LỜI CẢM ƠN

Em xin tỏ lòng biết ơn sâu sắc tới thầy giáo Nguyễn Trịnh Đông - người hướng dẫn trực tiếp, chỉ bảo tận tình, góp ý sâu sắc trong suốt quá trình học tập, nghiên cứu để em hoàn thành khóa luận này.

Em xin bày tỏ lòng biết ơn đến các thầy cô giáo trong bộ môn Công nghệ thông tin trường Đại học Dân lập Hải Phòng đã trực tiếp giảng dạy, góp ý, động viên em trong suốt bốn năm học qua.

Em xin gửi lời cảm ơn đến các thành viên lớp CT1002, những người bạn đã luôn ở bên cạnh động viên, tạo điều kiện thuận lợi và cùng em tìm hiểu, hoàn thành tốt khóa luận.

Cuối cùng em xin bày tỏ lòng biết ơn đến gia đình, và các bạn bè đã chia sẻ và động viên em hoàn thành khóa luận này.

Hải Phòng, *Ngày 09 tháng 07 năm 2010.*

Sinh viên

Phạm Ngọc Hùng

MỤC LỤC

LỜI MỞ ĐẦU

Sự phát triển của công nghệ thông tin và việc ứng dụng công nghệ thông tin trong nhiều lĩnh vực của đời sống, kinh tế xã hội trong nhiều năm qua cũng đồng nghĩa với lượng dữ liệu đã được các cơ quan thu thập và lưu trữ ngày một tích luỹ nhiều lên. Họ lưu trữ các dữ liệu này vì cho rằng trong nó ẩn chứa những giá trị nhất định nào đó. Tuy nhiên, theo thống kê thì chỉ có một lượng nhỏ của những dữ liệu này (khoảng từ 5% đến 10%) là luôn được phân tích, số còn lại họ không biết sẽ phải làm gì hoặc có thể làm gì với chúng nhưng họ vẫn tiếp tục thu thập rất tốn kém với ý nghĩ lo sợ rằng sẽ có cái gì đó quan trọng đã bị bỏ qua sau này có lúc cần đến nó. Mặt khác, trong môi trường cạnh tranh, người ta ngày càng cần có nhiều thông tin với tốc độ nhanh để trợ giúp việc ra quyết định và ngày càng có nhiều câu hỏi mang tính chất định tính cần phải trả lời dựa trên một khối lượng dữ liệu khổng lồ đã có. Với những lý do như vậy, các phương pháp quản trị và khai thác cơ sở dữ liệu truyền thống ngày càng không đáp ứng được thực tế đã làm phát triển một khuynh hướng kỹ thuật mới đó là Kỹ thuật phát hiện tri thức và khai phá dữ liệu (KDD - Knowledge Discovery and Data Mining).

Kỹ thuật phát hiện tri thức và khai phá dữ liệu đã và đang được nghiên cứu, ứng dụng trong nhiều lĩnh vực khác nhau ở các nước trên thế giới, tại Việt Nam kỹ thuật này tương đối còn mới mẻ tuy nhiên cũng đang được nghiên cứu và dần đưa vào ứng dụng. . Trên cơ sở đó áp dụng vào bài khai phá dữ liệu thống kê dân số.

DANH SÁCH HÌNH VẼ

CHƯƠNG 1 : TỔNG QUAN VỀ DATA MINING

1.1 Tổng quan về Datamining

1.1.1 Giới thiệu chung về Datamining.

Data Mining là một lĩnh vực mới xuất hiện, nhằm tự động khai thác những thông tin, những tri thức có tính tiềm ẩn, hữu ích từ những CSDL lớn của các đơn vị, tổ chức, doanh nghiệp,…. từ đó làm thúc đẩy khả năng sản xuất, kinh doanh, cạnh tranh cho các đơn vị, tổ chức này. Các kết quả khoa học cùng những ứng dụng thành công trong khám phá tri thức, cho thấy, Data Mining là một lĩnh vực phát triển bền vững, mang lại nhiều lợi ích và có nhiều triển vọng, đồng thời có ưu thế hơn hẳn so với các công cụ phân tích dữ liệu truyền thống. Hiện nay, Data Mining đã ứng dụng ngày càng rộng rãi trong các lĩnh vực như : Thương mại, tài chính, điều trị y học, viễn thông, tin – sinh,….

Data Mining là một hướng nghiên cứu mới ra đời hơn một thập niên trở lại đây, các kỹ thuật chính được áp dụng trong lĩnh vực này phần lớn được thừa kế từ lĩnh vực CSDL, học máy, trí tuệ nhân tạo, lý thuyết thông tin, xác suất thống kê, và tính toán hiệu năng cao. Do sự phát triển nhanh của Data Mining về phạm vi áp dụng và các phương pháp tìm kiếm tri thức, nên đã có nhiều quan điểm khác nhau về Data Mining. Tuy nhiên, ở một mức trừu tượng nhất định, chúng ta định nghĩa Data Mining như sau :

Định nghĩa : *DATA MINING là một quá trình tìm kiếm, phát hiện các tri thức mới, tiềm ẩn, hữu dụng trong CSDL lớn.*

Khám phá tri thức trong CSDL (Knowledge Discovery in Databases - KDD) là mục tiêu chính của Data Mining, do vậy hai khái niệm Data Mining và KDD được các nhà khoa học trên hai lĩnh vực được xem là tương đương với nhau. Thế nhưng, nếu phân chia một cách chi tiết thì Data Mining là một bước chính trong quá trình KDD.

1.1.2 Quá trình khám phá tri thức trong CSDL.

Quá trình khám phá tri thức trong CSDL gồm các giai đoạn sau:

- *Trích chọn dữ liệu* : là bước trích chọn những tập dữ liệu cần được khai phá từ các tập dữ liệu lớn (databases, data warehouses, data repositories) ban đầu theo một số tiêu chí nhất định.

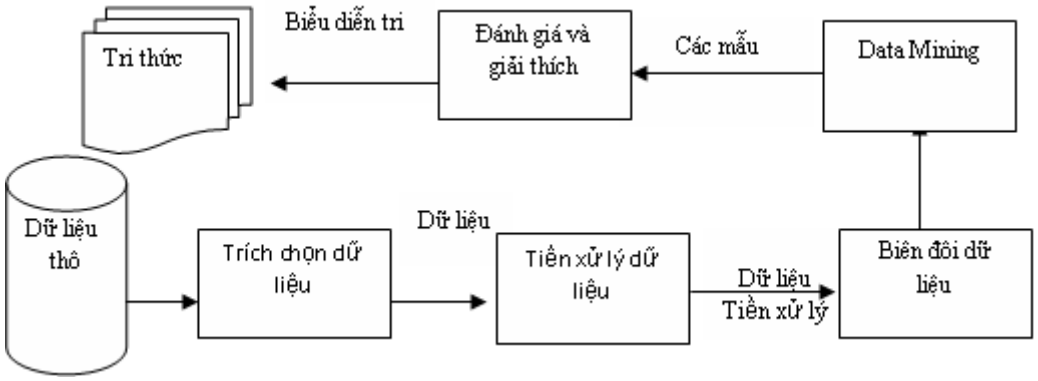
- *Tiền xử lý dữ liệu* : là bước làm sạch dữ liệu (xử lý với dữ liệu không đầy đủ, dữ liệu nhiễu, dữ liệu không nhất quán, .v.v.), rút gọn dữ liệu (sử dụng hàm nhóm

và tính tổng, các phương pháp nén dữ liệu, sử dụng histograms, lấy mẫu, .v.v.), rời rạc hóa dữ liệu (rời rạc hóa dựa vào histograms, dựa vào entropy, dựa vào phân khoảng, .v.v.). Sau bước này, dữ liệu sẽ nhất quán, đầy đủ, được rút gọn, và được rời rạc hóa.

- *Biến đổi dữ liệu* : đây là bước chuẩn hóa và làm mịn dữ liệu để đưa dữ liệu về dạng thuận lợi nhất nhằm phục vụ cho các kỹ thuật khai phá ở bước sau.

- *Data mining*: đây là bước áp dụng những kỹ thuật phân tích (phần nhiều là các kỹ thuật của học máy) nhằm để khai thác dữ liệu, trích chọn được những mẫu thông tin, những mối liên hệ đặc biệt trong dữ liệu. Đây được xem là bước quan trọng và tốn nhiều thời gian nhất của toàn quá trình KDD.

- *Đánh giá và biểu diễn tri thức* : những mẫu thông tin và mối liên hệ trong dữ liệu đã được khám phá ở bước trên được chuyển dạng và biểu diễn ở một dạng gần gũi với người sử dụng như đồ thị, cây, bảng biểu, luật, .v.v. Đồng thời bước này cũng đánh giá những tri thức khám phá được theo những tiêu chí nhất định.



