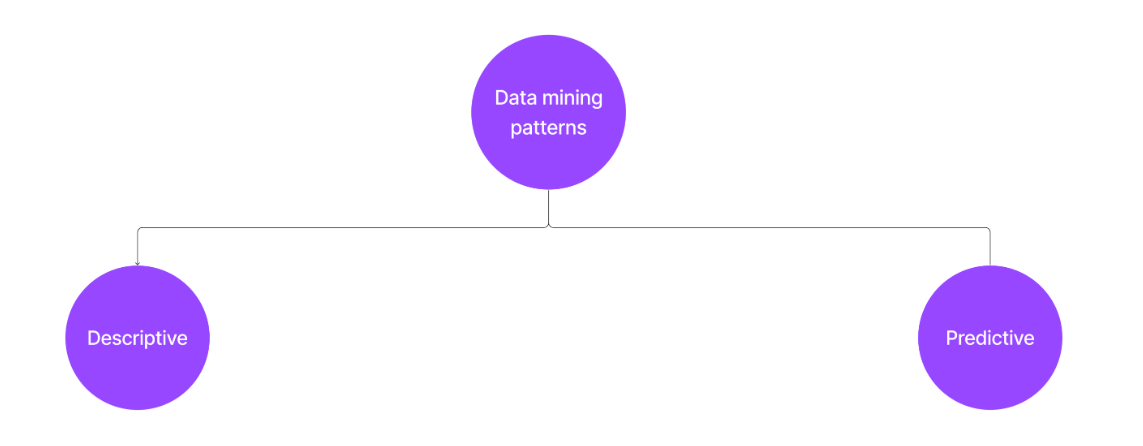
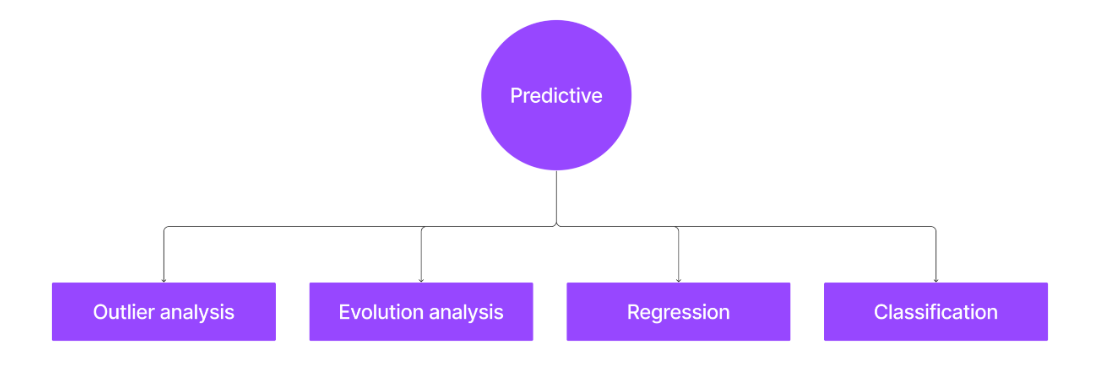
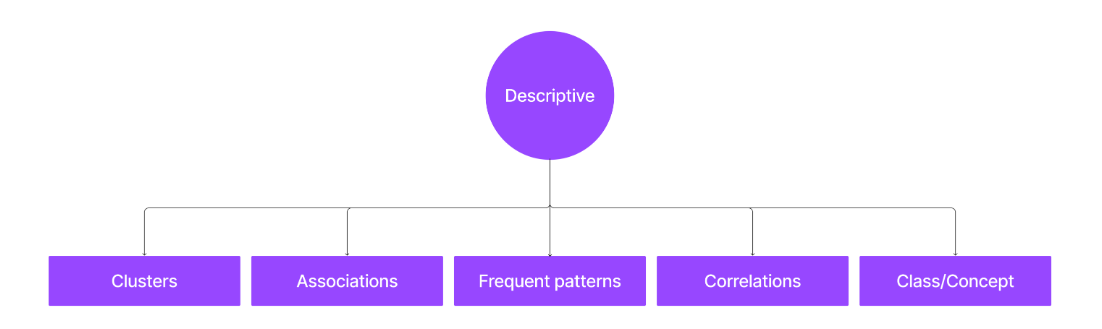
# 1.4.What Kinds of Patterns Can Be Mined? (Khải)

