

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 KHAI PHÁ
DỮ LIỆU VỚI PHẦN MỀM TANAGRA**

Giảng viên hướng dẫn: ThS. Nguyễn Thị Trần Lộc

Sinh viên thực hiện: Chung Hữu An – MSSV: 1821002711

Mã lớp HP: 2111112005903

TP. HCM, THÁNG 12 NĂM 2021

TRÍCH YẾU

Trong thời đại công nghiệp hiện nay, lợi thế chiến lược lớn nhất đến từ việc phân loại, sắp xếp, phân tích và khai thác dữ liệu từ mọi góc độ có thể. Tuy nhiên, không giống như tất cả các hoạt động liên quan đến dữ liệu, giá trị của các hoạt động khai thác dữ liệu được gắn trực tiếp với chất lượng và phạm vi dữ liệu có sẵn để khai thác, các doanh nghiệp còn phải tìm cách tối ưu hóa các thông tin và dữ liệu mình có được nhằm đạt được những chỉ tiêu, mong muốn mà doanh nghiệp đề ra. Chính vì lẽ đó việc khai phá dữ liệu hiệu quả, chính xác là một điều vô cùng trọng yếu và cần thiết trong việc vận hành một doanh nghiệp hiện nay. Vì hiểu được sự quan trọng của khai phá dữ liệu trong kinh doanh hiện nay, em đã quyết định thực hiện đề tài này, nhằm ứng dụng phần mềm vào việc khai phá dữ liệu, để có cái nhìn tổng quan hơn về việc khai phá dữ liệu và các hiệu quả nó mang lại. Phần mềm em chọn là Tanagra với thao tác đơn giản, dễ sử dụng trong việc thực hành và học tập, giúp em có những cái nhìn sơ lược đầu tiên về khai phá dữ liệu. Bên cạnh đó em chọn cơ sở dữ liệu Chẩn đoán Sức khỏe JPAC (Jayaramdas Patel Academic Centre) đã tiến hành khảo sát tình trạng suy gan cấp tính đối với người trưởng thành Ấn Độ (trên 20 tuổi) trên toàn đất nước, từ đó em sẽ có cái nhìn đa chiều hơn trong việc quản lý tình trạng người bị bệnh gan cấp tính ở Ấn Độ. Kết hợp tất cả yếu tố trên em đã hình thành ra đề tài : “**ỨNG DỤNG TANAGRA ĐỂ THỰC HIỆN KHAI PHÁ DỮ LIỆU KHẢO SÁT TÌNH TRẠNG SUY GAN CẤP TÍNH ĐỐI VỚI NGƯỜI TRƯỞNG THÀNH Ở ẤN ĐỘ**”.

MỤC LỤC

TRÍCH YẾU.....	i
MỤC LỤC	ii
LỜI CẢM ƠN	iv
DANH MỤC TỪ VIẾT TẮT	v
DANH MỤC THUẬT NGỮ ANH – VIỆT	vi
DANH MỤC CÁC HÌNH ẢNH.....	vii
DANH MỤC CÁC BẢNG BIỂU.....	x
DÃN NHẬP	Lỗi! Thẻ đánh dấu không được xác định.
❖ Mục tiêu của đồ án	Lỗi! Thẻ đánh dấu không được xác định.
❖ Phân công công việc	Lỗi! Thẻ đánh dấu không được xác định.
❖ Kế hoạch thực hiện đồ án.....	Lỗi! Thẻ đánh dấu không được xác định.
CHƯƠNG 1: TỔNG QUAN	1
1.1. Lý do hình thành đồ án	1
1.2. Mục tiêu đồ án	1
1.3. Dự kiến kết quả đạt được	1
CHƯƠNG 2: CƠ SỞ LÝ THUYẾT	3
2.1. Giới thiệu về khai phá dữ liệu.....	3
2.1.1 <i>Khái niệm</i>	3
2.1.2 <i>Vai trò của khai phá dữ liệu trong kinh doanh</i>	3
2.1.3 <i>Quy trình khai phá dữ liệu</i>	4
2.2. Kho dữ liệu.....	6
2.2.1 <i>Kiến trúc luồng dữ liệu</i>	6
2.2.2 <i>Kho dữ liệu và khai phá dữ liệu trong BI.....</i>	7
2.3. Các phương pháp trong khai phá dữ liệu	8
2.3.1 <i>Phương pháp phân lớp.....</i>	8
2.3.2 <i>Phương pháp gom cụm</i>	9
2.3.3 <i>Phương pháp luật kết hợp.....</i>	10
2.4. Giới thiệu về phần mềm sử dụng (Tanagra)	11
2.4.1 <i>Tổng quan về phần mềm Tanagra.....</i>	11
2.4.2 <i>Cách sử dụng phần mềm</i>	12

2.4.3	<i>Cách truyền dữ liệu vào chương trình</i>	15
2.4.4	<i>Cách thực hiện các thuật toán</i>	17
CHƯƠNG 3:	ỨNG DỤNG PHẦN MỀM TANAGRA	30
3.1.	Mô tả dữ liệu	30
3.1.1	<i>Mô tả chung</i>	30
3.1.2	<i>Mô tả chi tiết</i>	30
3.1.3	<i>Ưu điểm của cơ sở dữ liệu</i>	32
3.1.4	<i>Nhược điểm của cơ sở dữ liệu</i>	33
3.2.	Tiền xử lý dữ liệu	33
3.3.	Chuyển đổi dữ liệu thuộc biến liên tục (Discrete) thành dữ liệu thuộc biến rời rạc(Discrete)	34
3.4.	Xây dựng cây ra quyết định với Tanagra	38
3.4.1	<i>Cách thực hiện</i>	38
3.4.2	<i>Kết quả chạy thuật toán</i>	41
3.5.	Xây dựng thuật toán NaiveBayes với Tanagra	42
3.5.1	<i>Cách thực hiện</i>	43
3.5.2	<i>Kết quả chạy thuật toán</i>	46
3.6.	Xây dựng thuật toán gom cụm K-Means với Tanagra	46
3.6.1	<i>Cách thực hiện</i>	47
3.6.2	<i>Ý nghĩa kết quả của thuật toán K-Means</i>	51
3.7.	Xây dựng thuật toán luật kết hợp Apriori với Tanagra	52
3.7.1	<i>Cách thực hiện</i>	53
3.7.2	<i>Ý nghĩa kết quả của thuật toán Apriori</i>	57
CHƯƠNG 4:	KẾT LUẬN	59
4.1.	Những kết quả đạt được	59
4.2.	Hạn chế	59
TÀI LIỆU THAM KHẢO		60

LỜI CẢM ƠN

Đầu tiên em xin gửi lời cảm ơn đến cô **ThS. Nguyễn Thị Trần Lộc** – giảng viên khoa Công nghệ thông tin – Trường đại học Tài Chính Marketing đã tận tình giúp đỡ giảng dạy, giải đáp mọi thắc mắc của em trong suốt quá trình hoàn thành đồ án Khai Phá Dữ Liệu này.

Em xin bày tỏ lòng biết ơn sâu sắc đến cô đã tạo điều kiện tốt để em học tập và thuận lợi hoàn thành đồ án.

Tuy vậy, trong quá trình thực hiện kinh nghiệm của em còn hạn chế nên không tránh khỏi những thiếu sót trong quá trình hoàn thành, vì vậy em kính mong nhận được những ý kiến, những lời nhận xét chân thật nhất để em có thể bổ sung hoàn thiện đồ án của mình.

Cuối cùng chúc cô và toàn thể ban lãnh đạo nhà trường lời chúc sức khỏe và thành công trong sự nghiệp cuộc sống.

TP Hồ Chí Minh, tháng 12 năm 2021

Sinh viên thực hiện:

Chung Hữu An

DANH MỤC TỪ VIẾT TẮT

CNTT	Công nghệ thông tin
CSDL	Cơ sở dữ liệu
KPDL	Khai phá dữ liệu
PCDL	Phân cụm dữ liệu
OLAP	Online Analytical Processing

DANH MỤC THUẬT NGỮ ANH – VIỆT

Apriori	Thuật toán Apriori
Clustering Analysis	Phân tích theo cụm
Data mining	Khai phá dữ liệu
Decision tree	Cây ra quyết định
K-Means	Thuật toán K-Means
NaiveBayes	Thuật toán NaiveBayes
Streaming Systems	Hệ thống xử lý luồng

DANH MỤC CÁC HÌNH ẢNH

Hình 2. 1 Qui trình khai phá dữ liệu.....	4
Hình 2. 2 Phần mềm Tanagra	11
Hình 2. 3 Giao diện trang web Tanagra	13
Hình 2. 4 Nút Download	13
Hình 2. 5 Các phiên bản Tanagra	13
Hình 2. 6 Giao diện chính của phần mềm Tanagra	14
Hình 2. 7 Thanh công cụ của phần mềm Tanagra	15
Hình 2. 8 Khởi động Dataset	15
Hình 2. 9 Thiết lập nhập và xuất dữ liệu	16
Hình 2. 10 Bảng kết quả truyền dữ liệu vào	17
Hình 2. 11 Tạo Define cho cây ra quyết định	18
Hình 2. 13 Chọn Target cho cây ra quyết định	18
Hình 2. 14 Thêm thuật toán C-RT	19
Hình 2. 15 Xem kết quả thuật toán cây ra quyết định C-RT	20
Hình 2. 16 Tạo define cho NavieBayes.....	21
Hình 2. 17 Thêm thuật toán NavieBayes.....	22
Hình 2. 18 Xem kết quả thuật toán NavieBayes	23
Hình 2. 19 Tạo Define cho thuật toán K – Means.....	24
Hình 2. 20 Thêm component E UNIVARIATE CONT STAT	24
Hình 2. 21 Thêm component STANDARDIZE	25
Hình 2. 22 Tạo DEFINE 2	25
Hình 2. 23 Thêm thuật toán K – MEAN	26
Hình 2. 24 Xem kết quả thuật toán K – MEAN	26

Hình 2. 25 Tạo define cho luật kết hợp Apriori	27
Hình 2. 26 Thêm component A PRIORI MR	28
Hình 2. 27 Thiết lập thông số cho component A PRIORI MR	29
Hình 2. 28 Kết quả luật kết hợp Apriori.....	29
Hình 2. 29 Thêm thành phần Define Status vào Dataset Lỗi! Thẻ đánh dấu không được xác định.	
Hình 3. 1 Dữ liệu chưa được xử lý (1)	33
Hình 3. 2 Dữ liệu chưa được xử lý (2)	33
Hình 3. 3 Dữ liệu sau khi xử lý	34
Hình 3. 4 Thêm thành phần Define Status và Dataset.....	34
Hình 3. 5 Cho các dữ liệu cần chuyển đổi vào Input	35
Hình 3. 6 Kéo thả công cụ EqFreq Disc và Define	35
Hình 3. 7 Thiết lập khoảng phân rã dữ liệu	36
Hình 3. 8 Các dữ liệu đã được chuyển đổi	37
Hình 3. 9 Đẩy các giá trị sau khi chuyển đổi vào ô Input	38
Hình 3. 10 Thiết lập số khoản phân rã.....	39
Hình 3. 11 Thêm các thuộc tính cần thiết cho thuật toán(1)	39
Hình 3. 12 Thêm các thuộc tính cần thiết vào Input	40
Hình 3. 13 Thêm thuộc tính ALF vào Target.....	40
Hình 3. 14 Sử dụng công cụ End Tree	41
Hình 3. 15 Kết quả sau khi chạy thuật toán.....	41
Hình 3. 16 Kết quả chạy thuật toán (1)	42
Hình 3. 17 Kết quả chạy thuật toán 2	42
Hình 3. 18 Thiết lập số khoản phân rã.....	43

Hình 3. 19 Xác định các thuộc tính chạy thuật toán.....	44
Hình 3. 20 Thêm thành phần NAIVE BAYES (tab SPV LEARNING) vào sơ đồ.....	44
Hình 3. 21 Bấm vào menu View để nhận kết quả(1)	45
Hình 3. 22 Bấm vào menu View để nhận kết quả(2)	46
Hình 3. 23 Xác định thành phần cần phân tích	47
Hình 3. 24 Chèn thuật toán K – Means vào Define vừa tạo.....	48
Hình 3. 25 Thiết lập thông số cho thuật toán	49
Hình 3. 26 Kết quả của thuật toán gom cụm	50
Hình 3. 27 Mô hình hóa kết quả của thuật toán gom cụm Lỗi! Thẻ đánh dấu không được xác định.	
Hình 3. 28 Thiết lập số khoản phân rã.....	53
Hình 3. 29 Tạo thư mục Define để đồ dữ liệu cho component EqFreq Disc	54
Hình 3. 30 Thêm thuật toán A priori vào component EqFreq Disc	55
Hình 3. 31 Nhập thông số cho thuật toán	56
Hình 3. 32 Kết quả thuật toán (1) Lỗi! Thẻ đánh dấu không được xác định.	
Hình 3. 33 Kết quả thuật toán (2) Lỗi! Thẻ đánh dấu không được xác định.	

DANH MỤC CÁC BẢNG BIỂU

Bảng 1. Phân công công việc	Lỗi! Thẻ đánh dấu không được xác định.
Bảng 2. Kế hoạch thực hiện đề án	Lỗi! Thẻ đánh dấu không được xác định.
Bảng 3. 1 Mô tả chi tiết các thuộc tính.....	30

CHƯƠNG 1: TỔNG QUAN

1.1. Lý do hình thành đồ án

Trong thời đại 4.0, như hiện nay giới trẻ ngày càng có thú vui ăn chơi đặc biệt là uống các đồ uống có cồn như rượu, bia,... Làm ảnh hưởng rất lớn đến sức khỏe, cụ thể ở đây là bị suy gan, khiến nhiều người phải tử vong khi còn quá trẻ. Đây là một vấn đề cấp bách cần được nắm bắt chính xác các thông tin về nó, nhằm đưa ra các giải pháp để ngăn chặn tình trạng này. Nắm bắt được sự cấp thiết của việc này, em quyết định chọn đề tài “**ỨNG DỤNG PHẦN MỀM KHAI PHÁ TANAGRA VÀO PHÂN TÍCH DỮ LIỆU TÌNH TRẠNG SUY GAN CẤP TÍNH ĐỐI VỚI NGƯỜI TRƯỞNG THÀNH Ở ÂN ĐỘ**” nhằm phân tích chuyên sâu, cụ thể hơn về vấn đề suy gan ở độ tuổi quá trẻ. Qua đó có được một góc nhìn tổng quan về vấn đề này.

1.2. Mục tiêu đồ án

- Khái quát các khái niệm, vai trò của khai phá dữ liệu trong kinh tế, mô hình khai phá dữ liệu
- Vận dụng các phương pháp trong khai phá dữ liệu: phương pháp hồi quy tuyến tính, phương pháp gom cụm, phương pháp khai phá luật kết hợp bằng cây ra quyết định, phương pháp khai phá luật kết hợp vào bài toán kinh tế.
- Áp dụng thành thạo công cụ Tanagra trong việc triển khai quá trình khai phá dữ liệu cho các yêu cầu trong thực tiễn
- Nâng cao kỹ năng làm việc nhóm
- Nâng cao kỹ năng giải quyết vấn đề và ra quyết định

1.3. Dự kiến kết quả đạt được

Về mặt kiến thức

- Nắm rõ được các kiến thức căn bản của môn phân tích dữ liệu
- Nắm rõ được các yêu cầu của tiền khai phá dữ liệu
- Nắm rõ được các khái niệm trong môn phân tích dữ liệu
- Nắm rõ các thuật toán trong môn phân tích dữ liệu

Về mặt thực hành

- Nắm rõ cách thức xử lý tiền dữ liệu
- Áp dụng được các thuật toán đã được học vào phần mềm TANAGRA

Sản phẩm đề tài

- DataBase
- Project trong phần mềm TANAGRA
- File word đồ án

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1. Giới thiệu về khai phá dữ liệu

2.1.1 Khái niệm

Khai phá dữ liệu (Data Mining) 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 - Predictive Information) ẩn chứa trong khối lượng dữ liệu khổng lồ trong thực tế

2.1.2 Vai trò của khai phá dữ liệu trong kinh doanh

Không chỉ riêng trong kinh doanh, khai phá dữ liệu còn được ứng dụng trong rất nhiều ngành nghề, lĩnh vực khác như Marketing, quản trị chuỗi cung ứng (SCM), Logistic, sản xuất... Nhờ có khai phá dữ liệu, các tập dữ liệu khổng lồ được thu thập sẽ phát huy được tác dụng trong việc:

- **Ra quyết định tự động:**

Cho phép tự động phân tích dữ liệu và tự động hóa các quyết định mà không phải mất thời gian chờ quyết định của con người

- **Dự báo chính xác:**

Lập kế hoạch và cung cấp dự báo dựa trên các xu hướng trong quá khứ và điều kiện hiện tại

- **Giảm thiểu chi phí:**

Giúp các tổ chức kiểm soát các hoạt động sản xuất, bán hàng, quảng cáo... và phân bổ nguồn lực hợp lý

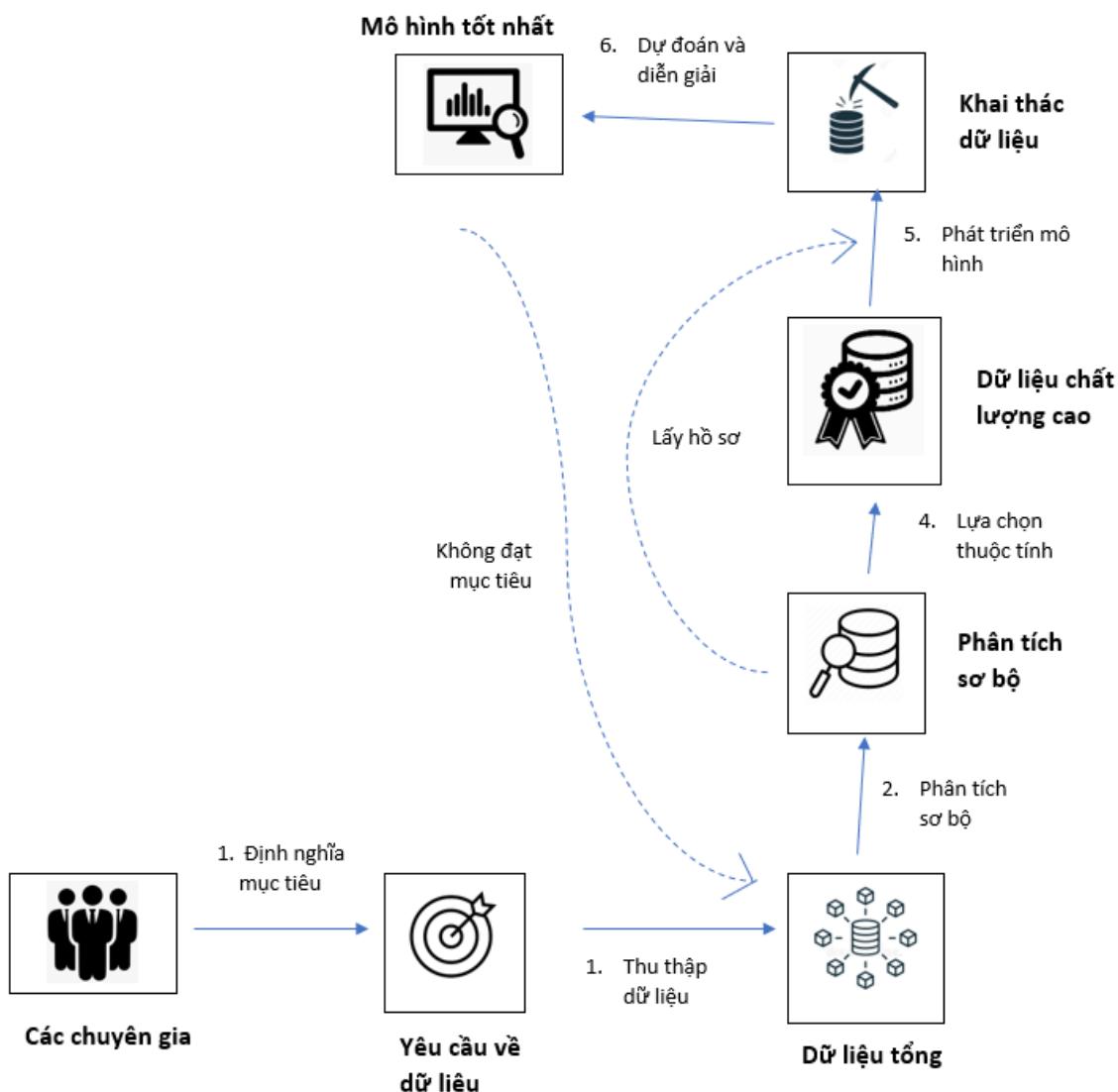
- **Thấu hiểu khách hàng:**

Khai phá dữ liệu giúp phân tích dữ liệu khách hàng, từ đó tìm các đặc điểm về sở thích, thói quen, hành vi... qua đó, xác định nhu cầu mỗi khách hàng một cách chính xác nhất.

2.1.3 Quy trình khai phá dữ liệu

Qui trình khai phá dữ liệu là một qui trình phức tạp bao gồm kho dữ liệu chuyên sâu cũng như các công nghệ tính toán. Hơn nữa, Data Mining không chỉ giới hạn trong việc trích xuất dữ liệu mà còn được sử dụng để chuyển đổi, làm sạch, tích hợp dữ liệu và phân tích mẫu.

Dưới đây là các bước trong quy trình khai phá dữ liệu



Hình 2. 1 Qui trình khai phá dữ liệu

- Bước 1: 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 cảm thấy 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
- **Bước 2: 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.
 - **Bước 3: 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.
 - **Bước 4: Lựa 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.
- **Bước 5: 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.
- **Bước 6: 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

- **Hệ thống xử lý luồng (Streaming Systems):**

Có hai dạng hệ thống xử lý luồng dữ liệu: hệ thống xử lý luồng đầu vào (stream ingestion systems) và hệ thống phân tích luồng dữ liệu (stream analytics systems).

- **Stream ingestion system:** thực hiện thu thập các luồng dữ liệu đầu vào; xử lý trực tuyến các dữ liệu từ nguồn phát sinh (access log, event log,...) và xuất ra hệ thống phân tích dữ liệu để tiến hành trích xuất các dữ liệu cần thiết.
- Stream analytics system: là hệ thống xử lý dữ liệu được thu nhận từ hệ thống thu thập đầu vào. Việc xử lý dữ liệu này được thực hiện trên các gói tin đi vào hệ thống mà không cần xử lý trên tập tin hoặc không cần lưu vào cơ sở dữ liệu trước khi phân tích.
- Hệ thống xử lý luồng đầu vào cung cấp nguồn dữ liệu cho hệ thống phân tích; kết quả đầu ra của hệ thống phân tích có thể chuyển ngược lại cho hệ thống phân tích đầu vào để tiếp tục xử lý hoặc được ghi vào hệ thống dữ liệu tĩnh (data at rest) để thực hiện lưu trữ.

2.2.2 Kho dữ liệu và khai phá dữ liệu trong BI

2.2.2.1. Kho dữ liệu

Kho dữ liệu (Data Warehouse) là kho lưu trữ quan trọng nhất cho các dữ liệu sẵn có để phát triển kiến trúc BI và các hệ thống hỗ trợ ra quyết định.

Thuật ngữ kho dữ liệu chỉ ra toàn bộ các hoạt động liên quan đến thiết kế, triển khai và sử dụng kho dữ liệu.

Có ba loại chính của dữ liệu đưa vào kho dữ liệu:

- **Dữ liệu nội bộ**
- **Dữ liệu bên ngoài**
- **Dữ liệu cá nhân**

2.2.2.2. Business intelligence(BI):

Business intelligence (BI) có thể được định nghĩa như là một tập hợp các mô hình toán học và phương pháp phân tích khai thác các dữ liệu có sẵn để tạo ra thông tin và kiến thức hữu ích cho các quá trình ra quyết định phức tạp

Mục đích chính của hệ thống BI là cung cấp kiến thức cho người làm việc các công cụ và phương pháp giúp cho phép họ ra quyết định hiệu quả và kịp thời

Hệ thống BI có thể được xem là sự kết hợp của 3 thành phần chính như sau:

- Data Warehouse (Kho dữ liệu): Chứa dữ liệu tổng hợp của doanh nghiệp
- Data Mining (Khai thác dữ liệu): Các kỹ thuật dùng để khai thác dữ liệu như phân loại (Classification), phân nhóm (Clustering), kết hợp (Association Rule), dự đoán (Prediction),...
- Business Analyst (Phân tích kinh Doanh): Quyết định chiến lược đối với hoạt động kinh doanh của doanh nghiệp.

Ta có thể thấy Business intelligence (BI) có mối quan hệ rất chặt chẽ với Data Warehouse và Data Mining. Vấn đề cốt lõi trong hệ thống BI là kho dữ liệu (Data Warehouse) và khai phá dữ liệu (Data Mining) vì dữ liệu dùng trong BI là dữ liệu tổng hợp (Nhiều nguồn, nhiều định dạng, phân tán và có tính lịch sử) đó là đặc trưng của kho dữ liệu. Đồng thời việc phân tích dữ liệu trong BI không phải là những phân tích đơn giản (Query, Filtering) mà là những kỹ thuật trong khai phá dữ liệu (Data Mining) dùng để phân loại (classification) phân cụm (Clustering), hay dự đoán (Prediction). Vì vậy BI có mối quan hệ rất chặt chẽ với Data Warehouse và Data Mining.