Hình 1: Các bước thực hiện trong quá trnh khám phá tri thức

1.1.3 Các kỹ thuật áp dụng trong Datamining

Nếu đứng trên quan điểm của học máy (Machine Learning), thì các kỹ thuật trong Data Mining, bao gồm :

❖ *Học có giám sát (Supervised learning)*:

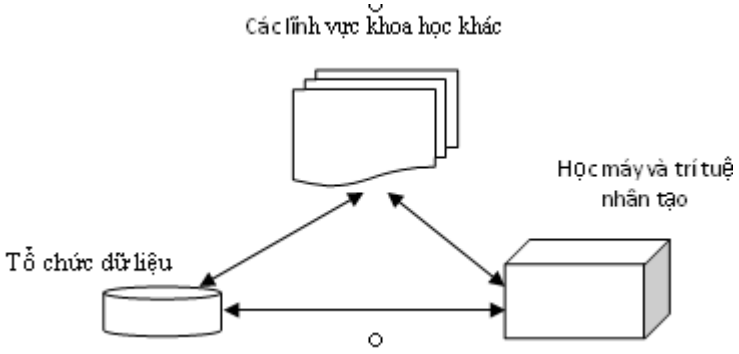
❖ *Học không có giám sát (Unsupervised learning):*

❖ *Học nửa giám sát (Semi - Supervised learning):*

Nếu căn cứ vào lớp các bài toán cần giải quyết, thì Data Mining bao gồm các kỹ thuật áp dụng sau :

- *Phân lớp và dự đoán (classification and prediction)*:

- *Luật kết hợp (association rules)*:
- *Phân tích chuỗi theo thời gian (sequential/ temporal patterns*
- *Phân cụm (clustering/ segmentation)*:
- *Mô tả khái niệm (concept description and summarization)*:



Hình 2: Các lĩnh vực liên quan đến Khám phá tri thức trong CSDL

1.1.4 Ứng dụng của Datamining

Khai phá dữ liệu có nhiều ứng dụng trong thực tế. Một trong số ứng dụng điển hình như:

Bảo hiểm

Tài chính và thị trường chứng khoán : phân tích tình hình tài chính và dự báo giá của các loại cổ phiếu trong thị trường chứng khoán. Danh mục vốn và giá, lãi suất, dữ liệu thẻ tín dụng, phát hiện gian lận…

Phân tích dũ liệu và hỗ trợ ra quyết định.

Điều trị và chăm sóc y tế : Một số thông tin về chuẩn đoán lưu bệnh trong các hệ thống quản lý bệnh viện. Phân tích mối liên hệ giữa triệu chứng bệnh, chuẩn đoán và phương pháp điều trị (chế độ dinh dưỡng, thuốc ..).

Sản xuất chế biến: Quy trình, phương pháp chế biến và xử lý sự cố

Text mining & Web mining: phân lớp văn bản và các trang web, tóm tắt văn bản …

Lĩnh vực khoa học: Quan sát thiên văn, dữ liệu gene, dữ liệu sinh vật học, tìm kiếm, so sánh các hệ gene và thông tin di truyền, mối liên hệ gene và một số bệnh di truyền.

Mạng viễn thông: Phân tích các cuộc gọi điện thoại và hệ thống giám sát lỗi, sự cố chất lượng dịch vụ…

Lĩnh vực xã hội: bài toán thống kê dân số, bài toán dự báo về dân số…để từ đó đưa ra cách khắc phục thích hợp nhất.

1.2 Phân cụm dữ liệu và các thuật toán về phân cụm dữ liệu.

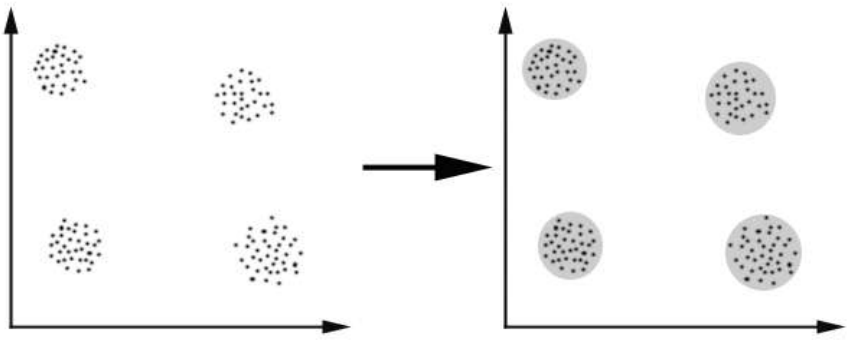
1.2.1.Giới thiệu chung về phân cụm dữ liệu.

Phân cụm dữ liệu là một lĩnh vực liên ngành và đang còn được phát triển mạnh mẽ như thống kê, học máy, nhận dạng, Data mining, … Ở một mức cơ bản nhất, người ta đã đưa ra định nghĩa PCDL như sau :

"*PCDL là một kỹ thuật trong DATA MINING, nhằm tìm kiếm, phát hiện các cụm, các mẫu dữ liệu tự nhiên tiềm ẩn, quan tâm trong tập dữ liệu lớn, từ đó cung cấp thông tin, tri thức hữu ích cho ra quyết định*"

Như vậy, PCDL là quá trình phân chia một tập dữ liệu ban đầu thành các cụm dữ liệu sao cho các phần tử trong một cụm "tương tự" (Similar) với nhau và các phần tử trong các cụm khác nhau sẽ "phi tương tự" (Dissimilar) với nhau. Số các cụm dữ liệu được phân ở đây có thể được xác định trước theo kinh nghiệm hoặc có thể được tự động xác định của phương pháp phân cụm.

Chúng ta có thể minh hoạ vấn đề phân cụm như hình 3 sau đây :



Hình 3: Mô phỏng vấn đề PCDL

Trong hình trên, sau khi phân cụm chúng ta thu được bốn cụm trong đó các phần tử "*gần nhau*" hay là "*tương tự*" thì được xếp vào một cụm, trong khi đó các phần tử "*xa nhau*" hay là "*phi tương tự*" thì chúng thuộc về các cụm khác nhau.

1.2.2 Một số thuật toán phân cụm dữ liệu.

a. Họ các thuật toán phân hoạch.

- Thuật toán k-means.
- Thuật toán PAM (Partioning Around Medoids).
- Thuật toán CLARA (Clustering LARge Applications).
- Thuật toán CLARANS (Clustering LARge ApplicatioNS).

b. Các thuật toán phân cụm phân cấp.

- Thuật toán BIRCH
- Thuật toán CURE

c. Các thuật toán phân cụm dựa trên mật độ.

- Thuật toán DBSCAN
- Thuật toán OPTICS
- Thuật toán DENCLUE

d. Một số thuật toán phân cụm dữ liệu đặc thù.

- Thuật toán STING
- Thuật toán CLIQUE
- Thuật toán EM

e. Phân cụm dữ liệu mờ.

- Thuật toán FCM
- Thuật toán εFCM

f. Phân cụm song song trên tập dữ liệu hỗn hợp.

- Thuật toán k- prototypes
- Thuật toán song song k - prototypes

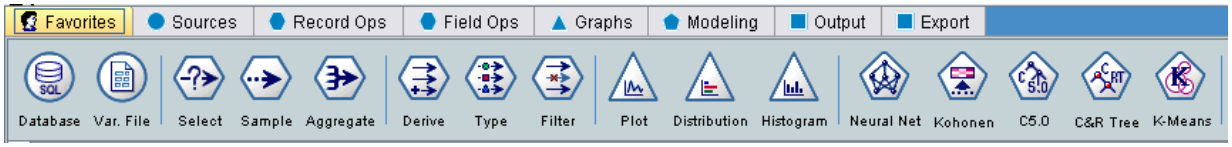
CHƯƠNG 2: PHẦN MỀM CLEMENTINE

2.1 Giới thiệu chung về Clementine.

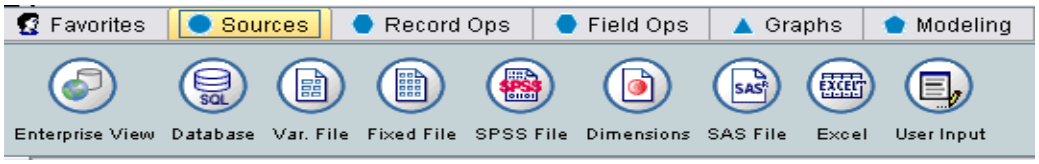
Clementine là một sản phẩm của SPSS inc, SPSS Clementine là một trong sản phẩm mô hình hóa quá trình khai phá dữ liệu, cho phép người dùng nhanh chóng phát triển các mô hình đã được dự đoán trước bằng cách sử dụng kinh nghiệm thực tế và triển khai chúng vào các lĩnh vực cụ thể được tốt hơn.