Có thể khai phá nhiều loại dữ liệu khác nhau trong khai phá dữ liệu. Tuy nhiên, dữ liệu phải có một mẫu để có được những thông tin hữu ích. Trước tiên ta tìm hiểu xem khai phá dữ liệu có những chức năng gì. Có một số chức năng (functionality) trong khai phá dữ liệu, bao gồm:

1. Đặc trưng hóa (characterization) và phân biệt (discrimination) (Mục 1.4.1)
2. Khai phá các mẫu phổ biến (frequent patterns), mối liên kết (associations) và tương quan (correlations) (Mục 1.4.2)
3. Phân loại (classification) và hồi quy (regression) (Mục 1.4.3)
4. Phân tích cụm (clustering analysis) (Mục 1.4.4)
5. Phân tích điểm dị thường (outlier analysis) (Mục 1.4.5)

Dựa trên chức năng trong khai phá dữ liệu, các mẫu có thể được phân loại thành hai loại đó là: mô tả (descriptive) và dự đoán (predictive).





Ngoài ra, Mục 1.4.6 xem xét những gì làm cho một mẫu hình trở nên hữu ích (interesting). Các mẫu hữu ích đại diện cho kiến thức.

## 1.4.1 Class/Concept Description: Characterization and Discrimination

Class/concept descriptions: là các mô tả về một nhóm đối tượng

Các mô tả này có thể được suy ra bằng cách sử dụng Data Characterization (đặc trưng hóa dữ liệu) hoặc Data Discrimination (phân biệt dữ liệu)

|  | **Data Characterization (đặc trưng hóa dữ liệu)** | **Data Discrimination (phân biệt dữ liệu)** |
| --- | --- | --- |
| Input | Một nhóm các đối tượng (object) dữ liệu | Một nhóm các mô tả (description) |
| Output | Mô tả (Description) chung về nhóm đối tượng này | Điểm giống và khác nhau giữa những đoạn mô tả (description) này |
| Biểu diễn | Sử dụng các loại biểu đồ thông dụng: biểu đồ tròn, biểu đồ đường, bảng đa chiều,... | |
|  | Nên có thêm những thể hiện về so sánh dữ liệu |
| Ví dụ | Ngữ cảnh: Công ty A chuyên cung cấp những thiết bị điện tử (laptop, chuột, bàn phím,..) | |
| Nhiệm vụ: Tóm tắt thông tin của những người trả cho những sản phẩm của công ty nhiều hơn 30 triệu / năm,  Kết quả: độ tuổi từ 25 → 35, có việc làm, có mức tín dụng cao. Bên cạnh đó, hệ thống khai phá dữ liệu còn có thể cho ra những kết quả ở khía cạnh khác như: nghề nghiệp cụ thể của từng người để cho ra thông tin về những mảng ngành nghề nào sử dụng dịch vụ của công ty | Nhiệm vụ: So sánh dữ liệu của 2 nhóm khách hàng:  - Mua hàng thường xuyên (từ 1 lần trở lên mỗi tháng)  - Hiếm khi mua hàng (Ít hơn 3 lần mỗi năm)  Kết quả:  - 80% khách hàng thường xuyên mua sản phẩm máy tính có độ tuổi từ 20 đến 40 và có trình độ đại học  - 60% khách hàng không thường xuyên mua những sản phẩm như vậy là người cao tuổi hoặc thanh thiếu niên và không có bằng đại học  Việc đào sâu vào những khía cạnh như nghề nghiệp hoặc mức thu nhập có thể giúp tìm ra những đặc điểm phân biệt hơn nữa giữa hai nhóm. |
| Phương pháp | Các phương pháp dùng cho Data Characterization cũng giống với Data Discrimination:  - Thu thập dữ liệu  - Trích xuất đặc trưng  - Lựa chọn đặc trưng quan trọng  - Xây dựng mô hình phân loại | |