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/đối tượng 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 (phần mô tả) của một lớp (class) và X là dữ liệu, đối tượng
 - Bước học: X trong tập huấn luyện, một trị y được cho trước với $X \Rightarrow$ xác định f
 - 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)
- Các giải thuật Phân lớp dữ liệu:
 - Phân lớp với cây quyết định (decision tree)
 - Phân lớp với mạng Bayesian
 - Phân lớp với mạng neural
 - Phân lớp với k phần tử lảng 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)

2.3.2 Phương pháp gom cụm

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 (Dissimilar) 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 giá hiệu quả của phân tích gom cụm, điều này phụ thuộc vào mục đích của gom cụm như:

- Data reduction
 - “Natural clusters”
 - “Useful” clusters
 - Outlier detection
- **Ý nghĩa của phương pháp gom cụm:**
 - Hỗ trợ giai đoạn tiền xử lý dữ liệu
 - Mô tả sự phân bố dữ liệu, đối tượng
 - Nhận dạng mẫu
 - Phân tích dữ liệu không gian
 - Xử lý ảnh
 - Phân mảnh thị trường
 - Gom cụm tài liệu

2.3.3 Phương pháp luật kết hợp

Khai phá luật kết hợp được mô tả như sự tương quan của các sự kiện những sự kiện xuất hiện thường xuyên một cách đồng thời. Nhiệm vụ chính của khai phá luật kết hợp là phát hiện ra các tập con cùng xuất hiện trong một khối lượng giao dịch lớn của một cơ sở dữ liệu cho trước.

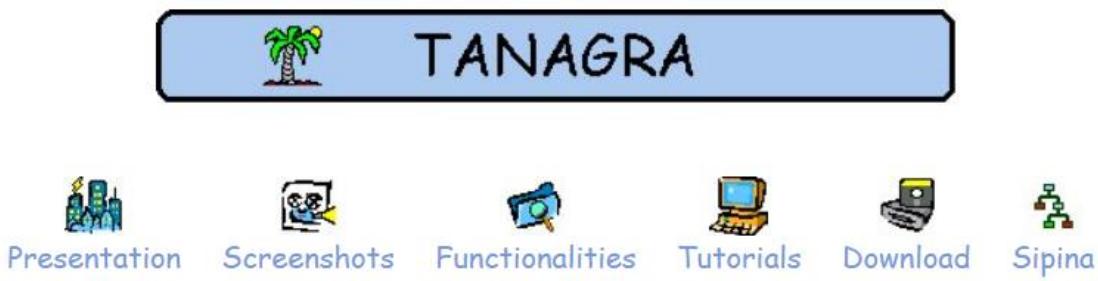
- **Một số thuật toán phát hiện luật kết hợp**
 - Thuật toán Apriori
 - Thuật toán Apriori-TID
 - Thuật toán Apriori-Hybrid
- **Phân loại các loại luật kết hợp**
 - Boolean association rule (luật kết hợp luận lý)/quantitative association rule (luật kết hợp lượng số)
 - Single-dimensional association rule (luật kết hợp đơn chiều)/multidimensional association rule (luật kết hợp đa chiều)

- Single-level association 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)

2.4. Giới thiệu về phần mềm sử dụng (Tanagra)

2.4.1 Tổng quan về phần mềm Tanagra

Tanagra được tạo ra bởi Ricco Rakotomala vào tháng 12 năm 2003



Hình 2. 2 Phần mềm Tanagra

Tanagra là một phần mềm khai phá dữ liệu miễn phí, mã nguồn mở, phục vụ cho công việc học tập và nghiên cứu. Nó là công cụ phục vụ cho công việc khai phá dữ liệu từ phân tích dữ liệu thăm dò (exploratory), học thống kê (statistical learning), máy học (machine learning) và vùng cơ sở dữ liệu (databases area).

Mục đích chính của dự án Tanagra là để cho các nhà nghiên cứu và sinh viên dễ sử dụng các phần mềm liên quan đến khai phá dữ liệu (đặc biệt trong việc thiết kế GUI và cách sử dụng nó), và cho việc phân tích thiết thực và tổng hợp dữ liệu.

Mục đích thứ hai của Tanagra là để đề xuất với các nhà nghiên cứu một kiến trúc cho phép họ dễ dàng thêm các phương pháp khai thác dữ liệu của riêng họ, để so sánh kết quả của họ. Nhiều tác động của Tanagra như là nền tảng thử nghiệm nhằm để cho họ đi đến những thiết yếu của công việc của họ, và để họ có thể đối phó với các phần khó khăn trong việc quản lý dữ liệu.

Mục đích thứ ba và cuối cùng, trong việc mới làm quen với hướng phát triển, bao gồm trong việc khuyến khích một phương pháp để xây dựng các loại phần mềm. Người sử dụng cần tận dụng lợi thế của việc miễn phí mã nguồn, để tìm các sắp xếp của phần mềm này được xây dựng, các vấn đề cần tránh, các bước chính của dự án, có các công

cụ và thư viện mã để sử dụng. Bằng cách này, Tanagra có thể được coi như là một công cụ nghiệp vụ sư phạm cho việc học tập kỹ thuật lập trình.

- **Ưu điểm:**

- Mã nguồn mở
- Dung lượng phần mềm nhẹ
- Thao tác đơn giản
- Đa dạng chức năng
- Phù hợp với nhu cầu học thuật, thực hành và nghiên cứu

- **Hạn chế**

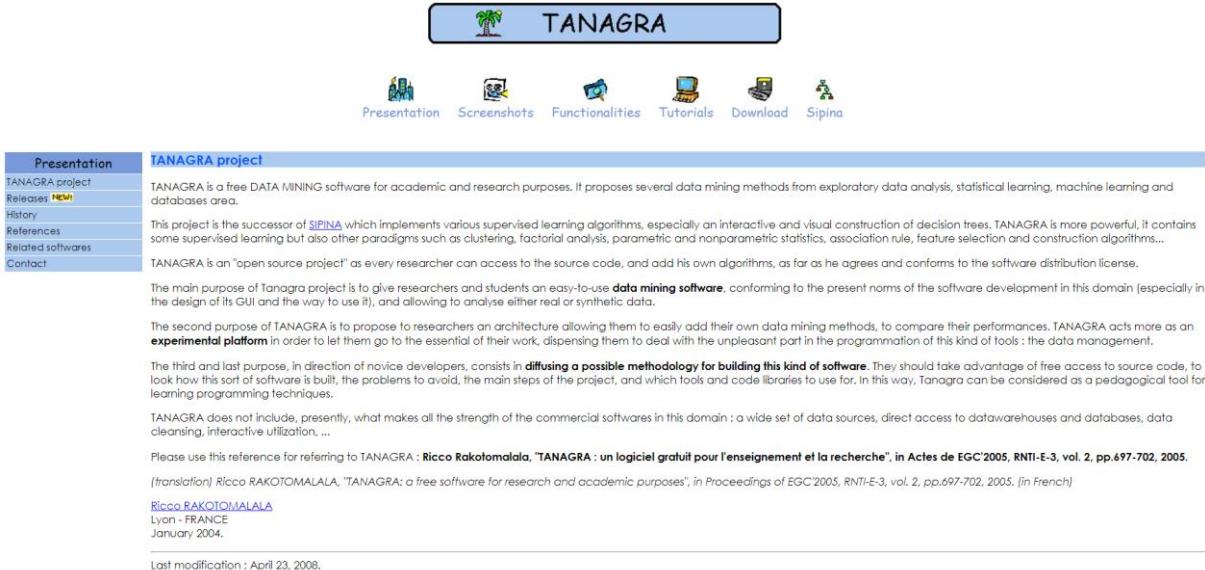
- Giao diện còn đơn giản
- Phù hợp với nhu cầu học tập hơn nhu cầu kinh doanh, quản lý
- TANAGRA hiện không bao gồm những điểm mạnh của một phần mềm phân mềm khai phá dữ liệu trong lĩnh vực thương mại : một bộ nguồn dữ liệu phong phú, truy cập trực tiếp vào kho dữ liệu và cơ sở dữ liệu, làm sạch dữ liệu, sử dụng tương tác, ...
- Tanagra chỉ hỗ trợ các file *.txt, *.arff, *.xls, *.dat, *.data.

2.4.2 Cách sử dụng phần mềm

2.4.2.1. Cách cài đặt phần mềm

Bước 1: Truy cập trang web: <http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html>

Ứng dụng Tanagra trong khai phá dữ liệu



The screenshot shows the Tanagra project website. At the top, there's a logo with a palm tree and the word "TANAGRA". Below the logo is a navigation bar with icons for Presentation, Screenshots, Functionalities, Tutorials, Download, and Sipina. On the left, there's a sidebar with a "Presentation" tab selected, showing links for TANAGRA project, Releases (with a "NEW" badge), History, References, Related softwares, and Contact. The main content area is titled "TANAGRA project" and discusses the software's purpose, its relationship to SIPINA, and its features like data mining methods, clustering, and association rule. It also mentions its open source nature and experimental platform status. A note at the bottom credits Ricco Rakotomalala for the work.

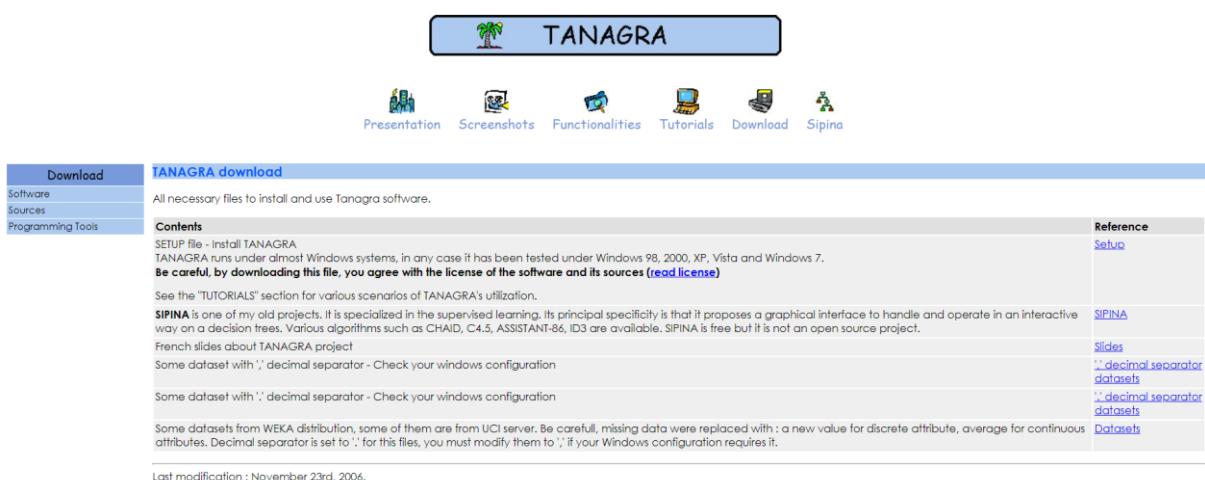
Hình 2. 3 Giao diện trang web Tanagra

Bước 2: Chọn vào download



Hình 2. 4 Nút Download

Bước 3: Chọn phiên bản và mục đích sử dụng sau đó tiến hành download



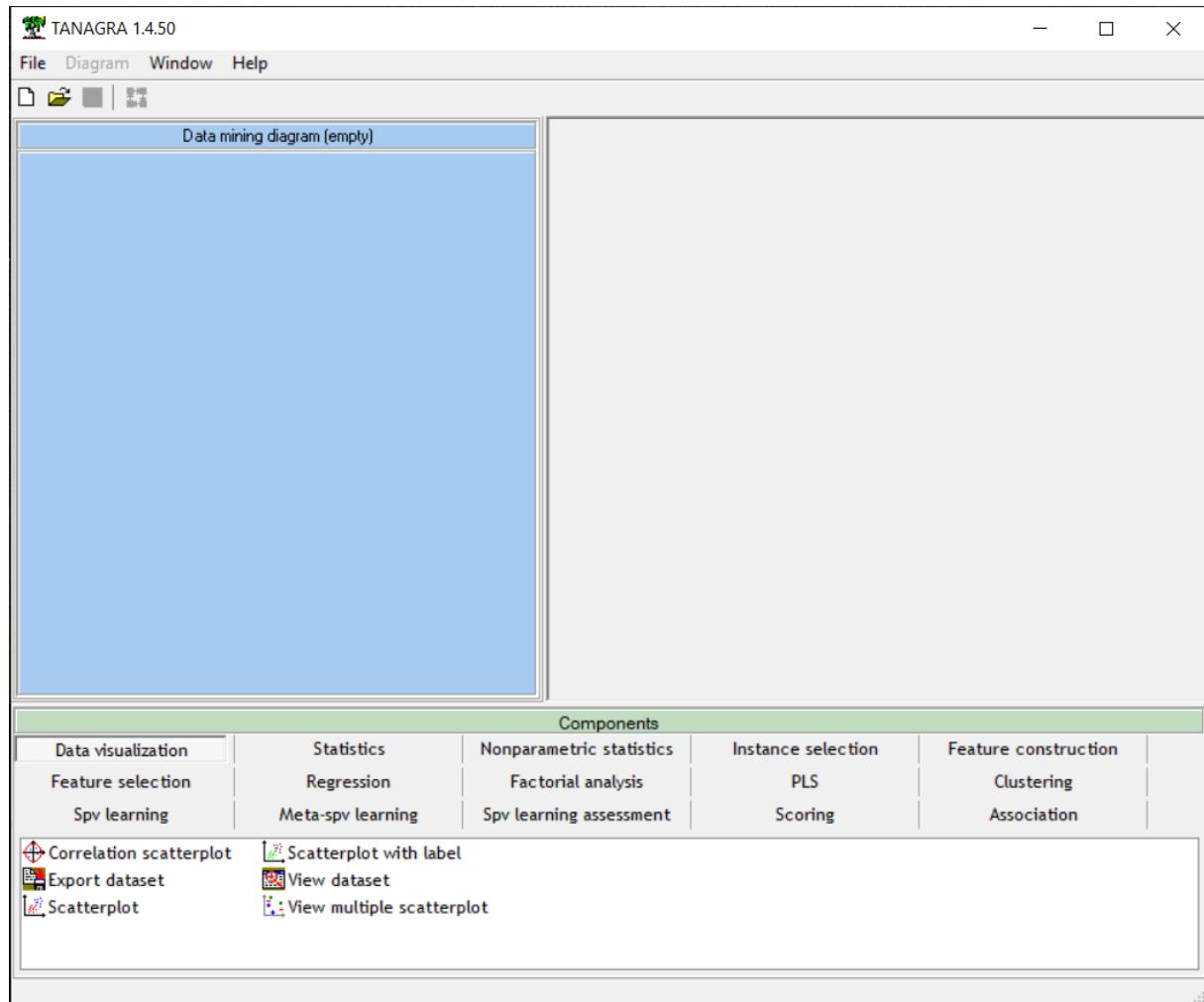
The screenshot shows the "Download" section of the Tanagra website. It includes a "Software" link, a "Sources" link, and a "Programming Tools" link. The main content area is titled "TANAGRA download" and contains a note about installing the software. It lists "SETUP file - Install TANAGRA" and "SIPINA" as one of the old projects. There are sections for "Contents" (including "French slides about TANAGRA project" and "Some dataset with ';' decimal separator - Check your windows configuration") and "Reference" (links for "Setup", "Slides", "'; decimal separator datasets", and "Datasets"). A note at the bottom discusses Windows configuration for datasets.

Hình 2. 5 Các phiên bản Tanagra

Bước 4: Cài đặt cho máy tính và sử dụng

2.4.2.2. Giao diện Tanagra

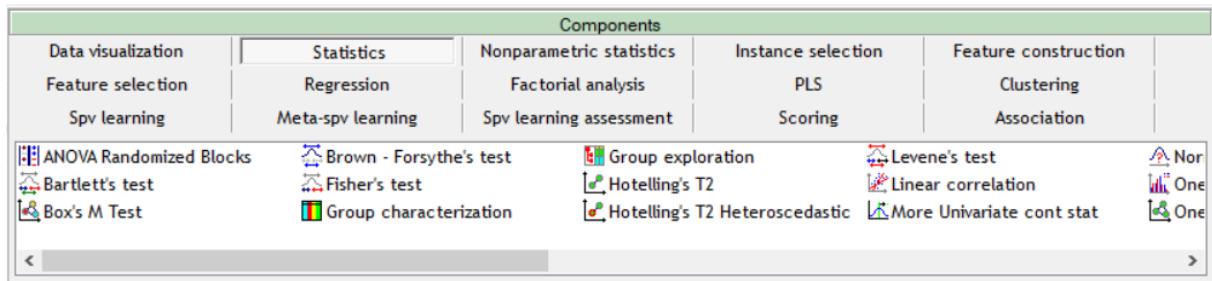
Giao diện chính:



Hình 2. 6 Giao diện chính của phần mềm Tanagra

2.4.2.3. Giao diện thanh công cụ

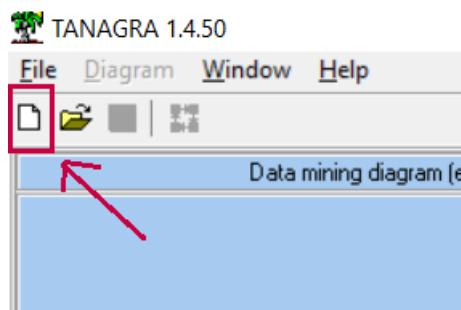
Mỗi nút ở thanh công cụ biểu diễn cho 1 thuật toán khai phá dữ liệu, khi kết hợp 2 hay nhiều công cụ sẽ tạo nên 1 bài toán khai phá dữ liệu mà người dùng mong muốn



Hình 2. 7 Thanh công cụ của phần mềm Tanagra

2.4.3 Cách truyền dữ liệu vào chương trình

Trước tiên ta sẽ khởi động chương trình Tanagra. Chọn vào biểu tượng để khởi động một dataset.



Hình 2. 8 Khởi động Dataset

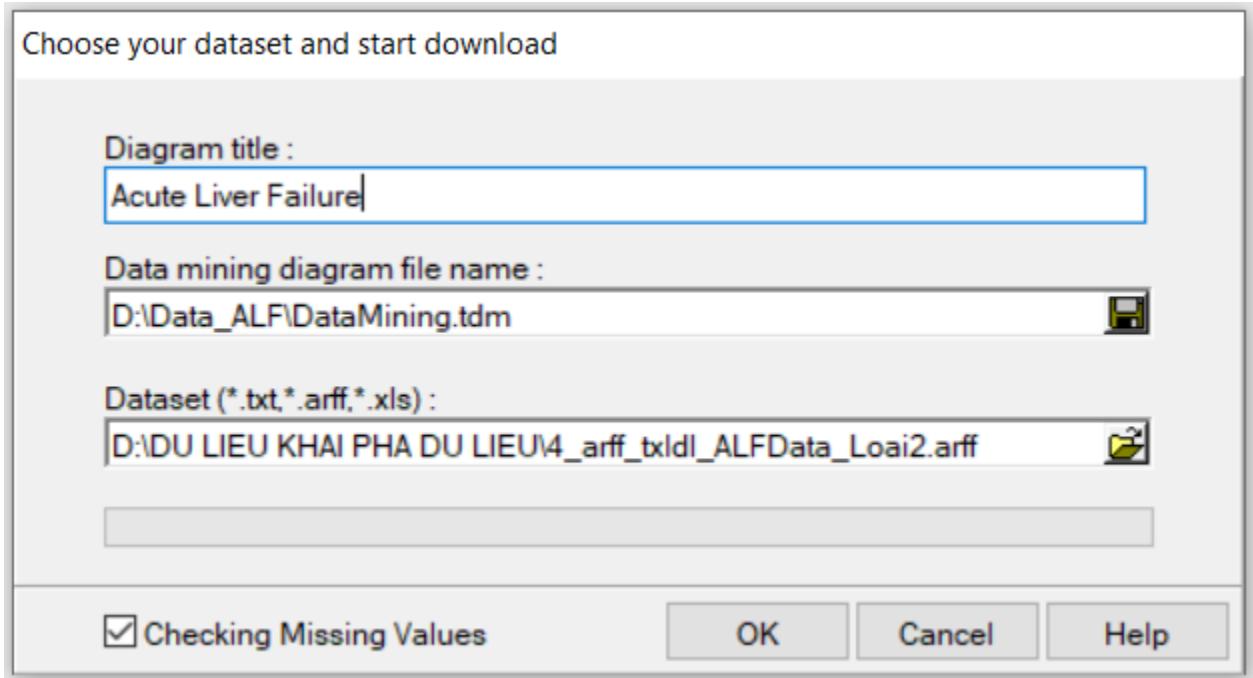
Bảng Setting Dataset xuất hiện, tại đây ta sẽ thiết lập những thông tin cơ bản về dataset:

Diagram titile: tên các sơ đồ của dataset

Data mining diagram file name: nơi lưu file

Dataset (*.txt, *.arff, *.xls): File data mà bạn cần import vào dataset

Ở ô tick chọn **Checking Missing Values**. Khi ta tích vào thì dữ liệu truyền vào sẽ được chương trình kiểm tra có thiếu sót hay không.



Hình 2. 9 Thiết lập nhập và xuất dữ liệu

Sau khi hoàn tất thiết lập, ta chọn OK để truyền dữ liệu vào phần mềm.

Tanagra sẽ hiện kết quả truyền vào:

- [1]. **Continuous missing data handling:** “Dữ liệu truyền vào liên tục” không bị thiếu do dữ liệu đã được làm sạch ở khâu tiền xử lý dữ liệu
- [2]. **Datasource processing:** Thời gian xử lý và khối lượng dữ liệu
- [3]. **Dataset description :** Hiện số lượng thuộc tính và dòng dữ liệu truyền vào (như hình là 21 thuộc tính và 1610 dữ liệu)
- [4]. Bảng cuối cùng là thông tin chi tiết của các thuộc tính bao gồm: Tên thực thể, loại dữ liệu như Continue (Biến liên tục) và Discrete (Biến rời rạc, số lượng dữ liệu đó).

