**TRƯỜNG ĐẠI HỌC THỦY LỢI**

**KHOA CÔNG NGHỆ THÔNG TIN**

-----🙡🕮🙣-----

****

**ĐỀ TÀI:**

**Khai phá xu hướng mua sắm từ dữ liệu siêu thị ứng dụng luật kết hợp Apriori**

**Nhóm sinh viên thực hiện:**

**Giảng viên hướng dẫn:**

**Hà Nội, 2023**

[**MỞ ĐẦU**](#_gjdgxs) **2**

**CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU 4**

[**CHƯƠNG 2: TIỀN XỬ LÝ DỮ LIỆU**](#_2s8eyo1) **6**

[2.1 Làm sạch dữ liệu.](#_17dp8vu) 6

[2.2 Xóa dữ liệu không cần thiết](#_3rdcrjn) 8

[**CHƯƠNG 3: KHAI PHÁ DỮ LIỆU BẰNG LUẬT KẾT HỢP APRIORI**](#_35nkun2) **10**

[3.1 Giới thiệu về luật kết hợp.](#_1ksv4uv) 10

[3.2 Phương pháp luật kết hợp Apriori](#_2jxsxqh) 11

[**CHƯƠNG 4: KẾT LUẬN**](#_3j2qqm3) **14**

[3.1 Kết luận.](#_1y810tw) 14

[**TÀI LIỆU THAM KHẢO**](#_2xcytpi) **15**

# **MỞ ĐẦU**

Trong những năm gần đây cùng với phát triển nhanh chóng của khoa học kỹ thuật là sự bùng nổ về tri thức. Kho dữ liệu, nguồn tri thức của nhân loại cũng trở nên đồ sộ, vô tận làm cho vấn đề khai thác các nguồn tri thức đó ngày càng trở nên nóng bỏng và đặt ra thách thức lớn cho nền công nghệ thông tin thế giới.

Nhu cầu về tìm kiếm và xử lý thông tin, cùng với yêu cầu về khả năng kịp thời khai thác chúng để mạng lại những năng suất và chất lượng cho công tác quản lý, hoạt động kinh doanh... đã trở nên cấp thiết trong xã hội hiện đại. Để đáp ứng phần nào yêu cầu này, người ta đã xây dựng các công cụ tìm kiếm và xử lý thông tin nhằm giúp cho người dùng tìm kiếm được các thông tin cần thiết cho mình.

Với các phương pháp khai thác cơ sở dữ liệu truyền thống chưa đáp ứng được các yêu cầu đó. Để giải quyết vấn đề này, một hướng đi mới đó là nghiên cứu và áp dụng kỹ thuật khai phá dữ liệu và khám phá tri thức trong môi trường Web. Do đó, việc nghiên cứu các mô hình dữ liệu mới và áp dụng các phương pháp khai phá dữ liệu trong khai phá tài nguyên Web là một xu thế tất yếu vừa có ý nghĩa khoa học vừa mang ý nghĩa thực tiễn cao.

Vì vậy chúng em chọn đề tài: “***Khai phá xu hướng mua sắm từ dữ liệu siêu thị ứng dụng luật kết hợp Apriori”***, để làm báo cáo kết thúc môn học của mình

**Báo cáo gồm 4 chương:**

Chương 1: Tổng quan về khai phá dữ liệu.

Chương 2: Tiền xử lý dữ liệu.

Chương 3; Khai phá dữ liệu bằng luật kết hợp Apriori.

Chương 4. Kết luận và hướng phát triển.

**Phân công công việc:**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

# **CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU**

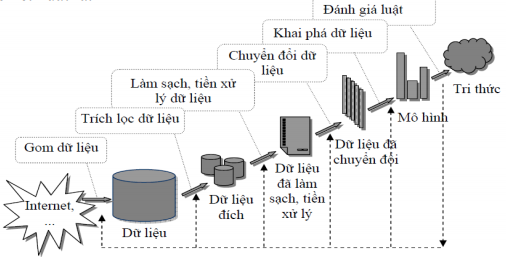
* 1. **Phát hiện tri thức và khai phá dữ liệu.**

Phát hiện tri thức (*Knowledge Discovery*) trong các cơ sở dữ liệu là một qui trình nhận biết các mẫu hoặc các mô hình trong dữ liệu với các tính năng: hợp thức, mới, khả ích, và có thể hiểu được.

Khai phá dữ liệu (Data mining) được định nghĩa như sau: “*Data mining là một quá trình tìm kiếm, phát hiện các tri thức mới, tiềm ẩn, hữu dụng trong CSDL lớn*”.

Khai phá dữ liệu có thể được sử dụng cho các lĩnh vực y tế, phân tích thị trường, xây dựng ... có thể được xem như là kết quả của sự tiến triển tự nhiên của công nghệ thông tin.

