

BỘ TÀI CHÍNH  
TRƯỜNG ĐẠI HỌC TÀI CHÍNH – MARKETING  
KHOA CÔNG NGHỆ THÔNG TIN



**ĐỒ ÁN MÔN HỌC  
KHAI PHÁ DỮ LIỆU**

Tên đề tài:

**ỨNG DỤNG IBM SPSS MODELER ĐỂ  
THỰC HIỆN KHAI PHÁ DỮ LIỆU  
NHẬN DIỆN XU HƯỚNG MUA HÀNG  
CỦA KHÁCH HÀNG**

Giảng viên hướng dẫn: ThS. Nguyễn Thị Trần Lộc

Danh sách nhóm sinh viên thực hiện: Nhóm 3

1. Nguyễn Tấn Tài – 2021010274
2. Nguyễn Minh Tín – 2021010317

Mã lớp học phần: 2311112005902

TP. HCM, THÁNG 4 NĂM 2023

## TRÍCH YẾU

Dữ liệu và chuyển nó thành một cấu trúc dễ hiểu để sử dụng tiếp. Ngoài bước phân tích thô, nó còn liên quan tới cơ sở dữ liệu và các khía cạnh quản lý dữ liệu, xử lý dữ liệu trước, suy xét mô hình và suy luận thống kê, các thước đo thú vị, các cân nhắc phức tạp, xuất kết quả về các cấu trúc được phát hiện, hiện hình hóa và cập nhật trực tuyến. Khai thác dữ liệu là bước phân tích của quá trình "khám phá kiến thức trong cơ sở dữ liệu" hoặc KDD.

Tuy nhiên, vẫn có các mối lo ngại về tính riêng tư gắn với việc khai thác dữ liệu. Ví dụ, nếu một ông chủ có quyền truy xuất vào các hồ sơ y tế, họ có thể loại những người có bệnh tiểu đường hay bệnh tim. Việc loại ra những nhân viên như vậy sẽ cắt giảm chi phí bảo hiểm, nhưng tạo ra các vấn đề về tính hợp pháp và đạo đức. Khai thác dữ liệu các tập dữ liệu thương mại hay chính phủ cho các mục đích áp đặt luật pháp và an ninh quốc gia cũng là những mối lo ngại về tính riêng tư đang tăng cao.

Có nhiều cách sử dụng hợp lý với khai thác dữ liệu. Ví dụ, một CSDL các mô tả về thuốc được thực hiện bởi một nhóm người có thể được dùng để tìm kiếm sự kết hợp của các loại thuốc tạo ra các phản ứng (hóa học) khác nhau. Vì việc kết hợp có thể chỉ xảy ra trong một phần 1000 người, một trường hợp đơn lẻ là rất khó phát hiện. Một dự án liên quan đến y tế như vậy có thể giúp giảm số lượng phản ứng của thuốc và có khả năng cứu sống con người. Không may mắn là, vẫn có khả năng lạm dụng đối với một CSDL như vậy.

Về cơ bản, khai thác dữ liệu đưa ra các thông tin mà sẽ không có sẵn được. Nó phải được chuyển đổi sang một dạng khác để trở nên có nghĩa. Khi dữ liệu thu thập được liên quan đến các cá nhân, thì có nhiều câu hỏi đặt ra liên quan đến tính riêng tư, tính hợp pháp, và đạo đức.

Cũng vì vậy mà nhóm chúng em chọn đề tài “Ứng dụng IBM SPSS Modeler để thực hiện khai phá dữ liệu trong nhận diện xu hướng mua hàng của khách hàng.”, để từ đó tích hợp chương trình khách hàng thân thiết nhằm xây dựng mối quan hệ dài lâu với khách hàng.

## MỤC LỤC

<b>TRÍCH YẾU .....</b>	<b>i</b>
<b>MỤC LỤC .....</b>	<b>ii</b>
<b>LỜI CẢM ƠN .....</b>	<b>iv</b>
<b>NHẬN XÉT VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN 1 .....</b>	<b>v</b>
<b>NHẬN XÉT VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN 2 .....</b>	<b>vi</b>
<b>DANH MỤC TỪ VIẾT TẮT.....</b>	<b>vii</b>
<b>DANH MỤC THUẬT NGỮ ANH – VIỆT .....</b>	<b>viii</b>
<b>DANH MỤC CÁC HÌNH ẢNH .....</b>	<b>xi</b>
<b>DANH MỤC CÁC BẢNG BIỂU .....</b>	<b>xiii</b>
<b>DÃN NHẬP .....</b>	<b>1</b>
❖ Mục tiêu của đồ án .....	1
❖ Phân công công việc.....	1
❖ Kế hoạch thực hiện đồ án .....	2
<b>CHƯƠNG 1. TỔNG QUAN.....</b>	<b>3</b>
1.1. Lý do hình thành đồ án .....	3
1.2. Mục tiêu đồ án.....	3
1.3. Dự kiến kết quả đạt được .....	3
<b>CHƯƠNG 2. CƠ SỞ LÝ THUYẾT .....</b>	<b>5</b>
2.1. Giới thiệu về khai phá dữ liệu.....	5
2.1.1. Khái niệm .....	5
2.1.2. Vai trò của khai phá dữ liệu trong kinh doanh.....	6
2.1.3. Quy trình khai phá dữ liệu .....	9
2.2. Kho dữ liệu .....	12
2.2.1. Kiến trúc luồng dữ liệu .....	12
2.2.2. Kho dữ liệu và khai phá dữ liệu trong BI.....	17
2.3. Các phương pháp trong khai phá dữ liệu.....	20
2.3.1. Phương pháp phân lớp. ....	20
2.3.2. Phương pháp gom cụm. ....	24
2.3.3. Phương pháp luật kết hợp. ....	32
2.4. Giới thiệu về phần mềm sử dụng (IBM SPSS Modeler) .....	38

2.4.1.	Tổng quan về phần mềm IBM SPSS Modeler.....	38
2.4.1.1	Ưu điểm. ....	39
2.4.1.2	Hạn chế.....	39
2.4.2.	Cách sử dụng phần mềm. ....	40
2.4.2.1.	Giới thiệu giao diện. ....	40
2.4.2.2.	Cách thức tiến hành các thuật toán.....	49
<b>CHƯƠNG 3.</b>	<b>ỨNG DỤNG PHẦN MỀM IBM SPSS MODELER.....</b>	<b>59</b>
3.1.	Thông tin về dữ liệu. ....	59
3.2.	Tiền xử lý dữ liệu. ....	61
3.3.	Kết luận của mỗi thuật toán.....	62
3.3.1.	Thuật toán phân lớp.....	62
3.3.2.	Thuật toán gom cụm.....	65
3.3.3.	Thuật toán kết hợp.....	75
<b>CHƯƠNG 4.</b>	<b>KẾT LUẬN.....</b>	<b>79</b>
4.1.	Những kết quả đạt được của đồ án. ....	79
4.2.	Nhược điểm của đồ án. ....	79
4.3.	Hướng phát triển cho đồ án. ....	79
<b>TÀI LIỆU THAM KHẢO .....</b>		<b>80</b>

## LỜI CẢM ƠN

Lời đầu tiên, chúng em xin phép gửi lời cảm ơn chân thành và tri ân sâu sắc đến quý thầy, cô là giảng viên của trường Đại học Tài chính – Marketing nói chung và khoa Công nghệ thông tin nói riêng đã giúp đỡ chúng em hoàn thành tốt bài báo cáo đồ án lần này. Đặc biệt, chúng em xin gửi lời cảm ơn chân thành đến cô Nguyễn Thị Trần Lộc – Giảng viên hướng dẫn môn Khai Phá Dữ Liệu. Trong quá trình chúng em học tập và tìm hiểu môn học này, cô đã quan tâm và hướng dẫn tận tình, cung cấp tài liệu tham khảo và truyền đạt cho chúng em những kiến thức vô cùng bổ ích. Bên cạnh việc giảng dạy trên lớp, cô đã luôn tạo điều kiện thuận lợi nhất, giải thích các thắc mắc, góp ý và sửa chữa những phần chúng em còn thiếu sót để mang lại một thành quả tốt nhất.

Tuy vậy, với kiến thức và kinh nghiệm còn hạn chế của những sinh viên nên đồ án vẫn còn nhiều những thiếu sót. Mong thầy/cô cảm thông và thấu hiểu. Chúng em luôn sẵn sàng nghe và nhận những góp ý của thầy cô để có thể hoàn thiện hơn nữa.

Lời cuối cùng, chúng em kính chúc cô luôn có thật nhiều sức khỏe và thành công trong sự nghiệp giảng dạy của mình. Chúng em xin chân thành cảm ơn!

**Sinh viên thực hiện**  
Nguyễn Tân Tài  
Nguyễn Minh Tín

## **NHẬN XÉT VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN 1**

Điểm số:.....

Điểm chữ: .....

Tp. Hồ Chí Minh, ngày... tháng... năm 202...

Giảng viên

## **NHẬN XÉT VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN 2**

Điểm số:.....

Điểm chữ: .....

Tp. Hồ Chí Minh, ngày... tháng... năm 202...

Giảng viên

## DANH MỤC TỪ VIẾT TẮT

Từ viết tắt	Điễn giải
CRM	Hệ quản trị quan hệ khách hàng
ETL	Extract, Transform, Load
OLAP	On-Line Analytical Processing

## **DANH MỤC THUẬT NGỮ ANH – VIỆT**

<b>Từ tiếng Anh</b>	<b>Dịch sang tiếng Việt</b>
Data mining	Khai phá dữ liệu
Data set	Bộ dữ liệu
Prediction Information	Thông tin dự đoán
Interpretation	Điễn giải
Predictive	Dự đoán
Statistics	Thống kê
Machine Learning	Máy học
Database	Cơ sở dữ liệu
Visualization	Trực quan hóa
Association	Kết hợp
Clustering	Gom cụm
Classifying	Phân lớp
Contingent Claim	Phân tích giá trị phụ thuộc
Fraud Detection	Phát hiện gian lận
Intrusion Detection	Phát hiện xâm nhập bất hợp pháp
Analysis of Stream data	Phân tích dòng dữ liệu
Training set	Tập huấn luyện

Test set	Tập kiểm tra
Data - Mart	Kho dữ liệu
Business Intelligence	Hệ thống thông tin quản trị thông minh
Dashboard	Báo cáo quản trị
Data warehouse	Kho dữ liệu
Acceptable	Có thể chấp nhận được
Class	Lớp
Supervised Learning	Học có giám sát
Decision Tree	Cây ra quyết định
K - nearest neighbor	K phần tử láng giềng gần nhất
Case - Based Reasoning	Suy diễn dựa trên tình huống
Genetic Algorithms	Phân lớp dựa trên tiến hóa gen
Rough sets	Lý thuyết tập thô
Fuzzy sets	Lý thuyết tập mờ
Predictive model	Mô hình dự báo
Series of rules	Dãy các luật
Category attribute	Thuộc tính phân lớp
Dependant variable	Biến phụ thuộc
Fegression tree	Cây hồi quy

Outliers	Phần tử bất thường
Noisy data	Giảm thiểu nhiễu
Similar	Tương tự
Dissimilar	Không tương tự
Data preprocessing	Hỗ trợ giai đoạn tiền xử lý dữ liệu
Data distribution	Mô tả sự phân bố dữ liệu/đối tượng
Pattern recognition	Nhận dạng mẫu
Spatial data analysis	Phân tích dữ liệu không gian
Image processing	Xử lý ảnh
Market segmentation	Phân mảnh thị trường
Document clustering	Gom cụm tài liệu

## DANH MỤC CÁC HÌNH ẢNH

<b>Hình 2. 1.</b> Các bước tổng quan về khai phá dữ liệu.....	5
<b>Hình 2. 2.</b> Các lĩnh vực liên quan đến khai phá dữ liệu.....	6
<b>Hình 2. 3.</b> Quy trình khai phá dữ liệu. ....	9
<b>Hình 2. 4.</b> Quá trình thu thập và hợp nhất dữ liệu. ....	10
<b>Hình 2. 5.</b> Mô hình phát triển và xác nhận. ....	12
<b>Hình 2. 6.</b> Kiến trúc luồng dữ liệu với Stage, ODS, DDS và MDB. ....	13
<b>Hình 2. 7.</b> Kiến trúc DDS đơn. ....	13
<b>Hình 2. 8.</b> Kiến trúc NDS + DDS. ....	15
<b>Hình 2. 9.</b> Kiến trúc ODS + DDS. ....	16
<b>Hình 2. 10.</b> Mô hình Online Analytical Processing. ....	18
<b>Hình 2. 11.</b> Mô hình Online Analytical Mining. ....	19
<b>Hình 2. 12.</b> Mô hình từ OLAP đến OLAM. ....	19
<b>Hình 2. 13.</b> Kiến trúc mức cao của hệ thống BI. ....	20
<b>Hình 2. 14.</b> Phân lớp dữ liệu dạng học có giám sát. ....	21
<b>Hình 2. 15.</b> Cấu trúc cây quyết định. ....	23
<b>Hình 2. 16.</b> Ví dụ của cây ra quyết định. ....	24
<b>Hình 2. 17.</b> Ví dụ về phương pháp gom cụm. ....	25
<b>Hình 2. 18.</b> Phân tích chi tiết gom cụm. ....	26
<b>Hình 2. 19.</b> Quá trình gom cụm dữ liệu. ....	26
<b>Hình 2. 20.</b> Hàm đo độ tương tự. ....	29
<b>Hình 2. 21.</b> Công thức tính khoảng cách giữa các đối tượng. ....	30
<b>Hình 2. 22.</b> Công thức cập nhật lại trọng tâm. ....	30
<b>Hình 2. 23.</b> Thuật toán K-Means. ....	30
<b>Hình 2. 24.</b> Gán điểm cho các cụm. ....	31
<b>Hình 2. 25.</b> Quá trình khai phá luật kết hợp. ....	32
<b>Hình 2. 26.</b> Công thức tính support của luật $X \rightarrow Y$ . ....	33
<b>Hình 2. 27.</b> Công thức để tính độ tin cậy của luật kết hợp $X \rightarrow Y$ . ....	34
<b>Hình 2. 28.</b> Maximal frequent itemsets. ....	36
<b>Hình 2. 29.</b> Gõ tìm IBM SPSS Modeler. ....	41
<b>Hình 2. 30.</b> Điền đầy đủ thông tin cần thiết. ....	42
<b>Hình 2. 31.</b> Nhập mã xác nhận từ Mail hoặc Telephone. ....	42
<b>Hình 2. 32.</b> Tải IBM SPSS Modeler về máy tính. ....	43
<b>Hình 2. 33.</b> Chạy cài đặt IBM SPSS Modeler. ....	43
<b>Hình 2. 34.</b> Giao diện đăng nhập vào IBM SPSS Modeler. ....	44
<b>Hình 2. 35.</b> Giao diện chính của IBM SPSS Modeler. ....	44
<b>Hình 2. 36.</b> Ribbon. ....	45
<b>Hình 2. 37.</b> Toolbar. ....	45
<b>Hình 2. 38.</b> Stream Window. ....	46
<b>Hình 2. 39.</b> Node Palette. ....	46
<b>Hình 2. 40.</b> Output Window. ....	47

<b>Hình 2. 41.</b> Models Viewer.....	47
<b>Hình 2. 42.</b> CRISP- DM.....	48
<b>Hình 2. 43.</b> Classes .....	49
<b>Hình 2. 44.</b> Tải bộ dữ liệu lên phần mềm. ....	50
<b>Hình 2. 45.</b> Kết nối bộ dữ liệu với Node Table. ....	50
<b>Hình 2. 46.</b> Xem lại dữ liệu trong Node Table. ....	51
<b>Hình 2. 47.</b> Chọn các biến dùng để phân tích. ....	51
<b>Hình 2. 48.</b> Khởi tạo cây ra quyết định. ....	52
<b>Hình 2. 49.</b> Cây ra quyết định. ....	52
<b>Hình 2. 50.</b> Tải dữ liệu lên phần mềm. ....	53
<b>Hình 2. 51.</b> Chọn các biến dùng để phân tích. ....	54
<b>Hình 2. 52.</b> Chọn số cụm muốn khởi tạo. ....	54
<b>Hình 2. 53.</b> Bảng thể hiện các kết quả của thuật toán K-Means.....	55
<b>Hình 2. 54.</b> Tải bộ dữ liệu lên phần mềm. ....	56
Hình 2. 55. Chọn các biến dùng để phân tích.....	56
<b>Hình 2. 56.</b> Xét độ hỗ trợ và độ tin cậy.....	57
<b>Hình 2. 57.</b> Bảng biểu diễn kết quả của thuật toán Apriori. ....	58

## DANH MỤC CÁC BẢNG BIỂU

Bảng 0-1 Phân công công việc .....	1
Bảng 0-2 Kế hoạch thực hiện đồ án .....	2
Bảng 3-1 Bảng thể hiện dữ liệu chi tiết.....	59

## DẪN NHẬP

Ngày trước đến nay, việc mua sắm hàng hóa đã trở thành một điều thiết yếu trong cuộc sống của mỗi con người. Ở các trung tâm mua sắm, siêu thị, cửa hàng tiện lợi... lượng hàng hóa được bán ra trong ngày, tháng, năm là vô cùng nhiều. Do đó, để tổng hợp các mặt hàng thiết yếu, việc thu thập dữ liệu về số lượng các mặt hàng thường xuyên được người dùng tiêu thụ, khai phá dữ liệu để tìm ra quy luật, hay nói đúng hơn là tính xác xuất các mặt hàng mà người tiêu dùng quan tâm hoặc thường xuyên mua cùng với nhau là hoàn toàn cần thiết. Từ đó, công ty xây dựng phần mềm đưa ra các ưu đãi cho các khách hàng thân thiết dựa trên các báo cáo về xu hướng mua hàng. Đó chính là mục đích cũng như lý do nhóm chúng em quyết định chọn đề tài này.

### ❖ Mục tiêu của đồ án

- Ứng dụng phần mềm IBM SPSS Modeler vào công việc khai phá dữ liệu.
- Biết cách làm việc nhóm.

### ❖ Phân công công việc

Bảng 0-1 Phân công công việc

STT	Họ tên SV	Công việc thực hiện
1	Nguyễn Minh Tín	Tìm kiếm thông tin bộ dữ liệu, xây dựng nội dung, thực hiện chạy chương trình rút ra kết luận, kiểm tra nội dung, chỉnh sửa đồ án, hoàn tất đồ án.
2	Nguyễn Tân Tài	Tìm kiếm thông tin phần mềm, xây dựng nội dung, thực hiện chạy chương trình rút ra kết luận, kiểm tra nội dung, chỉnh sửa nội dung đồ án, hoàn tất đồ án.

❖ Kế hoạch thực hiện đồ án

Bảng 0-2 Kế hoạch thực hiện đồ án

STT	Tên công việc thực hiện	SV thực hiện	Ghi chú
1	Chia nhóm, chọn đề tài.	Cả nhóm	Tốt
2	Thảo luận, phân công công việc.	Cả nhóm	Tốt
3	Đưa ra hướng xây dựng đề tài cụ thể.	Cả nhóm	Tốt
4	Tìm và giải nghĩa bộ dữ liệu	Nguyễn Minh Tín	Tốt
5	Tìm và hướng dẫn sử dụng phần mềm IBM SPSS Modeler	Nguyễn Tân Tài	Tốt
6	Chạy dữ liệu và đưa ra kết quả cuối cùng	Cả nhóm	Tốt
7	Chỉnh sửa file word, excel và hoàn thiện đồ án.	Cả nhóm	Tốt

# CHƯƠNG 1. TỔNG QUAN

## 1.1. Lý do hình thành đồ án

Ngày trước đến nay, việc mua sắm hàng hóa đã trở thành một điều thiết yếu trong cuộc sống của mỗi con người. Ở các trung tâm mua sắm, siêu thị, cửa hàng tiện lợi... lượng hàng hóa được bán ra trong ngày, tháng, năm là vô cùng nhiều. Do đó, để tổng hợp các mặt hàng thiết yếu, việc thu thập dữ liệu về số lượng các mặt hàng thường xuyên được người dùng tiêu thụ, khai phá dữ liệu để tìm ra quy luật, hay nói đúng hơn là tính xác xuất các mặt hàng mà người tiêu dùng quan tâm hoặc thường xuyên mua cùng với nhau là hoàn toàn cần thiết. Từ đó công ty xây dựng phần mềm đưa ra các ưu đãi cho các khách hàng thân thiết dựa trên các báo cáo về xu hướng mua hàng. Đó chính là mục đích cũng như lý do nhóm chúng em quyết định chọn đề tài này.

## 1.2. Mục tiêu đồ án

Đề tài tập chung vào nghiên cứu kỹ thuật Association (Kết hợp), Clustering (Phân cụm), Classifying (Phân loại) trong khai phá dữ liệu, từ đó nắm bắt được những giải thuật làm tiền đề cho nghiên cứu và sử dụng ứng dụng một cách cụ thể. Sau khi phân tích đặc điểm của dữ liệu thu nhập được và lựa chọn giải thuật phù hợp với dữ liệu, việc xây dựng và đánh giá chất lượng, độ hiệu quả của hệ thống cũng là mục tiêu chính của đề tài.

## 1.3. Dự kiến kết quả đạt được

- Phân tích đặc trưng khách hàng: Dự án có thể giúp phân tích các đặc trưng của khách hàng để hiểu hành vi mua hàng của họ. Những đặc trưng này có thể bao gồm độ tuổi, giới tính, thu nhập, phương thức thanh toán, v.v.
- Phân tích xu hướng mua hàng: Dự án có thể giúp phân tích xu hướng mua hàng của khách hàng để hiểu những sản phẩm nào được ưa chuộng và những sản phẩm nào ít được quan tâm. Những xu hướng này có thể bao gồm khi mua loại hàng này thì họ mua kèm với những mặt hàng nào v.v.
- Dự đoán hành vi mua hàng: Dự án có thể giúp dự đoán hành vi mua hàng của khách hàng trong tương lai. Những dự đoán này có thể giúp các doanh nghiệp đưa ra các chiến lược bán hàng hiệu quả hơn.

- Đề xuất sản phẩm: Dự án có thể giúp đề xuất những sản phẩm phù hợp với từng khách hàng dựa trên hành vi mua hàng của họ. Những đề xuất này có thể giúp các doanh nghiệp tăng doanh số và tăng khách hàng trung thành.

## CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

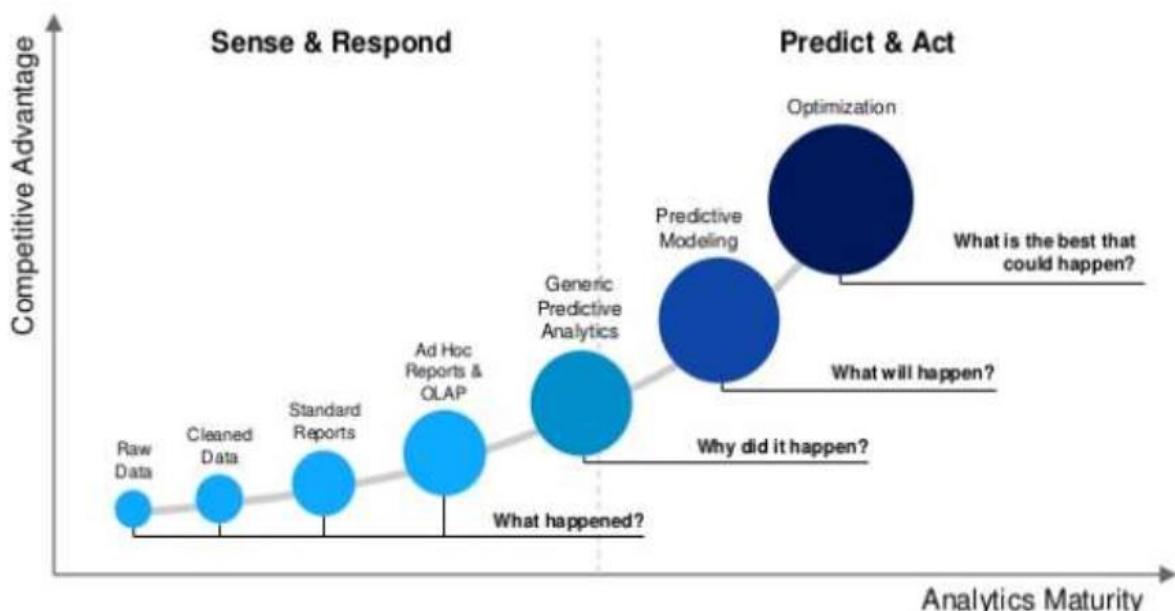
### 2.1. Giới thiệu về khai phá dữ liệu

#### 2.1.1. Khái niệm

**Khai phá dữ liệu** (datamining) là quá trình tìm kiếm các mẫu từ tập dữ liệu lớn (Data Set) và phân tích dữ liệu từ những quan điểm khác nhau. Việc này cho phép người dùng trong doanh nghiệp có thể phân tích dữ liệu từ nhiều góc độ khác nhau và tóm tắt các mối quan hệ xác định (relationship) để đưa ra các quyết định và giải quyết vấn đề.

Về bản chất, khai phá dữ liệu là quá trình tự động trích xuất thông tin có giá trị (Thông tin dự đoán – Prediction Information) ẩn chứa trong khối lượng dữ liệu khổng lồ trong thực tế.

Các hoạt động khai phá dữ liệu có thể được chia làm hai dòng nghiên cứu chính, theo mục đích chính của nghiên cứu: **diễn giải** (interpretation) và **dự đoán** (predictive)



Hình 2. 1. Các bước tổng quan về khai phá dữ liệu.

- **Diễn giải** (Interpretation)
  - Mục đích của việc diễn giải là xác định các mẫu thông thường trong dữ liệu và thể hiện chúng thông qua các quy tắc và tiêu chuẩn mà các chuyên gia trong lĩnh vực ứng dụng có thể dễ dàng hiểu được.
  - Các quy tắc tạo ra phải là nguyên gốc và không tầm thường để thực sự làm tăng mức độ hiểu biết và hiểu biết về hệ thống quan tâm.
- **Dự đoán** (Predictive): Dự đoán giá trị mà một biến ngẫu nhiên sẽ giả định trong tương lai hoặc để ước tính khả năng xảy ra các sự kiện trong tương lai.

Việc tiến hành khám phá dữ liệu có liên quan chặt chẽ đến các lĩnh vực sau:

- **Statistics** (Thống kê): Kiểm định mô hình và đánh giá tri thức phát hiện ra.
- **Machine Learning** (Máy học): Nghiên cứu xây dựng các giải thuật trên nền tảng của trí tuệ nhân tạo giúp cho máy tính có thể suy luận, dự đoán kết quả tương lai thông qua quá trình huấn luyện từ dữ liệu lịch sử.
- Việc tiến hành khám phá dữ liệu có liên quan chặt chẽ đến các lĩnh vực:
- **Database** (Cơ sở dữ liệu): Công nghệ quản trị dữ liệu, nhất là kho dữ liệu - data warehouse
- **Visualization** (Trực quan hóa): Giúp dữ liệu dễ hiểu, dễ sử dụng như chart, map
- Nhiệm vụ của khai phá dữ liệu:
- **Association** (Kết hợp): Tìm mồi quan hệ giữa các biến.
- **Clustering** (Phân cụm): Xác định mối quan hệ hợp lý trong các sản phẩm và nhóm chúng lại với nhau.
- **Classifying** (Phân loại): Liên quan đến việc áp dụng một mô hình được biết đến với các dữ liệu mới.

	<b>Statistics</b>	<b>Data Mining</b>	<b>Big Data</b>
<b>Structure</b>	structured	structured	unstructured
<b>Size</b>	small	large	very large
<b>Generation</b>	planned	transactional	behavioral
<b>Aim</b>	understand	optimize business	generate business
<b>Privacy Issues</b>	non	minor	huge
<b>Founded On</b>	concepts & theory	technology & tool	technology & tools
<b>Marketing</b>	bad	good	perfect

*Hình 2.2. Các lĩnh vực liên quan đến khai phá dữ liệu*

### 2.1.2. Vai trò của khai phá dữ liệu trong kinh doanh.

- ❖ *Phân tích và quản lý thị trường.*

- Có vai trò rất quan trọng trong ngành công nghiệp bán lẻ, do dữ liệu thu thập từ lĩnh vực này rất lớn từ doanh số bán hàng, lịch sử mua hàng của khách hàng, vận chuyển hàng hóa, tiêu thụ và dịch vụ.
- Khối lượng dữ liệu từ ngành công nghiệp này sẽ tiếp tục tăng lên nhanh chóng và dễ dàng thu thập bởi tính sẵn có trên môi trường Web.
- Ứng dụng khai phá dữ liệu nhằm xây dựng mô hình giúp xác định xu hướng mua hàng của khách hàng, giúp doanh nghiệp cải thiện chất lượng sản phẩm dịch vụ nhằm nâng cao sự hài lòng của khách hàng và giữ chân khách hàng tốt.
- Khai phá dữ liệu trên kho dữ liệu khách hàng
- Phân tích đa chiều trên kho dữ liệu khách hàng về doanh số bán hàng, khách hàng, sản phẩm, thời gian và khu vực.
- Phân tích hiệu quả của các chiến dịch bán hàng, Marketing.
- Quản trị mối quan hệ khách hàng (CRM).
- Giới thiệu và tư vấn sản phẩm phù hợp cho khách hàng.
- Khai thác hồ sơ khách hàng: phân tích xem liệu khách hàng A có phải là khách hàng tiềm năng sẽ mua sản phẩm B hay không?
- Xác định yêu cầu khách hàng: xem coi sản phẩm nào là sản phẩm phù hợp nhất cho từng đối tượng khách hàng khác nhau.
- Tiếp thị mục tiêu: phân tích xem khách hàng nào có sở thích, thu nhập, hoặc độ tuổi gần giống nhau sẽ được gom thành một nhóm. Với mỗi nhóm khách hàng khác nhau, các doanh nghiệp sẽ dễ dàng hơn trong việc tiếp thị và cung cấp sản phẩm phù hợp.

❖ ***Phân tích doanh nghiệp và quản lý rủi ro.***

Lập kế hoạch tài chính và đánh giá tài sản: bao gồm phân tích và dự đoán dòng chảy của đồng tiền, phân tích giá trị phụ thuộc (contingent claim) để thẩm định tài sản.

❖ ***Phát hiện gian lận – fraud detection.***

- Phát hiện gian lận thường được sử dụng trong các lĩnh vực như ngân hàng, với dịch vụ thẻ tín dụng, hoặc là viễn thông. Đối với viễn thông, khai phá dữ liệu từ các cuộc gọi lừa đảo sẽ giúp các cơ quan điều tra xác định thời

gian gọi, gọi trong bao lâu, gọi đến ai, ai là người gọi, từ đó sẽ xác định được thủ phạm và truy bắt chúng.

- Trong lĩnh vực tài chính, bán hàng, nó còn được dùng để phân tích các mô hình, xem coi có trường hợp nào đi chệch khỏi quỹ đạo hay không, doanh thu bán hàng có khác với chỉ tiêu dự kiến hay không.

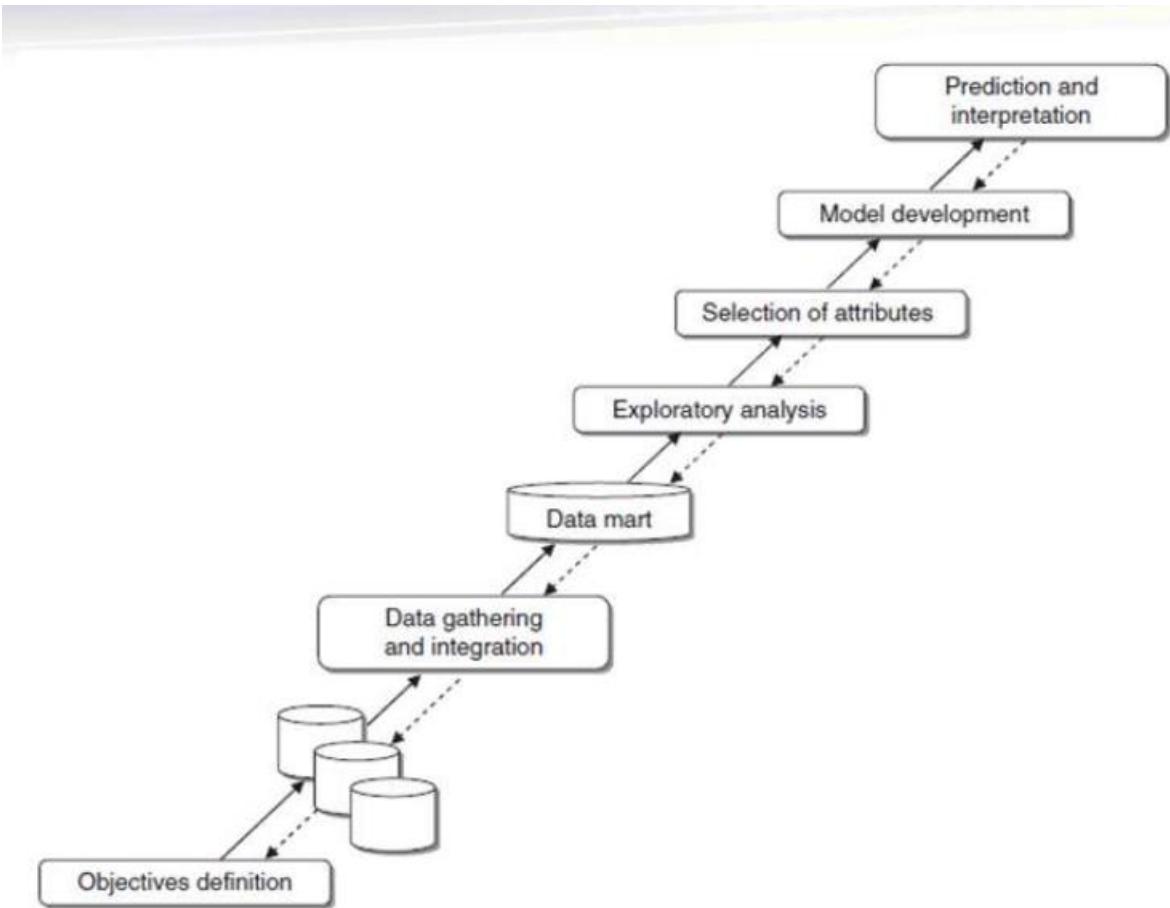
❖ **Phát hiện xâm nhập bất hợp pháp (intrusion detection).**

- Xâm nhập bất hợp pháp là những hành động đe dọa tính toàn vẹn, bảo mật và tính sẵn sàng của tài nguyên mạng.
- Trong thế giới của kết nối, bảo mật đã trở thành vấn đề lớn đối với tồn tại của hệ thống. Với sự phát triển của internet và sự sẵn có của các công cụ, thủ thuật trợ giúp cho xâm nhập và tấn công mạng, yêu cầu kiểm soát truy cập bất hợp pháp là yếu tố rất quan trọng đảm bảo cho sự ổn định của hệ thống.
- Phát triển các thuật toán khai phá dữ liệu để phát hiện xâm nhập.
- Phân tích kết hợp, tương quan và khác biệt để phát hiện xâm nhập.
- Phân tích dòng dữ liệu (Analysis of Stream data) để phát hiện bất thường.