Database : D:\DU LIEU KHAI PHA DU LIEU\4_arff_txid_ALFData_Loai2.arff

Download information

Continuous missing data handling	1
none	
Datasource processing	2
Computation time	0 ms
Allocated memory	146 KB

Dataset description 3

21 attribute(s)
1610 example(s)

Attribute	Category	Informations
Age	Continue	-
Gender	Discrete	2 values
Obesity	Continue	-
Maximum_Blood_Pressure	Continue	-
Minimum_Blood_Pressure	Continue	-
Good_Cholesterol	Continue	-
Bad_Cholesterol	Continue	-

4

Hình 2. 10 Bảng kết quả truyền dữ liệu vào

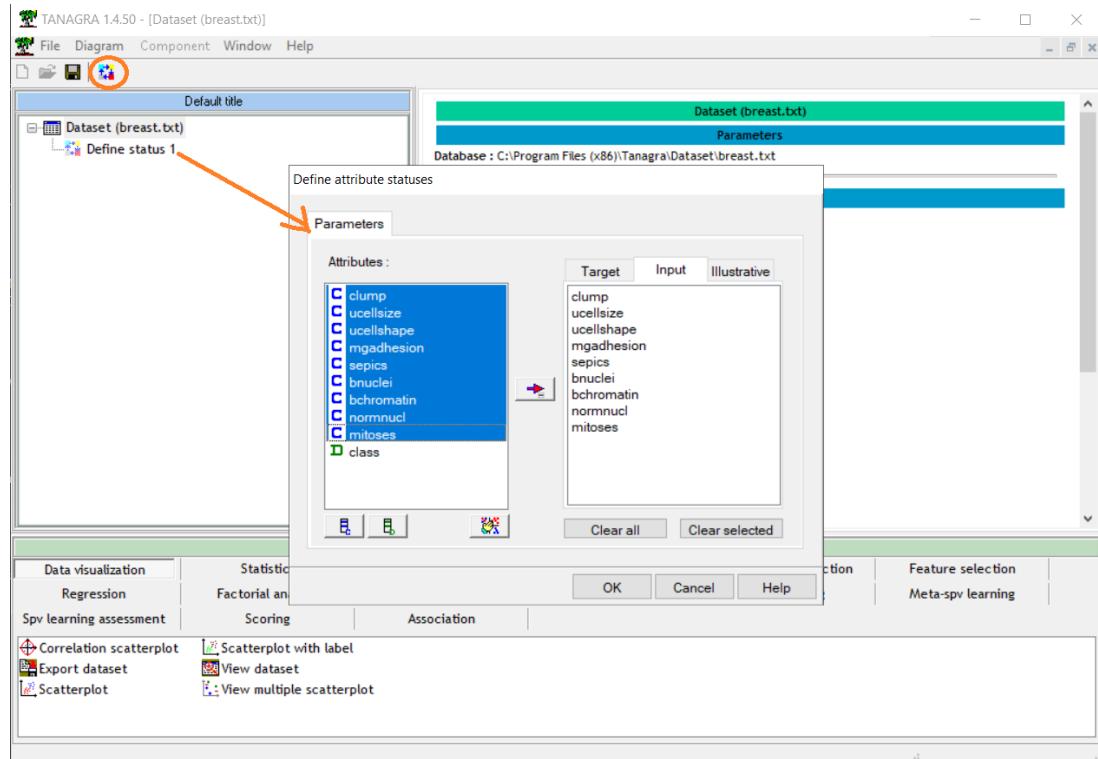
2.4.4 Cách thức tiến hành các thuật toán

2.4.4.1. Cây ra quyết định

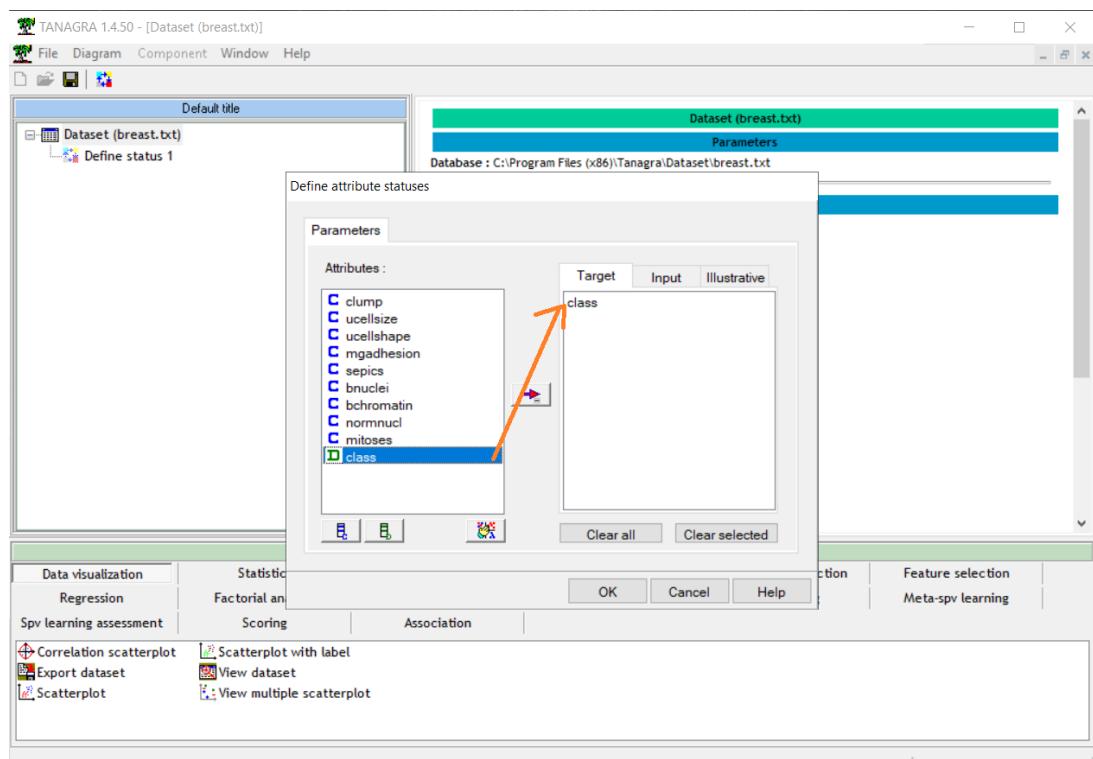
o Xác định các dữ liệu cần thiết

Chúng ta thêm thư mục DEFINE từ toolbar. Sau đó chọn các thuộc tính cần phân tích vào thư mục DEFINE.

Ứng dụng Tanagra trong khai phá dữ liệu

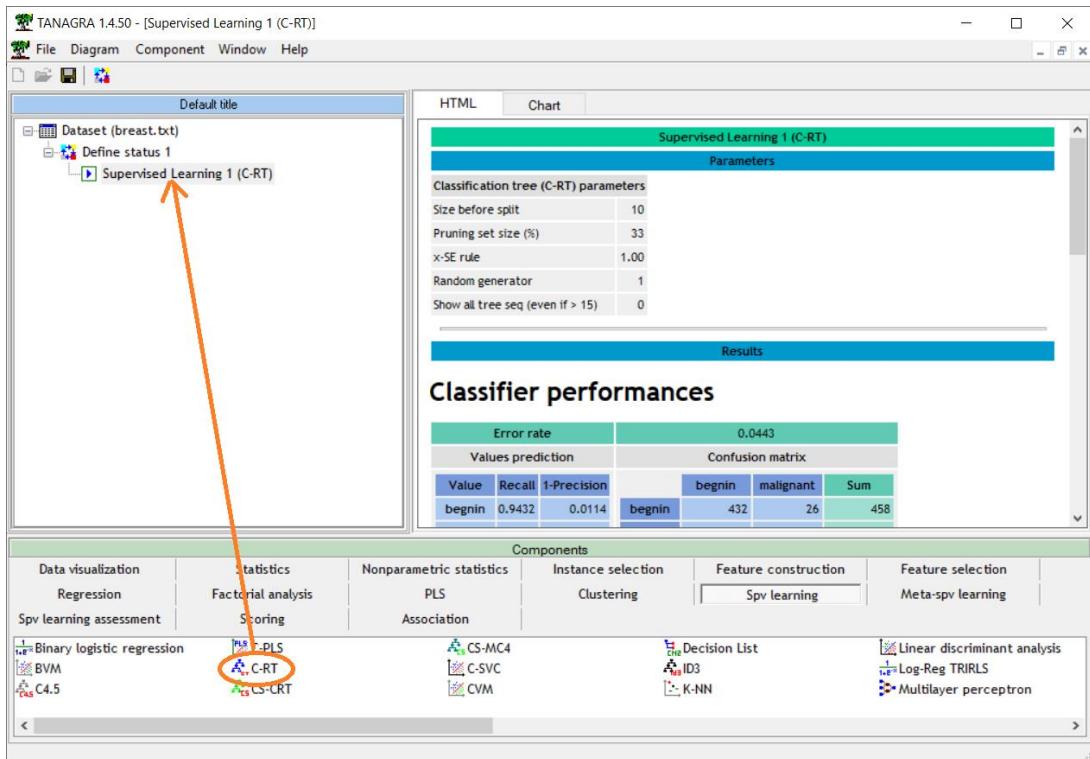


Hình 2. 11 Tạo Define cho cây ra quyết định



Hình 2. 12 Chọn Target cho cây ra quyết định

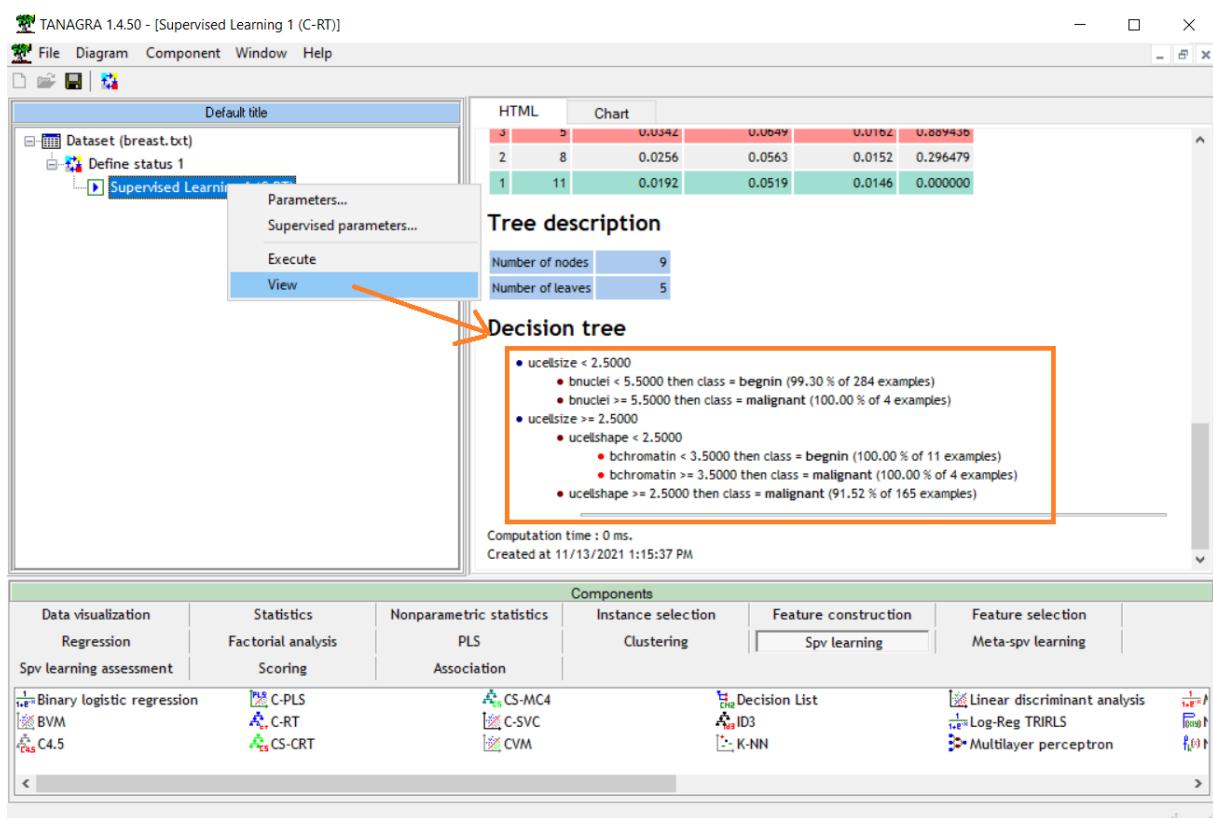
- Thêm thuật toán C-RT, hệ thống sẽ tự tạo thư mục Supervised Learning



Hình 2. 13 Thêm thuật toán C-RT

- Xem kết quả

Click chuột phải, sau đó chọn view để xem kết quả thuật toán



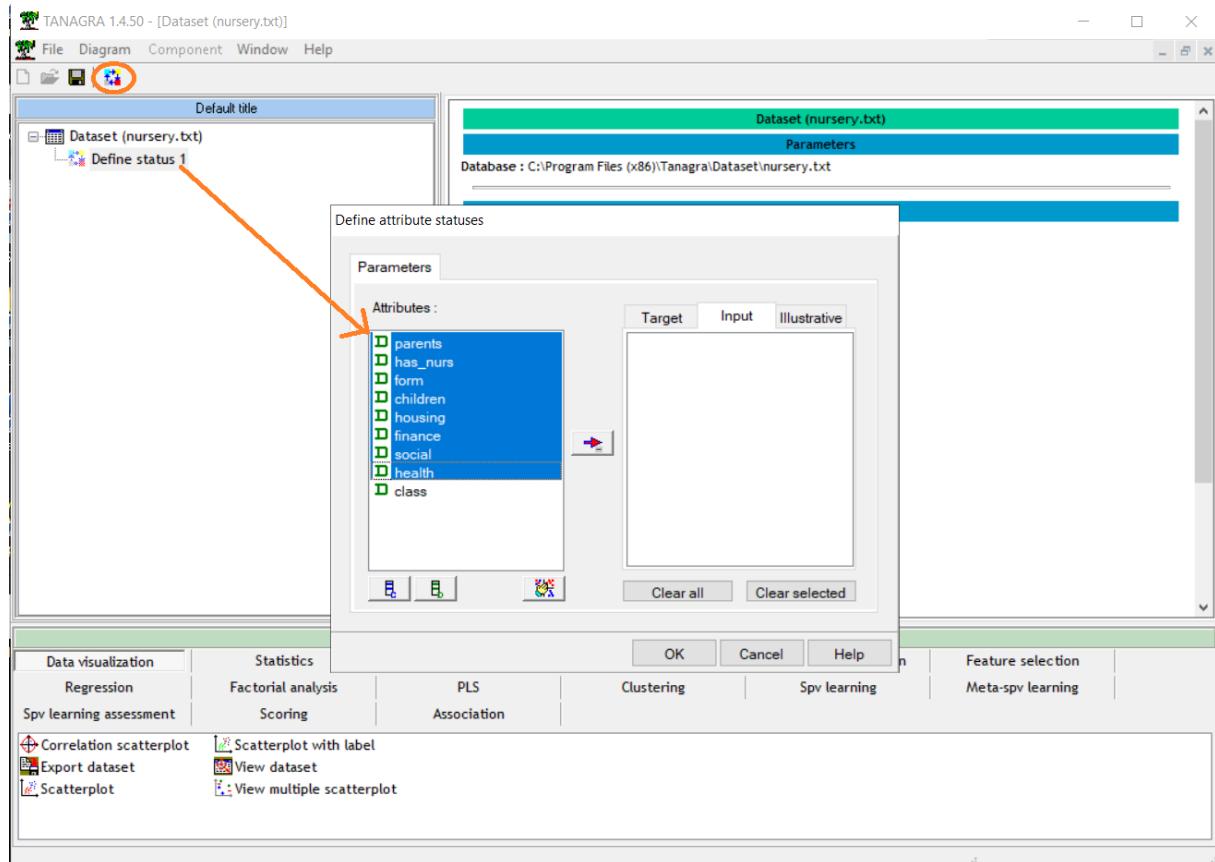
Hình 2. 14 Xem kết quả thuật toán cây ra quyết định C-RT

2.4.4.2. Thuật toán NaiveBayes

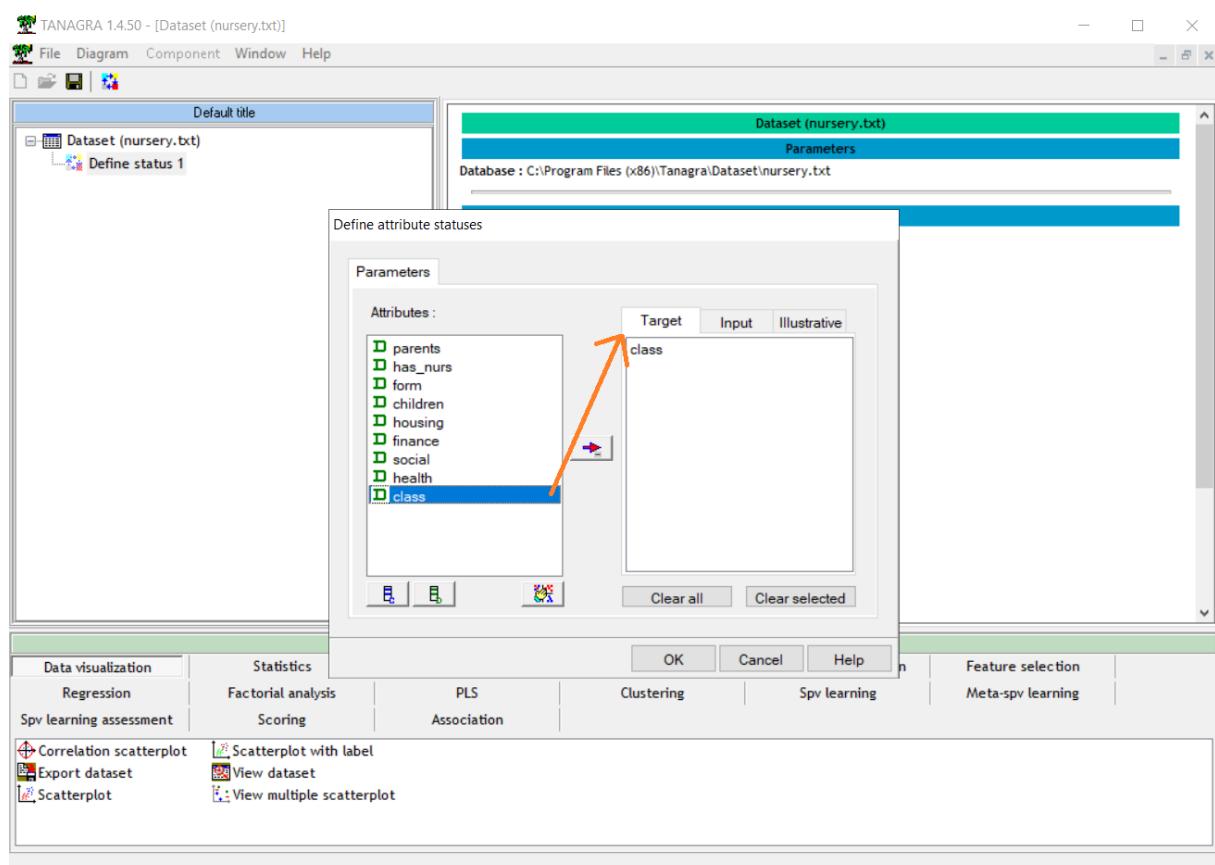
o Xác định dữ liệu cần phân tích

Tạo thư mục DEFINE sau đó đổ dữ liệu cần phân tích vào thư mục DEFINE

Ứng dụng Tanagra trong khai phá dữ liệu

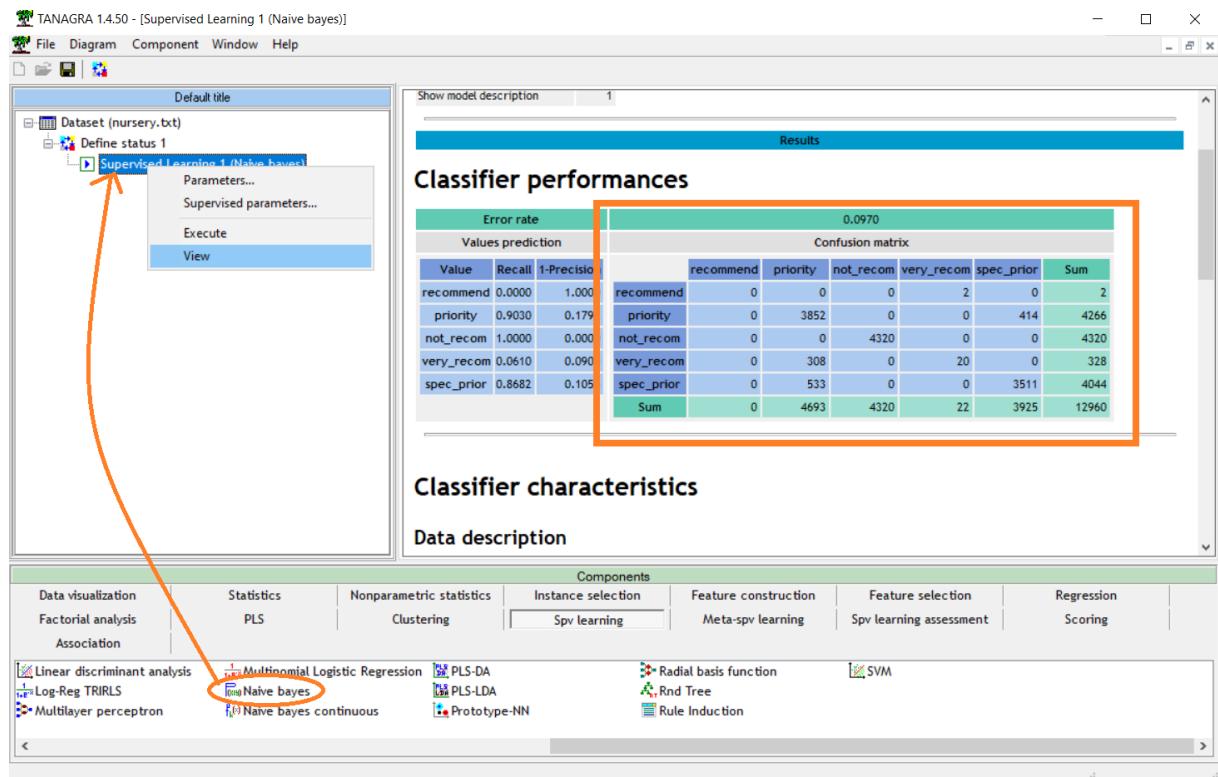


Hình 2. 15 Tạo define cho NavieBayes



HÌNH 2-12 Chọn Target cho NavieBayes

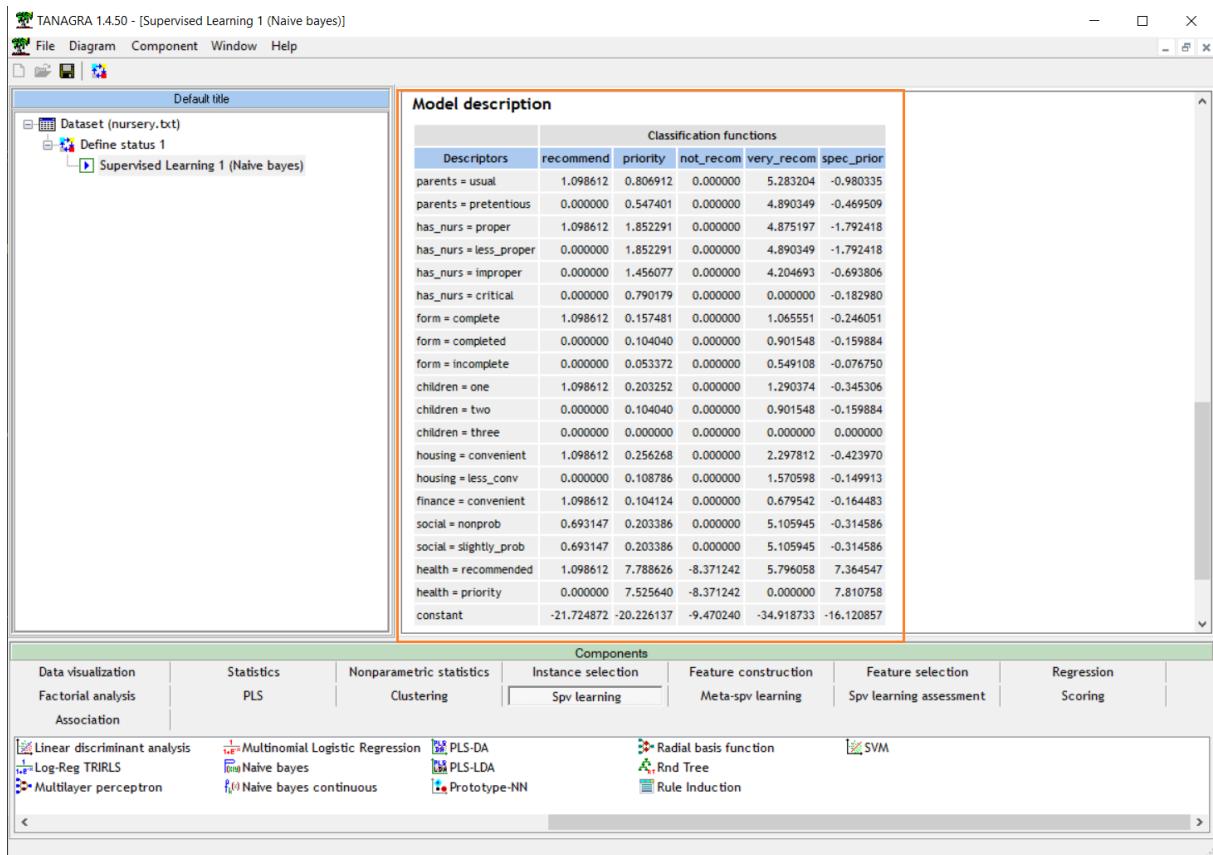
- Thêm thuật toán NaiveBayes, hệ thống sẽ tự tạo thư mục Supervised Learning



Hình 2. 16 Thêm thuật toán NavieBayes

- Xem kết quả

Ứng dụng Tanagra trong khai phá dữ liệu



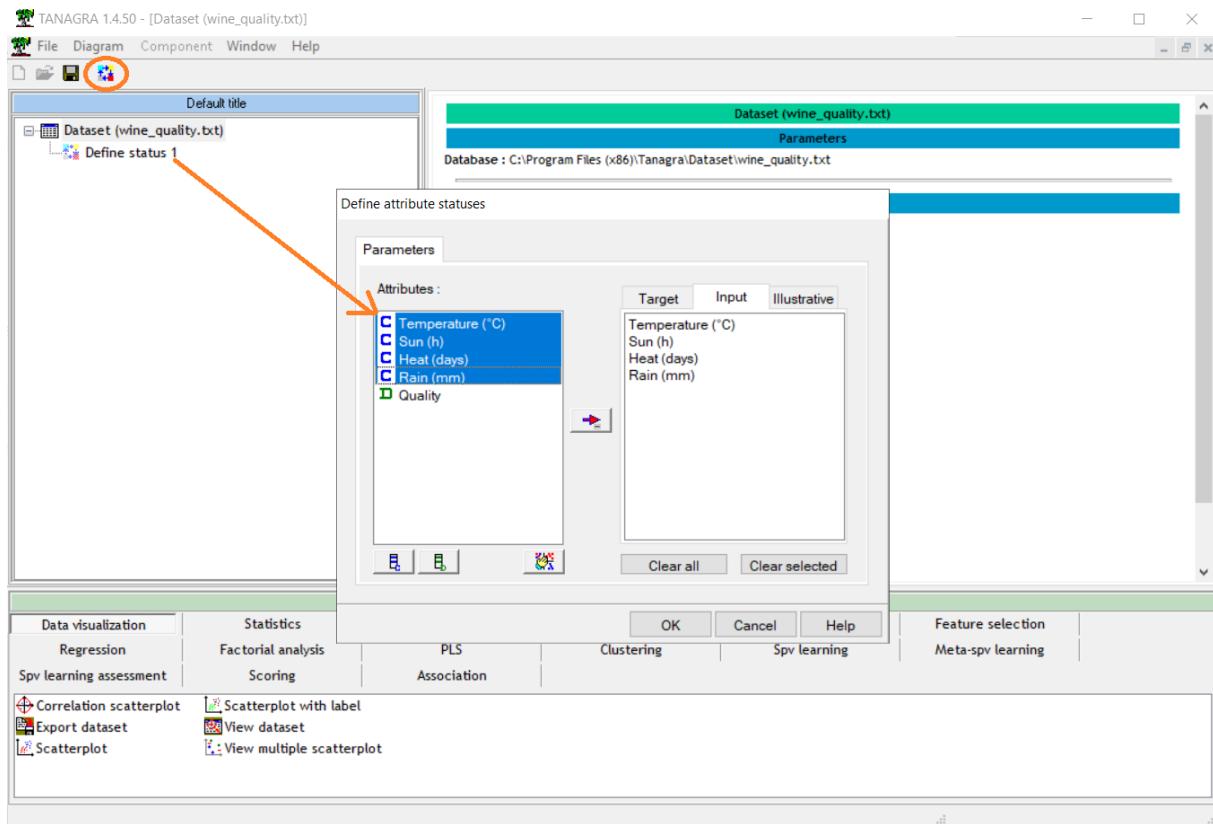
Hình 2. 17 Xem kết quả thuật toán NavieBayes

