**CHƯƠNG 2: PHÂN TÍCH TẬP THÔ**

**2.1 Phân tích bài toán:**

Input:

- Nhập data từ file csv chứa dữ liệu đã chuẩn bị trước đó.

- Data có dạng: cột 0: chứa ID, cột 1 đến cột n-1: là các cột chứa điều kiện, cột n là cột chứa giá trị quyết định

-Data có thể tự tạo hoặc chuẩn hóa từ các dataset có sẵn

-Người dùng tự lựa chọn tập điều kiện B và tập giá trị X trên giao diện GUI của chương trình

Output:

-Tìm phục thuộc của tập C (tập quyết định) vào tập B

-Tìm xấp xỉ của tập X qua B

-Tìm các reduct rút gọn nếu có

**2.2 Các giải thuật:**

***2.2.1 Tìm xấp xỉ phụ thuộc tập C vào B***

Giả sử C= X1 U X2 U…Xn



Nếu : C phụ thuộc vào B

Nếu : C không phụ thuộc vào B

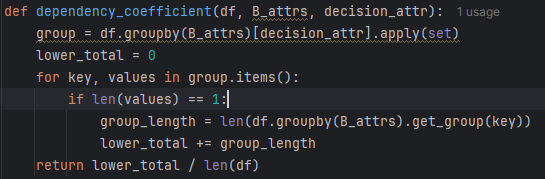
Ý tưởng giải thuật:

-Tìm tập IND(B) bằng cách nhóm các đối tượng theo tập điều kiện

-Duyệt qua tập IND(B), nếu ứng với một giá trị trong tập IND(B) có đến >1 giá trị tương ứng trong tập C (nghĩa là các điều kiện giống nhau nhưng quyết định khác nhau) thì giá trị đó thuộc vùng biên, vùng không ảnh hưởng đến quyết định nên sẽ bỏ qua

-Các giá trị phù hợp còn lại trong tập IND(B) sẽ giúp ta tính được 

Qua đó tính được hệ số *k* và rút ra kết luận phụ thuộc của C vào B



***2.2.2 Tìm xấp xỉ tập X qua B:***

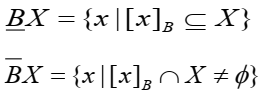
Muốn tìm xấp xỉ tập X qua B ta cần tìm Lower(B,X) và Upber(B,X) qua đó tính độ chính xác *a:*