❖ **Ngoài ra, khai phá dữ liệu còn được ứng dụng rộng rãi trong các lĩnh vực khác nhau:**

- Ngân hàng
- Y tế
- Viễn thông
- Bảo hiểm
- Địa chất học
- Thể thao
- Tin học
- Sinh học

### 2.1.3. Quy trình khai thác dữ liệu



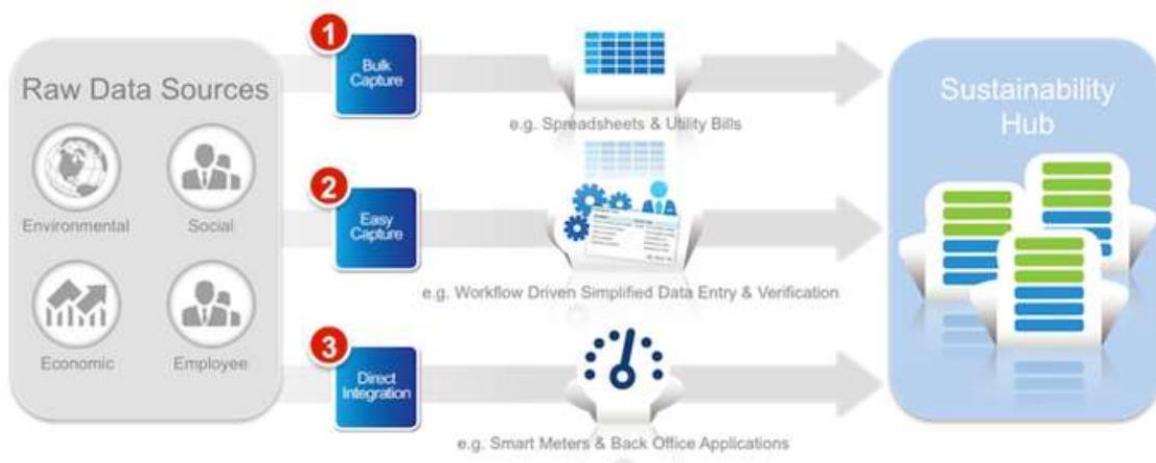
*Hình 2. 3. Quy trình khai thác dữ liệu.*

#### ❖ Định nghĩa các mục tiêu.

- Các phân tích khai thác dữ liệu được thực hiện trong các lĩnh vực ứng dụng cụ thể và nhằm cung cấp cho các nhà ra quyết định những kiến thức hữu ích.
- Các chuyên gia đòi hỏi phải có trực giác và năng lực để xây dựng các mục tiêu điều tra có thể xác định được và xác định rõ ràng.
- Nếu vấn đề đang bàn cãi không được xác định và xác định một cách đầy đủ, người ta có thể có nguy cơ cần trở bất kỳ nỗ lực trong tương lai nào trong các hoạt động khai thác dữ liệu.
- Việc xác định các mục tiêu sẽ được lợi từ sự hợp tác chặt chẽ giữa các chuyên gia trong lĩnh vực ứng dụng và các nhà phân tích khai thác dữ liệu.

#### ❖ Thu thập và hợp nhất dữ liệu.

- Khi các mục tiêu của cuộc điều tra đã được xác định, bắt đầu thu thập dữ liệu. Dữ liệu có thể đến từ các nguồn khác nhau và do đó có thể yêu cầu hợp nhất.
- Nguồn dữ liệu có thể là nội bộ, bên ngoài hoặc kết hợp cả hai. Việc tích hợp các nguồn dữ liệu khác nhau có thể được đề xuất bởi nhu cầu làm phong phú thêm dữ liệu với các tham số mô tả mới, chẳng hạn như các biến về tiếp thị địa lý hoặc với các danh sách tên khách hàng tiềm năng, khách hàng tiềm năng hiện chưa có trong hệ thống thông tin của công ty.
- Trong một số trường hợp, các nguồn dữ liệu đã được cấu trúc trong các kho dữ liệu và các trung tâm dữ liệu cho các phân tích của OLAP và nói chung là cho các hoạt động hỗ trợ ra quyết định.



**Hình 2. 4. Quá trình thu thập và hợp nhất dữ liệu.**

#### ❖ **Phân tích nghiên cứu.**

- Trong giai đoạn thứ ba của quá trình khai thác dữ liệu, một phân tích sơ bộ về dữ liệu được thực hiện với mục đích làm quen với các thông tin hiện có và thực hiện việc làm sạch dữ liệu.
- Bước này sẽ loại bỏ nhiều và dữ liệu không nhất quán.
- Thông thường, dữ liệu được lưu trữ trong kho dữ liệu được xử lý ở thời gian tải theo cách để loại bỏ bất kỳ sự không nhất quán về cú pháp.

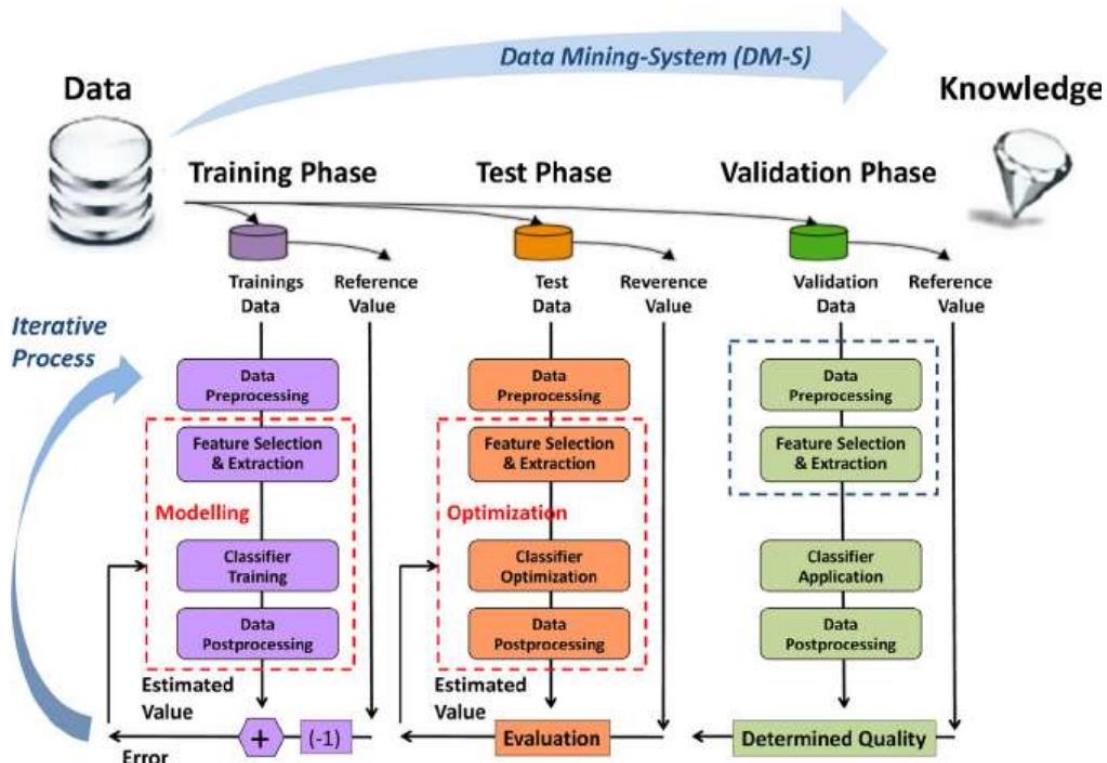
#### ❖ **Lưu chọn thuộc tính.**

- Trong giai đoạn tiếp theo, sự liên quan của các thuộc tính khác nhau được đánh giá liên quan đến các mục tiêu của phân tích.

- Các thuộc tính chứng tỏ ít được sử dụng sẽ bị xóa, để làm sạch các thông tin không liên quan từ bộ dữ liệu.
- Các thuộc tính mới thu được từ các biến ban đầu thông qua các phép biến đổi thích hợp được đưa vào bộ dữ liệu.

❖ **Mô hình phát triển và xác nhận.**

- Một khi bộ dữ liệu chất lượng cao đã được lắp ráp và có thể được làm phong phú với các thuộc tính mới được xác định, có thể phát triển các mô hình nhận diện và dự báo.
- Thông thường việc đào tạo các mô hình được thực hiện bằng cách sử dụng một mẫu các hồ sơ trích ra từ bộ dữ liệu ban đầu.
- Độ chính xác dự đoán của từng mô hình được tạo ra có thể được đánh giá bằng cách sử dụng phần còn lại của dữ liệu.
- Tập dữ liệu hiện có được chia thành hai tập con. Đầu tiên tạo thành tập huấn luyện (training set) và được sử dụng để xác định một mô hình học cụ thể trong mô hình các mô hình đã chọn. Thông thường cỡ mẫu của tập huấn luyện được chọn là tương đối nhỏ, mặc dù có ý nghĩa thống kê từ quan điểm thống kê, vài ngàn quan sát.
- Tập con thứ hai là tập kiểm tra (test set) và được sử dụng để đánh giá độ chính xác của các mô hình thay thế được tạo ra trong giai đoạn đào tạo để xác định mô hình tốt nhất cho dự đoán trong tương lai.



**Hình 2. 5. Mô hình phát triển và xác nhận.**

❖ **Dự đoán và diễn giải.**

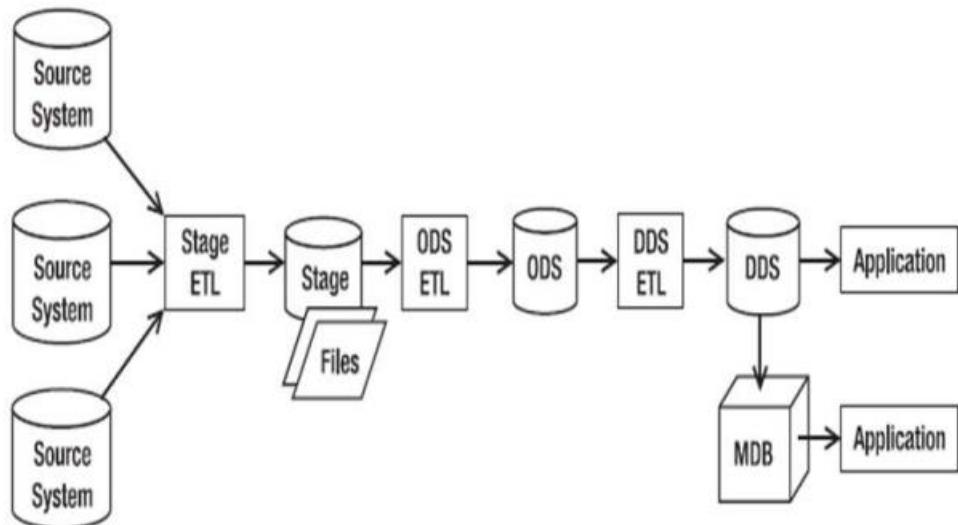
- Sau khi kết thúc quá trình khai thác dữ liệu, mô hình được lựa chọn giữa những người tạo ra trong giai đoạn phát triển nên được thực hiện và sử dụng để đạt được các mục tiêu ban đầu được xác định.
- Hơn nữa, cần kết hợp chặt chẽ vào các thủ tục hỗ trợ các quá trình ra quyết định để các nhân viên có thể sử dụng nó rút ra những dự đoán và thu thập kiến thức sâu hơn về hiện tượng quan tâm.

## 2.2. Kho dữ liệu

### 2.2.1. Kiến trúc luồng dữ liệu

❖ **Kho dữ liệu có rất nhiều loại kiến trúc.**

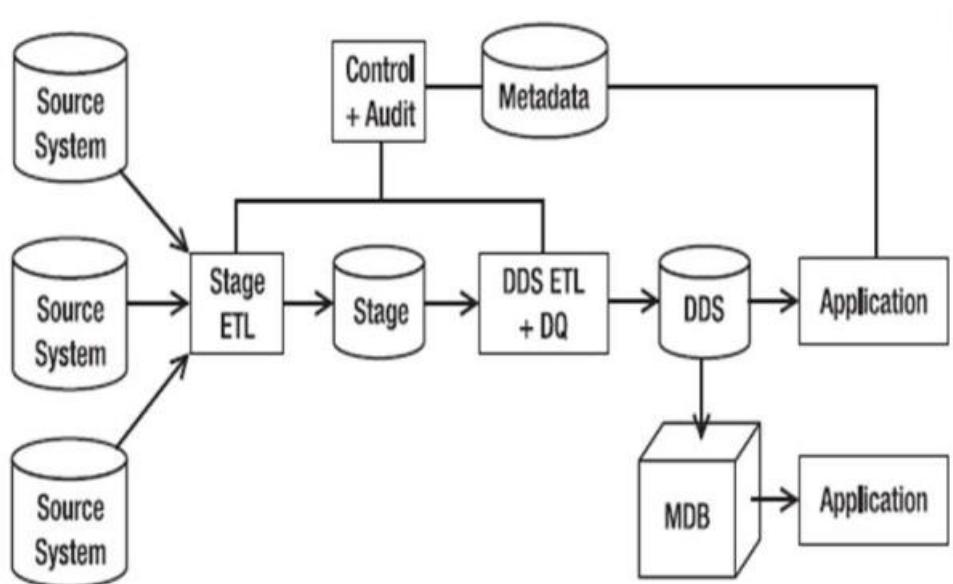
- Đơn giản nhất: chỉ gồm một kho dữ liệu đầu cuối.
- Rất phức tạp: bao gồm nhiều kho dữ liệu trung gian, được sử dụng trong những hệ thống lớn.



*Hình 2. 6. Kiến trúc luồng dữ liệu với Stage, ODS, DDS và MDB.*

- ❖ *Tuy nhiên, hầu hết các kiến trúc đều dựa trên 3 kiến trúc chung phổ biến sau:*

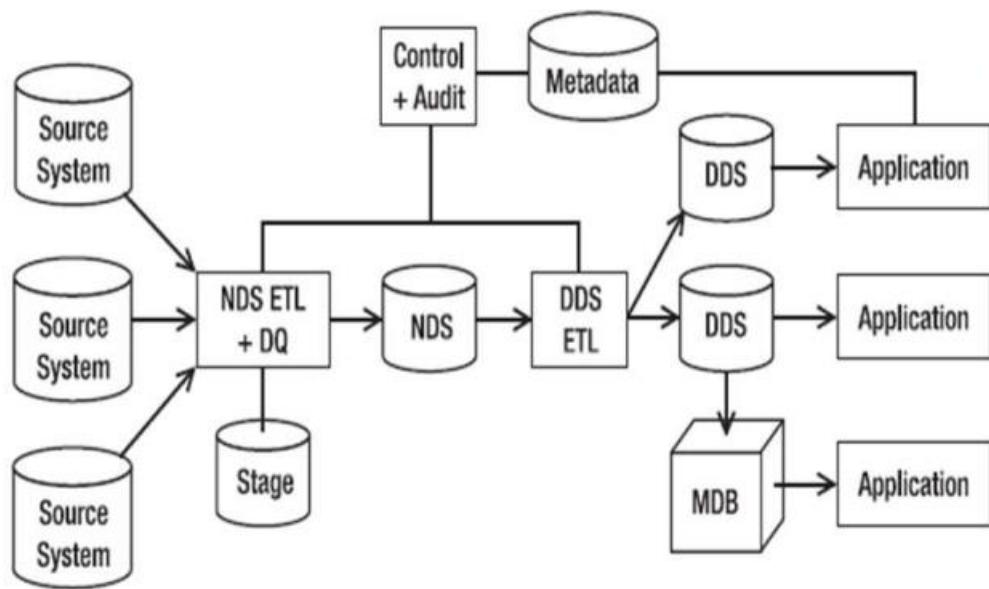
- *Kiến trúc DDS đơn (single DDS)*



*Hình 2. 7. Kiến trúc DDS đơn.*

- Kiến trúc DDS đơn là một trong những dạng kiến trúc đơn giản nhất của kho dữ liệu. Kiến trúc này có thành phần chính là một kho dữ liệu trung tâm.
- Dữ liệu từ nhiều hệ thống nguồn được nạp vào vùng xử lý thông qua một gói ETL.

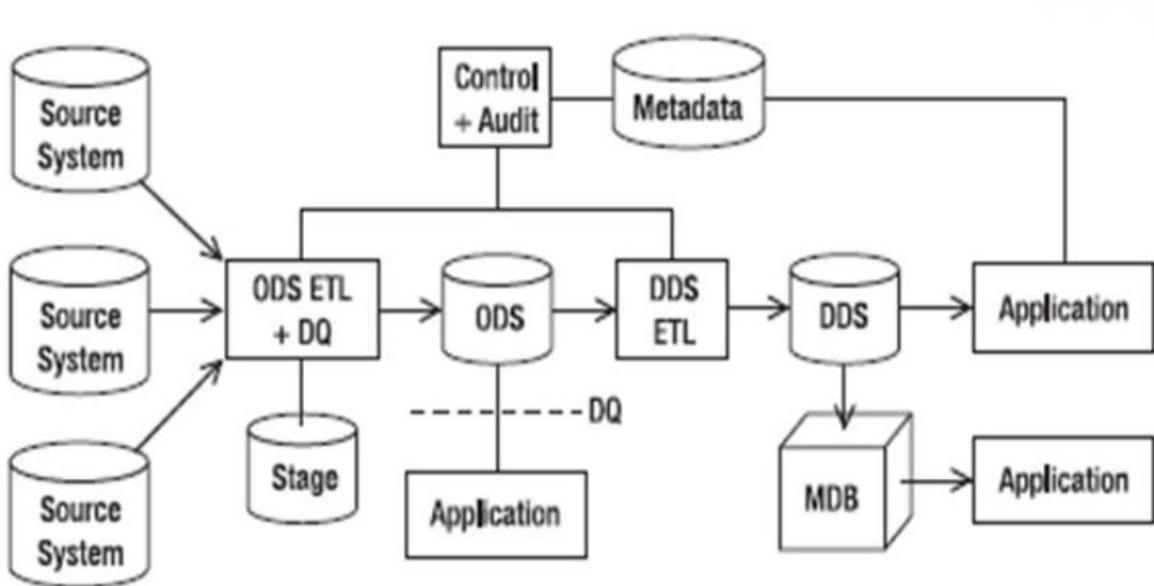
- Gói ETL này sẽ rút trích dữ liệu từ nhiều nguồn khác nhau, thực hiện một số phép biến đổi dữ liệu đơn giản. Dữ liệu sau đó được chứa trong vùng xử lý.
- Dữ liệu trong vùng xử lý sau khi được xử lý sơ bộ sẽ được biến đổi thông qua một gói ETL khác để đưa vào kho dữ liệu đầu cuối. Quá trình biến đổi này bao gồm nhiều công đoạn:
  - Việc làm sạch.
  - Chuẩn hóa dữ liệu.
  - Quản lý chất lượng và lịch sử thay đổi của dữ liệu.
- Kho dữ liệu đầu cuối chứa những dữ liệu đã được biến đổi, chuẩn hóa, và lưu trữ dưới dạng mô hình đa chiều, sẵn sàng phục vụ cho các ứng dụng đầu cuối.
- *Ưu điểm:*
  - Kiến trúc đơn giản.
  - Ít công đoạn xử lý.
  - Thuận lợi khi xây dựng những kho dữ liệu nhỏ.
- *Nhược điểm:*
  - Không hỗ trợ việc tạo ra nhiều kho dữ liệu phục vụ cho nhiều mục đích khác nhau dựa trên dữ liệu sẵn có. Nếu có nhu cầu chỉ cần sử dụng một phần của kho dữ liệu (data - mart) thì phải xây dựng một gói ETL khác phục vụ quá trình này.
  - Không tái sử dụng được gói ETL đã làm. Mỗi một quy trình rút trích - biến đổi - nạp cho từng thành phần trong kho dữ liệu đầu cuối được thực hiện độc lập. Việc này gây khó khăn cho việc xây dựng những kho dữ liệu lớn.
    - *Kiến trúc NDS+DDS.*



**Hình 2. 8. Kiến trúc NDS + DDS.**

- Đây là một kiến trúc khá phổ biến. Kiến trúc này tương tự như kiến trúc DDS đơn, nhưng có thêm một vùng chứa dữ liệu trung gian là vùng chứa dữ liệu chuẩn hoá NDS.
- Dữ liệu sau khi được làm sạch, thay vì đưa thẳng vào kho dữ liệu đầu cuối, nó được lưu trong vùng chứa dữ liệu trung gian.
- Vùng chứa dữ liệu trung gian đóng vai trò như là một cơ sở dữ liệu tập trung, đã được chuẩn hoá, bao gồm cả dữ liệu lịch sử.
- Việc nạp vào kho dữ liệu đầu cuối sẽ không cần qua công đoạn làm sạch và quản lý chất lượng dữ liệu nữa.
- *Ưu điểm:*
  - Lưu trữ dữ liệu tập trung đã được làm sạch.
  - Chứa dữ liệu lịch sử.
  - Sẵn sàng cho việc nạp vào nhiều kho dữ liệu đầu cuối.
  - Tái sử dụng được các gói ETL.
- *Nhược điểm:*
  - Kiến trúc phức tạp.
  - Tốn thêm không gian lưu trữ.
  - Thời gian thực hiện một chu kỳ nạp dữ liệu lâu hơn so với kiến trúc DDS đơn.
  - Vùng chứa dữ liệu trung gian không được tận dụng vào mục đích khác.

- *Kiến trúc ODS+DDS.*



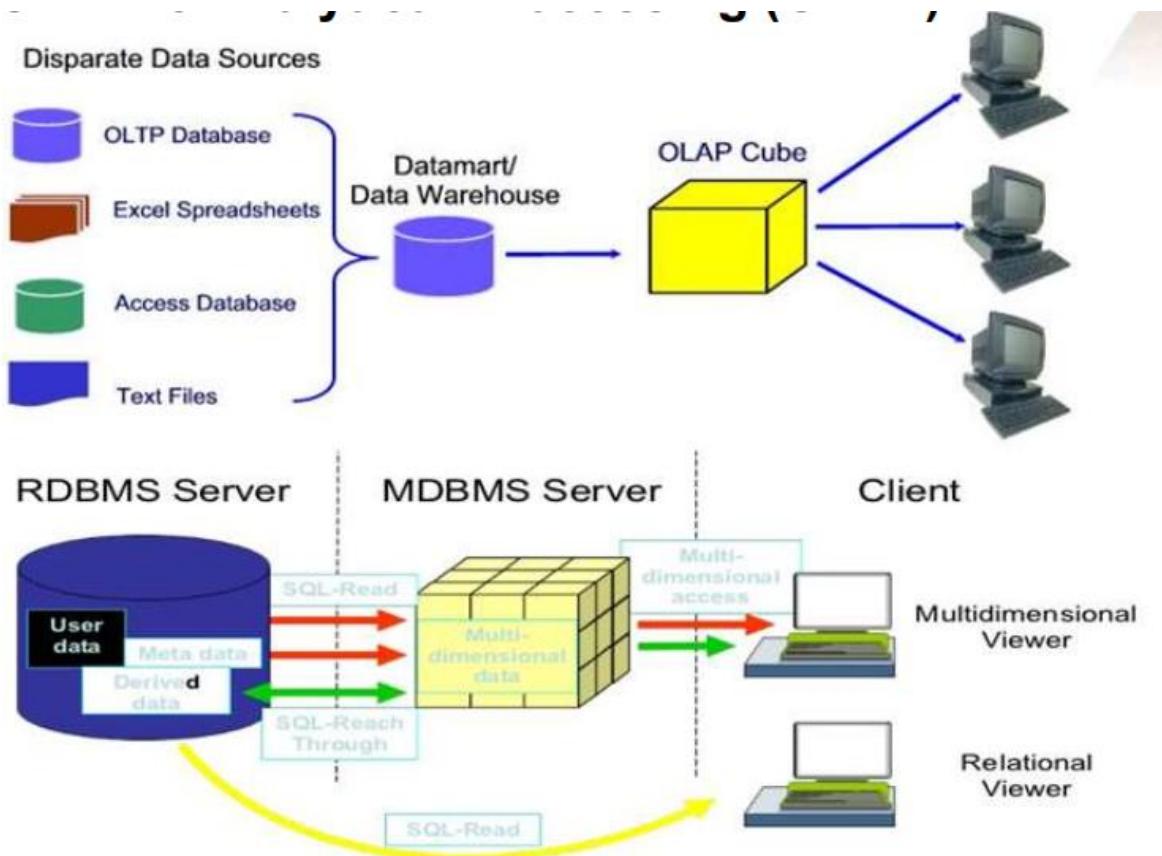
**Hình 2. 9. Kiến trúc ODS + DDS.**

- Kiến trúc này có nhiều điểm tương đồng với kiến trúc NDS+DDS. Như trong hình vẽ, thay vì sử dụng một vùng dữ liệu chuẩn hoá làm vùng dữ liệu trung gian, người ta sử dụng một vùng dữ liệu hoạt động thay cho nó.
- Vùng dữ liệu hoạt động này cũng là một cơ sở dữ liệu dạng chuẩn hoá cao.
- Tuy nhiên, nó không lưu dữ liệu lịch sử. Vùng dữ liệu hoạt động có cấu trúc nghiêng về dạng cơ sở dữ liệu phục vụ giao tác (OLTP) nhiều hơn. Nó đóng vai trò như là một cơ sở dữ liệu tập trung mà ở đó, ứng dụng đầu cuối cho phép khai thác trên nó.
- **Ưu điểm:**
  - Lưu trữ dữ liệu tập trung đã được làm sạch.
  - Tận dụng làm cơ sở dữ liệu tập trung phục vụ giao tác cho ứng dụng đầu cuối.
- **Nhược điểm:**
  - Không chứa dữ liệu lịch sử.
  - Các gói ETL để đưa dữ liệu từ vùng dữ liệu hoạt động vào kho dữ liệu đầu cuối phức tạp hơn.
  - Vùng dữ liệu hoạt động có thể bị gián đoạn khi nạp kho dữ liệu.
  - Không tái sử dụng được các gói ETL.

### 2.2.2. Kho dữ liệu và khai phá dữ liệu trong BI.

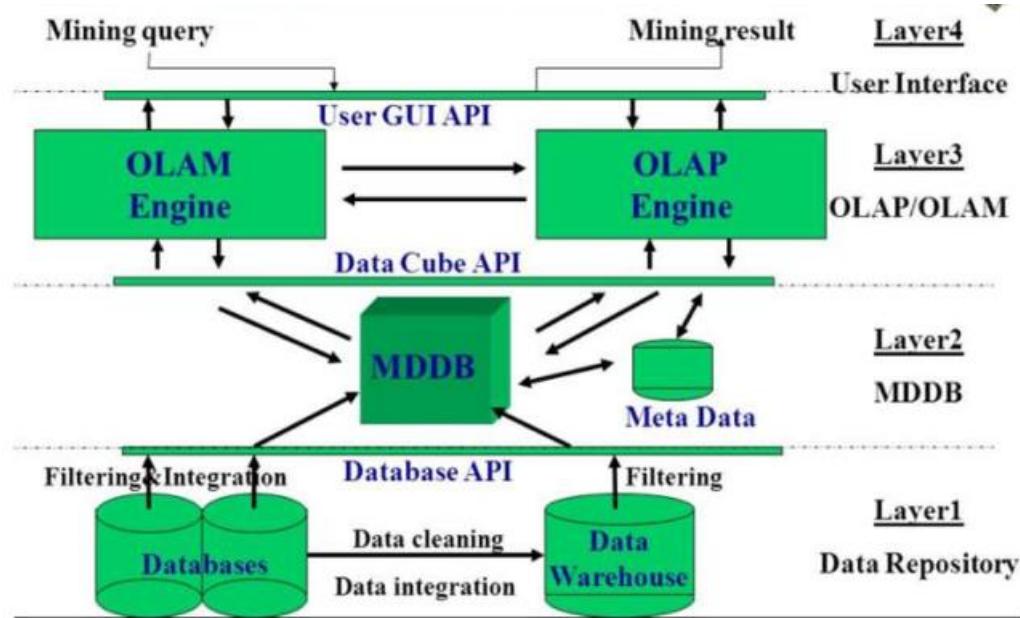
❖ Có ba loại ứng dụng kho dữ liệu: xử lý thông tin, xử lý phân tích và khai thác dữ liệu.

- **Xử lý thông tin** hỗ trợ truy vấn, phân tích thông kê cơ bản và báo cáo sử dụng crosstab, bảng, biểu đồ hoặc đồ thị. Xu hướng hiện tại trong xử lý thông tin kho dữ liệu là xây dựng các công cụ truy cập dựa trên web có chi phí thấp sau đó được tích hợp với các trình duyệt Web.
- **Xử lý phân tích** hỗ trợ các hoạt động OLAP cơ bản, bao gồm slice-and-dice, drill - down, roll - up, và pivoting. Nó thường hoạt động trên dữ liệu lịch sử trong cả hai dạng tóm tắt và chi tiết. Sức mạnh chính của xử lý phân tích trực tuyến đối với quá trình xử lý thông tin là phân tích số liệu dữ liệu kho dữ liệu theo chiều sâu.
- **Khai phá dữ liệu** hỗ trợ khám phá kiến thức bằng cách tìm kiếm các mẫu ẩn và các hiệp hội, xây dựng các mô hình phân tích, thực hiện phân loại và dự đoán, và trình bày các kết quả khai thác bằng các công cụ trực quan hóa.
- **On-Line Analytical Processing (OLAP)**: là một công nghệ được sử dụng để sắp xếp các cơ sở dữ liệu doanh nghiệp lớn và hỗ trợ nghiệp vụ thông minh. Cơ sở dữ liệu OLAP được chia thành một hoặc nhiều cube, đồng thời, mỗi cube được người quản trị cube sắp xếp và thiết kế sao cho phù hợp với cách bạn truy xuất và phân tích dữ liệu để tạo và sử dụng các báo cáo và báo cáo PivotChart, PivotTable mà bạn cần dễ dàng hơn.

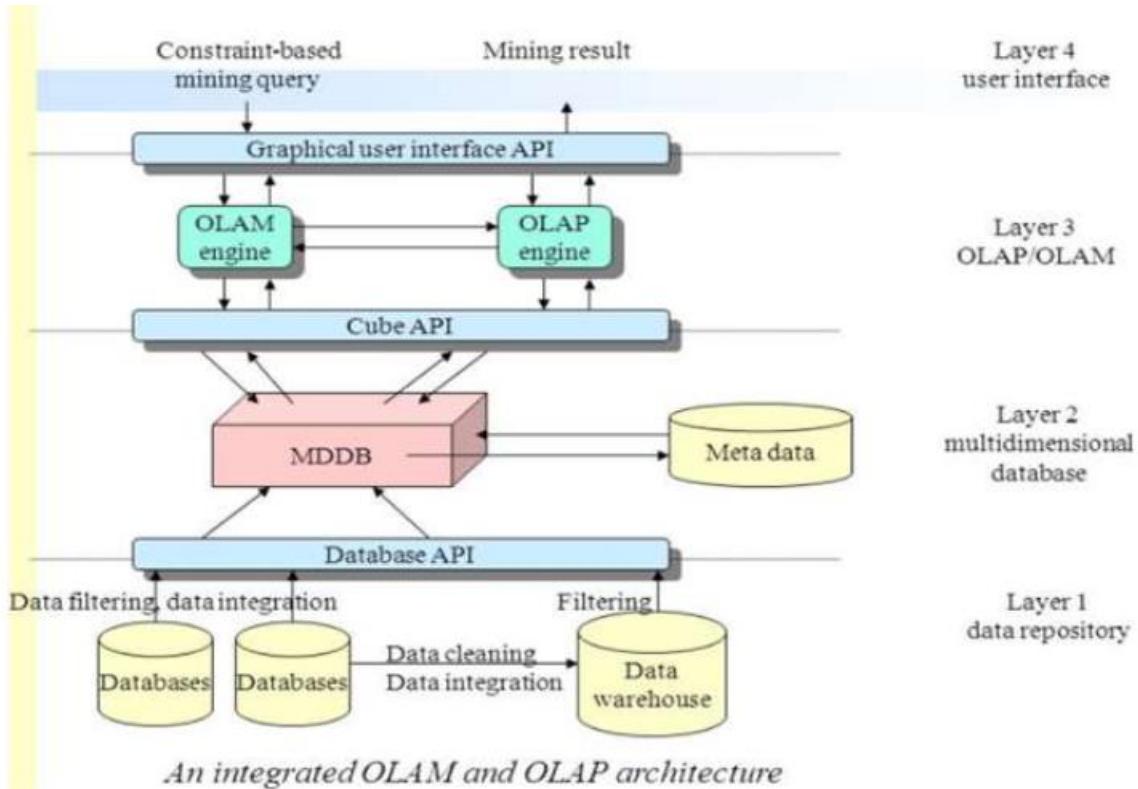


*Hình 2. 10. Mô hình Online Analytical Processing.*

- **On-Line Analytical Mining (OLAM):** hay còn gọi là OLAP mining: tích hợp xử lý phân tích trực tuyến (OLAP) với khai thác dữ liệu và kiến thức về khai phá dữ liệu trong cơ sở dữ liệu đa chiều.