2.4.4.3. Thuật toán gom cụm K-Means

o Xác định dữ liệu cần phân tích

Tạo DEFINE, để dữ liệu cần phân tích vào DEFINE

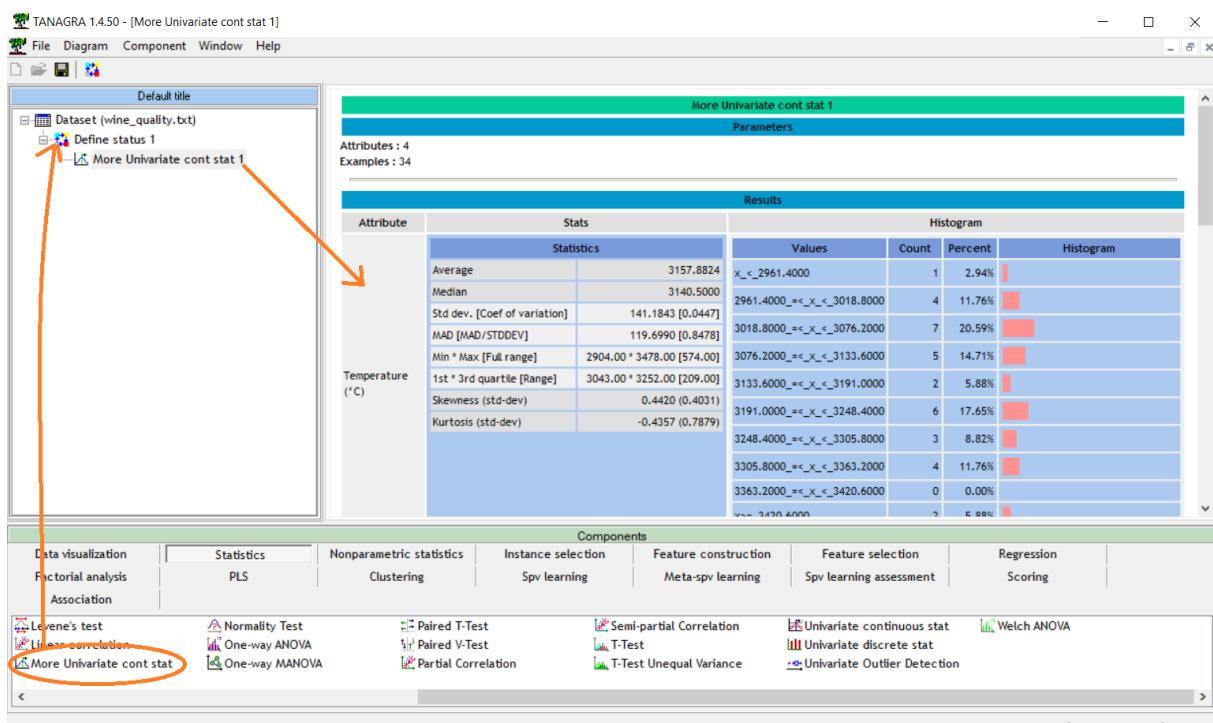
Ứng dụng Tanagra trong khai phá dữ liệu



Hình 2. 18 Tạo Define cho thuật toán K – Means

Thêm component E UNIVARIATE CONT STAT

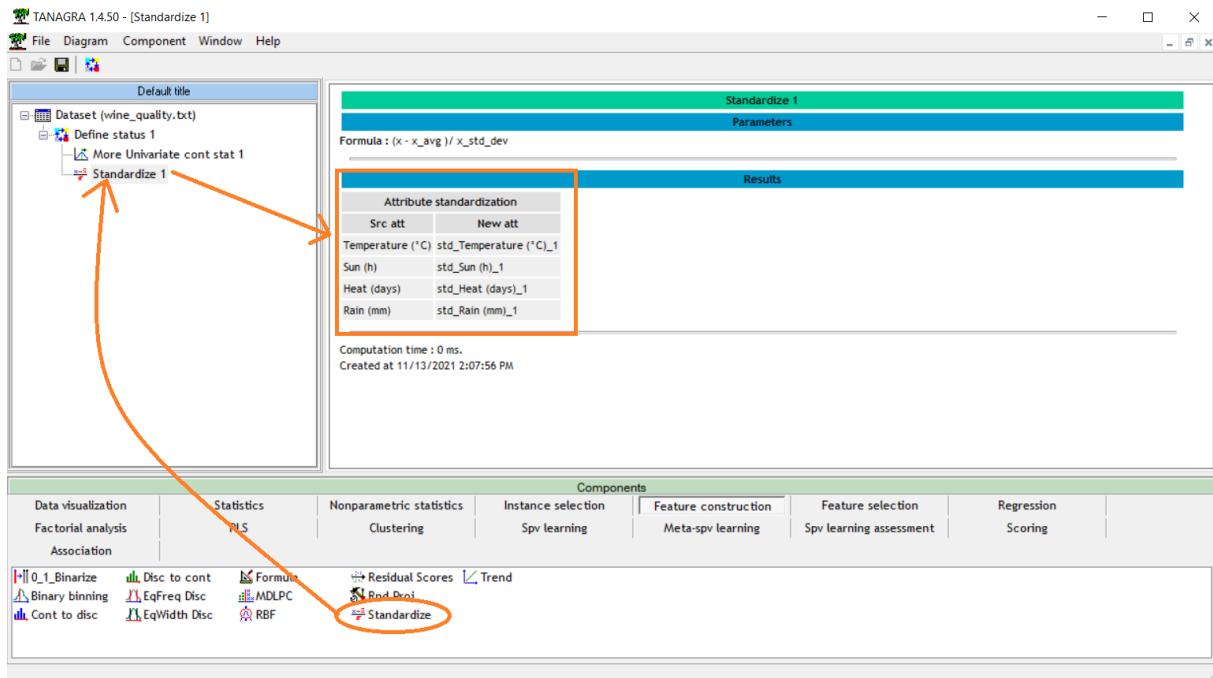
Thêm component E UNIVARIATE CONT STAT vào thư mục DEFINE vừa tạo



Hình 2. 19 Thêm component E UNIVARIATE CONT STAT

- Chuẩn hóa dữ liệu

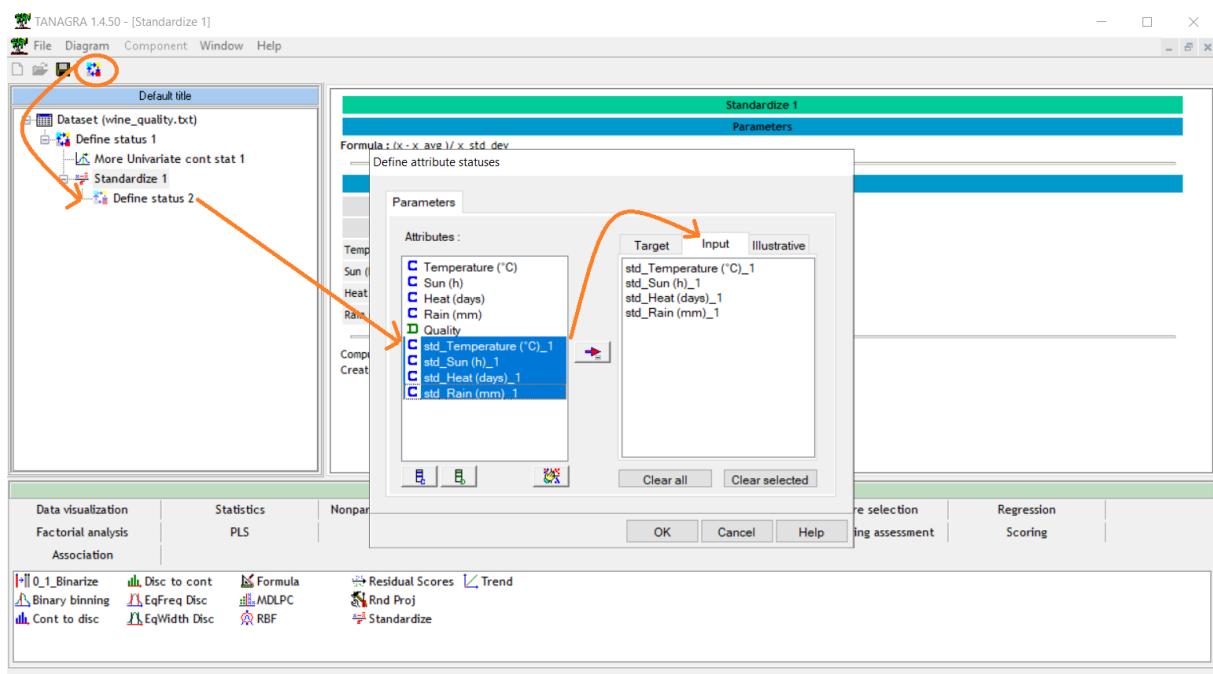
Thêm component STANDARDIZE vào thư mục DEFINE để chuẩn hóa dữ liệu



Hình 2. 20 Thêm component STANDARDIZE

- Tạo DEFINE 2

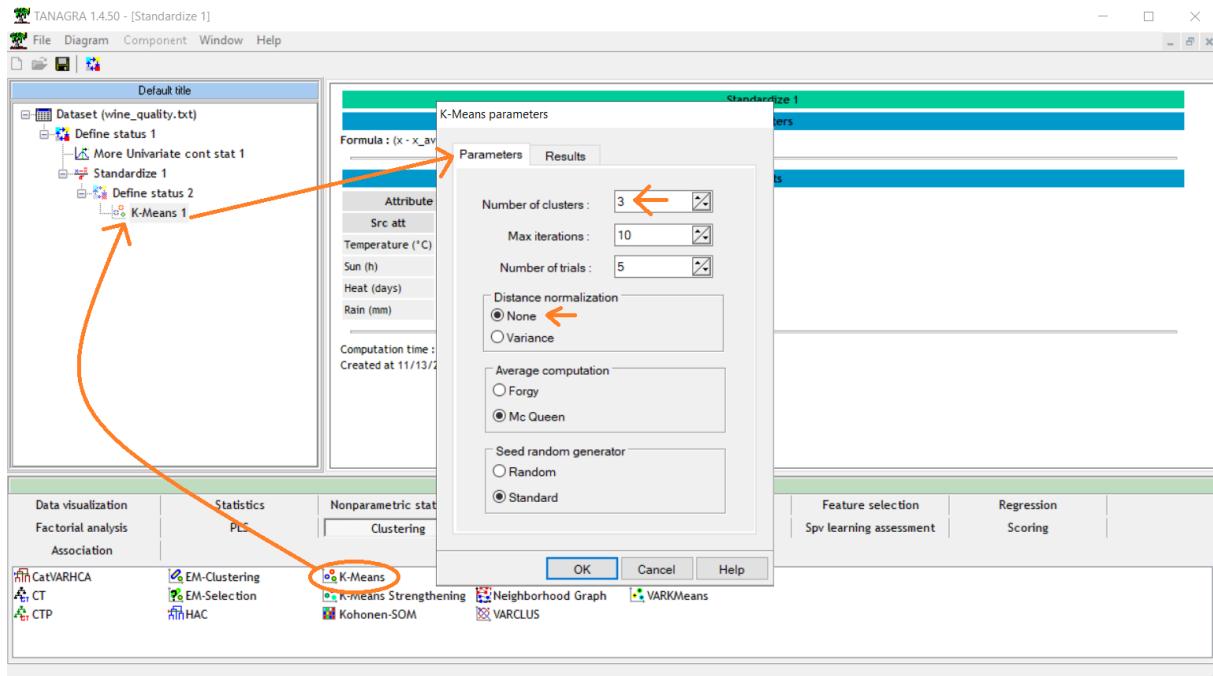
Tạo DEFINE 2 để đồ dữ liệu đã được chuẩn hóa vào DEFINE 2



Hình 2. 21 Tạo DEFINE 2

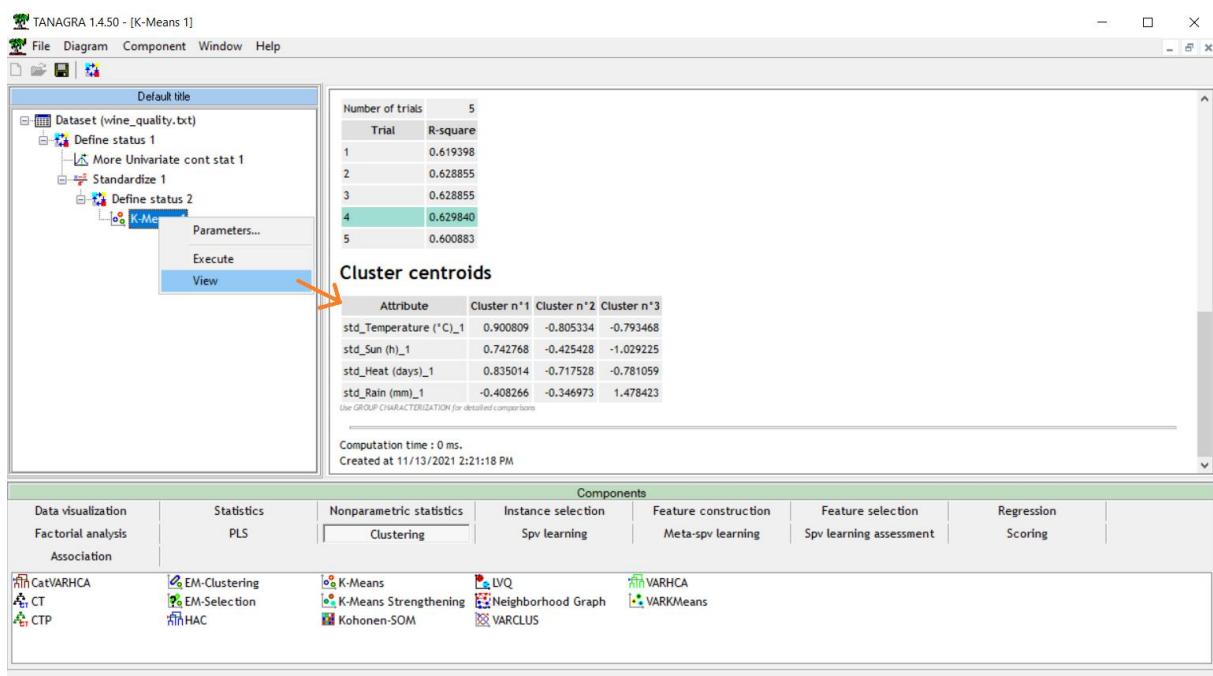
○ Tạo thuật toán K – MEAN

Thêm thuật toán K – MEAN vào thư mục DEFINE 2, nhập các thông số cần thiết:



Hình 2. 22 Thêm thuật toán K – MEAN

○ Xem kết quả

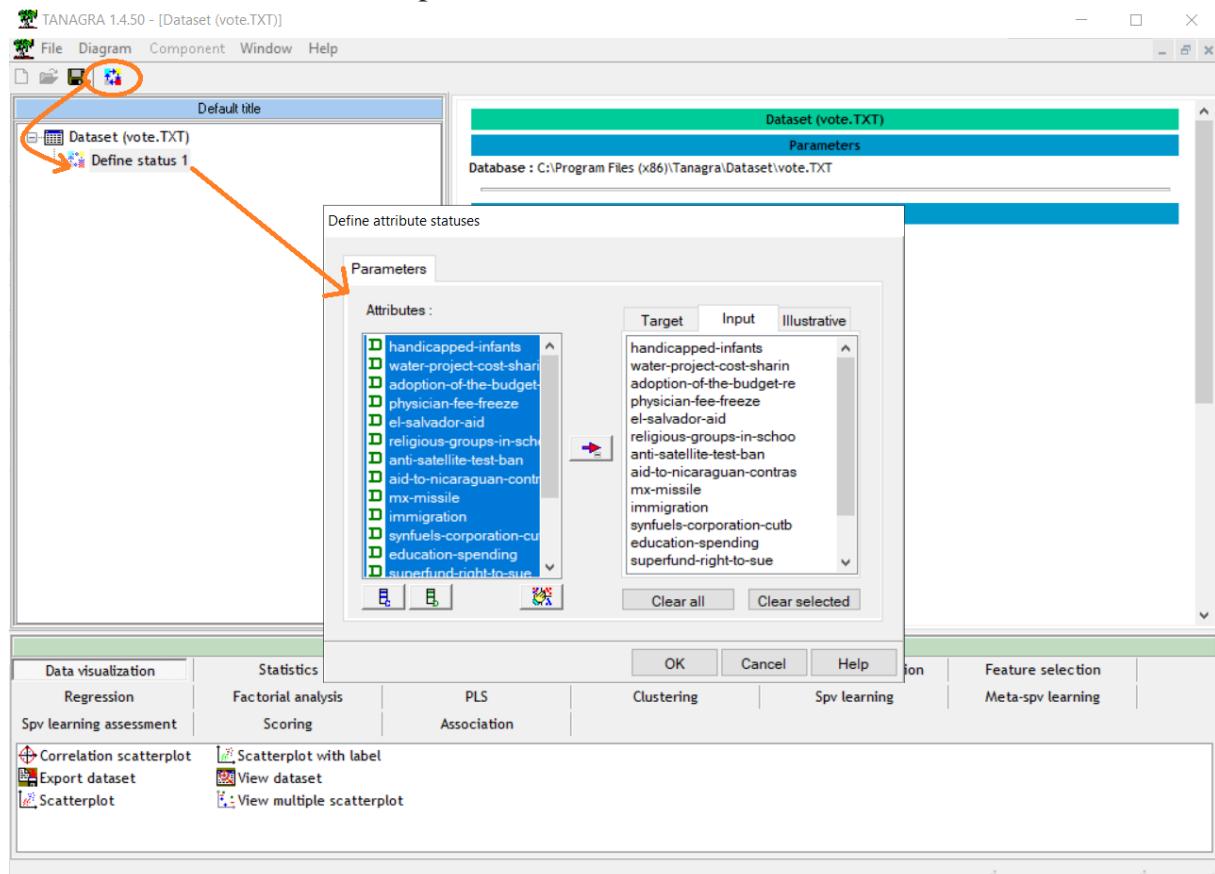


Hình 2. 23 Xem kết quả thuật toán K – MEAN

2.4.4.4. Luật kết hợp Apriori

- Xác định dữ liệu cần phân tích

Tạo DEFINE, để dữ liệu cần phân tích vào DEFINE

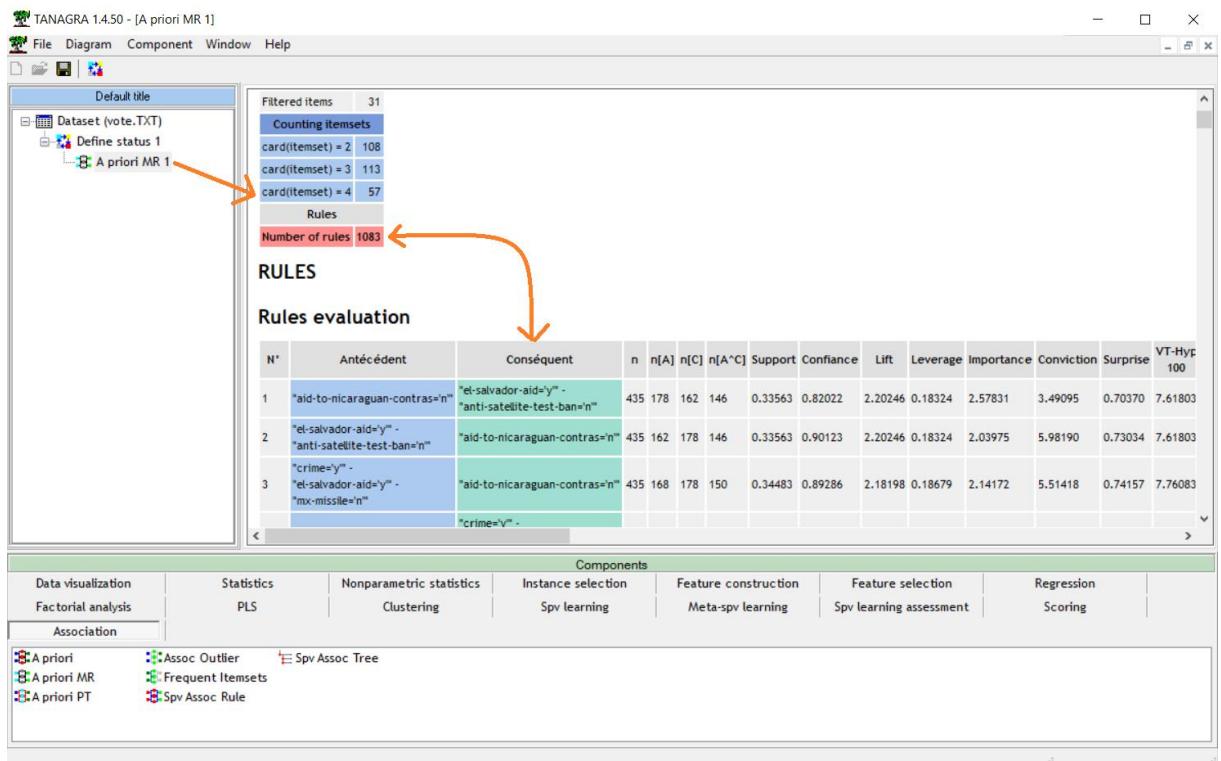


Hình 2. 24 Tạo define cho luật kết hợp Apriori

- Thêm component A PRIORI MR

Thêm component A PRIORI MR vào DEFINE vừa tạo

Ứng dụng Tanagra trong khai phá dữ liệu

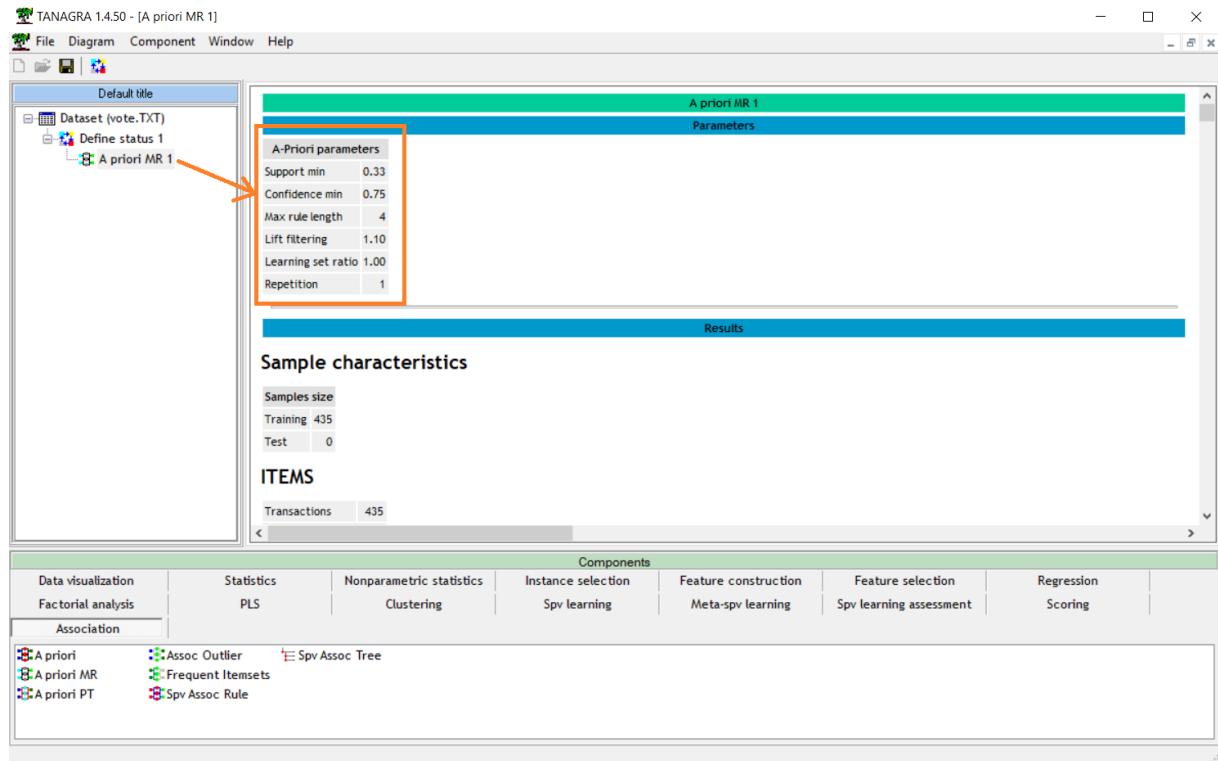


Hình 2. 25 Thêm component A PRIORI MR

○ Thiết lập thông số cho component A PRIORI MR

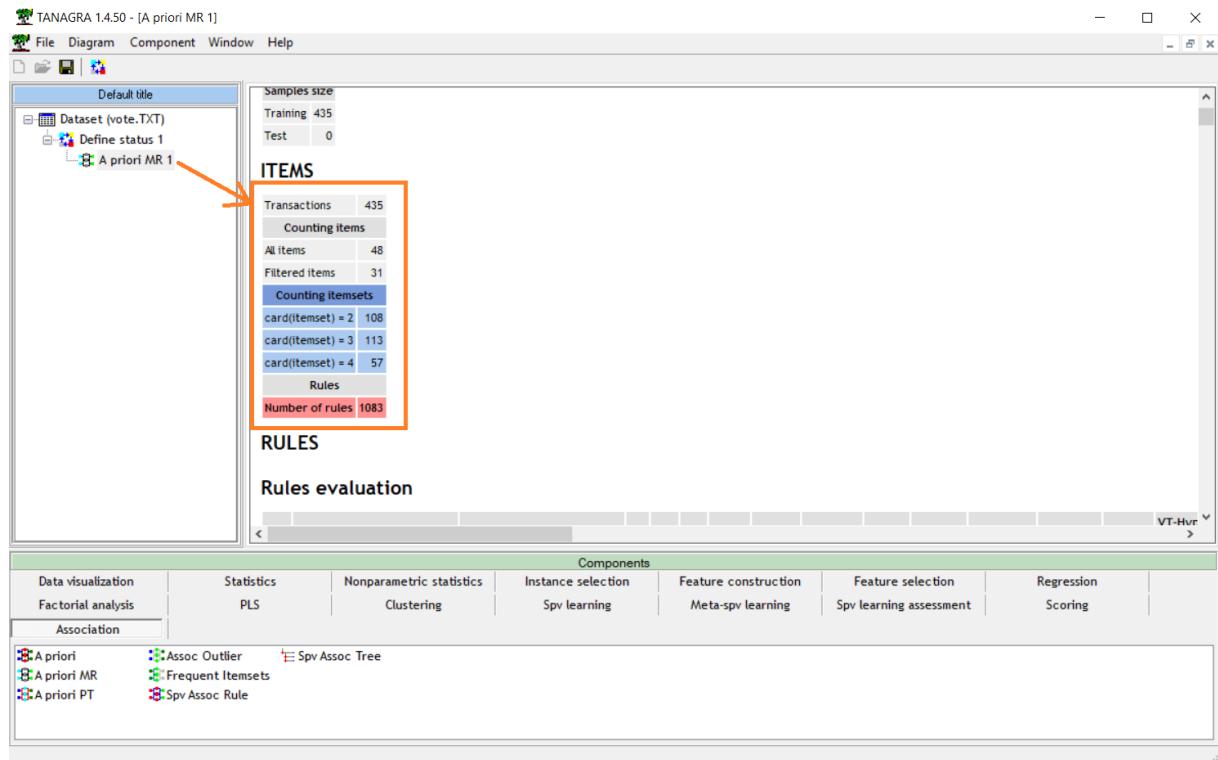
Thiết lập thông số cho component A PRIORI MR sẽ cho ra số lượng luật kết hợp khác nhau