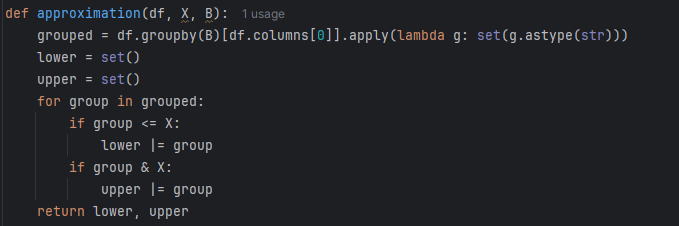


*α* = 1: X là rõ so với B

*α* < 1: X là thô so với B

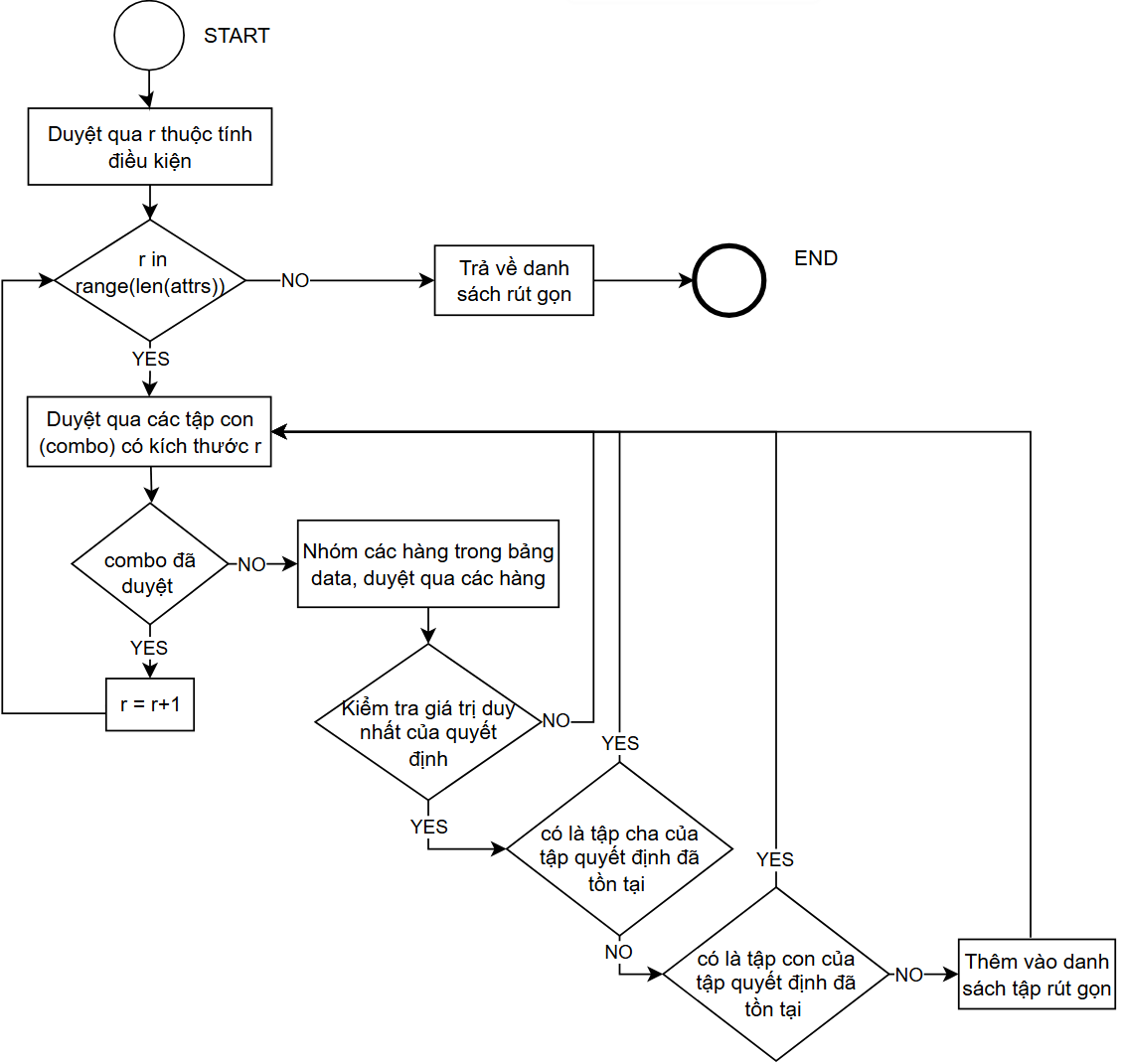
Ta viết code trực tiếp dựa trên định nghĩa của *Lower(B,X)* và *Upber(B,X)* :



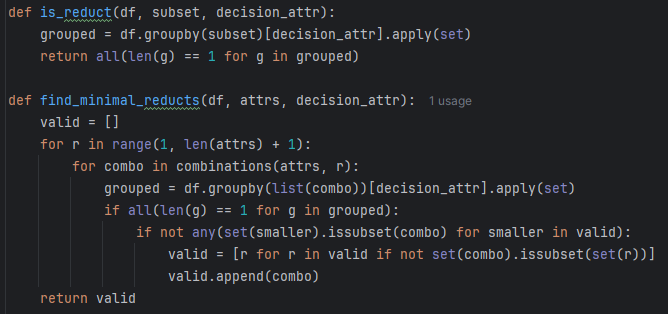


***2.2.3 Tìm các reduct rút gọn nếu có:***

Sơ đồ giải thuật:

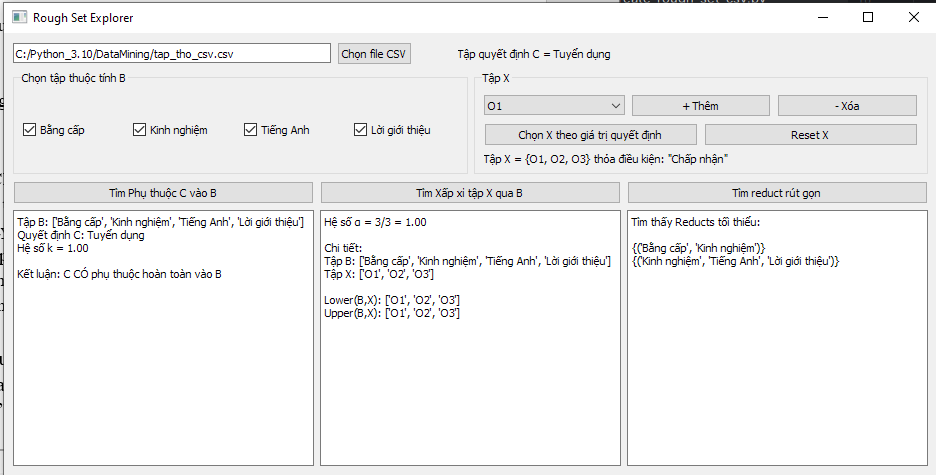


Code chương trình:



**2.3 Tạo GUI và chạy thử chương trình:**

***2.3.1 Giao diện của chương trình:***



Giao diện khá đơn giản dễ dùng, người dùng chỉ cần chọn thuộc tính B cần thiết, chọn tập X thủ công hoặc dựa theo giá trị của tập quyết định, sau đó chạy chương trình để giải các bài toán tương ứng. Kết quả hiển thị khu vực bên dưới mỗi nút chức năng.