Phần mềm Clementine gồm có 8 tab chính:

1. Tab ***Favorites***: Chỉnh sửa các nút được lựa chọn mặc định.

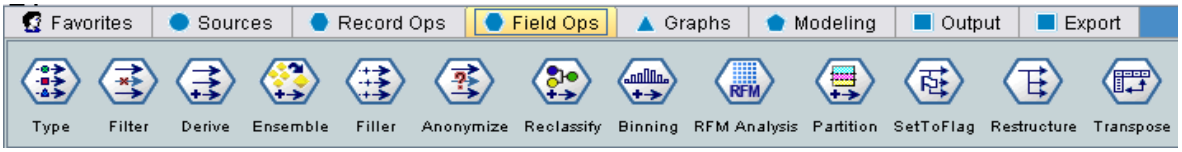


2. Tab ***Source*** : Nguồn dữ liệu, nhập dữ liệu vào Clementine.

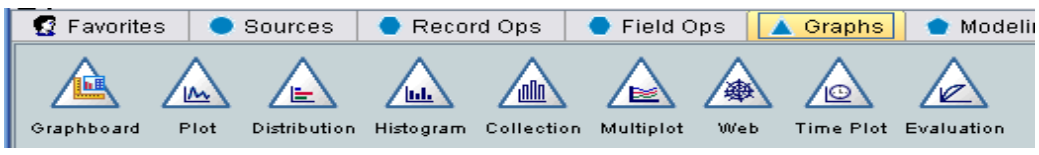


3. Tab ***Record Ops*** : Thực hiện các thao tác trên bản ghi dữ liệu như ; lựa chọn, trộn, thêm..... trường dữ liệu.

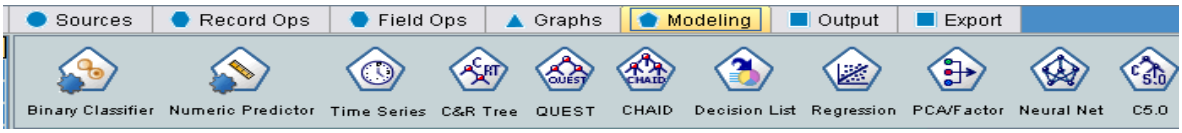
4. Tab ***Field Ops*** : Thực hiện các thao tác trên các trường dữ liệu như lọc, chuyển hóa trường dữ liệu mới, xác định kiểu dữ liệu...



5. Tab ***Graphs***(đồ thị) : bao gồm các nút hiển thị đồ họa trước và sau khi Modeling (mô hình hóa) gồm các nút Plot, nút Web, nút Histogram, biểu đồ đánh giá...



6. Tab ***Modeling*** : Mô hình hóa các thuật toán trong Clementine chẳng hạn như ; nút K-means, C&R Tree, C5.0, Sequence...



7. Tab ***Output*** :Xuất dữ liệu đầu ra dưới dạng như bản báo cáo (report), SPSS,...

8. Tab ***Export*** : Xuất dữ liệu đầu ra dưới dạng chẳng hạn như Exel, SPSS,...

2.2 Quá trình xử lý dữ liệu trong Clementine.

Clementine là phần mềm ứng dụng trong khai phá dữ liệu. Do đó quy trình xử lý dữ liệu được thể hiện như sau:

1. Nguồn dữ liệu:

Nguồn dữ liệu bao gồm tập dữ liệu với rất nhiều các định dạng giúp người sử dụng dễ dàng đưa dữ liệu của mình vào để xử lý như là ; Exel, SPSS, SQL,...

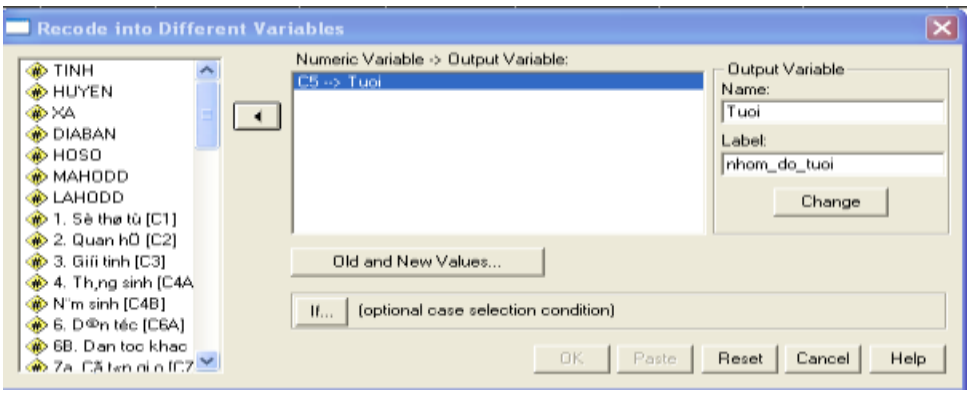
Nguồn dữ liệu hay còn gọi là dữ liệu thô nghĩa là dữ liệu chưa qua quá trình tinh chỉnh, là nguồn dữ liệu gốc, nguồn dữ liệu ban đầu.

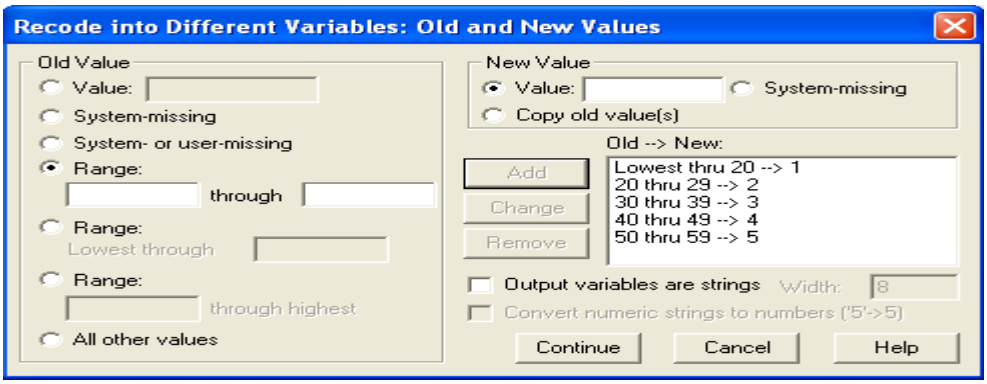
2. Trích chọn dữ liệu:

Chọn nguồn dữ liệu phù hợp nhất với yêu cầu bài toán đặt ra. Dữ liệu được chọn phải chứa những thông tin đầy đủ liên quan đến yêu cầu cần đặt ra, phải thỏa mãn các tiêu chí nhất định nào đó.

3. Tiền xử lý dữ liệu:

Tiền xử lý dữ liệu là quá trình tinh chỉnh dữ liệu, chỉnh sửa dữ liệu, dữ liệu có thể được xử lý trong SPSS trước khi được đưa vào khai thác.



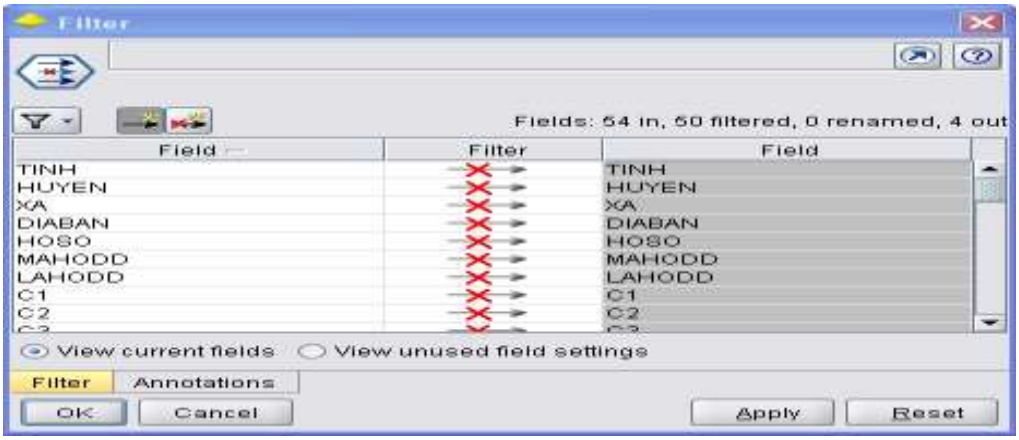


Tiền xử lý dữ liệu là quá trình làm sạch dữ liệu (xử lý với dữ liệu không đầy đủ, dữ liệu nhiễu, dữ liệu không nhất quán, .v.v.), rút gọn dữ liệu (sử dụng hàm nhóm và tính tổng, các phương pháp nén dữ liệu, sử dụng histograms, lấy mẫu, .v.v.), rời rạc hóa dữ liệu (rời rạc hóa dựa vào histograms, dựa vào entropy, dựa vào phân khoảng, .v.v.). Sau bước này, dữ liệu sẽ nhất quán, đầy đủ, được rút gọn, và được rời rạc hóa.