Hình 2. 11. Mô hình Online Analytical Mining.

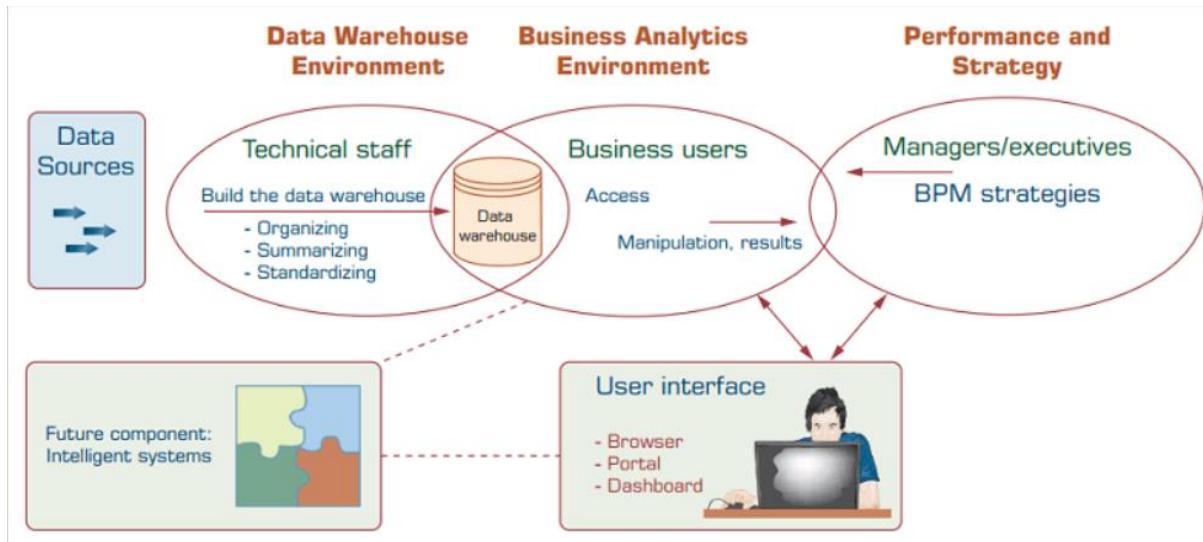


Hình 2. 12. Mô hình từ OLAP đến OLAM.

#### ❖ Khai phá dữ liệu trong BI.

**Hệ thống thông tin quản trị thông minh (Business Intelligence – BI)** là một hệ thống giúp các nhà quản lý công cụ và một phương pháp mới điều hành

doanh nghiệp như đã trình bày trong bài trước. Để có thể trình bày được thông tin trên các **báo cáo quản trị (dashboard)** thì cần có nguồn cung cấp thông tin đó – đó chính là **kho dữ liệu (Data warehouse)**. Vị trí của **kho dữ liệu** được minh họa ở Figure 1. Phía bên phải (hình oval bên phải) là đối tượng thụ hưởng của hệ thống – những người sẽ phân tích thông tin để đưa ra các kế hoạch dài hạn hay điều hành ngắn hạn.



**Hình 2. 13. Kiến trúc mức cao của hệ thống BI.**

Để có thể đưa ra được các thông tin có tính hệ thống, phù hợp với nghiệp vụ kinh doanh của doanh nghiệp thì cần có đội ngũ nghiệp vụ (hình oval ở giữa), chịu trách nhiệm xây dựng các báo cáo quản trị từ **kho dữ liệu**. Cuối cùng để có thể lấy được dữ liệu và đưa vào **kho dữ liệu** theo nhu cầu nghiệp vụ thì cần có đội ngũ kỹ thuật (hình oval bên trái).

Ngoài ra có thể có các hệ thống thông minh (hình vuông góc dưới bên trái) có thể khai thác dữ liệu từ **kho dữ liệu** nhằm hỗ trợ quản lý ra quyết định.

### 2.3. Các phương pháp trong khai phá dữ liệu

#### 2.3.1. Phương pháp phân lớp.

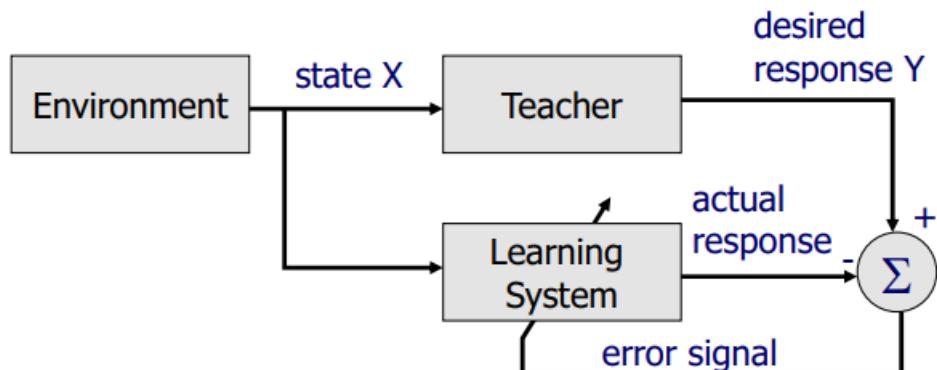
##### ❖ *Phân lớp dữ liệu (Classification)*

Dạng phân tích dữ liệu nhằm rút trích các mô hình mô tả các lớp dữ liệu hoặc dự đoán xu hướng dữ liệu

Quá trình gồm hai bước:

- Bước học (giai đoạn huấn luyện): xây dựng bộ Phân lớp (classifier) bằng việc phân tích/học tập huấn luyện.
- Bước Phân lớp (classification): Phân lớp dữ liệu/đòi tương mới nếu độ chính xác của bộ Phân lớp được đánh giá là có thể chấp nhận được (acceptable).  
 $y = f(X)$  với  $y$  là nhãn (phản mô tả) của một lớp (class) và  $X$  là dữ liệu/đòi tương.
- Bước học:  $X$  trong tập huấn luyện, một trị  $y$  được cho trước với  $X \rightarrow$  xác định
- Bước Phân lớp: đánh giá  $f$  với  $(X', y)$  và  $X' \Leftrightarrow$  Mọi  $X$  trong tập huấn luyện; nếu acceptable thì dùng  $f$  để xác định  $y'$  cho  $X'$  (mới)

Dạng học có giám sát (supervised learning)



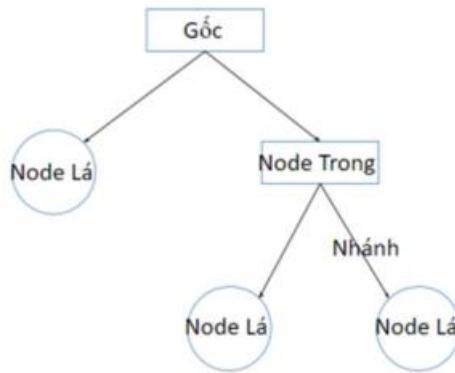
Hình 2. 14. Phân lớp dữ liệu dạng học có giám sát.

❖ Các giải thuật phân lớp dữ liệu.

- Phân lớp với cây ra quyết định (decision tree)
- Phân lớp với mạng Bayesian
- Phân lớp với mạng Neural
- Phân lớp với k phần tử láng giềng gần nhất (k-nearest neighbor)
- Phân lớp với suy diễn dựa trên tình huống (case-based reasoning)
- Phân lớp dựa trên tiến hóa gen (genetic algorithms)
- Phân lớp với lý thuyết tập thô (rough sets)
- Phân lớp với lý thuyết tập mờ (fuzzy sets) ...

Trong đồ án này ta sẽ chỉ tập trung phân tích Phân lớp với cây ra quyết định (decision tree)

- ❖ **Cây quyết định** là một kiểu mô hình dự báo (predictive model), nghĩa là một ánh xạ từ các quan sát về một sự vật/hiện tượng tới các kết luận về giá trị mục tiêu của sự vật hiện tượng.
  - Mỗi một nút trong (internal node) tương ứng với một biến
  - Đường nối giữa nó với nút con của nó thể hiện một giá trị cụ thể cho biến đó.
  - Mỗi nút lá đại diện cho giá trị dự đoán của biến mục tiêu, cho trước các giá trị của các biến được biểu diễn bởi đường đi từ nút gốc tới nút lá đó.
- ❖ Một cây quyết định có thể được học bằng cách chia tập hợp nguồn thành các tập con dựa theo một kiểm tra giá trị thuộc tính
- ❖ Quá trình này được lặp lại một cách đệ quy cho mỗi tập con dẫn xuất.
- ❖ Quá trình đệ quy hoàn thành khi không thể tiếp tục thực hiện việc chia tách được nữa, hay khi một phân loại đơn có thể áp dụng cho từng phần tử của tập con dẫn xuất. Một bộ phân loại rừng ngẫu nhiên (random forest) sử dụng một số cây quyết định để có thể cải thiện tỉ lệ phân loại.
- ❖ Cây quyết định (Decision Tree) là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật (series of rules).
- ❖ Các thuộc tính của đối tượng (ngoại trừ thuộc tính phân lớp - Category attribute) có thể thuộc các kiểu dữ liệu khác nhau (Binary, Nominal, ordinal, quantitative values) trong khi đó thuộc tính phân lớp phải có kiểu dữ liệu là Binary hoặc Ordinal.
- ❖ Tóm lại, cho dữ liệu về các đối tượng gồm các thuộc tính cùng với lớp (classes) của nó, cây quyết định sẽ sinh ra các luật để dự đoán lớp của các đối tượng chưa biết (unseen data)
- ❖ **Đặc điểm của cây quyết định:** là một cây có cấu trúc, trong đó:
  - Root (Gốc): Là nút trên cùng của cây.
  - Node nội (trong): nút trung gian trên một thuộc tính đơn (hình Oval).
  - Nhánh: Biểu diễn các kết quả của kiểm tra trên nút.
  - Node lá: Biểu diễn lớp hay sự phân phối lớp (hình vuông hoặc chữ nhật)



**Hình 2. 15.** Cấu trúc cây quyết định.

❖ **Đặc điểm của cây quyết định.**

- Cây quyết định cũng là một phương tiện có tính mô tả dành cho việc tính toán các xác suất có điều kiện.
- Cây quyết định có thể được mô tả như là sự kết hợp của các kỹ thuật toán học và tính toán nhằm hỗ trợ việc mô tả, phân loại và tổng quát hóa một tập dữ liệu cho trước.
- Dữ liệu được cho dưới dạng các bản ghi có dạng:

$$(x, y) = (x_1, x_2, \dots, x_k, y)$$

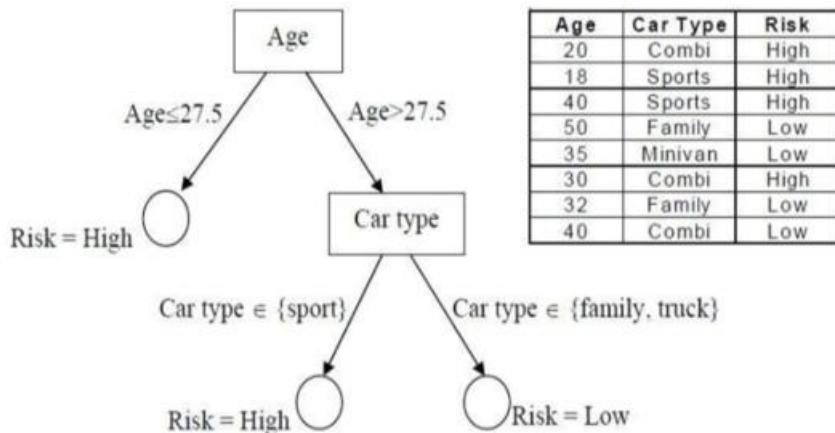
- Biến phụ thuộc (dependant variable) y là biến mà chúng ta cần tìm hiểu, phân loại hay tổng quát hóa.  $x_1, x_2, x_3 \dots$  là các biến sẽ giúp ta thực hiện công việc đó.

❖ **Phân loại cây quyết định.**

- Cây hồi quy (Fegression tree) ước lượng các hàm giá có giá trị là số thực thay vì được sử dụng cho các nhiệm vụ phân loại. (ví dụ: ước tính giá một ngôi nhà hoặc khoảng thời gian một bệnh nhân nằm viện)
- Cây phân loại (Classification tree), nếu y là một biến phân loại như: giới tính (nam hay nữ), kết quả của một trận đấu (thắng hay thua).

❖ **Cây quyết định (decision tree) = mô hình phân lớp.**

- Node nội: phép kiểm thử (test) trên một thuộc tính.
- Node lá: nhãn/mô tả của một lớp (class label)
- Nhánh từ một node nội: kết quả của một phép thử trên thuộc tính tương ứng.



**Hình 2.16.** Ví dụ của cây ra quyết định.

- ❖ **Phát triển cây quyết định:** đi từ gốc, đến các nhánh, phát triển quy nạp theo hình thức chia để trị.

**Bước 1:** Chọn thuộc tính “tố” nhất bằng một độ đo đã định trước

**Bước 2:** Phát triển cây bằng việc thêm các nhánh tương ứng với từng giá trị của thuộc tính đã chọn

**Bước 3:** Sắp xếp, phân chia tập dữ liệu đào tạo tới node con

**Bước 4:** Nếu các ví dụ được phân lớp rõ ràng thì dừng.

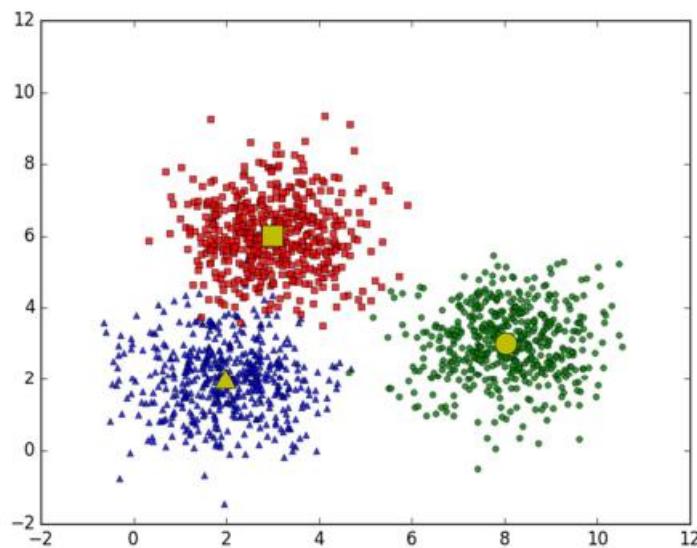
**Ngược lại:** lặp lại bước 1 tới bước 4 cho từng node con.

- ❖ **Cắt tia cây:** nhằm đơn giản hóa, khái quát hóa cây, tăng độ chính xác.

### 2.3.2. Phương pháp gom cụm.

- ❖ Nhận diện phần tử bất thường (outliers) và giảm thiểu nhiễu (noisy data)
- ❖ Gom cụm là kỹ thuật rất quan trọng trong khai phá dữ liệu, nó thuộc lớp các phương pháp **Unsupervised Learning** trong Machine Learning.
- ❖ Có rất nhiều định nghĩa khác nhau về kỹ thuật này, nhưng về bản chất ta có thể hiểu gom cụm là các qui trình tìm cách nhóm các đối tượng đã cho vào các cụm (clusters), Sao cho các đối tượng trong cùng 1 cụm tương tự (similar) nhau và các đối tượng khác cụm thì không tương tự (Disimilar) nhau.

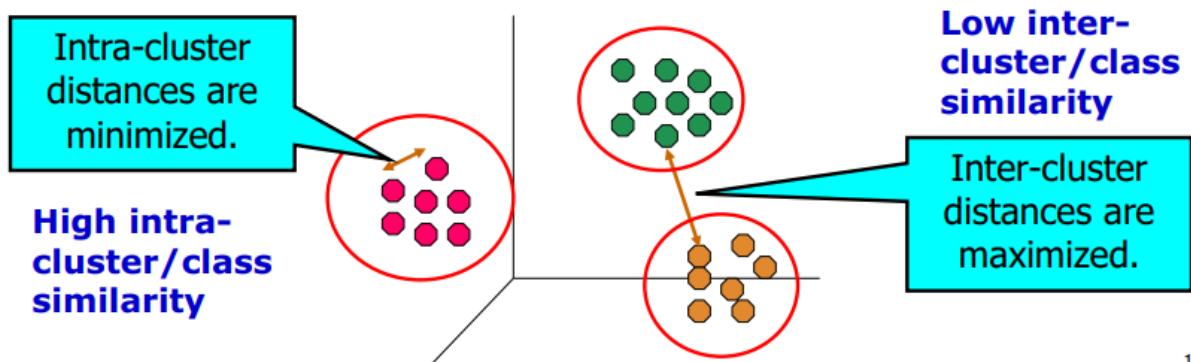
- ❖ Mục đích của gom cụm là tìm ra bản chất bên trong các nhóm của dữ liệu. Các thuật toán gom cụm (Clustering Algorithms) đều sinh ra các cụm (clusters).
- ❖ Tuy nhiên, không có tiêu chí nào là được xem là tốt nhất để đánh hiệu của của phân tích gom cụm, điều này phụ thuộc vào mục đích của gom cụm như:
  - “Data” reduction
  - “Natural” clusters
  - “Useful” clusters
  - “Outlier” detection



**Hình 2. 17.** Ví dụ về phương pháp gom cụm.

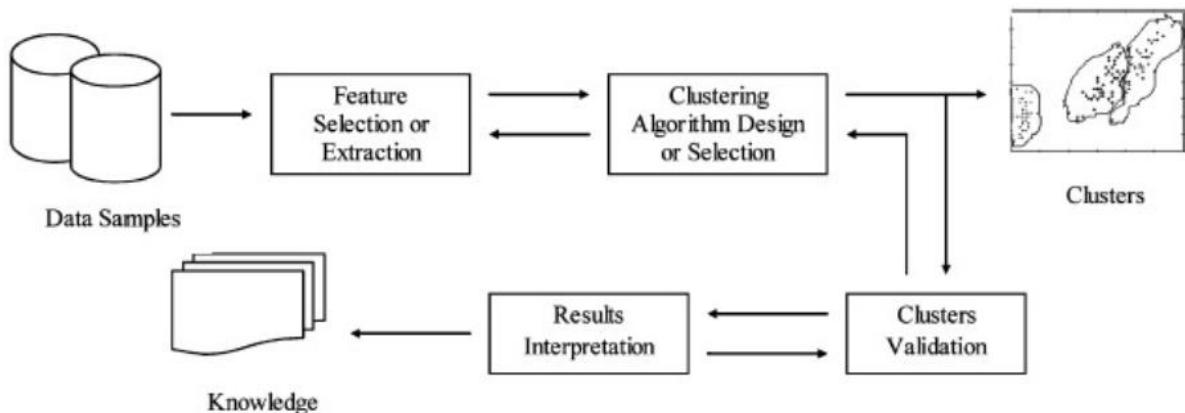
- ❖ Hỗ trợ giai đoạn tiền xử lý dữ liệu (data preprocessing);
- ❖ Mô tả sự phân bò dữ liệu/đối tượng (data distribution);
- ❖ Nhận dạng mẫu (pattern recognition);
- ❖ Phân tích dữ liệu không gian (spatial data analysis);
- ❖ Xử lý ảnh (image processing);
- ❖ Phân mảnh thị trường (market segmentation);
- ❖ Gom cụm tài liệu (WVW) document clustering);
- ❖ ...
- ❖ Gom cụm.
  - Quá trình gom nhóm/cụm dữ liệu/đối tượng vào các lớp/cụm.

- Các đối tượng trong cùng một cụm tương tự với nhau hơn so với đối tượng ở các cụm khác.
  - Obj1, Obj2 ở cụm C1; Obj3 ở cụm C2  $\Rightarrow$  Obj1 tương tự Obj2 hơn so vớiObj3.



Hình 2. 18. Phân tích chi tiết gom cụm.

❖ Quá trình gom cụm dữ liệu.



Hình 2. 19. Quá trình gom cụm dữ liệu.

❖ Các yêu cầu tiêu biểu về việc gom cụm dữ liệu.

- Khả năng co giãn về tập dữ liệu (scalability)
- Khả năng xử lý nhiều kiểu thuộc tính khác nhau (different types of attributes)
- Khả năng khám phá các cụm với hình dạng tùy ý (clusters with arbitrary shape)
- Tối thiểu hóa yêu cầu về tri thức miền trong việc xác định các thông số nhập (domain knowledge for input parameters)

- Khả năng xử lý dữ liệu có nhiễu (noisy data)
- Khả năng gom cụm tăng dần và độc lập với thứ tự của dữ liệu nhập (incremental clustering and insensitivity to the order of input records)
- Khả năng xử lý dữ liệu đa chiều (high dimensionality)
- Khả năng gom cụm dựa trên ràng buộc (constraint-based clustering)
- Khả diễn và khả dụng (interpretability and usability)

❖ **Phân loại các phương pháp gom cụm dữ liệu tiêu biểu.**

- Phân hoạch (partitioning): các phân hoạch được tạo ra và đánh giá theo một tiêu chí nào đó.
- Phân cấp (hierarchical): phân rã tập dữ liệu/đối tượng có thứ tự phân cấp theo một tiêu chí nào đó.
- Dựa trên mật độ (density-based): dựa trên connectivity and density functions.
- Dựa trên lưới (grid-based): dựa trên a multiple-level granularity structure.
- Dựa trên mô hình (model-based): một mô hình giả thuyết được đưa ra cho mỗi cụm; sau đó hiệu chỉnh các thông số để mô hình phù hợp với cụm dữ liệu/đối tượng nhất.

❖ **Các phương pháp đánh giá việc gom cụm dữ liệu.**

- Đánh giá ngoại (external validation)
  - Đánh giá kết quả gom cụm dựa vào cấu trúc được chỉ định trước cho tập dữ liệu
- Đánh giá nội (internal validation)
  - Đánh giá kết quả gom cụm theo số lượng các vector của chính tập dữ liệu (ma trận gần — proximity matrix)
- Đánh giá tương đối (relative validation)
  - Đánh giá kết quả gom cụm bằng việc so sánh các kết quả gom cụm khác ứng với các bộ trị thông số khác nhau

→ Tiêu chí cho việc đánh giá và chọn kết quả gom cụm tối ưu.

- Độ nén (compactness): các đối tượng trong cụm nên gần nhau.
- Độ phân tách (separation): các cụm nên xa nhau.

- Đánh giá ngoại (external validation)
  - Độ đo: Rand statistic, Jaccard coefficient, Folkes and Mallows index...
- Đánh giá nội (internal validation)
  - Độ đo: Huberts I statistic, Silhouette index, Dunn's index, ...
- Đánh giá tương đối (relative validation)
- Các độ đo đánh giá ngoại (external validation measures – contingency matrix)
- ❖ Phân 1 tập dữ liệu có n phần tử cho trước thành k tập con dữ liệu ( $k < n$ ), mỗi tập con biểu diễn 1 cụm.
- ❖ Các cụm hình thành trên cơ sở làm tối ưu giá trị hàm đo độ tương tự sao cho:
  - Các đối tượng trong 1 cụm là tương tự.
  - Các đối tượng trong các cụm khác nhau là không tương tự nhau.
- ❖ Đặc điểm:
  - Mỗi đối tượng chỉ thuộc về 1 cụm.
  - Mỗi cụm có tối thiểu 1 đối tượng.
- ❖ Một số thuật toán điển hình : K-mean, PAM, CLARA,...

Trong đồ án này ta sẽ chỉ tập trung phân tích Gom cụm với thuật toán **K-means**.

- ❖ **K-Means** là thuật toán rất quan trọng và được sử dụng phổ biến trong kỹ thuật gom cụm.
- ❖ Tư tưởng chính của thuật toán K-Means là tìm cách phân nhóm các đối tượng (objects) đã cho vào K cụm (K là số các cụm được xác định trước, K nguyên dương) sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm nhóm (centroid ) là nhỏ nhất.
- ❖ Thuật toán K-means.

Phát biểu bài toán:

- Input
  - Tập các đối tượng  $X = \{x_i | i = 1, 2, \dots, N\}$ ,
  - Số cụm: K
- Output

- Các cụm Ci ( $i = 1 / K$ ) tách rời và hàm tiêu chuẩn E đạt giá trị tối thiểu.
- Thuật toán hoạt động trên 1 tập vectơ d chiều, tập dữ liệu X gồm N phần tử:

$$X = \{x_i | i = 1, 2, \dots, N\},$$

- K-Mean lặp lại nhiều lần quá trình:
  - Gán dữ liệu.
  - Cập nhật lại vị trí trọng tâm.
- Quá trình lặp dừng lại khi trọng tâm hội tụ và mỗi đối tượng là 1 bộ phận của 1 cụm.
- Hàm đo độ tương tự sử dụng khoảng cách Euclidean trong đó  $c_j$  là trọng tâm của cụm  $C_j$

$$E = \sum_{i=1}^N \sum_{x_i \in C_j} (\|x_i - c_j\|^2)$$

*Hình 2. 20. Hàm đo độ tương tự.*

- Hàm trên không âm, giảm khi có một sự thay đổi trong 1 trong 2 bước: gán dữ liệu và định lại vị trí tâm.
- ❖ **Thuật toán K-Means thực hiện qua các bước chính sau:**

### Bước 1 - Khởi tạo

Chọn ngẫu nhiên K tâm (centroid) cho K cụm (cluster).

Mỗi cụm được đại diện bằng các tâm của cụm.

K trọng tâm  $\{c_i\}$  ( $i = 1/K$ ).

**Bước 2** - Tính khoảng cách giữa các đối tượng (objects) đến K tâm (thường dùng khoảng cách Euclidean)

$$S_i^{(t)} = \{x_j : \|x_j - c_i^{(t)}\| \leq \|x_j - c_i^*\| \text{ for all } i^* = 1, \dots, k\}$$

**Hình 2. 21.** Công thức tính khoảng cách giữa các đối tượng.

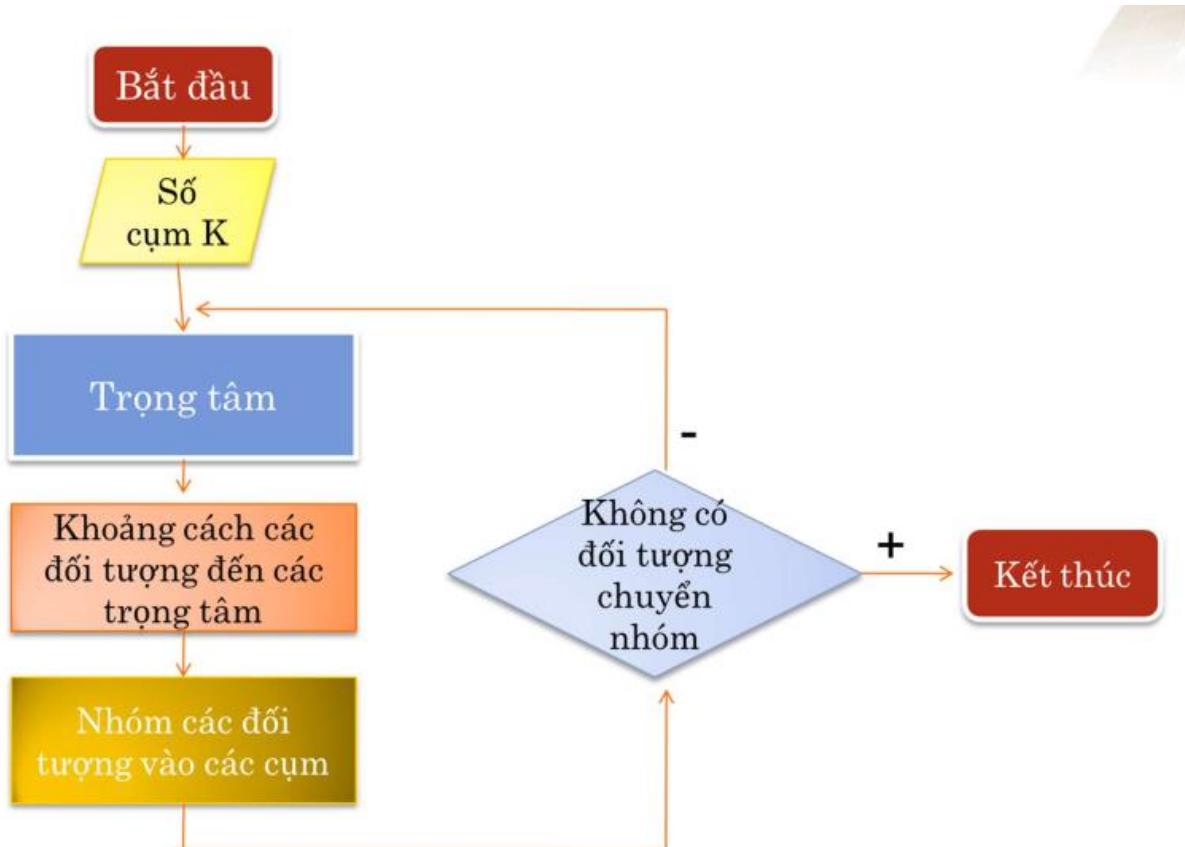
**Bước 3** – Cập nhật lại trọng tâm.

$$c_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

**Hình 2. 22.** Công thức cập nhật lại trọng tâm.

**Bước 4** – Điều kiện dừng

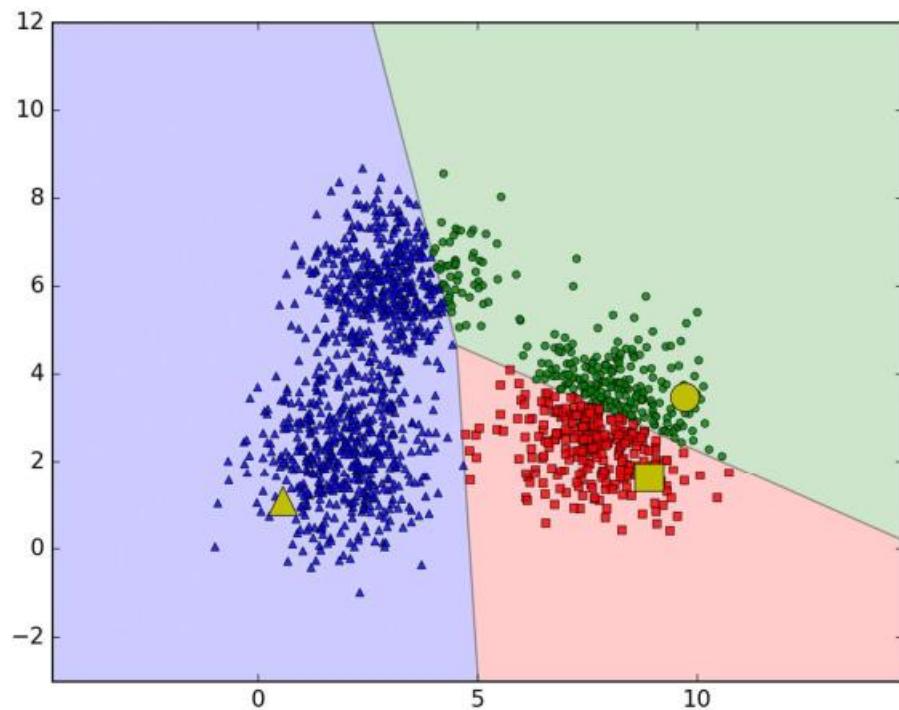
Lặp lại bước 2 và 3 cho tới khi không có sự thay đổi trọng tâm của cụm.



**Hình 2. 23.** Thuật toán K-Means.

❖ **Đặc điểm của giải thuật K-Means.**

- Sau càng nhiều vòng lặp, các tâm càng di chuyển chậm dần, và tổng khoảng cách từ mỗi điểm trong cụm tới tâm cụm lại càng nhỏ đi.
  - Quá trình sẽ kết thúc cho tới khi hàm tổng khoảng cách hội tụ (tức là không có sự thay đổi nào xảy ra ở giai đoạn gán nữa).
  - Tọa độ tâm vẫn sẽ bằng trung bình cộng các điểm hiện tại trong cụm, hay nói cách khác tâm sẽ không còn di chuyển tiếp nữa.
- ⇒ **thuật toán K-Means** chỉ đảm bảo được quá trình này sẽ đưa hàm tổng khoảng cách hội tụ tới điểm cực tiểu địa phương, chứ **KHÔNG** chắc chắn đó là giá trị nhỏ nhất của toàn bộ hàm số.
- Tùy vào các center ban đầu mà thuật toán có thể có tốc độ hội tụ rất chậm

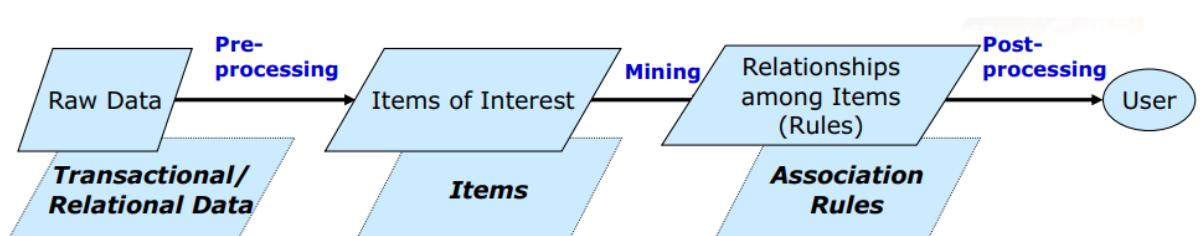


*Hình 2. 24. Gán điểm cho các cụm.*

❖ **Ưu điểm của thuật toán K-Means.**

- Độ phức tạp:  $O(K, X, I)$  với  $I$ : số lần lặp
- Có khả năng mở rộng, có thể dễ dàng sửa đổi với những dữ liệu mới.
- Bảo đảm hội tụ sau 1 số bước lặp hữu hạn.

