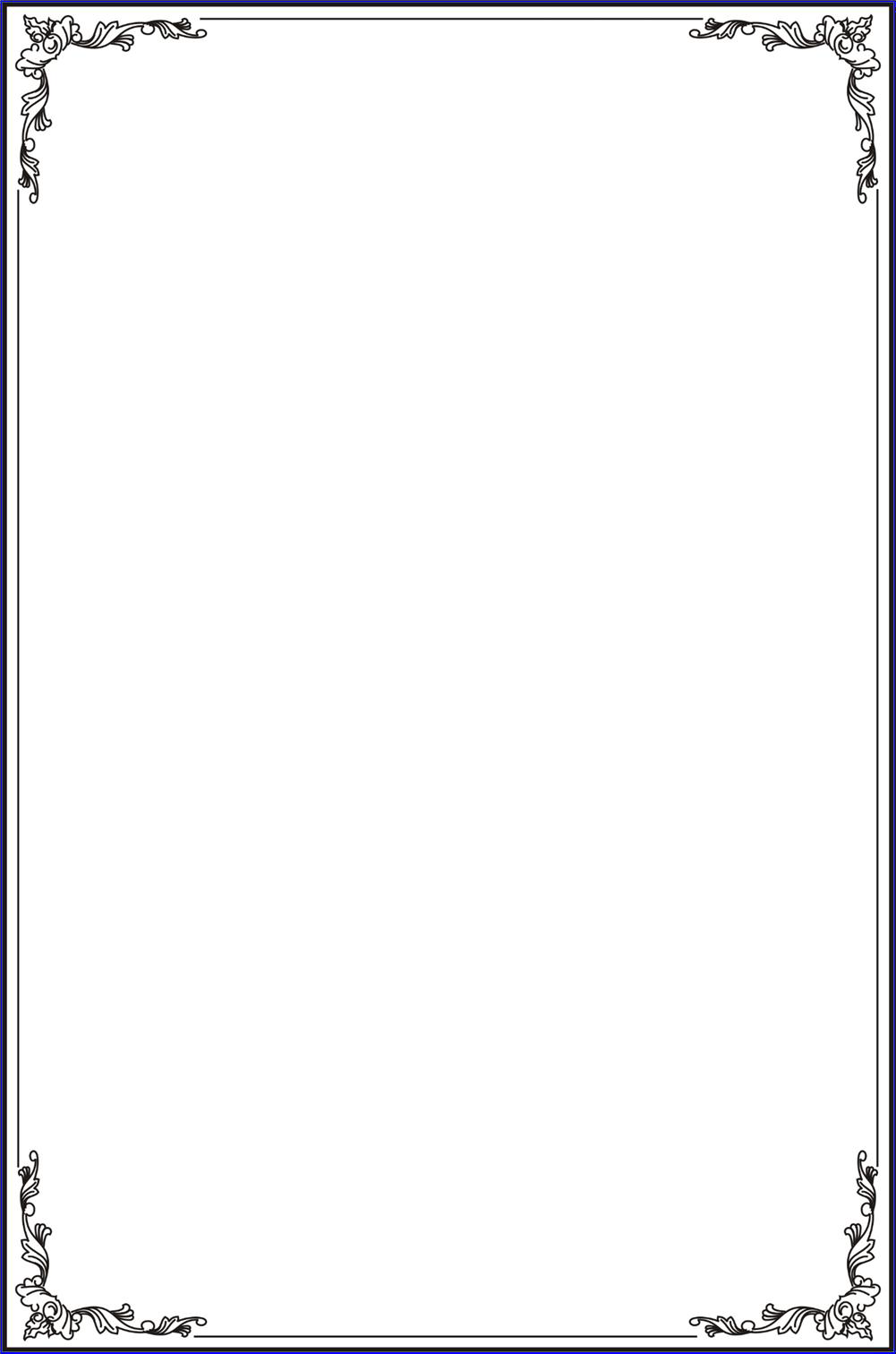
**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHI MINH**



**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: DAWH430784\_01**

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

**Nhóm sinh viên thực hiện: Nhóm 3**

**Nguyễn Võ Song Toàn – 20133097**

**Lê Vũ Thế Khôi – 20133057**

**Ngô Văn Khánh – 20133056**

TP Hồ Chí Minh, tháng 05 năm 2023

[LỜI CẢM ƠN 1](#_Toc134969495)

[CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI 2](#_Toc134969496)

[1.1. Giới thiệu đề tài 2](#_Toc134969497)

[1.1.1. Lý do chọn đề tài 2](#_Toc134969498)

[1.1.2. Nội dung đề tài 3](#_Toc134969499)