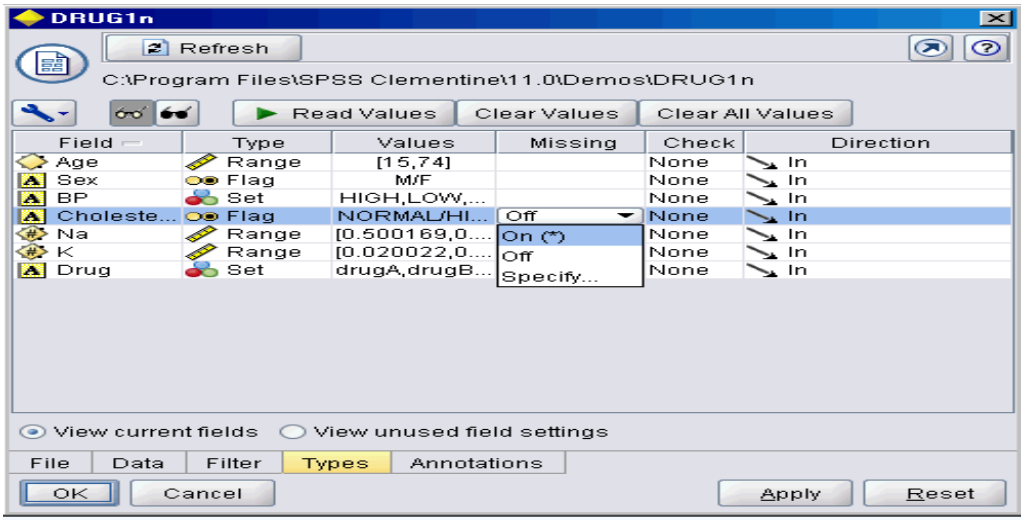
4. Biến đổi dữ liệu:

Biến đổi dữ liệu là quá trình chuẩn hóa và làm mịn dữ liệu để đưa dữ liệu về dạng ngắn gọn và đơn giản giúp giải quyết bài toán một cách nhanh nhất.

Biến đổi dữ liệu ban đầu thành các dữ liệu chuẩn nhất, có thể thêm các trường dữ liệu cần thiết hoặc bỏ đi các trường dữ liệu không cần thiết.



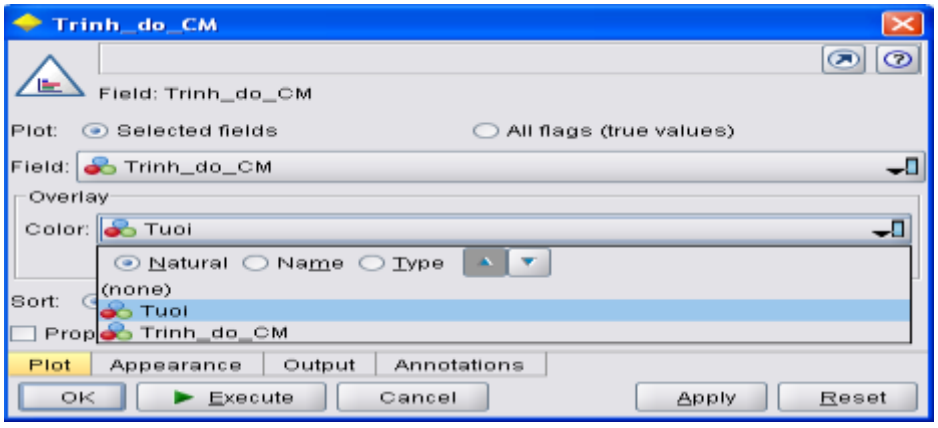
Biến đổi dữ liệu thành các dữ liệu mới với những thuộc tính mới và các trường dữ liệu mới.



Có thể biến đổi thành các loại dữ liệu sau: Range(khoảng cách, hàng), Default (mặc định), Flag (dạng cờ), Set (tập hợp), Ordered Set, Typeless, Discrete …

5. Khai phá dữ liệu.

Đây được xem là bước quan trọng và tốn nhiều thời gian nhất của toàn quá trình.



Áp dụng các kỹ thuật phân tích để khai phá dữ liệu.

Trong quá trình này sử dụng các thuật toán phân hoạch, các thuật toán phân cụm phân cấp để khai phá dữ liệu…như thuật toán K-means, PAM, CLARA, BIRCH,….

6. Đánh giá và biểu diễn tri thức.

Đây là kết quả của toàn bộ quá trình. Kết quả được thể hiện dưới các dạng khác nhau như bảng biểu (Exel, Table, Custom Table..), dạng cây (C&R Tree, C5.0…), hay dưới dạng đồ thị (Graphboard, Plot, Distribution, histogram, collection, multiplot, Web, Timelot, Evaluation …) giúp đưa ra kết quả gần gũi với người sử dụng, có cái nhìn trực quan hơn đối với yêu cầu được đặt ra.

Sau khi kết quả được đưa ra thì đánh giá kết quả đó có đúng yêu cầu của bài toán không, có thỏa mãn tiêu chí hay điều kiện nhất định nào đó hay không.

Ví dụ minh họa:

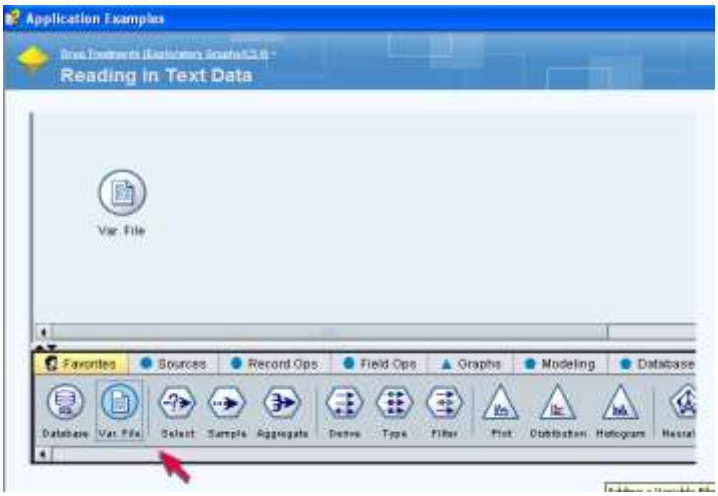
Trong ví dụ này, hãy hình dung rằng bạn là một nhà nghiên cứu y tế. Bạn đã thu thập dữ liệu về một danh sách các bệnh nhân, tất cả đều bị bệnh tương tự nhau. Trong khóa học của họ về điều trị, mỗi bệnh nhân đáp ứng một trong năm loại thuốc. Một phần của công việc của bạn là sử dụng khai phá dữ liệu để tìm ra thuốc có thể là thích hợp nhất cho một bệnh nhân trong tương lai với các bệnh như nhau.

Ví dụ này sử dụng các dòng có tên ***druglearn.str***, có sự tham chiếu các dữ liệu tập tin có tên DRUG1n.

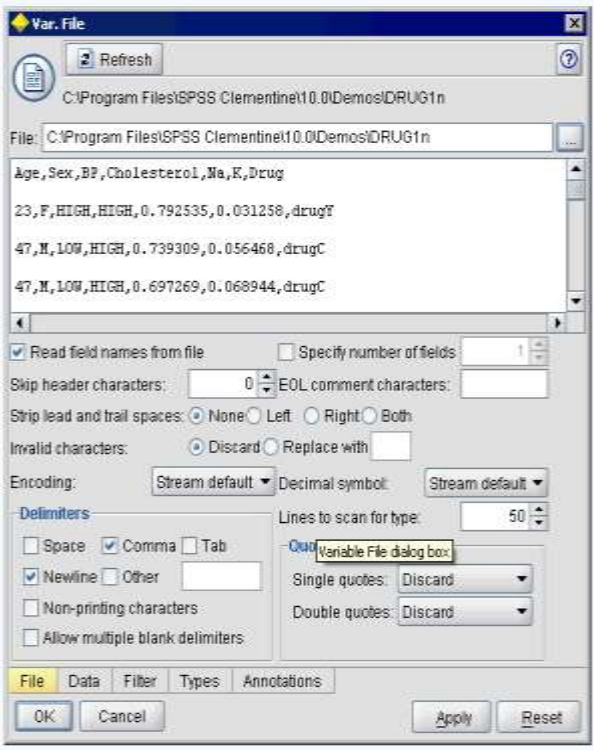
Các trường dữ liệu được sử dụng trong bản demo này là:

Trường dữ liệu	**Mô tả**
Age	Tuổi (dạng số)
Sex	Giới tính : M - Nam, F – Nữ
BP	Huyết áp : HIGH, NORMAL, hoặc LOW
Cholesterol	Nồng độ Cholesterol : NORMAL hoặc HIGH
Na	Nồng độ Natri trong máu
K	Nồng độ Kali trong máu
Drug	Thuốc