Ứng dụng Tanagra trong khai phá dữ liệu



Hình 2. 26 Thiết lập thông số cho component A PRIORI MR

○ Kết quả sau khi thiết lập



Hình 2. 27 Kết quả luật kết hợp Apriori

CHƯƠNG 3: ÚNG DỤNG PHẦN MỀM TANAGRA

3.1. Mô tả dữ liệu

3.1.1 Mô tả chung

- Sơ lược về dữ liệu**

Trung tâm Chẩn đoán Sức khỏe JPAC (Jayaramdas Patel Academic Centre) đã tiến hành khảo sát tình trạng suy gan cấp tính đối với người trưởng thành Ấn Độ (trên 20 tuổi) trên toàn đất nước. Với đội ngũ nhân viên, y bác sĩ chuyên nghiệp, JPAC đã thu thập được nhiều thông tin nhân khẩu học và sức khỏe bằng hình thức phỏng vấn trực tiếp các đối tượng được tham gia khảo sát.

- Về kích thước dữ liệu**

Dữ liệu có đến hơn 8700 dòng và có 30 cột. Các thuộc tính đa dạng và xoay quanh sức khỏe, tiền sử bệnh lý và lối sống sinh hoạt của đối tượng.

- Mục tiêu sau khi phân tích dữ liệu**

Sau khi phân tích dữ liệu, chúng ta có thể rút ra nhiều kết luận về những tác nhân, hành vi gây ra bệnh suy gan cấp tính đối với một người. Từ đó chúng ta có thể tìm và đưa ra được phương án hợp lý để có thể phòng ngừa căn bệnh nguy hiểm này.

3.1.2 Mô tả chi tiết

Bảng 3. 1 Mô tả chi tiết các thuộc tính

STT	Tên thuộc tính	Dịch nghĩa	Tập giá trị
1	Age	Tuổi	Từ 20 đến 85 tuổi
2	Gender	Giới tính	Nam, nữ
3	Region	Khu vực sống	Phía đông, phía tây, phía nam, phía bắc nước Ấn Độ
4	Weight	Cân nặng	Từ 33.7 kg đến 193kg
5	Height	Chiều cao	Từ 130cm đến 200cm
6	BMI	Chỉ số BMI	Từ 12 đến 66

7	Obesity	Béo phì	Có, không
8	Waist	Số đo vòng 2	Từ 58.5 đến 173.4
9	Maximum Blood Pressure	Huyết áp tối đa	Từ 72 đến 233
10	Minimum Blood Pressure	Huyết áp tối thiểu	Từ 10 đến 132
11	Good Cholestorol (LDL Cholestorol)	Hàm lượng Cholestorol xấu	Từ 8 đến 160
12	Bad Cholestorol (HDL Cholestorol)	Hàm lượng Cholestorol tốt	Từ 27 đến 684
13	Total Cholestorol	Hàm lượng Cholestorol toàn phần	Từ 72 đến 727
14	Dyslipidemia	Rối loạn lipid máu	Có hoặc không
15	PVD(Peripheral vascular disease)	Bệnh mạch máu ngoại biên	Có hoặc không
16	Physical Activity	Hoạt động ngoài trời	Tối thiểu 0 hoạt động Tối đa 4 hoạt động
17	Education	Học thức	Có hoặc không
18	Unmarried	Độc thân	Có hoặc không
19	Income	Thu nhập	Có hoặc không
20	Source of Care	Nơi khám bệnh thường xuyên	Bệnh viện nhà nước, bệnh viện tư nhân, trạm y tế địa phương
21	PoorVision	Có vấn đề về mắt	Có hoặc không

22	Alcohol Consumption	Sử dụng thức uống có cồn	Có hoặc không
23	HyperTension	Bị tăng huyết áp	Có hoặc không
24	Family HyperTension	Gia đình có tiền sử tăng huyết áp	Có hoặc không
25	Diabetes	Bị tiểu đường	Có hoặc không
26	Family Diabetes	Gia đình có tiền sử bệnh tiểu đường	Có hoặc không
27	Hepatitis	Viêm gan	Có hoặc không
28	Family Hepatitis	Gia đình có tiền sử bệnh viêm gan	Có hoặc không
29	Chronic Fatigue	Hội chứng mệt mỏi mãn tính	Có hoặc không
30	ALF (Acute Liver Failure)	Suy gan cấp tính	Có hoặc không

3.1.3 Ưu điểm của cơ sở dữ liệu

- Cơ sở dữ liệu lớn và đa dạng các thuộc tính
- Cơ sở dữ liệu được trung tâm JPAC khảo sát trực tiếp với các đối tượng nên có độ uy tín cao.
- Quá trình khảo sát dữ liệu rất nghiêm ngặt, trải qua nhiều khâu như đo thể trạng, đo máu, kiểm tra các bệnh lý khác của đối tượng nên dữ liệu này có độ chi tiết cao.
- Dù dữ liệu lớn, nhưng không khó phân tích và áp dụng các thuật toán khai phá dữ liệu.

3.1.4 Nhược điểm của cơ sở dữ liệu

- Dữ liệu có vài thuộc tính và và dòng còn lỗi, chưa thể phân tích và áp dụng các thuật toán khai phá dữ liệu
- Dữ liệu không liên quan đến chuyên ngành công nghệ thông tin
- Dữ liệu về bệnh nhân mắc bệnh suy gan còn ít so với những người không có bệnh lý này (444 người mắc bệnh – 1167 người không mắc bệnh) .

Age	Gender	Region	Weight	Height	Body Mass Index	Obesity	Waist	Maximum Blood Pressure	Minimum Blood Pressure	Good Cholesterol	Bad Cholesterol	Total Cholesterol	Dyslipidem
65 M	east	56	162.1		21.31	0	83.6	135	71	48	249	297	
36 M	south	60.2	162.2		22.88	0	76.6	96	52	31	135	166	
66 M	east	83.9	162.5		31.77	1	113.2	115	57	44	211	255	
54 M	east	69.4	160.5		26.94	0	77.9	110	57	74	156	230	
63 M	north	73.1	159.2		28.84	0	89.3	132	73	67	154	221	
26 F	east	119.3	193.2		31.96	1	117.9	129	70	43	159	202	
66 F	north	85.1	172.1		28.73	0	99.2	137	92	41	143	184	
59 M	east	69.9	160.9		27	0	101.5	124	73	43	140	183	
53 M	east	75.2	174.1		24.81	0	85.6	110	74	62	110	172	
78 M	north	47.6	155.3		19.74	0	70.3	170	78	105	90	195	
47 F	east	99.6	188.2		28.12	0	95.1			63	162	225	
47 M	south	49	155.3		20.32	0	78.6	146	87	76	133	209	
62 F	south	56.1	165.5		20.48	0	78.7	201	119	55	171	226	
36 F	south	78.8	183.8		23.33	0	86.8	108	62	48	124	172	
60 M	south	68.3	146.7		31.74	1	88.5	153	77	38	141	179	
30 M	south	68.3	157.8		27.43	0	89.1	105	72	73	107	180	
47 M	south	62.1	149.6		27.75	0	89.5	116	71	49	135	184	
53 M	east	94.9	178		29.95	0	104.8	112	68	68	109	177	
28 M	east	65	156.8		26.44	0	85.9	105	69	56	219	275	
30 M	south	109.8	166.6		39.56	1	119.4	108	40	58	206	264	
52 F	east	96.8	186.4		27.86	0	105.9	133	77	70	108	178	
24 M	east	76.8	155.1		31.93	1	86.2	120	78	57	140	197	
38 M	south	67	153.7		28.36	0	90	144	90	48	163	211	
85 M	south	38.5						135		66	99	165	

Hình 3. 1 Dữ liệu chưa được xử lý (1)

PVD	Physical Activity	Education	Unmarried	Income	Source of Care	PoorVision	Alcohol Consumption	HyperTension	Family HyperTension	Diabetes	Family Diabetes	Hepatitis	Family Hepatitis	Chronic Fatigue	ALF
0	3	0	0	1	Government Hospital	0	1	0	0	0	1	1	0	0	1
0	3	0	0	1	Never Consulted	0	0	0	0	0	0	0	0	0	0
0	1	0	1	0	Never Consulted	0	1	0	0	1	0	0	0	0	0
0	2	1	0	0	Private Hospital	0	1	0	0	0	0	0	0	0	0
0	1	0	0	0	clinic	0	0	1	0	0	0	0	0	0	0
0	2	1	0	0	Private Hospital	0	0	0	1	0	0	0	0	0	0
0	3	1	0	0	Private Hospital	0	0	1	0	0	0	0	0	0	0
0	2	1	1	0	Private Hospital	0	0	0	1	0	1	1	0	0	0
0	1	1	0	0	1	0	1	1	1	1	0	0	0	0	0
0	1	0	1	0	Private Hospital	0	1	0	0	0	1	1	0	0	0
0	3	1	0	1	Private Hospital	0	1	0	0	0	0	0	0	0	0
0	3	0	0	1	Private Hospital	0	0	1	0	0	0	0	0	0	0
0	4	0	0	0	Private Hospital	0	1	1	0	0	0	0	0	0	0
0	2	1	1	0	clinic	0	0	0	0	0	0	1	0	0	0
0	2	0	0	0	Private Hospital	0	1	1	0	0	0	1	0	0	0
0	3	1	1	0	Private Hospital	0	1	0	1	0	1	0	0	0	0
0	1	0	0	0	Private Hospital	0	0	0	1	0	1	0	0	0	0
0	1	0	0	1	Private Hospital	0	0	0	1	0	1	0	0	0	0
0	1	0	0	0	Never Consulted	0	0	0	1	0	0	0	0	0	0
0	3	1	0	0	Never Consulted	0	0	0	1	0	0	0	0	0	0
0	2	0	1	0	Private Hospital	0	0	0	1	0	1	0	1	0	0
0	1	1	0	0	Private Hospital	0	1	1	1	1	0	0	0	0	0
0	2	1	0	1	Private Hospital	0	1	0	0	0	1	0	0	0	0
0	1	0	0	0	Private Hospital	0	0	1	1	1	0	1	0	0	0
0	1	0	1	1	Private Hospital	0	0	0	0	0	0	0	0	0	0
0	2	0	1	0	Never Consulted	0	0	0	1	0	0	0	0	0	0
0	3	1	0	1	Private Hospital	0	1	1	0	0	1	0	0	0	0
0	2	0	1	0	clinic	0	1	1	0	1	0	1	0	0	0
0	2	0	0	1	Private Hospital	0	1	0	0	0	1	0	0	0	0
0	2	0	0	1	1	0	1	0	1	0	1	0	0	0	0
0	2	0	0	0	Private Hospital	0	1	0	0	1	0	1	0	0	0

Hình 3. 2 Dữ liệu chưa được xử lý (2)

3.2. Tiền xử lý dữ liệu

- Loại bỏ các thuộc tính không phục vụ hoặc đáp ứng cho việc khai phá dữ liệu:
- Region, Weight, Height, BMI, Waist, Income, Family HyperTension, Family Diabetes, Family Hepatitis
- Các thuộc tính có giá trị rỗng trong Dataset cũng tiến hành loại bỏ

- File tiền xử lý dữ liệu được đặt tên viết liền, không ký tự đặc biệt, không dấu câu.
- Chuyển đổi file thành file arff thông qua công cụ Weka

Age	Gender	Obesity	Maximum_Blood_Pressure	Minimum_Blood_Pressure	Good_Cholesterol	Bad_Cholesterol	Total_Cholesterol	Dyslipidemia	PVD	Physical_J_Education	Unmarries	Source_of_Care	PoorVisor	Alcohol_C	HyperTension	Diabetes	Hepatitis_C	Chronic_Fa	ALF
20	M	0	112	76	67	131	198	0	0	2	0	1	Never Counselled	0	0	0	0	0	no
20	F	0	112	70	46	117	163	0	0	2	1	1	Private Hospital	0	0	0	0	0	no
20	F	0	111	57	43	126	169	0	0	4	0	1	Private Hospital	0	0	0	0	0	no
20	F	1	113	52	37	163	200	0	0	2	0	1	Private Hospital	0	0	0	0	0	no
20	M	0	100	55	52	140	192	0	0	2	0	1	Never Counselled	0	0	0	0	0	no
20	M	1	94	54	31	161	192	0	0	2	0	1	Never Counselled	0	1	0	0	0	no
20	M	0	111	59	45	126	171	0	0	2	0	1	clinic	0	0	0	0	0	no
20	F	0	105	72	44	114	158	0	0	2	0	1	Private Hospital	0	0	0	0	0	no
20	M	1	103	72	41	132	173	0	0	1	0	1	clinic	0	0	0	0	0	no
20	M	1	107	41	52	90	142	0	0	2	0	1	Never Counselled	1	0	0	0	0	no
20	F	1	121	57	44	109	153	0	0	2	0	1	Government Hospital	0	1	0	0	0	no
20	F	0	118	78	55	141	196	0	0	1	0	1	Never Counselled	0	0	0	0	0	no
20	M	0	104	54	64	106	170	0	0	2	0	0	Never Consulted	0	0	0	0	0	no
20	F	0	104	59	49	131	180	0	0	3	0	1	Clinic	0	1	0	0	0	no
20	M	0	102	28	66	203	269	0	0	1	0	0	Clinic	0	0	0	0	0	no
20	M	1	120	74	49	117	166	0	0	1	0	1	Clinic	0	0	0	0	0	no
20	M	0	106	71	51	113	164	0	0	2	1	1	Private Hospital	1	0	0	0	0	no
20	F	0	122	78	23	118	141	1	0	3	0	1	Never Counselled	0	0	0	0	0	no
20	M	0	90	62	79	122	201	0	0	2	0	0	Never Consulted	0	0	0	0	0	no
20	F	0	98	70	43	91	134	0	0	1	1	1	Never Counselled	0	1	0	0	0	no
20	F	0	120	76	44	105	149	0	0	2	0	1	Private Hospital	0	0	0	0	0	no

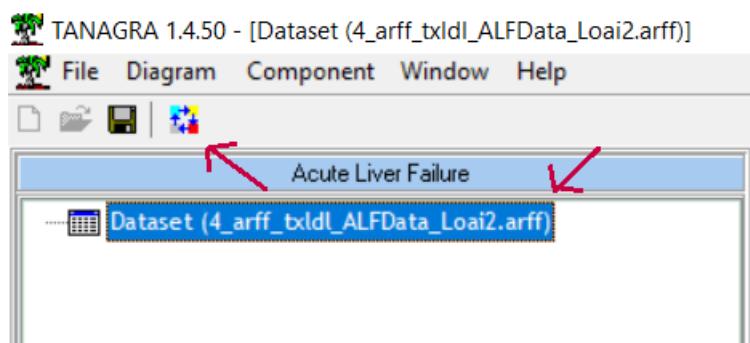
Hình 3. 3 Dữ liệu sau khi xử lý

3.3. Chuyển đổi dữ liệu thuộc biến liên tục (Discrete) thành dữ liệu thuộc biến rời rạc(Discrete)

Sử dụng EqFreq Disc sẽ chuyển đổi kiểu dữ liệu từ dữ liệu biến liên tục thành dữ liệu biến rời rạc nếu các dữ liệu biến liên tục quá nhiều, khiến chương trình không thể thực hiện được các thuật toán. Bằng cách sử dụng EqFreq Disc, ta sẽ cắt khoảng dữ liệu đó ra thành nhiều mảnh rời rạc. Từ đó chúng ta có thể nhìn sơ lược hoặc tổng quan hơn về thuật toán mà chúng ta dự định khai phá (Tuy nhiên nếu chúng ta thiết lập tham số cắt quá lớn có thể làm tăng phương sai, gây hạn chế cho thuật toán)

- **Cách thực hiện:**

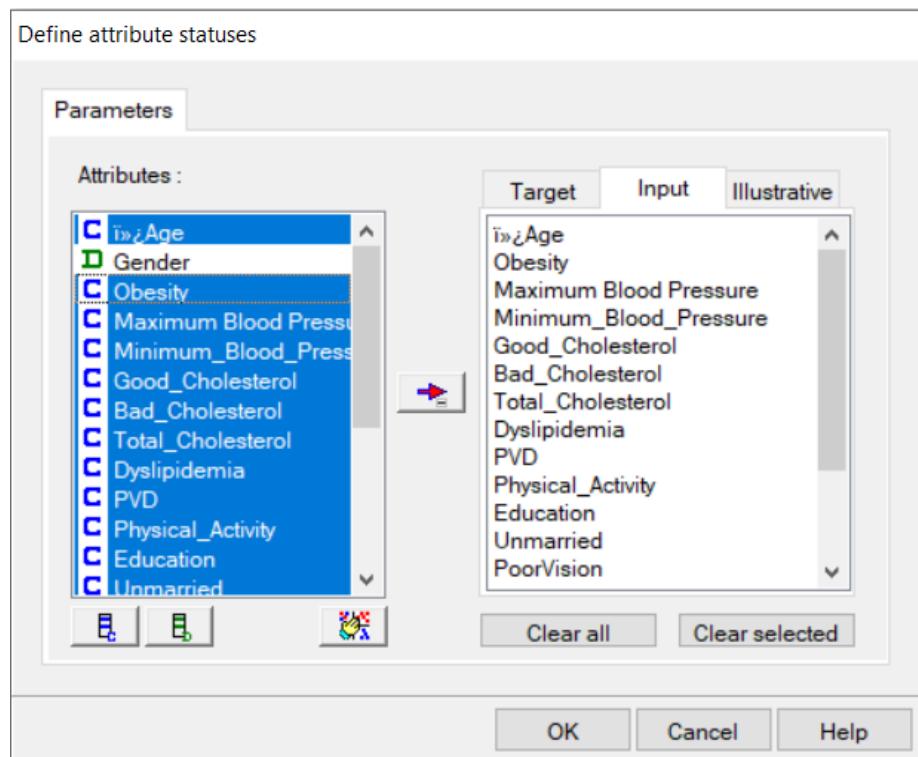
Trước tiên ta sẽ khởi động dữ liệu như ở mục 3.3. Sau đó ta nhấp chọn vào Dataset ta cần thực hiện và chọn vào biểu tượng để thêm thành phần Define Status vào Dataset.



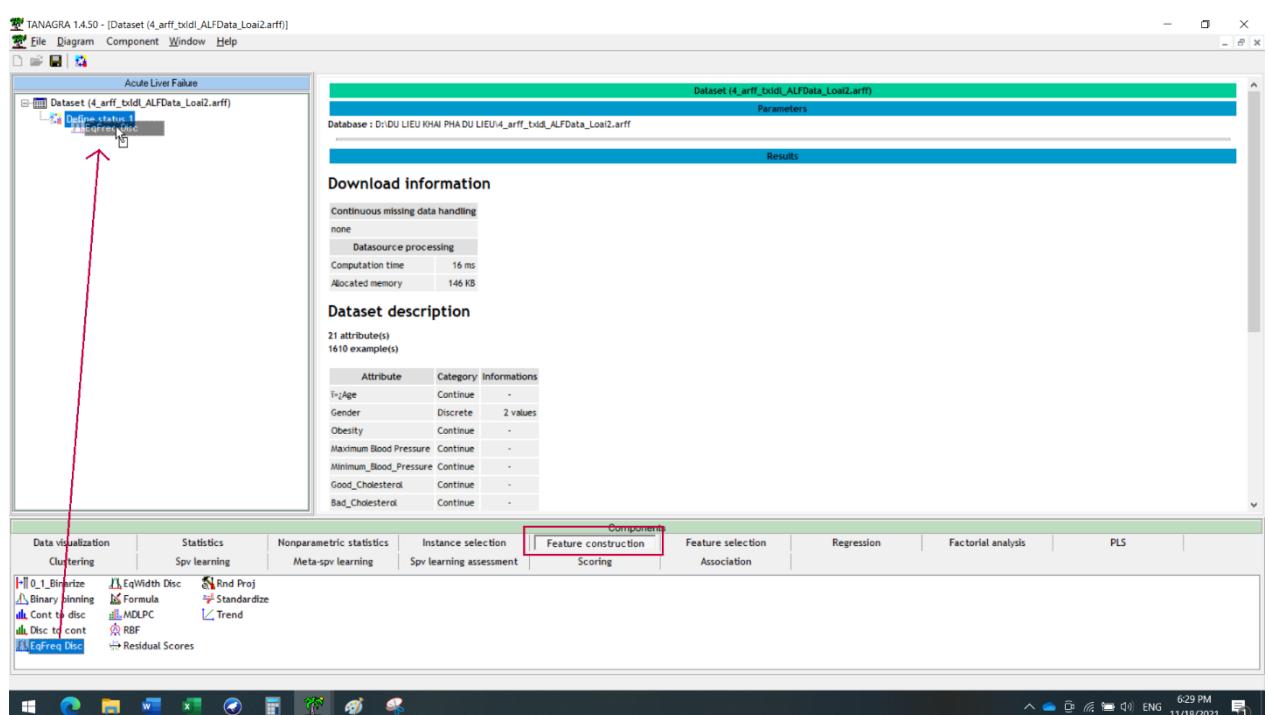
Hình 3. 4 Thêm thành phần Define Status và Dataset

Ứng dụng Tanagra trong khai phá dữ liệu

Hộp Define attribute statuses hiện ra, ta chọn vào các thuộc tính cần chuyển đổi (Có biểu tượng chữ C màu xanh bên cạnh) và chọn vào biểu tượng để dán vào ô Input



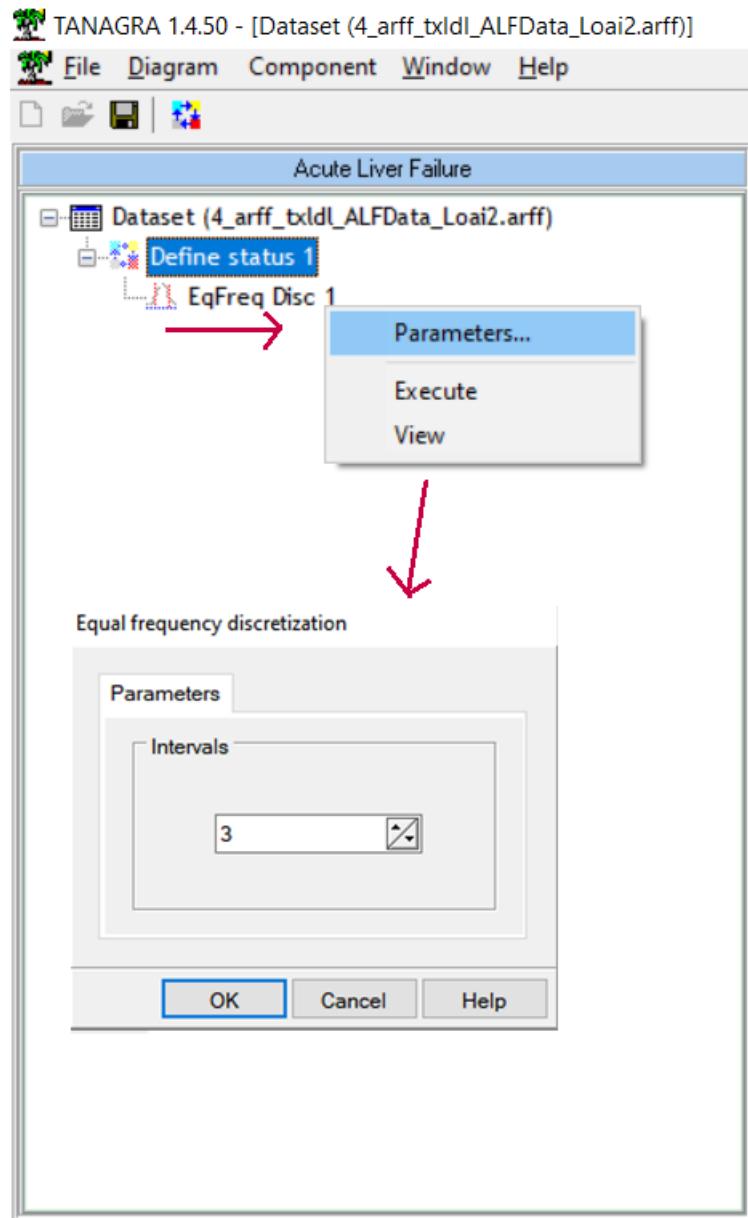
Hình 3. 5 Cho các dữ liệu cần chuyển đổi vào Input



Hình 3.6 Kéo thả công cụ EqFreq Disc và Define

Như hình trên, ta chọn vào ô Feature construction. Chọn vào công cụ EqFreq Disc và kéo thả vào Define.