**1.2 Quy trình khám phá tri thức trong CSDL.**



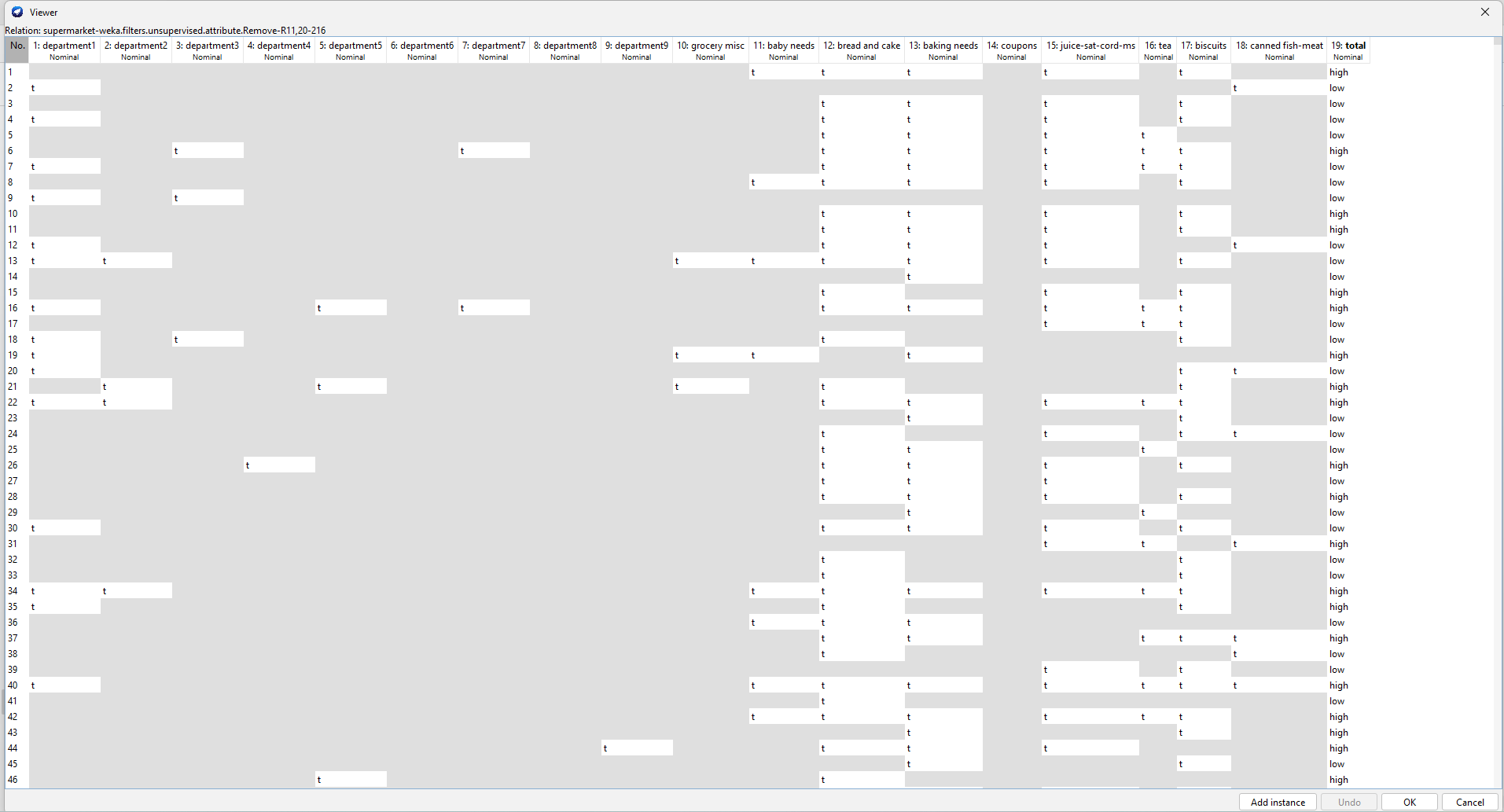
* 1. **Mô tả bài toán chuẩn đoán bệnh.**

**1.3.1 Tổng quan bài toán.**

* Dataset gồm các mô tả về các thuộc tính tương ứng với các món hàng trong siêu thị. Áp dụng khai phá luật kết hợp Apriori để khai phá dữ liệu.

**1.3.2 Phân tích dữ liệu thô.**

* *Hiểu dữ liệu:* Dữ liệu xu hướng mua sắm của một siêu thị với các số liệu về các món hàng
* *Dữ liệu gồm:* Dữ liệu bao gồm 4627 bản ghi cùng 19 thuộc tính



+ Các thuộc tính: deparment[1-9], grocery misc, baby needs, bread and cake, baking needs, coupons, juice-sat-cord-ms, tea, biscuits, total.