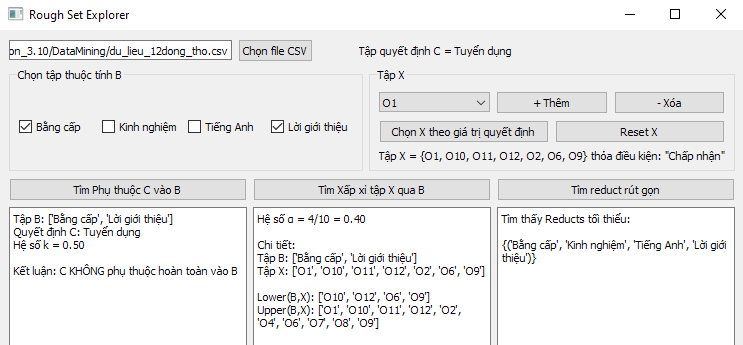
***2.3.2 Chạy thử chương trình:***

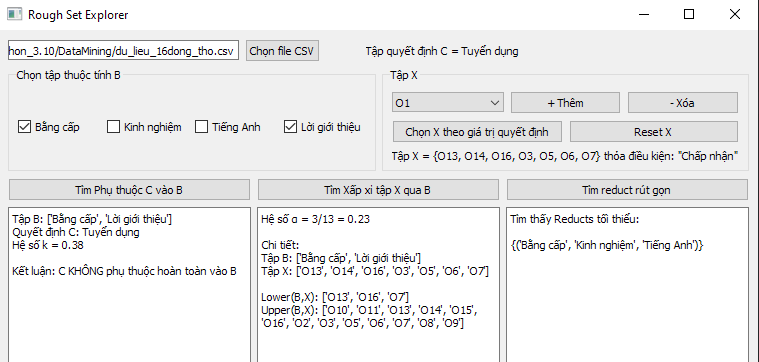
-Chạy thử với bài toán tuyển dụng cho kết quả giống như tính thủ công

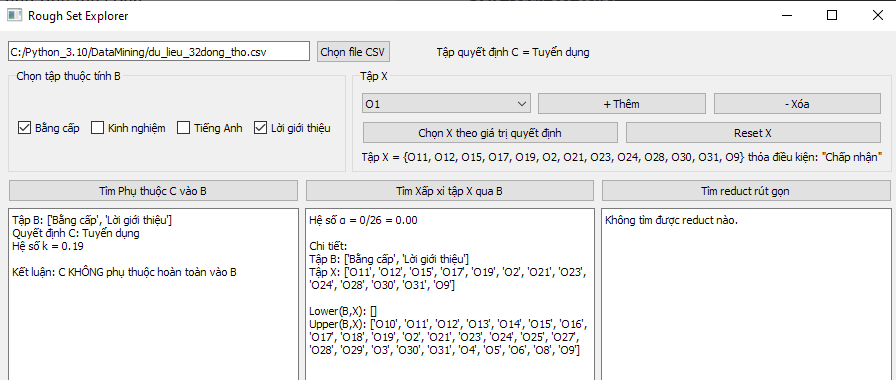
-Từ data gồm 8 dòng của bài toán tuyển dụng, thử trộn lẫn và nhân bản lên thành data 12 dòng, 16 dòng, 32 dòng… để kiểm tra tính chất tập thô

-Nhận thấy khi số dòng tăng lên thì phụ thuộc của C vào B và xấp xỉ X qua B đều giảm.

-Lý giải: do tập B được sinh ra từ tổ hợp đã có chứ không phải dữ liệu mới nên sẽ tạo ra nhiều giá trị trùng lắp khiến tập B “thô” hơn, tập điều kiện lúc này không thể phân biệt rõ tập quyết định.