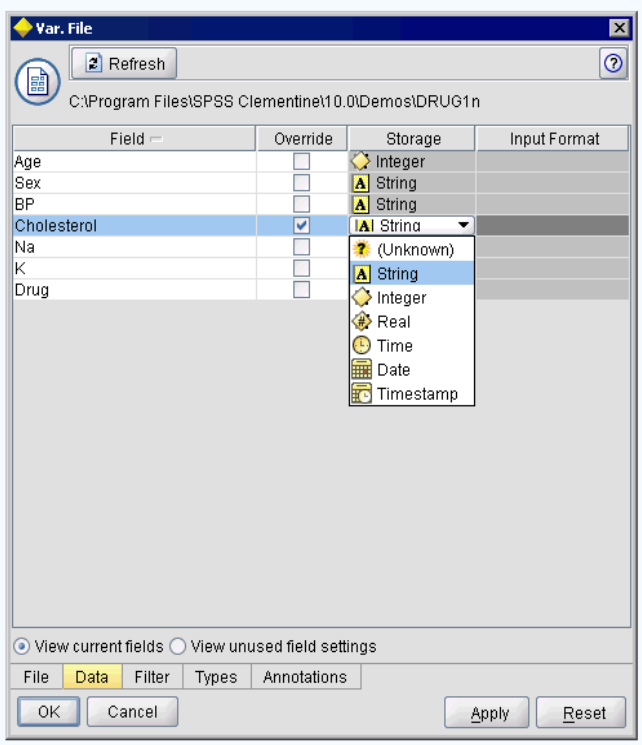
Để đọc dữ liệu ta sử dụng một nút ***Var. File***. Bạn có thể thêm một nút ***Var.File*** từ bảng màu - hoặc nhấn vào tab ***Sources*** để tìm nút theo yêu cầu . Tiếp theo, nhấp đúp vào nút vừa được đặt để mở hộp thoại của nó.



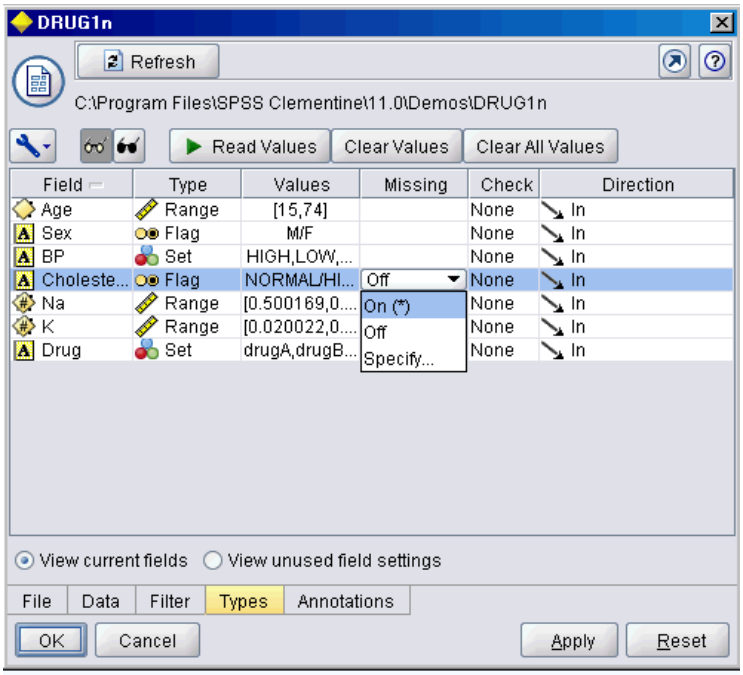
Nhấp vào nút (hình vuông) ngay bên phải của hộp ***Var.File*** để duyệt đến thư mục cần chọn. Mở thư mục Demo và chọn tập tin gọi là DRUG1n.



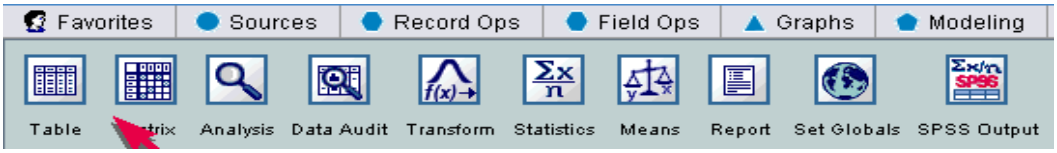
Nhấp vào tab ***Data*** để ghi đè lên và thay đổi giá trị cho một tập tin.



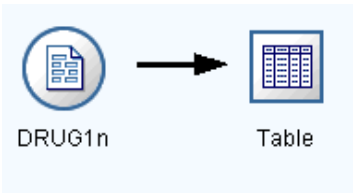
Nút ***Type*** thể hiện về các loại trường trong dữ liệu .Chọn ***Read Values*** để xem các giá trị thực tế cho từng tập tin.



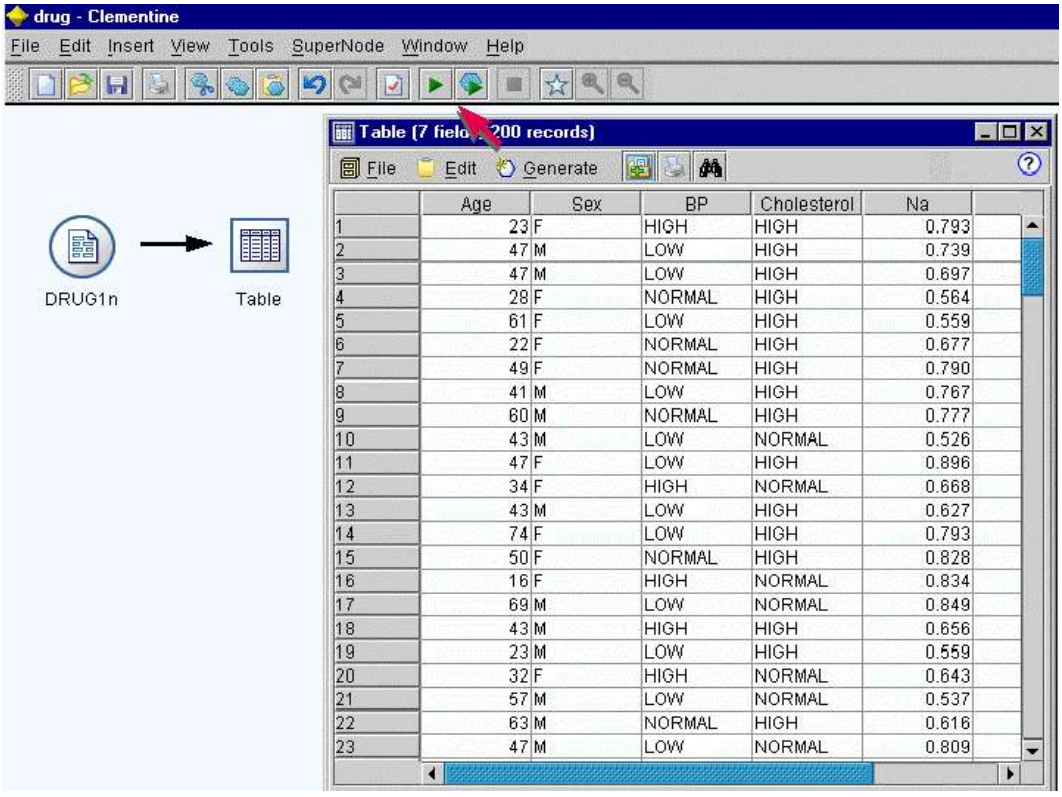
Khi đã tải tập tin dữ liệu.Để tạo bảng lưu trữ các dữ liệu đó thì nhấp đúp vào nút ***Table*** trong bảng màu hoặc kéo và thả nó vào bài.



Nhấn đúp chuột vào nút ***Tabel*** từ bảng màu, nó sẽ tự động kết nối nó với tập dữ liệu gốc.