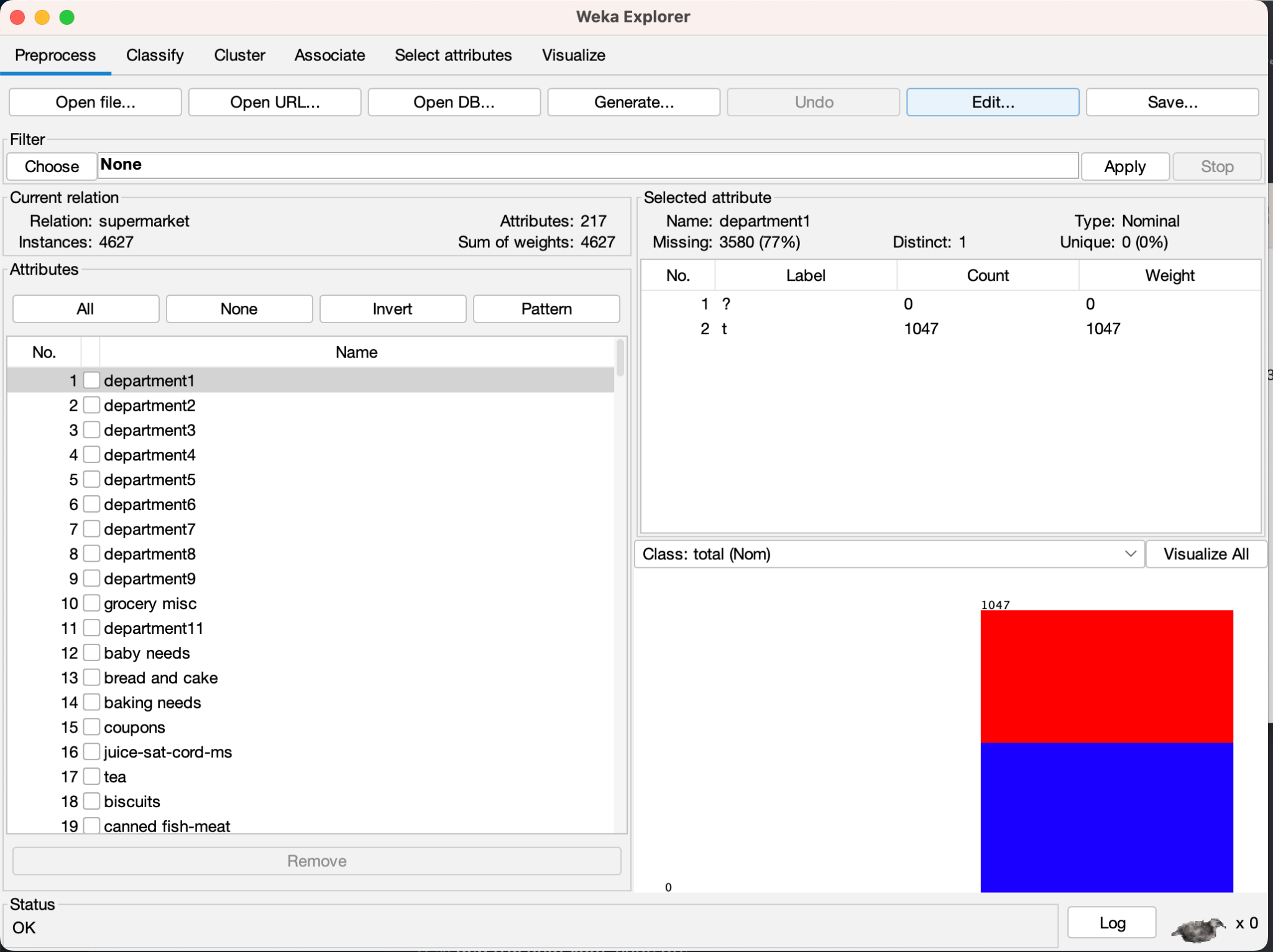
# **CHƯƠNG 2: TIỀN XỬ LÝ DỮ LIỆU**

## **2.1 Làm sạch dữ liệu.**

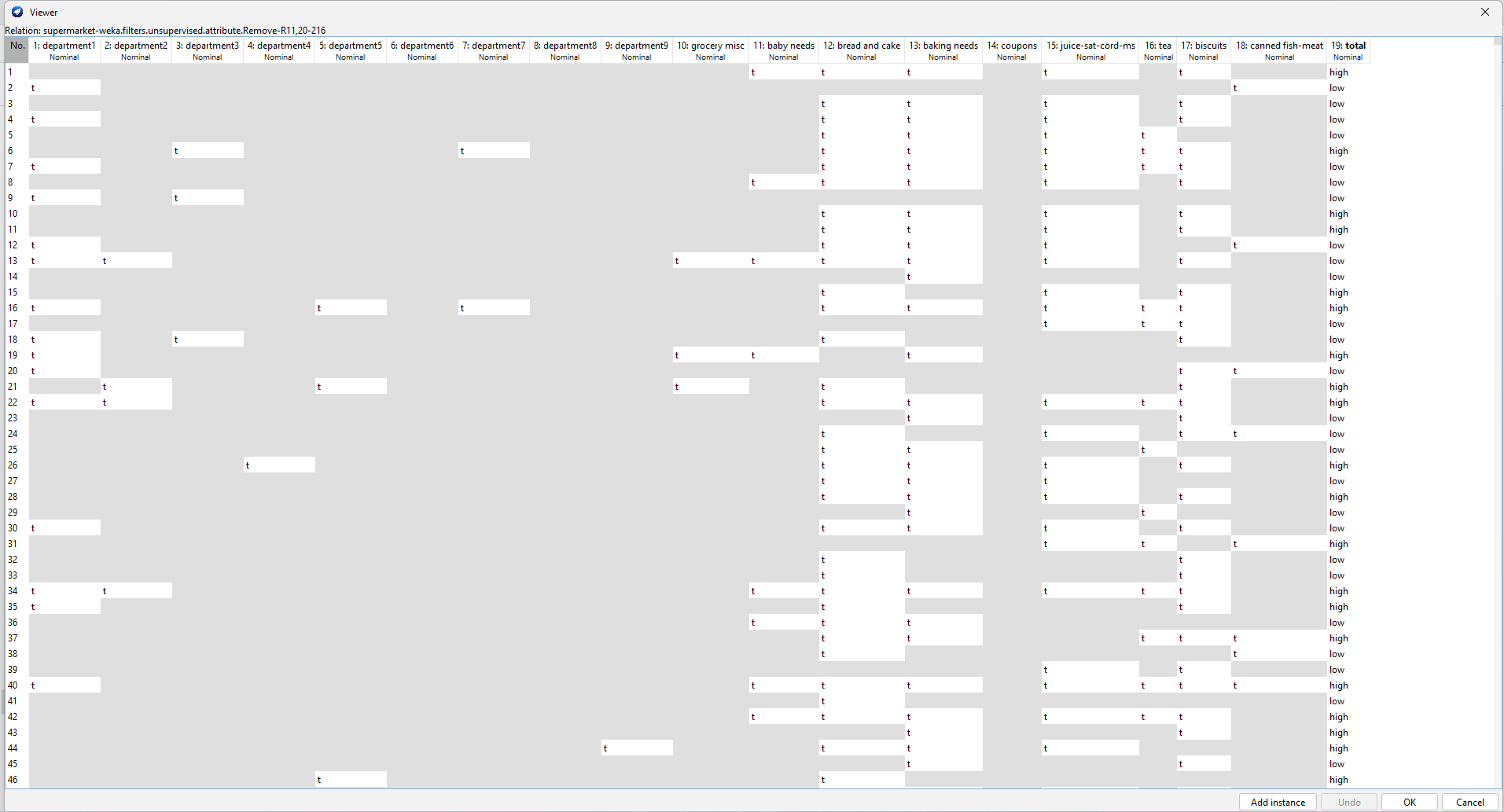
Là quá trình nhận dạng dữ liệu đã có để tiến hành xử lý các dữ liệu bị thiếu (missing data) xử lý dữ liệu bị nhiễu (noisy data) và không nhất quán. [2]