-Để kiểm tra thử thuật toán của chương trình trên bộ dữ liệu lớn, nhiều dòng mà không thể tính thủ công ta chọn 2 tập dataset :

**Acute Inflammations dataset**

Nguồn: UCI Machine Learning Repository  
<https://archive.ics.uci.edu/ml/datasets/Acute+Inflammations>

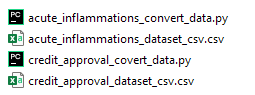
| **Thông tin** | **Chi tiết** |
| --- | --- |
| Số dòng | 120 bệnh nhân |
| Số thuộc tính | 6 thuộc tính + 2 thuộc tính quyết định |
| Dạng thuộc tính | Boolean (yes/no) |
| Loại dữ liệu | Phân loại đơn giản |
| Mục tiêu | Dự đoán viêm bàng quang hoặc viêm thận |

**Credit Approval dataset**

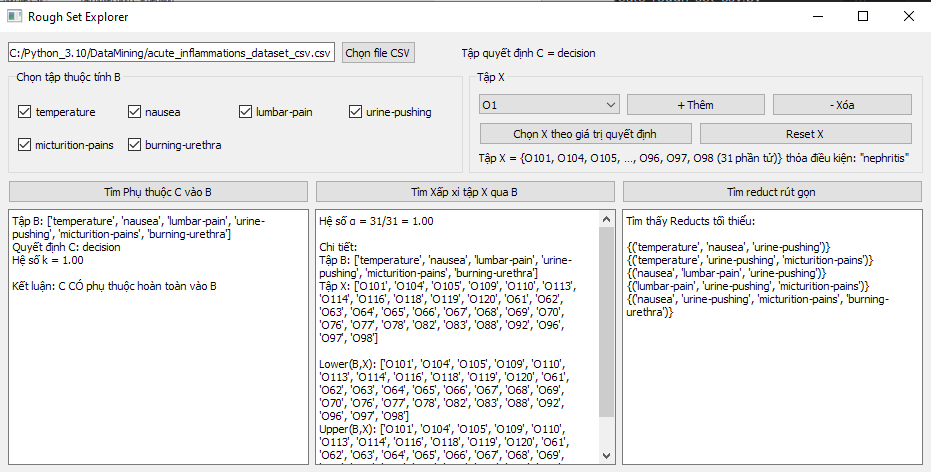
Nguồn: UCI Machine Learning Repository  
<https://archive.ics.uci.edu/ml/datasets/Credit+Approval>

| **Thông tin** | **Chi tiết** |
| --- | --- |
| Số dòng | 690 yêu cầu cấp thẻ tín dụng |
| Số thuộc tính | 15 thuộc tính + 1 quyết định |
| Loại thuộc tính | Hỗn hợp: số thực + phân loại |
| Loại dữ liệu | Thực tế, có nhiễu (missing, ?...) |
| Mục tiêu | Phê duyệt hoặc từ chối hồ sơ xin thẻ tín dụng |

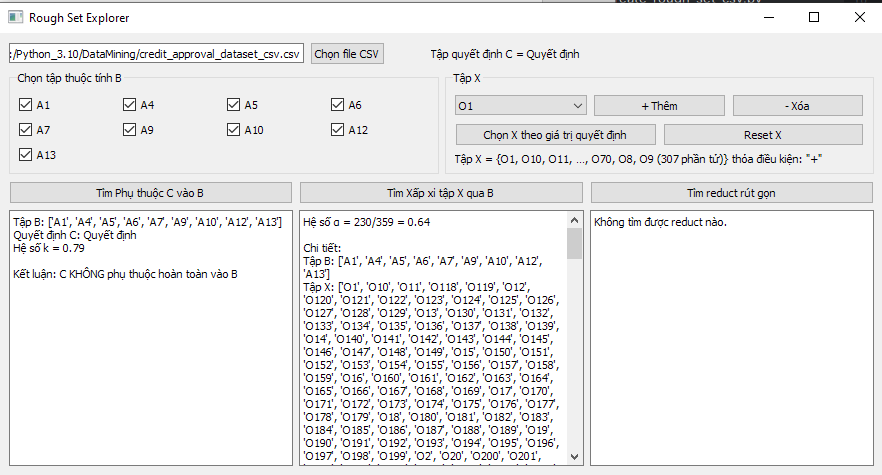
-Để cho phù hợp với input của chương trình ta viết các hàm để chuẩn hóa dữ liệu từ tập dataset rồi lưu lại dạng file csv



-Kết quả chạy thử file “acute\_inflammations\_dataset\_csv.csv”:



-Kết quả chạy thử file “credit\_approval\_dataset\_csv.csv”:

  
-Nhận xét: với tập Acute Inflammations dataset với 120 dòng cùng 6 cột điều kiện có cấu trúc dữ liệu chuẩn, sạch nên khi chạy chương trình ra kết quả chuẩn.

Tập Credit Approval dataset với 690 dòng cùng 16 cột điều kiện có dữ liệu còn thô nên khi chạy chương trình cho kết quả chưa chuẩn.