[1.2. Mô tả dữ liệu gốc 4](#_Toc134969500)

[1.3. Xử lý dữ liệu 5](#_Toc134969501)

[CHƯƠNG : KẾT LUẬN 56](#_Toc134969526)

[4.1. Môi trường làm việc 56](#_Toc134969527)

[4.2. Kết quả đạt được 56](#_Toc134969528)

[4.3. Hướng phát triển 56](#_Toc134969529)

# 

# 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 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

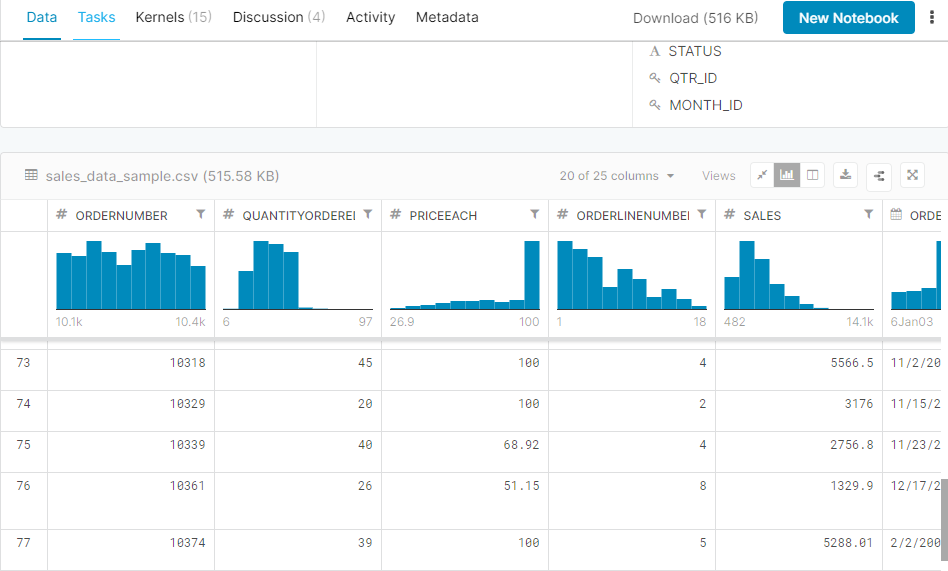
Ngày nay, các nghiên cứu cho thấy chiến lược marketing “cá nhân hóa” (personalized marketing) có tác động mạnh đến hành vi người tiêu dùng. Khi Facebook và Instagram có thể cho chúng ta thấy những quảng cáo liên quan dựa trên những thông tin chúng ta chia sẻ hằng ngày; thì các nhà bán hàng cũng có thể 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 |