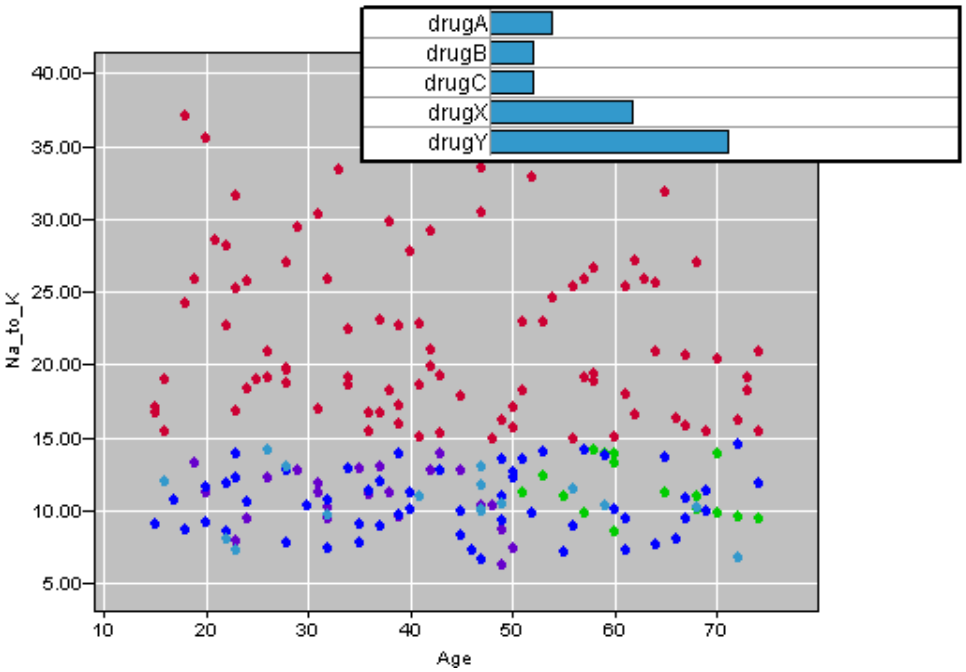


Để xem bảng, nhấn vào nút mũi tên màu xanh trên thanh công cụ để thực thi, hoặc kích chuột phải vào nút ***Table*** và chọn ***Execute.***

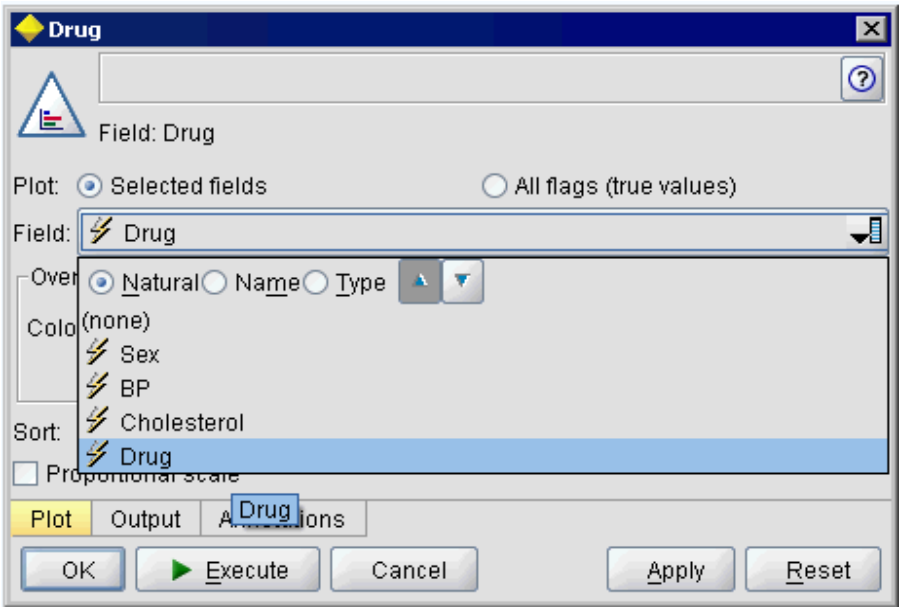


	Age	Sex	BP	Cholesterol	Na
1	23	F	HIGH	HIGH	0.793
2	47	M	LOW	HIGH	0.739
3	47	M	LOW	HIGH	0.697
4	28	F	NORMAL	HIGH	0.564
5	61	F	LOW	HIGH	0.559
6	22	F	NORMAL	HIGH	0.677
7	49	F	NORMAL	HIGH	0.790
8	41	M	LOW	HIGH	0.767
9	60	M	NORMAL	HIGH	0.777
10	43	M	LOW	NORMAL	0.526
11	47	F	LOW	HIGH	0.896
12	34	F	HIGH	NORMAL	0.668
13	43	M	LOW	HIGH	0.627
14	74	F	LOW	HIGH	0.793
15	50	F	NORMAL	HIGH	0.828
16	16	F	HIGH	NORMAL	0.834
17	69	M	LOW	NORMAL	0.849
18	43	M	HIGH	HIGH	0.656
19	23	M	LOW	HIGH	0.559
20	32	F	HIGH	NORMAL	0.643
21	57	M	LOW	NORMAL	0.537
22	63	M	NORMAL	HIGH	0.616
23	47	M	LOW	NORMAL	0.809

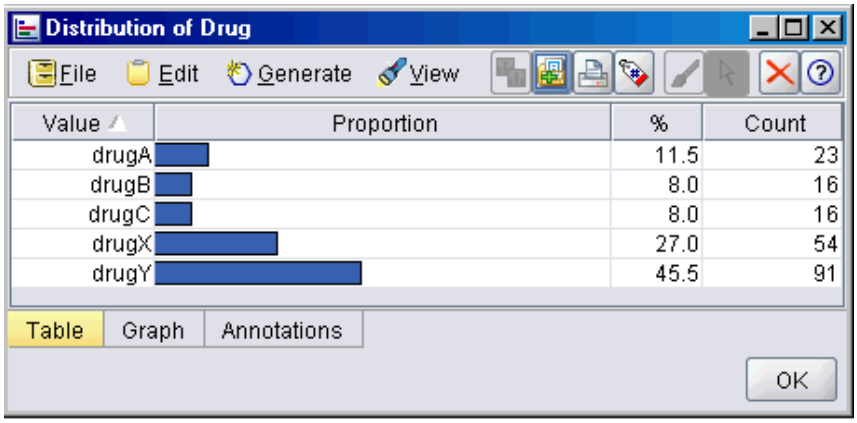
Trong quá trình khai thác dữ liệu, để có cái nhìn trực quan và dễ dàng hơn, Clementine cung cấp một số loại đồ thị khác nhau để lựa chọn, tùy thuộc vào loại dữ liệu mà bạn muốn hiển thị. Ví dụ, để tìm ra những tỷ lệ bệnh nhân phù hợp với từng loại thuốc, ta sử dụng một nút ***Distribution*** (phân phối).



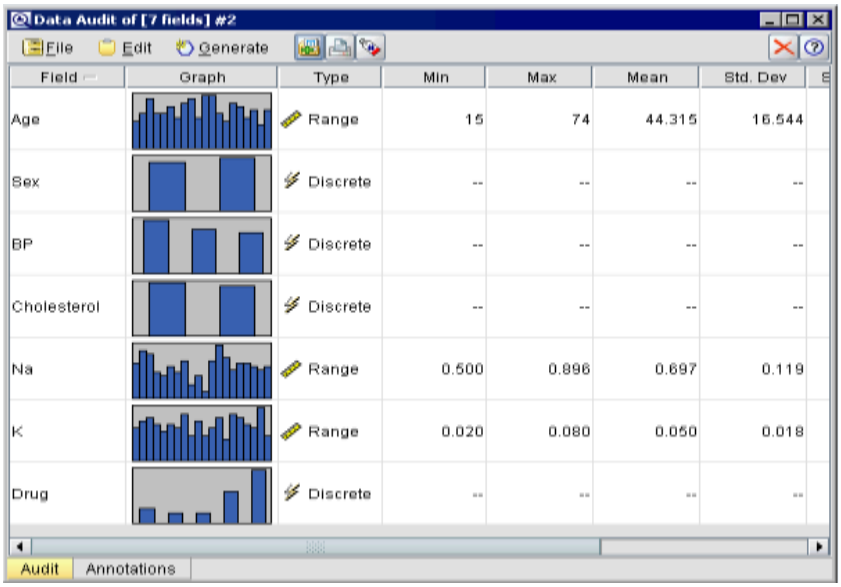
Thêm một nút ***Distribution*** (phân phối) và kết nối nó với nút nguồn, sau đó kích đúp vào nút để chỉnh sửa các tùy chọn cho hiển thị. Chọn tập *Drug* (thuốc) muốn hiển thị. Sau đó, bấm ***Execute*** (thực hiện) từ hộp thoại.



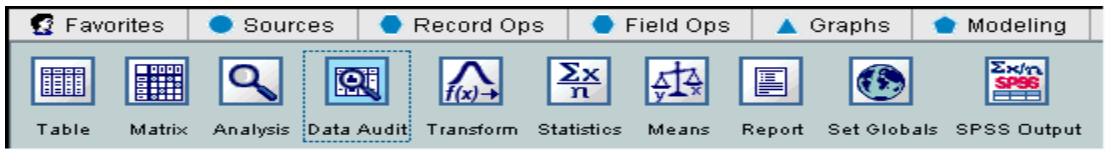
Đồ thị kết quả sẽ giúp bạn nhìn thấy tỷ lệ của dữ liệu. Nó cho thấy rằng bệnh nhân thường dùng thuốc Y và dùng thuốc B và C là ít nhất .