- Luôn có K cụm dữ liệu
  - Luôn có ít nhất 1 điểm dữ liệu trong 1 cụm dữ liệu.
  - Các cụm không phân cấp và không bị chồng chéo dữ liệu lên nhau.
  - Mọi thành viên của 1 cụm là gần với chính cụm đó hơn bất cứ 1 cụm nào khác.
- ❖ Nhược điểm của thuật toán K-Means.
- Không có khả năng tìm ra các cụm không lồi hoặc các cụm có hình dạng phức tạp.
  - Khó khăn trong việc xác định các trọng tâm cụm ban đầu
    - Chọn ngẫu nhiên các trung tâm cụm lúc khởi tạo.
    - Độ hội tụ của thuật toán phụ thuộc vào việc khởi tạo các vector trung tâm cụm.
  - Khó để chọn ra được số lượng cụm tối ưu ngay từ đầu, mà phải qua nhiều lần thử để tìm ra được số lượng cụm tối ưu.
  - Rất nhạy cảm với nhiễu và các phần tử ngoại lai trong dữ liệu.
  - Không phải lúc nào mỗi đối tượng cũng chỉ thuộc về 1 cụm, chỉ phù hợp với đường biên giữa các cụm rõ.
- ❖ Biến thể của thuật toán K-Means.
- Thuật toán K-medoid.
  - Thuật toán Fuzzy c-mean (FCM)
- 2.3.3. Phương pháp luật kết hợp.**
- ❖ Quá trình khai phá luật kết hợp.



Hình 2. 25. Quá trình khai phá luật kết hợp.

❖ Các khái niệm cơ bản.

- Item (phản tử)

- Các phần tử, mẫu, đối tượng đang được quan tâm.
- $J = \{I1, I2, \dots, Im\}$ : tập tất cả m phân tử có thể có trong tệp dữ liệu.
- *Itemset (tập phân tử)*
  - Tập hợp các items.
  - Một itemset có k items gọi là k-itemset.
- *Transaction (giao dịch)*
  - Lần thực hiện tương tác với hệ thống (ví dụ: giao dịch “khách hàng mua hàng”)
  - Liên hệ với một tập T gồm các phân tử được giao dịch.
- *Association (sự kết hợp) và association rule (luật kết hợp)*
  - Sự kết hợp: các phân tử cùng xuất hiện với nhau trong một hay nhiều giao dịch.
    - Thể hiện mi liên hệ giữa các phân tử/các tập phân tử
  - Luật kết hợp: qui tắc kết hợp có điều kiện giữa các tập phân tử.
    - Thể hiện mi liên hệ (có điều kiện) giữa các tập phân tử.
    - Cho X và Y là các tập phân tử, luật kết hợp giữa X và Y là  $X \rightarrow Y$ .
- ⇒ Y xuất hiện trong điều kiện X xuất hiện.
- *Support (độ hỗ trợ)*
  - Độ hỗ trợ (Support) của luật kết hợp  $X \rightarrow Y$  là tần suất của giao dịch chứa tất cả các items trong cả hai tập X và Y.
  - Ví dụ, support của luật  $X \rightarrow Y$  là 5% có nghĩa là 5% các giao dịch X và Y được mua cùng nhau.
  - Công thức để tính support của luật  $X \rightarrow Y$  với N là tổng số giao dịch như sau:

$$support(X \rightarrow Y) = P(X \cup Y) = \frac{n(X \cup Y)}{N}$$

**Hình 2. 26.** Công thức tính support của luật  $X \rightarrow Y$ .

- *Confidence (độ tin cậy)*

- Độ tin cậy (Confidence) của luật kết hợp  $X \rightarrow Y$  là xác suất xảy ra  $Y$  khi đã biết  $X$ .
- Ví dụ độ tin cậy của luật kết hợp  $\{\text{Apple}\} \rightarrow \text{Banana}$  là 80% có nghĩa là 90% khách hàng mua Apple cũng mua Banana.
- Công thức để tính độ tin cậy của luật kết hợp  $X \rightarrow Y$  là xác suất có điều kiện  $Y$  khi đã biết  $X$  với  $n(X)$  là số giao dịch chứa  $X$  như sau :

$$\text{confidence}(X \rightarrow Y) = P(Y|X) = \frac{n(X \cup Y)}{n(X)}$$

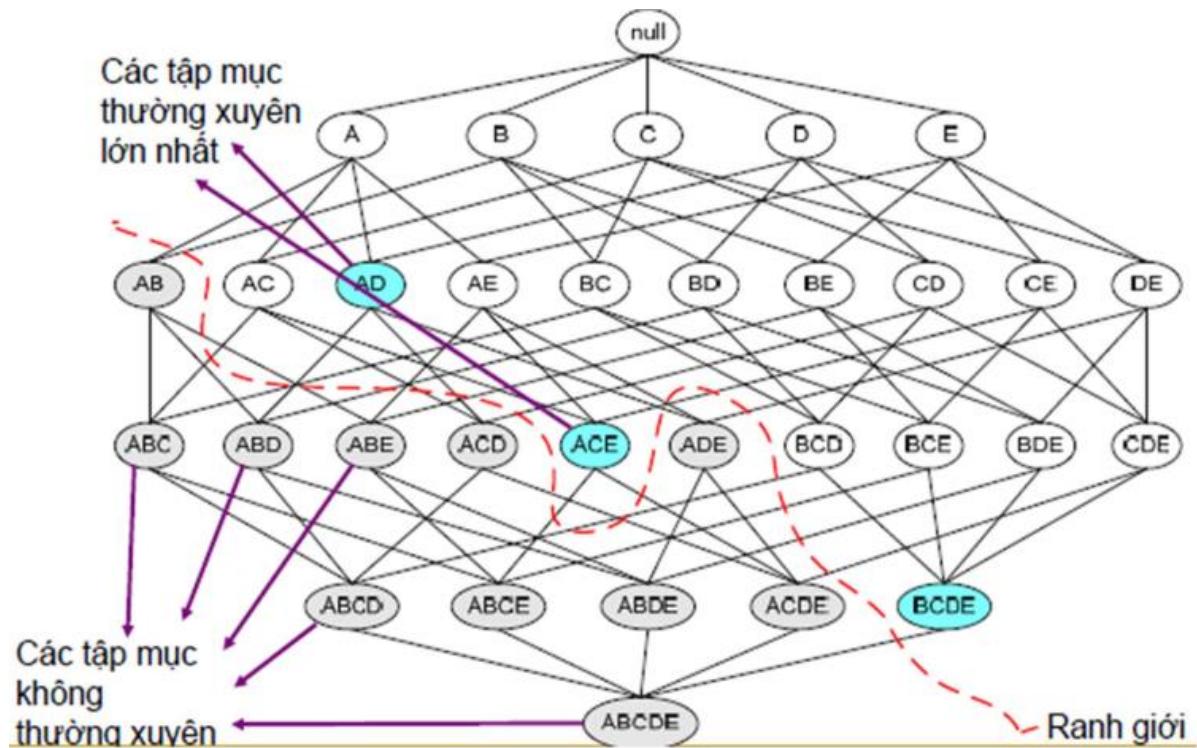
**Hình 2. 27.** Công thức để tính độ tin cậy của luật kết hợp  $X \rightarrow Y$ .

- *Frequent itemset (tập phần tử phổ biến/thường xuyên)*
  - Cho tập mục  $X \subseteq I$  với ngưỡng hỗ trợ tối thiểu (minimum support)  $\text{minsup} \in [0,1]$  (được xác định trước bởi người sử dụng).
  - $X$  được gọi là tập mục thường xuyên (frequent itemset hoặc large itemset) với độ hỗ trợ tối thiểu  $\text{minsup}$  nếu  $\text{sup}(X) \geq \text{minsup}$ , ngược lại  $X$  gọi là tập mục không thường xuyên.
- *Strong association rule (luật kết hợp mạnh)*
  - Để thu được các luật kết hợp, ta thường áp dụng 2 tiêu chí: minimum support (min\_sup) và minimum confidence (min\_conf)
  - Các luật thỏa mãn có support và confidence thỏa mãn (lớn hơn hoặc bằng) cả Minimum support và Minimum confidence gọi là các luật mạnh (Strong Rule)
  - Minimum support và Minimum confidence gọi là các giá trị ngưỡng (threshold) và phải xác định trước khi sinh các luật kết hợp.

#### ❖ Phân loại luật kết hợp.

- Boolean association rule (luật kết hợp luận lý)/ quantitative association rule (luật kết hợp lượng số)
- Single-dimensional association rule (luật kết hợp đơn chiều)/ multidimensional association rule (luật kết hợp đa chiều)

- Single-level associalion rule (luật kết hợp đơn mức)/ multilevel association rule (luật kết hợp đa mức)
  - Association rule (luật kết hợp)/ correlation rule (luật tương quan thống kê)
- ❖ Dạng luật:  $X \rightarrow Y$  [support, confident]
- Cho trước minimum support threshold (min\_sup), minimum confidence threshold (cmin\_conf)
  - X và Y là các itemsets.
    - Frequent itemsets/subsequences/substructures.
    - Closed frequent itemsets.
    - Maximal frequent itemsets.
    - Constrained frequent itemsets.
    - Approximate frequent itemsets.
    - Top-k frequent itemsets.
- ❖ Frequent itemsets/ subsequences/ substructures.
- Itemset/ subsequence/ substructure. X là frequent nếu support (X)  $\geq$  min\_sup.
  - Itemsets: tập các items.
  - Subsequences: chuỗi tuần tự các events/items.
  - Substructures: các tiêu cấu trúc (graph, lattice, tree, sequence, set,...)
- ❖ Maximal frequent itemsets: một tập mục thường xuyên là lớn nhất nếu mọi tập cha (superset) của nó đều là tập mục không thường xuyên.



Hình 2. 28. Maximal frequent itemsets.

- ❖ **Closed frequent itemsets:** một tập mục thường xuyên là đóng không có tập cha nào của nó có cùng độ hỗ trợ với nó.
- ❖ **Constrained frequent itemsets.**
  - Frequent itemsets thỏa các ràng buộc do người dùng định nghĩa.
- ❖ **Approximate frequent itemsets.**
  - Frequent itemsets dẫn ra support (xấp xỉ) cho các frequent itemsets sẽ được khai phá.
- ❖ **Top-k frequent itemsets.**
  - Frequent itemsets có nhiều nhất k phần tử với k do người dùng chỉ định.
- ❖ Luật kết hợp luận lý, đơn mức, đơn chiều giữa các tập phần tử phổ biến:  $X \rightarrow Y$  [support, confidence]
  - X và Y là các frequent itemsets
    - Single-dimensional.
    - Single-level.
    - Boolean.
  - $\text{Support}(X \rightarrow Y) = \text{Support}(X \cup Y) \geq \text{min\_sup}$ .

- $\text{Confidence}(X>2Y) = \text{Support}(X \cup Y)/\text{Support}(X) = P(B|A) \geq \text{min\_conf.}$

Trong đồ án này ta sẽ chỉ tập trung phân tích Kết hợp với giải thuật Apriori.

❖ **Giải thuật Apriori:**

- Sinh ra tất cả các tập mục thường xuyên mức 1 (frequent 1-1 itemsets): các tập mục thường xuyên chỉ chứa 1 mục.
- Gán  $k = 1$ .
- Lặp lại, cho đến khi không có thêm bất kỳ tập mục thường xuyên nào mới.
  - Từ các tập mục thường xuyên mức  $k$  (chứa  $k$  mục), sinh ra các tập mục mức  $(k+1)$  cần xét.
  - Loại bỏ các tập mục mức  $(k+1)$  chứa các tập con là các tập mục không thường xuyên mức  $k$ .
  - Tính độ hỗ trợ của mỗi tập mục mức  $(k+1)$ , bằng cách duyệt qua tất cả các giao dịch.
  - Loại bỏ các tập mục không thường xuyên mức  $(k+1)$
  - Thu được các tập mục thường xuyên mức  $(k+1)$

❖ **Nguyên tắc của giải thuật Apriori** – loại bỏ (prunning) dựa trên độ hỗ trợ.

- Nếu một tập mục là thường xuyên, thì tất cả các tập con (subsets) của nó đều là các tập mục thường xuyên.
- Nếu một tập mục là không thường xuyên (not frequent), thì tất cả các tập cha (supersets) của nó đều là các tập mục không thường xuyên.

❖ **Nguyên tắc của giải thuật Apriori** dựa trên đặc tính không đơn điệu (anti-monotone) của độ hỗ trợ.

- $\forall X, Y: (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$
- Độ hỗ trợ của một tập mục nhỏ hơn độ hỗ trợ của các tập con của nó.

❖ **Các cải tiến của giải thuật Apriori.**

❖ **Kỹ thuật dựa trên bảng băm** (hash-based technique)

- Một  $k$ -itemset ứng với hashing bucket count nhỏ hơn minimum support threshold không là một frequent itemset.

❖ **Giảm giao dịch** (transaction reduction)

- Một giao dịch không chứa frequent k-itemset nào thì không cần được kiểm tra ở các lần sau (cho k+1-itemset).

❖ **Phân hoạch** (partitioning)

- Một itemset phải frequent trong ít nhất một phân hoạch thì mới có thể frequent trong toàn bộ tập dữ liệu.

❖ **Lấy mẫu** (sampling)

- Khai phá chỉ tập con dữ liệu cho trước với một trị support threshold nhỏ hơn và cần một phương pháp để xác định tính toàn diện (completeness).

❖ **Đếm itemset động** (dynamic itemset counting)

- Chỉ thêm các itemsets dự tuyển khi tất cả các tập con của chúng được dự đoán là frequent.

## 2.4. Giới thiệu về phần mềm sử dụng (IBM SPSS Modeler)

### 2.4.1. Tổng quan về phần mềm IBM SPSS Modeler.

**IBM SPSS Modeler** là một phần mềm phân tích dữ liệu và khai thác dữ liệu được phát triển bởi IBM. Nó cung cấp cho người dùng các công cụ để thực hiện các tác vụ phân tích dữ liệu và khai thác dữ liệu phức tạp, bao gồm việc tạo mô hình dự đoán, phân tích định lượng, phân tích chuỗi thời gian, phân tích quyết định và phân tích khách hàng. Ngoài ra, IBM SPSS Modeler còn cung cấp khả năng tích hợp với các công cụ khác như Excel, Python và R để cung cấp cho người dùng các công cụ phân tích dữ liệu linh hoạt hơn.

Nó cũng cung cấp các công cụ trực quan và dễ sử dụng để thực hiện các tác vụ phân tích này, cho phép người dùng xử lý dữ liệu lớn và phức tạp một cách hiệu quả. Phần mềm này có khả năng xử lý hàng trăm triệu hàng hoặc cột và cho phép người dùng khai thác dữ liệu một cách hiệu quả hơn.

**IBM SPSS Modeler** được sử dụng rộng rãi trong nhiều lĩnh vực, bao gồm kinh doanh, tài chính, y tế, giáo dục và chính phủ. Với giao diện người dùng trực quan và dễ sử dụng, nó cho phép người dùng tìm hiểu và sử dụng các tính năng phân tích dữ liệu một cách nhanh chóng và dễ dàng.

#### **2.4.1.1 Ưu điểm.**

**IBM SPSS Modeler** có nhiều ưu điểm như sau:

- **Khả năng xử lý dữ liệu lớn:** IBM SPSS Modeler có thể xử lý dữ liệu lớn và phức tạp, cho phép người dùng phân tích các tập dữ liệu lớn một cách hiệu quả.
- **Các công cụ phân tích mạnh mẽ:** IBM SPSS Modeler cung cấp các công cụ phân tích mạnh mẽ để thực hiện các tác vụ phân tích dữ liệu và khai thác dữ liệu phức tạp như phân tích định lượng, phân tích chuỗi thời gian, phân tích quyết định, phân tích khách hàng và tạo mô hình dự đoán.
- **Giao diện người dùng trực quan:** IBM SPSS Modeler có giao diện người dùng trực quan, cho phép người dùng dễ dàng thực hiện các tác vụ phân tích dữ liệu một cách nhanh chóng và dễ dàng.
- **Tích hợp với các công cụ khác:** IBM SPSS Modeler có khả năng tích hợp với các công cụ khác, cho phép người dùng kết hợp các kết quả phân tích với các công cụ khác như Excel và SQL.
- **Hỗ trợ đa nền tảng:** IBM SPSS Modeler có thể chạy trên nhiều hệ điều hành khác nhau và hỗ trợ nhiều ngôn ngữ lập trình, cho phép người dùng sử dụng nó trên nhiều nền tảng khác nhau.

Tóm lại, IBM SPSS Modeler là một công cụ phân tích dữ liệu mạnh mẽ và linh hoạt, có thể giúp người dùng phân tích và khai thác dữ liệu một cách hiệu quả để tạo ra các kết quả có ý nghĩa cho doanh nghiệp hoặc tổ chức.

#### **2.4.1.2 Hạn chế.**

Tuy có nhiều ưu điểm, **IBM SPSS Modeler** vẫn còn một số hạn chế như sau:

- **Giá thành cao:** IBM SPSS Modeler là một phần mềm có giá thành cao, đặc biệt là đối với các tổ chức hoặc doanh nghiệp nhỏ. Bản dùng thử của nó có thời hạn là 30 ngày kể từ ngày cài đặt thành công phần mềm.
- **Khả năng phân tích dữ liệu không linh hoạt:** Mặc dù IBM SPSS Modeler cung cấp các công cụ phân tích mạnh mẽ, nhưng không phải tất cả các loại phân tích dữ liệu đều được hỗ trợ.

- **Khó khăn trong việc cài đặt và sử dụng:** IBM SPSS Modeler có giao diện phức tạp và cần thời gian để học cách sử dụng các tính năng và công cụ của nó.
- **Khả năng tích hợp với các hệ thống khác hạn chế:** IBM SPSS Modeler có thể gặp khó khăn trong việc tích hợp với các hệ thống khác, đặc biệt là đối với các hệ thống phân tích dữ liệu khác.
- **Khả năng mở rộng hạn chế:** IBM SPSS Modeler không cung cấp cho người dùng các tính năng mở rộng để tùy chỉnh hoặc phát triển thêm các tính năng phân tích dữ liệu mới.

Tóm lại, IBM SPSS Modeler là một công cụ phân tích dữ liệu mạnh mẽ, nhưng nó cũng có một số hạn chế, đặc biệt là đối với các tổ chức hoặc doanh nghiệp nhỏ.

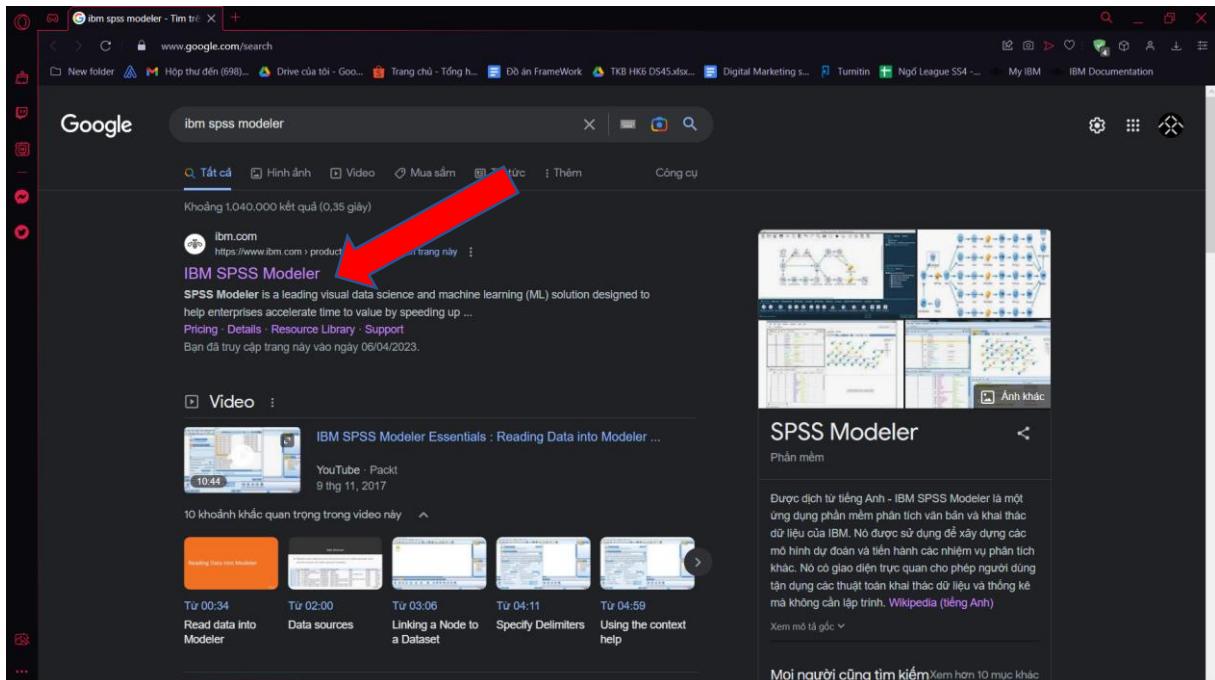
#### **2.4.2. Cách sử dụng phần mềm.**

##### **2.4.2.1. Giới thiệu giao diện.**

###### **❖ Cách cài đặt phần mềm.**

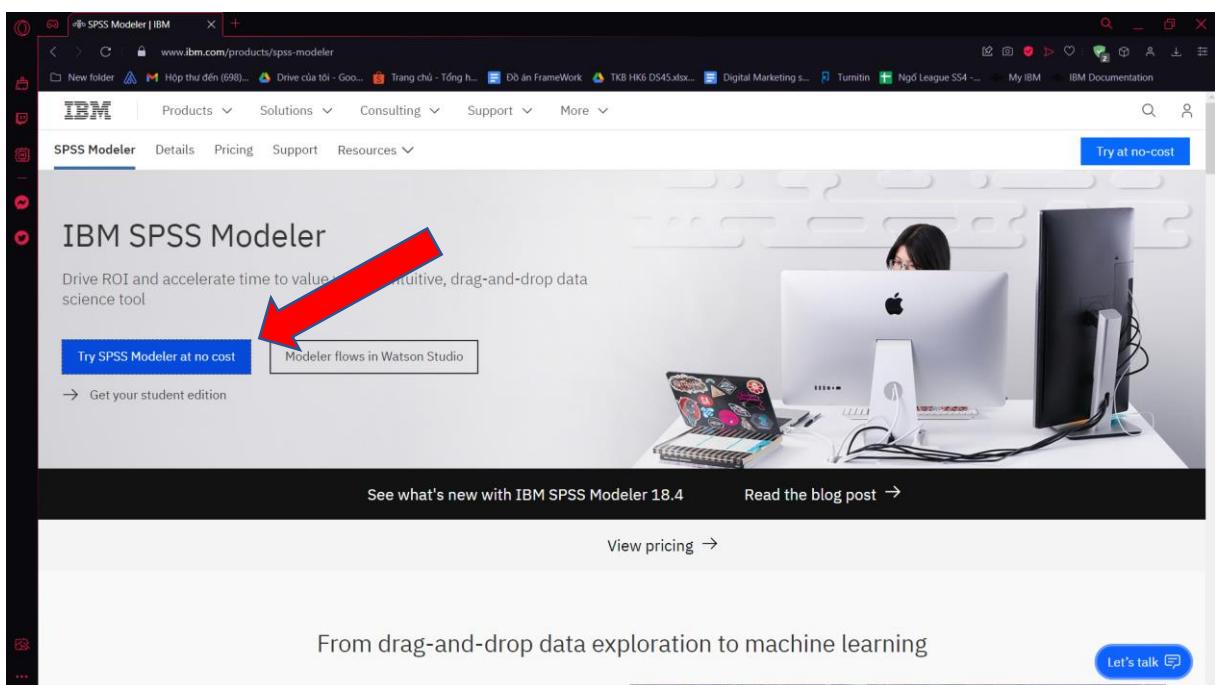
Vì không có chi phí để sử dụng phần mềm hoàn chỉnh nên nhóm chúng em sẽ chỉ trình bày về cách cài đặt bản dùng thử 30 ngày của IBM SPSS Modeler.

**Bước 1:** Vào trình duyệt bất kỳ gõ tìm IBM SPSS Modeler, bạn sẽ thấy kết quả được hiển thị ngay trên màn hình, sau đó nhấn chọn IBM SPSS Modeler.

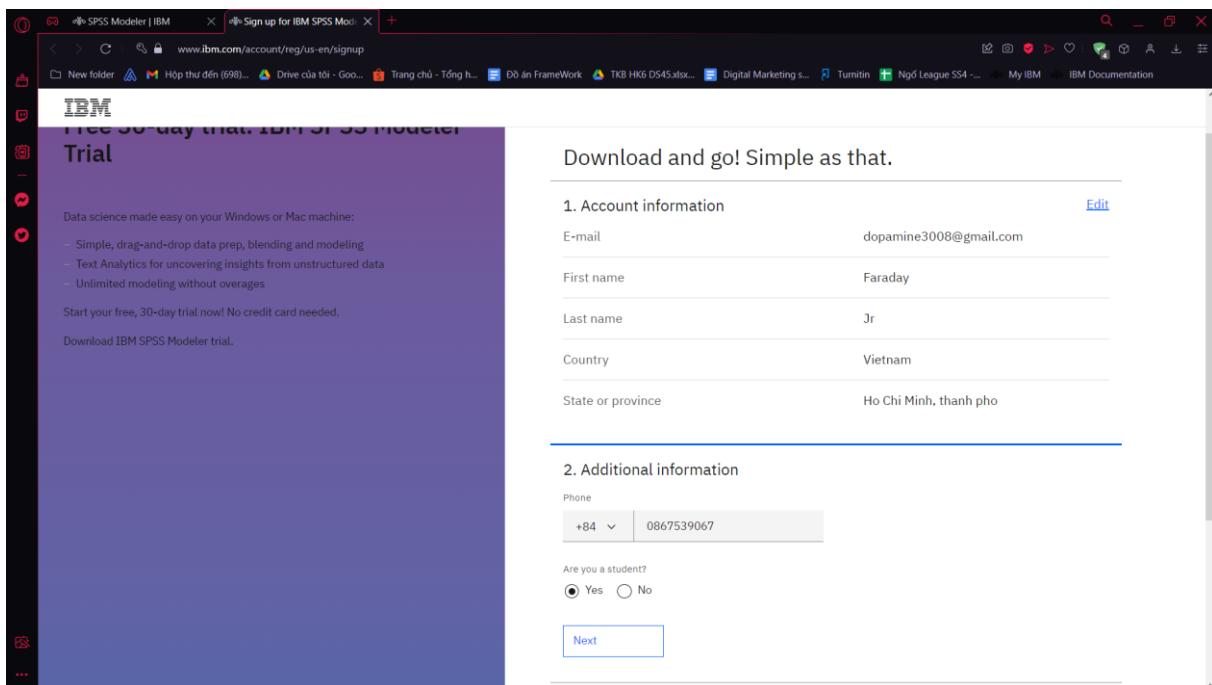


**Hình 2. 29. Gõ tìm IBM SPSS Modeler.**

**Bước 2:** Chọn Try SPSS Modeler at no cost.

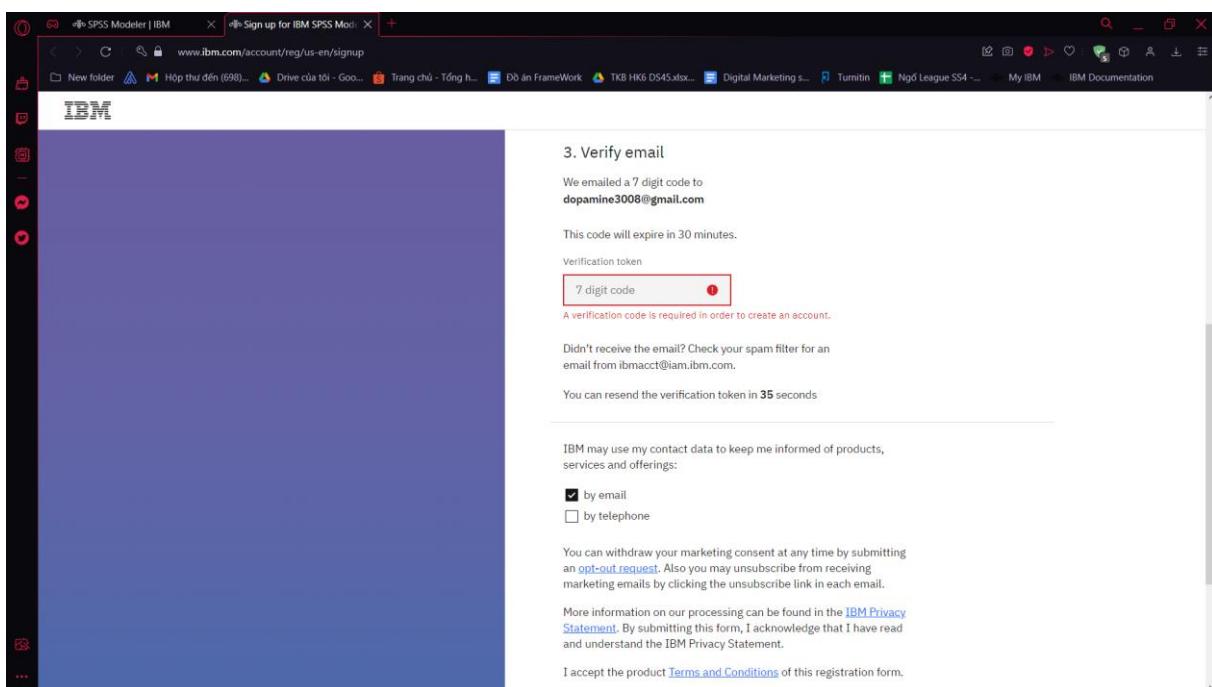


**Bước 3:** Điền đầy đủ thông tin cần thiết vào phần Account information và Addition infomation sau đó bấm Next.



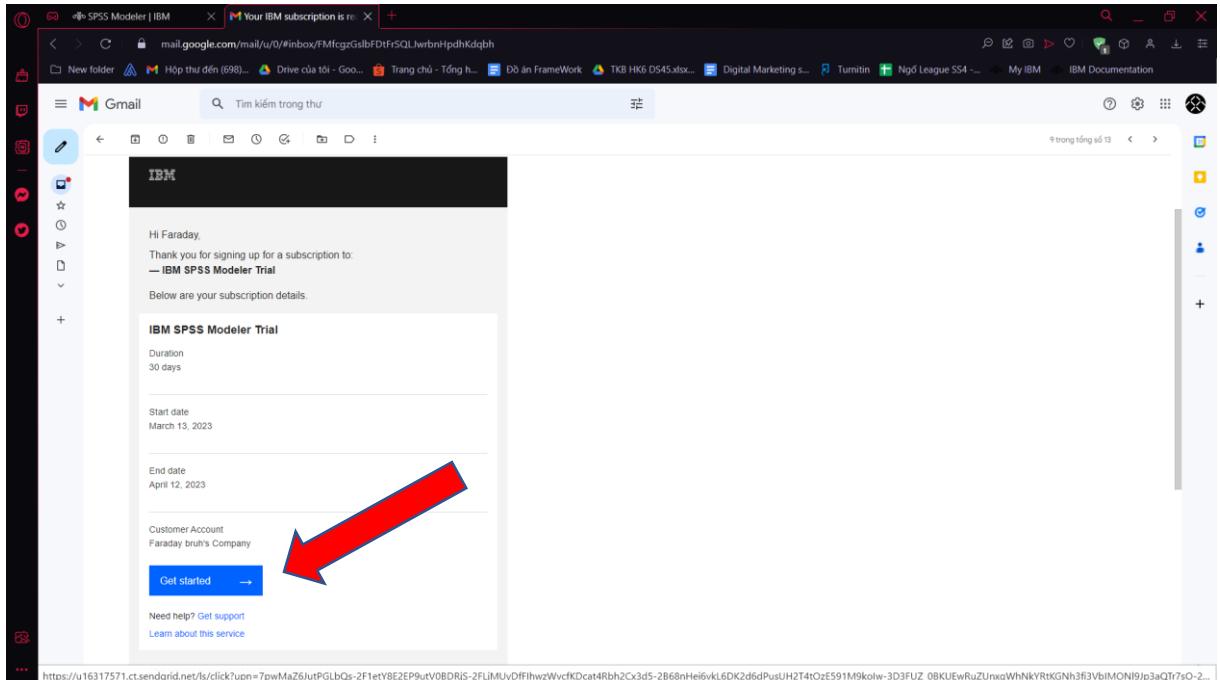
**Hình 2. 30.** *Điền đầy đủ thông tin cần thiết.*

**Bước 4:** Xác nhận thông tin của bạn qua Gmail hoặc Telephone.



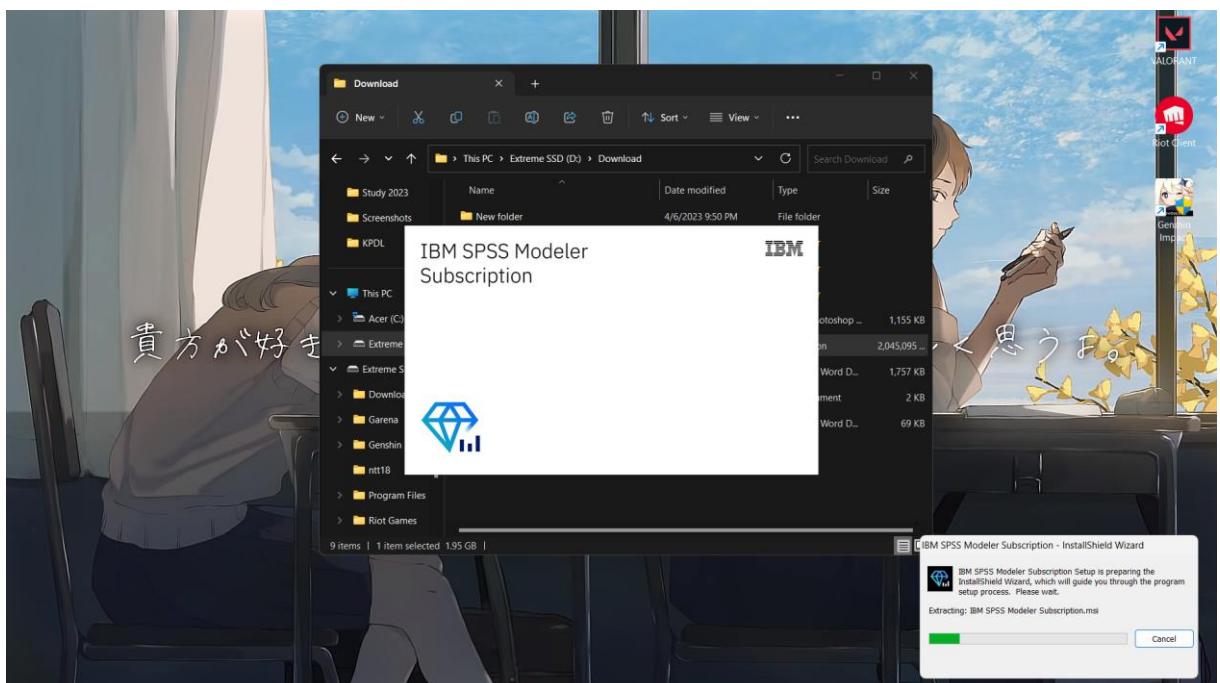
**Hình 2. 31.** *Nhập mã xác nhận từ Mail hoặc Telephone.*

**Bước 5:** Sau khi xác nhận thành công IBM sẽ gửi mail đính kèm thông báo hạn sử dụng của phần mềm, tên người dùng đã đăng ký cùng với đường link để tải phần mềm. Ở đây chọn Get started.