Sau đó ta nhấn chuột phải vào EqFreq Disc, chọn Parameter để thiết lập số lượng khoảng phân rã dữ liệu của thuộc tính đó(như trong hình, các biến được sắp xếp rời rạc thành 3 khoảng). Sau đó chọn OK



Hình 3. 7 Thiết lập khoảng phân rã dữ liệu

Nhấp chuột phải vào Eq Freq Disc, chọn View để xem kết quả. Ta thấy, các giá trị chuyển đổi sẽ có ký tự d_eqF ở đầu

Data description

Attributes discretized	18
Examples	1610

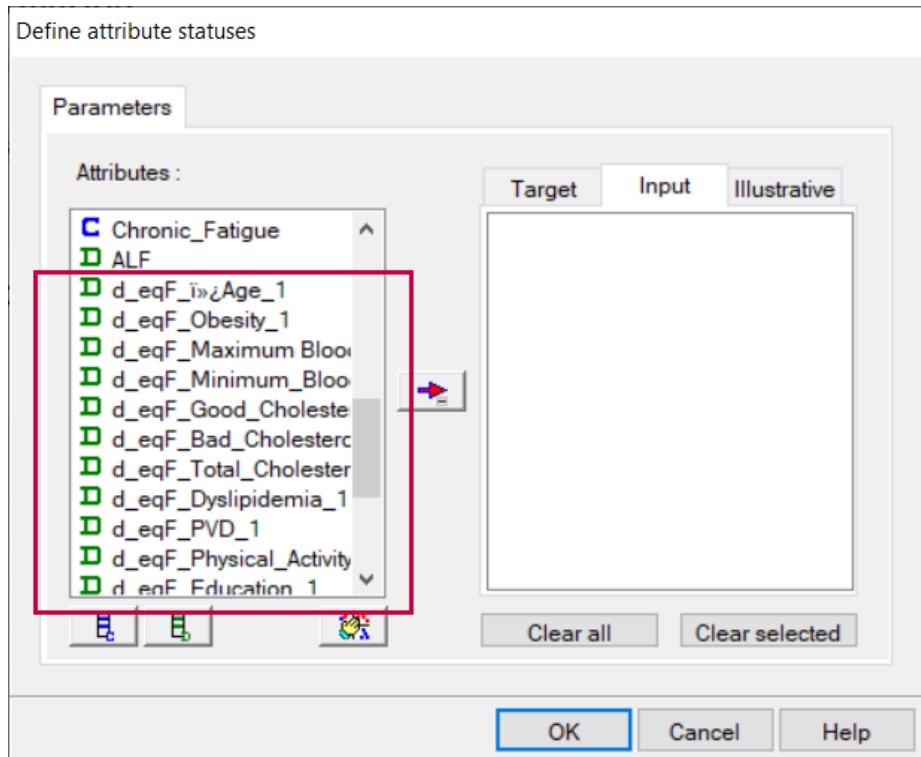
Generated attributes

Các điểm cắt

Source	New att	Intervals	Cut points
Age	d_eqF_Age_1	3	(42.0000 ; 66.0000)
Obesity	d_eqF_Obesity_1	1	0
Maximum Blood Pressure	d_eqF_Maximum_Blood_Pressure_1	3	(116.0000 ; 134.0000)
Minimum_Blood_Pressure	d_eqF_Minimum_Blood_Pressure_1	3	(66.0000 ; 76.0000)
Good_Cholesterol	d_eqF_Good_Cholesterol_1	3	(43.0000 ; 56.0000)
Bad_Cholesterol	d_eqF_Bad_Cholesterol_1	3	(133.0000 ; 167.0000)
Total_Cholesterol	d_eqF_Total_Cholesterol_1	3	(185.0000 ; 221.0000)
Dyslipidemia	d_eqF_Dyslipidemia_1	1	0
PVD	d_eqF_PVD_1	1	0
Physical_Activity	d_eqF_Physical_Activity_1	2	(2.0000)
Education	d_eqF_Education_1	2	(1.0000)
Unmarried	d_eqF_Unmarried_1	2	(1.0000)
PoorVision	d_eqF_PoorVision_1	1	0
Alcohol_Consumption	d_eqF_Alcohol_Consumption_1	1	0
HyperTension	d_eqF_HyperTension_1	2	(1.0000)
Diabetes	d_eqF_Diabetes_1	1	0
Hepatitis	d_eqF_Hepatitis_1	1	0
Chronic_Fatigue	d_eqF_Chronic_Fatigue_1	1	0

Hình 3. 8 Các dữ liệu đã được chuyển đổi

Thực hiện Define lần nữa, sau đó ta có thể sử dụng các giá trị sau khi chuyển đổi để tiến hành thực hiện các thuật toán phù hợp.



Hình 3. 9 Đẩy các giá trị sau khi chuyển đổi vào ô Input

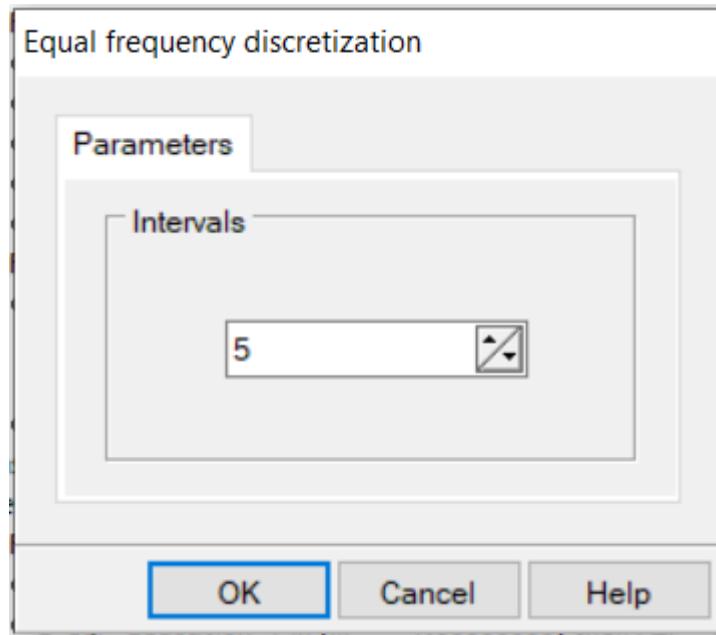
3.4. Xây dựng cây ra quyết định với Tanagra

Mục tiêu của thuật toán cây ra quyết định trên tập dữ liệu này là cho chúng ta có cái nhìn tổng quát về các thuộc tính nào quyết định trực tiếp đến kết quả suy gan cấp tính trên những người trưởng thành ở ẩn độ.

3.4.1 Cách thực hiện

Bước 1: Chuyển đổi các dữ liệu cần thiết.

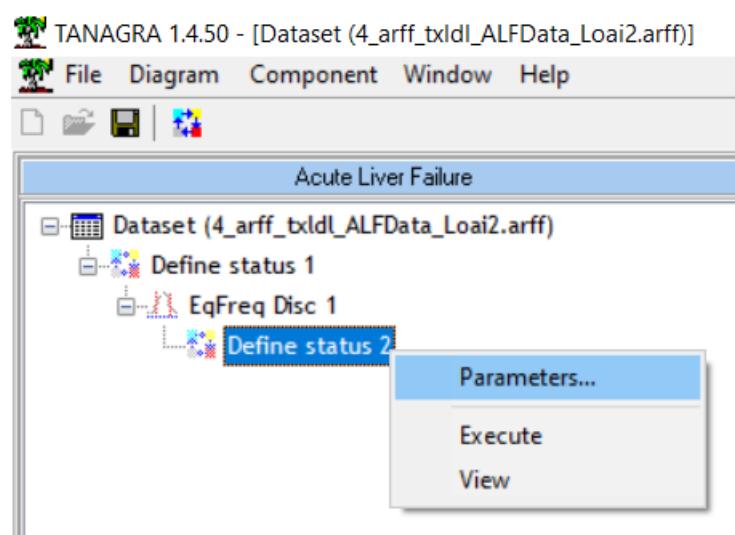
Trước tiên ta sẽ chuyển đổi dữ liệu như ở mục 3.3 (Ở đây ta sẽ chọn số lượng khoản phân rã bằng 5)



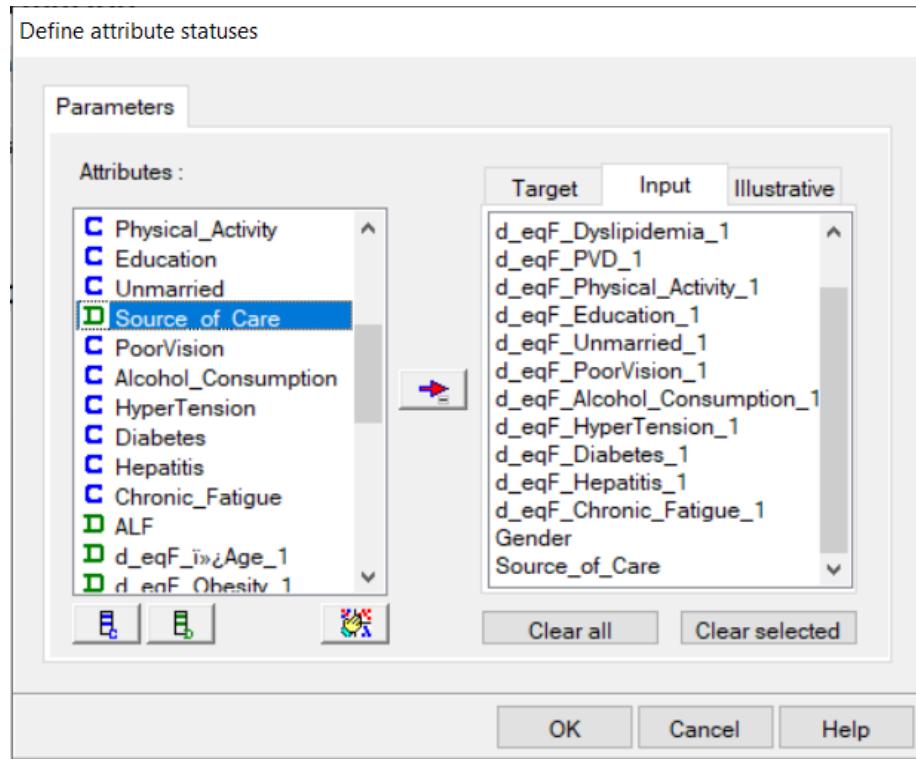
Hình 3. 10 Thiết lập số khoản phân rã

Bước 2: Xác định các thuộc tính chạy thuật toán

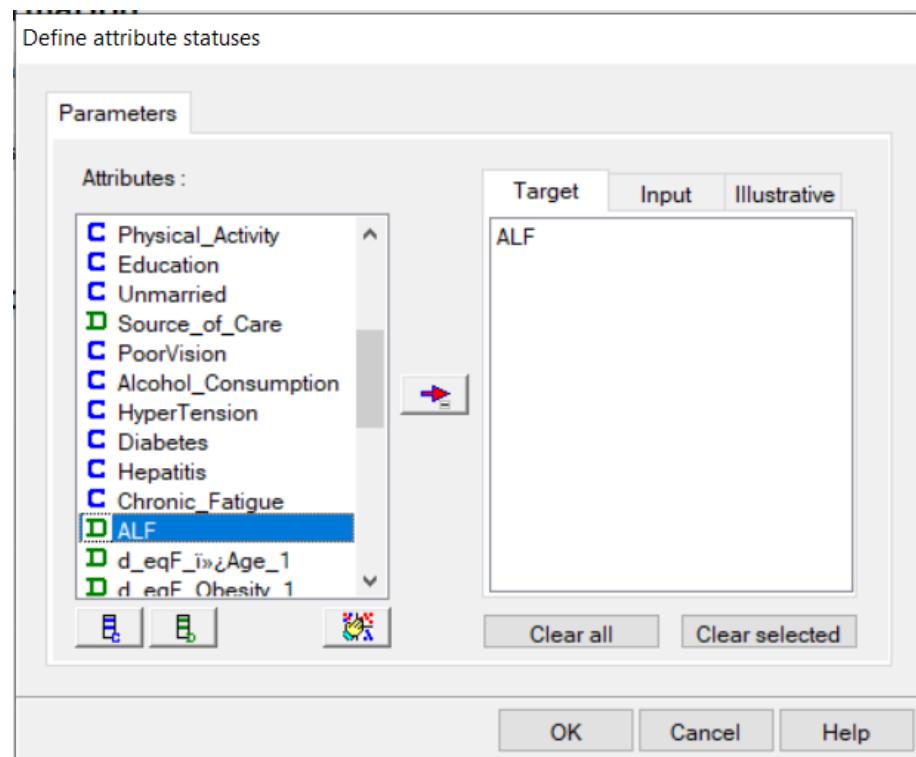
Sau đó ta chọn Define Status 2 → Parameters.. để thêm các thuộc tính cần thiết cho thuật toán



Hình 3. 11 Thêm các thuộc tính cần thiết cho thuật toán(1)



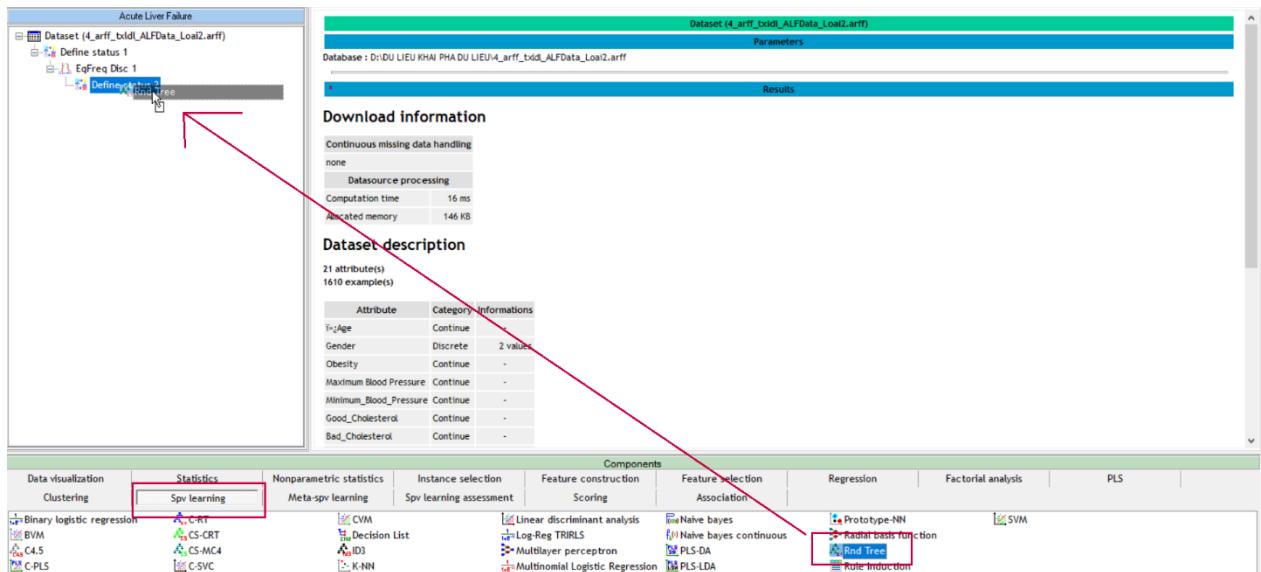
Hình 3. 12 Thêm các thuộc tính cần thiết vào Input



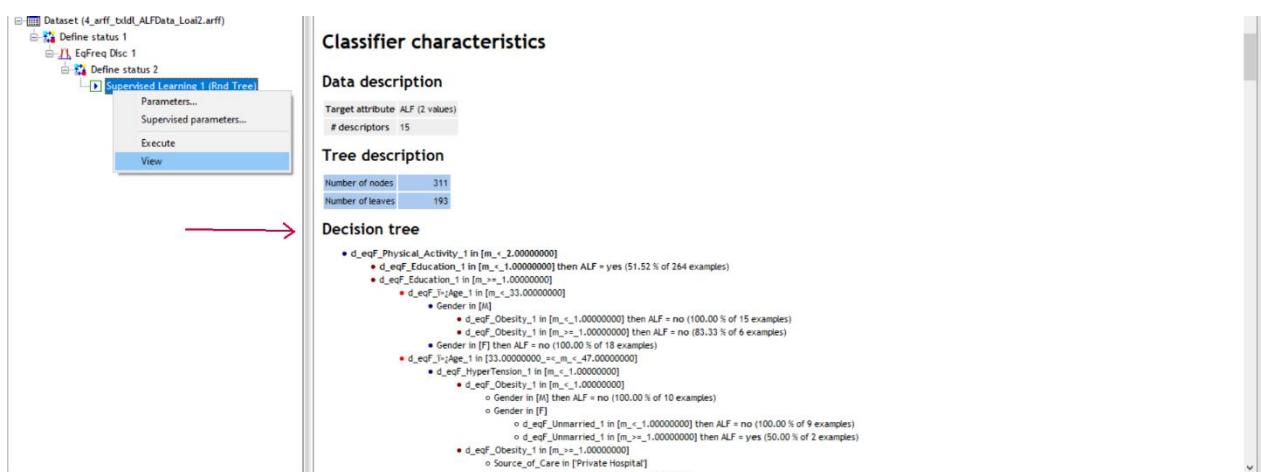
Hình 3. 13 Thêm thuộc tính ALF vào Target

Bước 3: Sử dụng thuật toán Decision Tree

Sử dụng công cụ Rnd Tree để xây dựng cây ra quyết định.



Hình 3. 14 Sử dụng công cụ End Tree



Hình 3. 15 Kết quả sau khi chạy thuật toán

3.4.2 Kết quả chạy thuật toán

Cây quyết định hỗ trợ cho chúng ta ra được các quyết định trong việc các thuộc tính nào ảnh hưởng trực tiếp đến việc xếp loại ALF trong độ tuổi trưởng thành. Thì hình ảnh trên cho ta thấy được có đến 311 nodes và 193 nhánh hỗ trợ ra quyết định trong thuật toán ta vừa thực hiện trên.

Một số nhánh nổi bậc như sau:

[1]. Với những nữ giới có độ tuổi từ 33 cho đến 66 tuổi, tham gia trên 3 hoạt động thể thao, thăm khám tại các bệnh viện tư nhân, không bị béo phì, không bị tăng huyết áp, không sử dụng thức uống có cồn thì khả năng bị bệnh suy gan là gần như không có (lên đến 100% cho 14 trường hợp trong nhóm này)

- d_eqF_Physical_Activity_1 in [m_>= 3.00000000]
 - Source_of_Care in ['Private Hospital']
 - d_eqF_Obesity_1 in [m_<_1.00000000]
 - d_eqF_Alcohol_Consumption_1 in [m_<_1.00000000]
 - Gender in [M] then ALF = no (84.09 % of 44 examples)
 - Gender in [F]
 - d_eqF_HyperTension_1 in [m_<_1.00000000]
 - d_eqF_i>Age_1 in [m_<_33.00000000] then ALF = no (100.00 % of 9 examples)
 - d_eqF_i>Age_1 in [33.00000000_<_m_<_47.00000000] then ALF = no (100.00 % of 14 examples)
 - d_eqF_i>Age_1 in [47.00000000_<_m_<_62.00000000] then ALF = no (100.00 % of 6 examples)

Hình 3. 16 Kết quả chạy thuật toán (1)

[2]. Với những nam giới ít học vấn, chưa kết hôn, có tiêu thụ rượu và bị tăng huyết áp thì khả năng bị suy gan là khá cao (lên đến 70% cho 40 trường hợp) ở thanh màu đỏ. Nếu người đó không tiêu thụ rượu mà độ tuổi trên 85 thì khả năng bị suy gan còn cao hơn trường hợp trên (lên đến 83.08% cho 65 trường hợp) ở thanh màu xanh

- d_eqF_Education_1 in [m_<_1.00000000]
 - d_eqF_Unmarried_1 in [m_>= _1.00000000]
 - Gender in [M]
 - d_eqF_Alcohol_Consumption_1 in [m_<_1.00000000]
 - d_eqF_i>Age_1 in [m_<_33.00000000] then ALF = no (100.00 % of 34 examples)
 - d_eqF_i>Age_1 in [33.00000000_<_m_<_47.00000000]
 - d_eqF_Obesity_1 in [m_<_1.00000000]
 - d_eqF_Physical_Activity_1 in [m_<_2.00000000] then ALF = no (100.00 % of 2 examples)
 - d_eqF_Physical_Activity_1 in [2.00000000_<_m_<_3.00000000] then ALF = no (87.50 % of 8 examples)
 - d_eqF_Physical_Activity_1 in [m_>= _3.00000000] then ALF = no (100.00 % of 3 examples)
 - d_eqF_Obesity_1 in [m_>= _1.00000000] then ALF = no (100.00 % of 9 examples)
 - d_eqF_i>Age_1 in [47.00000000_<_m_<_62.00000000] then ALF = no (83.33 % of 30 examples)
 - d_eqF_i>Age_1 in [62.00000000_<_m_<_75.00000000]
 - d_eqF_HyperTension_1 in [m_<_1.00000000] then ALF = no (75.00 % of 8 examples)
 - d_eqF_HyperTension_1 in [m_>= _1.00000000] then ALF = no (61.11 % of 36 examples)
 - d_eqF_i>Age_1 in [m_>= _75.00000000] then ALF = yes (83.08 % of 65 examples)
 - d_eqF_Alcohol_Consumption_1 in [m_>= _1.00000000]
 - d_eqF_HyperTension_1 in [m_<_1.00000000] then ALF = no (65.22 % of 23 examples)
 - d_eqF_HyperTension_1 in [m_>= _1.00000000] then ALF = yes (75.00 % of 40 examples)

Hình 3. 17 Kết quả chạy thuật toán 2

3.5. Xây dựng thuật toán NaiveBayes với Tanagra

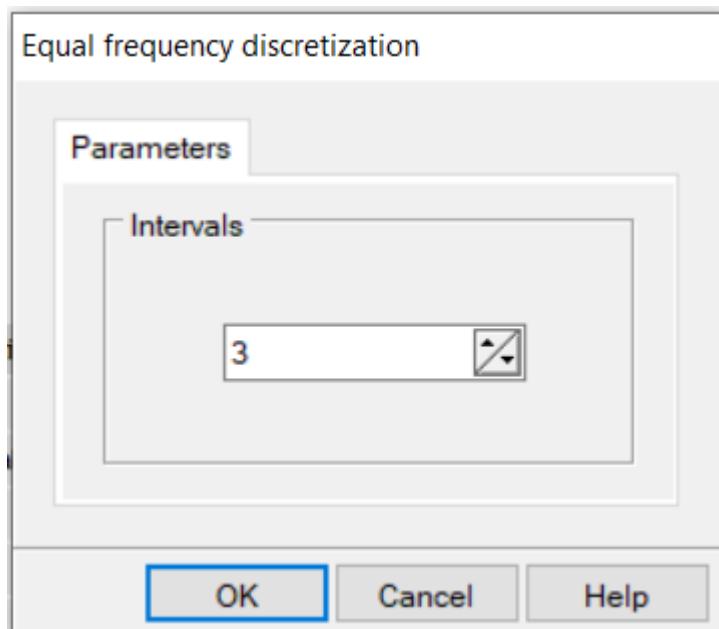
Bộ phân lớp Bayes là một giải thuật thuộc lớp giải thuật thống kê, nó có thể dự đoán xác suất của một phần tử dữ liệu thuộc vào một lớp là bao nhiêu. Trong tập dữ liệu này

NaiveBayes sẽ phân tích được xác xuất các thuộc tính trên phần tử dữ liệu sẽ thuộc vào lớp ALF yes/no là bao nhiêu.