## 1.4.2. Mining Frequent Patterns, Associations, and Correlations

Mining Frequent Patterns, Associations, and Correlations là một chủ đề quan trọng trong học máy và phân tích dữ liệu, liên quan đến việc phát hiện các mẫu (patterns) phổ biến, các mối liên kết (associations) và các tương quan (correlations) trong dữ liệu.

Cụ thể, quá trình này bao gồm các bước sau:

1. Phát hiện Mẫu Phổ Biến (Frequent Patterns):
   * Xác định các mẫu (ví dụ: tập hợp các mục, chuỗi các sự kiện) xuất hiện thường xuyên trong dữ liệu.
   * Ví dụ: Các sản phẩm thường được mua cùng nhau.
2. Khai thác Các Liên Kết (Association Rule Mining):
   * Tìm ra các quy tắc liên kết (association rules) giữa các mẫu phổ biến.
   * Ví dụ: "Khách hàng mua sản phẩm A thì có 80% khả năng cũng mua sản phẩm B".
3. Phát hiện Các Tương Quan (Correlation Analysis):
   * Xác định các mối tương quan (tương quan dương hoặc âm) giữa các biến số trong dữ liệu.
   * Ví dụ: Tìm mối tương quan giữa tuổi và thu nhập của khách hàng.

Các kỹ thuật phổ biến để thực hiện các nhiệm vụ này bao gồm:

* Algorithms như Apriori, FP-Growth để khai thác mẫu phổ biến.
* Các phương pháp như Association Rule Mining, Correlation Analysis.

Các ứng dụng của việc khai thác Frequent Patterns, Associations và Correlations bao gồm:

* Phân tích hành vi người dùng (ví dụ: khuyến nghị sản phẩm)
* Phát hiện gian lận (ví dụ: giao dịch nghi ngờ)
* Phân tích tài chính (ví dụ: dự đoán xu hướng thị trường)
* Phân tích dữ liệu y tế (ví dụ: quan hệ giữa các triệu chứng bệnh)

· Các mẫu phổ biến: là các mẫu mà thường xuyên xuất hiện trong dữ liệu

· Có nhiều loại frequent patterns

o Frequent itemsets: tập các mục thường xuất hiện cùng nhau trong 1 tập dữ liệu. Ví dụ là sữa và bánh thường được mua cùng nhau bới khách hàng ở cửa hàng tiện lợi

o Frequent subsequences (sequential patterns): tập nó xảy ra tuần tự. Ví dụ khách hàng có xu hướng mua laptop sau đó mua máy ảnh kỹ thuật số sau đó đến thẻ nhớ

o Frequent substructures: là sự kết hợp của Frequent itemsets và Frequent subsequences và có thể được biểu diễn dưới dạng cậy, đồ thị, lưới

· Khai thác các mẫu thường xuyên dẫn đến việc khám phá ra các mối liên kết và tương quan thú vị trong dữ liệu

· **Association:** là một kỹ thuật được sử dụng trong khai thác dữ liệu để xác định mối quan hệ hoặc sự đồng hiện giữa các mục trong một tập dữ liệu

· **Correlation Analysis** là một kỹ thuật khai thác dữ liệu được sử dụng để xác định mức độ mà hai hoặc nhiều biến có liên quan hoặc liên kết với nhau

· Association analysis.