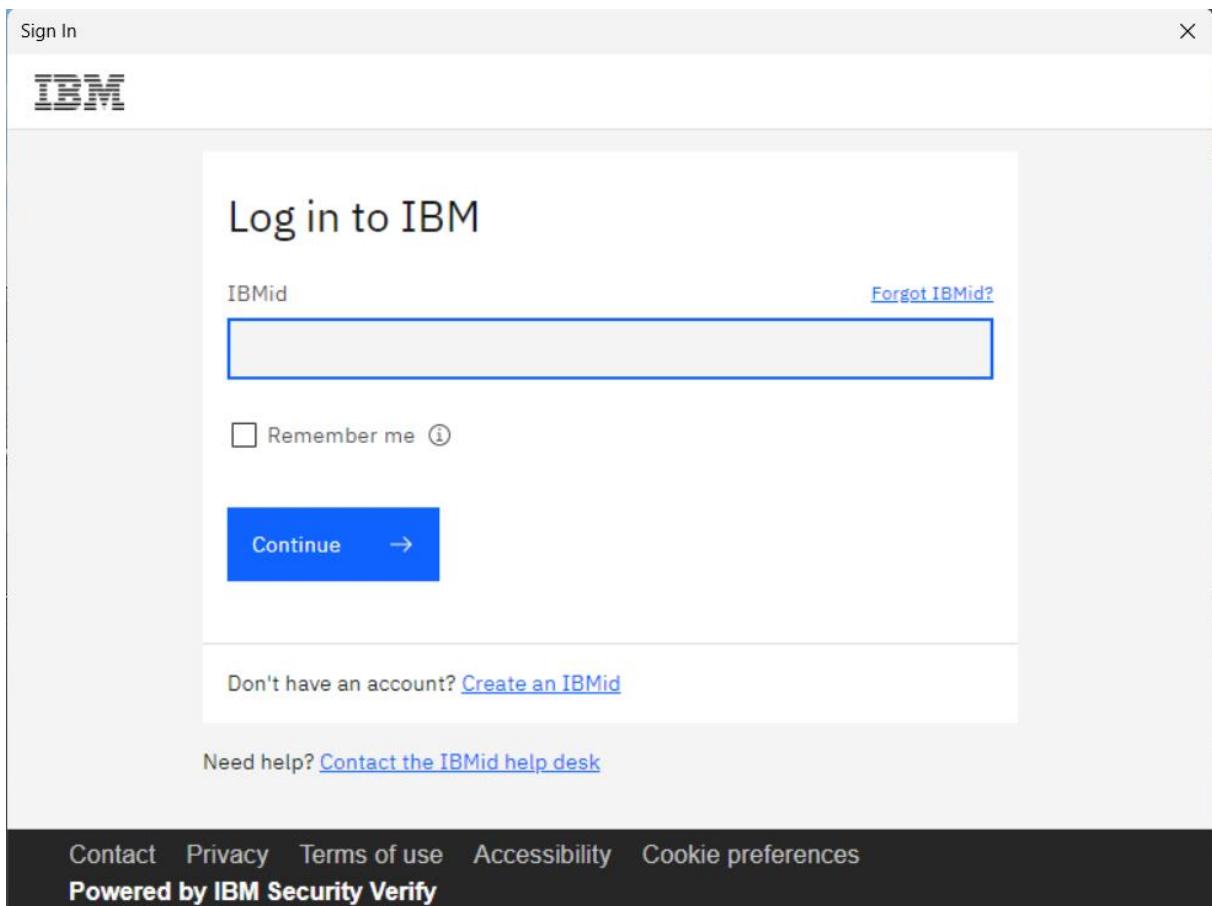
**Hình 2. 32.** Tải IBM SPSS Modeler về máy tính.

**Bước 6:** Sau khi đã tải về thành công Setup của IBM SPSS Modeler, ta tiến hành chạy cài đặt IBM SPSS Modeler trên thiết bị.



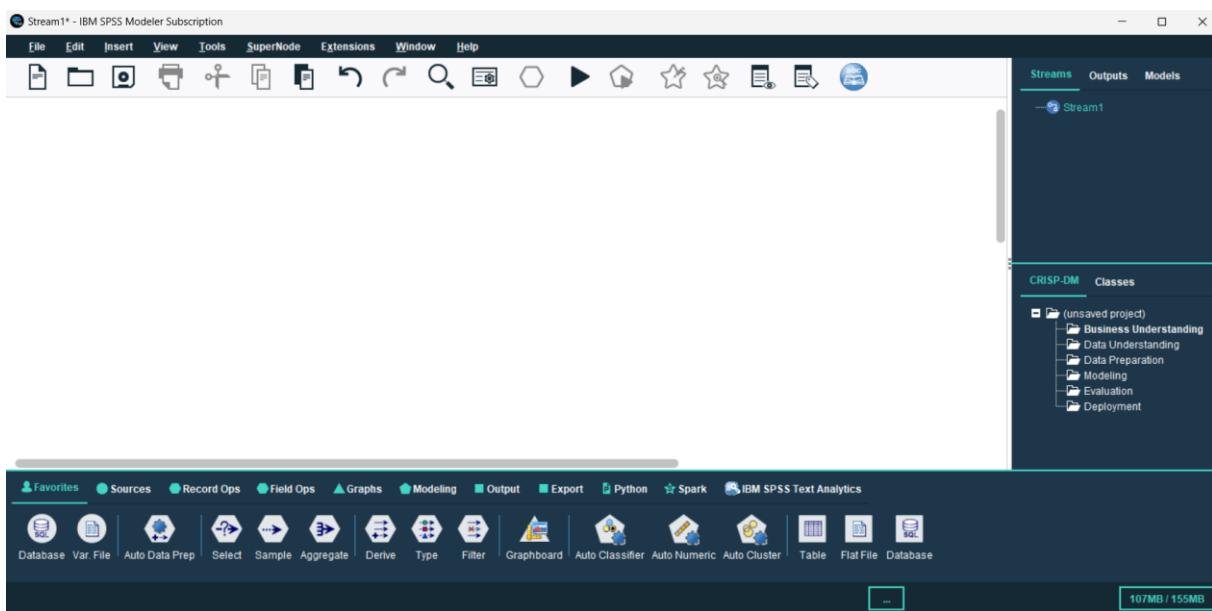
**Hình 2. 33.** Chạy cài đặt IBM SPSS Modeler.

**Bước 7:** Sau khi cài đặt thành công, từ đây bạn đã có thể đăng nhập bằng tài khoản đã đăng ký và sử dụng một cách bình thường.



*Hình 2. 34. Giao diện đăng nhập vào IBM SPSS Modeler.*

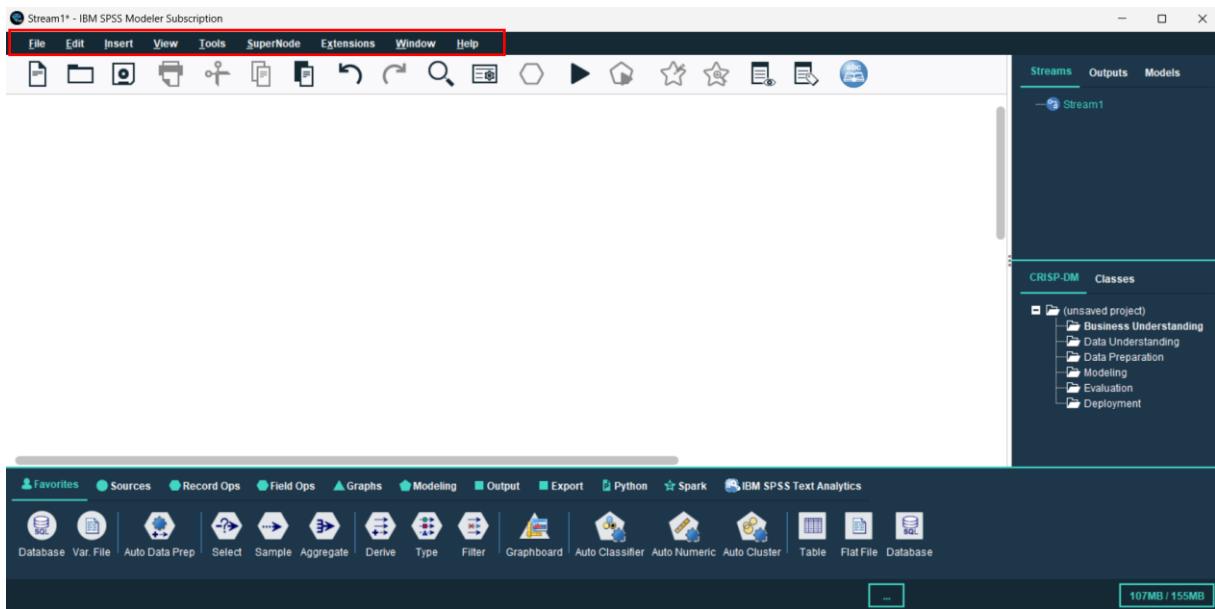
❖ *Giới thiệu giao diện.*



*Hình 2. 35. Giao diện chính của IBM SPSS Modeler.*

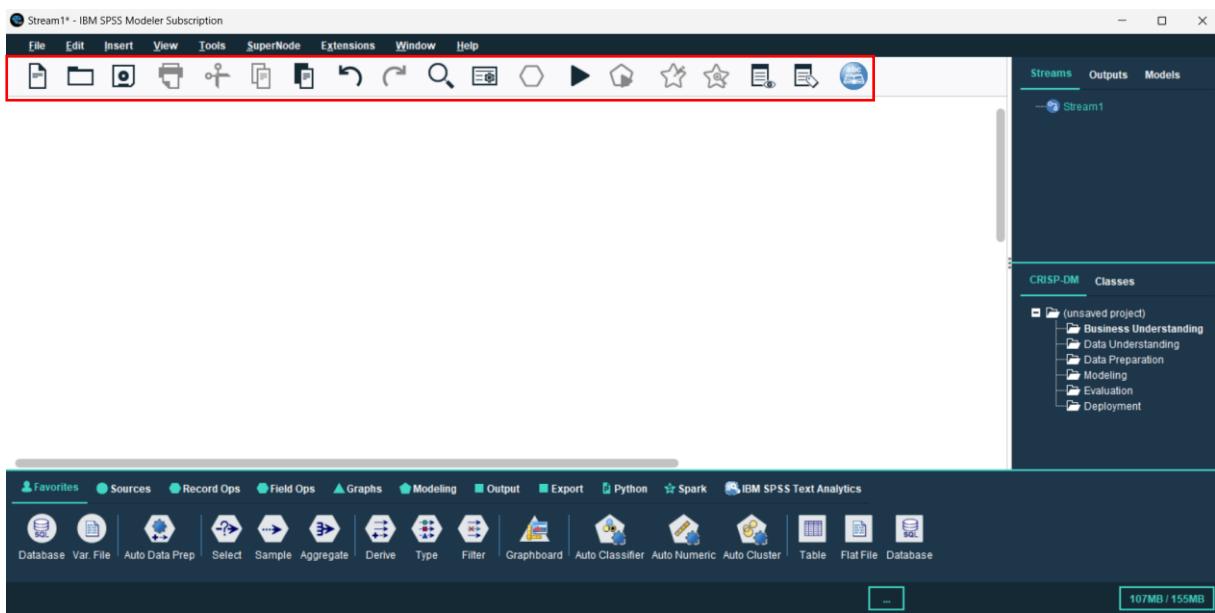
Giao diện của IBM SPSS Modeler bao gồm các thành phần sau:

**Ribbon:** Thanh công cụ ở đầu trang chứa các nút lệnh để thực hiện các tác vụ phổ biến như mở tập tin, lưu tập tin, thêm nút, chạy mô hình, và xuất kết quả.



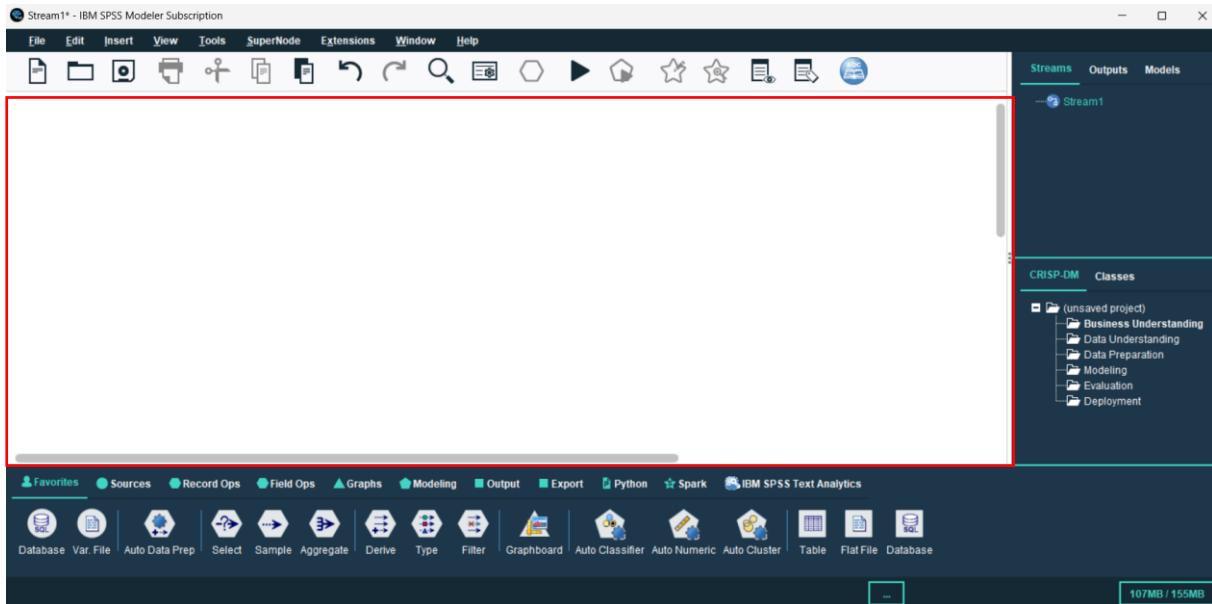
*Hình 2. 36. Ribbon.*

**Toolbar:** Thanh công cụ nằm ở đầu Stream Window để thực hiện các tác vụ như thêm nút, chạy mô hình, và chỉnh sửa các thuộc tính của nút.



*Hình 2. 37. Toolbar.*

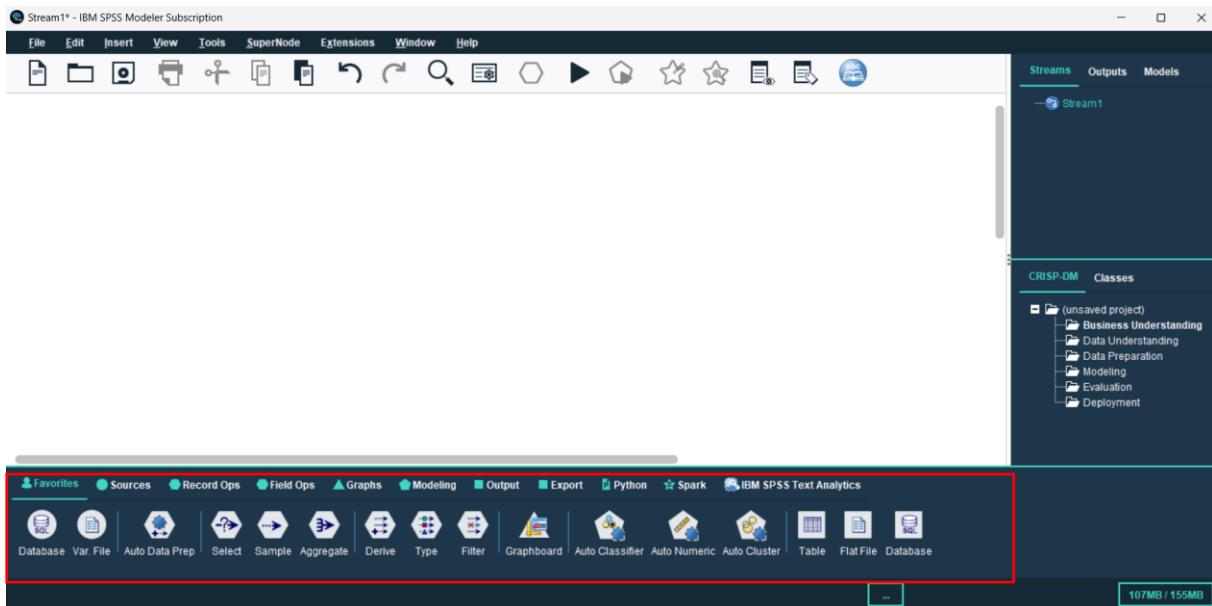
**Stream Window:** Cửa sổ chính hiển thị dòng chảy của các nút được sử dụng để xây dựng mô hình, các kết nối giữa chúng, và các tham số được cấu hình cho mỗi nút.



Hình 2. 38. Stream Window.

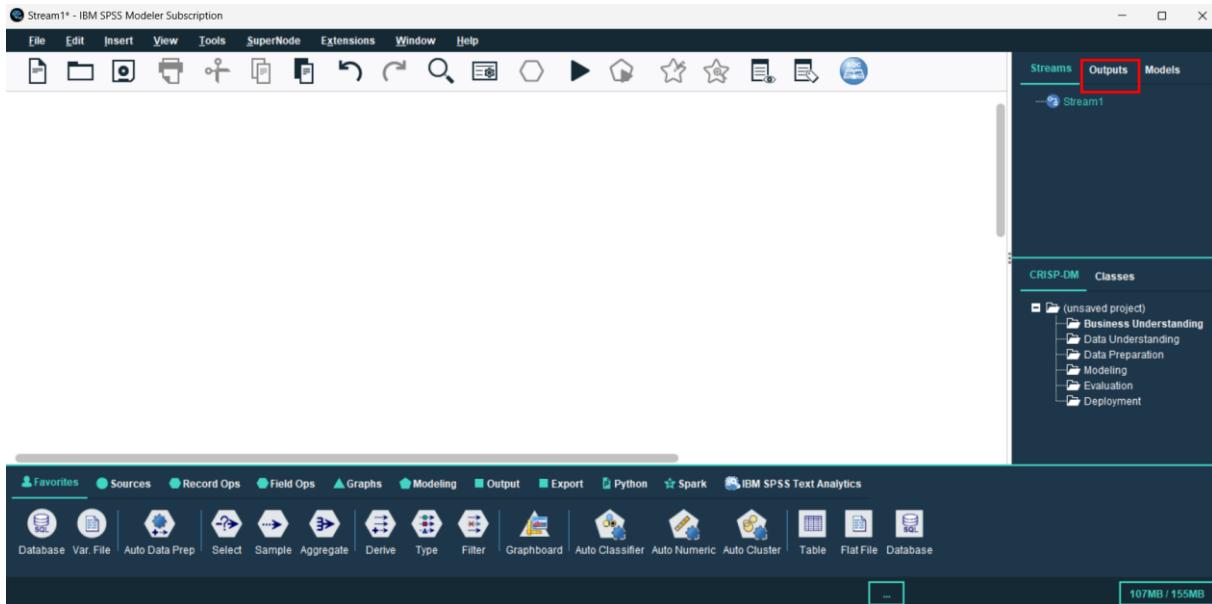
**Node Palette:** Bảng điều khiển chứa các nút để thêm vào Stream Window.

Các nút này có thể được sắp xếp theo danh mục hoặc tìm kiếm theo từ khóa.



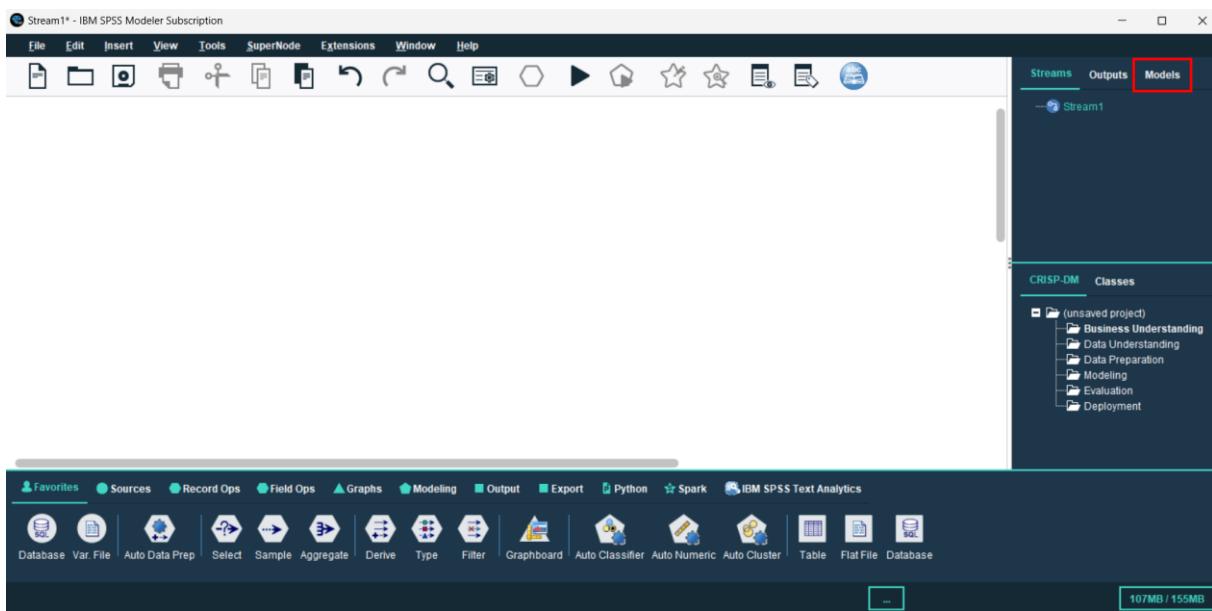
Hình 2. 39. Node Palette.

**Output Window:** Cửa sổ này hiển thị các kết quả đầu ra được tạo bởi các nút trong Stream Window. Kết quả này bao gồm các bảng, biểu đồ, và các thông tin khác về các quá trình và kết quả của mô hình.



*Hình 2. 40. Output Window.*

**Model Viewer:** Cửa sổ này hiển thị một trình xem đồ họa cho mô hình được tạo bởi các nút trong Stream Window. Trình xem này cung cấp một cách tiện lợi để xem các thuộc tính của mô hình và hiểu các mối quan hệ giữa các biến.



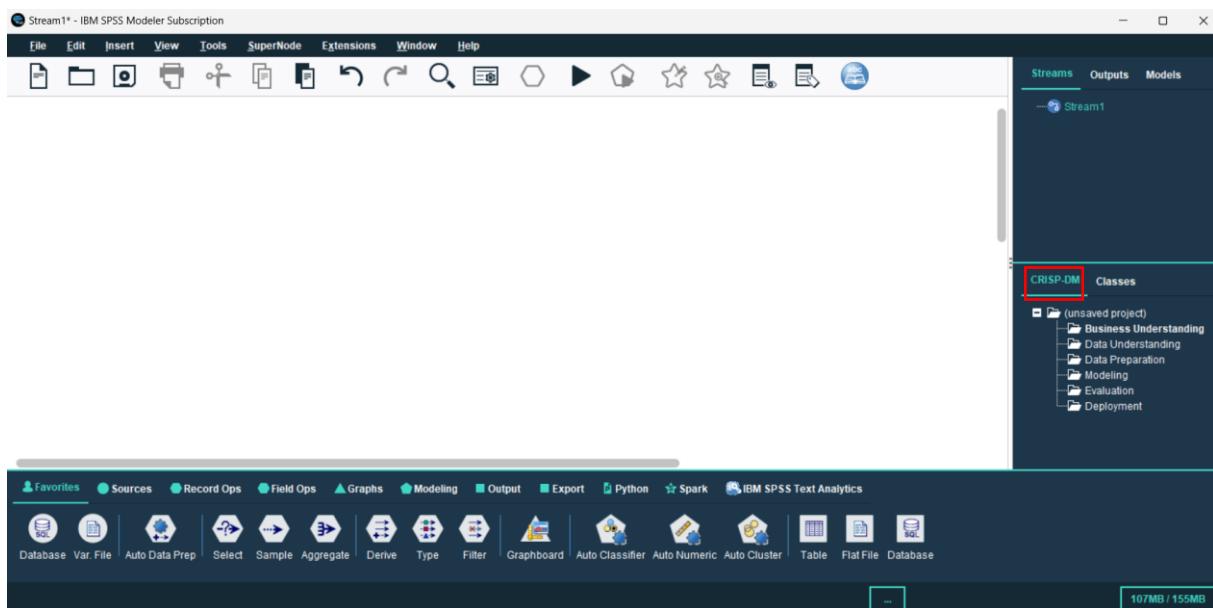
*Hình 2. 41. Models Viewer.*

**CRISP-DM (Cross Industry Standard Process for Data Mining)** là một chuẩn mực quản lý dự án cho các hoạt động khai thác dữ liệu. IBM SPSS Modeler được thiết kế để hỗ trợ và thực thi chuẩn mực CRISP-DM.

**CRISP-DM** cung cấp một khung làm việc cho các hoạt động khai thác dữ liệu, bao gồm các giai đoạn phân tích, thiết kế, triển khai và theo dõi. Nó tập trung vào các giai đoạn chủ yếu của quá trình khai thác dữ liệu, bao gồm:

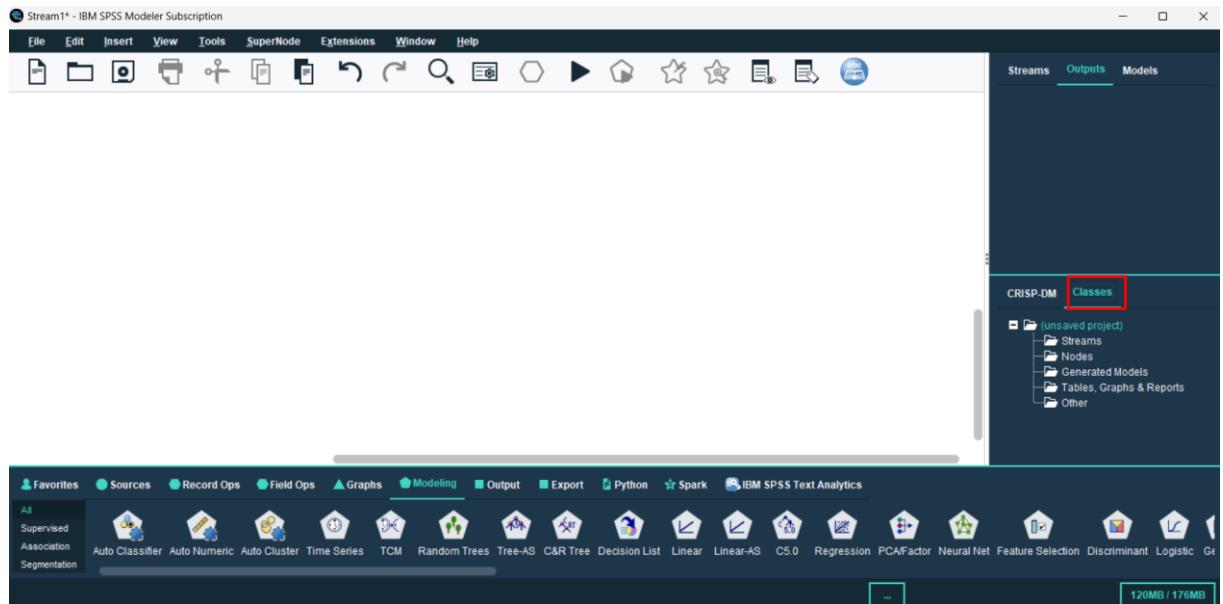
1. **Understanding:** Hiểu rõ yêu cầu và mục tiêu của khách hàng.
2. **Data Preparation:** Chuẩn bị dữ liệu để sử dụng cho việc khai thác dữ liệu.
3. **Modeling:** Xây dựng mô hình dựa trên dữ liệu đã được chuẩn bị.
4. **Evaluation:** Đánh giá mô hình và đảm bảo rằng nó đáp ứng các yêu cầu và mục tiêu của khách hàng.
5. **Deployment:** Đưa mô hình vào sử dụng trong môi trường sản xuất.
6. **Monitoring:** Theo dõi và duy trì mô hình trong suốt quá trình sử dụng.

**IBM SPSS Modeler** cung cấp các công cụ và chức năng để hỗ trợ các hoạt động trong CRISP-DM, bao gồm các công cụ để khám phá dữ liệu, xử lý dữ liệu, xây dựng mô hình, đánh giá mô hình, triển khai mô hình và theo dõi hiệu suất của mô hình. Sử dụng CRISP-DM trong IBM SPSS Modeler giúp đảm bảo rằng các dự án khai thác dữ liệu được thực hiện đúng cách và đáp ứng các yêu cầu của khách hàng.



Hình 2. 42. CRISP- DM.

**Classes:** Đơn giản là nơi lưu trữ các nút, model, các bảng, biểu đồ, báo cáo trong quá trình thao tác trên phần mềm.



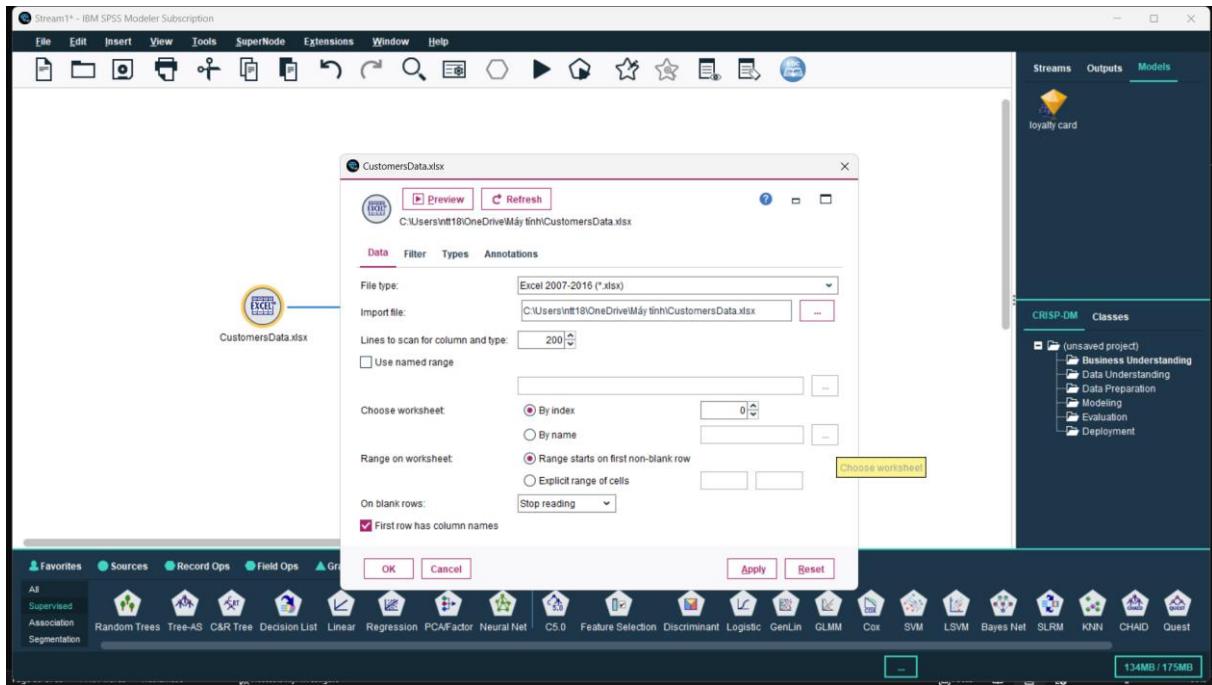
Hình 2. 43. Classes

#### 2.4.2.2. Cách thức tiến hành các thuật toán.

Ở đồ án lần này, ta sẽ chỉ tập trung vào ba thuật toán chính:

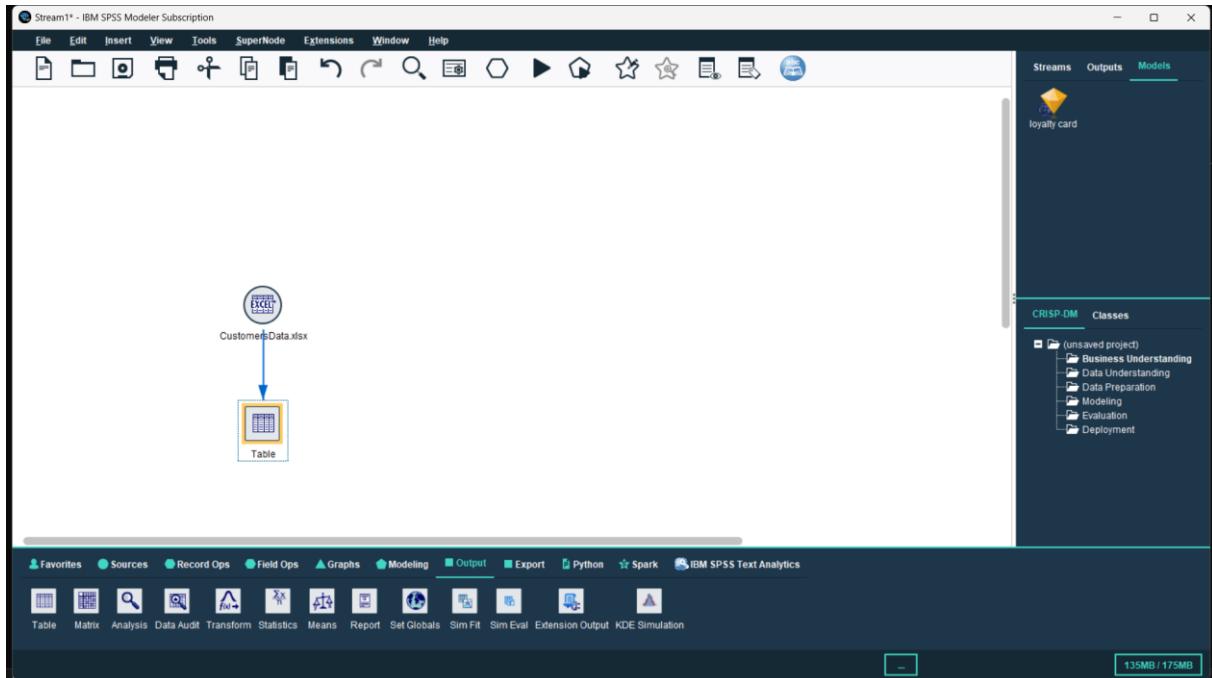
- ❖ *Phân lớp – Cây ra quyết định.*

**Bước 1:** Tải bộ dữ liệu lên phần mềm. Ví dụ ta có dữ liệu từ Excel: ở phần Node Pallete, vào phần Sources → Excel. Chuột phải hoặc double click vào Node Excel để nạp dữ liệu. Sau khi đã chọn dữ liệu mong muốn ta chọn Apply → OK.