1. Xử lý dữ liệu bị thiếu (missing data)
2. Xử lý dữ liệu nhiễu, không nhất quán (inconsistent data).

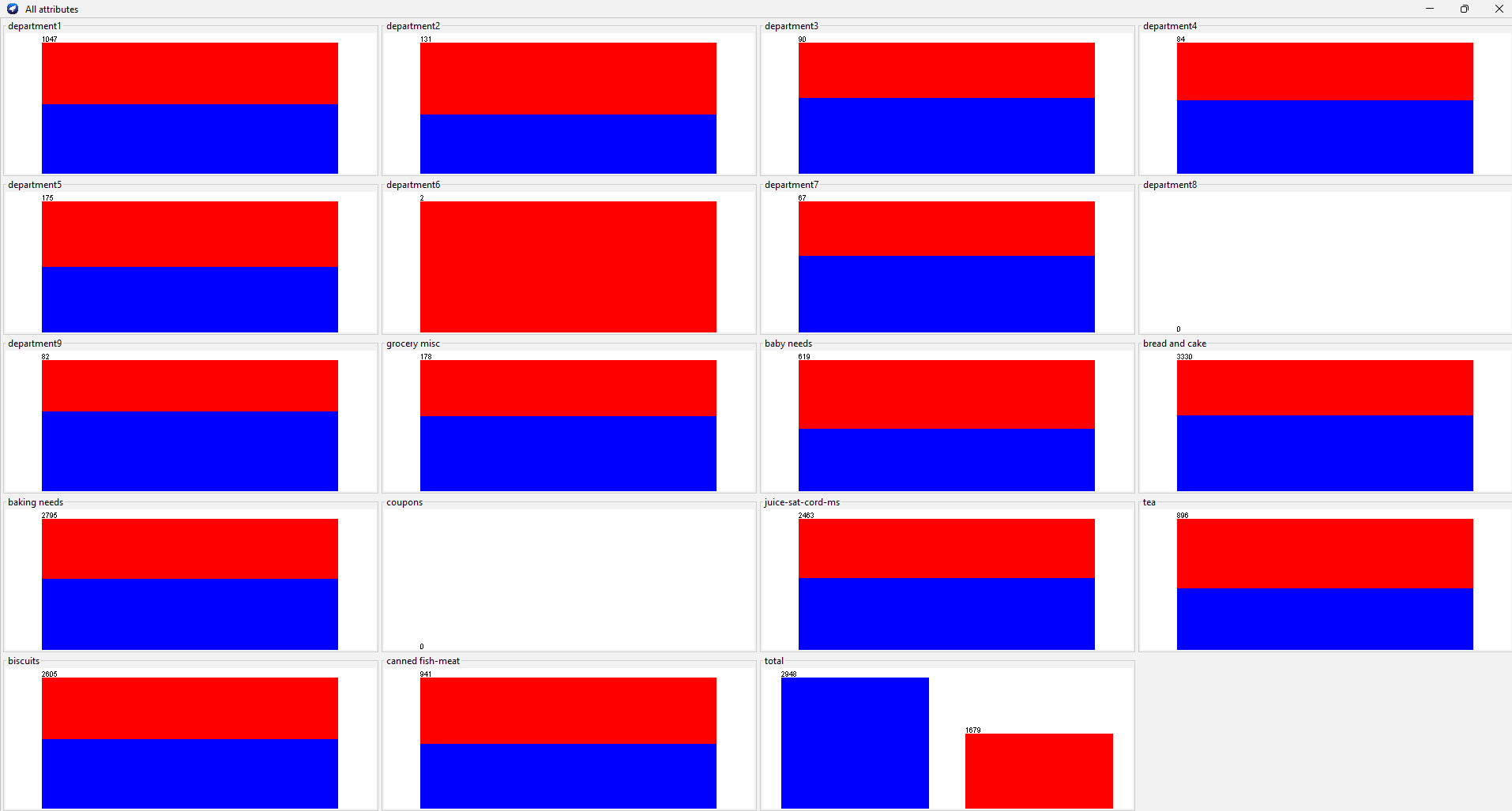
Ta thấy có các trường dữ liệu bị thiếu



**Nhận xét** : Dữ liệu có 2 dạng ? và t . ? này không phải là dữ liệu bị thiếu mà ? tương đương với dữ liệu không mua . t tương đương với dữ liệu đã mua . Vì vậy ta không cần tiền xử lí dữ liệu bị thiếu . Ta sẽ giữ nguyên tập dữ liệu như ban đầu và bắt đầu triển khai tiếp .