o Giả sử, với tư cách là giám đốc tiếp thị tại AllElectronics, bạn muốn biết những mặt hàng nào thường được mua cùng nhau. Một ví dụ về quy tắc như vậy, được khai thác từ cơ sở dữ liệu giao dịch của AllElectronics, là

buys(X, “computer”) ⇒ buys(X, “software”) [support = 1%,confidence = 50%] (single-dimensional association rules)

o X : khách hàng

o Confidence (độ tin cậy): độ tin cậy 50% có nghĩa là cứ một khách hàng mua máy tính thì sẽ có 50% cơ hội rằng khách sẽ mua phần mềm (cho biết độ tin cậy của quy tắc đó trong tập dữ liệu)

o Support: support = 1%, có nghĩa là 1% trong tổng số giao dịch được phân tích cho thấy máy tính và phần mềm được mua cùng nhau (cho biết mức độ phổ biến của quy tắc trong tập dữ liệu)

o Có thể viết tắt là computer => software[1%,50%]

o Có thể kết hợp nhiều vị ngữ (predicate)

age(X, “20..29”) ∧ income(X, “40K..49K”) ⇒ buys(X, “laptop”)[ support = 2%, confidence = 60%] (multidimensional association rule)

Nghĩa là: khách hàng có độ tuổi từ 20 đến 29 và thu nhập từ $40,000 đến $49,000 đã mua một chiếc laptop tại AllElectronics và có 60% khả năng một khách hàng trong độ tuổi và thu nhập trên sẽ mua laptop tại AllElectronics

o Thông thường, **association rules** bị loại bỏ vì chúng không thỏa mãn **minimum support threshold** và **minimum confidence threshold**. Có thể thực hiện phân tích bổ sung để khám phá mối tương quan thống kê thú vị giữa các cặp thuộc tính–giá trị liên quan

Kham khảo: [Association and Correlation in Data Mining - Scaler Topics](https://www.scaler.com/topics/data-mining-tutorial/association-and-correlation-in-data-mining/)

## 1.4.3.Classification and Regression for Predictive Analysis

· Classification (phân loại): là quá trình tìm kiếm một mô hình (hoặc chức năng) mô tả và phân biệt các lớp dữ liệu hoặc khái niệm. Mô hình được suy ra dựa trên việc phân tích một tập dữ liệu đào tạo (tức là các đối tượng dữ liệu mà nhãn lớp được biết đến). Mô hình được sử dụng để dự đoán nhãn lớp của các đối tượng mà nhãn lớp không được biết đến.

· Mô hình hồi quy các hàm có giá trị liên tục. Nghĩa là, hồi quy được sử dụng để dự đoán các giá trị dữ liệu số bị thiếu hoặc không có sẵn thay vì các nhãn lớp (rời rạc). Thuật ngữ dự đoán đề cập đến cả dự đoán số và dự đoán nhãn lớp. Phân tích hồi quy là một phương pháp thống kê thường được sử dụng nhất để dự đoán số, mặc dù cũng có các phương pháp khác. Hồi quy cũng bao gồm việc xác định các xu hướng phân phối dựa trên dữ liệu có sẵn.

· Ví dụ

**Phân loại**

Giả sử bạn là quản lý bán hàng của cửa hàng AllElectronics và muốn phân loại một lượng lớn sản phẩm trong cửa hàng dựa trên ba loại phản ứng với một chiến dịch quảng cáo: phản ứng tốt, phản ứng trung bình và không phản ứng. Bạn muốn xây dựng một mô hình cho mỗi loại dựa trên các đặc tính mô tả của sản phẩm như giá, thương hiệu, nơi sản xuất, loại và danh mục. Phân loại kết quả phải phân biệt tối đa mỗi loại với các loại khác, tạo ra một bức tranh tổng quan về tập dữ liệu.