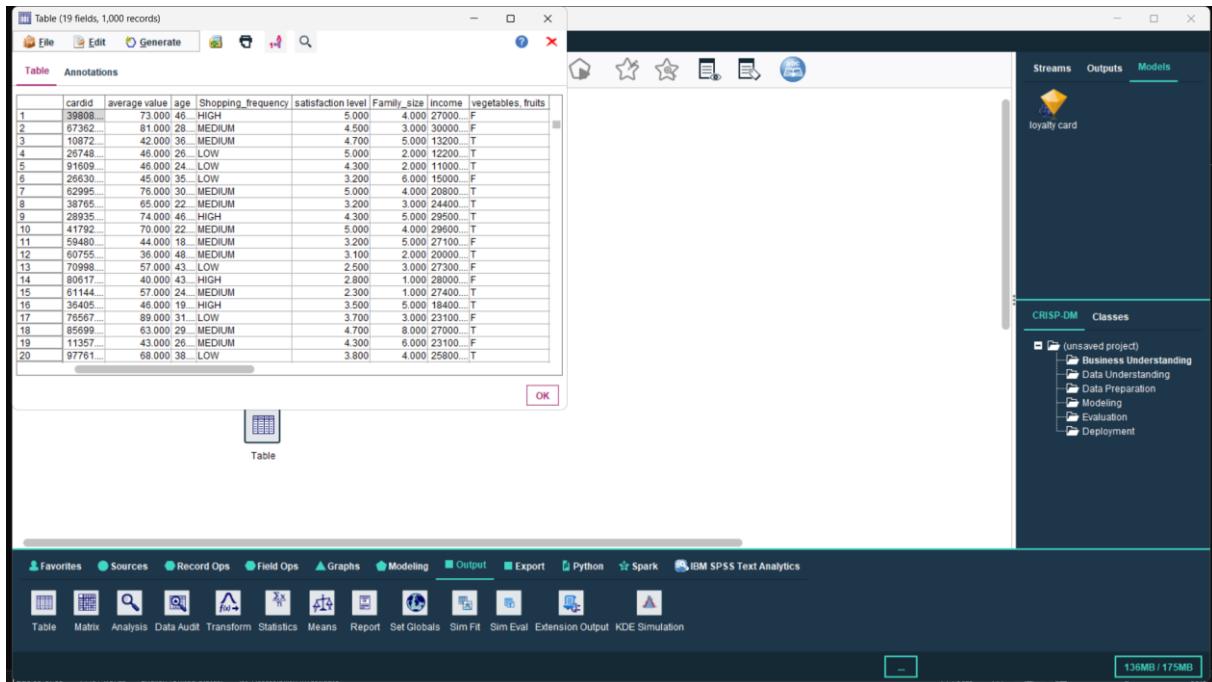
**Hình 2. 44.** Tải bộ dữ liệu lên phần mềm.

**Bước 2:** Vào mục Output ở Node Pallete, chọn Table. Sau đó ở mục Excel ta chuột phải → chọn Connect và chọn Table để kết nối 2 Node lại với nhau.



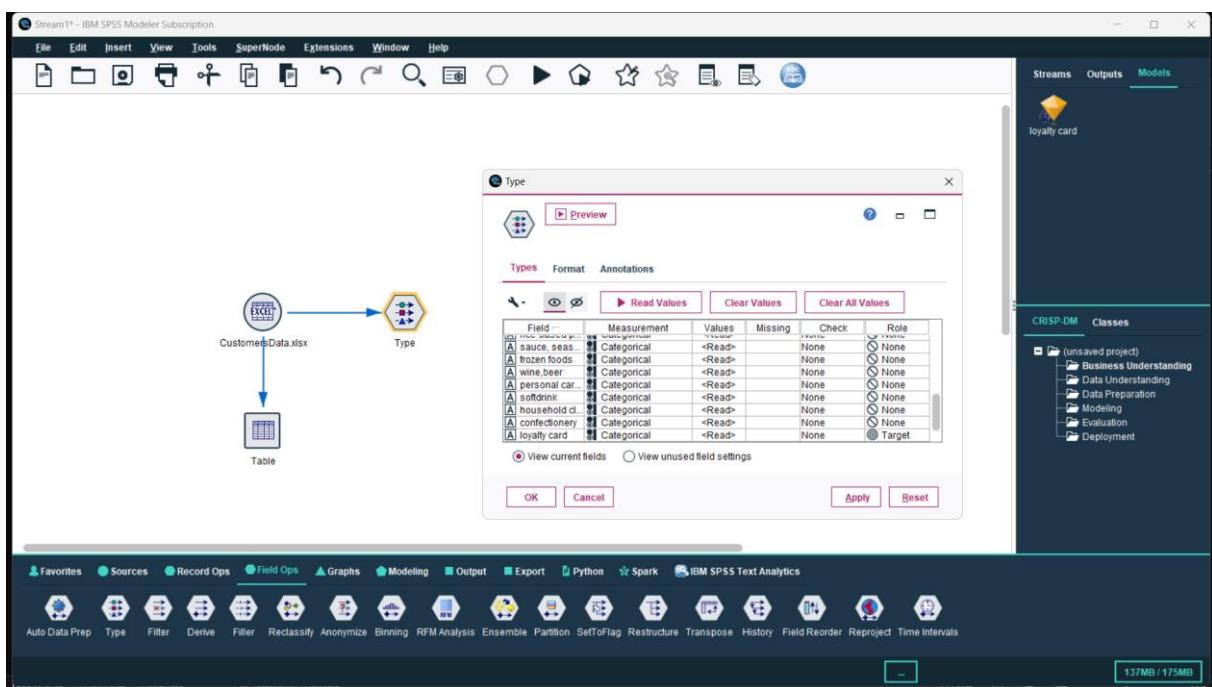
**Hình 2. 45.** Kết nối bộ dữ liệu với Node Table.

**Bước 3:** Double Click vào Table, chọn Apply → Run, quá trình này giúp bạn có thể xem được bộ dữ liệu của mình một cách thuận tiện.



**Hình 2. 46.** Xem lại dữ liệu trong Node Table.

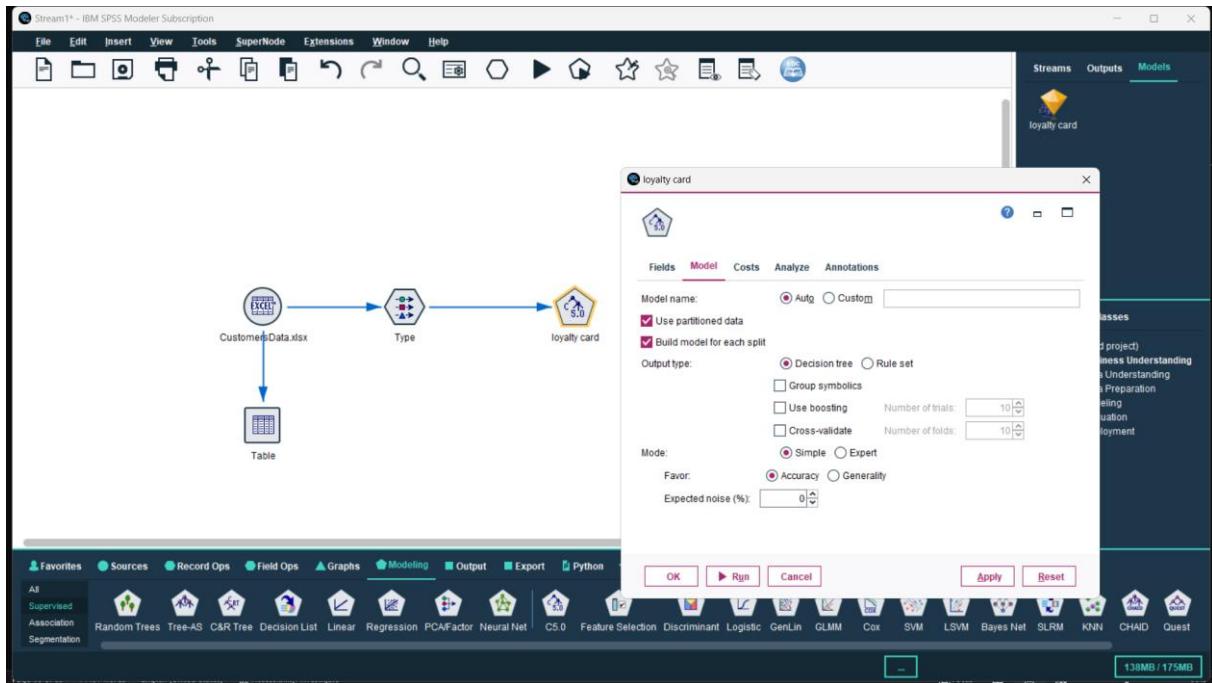
**Bước 4:** Tiếp tục ở phần Node Pallete, vào mục Field Ops chọn Type. Sau đó kết nối nó với Excel như ở bước 1. Quá trình này cho phép bạn chọn một hay nhiều biến trong dữ liệu để sử dụng trong quá trình phân tích.



**Hình 2. 47.** Chọn các biến dùng để phân tích.

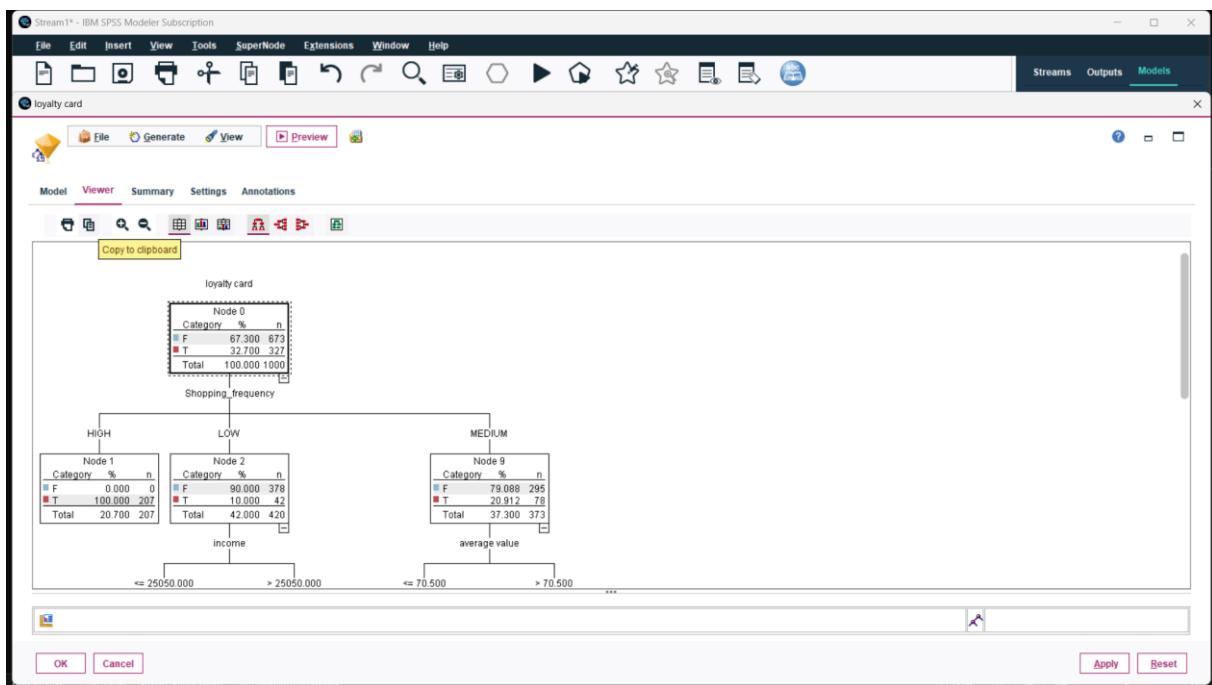
**Bước 5:** Vào phần Modeling chọn C5.0 và kết nối với Type như các bước ở trên. Chọn Use custom field assignments. Sau đó, chọn một biến để làm Mục

tiêu phân tích, và chọn những biến để Dự đoán. Chọn Apply → Run để khởi tạo cây ra quyết định.



Hình 2. 48. Khởi tạo cây ra quyết định.

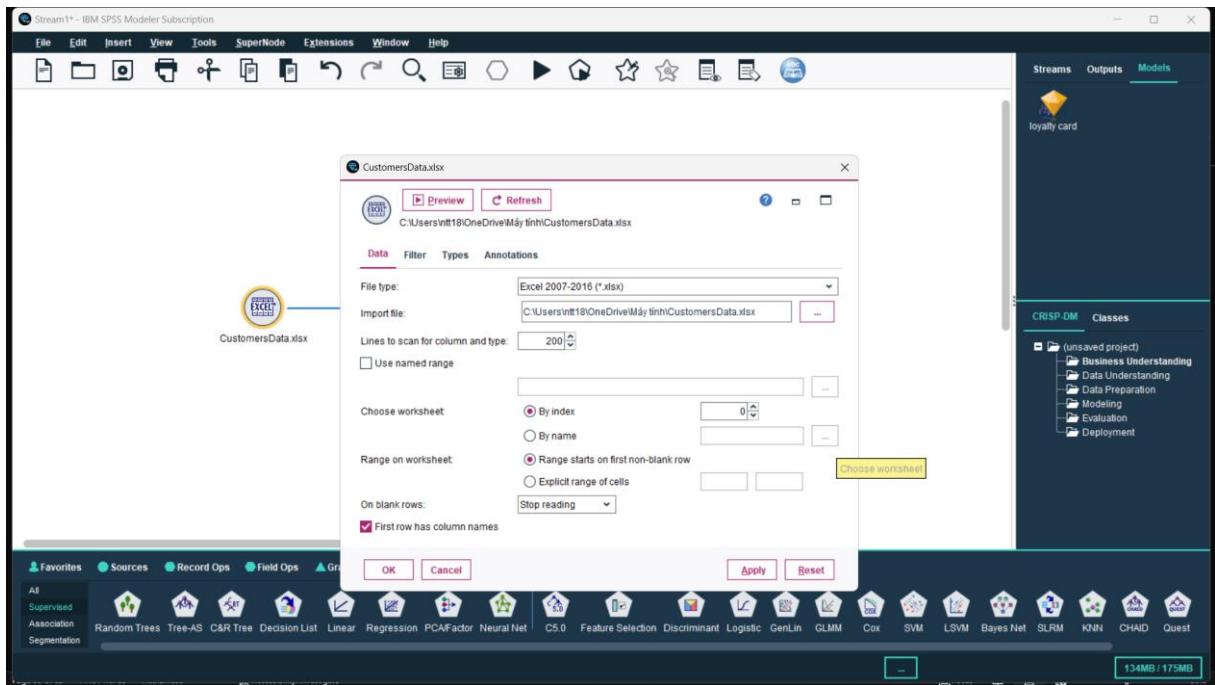
**Bước 6:** Sau khi khởi tạo thành công trên màn hình Stream sẽ xuất hiện một khói đa giác màu vàng. Double Click vào để xem kết quả vừa khởi tạo.



Hình 2. 49. Cây ra quyết định.

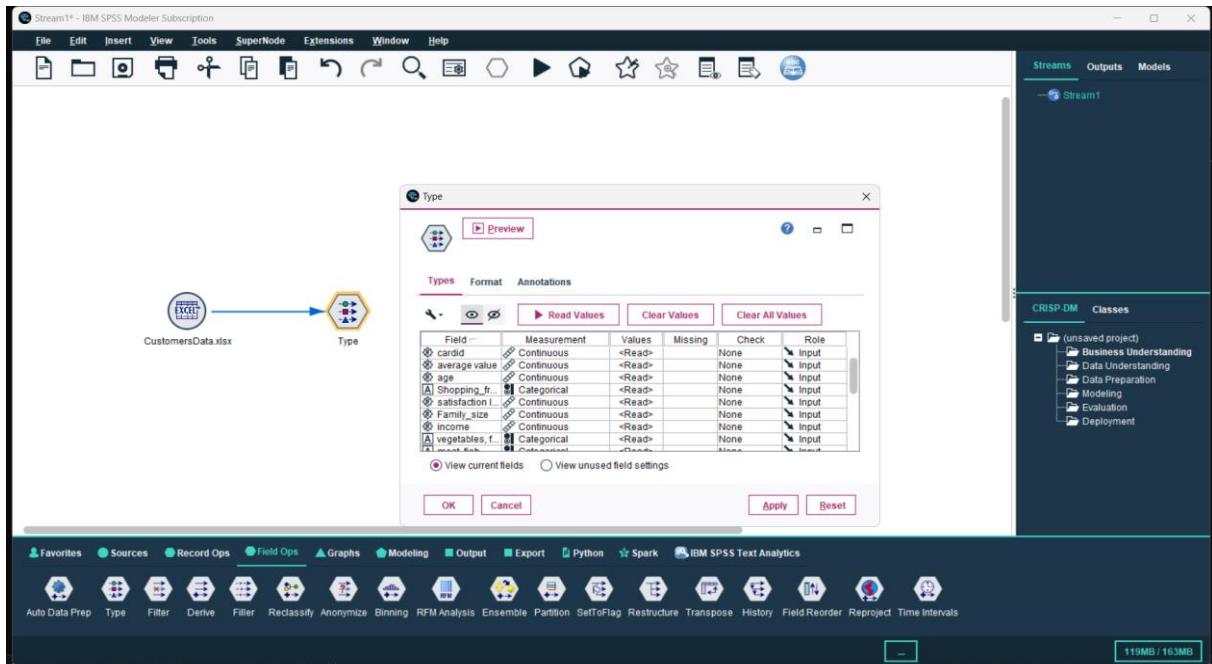
❖ Gom cụm – K-Means.

**Bước 1:** Tải bộ dữ liệu lên phần mềm. Ví dụ ta có dữ liệu từ Excel: ở phần Node Pallete, vào phần Sources → Excel. Chuột phải hoặc double click vào Node Excel để nạp dữ liệu. Sau khi đã chọn dữ liệu mong muốn ta chọn Apply → OK.



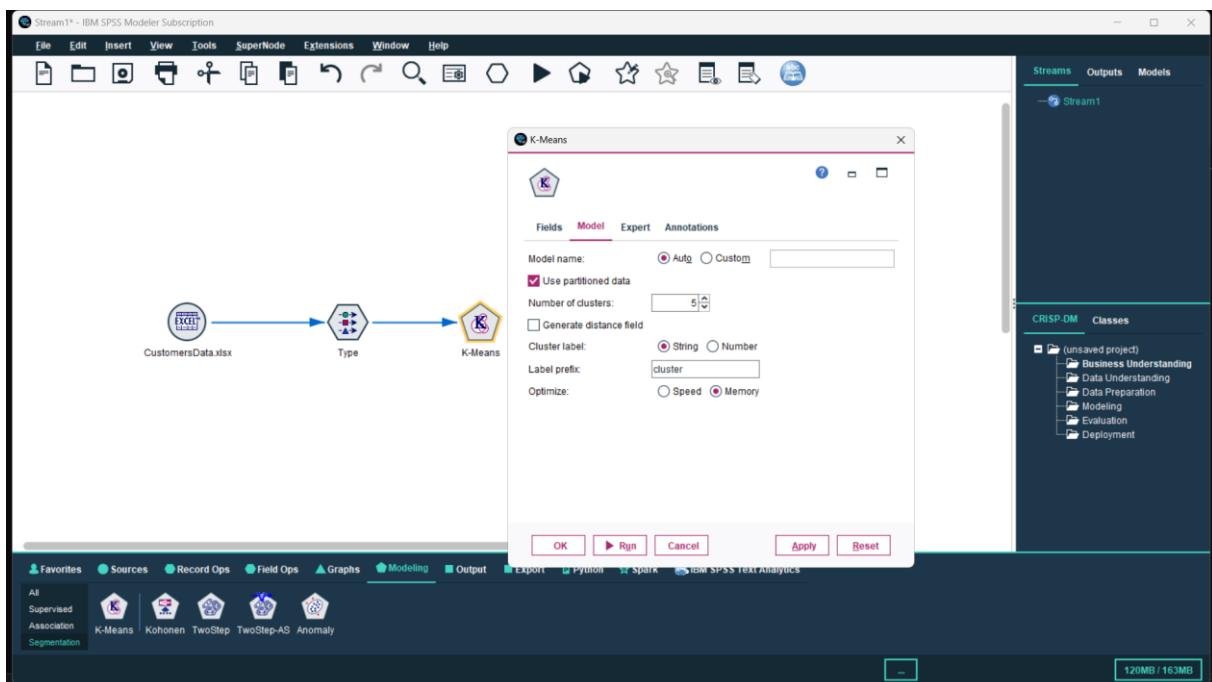
**Hình 2. 50.** Tải dữ liệu lên phần mềm.

**Bước 2:** Tiếp tục ở phần Node Pallete, vào mục Field Ops chọn Type. Sau đó kết nối nó với Excel như ở bước 1. Quá trình này cho phép bạn chọn một hay nhiều biến trong dữ liệu để sử dụng trong quá trình phân tích.



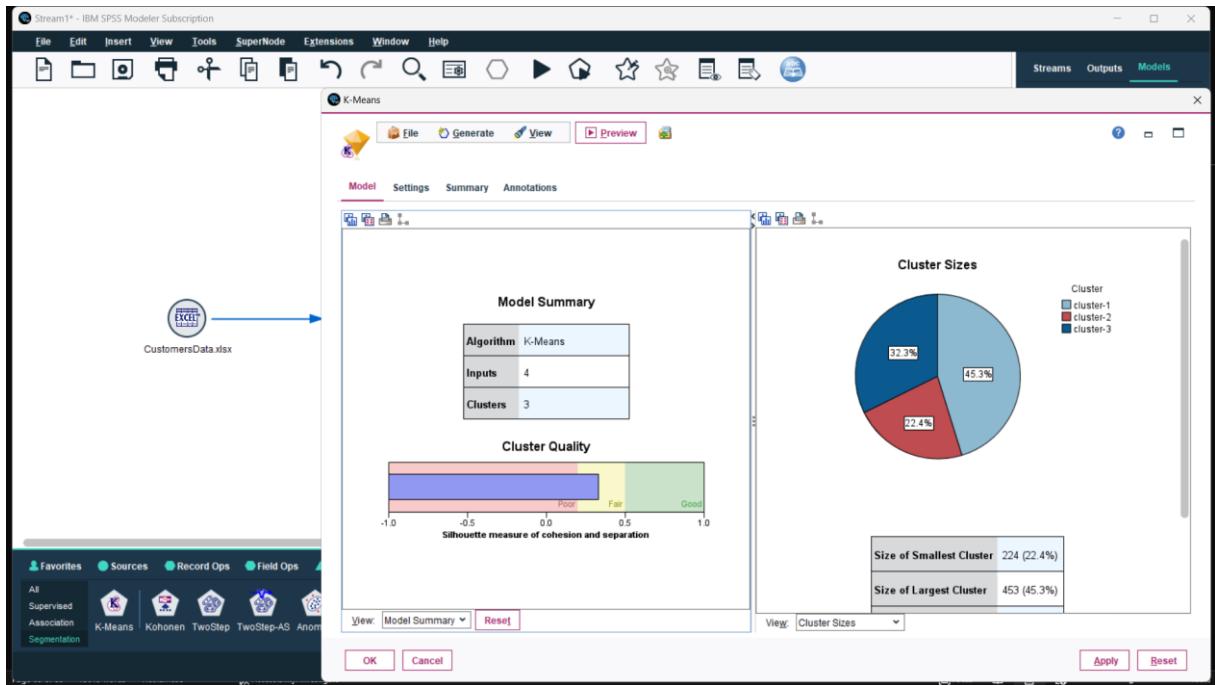
Hình 2. 51. Chọn các biến dùng để phân tích.

**Bước 3:** Vào phần Modeling chọn K-Means và kết nối với Type như các bước ở trên. Sau đó, chọn số cụm muốn tạo. Chọn Apply → Run để khởi tạo thuật toán K-Means.



Hình 2. 52. Chọn số cụm muốn khởi tạo.

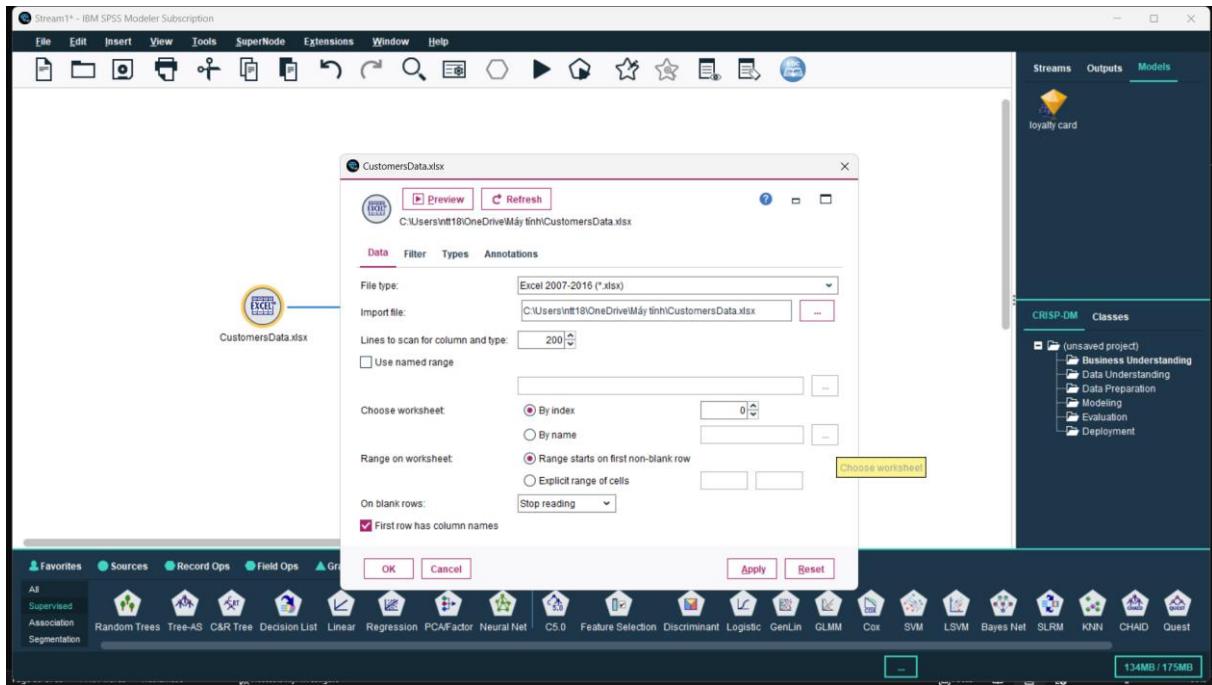
**Bước 4:** Sau khi khởi tạo thành công trên màn hình Stream sẽ xuất hiện một khôi đa giác màu vàng. Double Click vào để xem kết quả vừa khởi tạo.



**Hình 2. 53.** Bảng thẻ hiện các kết quả của thuật toán K-Means

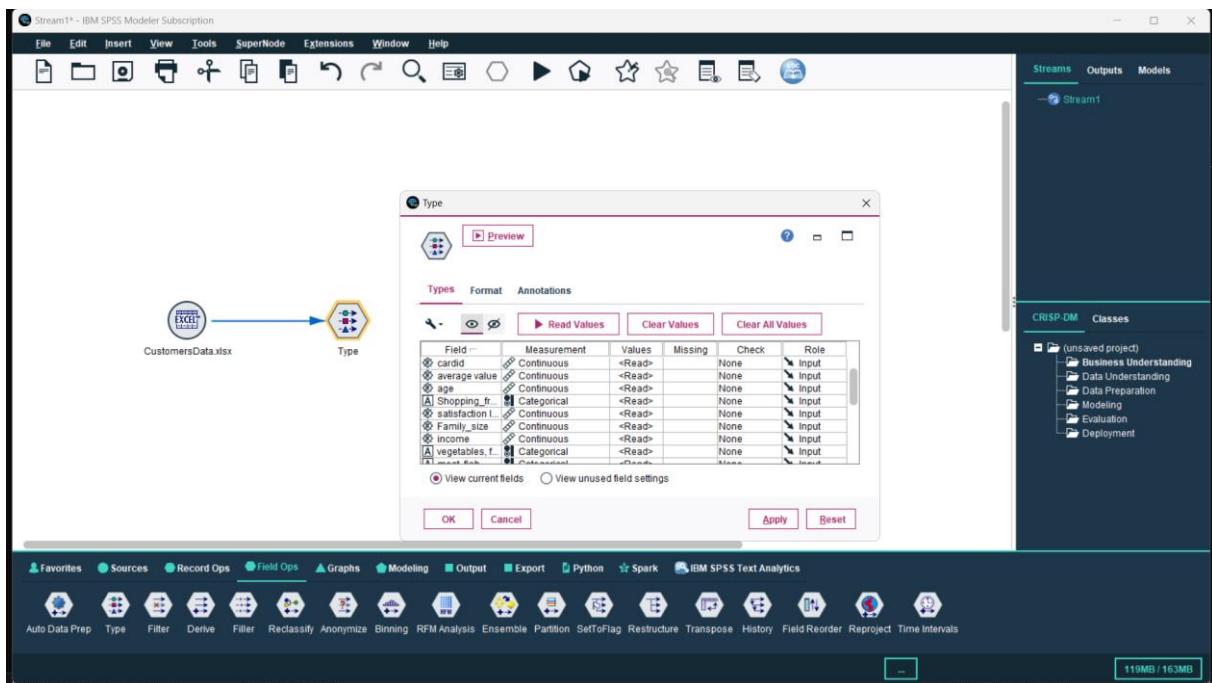
❖ **Kết hợp – Apriori.**

**Bước 1:** Tải bộ dữ liệu lên phần mềm. Ví dụ ta có dữ liệu từ Excel: ở phần Node Pallete, vào phần Sources → Excel. Chuột phải hoặc double click vào Node Excel để nạp dữ liệu. Sau khi đã chọn dữ liệu mong muốn ta chọn Apply → OK.



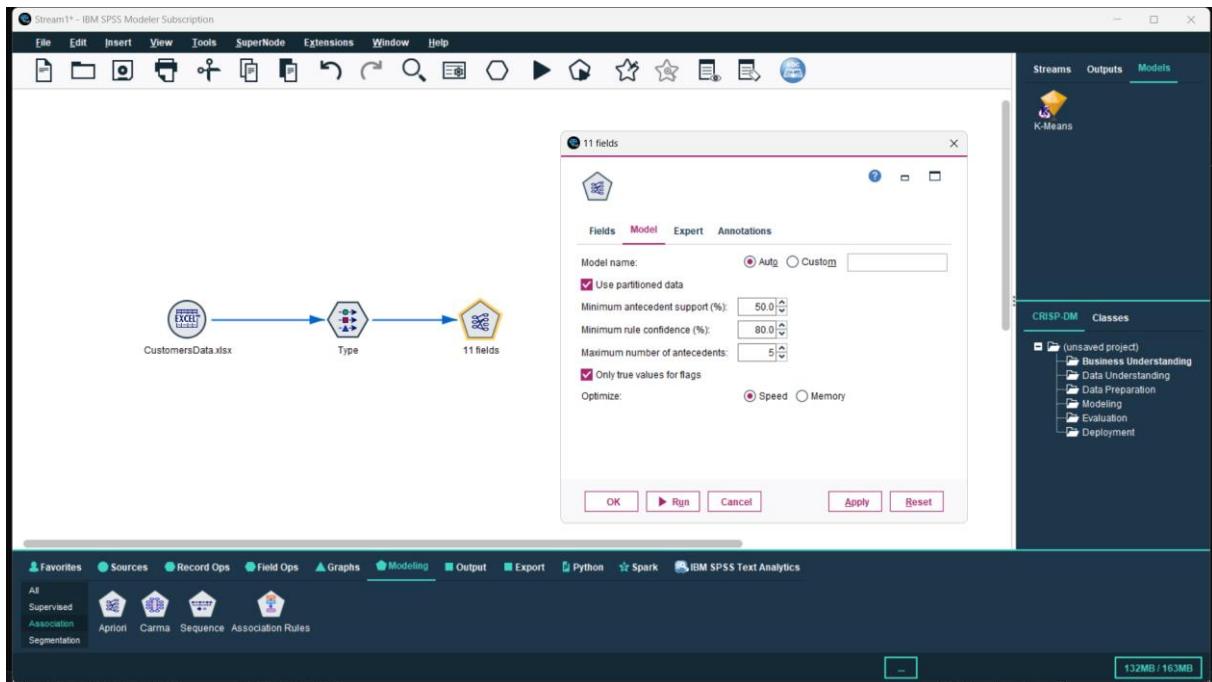
**Hình 2. 54.** Tải bộ dữ liệu lên phần mềm.