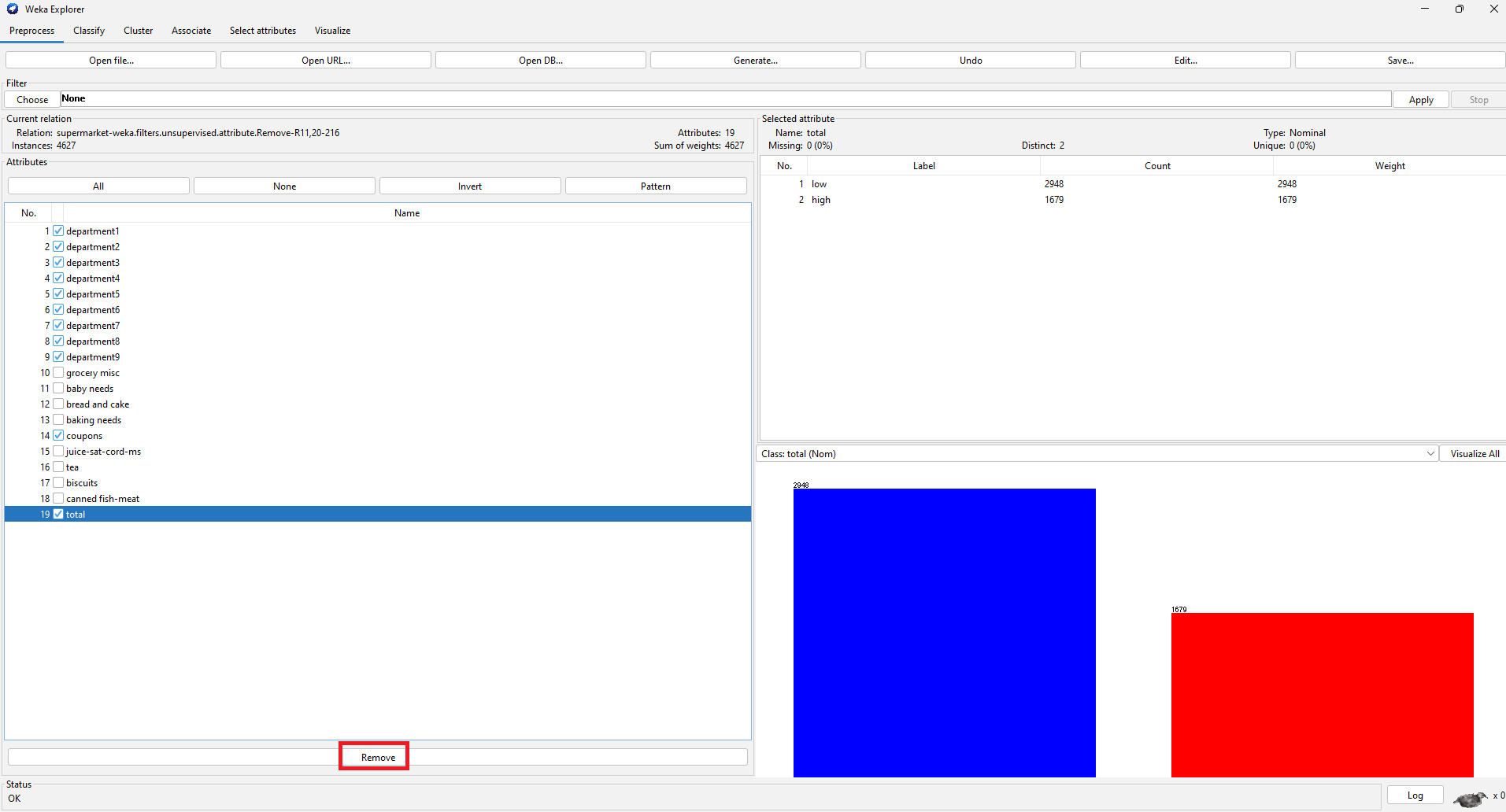
Sơ đồ của tất cả các thuộc tính qua visualize all trên weka



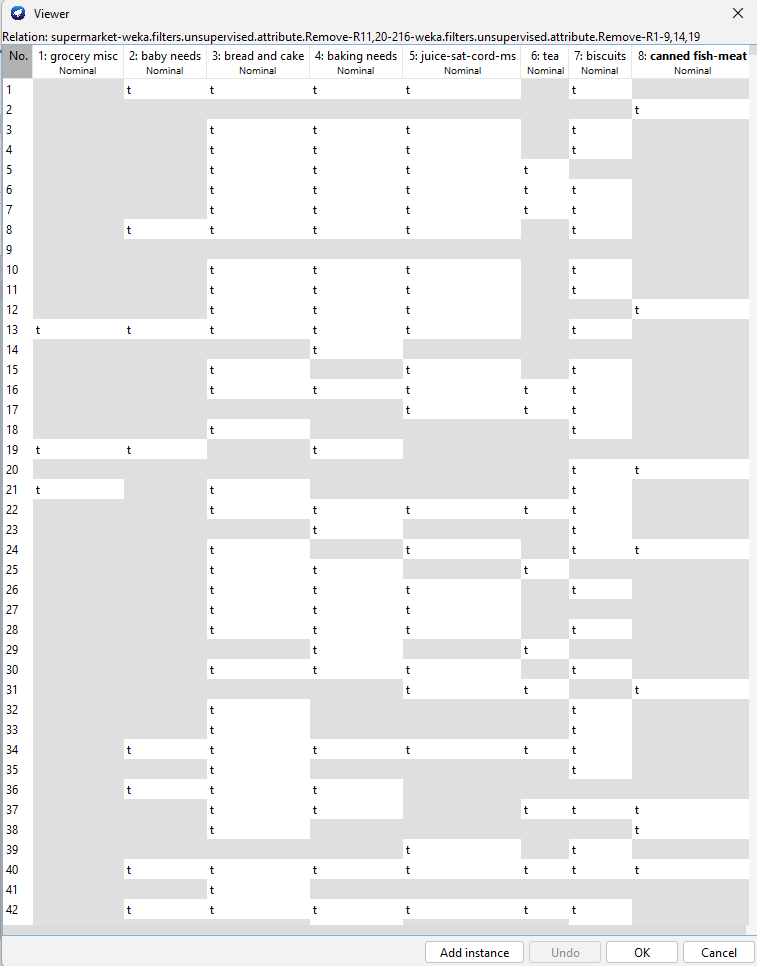
=> Department là các trường gian hàng không xét đến và thuộc tính coupons là thuộc tính rỗng , nên ta sẽ xoá đi .