3.5.1 Cách thực hiện

Bước 1: Chuyển đổi các dữ liệu cần thiết.

Trước tiên ta sẽ chuyển đổi dữ liệu như ở mục 3.3 (Ở đây ta sẽ chọn số lượng khoản phân rã bằng 3)

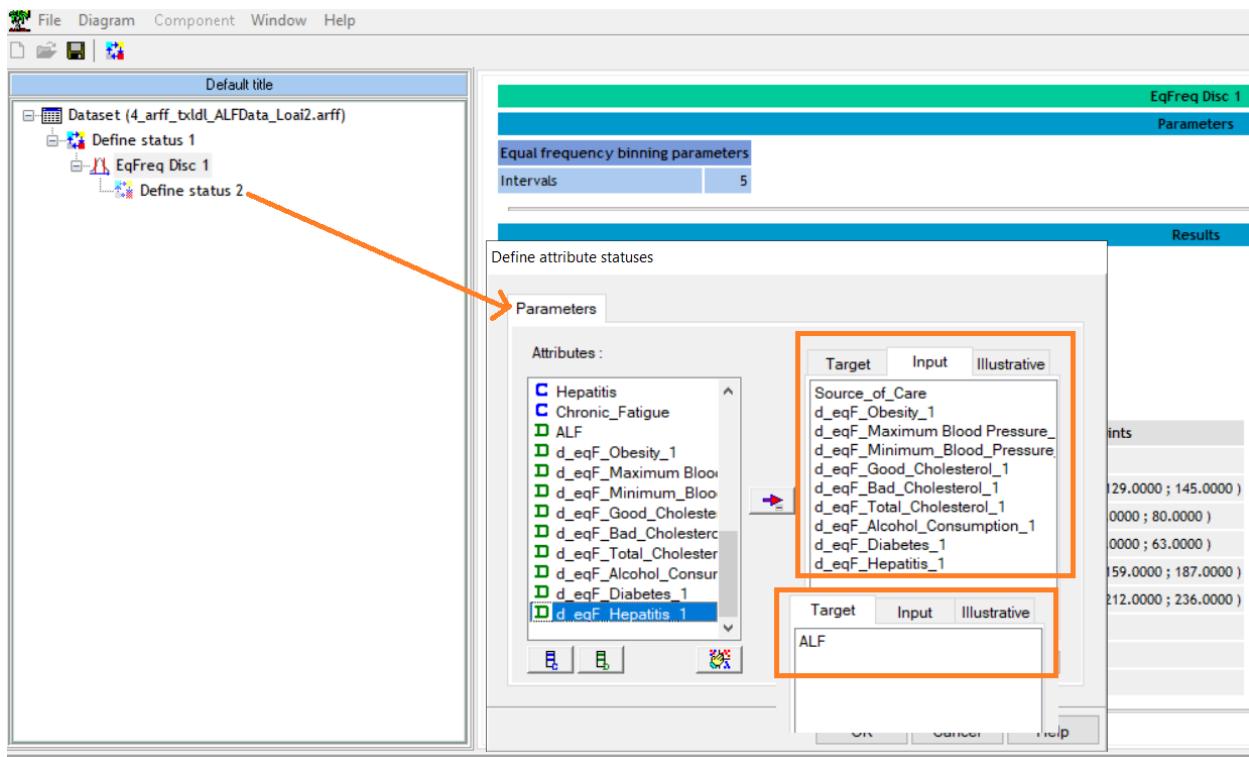


Hình 3. 18 Thiết lập số khoản phân rã

Bước 2: Xác định các thuộc tính chạy thuật toán

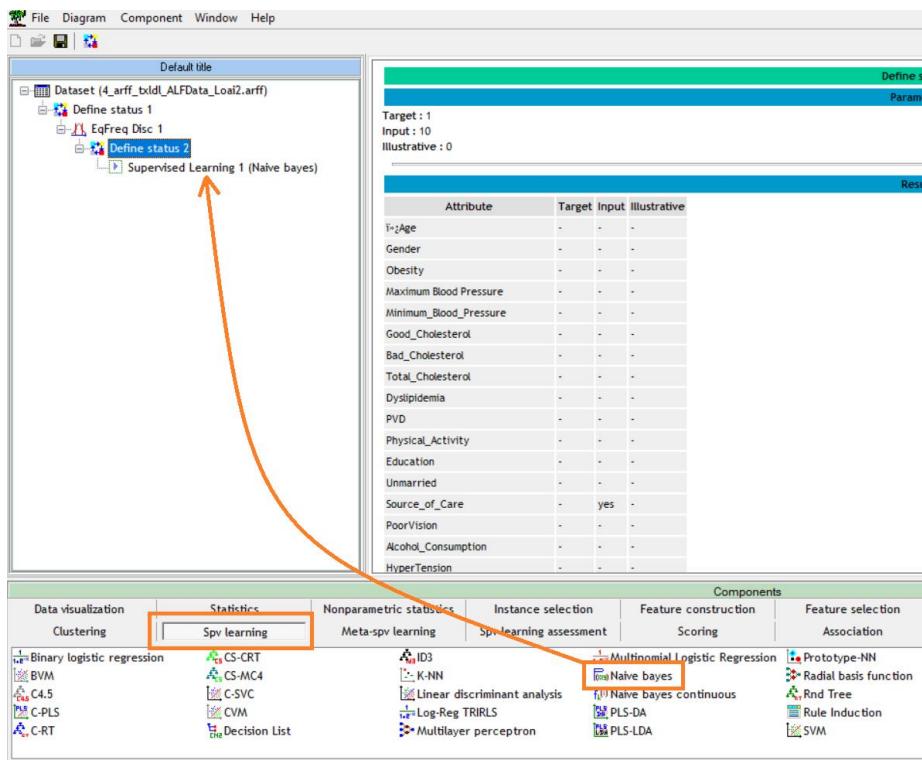
Sau đó ta chọn Define Status 2 → Parameters.. để thêm các thuộc tính cần thiết cho thuật toán

Ứng dụng Tanagra trong khai phá dữ liệu



Hình 3. 19 Xác định các thuộc tính chạy thuật toán

Bước 3: Sử dụng thuật toán NaiveBayes



Hình 3. 20 Thêm thành phần NAIVE BAYES (tab SPV LEARNING) vào sơ đồ.

Bước 4: Bấm vào menu View để nhận kết quả:

Classifier characteristics

Data description

Target attribute	ALF (2 values)
# descriptors	10

Prior distribution of class attribute "ALF"

Values	Count	Percent	Histogram
yes	442	27.45 %	
no	1168	72.55 %	

Model description

Descriptors	Classification functions	
	yes	no
Source_of_Care = 'Private Hospital'	3.028522	1.167714
Source_of_Care = clinic	1.876917	0.182322
Source_of_Care = 'Gouvernement Hospital'	0.427444	-1.195876
d_eqF_Obesity_1 = m_<_1.00000000	0.613104	0.855451
d_eqF_Maximum Blood Pressure_1 = m_<_110.00000000_=<_m_<_119.00000000	-1.655423	0.531003
d_eqF_Maximum Blood Pressure_1 = 110.00000000_=<_m_<_129.00000000	-1.375121	0.556882
d_eqF_Maximum Blood Pressure_1 = 119.00000000_=<_m_<_129.00000000	-0.933288	0.485018

Hình 3. 21 Bấm vào menu View để nhận kết quả(1)

d_eqF_Maximum_Blood_Pressure_1 = 129.000000000_=<_m_<_145.000000000	-0.394292	0.337382
d_eqF_Minimum_Blood_Pressure_1 = m_<_61.000000000	0.513354	-0.398423
d_eqF_Minimum_Blood_Pressure_1 = 61.000000000_=<_m_<_69.000000000	0.096460	-0.177387
d_eqF_Minimum_Blood_Pressure_1 = 69.000000000_=<_m_<_74.000000000	-0.078988	-0.207368
d_eqF_Minimum_Blood_Pressure_1 = 74.000000000_=<_m_<_80.000000000	-0.038715	-0.194408
d_eqF_Good_Cholesterol_1 = m_<_39.000000000	0.310155	-0.209973
d_eqF_Good_Cholesterol_1 = 39.000000000_=<_m_<_45.000000000	0.177983	-0.282644
d_eqF_Good_Cholesterol_1 = 45.000000000_=<_m_<_53.000000000	0.210071	-0.087011
d_eqF_Good_Cholesterol_1 = 53.000000000_=<_m_<_63.000000000	0.012903	-0.038615
d_eqF_Bad_Cholesterol_1 = m_<_118.000000000	-0.246860	-0.012903
d_eqF_Bad_Cholesterol_1 = 118.000000000_=<_m_<_139.000000000	-0.194900	0.062132
d_eqF_Bad_Cholesterol_1 = 139.000000000_=<_m_<_159.000000000	-0.075712	-0.012903
d_eqF_Bad_Cholesterol_1 = 159.000000000_=<_m_<_187.000000000	0.117783	-0.025975
d_eqF_Total_Cholesterol_1 = m_<_170.000000000	-0.209092	-0.017622
d_eqF_Total_Cholesterol_1 = 170.000000000_=<_m_<_191.000000000	-0.152580	0.030110
d_eqF_Total_Cholesterol_1 = 191.000000000_=<_m_<_212.000000000	-0.281412	0.099667
d_eqF_Total_Cholesterol_1 = 212.000000000_=<_m_<_236.000000000	-0.232622	0.004357
d_eqF_Alcohol_Consumption_1 = m_<_1.000000000	0.355034	0.925883
constant	-14.006206	-12.704009

Hình 3. 22 Bấm vào menu View để nhận kết quả(2)

3.5.2 Kết quả chạy thuật toán

Kết quả trả về cho ta biết được xác suất xuất hiện của từng giá trị trong lớp yes hoặc no của lớp ALF chi tiết các giá trị có thể quan sát trong hình trên.

3.6. Xây dựng thuật toán gom cụm K-Means với Tanagra

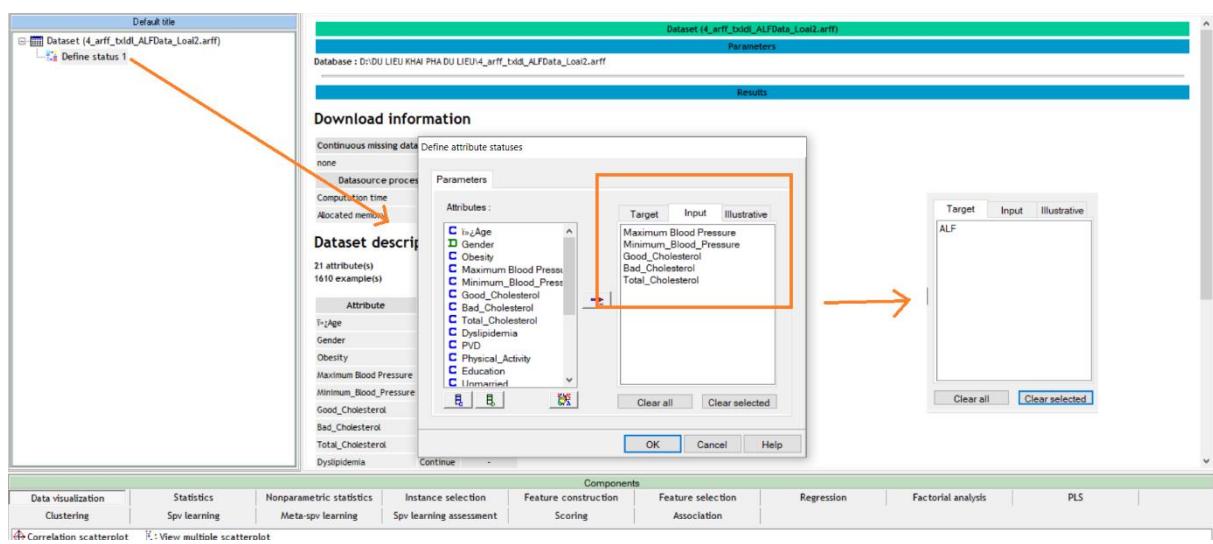
Mục tiêu của việc xây dựng thuật toán gom cụm ứng dụng trong dataset này là phân tích được những thuộc tính có giá trị tương đương nhau để gom lại thành các cụm. Từ đó thấy được các cụm này có các chỉ số về huyết áp, Cholesterol,.. như thế nào và cũng cho ta biết được số lượng các nhóm này. Qua đó đánh giá nguy cơ mắc bệnh suy gan cấp tính cho người trung niên.

3.6.1 Cách thực hiện

Bước 1: Xác định thành phần cần phân tích

Muốn có được một cái nhìn tổng quan về các đặc điểm chính của tập dữ liệu. Ta thêm một DEFINE STATUS thành phần vào sơ đồ. Đặt các biến liên tục vào INPUT, ALF vào Target

Đây là các biến hoạt động của phân tích, tức là chúng được sử dụng trong quá trình phân nhóm.

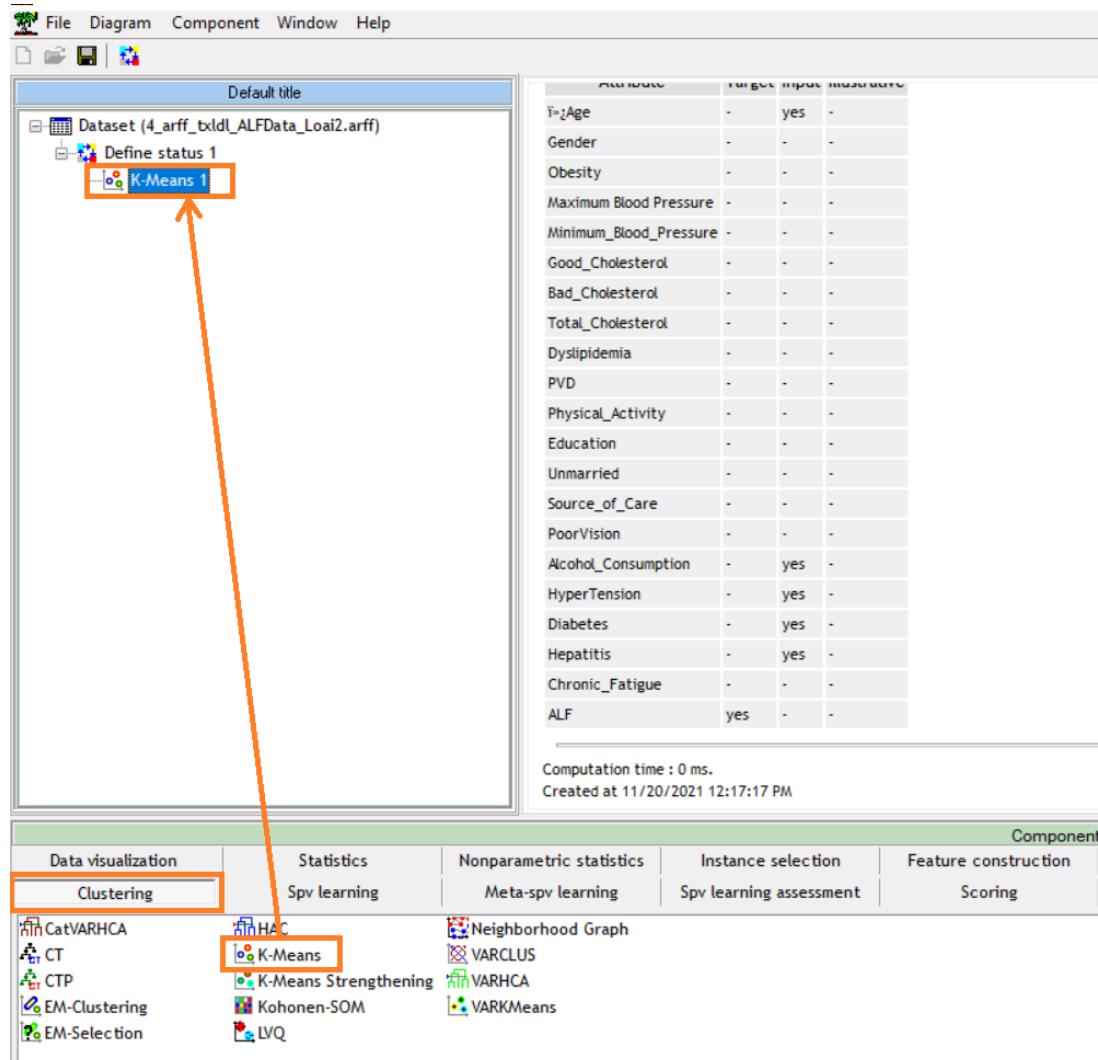


Hình 3. 23 Xác định thành phần cần phân tích

Bước 2 chèn thuật toán K – Means vào Define vừa tạo

Để chuẩn hóa các biến trước khi thực hiện phương pháp k - mean. Mục đích là để loại bỏ sự khác biệt của thang đo giữa các biến. Ta thêm thành phần STANDARDIZE (Tab FEATURE CONSTRUCTION) vào sơ đồ. Sau đó, chúng ta bấm vào menu View. Nhưng trên dataset hiện tại các thuộc tính có giá trị liên tục không quá lớn nên sẽ không sử dụng chuẩn hóa dữ liệu.

Chúng ta chèn thành phần K - MEANS (tab CLUSTERING) => vào Define Status

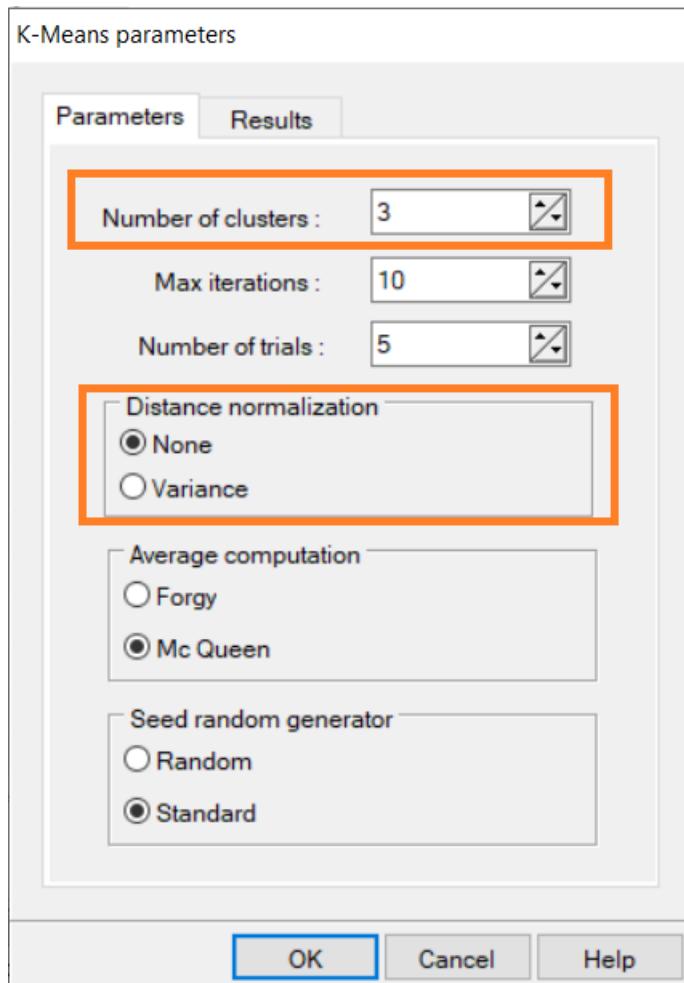


Hình 3. 24 Chèn thuật toán K – Means vào Define vừa tạo

Bước 3: Thiết lập thuật toán

Nhấp vào menu ngữ cảnh PARAMETERS => thiết lập các thông số sau:

- Number of Clusters: số cụm bạn muốn gom
- Distance normalization: None



Hình 3. 25 Thiết lập thông số cho thuật toán

Bước 4: Kết quả của thuật toán gom cụm:

Global evaluation

Within Sum of Squares	2763323.6935
Total Sum of Squares	6982829.4714
R-Square	0.6043

Cluster size and WSS

Clusters	3		
Cluster	Description	Size	WSS
cluster n°1	c_kmeans_1	729	1144049.4851
cluster n°2	c_kmeans_2	613	975448.2413
cluster n°3	c_kmeans_3	268	643825.9670

R-Square for each attempt

Number of trials	5
Trial	R-square
1	0.604269
2	0.603794
3	0.603740
4	0.603906
5	0.603754

Cluster centroids

Attribute	Cluster n°1	Cluster n°2	Cluster n°3
Maximum_Blood_Pressure	131.310014	123.601958	132.179104
Minimum_Blood_Pressure	72.160494	68.504078	71.246269
Good_Cholesterol	52.567901	50.876020	52.123134
Bad_Cholesterol	161.089163	114.590538	217.365672
Total_Cholesterol	213.657064	165.466558	269.488806

Hình 3. 26 Kết quả của thuật toán gom cụm

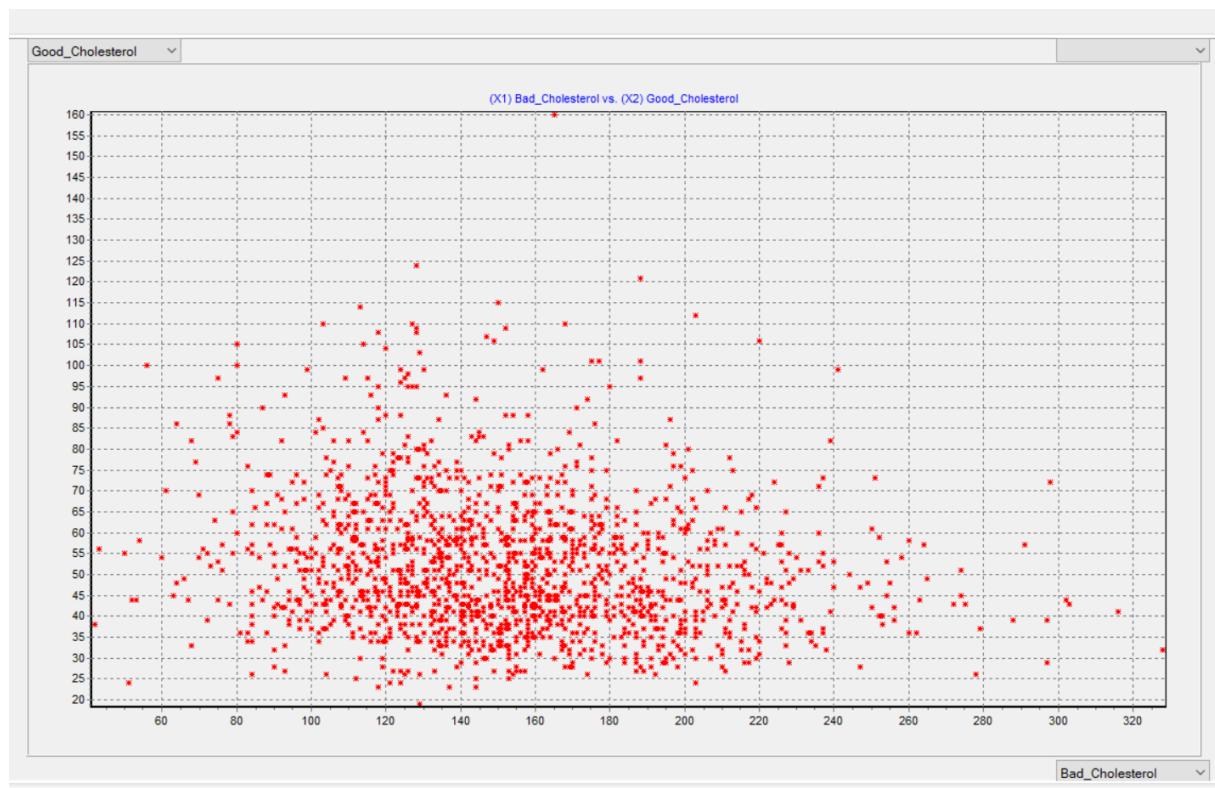
3.6.2 Ý nghĩa kết quả của thuật toán K-Means

Kết quả trả về cho ta thông tin về 3 cụm dữ liệu như sau:

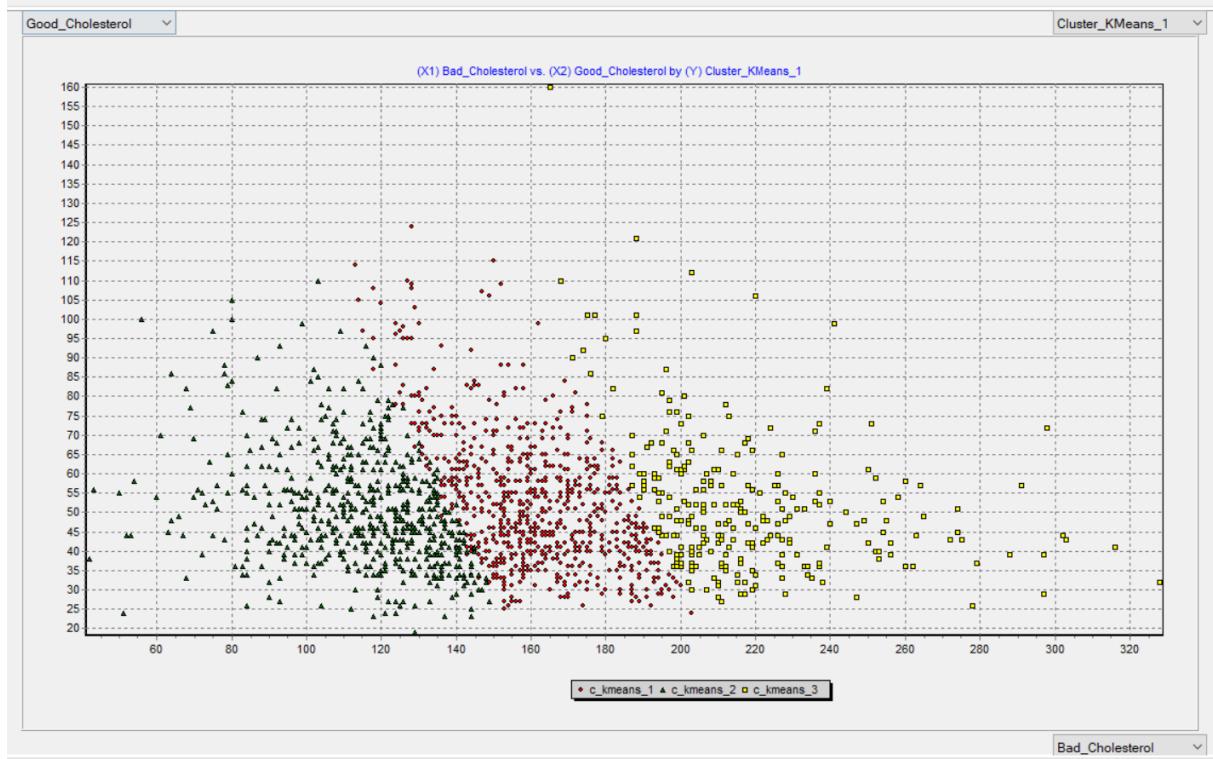
- Cụm 1: 729 người
- Cụm 2: 613 người
- Cụm 3: 286 người

Và một số điểm trung tâm của các thuộc tính thuộc các cụm như trên. Ví dụ như ta thấy cụm 1: Age sẽ có điểm trung tâm là 131.3100, Alcohol_Consumption: 72.160, HyperTension: 52.567, Diabetes: 161.089 , Hepatitis: 213.657.