**Bước 2:** Tiếp tục ở phần Node Pallette, vào mục Field Ops chọn Type. Sau đó, kết nối nó với Excel như ở bước 1. Quá trình này cho phép bạn chọn một hay nhiều biến trong dữ liệu để sử dụng trong quá trình phân tích.



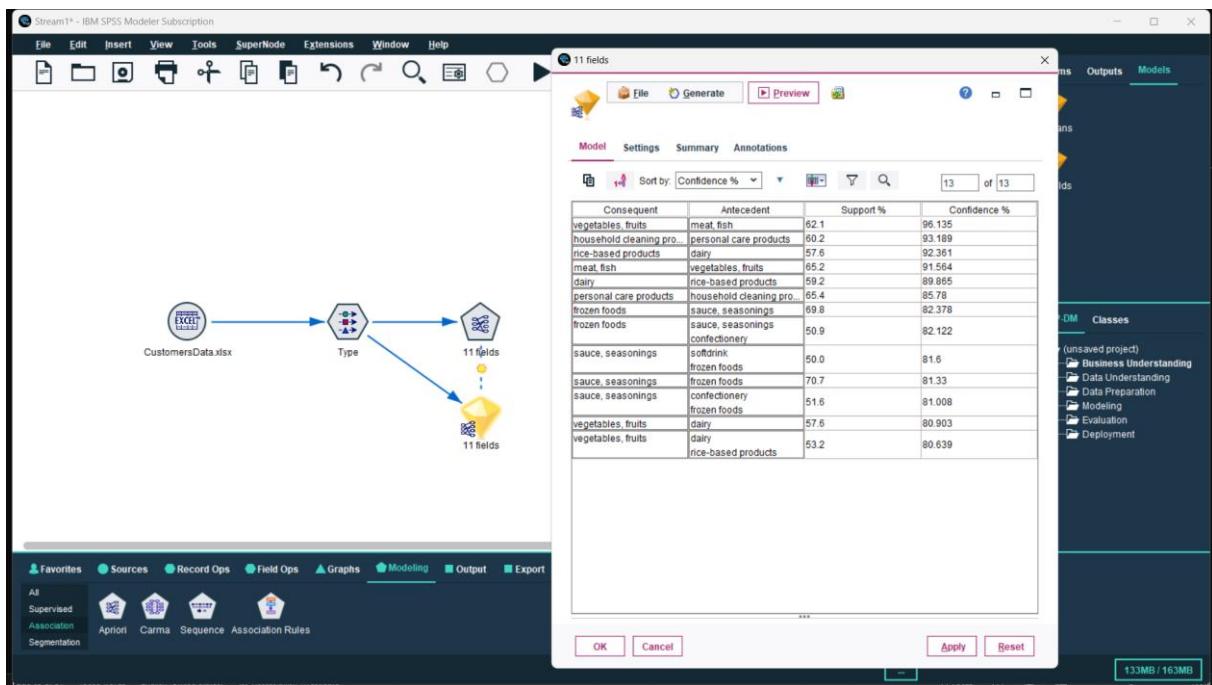
**Hình 2. 55.** Chọn các biến dùng để phân tích.

**Bước 3:** Vào phần Modeling chọn Apriori và kết nối với Type như các bước ở trên. Sau đó, chọn độ hỗ trợ và độ tin cậy mong muốn. Chọn Apply → Run để khởi tạo thuật toán Apriori.



**Hình 2. 56.** Xét độ hỗ trợ và độ tin cậy.

**Bước 4:** Sau khi khởi tạo thành công trên màn hình Stream sẽ xuất hiện một khói đa giác màu vàng. Double Click vào để xem kết quả vừa khởi tạo.



Hình 2. 57. Bảng biểu diễn kết quả của thuật toán Apriori.

## CHƯƠNG 3. ỨNG DỤNG PHẦN MỀM IBM SPSS MODELER

### 3.1. Thông tin về dữ liệu.

Bài toán và bộ dữ liệu của nhóm chúng em nói về việc phân tích và đánh giá chất lượng của tập dữ liệu về khách hàng trong lĩnh vực bán lẻ. Trong tập dữ liệu này, chúng em có các thuộc tính về thông tin cá nhân của khách hàng như tuổi, số lượng mua sắm, mức độ hài lòng, kích cỡ gia đình, thu nhập cũng như thông tin về các sản phẩm mà khách hàng thường xuyên mua sắm, bao gồm các loại rau củ, trái cây, thịt, cá, sữa, các sản phẩm từ gạo, gia vị, đồ đông lạnh, rượu bia, sản phẩm chăm sóc cá nhân, nước giải khát, sản phẩm làm sạch nhà cửa, sản phẩm kẹo cao su và cuối cùng là quyết định có cấp thẻ thành viên.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	cardid	average va	age	Shopping_satisfactor	Family_size	income	vegetables	meat,fish	dairy	rice-based	sauce,	frozen	wine,beer	personal c:	softdrink	household	confection	loyalty card
2	39808	73	46	HIGH	5	4	27000	F	T	F	T	T	T	T	F	F	F	T
3	67362	81	28	MEDIUM	4.5	3	30000	F	T	F	T	T	T	T	F	F	F	T
4	10872	42	36	MEDIUM	4.7	5	13200	T	F	F	F	F	T	T	F	F	T	F
5	26748	46	26	LOW	5	2	12200	T	F	F	T	T	T	F	T	F	T	F
6	91609	46	24	LOW	4.3	2	11000	T	F	T	T	T	T	F	T	T	T	F
7	26630	45	35	LOW	3.2	6	15000	F	T	F	T	T	T	F	F	T	T	F
8	62995	76	30	MEDIUM	5	4	20800	T	F	T	T	T	T	T	F	T	T	T
9	38765	65	22	MEDIUM	3.2	3	24400	T	F	T	T	T	F	T	T	T	T	F
10	28935	74	46	HIGH	4.3	5	29500	T	T	T	F	T	T	T	F	T	T	T
11	41792	70	22	MEDIUM	5	4	29600	T	F	F	F	T	T	T	F	F	T	T
12	59480	44	18	MEDIUM	3.2	5	27100	F	F	F	T	T	F	F	F	F	T	F
13	60755	36	48	MEDIUM	3.1	2	20000	T	F	T	T	T	T	F	F	F	T	F
14	70998	57	43	LOW	2.5	3	27300	F	F	F	F	T	T	F	F	F	T	F
15	80617	40	43	HIGH	2.8	1	28000	F	T	F	F	T	T	T	F	T	T	T
16	61144	57	24	MEDIUM	2.3	1	27400	T	T	T	T	T	T	F	F	F	T	F
17	36405	46	19	HIGH	3.5	5	18400	T	T	T	F	F	T	F	T	T	T	T
18	76567	89	31	LOW	3.7	3	23100	F	T	F	F	T	T	F	F	F	T	F
19	85699	63	29	MEDIUM	4.7	8	27000	T	F	T	T	T	T	F	F	F	T	F
20	11357	43	26	MEDIUM	4.3	6	23100	F	F	F	T	F	T	T	F	T	T	F

**Hình 3. 1. Dữ liệu dùng cho đồ án.**

Bộ dữ liệu bao gồm có 19 cột và 1000 dòng.

**Bảng 3-1 Bảng thể hiện dữ liệu chi tiết.**

STT	Tên dữ liệu	Tên tiếng Việt	Tập giá trị
1	cardid	Mã số khách hàng	6 chữ số
2	average value	Giá trị trung bình của mỗi lần mua hàng	2-3 chữ số
3	age	Tuổi của khách hàng	15-65

4	shopping_frequency	Tần suất mua sắm	HIGH – MEDIUM-LOW
5	satisfaction level	Mức độ hài lòng của khách hàng	1-5
6	family_size	Số lượng thành viên trong gia đình	1-7
7	income	Thu nhập của khách hàng	10000-50000
8	vegetables, fruits	Khách hàng có mua rau, hoa quả không	True (T) or False (F)
9	meat, fish	Khách hàng có mua thịt, cá không	True (T) or False (F)
10	dairy	Khách hàng có mua sản phẩm sữa không	True (T) or False (F)
11	rice-based products	Khách hàng có mua các sản phẩm từ gạo không	True (T) or False (F)
12	sauce, seasonings	Khách hàng có mua các loại sốt, gia vị không	True (T) or False (F)
13	frozen foods	Khách hàng có mua các sản phẩm đông lạnh không	True (T) or False (F)
14	wine, beer	Khách hàng có mua rượu, bia không	True (T) or False (F)
15	personal care products	Khách hàng có mua các sản phẩm chăm sóc cá nhân không	True (T) or False (F)

16	soft drink	Khách hàng có mua các sản phẩm nước giải khát không	True (T) or False (F)
17	household cleaning products,	Khách hàng có mua các sản phẩm dụng cụ vệ sinh gia đình không	True (T) or False (F)
18	confectionery	Khách hàng có mua các sản phẩm kẹo, đồ ngọt không	True (T) or False (F)
19	loyalty card	Liệu có cấp thẻ thành viên hay không	True (T) or False (F)

### 3.2. Tiền xử lý dữ liệu.

Bộ dữ liệu về hành vi mua hàng của khách hàng trong siêu thị cần được tiền xử lý để chuẩn bị cho việc phân tích và xây dựng mô hình.

Đầu tiên, chúng em kiểm tra và xử lý các giá trị thiếu sót hoặc không chính xác trong bộ dữ liệu. Ở đây chúng em sẽ kiểm dò lại một lược để xem bộ dữ liệu của mình có thiếu sót thông tin gì không, từ đó chúng em sẽ thêm vào hoặc chỉnh sửa để dữ liệu có thể chạy tốt các luật mà cô yêu cầu

Sau đó, chúng em chuẩn hóa các giá trị thuộc tính để các giá trị này có thể so sánh được với nhau. Bọn em chuyển giá trị tần suất của hàng của khách hàng từ dữ liệu số sang dữ liệu **Cao, Trung bình, Thấp** để dễ sử dụng vào thuật toán Phân lớp – Cây ra quyết định. Ngoài ra chúng em còn làm tròn thuật tính average value và income để dễ dàng cho việc áp dụng vào các thuật toán, giúp bài làm của bọn em trở nên trực quan hơn.

Tiếp theo, chúng em sẽ thêm/bớt một số đặc trưng mới từ các thuộc tính hiện có để giúp mô hình dự đoán tốt hơn. Ví dụ như là chúng em đã tạo thêm thuộc tính loyalty card để sử dụng có thuật toán Phân lớp – Cây ra quyết định. Bỏ đi một số thuộc tính như giới tính, phương thức thanh toán, có sở hữu nhà hay không và sửa lại một số thuộc tính khác.

Cuối cùng, chúng em sẽ lựa chọn các thuộc tính quan trọng nhất để sử dụng trong mô hình và loại bỏ các thuộc tính không cần thiết.

Qua đó, việc tiền xử lý dữ liệu sẽ giúp cho việc phân tích và xây dựng mô hình trở nên chính xác hơn và đưa ra các kết quả có giá trị hơn để giải quyết các vấn đề liên quan đến hành vi mua hàng của khách hàng.

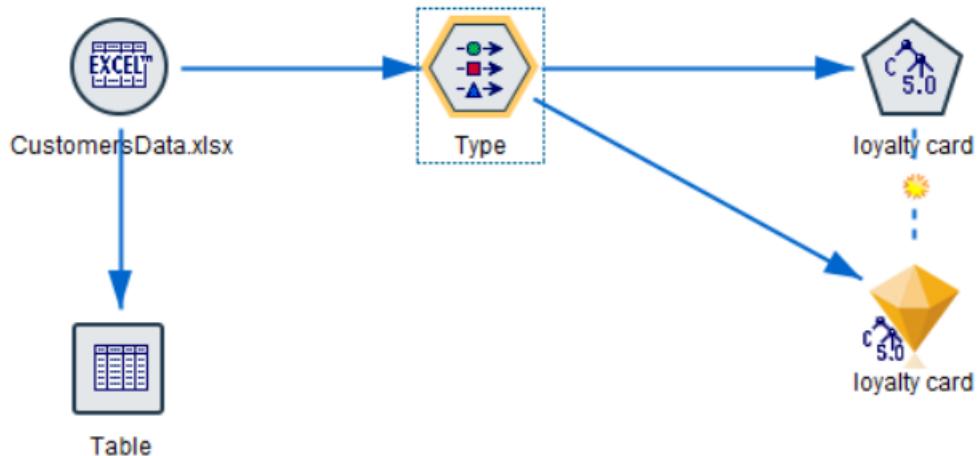
### 3.3. Kết luận của mỗi thuật toán.

#### 3.3.1. Thuật toán phân lớp.

Để thực hiện thuật toán phân lớp thì chúng em dùng 7 thuộc tính đó là shopping\_frequency, income, average value, family\_size, age, satisfaction level, loyalty card trong đó thuộc tính loyalty card làm thuộc tính quyết định.

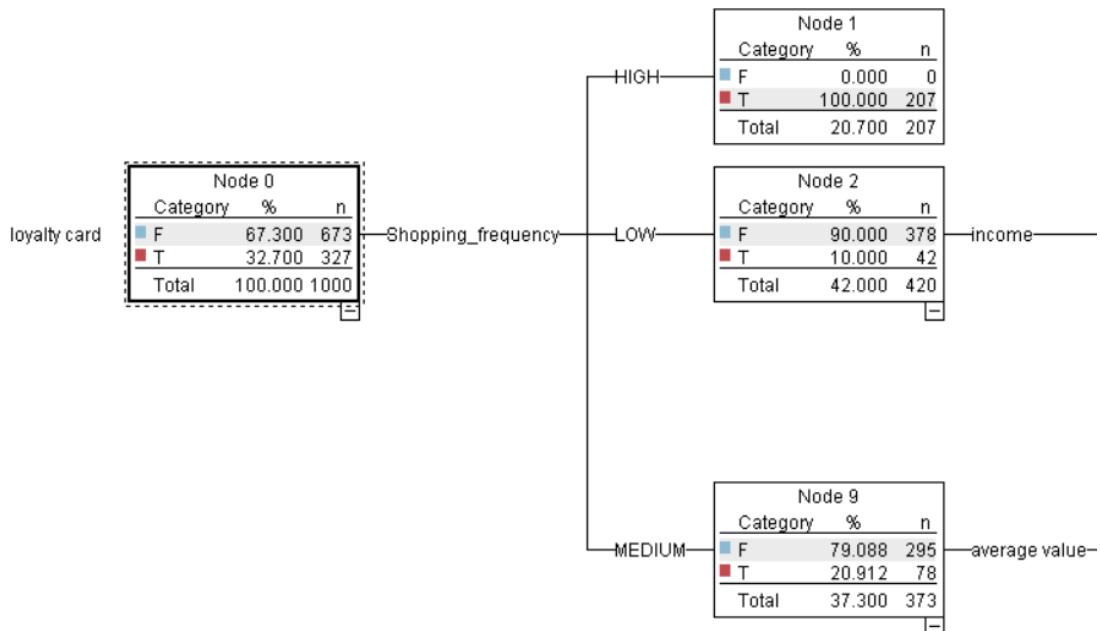
Field	Measurement	Values	Missing	Check	Role
cardid	Continuous	[10150.0,1...	None	<input type="radio"/> None	<input checked="" type="radio"/> Input
average value	Continuous	[20.0,90.0]	None	<input checked="" type="radio"/> None	<input checked="" type="radio"/> Input
age	Continuous	[16.0,50.0]	None	<input checked="" type="radio"/> None	<input checked="" type="radio"/> Input
Shopping_fr...	Nominal	HIGH,LO...	None	<input checked="" type="radio"/> None	<input checked="" type="radio"/> Input
satisfaction l...	Continuous	[2.3,5.0]	None	<input checked="" type="radio"/> None	<input checked="" type="radio"/> Input
Family_size	Continuous	[1.0,8.0]	None	<input checked="" type="radio"/> None	<input checked="" type="radio"/> Input
income	Continuous	[10200.0,3...	None	<input checked="" type="radio"/> None	<input checked="" type="radio"/> Input
vegetables, fr...	Flag	T/F	None	<input type="radio"/> None	<input type="radio"/> None
meat, fish	Flag	T/F	None	<input type="radio"/> None	<input type="radio"/> None
dairy	Flag	T/F	None	<input type="radio"/> None	<input type="radio"/> None
rice-based pr...	Flag	T/F	None	<input type="radio"/> None	<input type="radio"/> None
sauce, seas...	Flag	T/F	None	<input type="radio"/> None	<input type="radio"/> None
frozen foods	Flag	T/F	None	<input type="radio"/> None	<input type="radio"/> None
wine,beer	Flag	T/F	None	<input type="radio"/> None	<input type="radio"/> None
personal car...	Flag	T/F	None	<input type="radio"/> None	<input type="radio"/> None
softdrink	Flag	T/F	None	<input type="radio"/> None	<input type="radio"/> None
household cl...	Flag	T/F	None	<input type="radio"/> None	<input type="radio"/> None
confectionery	Flag	T/F	None	<input type="radio"/> None	<input type="radio"/> None
loyalty card	Flag	T/F	None	<input checked="" type="radio"/> Target	<input checked="" type="radio"/> Target

**Hình 3. 2.** Chính sửa thuộc tính cho thuật toán phân lớp.

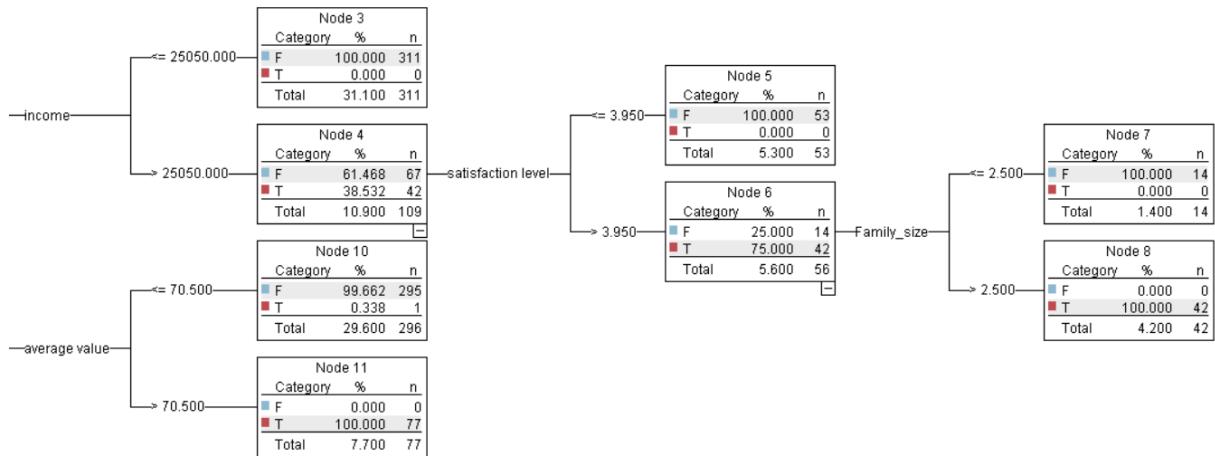


**Hình 3. 3.** Mô hình của thuật toán phân lớp.

Sau khi chạy thuật toán ta sẽ ra một cái cây:



**Hình 3. 4.** Kết quả của thuật toán phân lớp 1.



**Hình 3.5. Kết quả của thuật toán phân lớp 2.**

Nút gốc: shopping frequency

- ❖ Khi Shopping\_frequency = HIGH, loyalty card = "T"
- ❖ Khi Shopping\_frequency = MEDIUM
  - Nếu avarage value  $> 70.5$ , loyalty card = "T"
  - Nếu avarage value  $\leq 70.5$ , loyalty card = "F"
- ❖ Khi Shopping\_frequency = LOW
  - Nếu income  $\leq 25050$ , loyalty card = "F"
  - Nếu income  $> 25050$ , loyalty card
    - Nếu Satisfaction\_level  $\leq 3.95$ , loyalty card = "F"
    - Nếu Satisfaction\_level  $> 3.95$ 
      - Nếu Family\_size  $\leq 2.5$ , loyalty card = "F"
      - Nếu Family\_size  $> 2.5$ , loyalty card = "T"

Cây quyết định này được sử dụng để xác định liệu khách hàng có nên cấp thẻ thành viên hay không, dựa trên tần suất mua sắm của họ và một số thông tin khác về thu nhập, giá trị trung bình và mức độ hài lòng. Trong cây quyết định này, nút gốc là Shopping\_frequency, chỉ định tần suất mua sắm của khách hàng.

Nếu tần suất mua hàng của khách hàng là cao thì khách hàng sẽ được cấp thẻ thành viên.

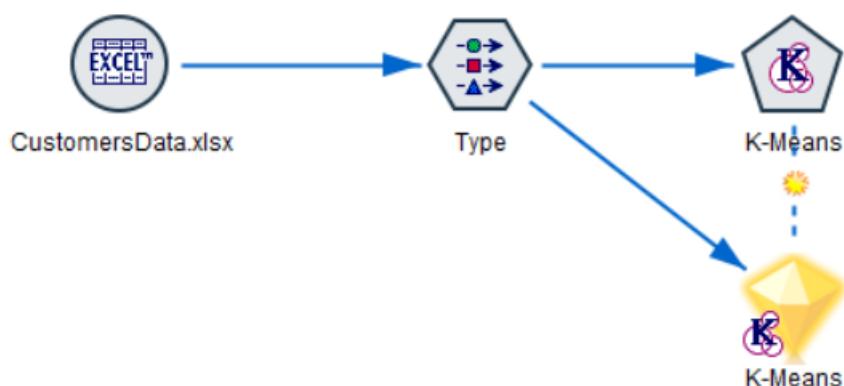
Nếu tần suất mua hàng của khách hàng là trung bình và giá trị trung bình của giỏ hàng cao hơn 70,5, khách hàng sẽ được cấp thẻ. Tuy nhiên, nếu giá trị trung bình của giỏ hàng thấp hơn hoặc bằng 70,5, khách hàng sẽ không được cấp thẻ.

Nếu tần suất mua hàng của khách hàng là thấp thì quyết định có phức tạp hơn. Nếu thu nhập của khách hàng nhỏ hơn hoặc bằng 25050, họ sẽ không được cấp thẻ. Tuy nhiên, nếu thu nhập lớn hơn 25050, quyết định sẽ phụ thuộc vào mức độ hài lòng và kích thước gia đình. Nếu độ hài lòng của khách hàng thấp hơn hoặc bằng 3,95, họ sẽ không được cấp thẻ. Nếu độ hài lòng của khách hàng lớn hơn 3,95 và kích thước gia đình nhỏ hơn hoặc bằng 2,5, khách hàng lại không được cấp tạo thẻ. Tuy nhiên, nếu kích thước gia đình lớn hơn 2,5, khách hàng sẽ được cấp thẻ thành viên.

Vì vậy, cây quyết định này có thể giúp các doanh nghiệp xác định xem khách hàng nào nên được cung cấp thẻ thành viên để tăng tối đa số lượng khách hàng trung thành của họ.

### 3.3.2. Thuật toán gom cụm.

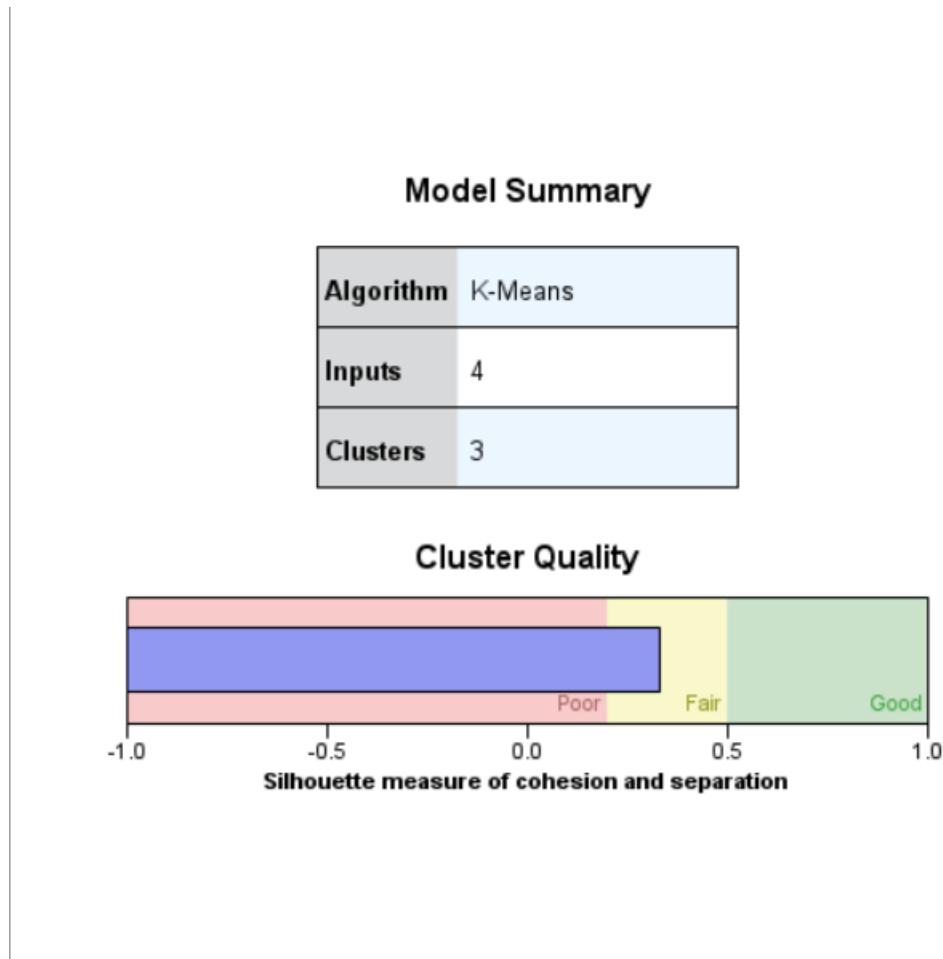
Thể hiện xu hướng mua hàng của khách hàng dựa trên 4 dữ liệu đầu vào bao gồm: thu nhập (income), tuổi (age), số thành viên trong gia đình (family\_size) và giá trị giỏ hàng (average value).



**Hình 3. 6. Mô hình của thuật toán gom cụm.**

Field	Measurement	Values	Missing	Check	Role
cardid	Continuous	[10150,0,1...]		None	None
average value	Continuous	[20.0,90.0]		None	Input
age	Continuous	[16.0,50.0]		None	Input
Shopping_fri...	Nominal	HIGH,LO...		None	None
satisfaction l...	Continuous	[2.3,5.0]		None	None
Family_size	Continuous	[1.0,8.0]		None	Input
income	Continuous	[10200,0,3...]		None	Input
A vegetables, fr...	Flag	T/F		None	None
A meat, fish	Flag	T/F		None	None
A dairy	Flag	T/F		None	None
A rice-based pr...	Flag	T/F		None	None
A sauce, seas...	Flag	T/F		None	None
A frozen foods	Flag	T/F		None	None
A wine,beer	Flag	T/F		None	None
A personal car...	Flag	T/F		None	None
A softdrink	Flag	T/F		None	None
A household cl...	Flag	T/F		None	None
A confectionery	Flag	T/F		None	None
A loyalty card	Flag	T/F		None	None

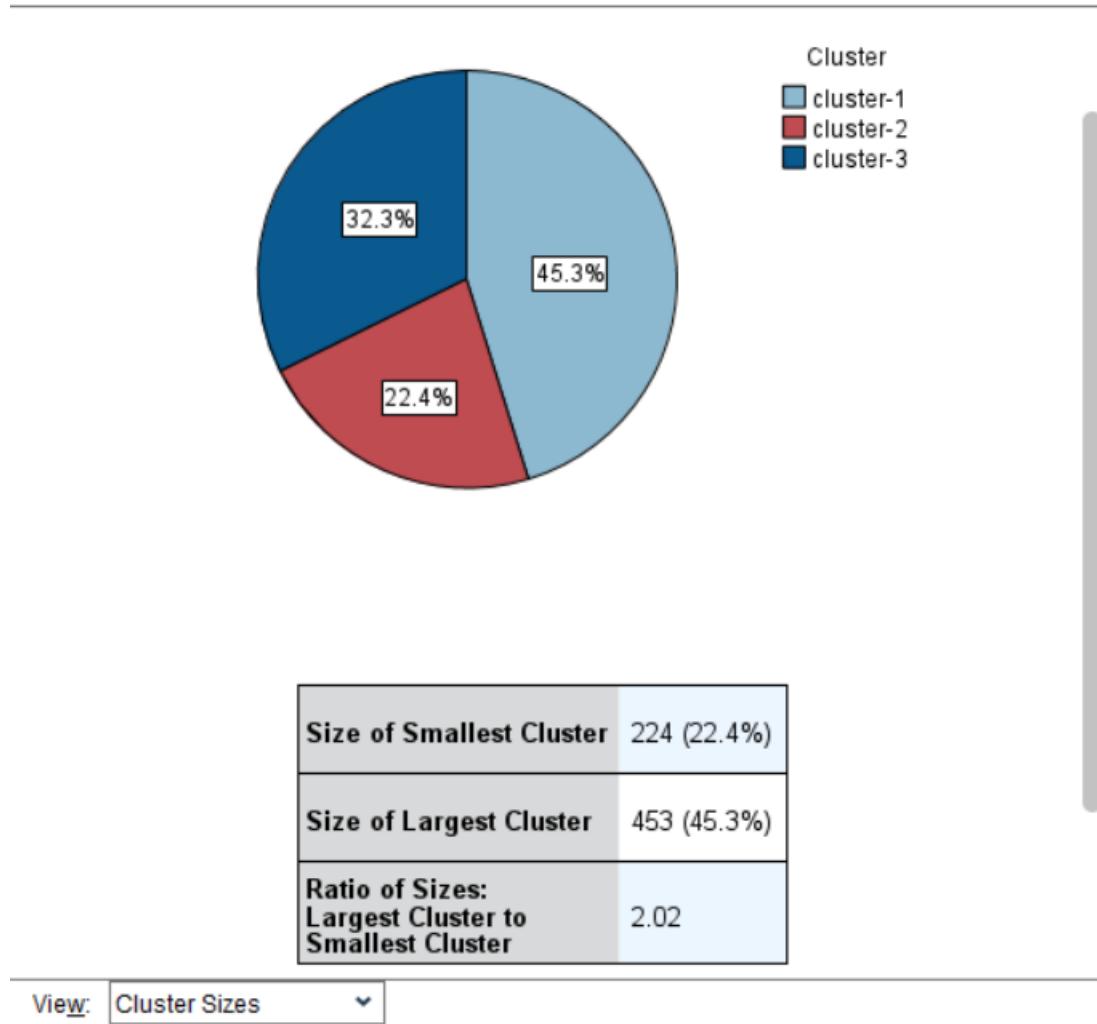
Hình 3. 7. Chính sửa thuộc tính cho thuật toán phân lớp.



Hình 3. 8. Model tổng quát của thuật toán gom cụm.

**Model Summary** cho biết thuật toán gom cụm của nhóm sử dụng là thuật toán K-Means, có số dữ liệu đầu vào là 4 và số cụm tạo ra là 3.

**Cluster Quality** thể hiện chất lượng được tạo ra bởi thuật toán phân cụm trong quá trình phân tích dữ liệu là 0.2



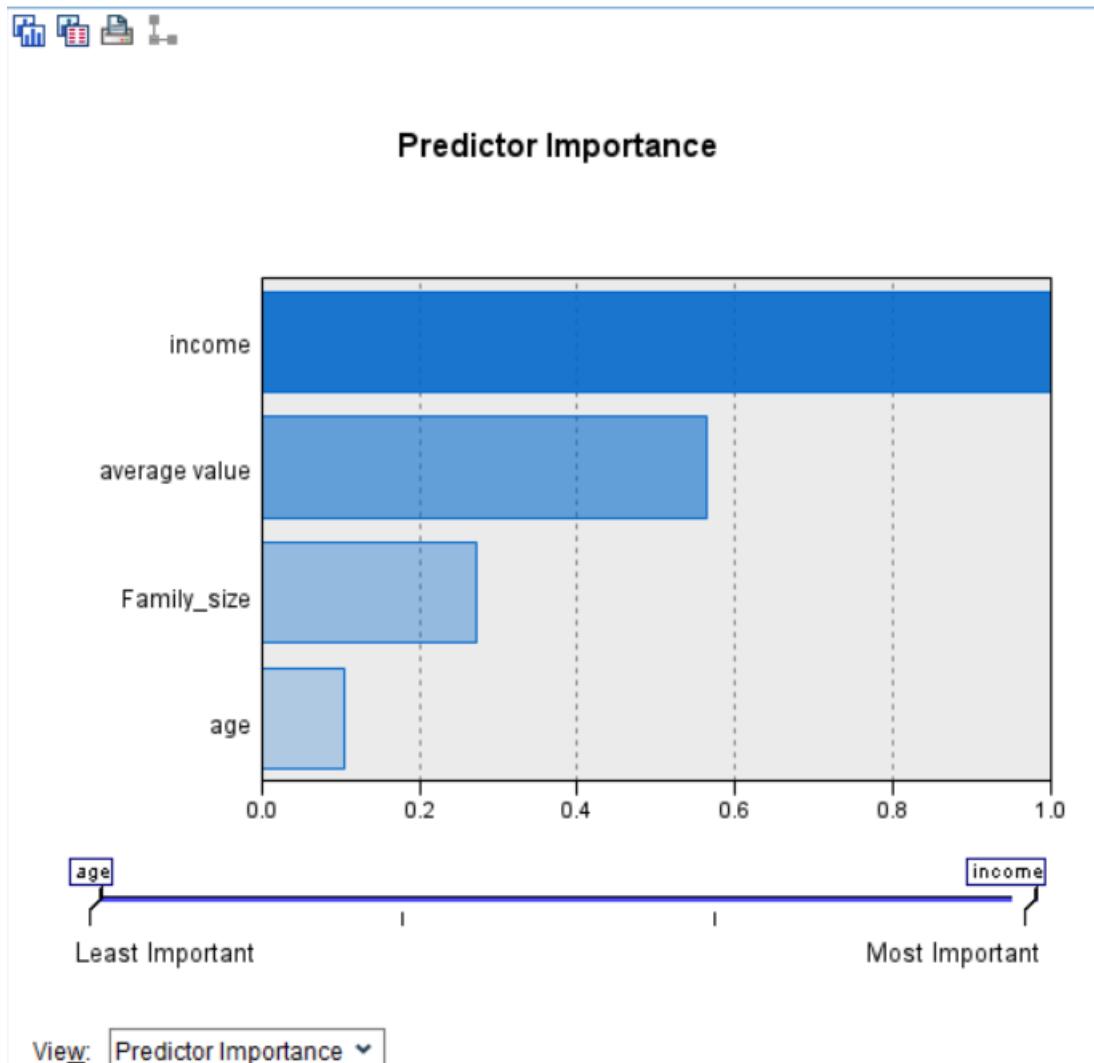
*Hình 3. 9. Cluster Sizes của thuật toán gom cụm.*

**Cluster Sizes** hiển thị một biểu đồ tròn chứa các cụm. Phần trăm kích thước của mỗi cụm được hiển thị trên mỗi phần; khi di chuột qua mỗi phần, số lượng trong phần đó sẽ được hiển thị.