## **2.2 Xóa dữ liệu không cần thiết**

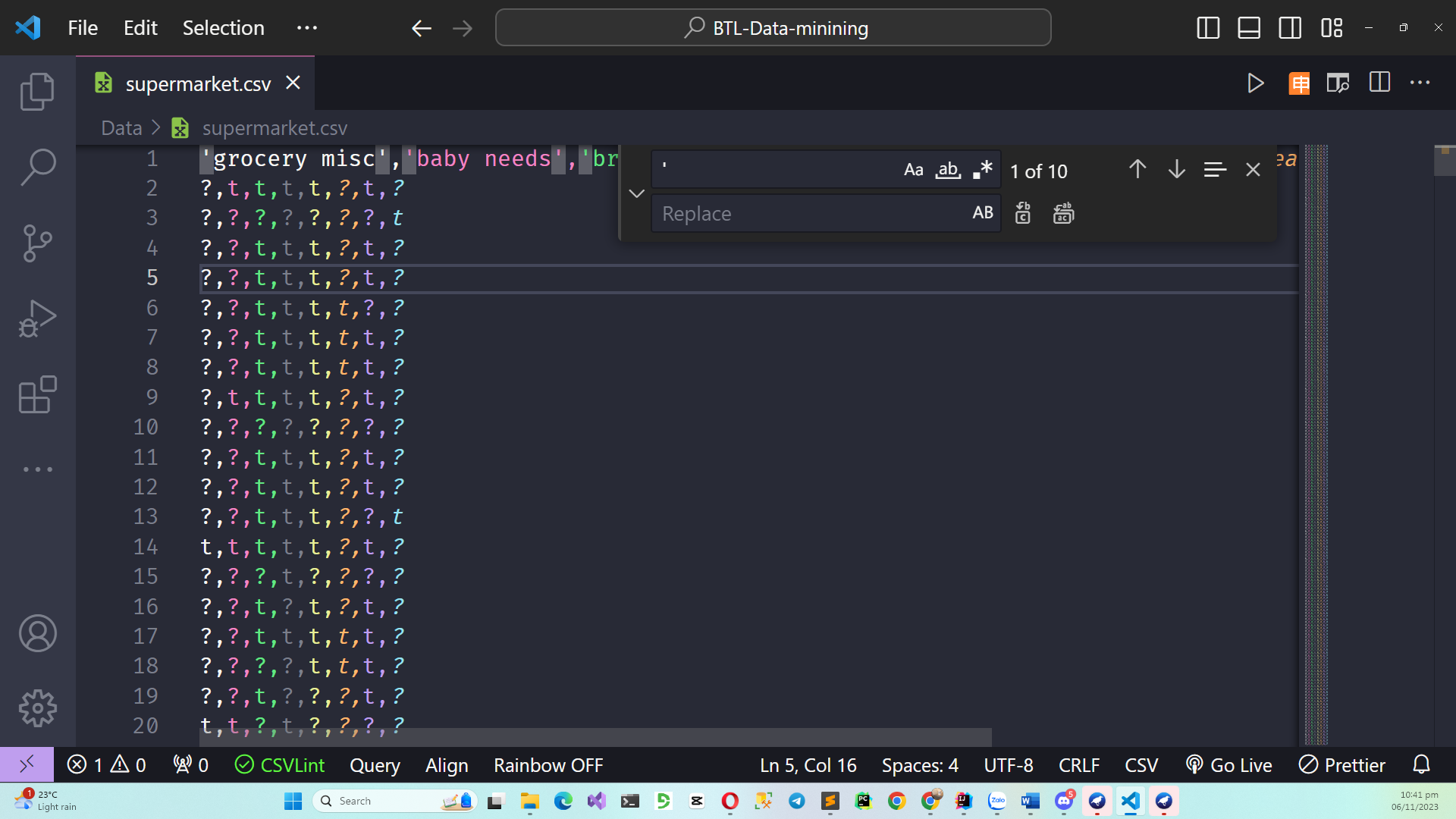
* Có 11 trường dữ liệu không cần thiết là department[1-9], Coupons và Total.



- Data sau khi tiền xử lý:



- Chuẩn hóa label :



# **CHƯƠNG 3: KHAI PHÁ DỮ LIỆU BẰNG LUẬT KẾT HỢP APRIORI**

## **3.1 Giới thiệu về luật kết hợp.**

* **Khái niệm:**

- Cho X, Y ⊆ I, X ∩ Y = ∅

- Định nghĩa: Luật kết hợp ký hiệu là X => Y chỉ ra mối ràng buộc của tập mục Y theo tập mục X, nghĩa là khi X xuất hiện trong D thì sẽ kéo theo sự xuất hiện của Y với một tỉ lệ nào đấy.

- Độ phổ biến (support) của một tập mục X trong CSDL D.

+ Ký hiệu supp(X)

+ Là tỉ lệ phần trăm giữa số giao dịch chứa X trên tổng số các giao dịch trong D: 𝑆𝑢𝑝𝑝 𝑋 =

+ Độ phổ biến của một luật kết hợp X => Y, ký hiệu 𝑆𝑢𝑝𝑝 𝑋 ⇒ 𝑌 , là độ phổ biến của tập mục 𝑋 ∪ 𝑌: 𝑆𝑢𝑝𝑝 𝑋 ⇒ 𝑌 = Supp(𝑋 ∪ 𝑌)

**-** Độ tin cậy (confidence): là tỉ lệ các giao dịch có chứa cả X và Y so với các giao dịch có chứa X: 𝐶𝑜𝑛𝑓 𝑋 ⇒ 𝑌 =

- Các luật có độ phổ biến lớn hơn ngưỡng minsupp và độ tin cậy lớn hơn ngưỡng minconf cho trước gọi là các luật mạnh (strong association rule) hay “luật có giá trị”:

+ supp(X => Y) ≥ minsupp và conf(X => Y) ≥ minconf

* **Tính chất của luật kết hợp:**

Tính chất 1: Nếu X ⊆ Y với X, Y là các tập mục thì supp(X) >= supp(Y)

=> hiển nhiên vì tất cả các giao dịch trong D hỗ trợ Y thì cũng hỗ trợ X

Tính chất 2: Một tập chứa một tập không phổ biến thì cũng là tập không phổ biến

- Nếu supp(X) < minsupp thì tập Y chứa X cũng không là tập phổ biến vì: supp(Y) <= supp(X) < minsup (theo tính chất 1)