Chi tiết ta quan sát biểu đồ bên dưới:



Hình 3. 27 Dữ liệu chưa phân cụm



Hình 3. 28 Dữ liệu sau khi phân cụm

Chúng ta thấy được khi hiển thị các cụm này trên biểu đồ 2 giá trị đó là Good_Cholesterol và Bad_Cholesterol thì có thể thấy rõ:

- Cụm 1: Người có lượng Bad_Cholesterol từ khoảng 60 - 140 và lượng Good_Cholesterol từ khoảng 25 - 100
- Cụm 2: Người có lượng Bad_Cholesterol từ khoảng 120 – 200 và lượng Good - Cholesterol từ khoảng 25 - 110
- Cụm 3: Người có lượng Bad_Cholesterol từ khoảng 180 – 300 và lượng Good - Cholesterol từ khoảng 25 - 100

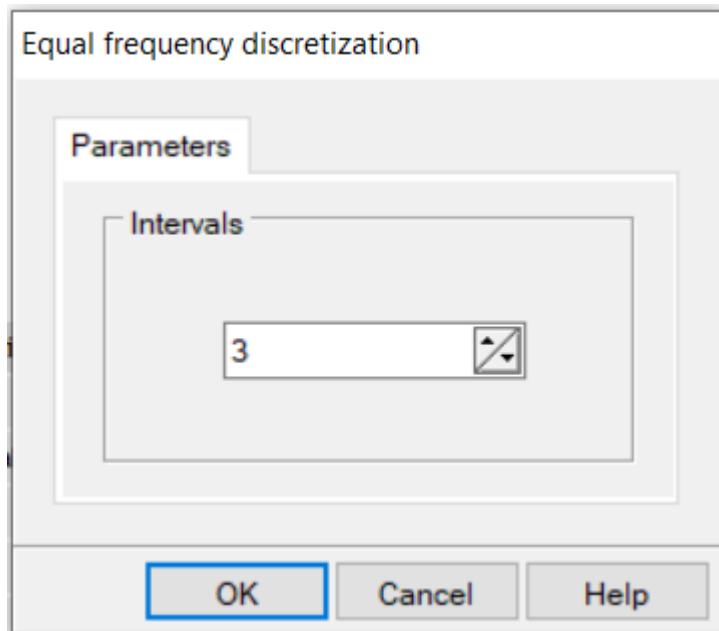
3.7. Xây dựng thuật toán luật kết hợp Apriori với Tanagra

Xây dựng thuật kết hợp trên tập dữ liệu này với mục đích tìm ra các quy luật về đánh giá mức mắc bệnh suy gan cấp tính (ALF) phụ thuộc vào những yếu tố nào? Và cũng phát hiện ra các quy luật để tìm ra các thuộc tính ảnh hưởng nhiều đến việc xếp loại ALF trên người trung niên ở ánh độ. Từ đó hướng đến các giải pháp để giảm thiểu số lượng mắc bệnh suy gan cấp tính.

3.7.1 Cách thực hiện

Bước 1: Chuyển đổi các dữ liệu cần thiết.

Trước tiên ta sẽ chuyển đổi dữ liệu như ở mục 3.3 (Ở đây ta sẽ chọn số lượng khoản phân rã bằng 3)

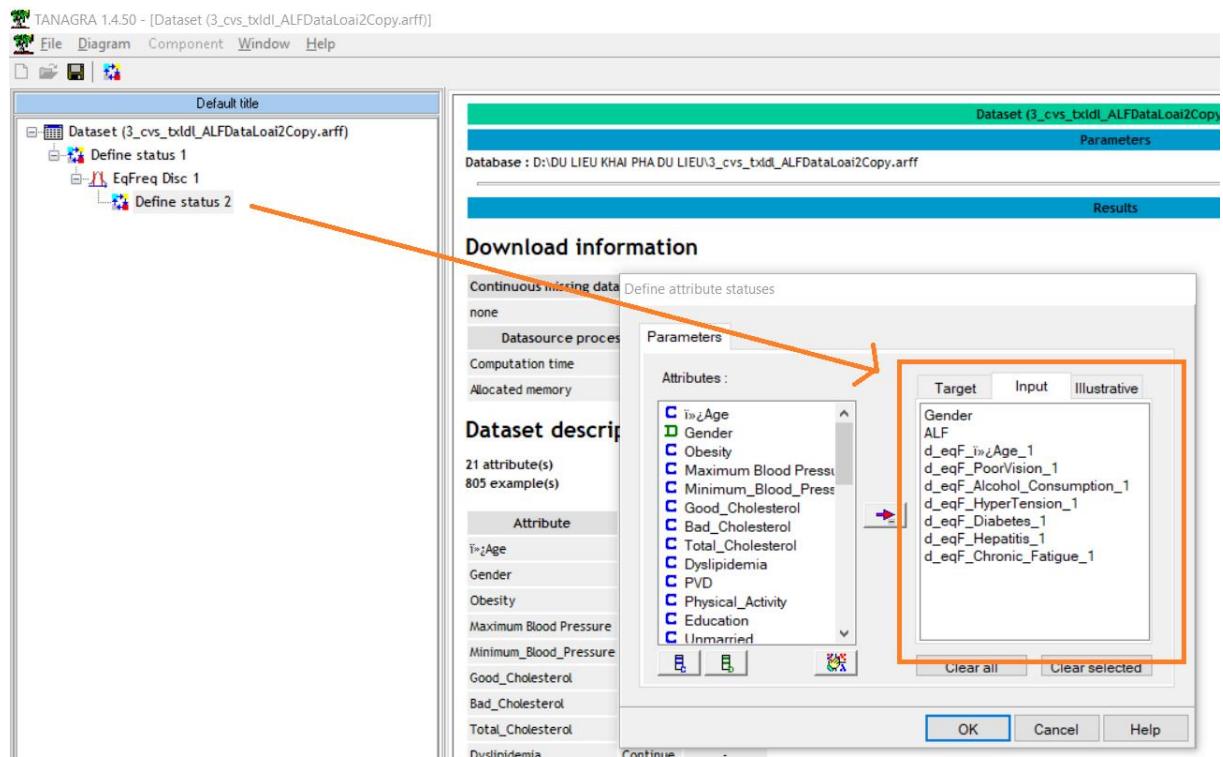


Hình 3. 29 Thiết lập số khoản phân rã

Bước 2: Xác định các thuộc tính cần phân tích

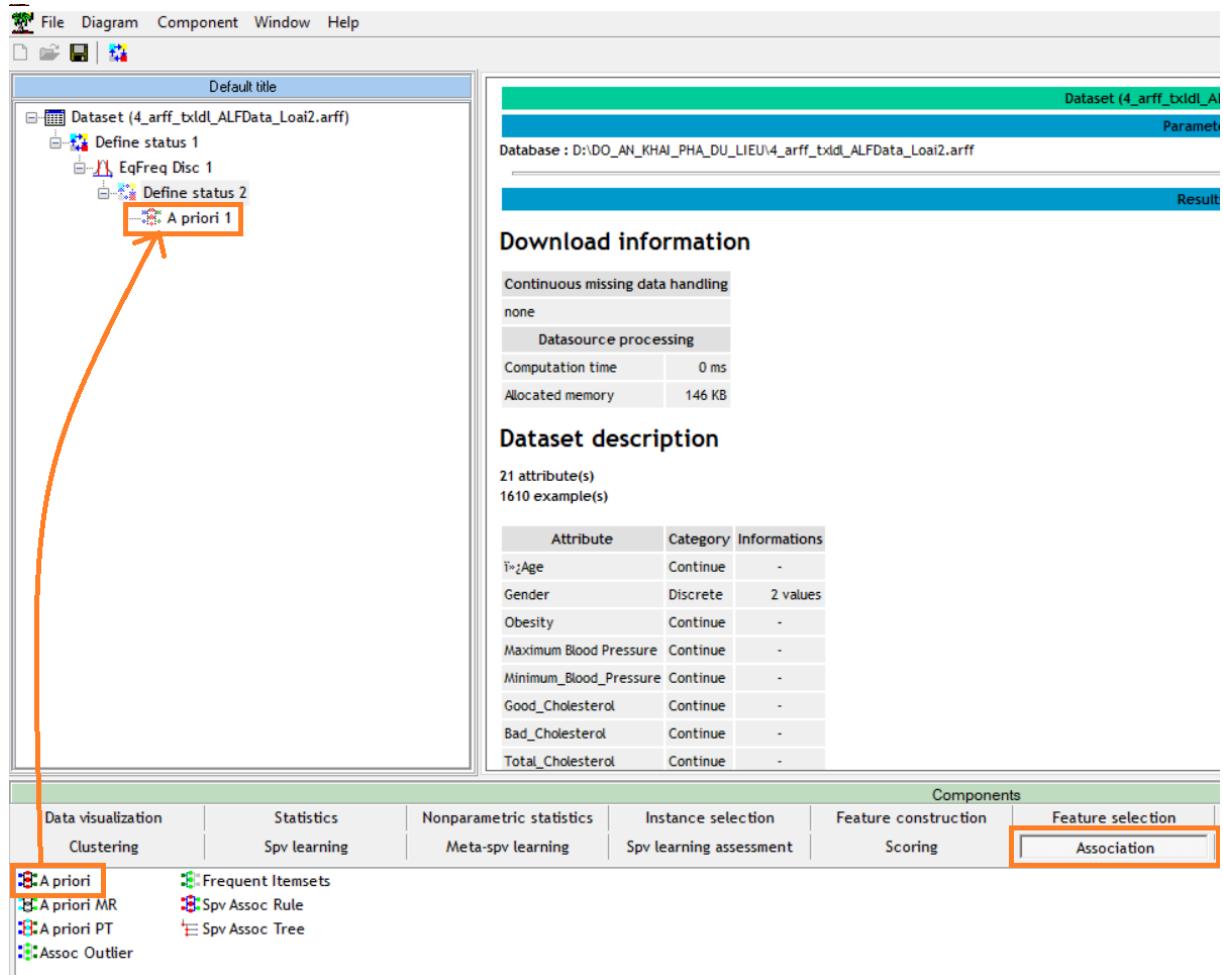
Tạo thư mục Define, đổ dữ liệu cần phân tích vào thư mục

Ứng dụng Tanagra trong khai phá dữ liệu



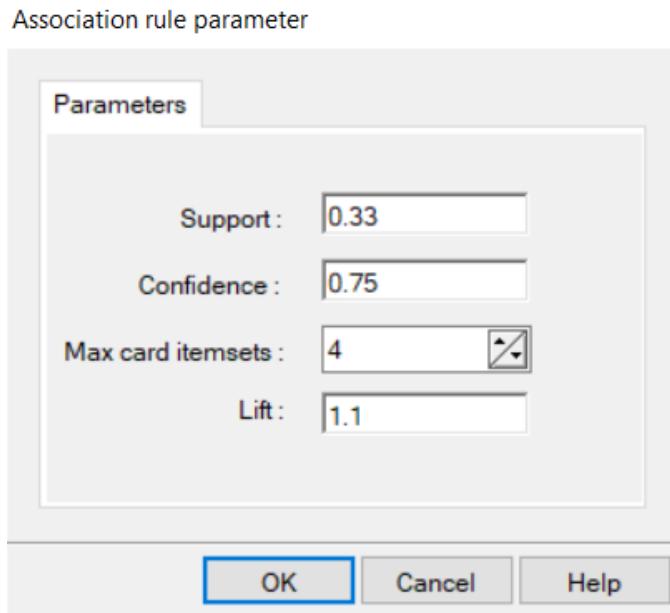
Hình 3. 30 Tạo thư mục Define để đỗ dữ liệu cho component EqFreq Disc

Bước 4: Thêm thuật toán A priori vào component EqFreq Disc



Hình 3.31 Thêm thuật toán A priori vào component EqFreq Disc

Bước 5: Nhập thông số cho thuật toán



Hình 3. 32 Nhập thông số cho thuật toán

Bước 6: Kết quả

Results					
ITEMS					
Transactions	1610				
Counting items					
All items	14				
Filtered items	12				
Counting itemsets					
card(itemset) = 2	48				
card(itemset) = 3	95				
card(itemset) = 4	105				
Rules					
Number of rules	51				

RULES					
N°	Antecedent	Consequent	Lift	Support (%)	Confidence (%)
1	"d_eqF_HyperTension_1=m_<_1.0000000"	"d_eqF_Hepatitis_1=_const_" - "d_eqF_Alcohol_Consumption_1=_const_" - "ALF=no"	1.23393	47.205	89.517
2	"d_eqF_Alcohol_Consumption_1=_const_" - "d_eqF_HyperTension_1=m_<_1.0000000"	"d_eqF_Hepatitis_1=_const_" - "ALF=no"	1.23393	47.205	89.517
3	"d_eqF_Hepatitis_1=_const_" - "d_eqF_HyperTension_1=m_<_1.0000000"	"d_eqF_Alcohol_Consumption_1=_const_" - "ALF=no"	1.23393	47.205	89.517
4	"d_eqF_Chronic_Fatigue_1=_const_" - "d_eqF_HyperTension_1=m_<_1.0000000"	"d_eqF_Diabetes_1=_const_" - "ALF=no"	1.23393	47.205	89.517
5	"d_eqF_Diabetes_1=_const_" - "d_eqF_HyperTension_1=m_<_1.0000000"	"d_eqF_Chronic_Fatigue_1=_const_" - "ALF=no"	1.23393	47.205	89.517
6	"d_eqF_Diabetes_1=_const_" - "d_eqF_Chronic_Fatigue_1=_const_" - "d_eqF_HyperTension_1=m_<_1.0000000"	"ALF=no"	1.23393	47.205	89.517
7	"d_eqF_Hepatitis_1=_const_" - "d_eqF_HyperTension_1=m_<_1.0000000"	"d_eqF_PoorVision_1=_const_" - "ALF=no"	1.23393	47.205	89.517
8	"d_eqF_Hepatitis_1=_const_" - "d_eqF_PoorVision_1=_const_" - "d_eqF_HyperTension_1=m_<_1.0000000"	"ALF=no"	1.23393	47.205	89.517
9	"d_eqF_HyperTension_1=m_<_1.0000000"	"d_eqF_Hepatitis_1=_const_" - "d_eqF_Chronic_Fatigue_1=_const_" - "ALF=no"	1.23393	47.205	89.517
10	"d_eqF_Hepatitis_1=_const_" - "d_eqF_Alcohol_Consumption_1=_const_" - "d_eqF_HyperTension_1=m_<_1.0000000"	"ALF=no"	1.23393	47.205	89.517

Hình 3. 33 Kết quả thuật toán (1)

Như hình 3. 33, ta thấy phần lớn kết quả trả về của ALF là No. Nên ta sẽ xóa bỏ những dòng mà ALF = No trong file .arff để xác định được chính xác những luật kết hợp cho ra kết quả ALF = Yes

The screenshot shows the Tanagra software interface with the following details:

- Default title:** Dataset (3_cvs_tidy_ALFDataLoai2Copy.arff)
- Dataset (3_cvs_tidy_ALFDataLoai2Copy.arff):**
 - Database: D:\DU LIEU KHAI PHA DU LIEU\3_cvs_tidy_ALFDataLoai2Copy.arff
 - Parameters
 - Results
 - Download information
 - Continuous missing data handling: none
 - Datasource processing
 - Computation time: 0 ms
 - Allocated memory: 92 kB
 - Dataset description:** 21 attribute(s) 805 example(s)
 - Attribute Category Informations
 - T_rAge Continue -
 - Gender Discrete 2 values
 - Obesity Continue -
 - Maximum Blood Pressure Continue -

Hình 3. 34 Số dòng đã giảm bớt chỉ còn 805 dòng

ITEMS

Transactions	805
Counting items	
All items	14
Filtered items	13
Counting itemsets	
card(itemset) = 2	53
card(itemset) = 3	105
card(itemset) = 4	115
Rules	
Number of rules	255

RULES

Number of rules : 255					
N°	Antecedent	Consequent	Lift	Support (%)	Confidence (%)
1	"d_eqF_i:>Age_1=m_>=_73.0000000"	"d_eqF_Chronic_Fatigue_1=_const_ - "ALF=yes"	1.82127	33.913	100.000
2	"d_eqF_i:>Age_1=m_>=_73.0000000"	"d_eqF_Diabetes_1=_const_ - "d_eqF_Alcohol_Consumption_1=_const_ - "ALF=yes"	1.82127	33.913	100.000
3	"d_eqF_Alcohol_Consumption_1=_const_ - "d_eqF_i:>Age_1=m_>=_73.0000000"	"d_eqF_Diabetes_1=_const_ - "ALF=yes"	1.82127	33.913	100.000
4	"d_eqF_Chronic_Fatigue_1=_const_ - "d_eqF_i:>Age_1=m_>=_73.0000000"	"ALF=yes"	1.82127	33.913	100.000
5	"d_eqF_PoorVision_1=_const_ - "d_eqF_i:>Age_1=m_>=_73.0000000"	"d_eqF_Hepatitis_1=_const_ - "ALF=yes"	1.82127	33.913	100.000
6	"d_eqF_i:>Age_1=m_>=_73.0000000"	"d_eqF_Hepatitis_1=_const_ - "d_eqF_PoorVision_1=_const_ - "ALF=yes"	1.82127	33.913	100.000
7	"d_eqF_i:>Age_1=m_>=_73.0000000"	"d_eqF_Alcohol_Consumption_1=_const_ - "d_eqF_PoorVision_1=_const_ - "ALF=yes"	1.82127	33.913	100.000
8	"d_eqF_Hepatitis_1=_const_ - "d_eqF_i:>Age_1=m_>=_73.0000000"	"d_eqF_PoorVision_1=_const_ - "ALF=yes"	1.82127	33.913	100.000

Hình 3. 35 Kết quả của thuật toán (2)

3.7.2 Ý nghĩa kết quả của thuật toán Apriori

- Với những kết quả ALF = No

Đánh giá kết quả của thuật toán Apriori cho ta thấy được 51 quy luật như sau:

Sau đây là một số đánh giá:

Đánh giá luật 42: Alcohol_Consumption = const (hằng số - có thể có hoặc không) & PoorVision = const & HyperTension < 1

42	"d_eqF_Alcohol_Consumption_1=_const_ - "d_eqF_PoorVision_1=_const_ - "d_eqF_HyperTension_1=m_<_1.0000000"	"ALF=no"	1.23393	47.205	89.517
----	---	----------	---------	--------	--------

→ ALF = no (với độ tin cậy khoản 89.517%, độ hỗ trợ 47.205%)

⇒ Đối tượng tiêu thụ rượu bia nhưng ko thường xuyên, bị rủi ro về tầm nhìn nhưng không bị tăng huyết áp thì sẽ không bị suy gan

Đánh giá luật 50: Chronic_Fatigue = const & HyperTension < 1 & HyperTension < 1

50 "d_eqF_Chronic_Fatigue_1=_const_" - "d_eqF_HyperTension_1=m_<_1.00000000" "d_eqF_Alcohol_Consumption_1=_const_" - "ALF=no" 1.23393 47.205 89.517

→ ALF = no và Alcohol_consumption = const (với độ tin cậy khoản 89.517%, độ hỗ trợ 47.205%)

⇒ Đối tượng có khả năng đang bị mệt mỏi kinh niên, có khả năng bị tăng huyết áp thì có khả năng tiêu thụ rượu nhưng sẽ không bị suy gan

- **Với những kết quả ALF = Yes**

Đánh giá kết quả của thuật toán Apriori cho ta thấy được 255 quy luật như sau:

Sau đây là một số đánh giá:

Đánh giá luật 171: Alcohol_Consumption = const & Chronic_Fatigue = const & HyperTension >= 1

171 "d_eqF_Alcohol_Consumption_1=_const_" - "d_eqF_Chronic_Fatigue_1=_const_" - "d_eqF_HyperTension_1=m_>_1.00000000" "ALF=yes" 1.61130 43.851 88.471

→ ALF = Yes (với độ tin cậy khoản 88.471%, độ hỗ trợ 43.851%)

⇒ Đối tượng có khả năng đang tiêu thụ rượu, có khả năng đang bị mệt mỏi kinh niên, bị tăng huyết áp thì sẽ bị suy gan

Đánh giá luật 171: Hepatitis = const & Poor_Vision= const & HyperTension >= 1

233 "d_eqF_Hepatitis_1=_const_" - "d_eqF_PoorVision_1=_const_" - "d_eqF_HyperTension_1=m_>_1.00000000" "ALF=yes" 1.61130 43.851 88.471

→ ALF = Yes (với độ tin cậy khoản 88.471%, độ hỗ trợ 43.851%)

⇒ Đối tượng có khả năng đang bị viêm gan, có khả năng tầm nhìn bị kém, đang bị tăng huyết áp thì sẽ bị suy gan

CHƯƠNG 4: KẾT LUẬN

4.1. Những kết quả đạt được

Vận dụng các cơ sở lý thuyết môn khai phá dữ liệu, kết quả đạt được:

- Tổng quan về khai phá dữ liệu
- Tìm hiểu về phần mềm Tanagra
- Tìm hiểu được các thuật toán Decision tree, Kmean, Apriori, ... trên Tanagra
- Tìm hiểu về tập dữ liệu khảo sát suy gan cấp tính ở độ tuổi trưởng thành
- Thực hiện được các thuật toán trên dữ liệu thực tế
- Biết cách làm việc nhóm với nhau.

4.2. Hạn chế

Một số hạn chế còn tồn động:

- Chưa khai phá hết tiềm năng ứng dụng
- Chưa áp dụng vào thực tiễn

4.1. Hướng phát triển

- Tiếp tục khai thác thêm dữ liệu khác nhau trên Tanagra
- Tiếp tục tìm hiểu thêm các thuật toán khác trên Tanagra
- Tiếp tục tìm hiểu kỹ hơn về các dữ liệu
- Có thể đồng nhất sử dụng 1 dữ liệu khi thực hiện

TÀI LIỆU THAM KHẢO

- [1] <http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html>
- [3] [TANAGRA - A free DATA MINING software for teaching and research \(univ-lyon2.fr\)](#)
- [2] Slide bài giảng chương 1,2,3,4 Khai phá dữ liệu của cô Nguyễn Thị Trần Lộc