Dưới biểu đồ, một bảng liệt kê thông tin về kích thước như sau:

- Kích thước của cụm nhỏ nhất (bao gồm cả số lượng và phần trăm của toàn bộ): 224 (22.4%)

- Kích thước của cụm lớn nhất (bao gồm cả số lượng và phần trăm của toàn bộ): 435 (45.3%)
- Tỷ lệ kích thước của cụm lớn nhất so với cụm nhỏ nhất là: 2.02

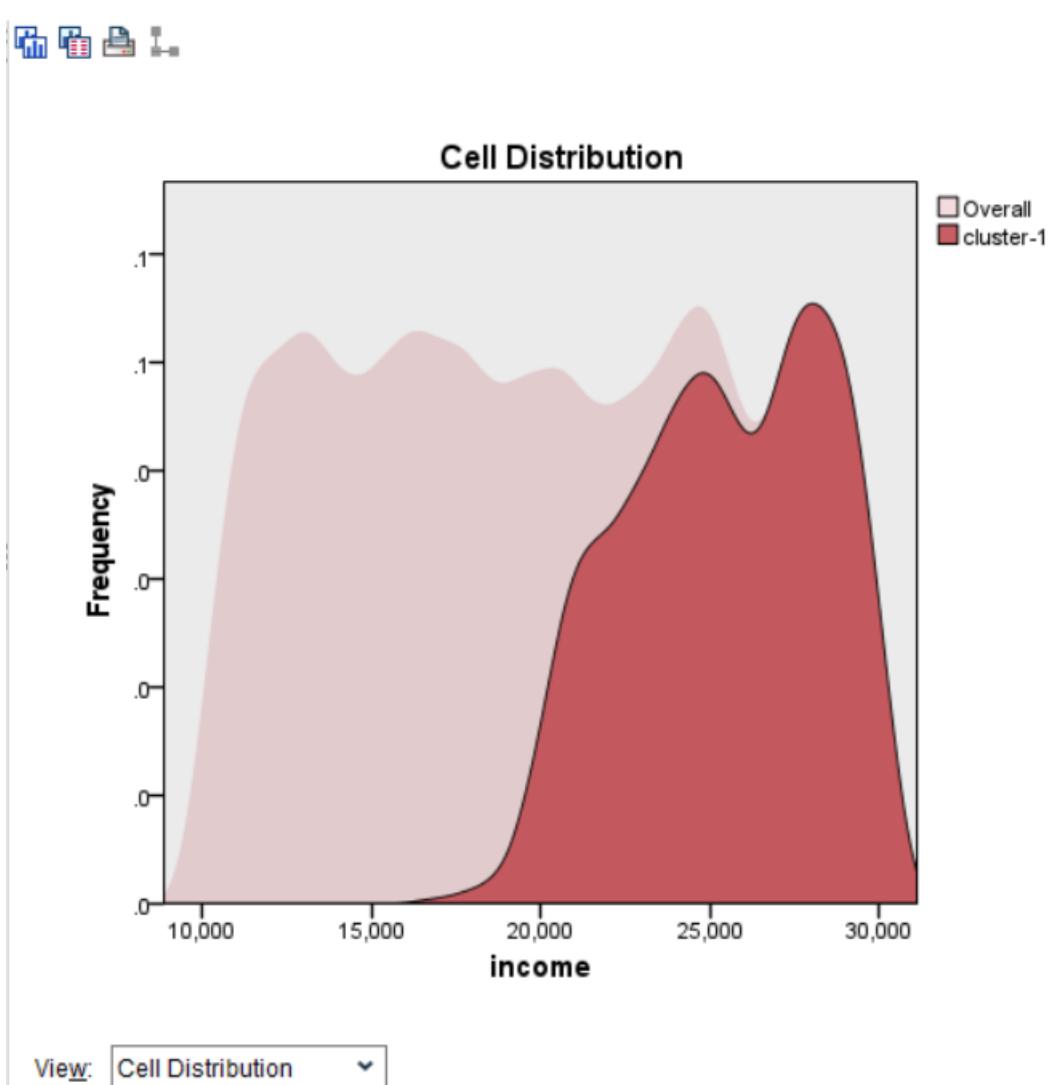


**Hình 3. 10.** Predictor Importance của thuật toán gom cụm.

**Predictor Importance** hiển thị sự quan trọng tương đối của mỗi trường để ước tính mô hình.

Ở đây ta có thể thấy:

- Thu nhập (income) có độ quan trọng: 100%
- Giá trị giỏ hàng (average value) có độ quan trọng: 56%
- Số thành viên trong gia đình (family\_size) có độ quan trọng: 27%
- Tuổi (age) có độ quan trọng: 11%



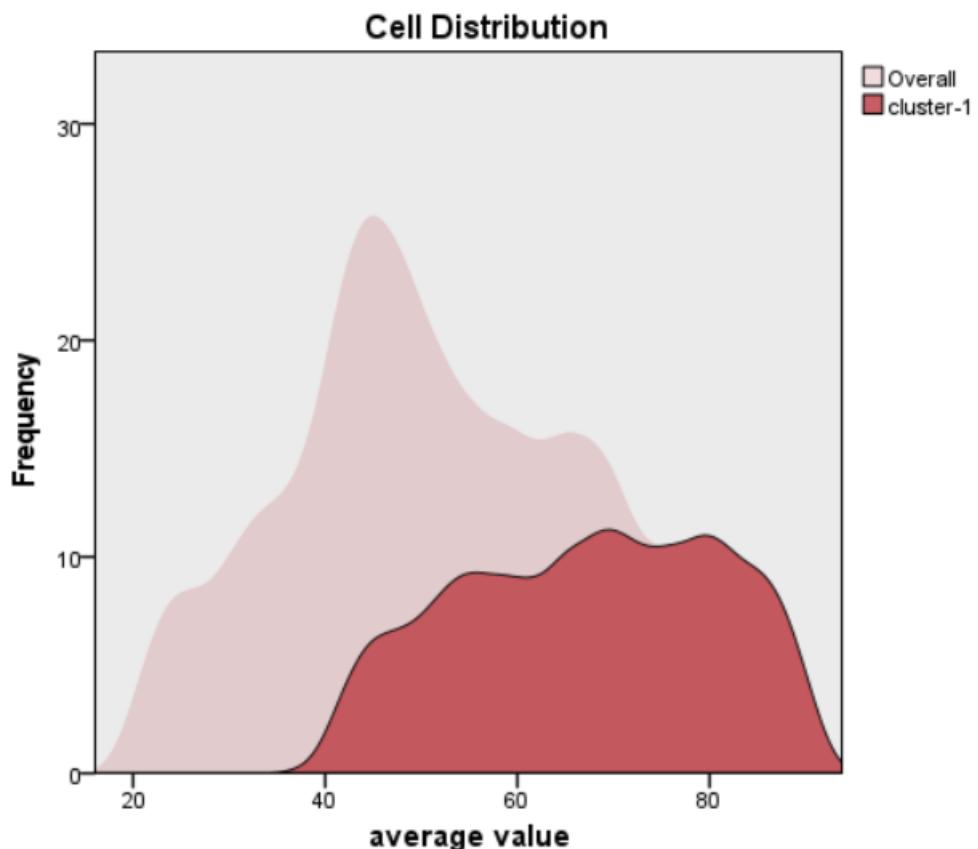
**Hình 3. 11.** Cell Distribution của thuật toán gom cụm.

**Cell Distribution** hiển thị một biểu đồ mở rộng, chi tiết hơn, về phân phối của dữ liệu cho bất kỳ ô tính năng nào mà bạn chọn trong bảng trên bảng điều khiển chính của mỗi cụm.

**Frequency** thể hiện tính thường xuyên của dữ liệu, hiển thị số lần xuất hiện của mỗi giá trị trong một biến. Nó có thể được sử dụng để tìm hiểu phân bố của một biến và cung cấp thông tin về tần suất xuất hiện của các giá trị riêng lẻ.

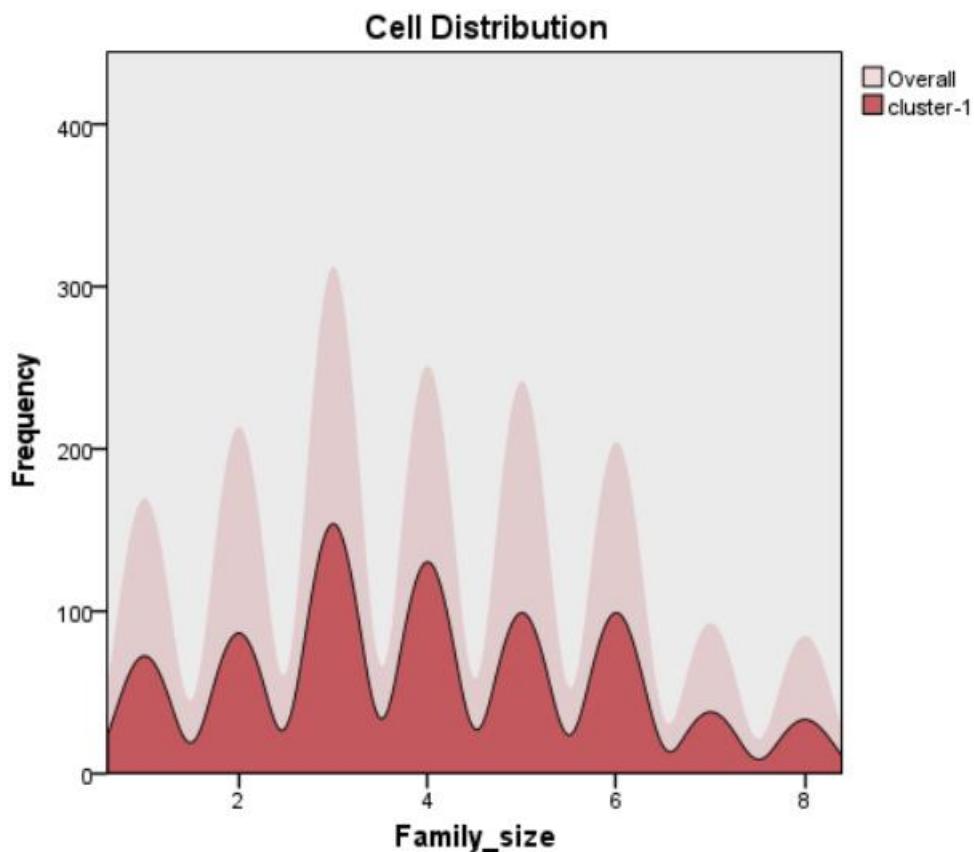
Ở đây nhóm chúng em sẽ phân tích cụm 1 làm mẫu:

Như hình trên cho ta thấy Cell Distribution của thu nhập (income) của cụm 1, nó cho ta thấy thu nhập của khách hàng ở cụm nằm ở mức khá cao (khoảng 13.000 đến hơn 30.000)



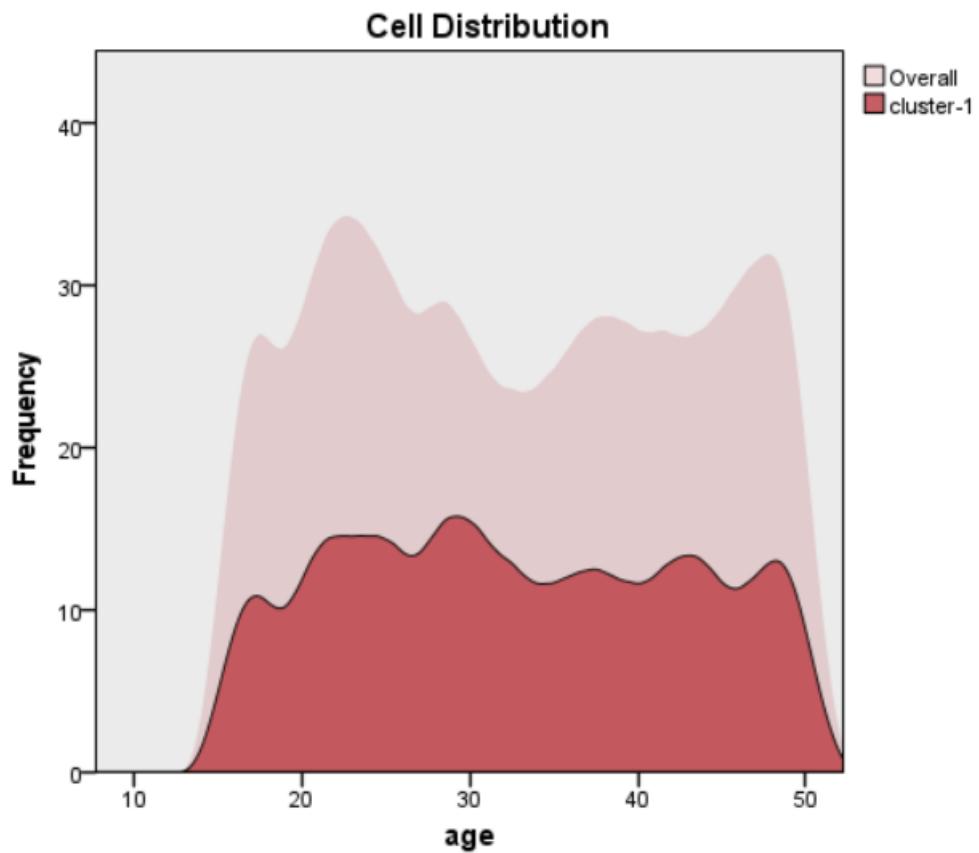
**Hình 3. 12.** Cell Distribution cho average value.

Cell Distribution của giá trị đơn hàng (average value) của cụm 1 cũng ở mức khá cao (khoảng 37 đến gần 90)



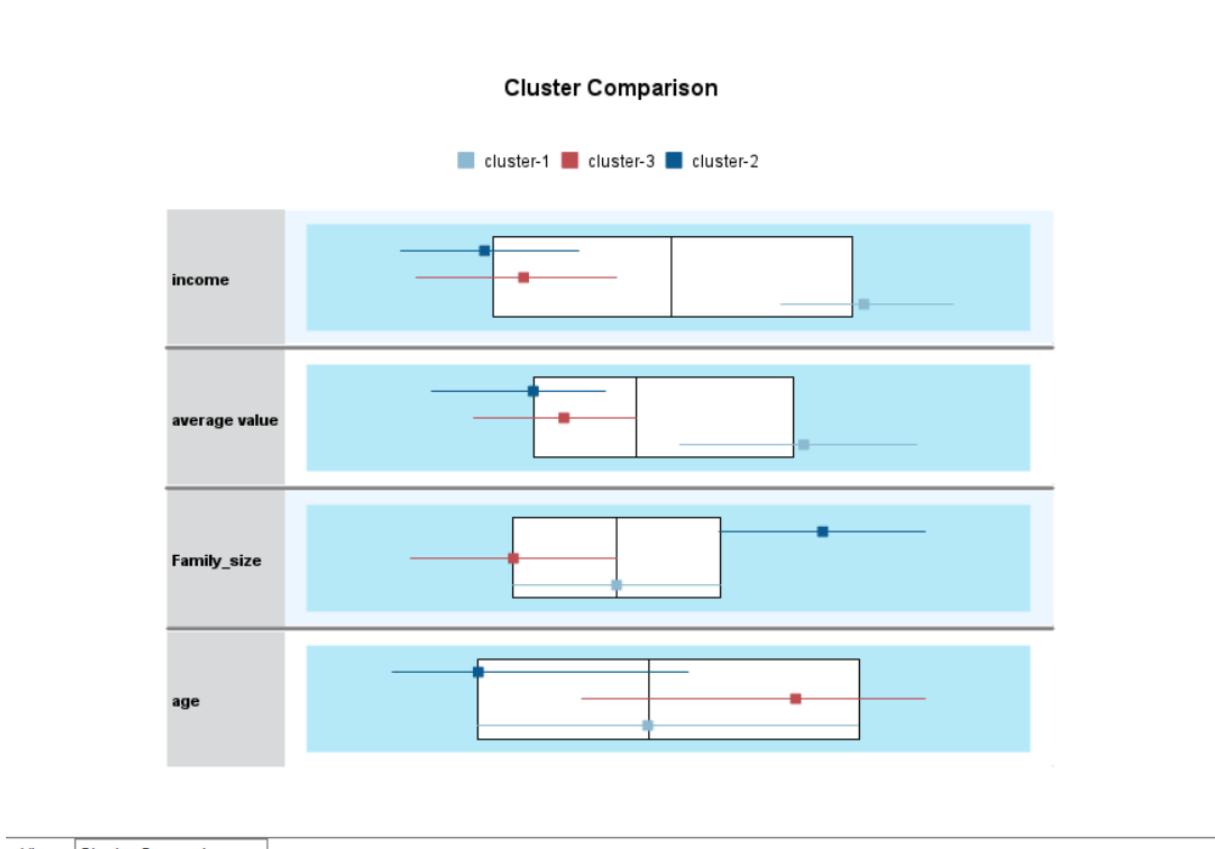
**Hình 3. 13.** Cell Distribution cho Family\_size.

Cell Distribution của số các thành viên trong gia đình (family\_size) của cụm 1 như biểu đồ cho ta thấy có thể nói là khá đồng đều (kéo dài từ độc thân đến hơn 8 thành viên).



**Hình 3. 14.** Cell Distribution cho age.

Cell Distribution của tuổi (age) của cụm 1 cũng được phân bổ khá đồng đều (khoảng 16 đến 52 tuổi)

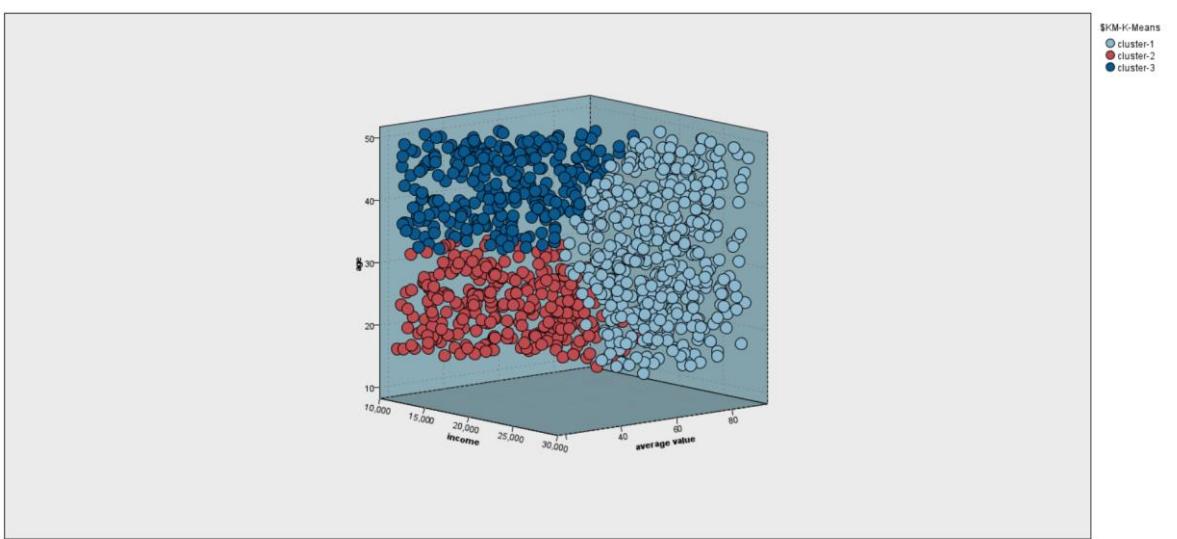


**Hình 3. 15. Cluster Comparison của thuật toán gom cụm.**

**Cluster Comparison** ché độ xem này giúp bạn hiểu rõ hơn về các yếu tố tạo nên các cụm; nó cũng cho phép bạn xem sự khác biệt giữa các cụm không chỉ so sánh với dữ liệu tổng thể, mà còn so sánh với nhau.

- **Cụm 1:** sẽ là nhóm khách hàng có trung bình thu nhập (income) cao (23,188 – 27,901) ; trung bình giá trị đơn hàng ( average value) cao (56 – 79); trung bình số thành viên trong gia đình (family\_size) được phân bố đồng đều (3 – 5) và trung bình tuổi (age) được phân bố đồng đều (24 – 41).
- **Cụm 2:** sẽ là nhóm khách hàng có trung bình thu nhập (income) thấp (12,770 – 17,643); trung bình giá trị đơn hàng ( average value) thấp (32 – 48,92); trung bình số thành viên trong gia đình (family\_size) được phân bố ở mức cao (4 – 6) và trung bình tuổi (age) ở mức khá trẻ (20 – 33,96)
- **Cụm 3:** sẽ là nhóm khách hàng có trung bình thu nhập (income) khá thấp (13,194 – 17,643); trung bình giá trị đơn hàng ( average value) khá thấp (36 – 51,89);

trung bình số thành viên trong gia đình (family\_size) được phân bố ở mức thấp (2 – 4) và trung bình tuổi (age) ở mức khá cao (28,94 – 45,06)



**Hình 3. 16.** Biểu đồ trực quan thuật toán gom cụm.

Biểu diễn dữ liệu giúp chúng ta trực quan hóa dữ liệu và hiểu rõ hơn về phân bố của nó.

**Cụm 1:** cho thấy sự đồng đều về giá trị đơn hàng của khách hàng bát kẽ thu nhập, độ tuổi và số thành viên trong gia đình.

**Cụm 2:** cho thấy xu hướng giá trị giỏ hàng từ thấp đến trung bình do thu nhập thấp và khách có độ tuổi khá thấp.

**Cụm 3:** cho thấy xu hướng giá trị giỏ hàng cao nhờ vào thu nhập từ trung bình đến cao và độ tuổi cao.

Từ biểu đồ trên ta có thể đưa ra xu hướng và dự báo dựa trên giá trị đơn hàng và độ tuổi, trong đó trục x biểu thị giá trị đơn hàng và trục y biểu thị độ tuổi. Ta nhận thấy rằng, có một xu hướng tăng dần của giá trị đơn hàng khi độ tuổi tăng lên, thì bạn có thể kết luận rằng khách hàng càng già thì càng có xu hướng mua nhiều sản phẩm có giá trị cao hơn. Bằng cách sử dụng kết quả dự báo, bạn có thể đưa ra các quyết định về chiến lược kinh doanh như tăng cường quảng cáo cho nhóm khách hàng già hơn, tăng cường sản xuất các sản phẩm có giá trị cao hơn, v.v.

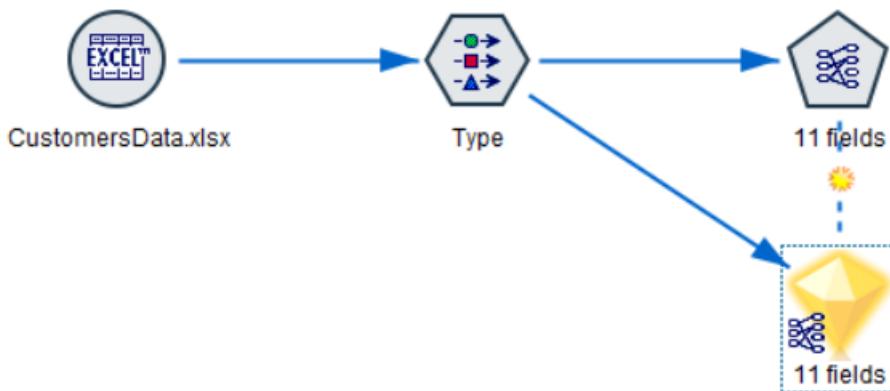
## Kết luận.

Kết luận của thuật toán K-means là các cụm khách hàng được phân chia dựa trên sự tương đồng về thu nhập, tuổi, số thành viên trong gia đình và giá trị giỏ hàng. Các cụm này có thể được sử dụng để tối ưu hóa chiến lược kinh doanh, chẳng hạn như định hướng các chiến dịch quảng cáo cho các nhóm đối tượng dựa trên độ tuổi, giảm giá nhóm sản phẩm dựa vào mức tiêu dùng hoặc tăng cường dịch vụ khách hàng tại các cụm khách hàng có giá trị cao hơn.

### 3.3.3. Thuật toán kết hợp.

Field	Measurement	Values	Missing	Check	Role
cardid	Continuous	[10150.0,1...		None	<input type="checkbox"/> None
average value	Continuous	[20.0,90.0]		None	<input type="checkbox"/> None
age	Continuous	[16.0,50.0]		None	<input type="checkbox"/> None
Shopping_fr...	Nominal	HIGH,LO...		None	<input type="checkbox"/> None
satisfaction l...	Continuous	[2.3,5.0]		None	<input type="checkbox"/> None
Family_size	Continuous	[1.0,8.0]		None	<input type="checkbox"/> None
income	Continuous	[10200.0,3...		None	<input type="checkbox"/> None
vegetables, fr...	Flag	T/F		None	<input checked="" type="checkbox"/> Both
meat, fish	Flag	T/F		None	<input checked="" type="checkbox"/> Both
dairy	Flag	T/F		None	<input checked="" type="checkbox"/> Both
rice-based pr...	Flag	T/F		None	<input checked="" type="checkbox"/> Both
sauce, seas...	Flag	T/F		None	<input checked="" type="checkbox"/> Both
frozen foods	Flag	T/F		None	<input checked="" type="checkbox"/> Both
wine,beer	Flag	T/F		None	<input checked="" type="checkbox"/> Both
personal car...	Flag	T/F		None	<input checked="" type="checkbox"/> Both
softdrink	Flag	T/F		None	<input checked="" type="checkbox"/> Both
household cl...	Flag	T/F		None	<input checked="" type="checkbox"/> Both
confectionery	Flag	T/F		None	<input checked="" type="checkbox"/> Both
loyalty card	Flag	T/F		None	<input type="checkbox"/> None

Hình 3. 17. Chính sửa thuộc tính cho thuật toán kết hợp.



**Hình 3. 18.** Mô hình của thuật toán gom cụm.

Model				Settings	Summary	Annotations
				Sort by:	Confidence %	
Consequent	Antecedent	Support %	Confidence %			
vegetables of fruits	meat, fish	62.1	96.135			
household cleaning pro...	personal care products ...	60.2	93.189			
rice-based products	dairy	57.6	92.361			
meat, fish	vegetables of fruits	65.2	91.564			
dairy	rice-based products	59.2	89.865			
personal care products ...	household cleaning pro...	65.4	85.78			
frozen foods	sauce, seasonings	69.8	82.378			
frozen foods	sauce, seasonings confectionery	50.9	82.122			
sauce, seasonings	softdrink frozen foods	50.0	81.6			
sauce, seasonings	frozen foods	70.7	81.33			
sauce, seasonings	frozen foods confectionery	51.6	81.008			
vegetables of fruits	dairy	57.6	80.903			
vegetables of fruits	dairy rice-based products	53.2	80.639			

**Hình 3. 19.** Kết quả của thuật toán gom cụm.

Từ các cặp dữ liệu kết hợp cho trước, chúng ta có thể nhận thấy một số thông tin về những mối quan hệ giữa các sản phẩm mà khách hàng thường hay mua sắm cùng nhau. Cụ thể, khách hàng thường hay mua rau củ và trái cây cũng thường mua thịt và cá (với tỷ lệ hỗ trợ là 62,1% và độ tin cậy là 96,135%). Ngược lại, khách hàng thường hay

mua sản phẩm vệ sinh gia đình thì cũng thường hay mua sản phẩm chăm sóc cá nhân (với tỷ lệ hỗ trợ là 60,2% và độ tin cậy là 93,189%).

Bên cạnh đó, khách hàng mua sản phẩm từ gạo thường cũng hay mua sản phẩm từ sữa (với tỷ lệ hỗ trợ là 59,2% và độ tin cậy là 89,865%). Trong khi đó, khách hàng mua sản phẩm đông lạnh thì thường có xu hướng mua sốt và gia vị (với tỷ lệ hỗ trợ là 69,8% và độ tin cậy là 82,378%), hoặc mua sốt, gia vị và kẹo (với tỷ lệ hỗ trợ là 50,9% và độ tin cậy là 82,122%).

Ngoài ra, phân tích còn cho thấy rằng giữa thịt và cá và rau củ và trái cây, cũng như giữa sản phẩm đông lạnh và sốt và gia vị có một mối liên hệ tương quan cao. Điều này có nghĩa là khách hàng mua thịt và cá cũng thường hay mua rau củ và trái cây (với tỷ lệ hỗ trợ là 65,2% và độ tin cậy là 91,564%), trong khi khách hàng mua sản phẩm đông lạnh thì thường hay mua sốt và gia vị (với tỷ lệ hỗ trợ là 70,7% và độ tin cậy là 81,33%).

Dựa trên phân tích kết hợp dữ liệu, cửa hàng có thể áp dụng các giải pháp sau để nâng cao doanh thu:

- Xây dựng gian hàng hoặc khu vực trưng bày sản phẩm được kết hợp với nhau, chẳng hạn như đặt sản phẩm rau củ và trái cây kế bên với thịt và cá, hoặc đặt sản phẩm từ sữa tại khu vực gần sản phẩm từ gạo, để tăng khả năng khách hàng mua các sản phẩm này cùng nhau.
- Tạo ra các gói sản phẩm combo đi kèm với giá ưu đãi, ví dụ như bán một gói bao gồm sản phẩm đông lạnh và sốt gia vị để khuyến khích khách hàng mua các sản phẩm này cùng nhau.
- Sử dụng các ứng dụng quản lý khách hàng để phân tích dữ liệu và hiểu rõ hơn về sở thích và nhu cầu của từng khách hàng. Dựa trên thông tin này, cửa hàng có thể gợi ý cho khách hàng mua sắm các sản phẩm kết hợp với nhau hoặc gửi thông báo khuyến mãi đến khách hàng khi có các sản phẩm mà họ thường hay mua.

- Chạy chiến dịch quảng cáo và marketing tập trung vào những sản phẩm kết hợp với nhau để khách hàng có thể nhận ra giá trị của việc mua các sản phẩm này cùng nhau.
- Nâng cao trải nghiệm mua sắm bằng cách cung cấp thông tin chi tiết và hướng dẫn về cách kết hợp các sản phẩm tốt nhất để khách hàng có thể quyết định mua sắm dễ dàng hơn.

Những giải pháp trên sẽ giúp tăng doanh thu cho cửa hàng bằng cách khuyến khích khách hàng mua sắm các sản phẩm kết hợp với nhau, từ đó tăng tỷ lệ chuyển đổi và giá trị đơn hàng.

## CHƯƠNG 4. KẾT LUẬN

### 4.1. Những kết quả đạt được của đồ án.

- Thực hiện đầy đủ ba thuật toán cơ bản theo yêu cầu của đồ án.
- Tối ưu hóa quy trình kinh doanh của mình, bằng cách phân tích các dữ liệu kinh doanh, nhóm chúng em có thể tìm ra cách cải thiện hoạt động kinh doanh và tăng hiệu quả.
- Giúp phân tích hiệu quả của các chiến lược hoặc quyết định đã được đưa ra trước đó. Bằng cách đánh giá và phân tích các dữ liệu đã được thu thập, có thể tìm ra các lợi thế và nhược điểm của các chiến lược hoặc quyết định đã được thực hiện.
- Tìm ra thông tin mới từ dữ liệu được thu thập. Điều này có thể giúp hiểu rõ hơn về các mối liên hệ giữa các biến hoặc phát hiện ra các xu hướng mới.
- Sử dụng các phương pháp khai phá dữ liệu phù hợp, giúp đưa ra các quyết định hoặc lập kế hoạch cho tương lai.

### 4.2. Nhược điểm của đồ án.

- Do thời gian nghiên cứu và thực hiện còn hạn chế, nên việc tìm hiểu về các thuật toán còn chưa được đầy đủ.
- Chưa tìm được dữ liệu phù hợp khiết cho các kết quả phân tích thiếu tính tổng quát.
- Còn nhiều chức năng của phần mềm IBM SPSS Modeler chưa được sử dụng một cách tối ưu.

### 4.3. Hướng phát triển cho đồ án.

- Thu thập những bộ dữ liệu lớn hơn để tiến hành thử nghiệm trên phần mềm.
- Áp dụng các phương pháp khai phá dữ liệu phù hợp hơn với mục tiêu của đồ án.
- Thực hành thêm nhiều thuật toán khác với các thuật toán đã được triển khai trong đồ án.
- Cải thiện và tối ưu hóa đồ án khai phá dữ liệu bằng các phương pháp khác để tăng tính chính xác và hiệu quả của kết quả.

## TÀI LIỆU THAM KHẢO

### Danh sách tài liệu.

[1] Nguyễn Thị Trần Lộc, Bài giảng Khai phá Dữ liệu.

### Danh sách website.

[2] <https://www.ibm.com/docs/en/spss-modeler/18.0.0?topic=mining-types-models>

[3] <https://insight.isb.edu.vn/6-ky-thuat-quan-trong-trong-khai-pha-du-lieu/>

[4] <https://support.microsoft.com/vi-vn/office/tổng-quan-về-xử-lý-phân-tích-trực-tuyến-olap-15d2cdde-f70b-4277-b009-ed732b75fdd6>