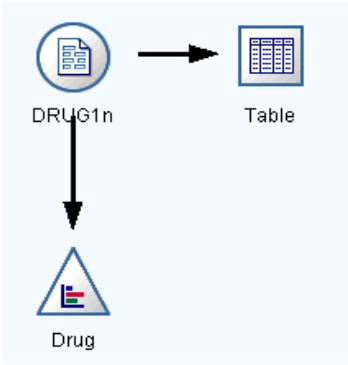
Ngoài ra, bạn có thể đính kèm và thực thi một nút ***Data Audit*** (Kiểm kê dữ liệu) giúp người xem dễ dàng nhìn thấy tỷ lệ trong đồ thị cho tất cả các trường cùng một lúc.



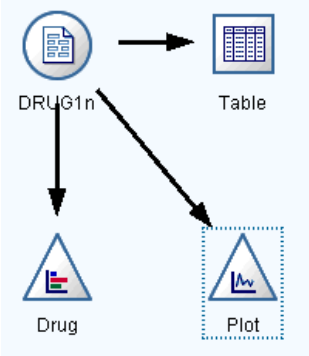
Nút ***Data Audit*** có sẵn trên tab ***Output.***



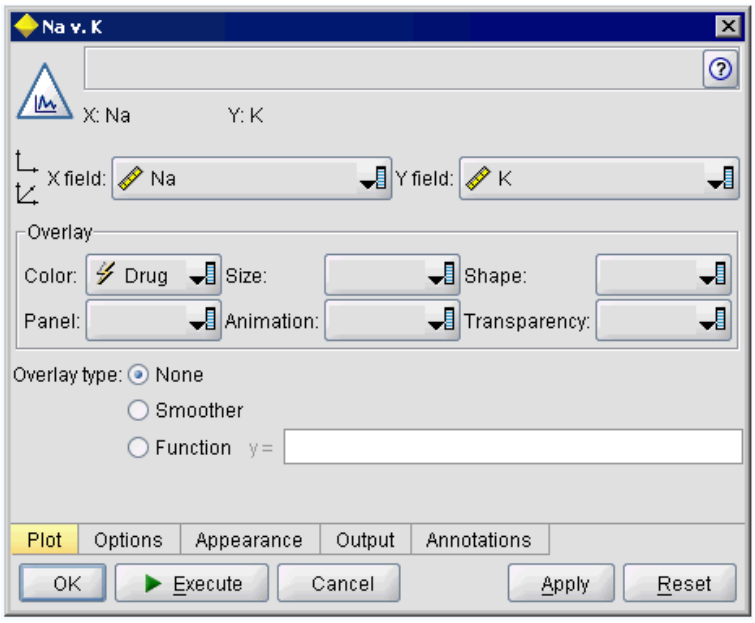
Bây giờ chúng ta hãy nhìn vào những yếu tố liên quan đến thuốc. Như chúng ta biết rằng nồng độ của natri và kali trong máu là những nhân tố quan trọng.



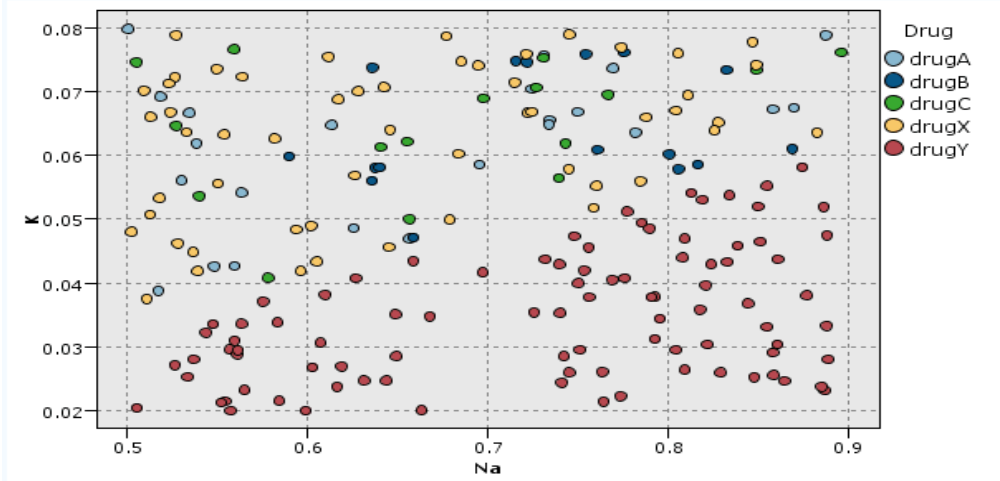
Đặt một nút ***Plot*** trong vùng làm việc và kết nối nó với nút nguồn, và nhấn đúp để chỉnh sửa các nút.



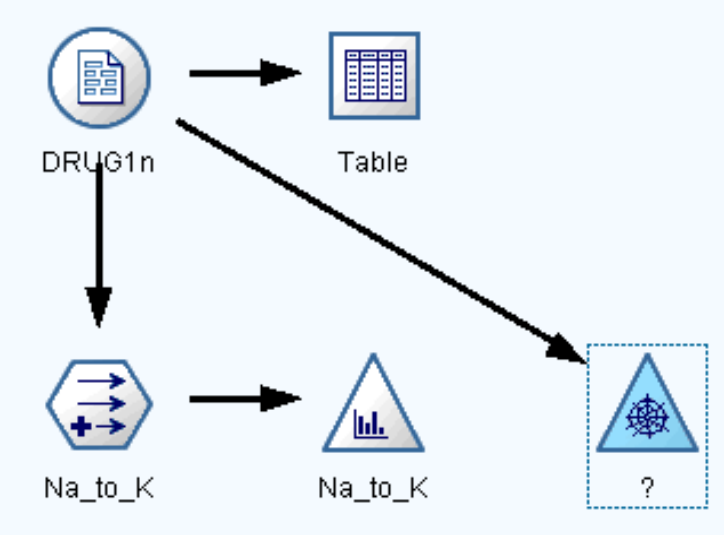
Trên nút ***Plot***, chọn Na là trường dữ liệu X, K là trường dữ liệu Y, Drug là trường dữ liệu che phủ. Sau đó, nhấn **Execute.**



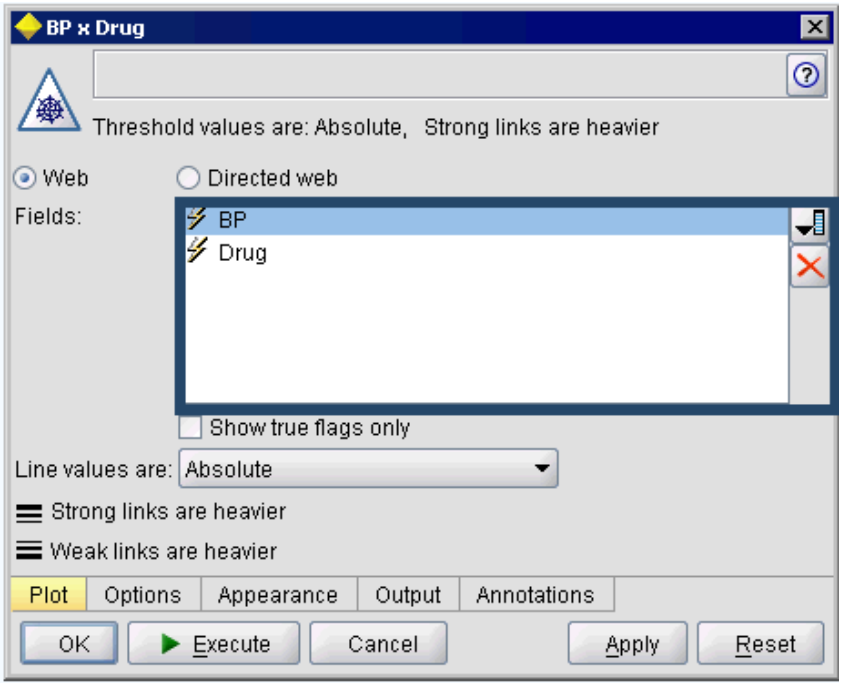
Kết quả cho thấy một tỷ lệ thuốc Y là lớn ở một vùng nhưng ở một vùng khác số lượng thuốc Y là ít . Đây là một tỷ lệ của natri (Na) với kali (K).