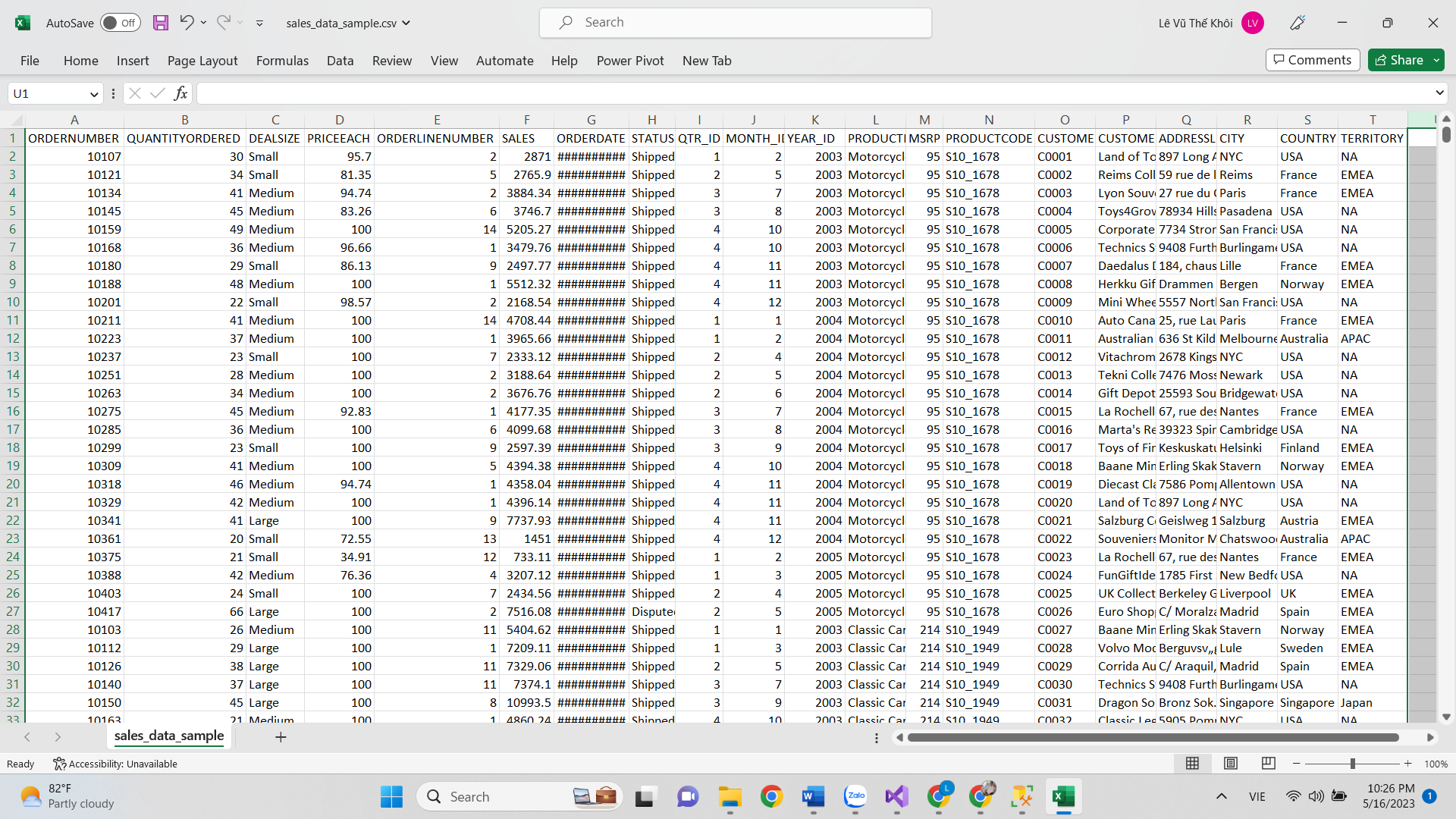
## Xử lý dữ liệu

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:



**1.4 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

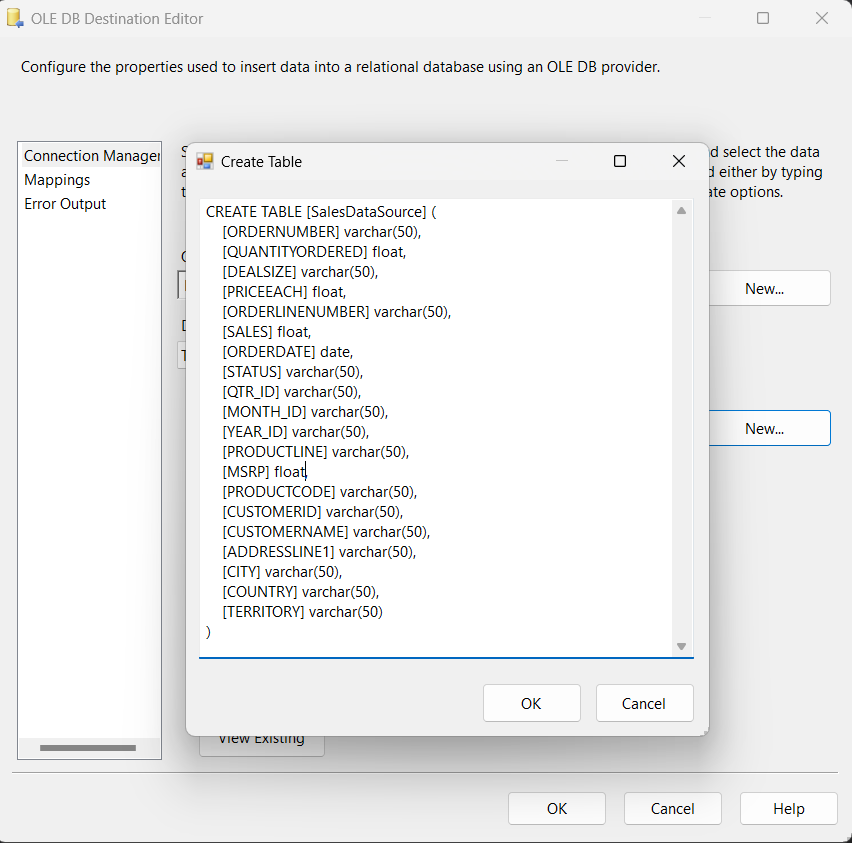
Description automatically generated

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

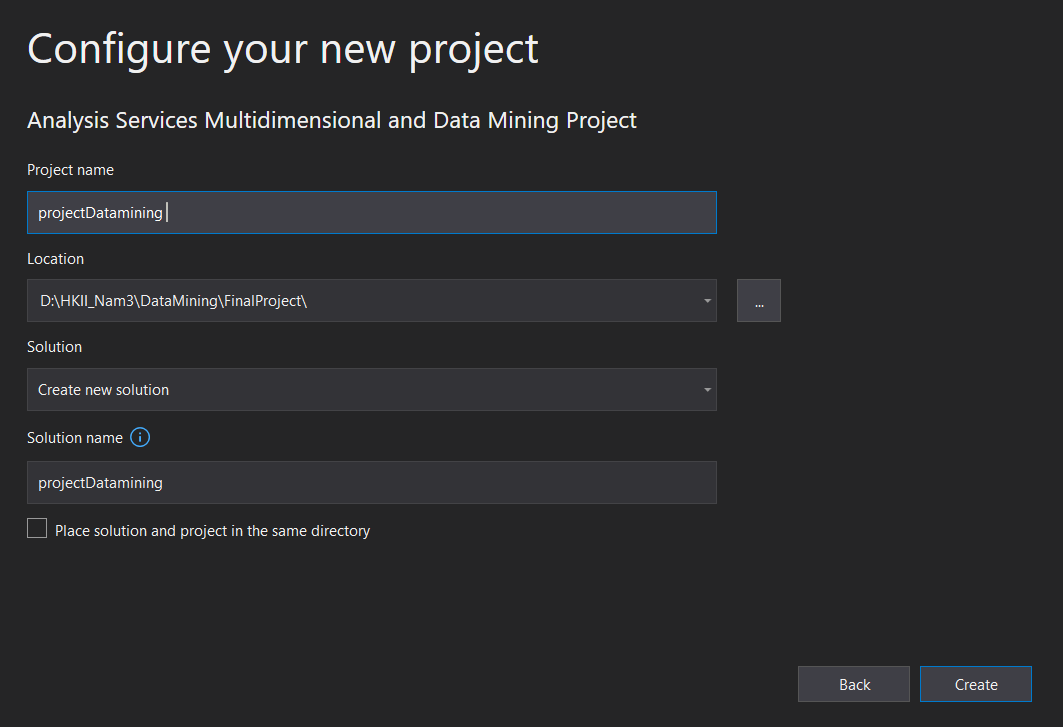
Tương tự với BankDataSource

A screenshot of a computer

Description automatically generated with medium confidence

**CHƯƠNG 2: 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

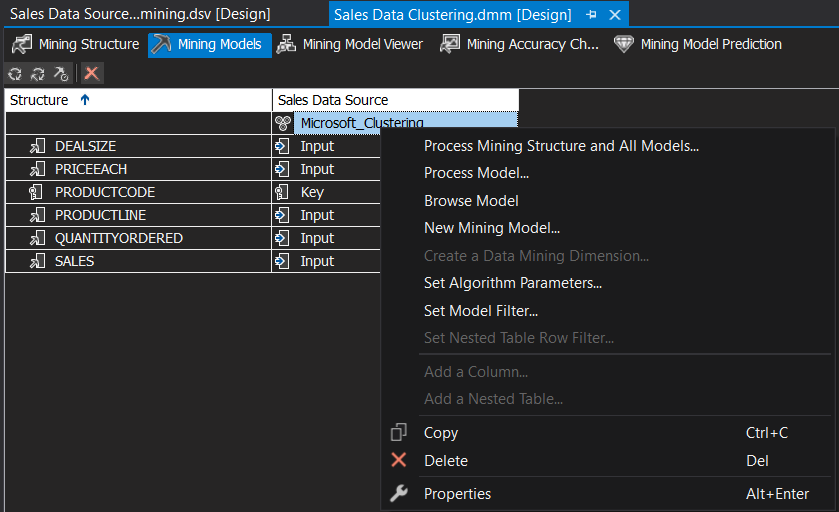
Description automatically generated

Click chuột phải vào project và chọn Deploy

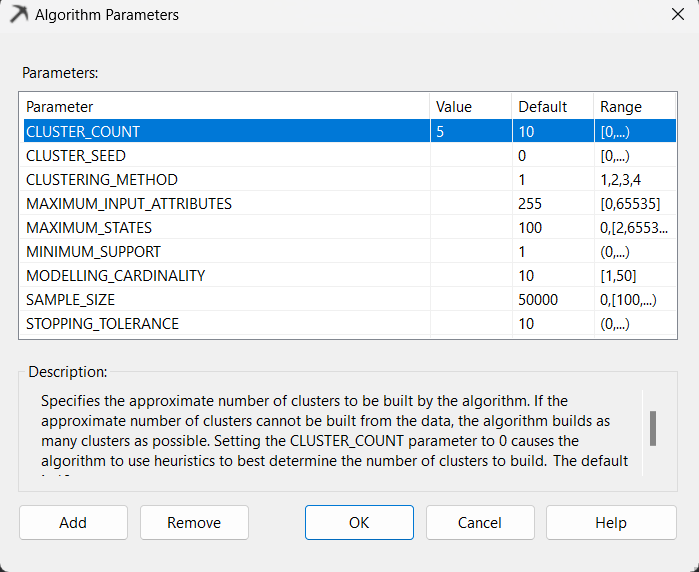
A screenshot of a computer

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

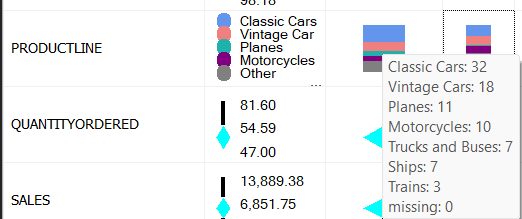