Giả sử phân loại kết quả được biểu diễn dưới dạng cây quyết định. Cây quyết định, ví dụ, có thể xác định giá là yếu tố duy nhất phân biệt tốt nhất ba loại. Cây có thể tiết lộ rằng, ngoài giá, các tính năng khác giúp phân biệt thêm các đối tượng của mỗi loại bao gồm thương hiệu và nơi sản xuất. Một cây quyết định như vậy có thể giúp bạn hiểu tác động của chiến dịch quảng cáo đã cho và thiết kế một chiến dịch hiệu quả hơn trong tương lai.

**Hồi quy**

Giả sử thay vì dự đoán nhãn phản hồi theo loại cho mỗi mặt hàng cửa hàng, bạn muốn dự đoán số tiền doanh thu mà mỗi mặt hàng sẽ tạo ra trong một đợt bán hàng sắp tới tại AllElectronics, dựa trên dữ liệu bán hàng trước đó. Đây là một ví dụ về phân tích hồi quy vì mô hình hồi quy được xây dựng sẽ dự đoán một hàm liên tục (hoặc giá trị được sắp xếp).

**Tóm lại:**

* **Phân loại** là việc gán một nhãn (ví dụ: tốt, trung bình, kém) cho một đối tượng dựa trên các đặc tính của nó.
* **Hồi quy** là việc dự đoán một giá trị số liên tục (ví dụ: doanh thu) dựa trên các đặc tính của đối tượng.

Cả hai kỹ thuật đều được sử dụng rộng rãi trong khai phá dữ liệu để tìm hiểu mối quan hệ giữa các biến và đưa ra dự đoán.

## 1.4.4. Cluster Analysis

Không như phân loại (classification) và hồi quy (regression), phân tích cụm (clustering) phân tích các đối tượng dữ liệu mà không cần tham khảo các nhãn lớp. Cluster Analysis (Phân tích cụm) là một kỹ thuật data mining phổ biến, được sử dụng để phát hiện các mẫu ẩn trong dữ liệu bằng cách chia dữ liệu thành các cụm (clusters) dựa trên tính tương tự của các dữ liệu. Mỗi cụm được hình thành như vậy có thể được xem là một lớp các đối tượng, từ đó có thể rút ra các quy tắc.

Ở đây là cách thức Cluster Analysis hoạt động trong data mining:

1. Mục tiêu:
   * Phân nhóm các đối tượng vào các cụm sao cho các đối tượng trong cùng nhóm có độ tương tự cao, còn các đối tượng ở các nhóm khác có độ tương tự thấp.
   * Tìm ra các cấu trúc ẩn trong dữ liệu thông qua việc phân chia dữ liệu vào các nhóm khác nhau.
2. Cách thức:
   * Xác định các đặc trưng (features) mô tả các đối tượng.
   * Tính toán độ tương tự hoặc độ khác biệt giữa các đối tượng dựa trên các đặc trưng này.
   * Sử dụng thuật toán phân cụm (clustering algorithm) để chia dữ liệu thành các nhóm dựa trên độ tương tự.
   * Các thuật toán phổ biến bao gồm K-Means, Hierarchical Clustering, DBSCAN, Gaussian Mixture Models.
3. Ứng dụng:
   * Phân khúc khách hàng trong marketing để thiết kế chiến lược tiếp thị phù hợp.
   * Phát hiện gian lận trong giao dịch tài chính.
   * Nhóm các bài báo tin tức có chủ đề tương tự.
   * Khám phá các nhóm gen tương đồng trong dữ liệu gene.

## 1.4.5. Outlier Analysis

1. Cách thức:
   * Xác định các đặc trưng (features) mô tả các quan sát.
   * Sử dụng các phương pháp thống kê và machine learning để phát hiện các quan sát ngoại biên, như:
     + Khoảng cách (distance-based): Xác định các quan sát cách xa trung tâm của dữ liệu.
     + Dựa trên mô hình (model-based): Xác định các quan sát không phù hợp với mô hình dự đoán.
     + Dựa trên mật độ (density-based): Xác định các quan sát nằm ở vùng có mật độ dữ liệu thấp.
   * Phân tích và hiểu rõ nguyên nhân dẫn đến các quan sát ngoại biên.

