**Phần weka**  
- Bài toán là phân tích mức độ hài lòng khách hàng dựa trên phân lớp  
- Phân biệt phân lớp và phân cụm:   
\* phân lớp có giám sát và không giám sát ( có nhãn và không nhãn )

Có giám sát => có tiêu chí để đánh giá => y nhãn => lịch sử quá khứ hoặc khảo sát  
không giám sát ngược lại

B1: Xác định dữ liệu  
- Dữ liệu thô  
- Tại sao nó là dữ liệu thô?   
1.Chưa qua xử lý  
2.Chưa được chuẩn hóa: Các giá trị thuộc tính không được chuyển đổi vào một miền giá trị nhất định và không được được định nghĩa trước   
3.Có nhiều dữ liệu nhiễu: là dữ liệu ảnh hưởng đến % của thuật toán khi chạy  
4.Không được nhất quán ví dụ: 1,8m 180cm = nhau nhưng mà 2 cách ghi khác nhau  
  
B2: Trích chọn dữ liệu

* Tại sao ID là dữ liệu dư thừa vì nó là thuộc tính duy nhất ko sử dụng cho để đánh giá dữ liệu nào hay ảnh hưởng đến dữ liệu nào

B3: Tiền xử lý dữ liệu  
- Cách xử lý dữ liệu nhiễu (MathExpression):   
+ Thay đổi các khoảng dữ liệu để chia các nhóm vào dữ liệu nhau vào nhau loại bỏ các missing test bằng cách chạy thuật toán để xem sự thay đổi của % trong thuật toán.   
+ Hồi quy : pp1: kẻ vuông góc từ 1 điểm kẻ vuông góc trục kia và kéo các trục vuông góc kia lại  
- Tại sao lại chia dữ liệu:  
+ Bài toán dùng là Naive Bayes và Apriori:  
- Naive Bayes là mối quan hệ giữa các thuộc tính   
nó không biểu hiện được các giá trị của thuộc tính đối với lớp  
- Apriori là

Giá trị liên tục không phù hợp với Apriori vì nó ko có miền giá trị mang ý nghĩa cụ thể. Quy về chuẩn chung để quy tập phổ biến

Tại vì là thương mại điện tử : không có độ chuẩn cụ thể: chỉ có Age là có độ chuẩn cụ thể còn lại tự chia ( vì khi chia thấy khoảng đấy đẹp )   
B4. Tích hợp dữ liệu

* Là quá trình trộn dữ liệu từ các nguồn khác nhau vào một kho dữ liệu sẵn sàng cho quá trình khai phá dữ liệu

B5: **Biến đổi dữ liệu (chuẩn hóa)**

* Chuyển tất cả dữ liệu về nominal để
* Normal : dữ liệu kiểu phân lớp
* String To Nominal: kiểu dữ liệu phân lớp
* Numeric: số
* **NumbericToNominal: số -> nominal  
    
    
  P(Xi|C): xác suất giá trị thuộc tính với mỗi lớp   
  P(P): tỷ lệ của lớp**

**Tri thức suy ra từ Apriori để làm gì:**

* Chiến lược tiếp thị: Nếu bạn đang làm việc trong lĩnh vực tiếp thị, bạn có thể sử dụng tri thức này để phân đoạn khách hàng và tạo ra các chiến dịch tiếp thị phù hợp. Ví dụ, nếu bạn muốn tiếp cận đối tượng khách hàng trẻ tuổi và tăng cường mức độ hài lòng của họ, bạn có thể tạo ra các chương trình khuyến mãi hoặc ưu đãi đặc biệt dành riêng cho nhóm này.
* Tối ưu hóa trải nghiệm khách hàng: Tri thức này có thể giúp bạn hiểu hơn về hành vi và sở thích của khách hàng. Với thông tin về giới tính, tuổi, mức độ hài lòng và mức chi tiêu, bạn có thể tùy chỉnh trải nghiệm khách hàng, cung cấp dịch vụ và sản phẩm phù hợp với từng nhóm khách hàng cụ thể.
* Quản lý tài chính cá nhân: Nếu bạn đang quản lý tài chính cá nhân, tri thức này có thể giúp bạn hiểu rõ hơn về mối quan hệ giữa mức chi tiêu, tuổi và giới tính. Bạn có thể sử dụng thông tin này để xác định các mục tiêu tài chính và phân bổ nguồn lực phù hợp cho từng nhóm khách hàng.
* Định hướng sản phẩm: Nếu bạn là một nhà kinh doanh hoặc nhà phát triển sản phẩm, tri thức này có thể giúp bạn hiểu nhu cầu và mong muốn của khách hàng. Bằng cách tìm hiểu về mối quan hệ giữa các yếu tố như tuổi, giới tính, mức chi tiêu và hài lòng, bạn có thể phát triển sản phẩm phù hợp và tạo ra giá trị cho khách hàng.