- PRICEEACH (Giá bán ra): Thấp nhất là 98.18$ và cao nhất là 100.00$, trung bình là 99.92$ +/- 0.58 là độ lệch chuẩn.

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



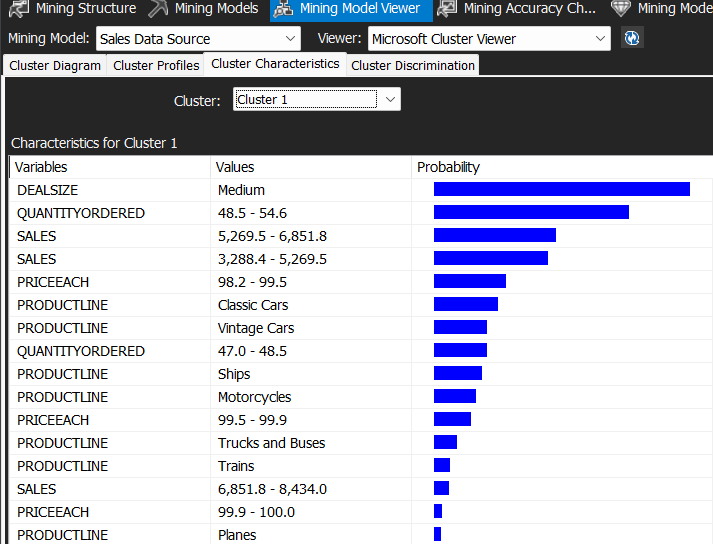
- QUANTITYORDERED (Số lượng đặt hàng): Với tổng giá trị trung bình của số lượng đặt hàng là 54.59 +/- 9.00. Số lượng cao nhất là ở cụm 3 và thấp nhất là cụm 1



- 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

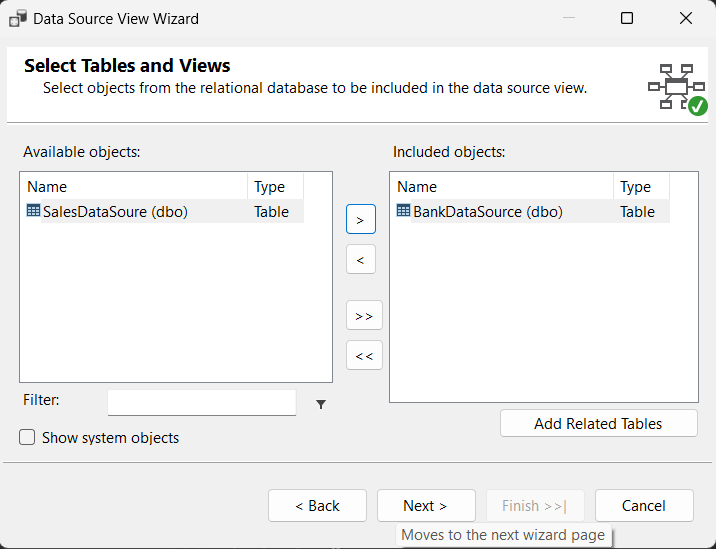
**2.2 Giải thuật Decesion Tree**

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

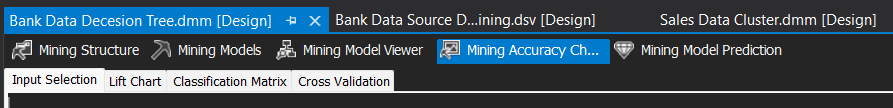
Description automatically generated

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).

## Kết quả đạt được

## Hướng phát triển

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