Một tập dữ liệu có thể chứa các đối tượng không tuân theo hành vi chung hoặc mô hình của dữ liệu. Những đối tượng dữ liệu này là các giá trị ngoại lai (outlier).

Nhiều phương pháp khai phá dữ liệu loại bỏ các ngoại lai vì coi chúng là nhiễu hoặc ngoại lệ. Tuy nhiên, trong một số ứng dụng (ví dụ: phát hiện gian lận), những sự kiện hiếm có thể thú vị hơn so với các sự kiện xảy ra thường xuyên hơn. Việc phân tích dữ liệu ngoại lệ được gọi là phân tích ngoại lệ hoặc khai phá dị thường.

* Ngoại lệ có thể được phát hiện bằng cách sử dụng các kiểm tra thống kê giả định một mô hình phân phối hoặc xác suất cho dữ liệu, hoặc sử dụng các thước đo khoảng cách, trong đó các đối tượng cách xa bất kỳ cụm nào khác được coi là ngoại lệ. Thay vì sử dụng các thước đo thống kê hoặc khoảng cách, các phương pháp dựa trên mật độ có thể xác định ngoại lệ trong một khu vực cục bộ, mặc dù chúng có vẻ bình thường khi nhìn từ quan điểm phân phối thống kê toàn cục.

Phân tích ngoại lệ. Phân tích ngoại lai có thể phát hiện việc sử dụng thẻ tín dụng gian lận bằng cách phát hiện các giao dịch mua với số tiền bất thường lớn so với các giao dịch thông thường được thực hiện bởi cùng một số tài khoản. Các giá trị ngoại lệ cũng có thể được phát hiện liên quan đến vị trí và loại hình mua sắm, hoặc tần suất mua sắm.

## 1.4.6. Are All Patterns Interesting?

* Không, chỉ một phần nhỏ các mẫu có khả năng được tạo ra thực sự hữu ích cho một người dùng nhất định.
* Một mẫu (pattern) thì có ích khi nó dễ dàng được nhận diện và hiểu biết bởi con người.
* Có một vài biện pháp khách quan để đánh giá độ hữu ích của mẫu. Những điều này dựa trên cấu trúc của các mẫu được phát hiện và các số liệu thống kê cơ bản của chúng.
* Một số mẫu có thể chỉ là ngẫu nhiên hoặc không có giá trị thực tế cho việc đưa ra quyết định.
* Một số lý do có thể khiến các mẫu không hữu ích:
* Tần suất cao nhưng không có ý nghĩa: một số mẫu có thể xuất hiện rất thường xuyên trong dữ liệu nhưng không cung cấp thông tin mới hay hữu ích. Chẳng hạn, nếu một cửa hàng phát hiện rằng hầu hết khách hàng đều mua sữa, điều này có thể không mang lại giá trị vì đây là một hành vi phổ biến và hiển nhiên.
* Mẫu không đáp ứng ngưỡng tối thiểu: một mẫu có thể không thú vị nếu nó không đáp ứng các tiêu chí cụ thể, chẳng hạn như ngưỡng hỗ trợ (support) hoặc độ tin cậy (confidence) tối thiểu. Những mẫu này có thể không đủ mạnh để được xem xét.
* Tính ngẫu nhiên hoặc lỗi dữ liệu: một số mẫu có thể là kết quả của dữ liệu nhiễu (noise) hoặc các sai lệch ngẫu nhiên trong dữ liệu. Những mẫu này không có ý nghĩa thực sự và không đáng để theo đuổi.
* Không phù hợp với mục tiêu phân tích: Một mẫu có thể không phù hợp hoặc không liên quan đến mục tiêu hoặc vấn đề mà người dùng đang cố gắng giải quyết. Ví dụ, một mẫu liên quan đến mùa bán hàng có thể không có giá trị trong phân tích về hành vi khách hàng cụ thể.
* Mẫu đã biết trước: một mẫu phát hiện được có thể là điều mà người dùng đã biết từ trước, và do đó không mang lại thông tin mới hay giúp ích trong việc ra quyết định.
* Các khái niệm khác:
* Khai phá dữ liệu: Là quá trình khám phá ra các mẫu và kiến thức thú vị từ những lượng dữ liệu lớn. Đây là một khái niệm rộng, bao gồm toàn bộ quá trình khám phá kiến thức, bao gồm các bước sau