Tính chất 3: Các tập con của tập phổ biến cũng là tập phổ biến

- Nếu Y là tập phổ biến trong D tức: supp(Y) >= minsup => mọi tập con A của Y cũng là phổ biến vì supp(A) >= supp(Y) >= minsup (theo tính chất 1)

- Trường hợp đặc biệt, nếu tập A= {i1,i2,..., ik} là tập phổ biến thì mọi tập con có (k-1) mục của nó cũng là phổ biến => ngược lại không đúng

## **3.2 Phương pháp luật kết hợp Apriori**

* Ý tưởng thuật toán Apriori**:**

- Thuật toán tỉa bớt những tập ứng cử viên có tập con không phổ biến trước khi tính độ hỗ trợ

- Phát hiện các tập mục phổ biến qua nhiều lần duyệt dữ liệu

- Kiểm tra tập 1-mục:

+ B1: Tính độ phổ biến của các tập mục, mỗi tập chỉ gồm 1 mục (thuộc tính) riêng biệt

+ B2: Với mỗi tập, kiểm tra có thỏa mãn điều kiện tập phổ biến (thỏa mãn độ hỗ trợ cực tiểu)

- Kiểm tra tập k-mục (k > 1):

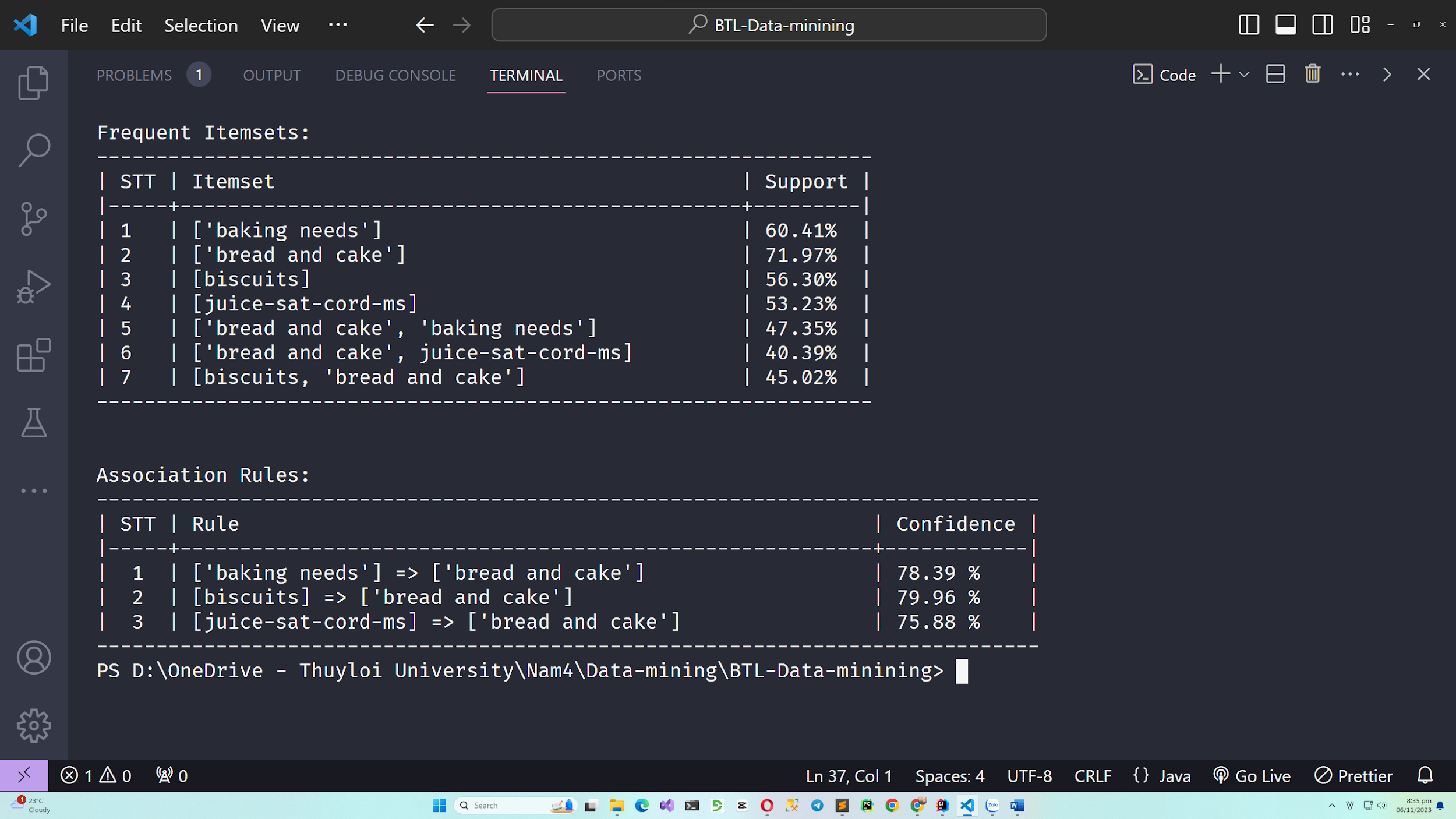
+ B1: Nếu k > tổng số mục => Kết thúc thuật toán: Các tập phổ biến có (k-1) mục

+B2: Sinh các tập mục ứng viên k-mục khác nhau dựa trên các tập phổ biến (k-1)-mục – tìm thấy trong lần duyệt (k-1)

+ B3: Tính độ hỗ trợ thực sự của các tập mục ứng cử k mục này và kiểm tra có thỏa mãn điều kiện tập phổ biến (thỏa mãn độ hỗ trợ cực tiểu)