1. Làm sạch dữ liệu - loại bỏ các dữ liệu ồn và không nhất quán
2. Tích hợp dữ liệu - kết hợp nhiều nguồn dữ liệu khác nhau
3. Lựa chọn dữ liệu - truy xuất dữ liệu liên quan cho tác vụ phân tích
4. Biến đổi dữ liệu - tổng hợp và tóm tắt dữ liệu thành các dạng phù hợp để khai thác
5. Khai phá dữ liệu - áp dụng các phương pháp thông minh để trích xuất các mẫu từ dữ liệu
6. Đánh giá mẫu - xác định những mẫu thú vị thật sự đại diện cho kiến thức có giá trị
7. Trình bày kiến thức - sử dụng các kỹ thuật trực quan hóa và biểu diễn khác để trình bày kiến thức được phát hiện cho người dùng.

*Mặc dù khai phá dữ liệu đôi khi chỉ được xem là một bước trong quá trình khám phá kiến thức tổng thể, nhưng trong thực tế, thuật ngữ "khai phá dữ liệu" và "khám phá kiến thức" thường được sử dụng thay thế cho nhau, vì khai phá dữ liệu bao gồm toàn bộ quá trình trích xuất kiến thức hữu ích từ những tập dữ liệu lớn. Đây là một lĩnh vực liên ngành, kết hợp các yếu tố từ thống kê, học máy, hệ thống cơ sở dữ liệu và các lĩnh vực khác.*

* Mẫu (Pattern) là gì?: Trong Data Mining, mẫu (pattern) là một cấu trúc, mối quan hệ hoặc xu hướng được tìm thấy trong dữ liệu
* Hãy tưởng tượng bạn có một bộ dữ liệu về các sản phẩm mà khách hàng mua trong một cửa hàng. Trong đó có thông tin về các sản phẩm được mua cùng nhau. Ví dụ, bạn nhận thấy rằng nhiều khách hàng thường mua bánh mì, sữa và trứng cùng với nhau. Điều này tạo thành một "mẫu" - một sự kết hợp các sản phẩm thường được mua cùng.

Mẫu này có thể giúp bạn:

* Hiểu được thói quen mua sắm của khách hàng
* Đề xuất các sản phẩm khác có thể khách hàng sẽ quan tâm
* Sắp xếp các sản phẩm trên kệ để khuyến khích mua các sản phẩm cùng nhau
* Như vậy, mẫu là những quy luật, mối liên kết, hoặc mối quan hệ được tìm thấy trong dữ liệu, và có thể mang lại những thông tin hữu ích cho việc ra quyết định hoặc hiểu rõ hơn về dữ liệu.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Các ví dụ về mẫu trong khai phá dữ liệu bao gồm:

* Các quy luật liên kết (association rules) như "Khách hàng mua bánh mì và sữa thường cũng mua trứng"
* Các chuỗi sự kiện cùng diễn ra (sequential patterns) như "Khách hàng thường mua sản phẩm X trước khi mua sản phẩm Y"
* Các nhóm khách hàng có hành vi tương đồng (clustering)
* Các yếu tố bất thường hoặc không phù hợp so với phần còn lại của dữ liệu (outliers)

Phát hiện và phân tích các mẫu trong dữ liệu là mục tiêu chính của khai phá dữ liệu, giúp tìm ra những kiến thức, hiểu biết và những thông tin hữu ích.