**3.3 Thực hiện Apriori trên java**

Kết quả thu được khi thực hiện code trên Java với minsupp = 0,4 , minconf = 0,7



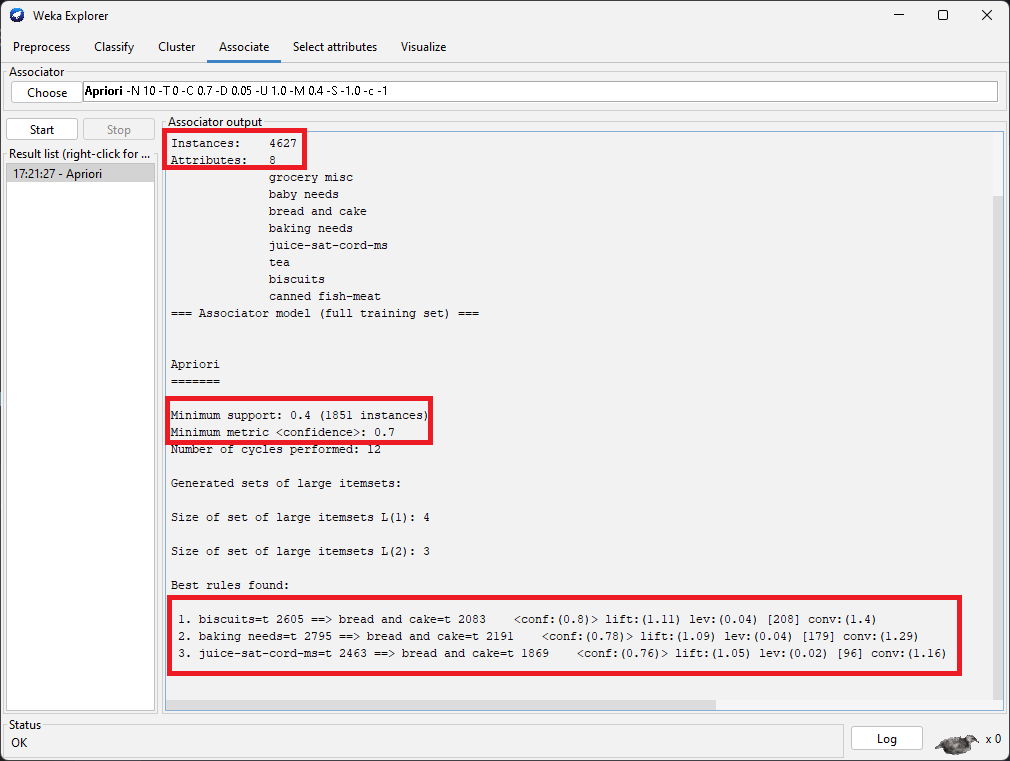
=> Đánh giá : Với ngưỡng minsupp = 0.4 và minconf = 0.7

1. 78% người mua ‘Baking need’ thì cũng mua ‘bread and cake’

2. 80% người mua ‘biscuits’ thì cũng mua ‘bread and cake’.

3. 76% người mua ‘juice-sat-cord-ms’ thì cũng mua ‘bread and cake’.

- Kiểm tra lại với phần mềm weka sử dụng luật kết hợp Apriori ta có kết quả:



-> Kết quả trùng khớp, vậy thuật toán code luật kết hợp Apriori trên java đã chính xác.

# **CHƯƠNG 4: KẾT LUẬN**

## **3.1 Kết luận.**

Thuật toán kết hợp Apriori là một trong những phần quan trọng của phân tích luật kết hợp (association rule mining) trong lĩnh vực học máy và khai thác dữ liệu. Nó được sử dụng để tìm kiếm các mẫu hoặc luật kết hợp trong dữ liệu, giúp tìm hiểu mối quan hệ giữa các mục hoặc sự kiện trong tập dữ liệu. Apriori có nhiều ứng dụng quan trọng trong thực tế, bao gồm: Y tế, thương mại, …

Chính vì vậy nhóm em đã hoàn thành đề tài “*Khai phá xu hướng mua sắm từ dữ liệu siêu thị ứng dụng luật kết hợp Apriori*”. **Nhóm em đã đạt được một số kết quả như sau:**

* Tìm hiểu tổng quan về khai phá dự liệu, bài toán luật kết hợp, phương pháp luật kết hợp Apriori để từ đó xây dựng mô hình khai phá xu hướng mua sắm.
* Thu thập dữ liệu các loại hàng hóa, tiền xử lý dữ liệu bằng excel và weka. Xây dựng nên mô hình luật kết hợp Apriori trên phần mềm weka.

**Tuy nhiên bài tập nhóm vẫn còn một số hạn chế:**

* Việc thu thập dữ liệu chưa đầy đủ, chi tiết với nhiều loại hàng hóa.
* Kết quả dự đoán tương đối cao nhưng vẫn chưa được tốt nhất.
  1. **Hướng phát triển.**
* Xây dựng, cải tiến mô hình khai phá xu hướng mua sắm bằng các phương pháp học máy khác như FP-growth, …
* Áp dụng mô hình học máy vào một số lĩnh vực khác và đi vào áp dụng thực.

Trong quá trình hoàn thành bài tập lớn, nhóm em đã cố gắng tìm hiểu và tham khảo các tài liệu liên quan. Tuy nhiên, thời gian có hạn nên chúng em sẽ không tránh khỏi những thiếu sót, rất mong nhận được sự đóng góp ý kiến của quý thầy cô và các bạn để báo cáo và kỹ năng của chúng em ngày được hoàn thiện hơn và có thế áp dụng được trong thực tiễn.

# **TÀI LIỆU THAM KHẢO**

1. "Data Mining: Concepts and Techniques" by Jiawei Han, Micheline Kamber, and Jian Pei

2. Introduction to Data Mining" by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar

3 Research Papers by R. Agrawal and R. Srikant

4. Online Courses and Tutorials