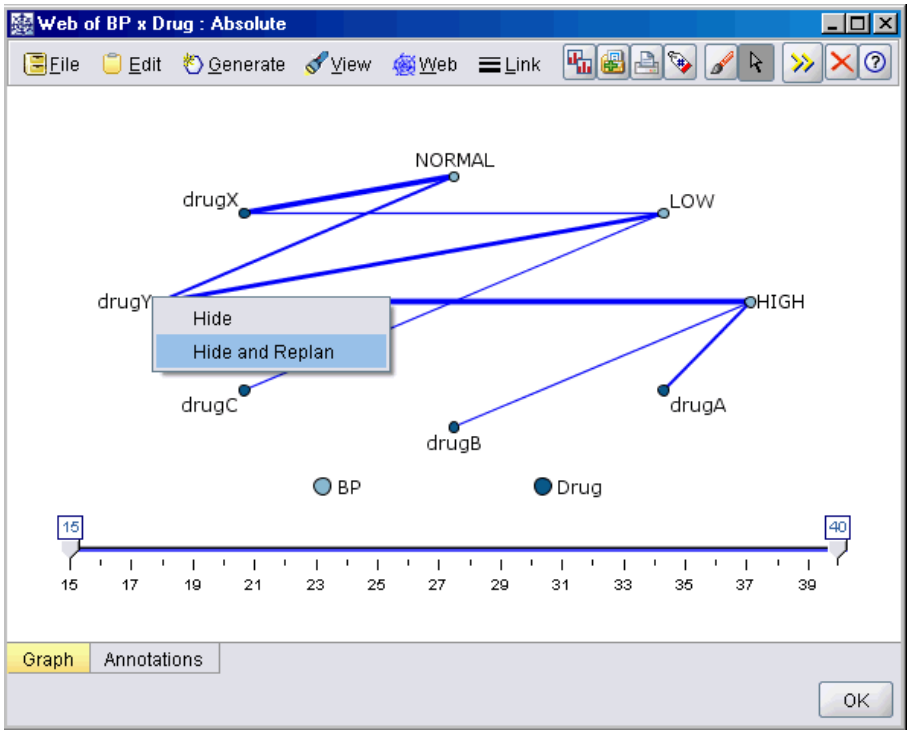
Từ các dữ liệu đã được tạo ra rõ ràng, chúng ta vẽ một đồ thị web. Bắt đầu bằng cách kết nối một nút ***Web*** sang nút ***Source* .**



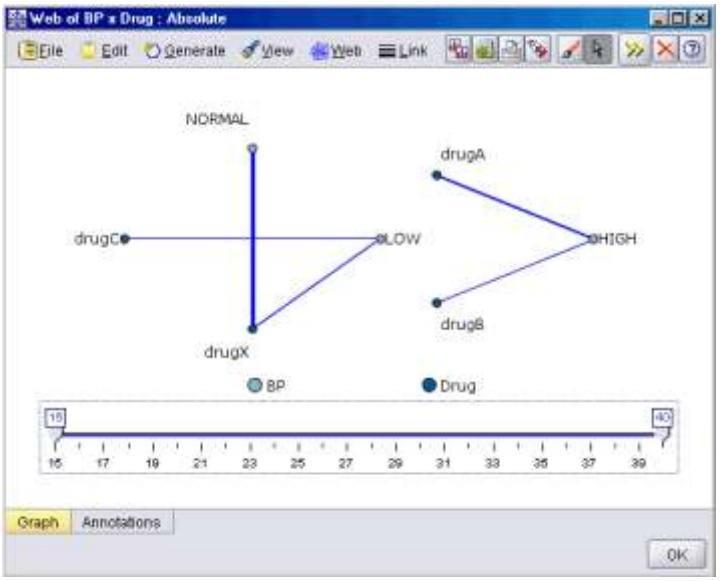
Trong hộp thoại Web, chọn BP (đối với huyết áp) và Drug(thuốc). Sau đó, nhấn ***Execute*** để chạy.



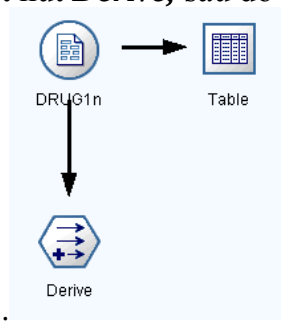
Ta thấy rằng thuốc Y là liên kết với tất cả ba cấp độ của huyết áp.



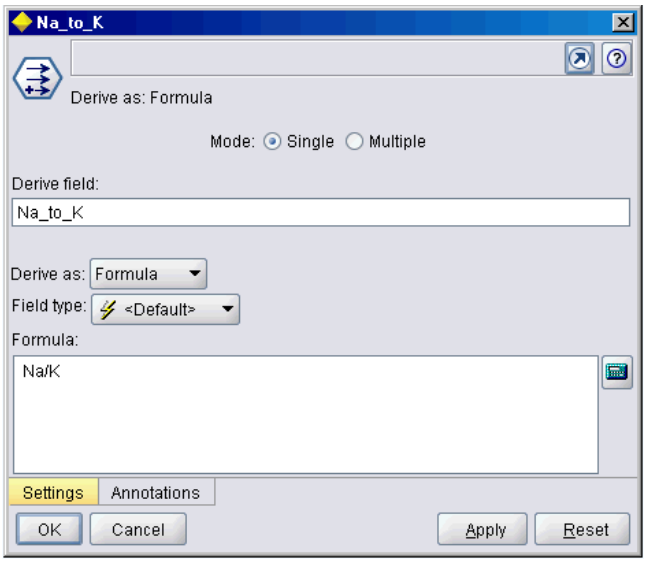
Thuốc Y và tất cả các liên kết của nó được ẩn. Có thể thấy rõ rằng chỉ có thuốc A và B có liên quan đến huyết áp cao. Chỉ có thuốc C và X có liên quan đến huyết áp thấp. Và huyết áp bình thường có liên quan với thuốc X.



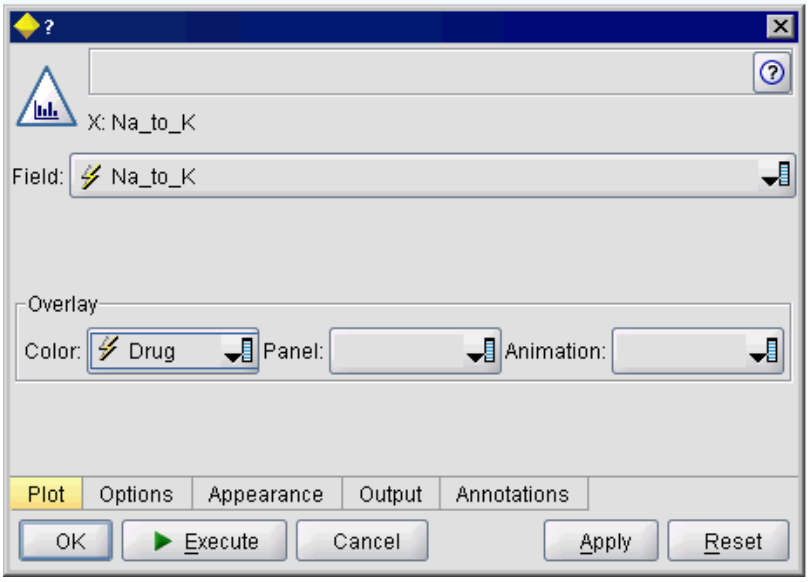
Bước tiếp theo ta chèn một nút ***Derive,*** sau đó kích đúp vào nút đó để chỉnh sửa



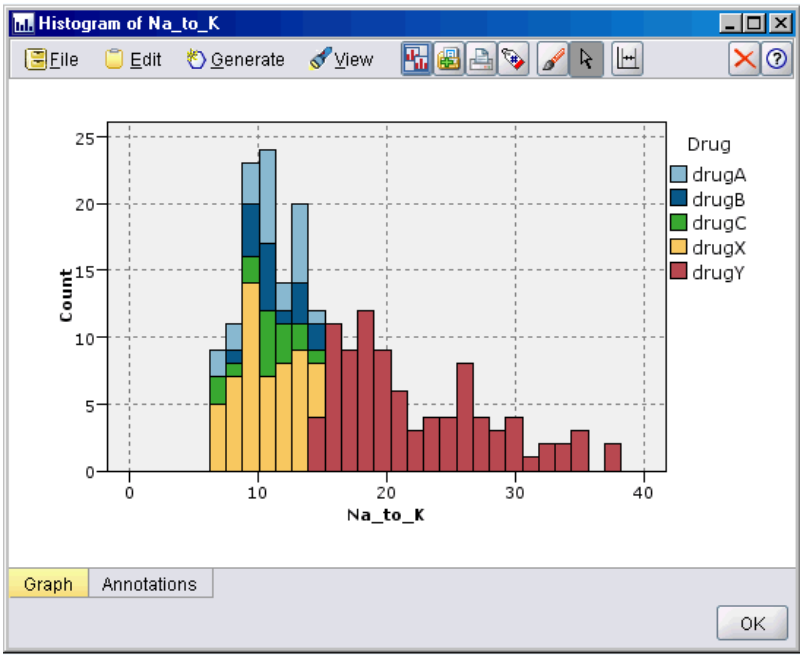
Tập tin mới có tên là Na_to_K . Vì có được những tập tin mới bằng cách chia giá trị của natri và kali (Na / K). Bạn cũng có thể tạo ra một lệnh bằng cách nhấp vào biểu tượng ngay bên phải của trường.



Kiểm tra sự phân bố của tập tin mới bằng cách gắn một nút ***Histogram*** sang nút nguồn gốc. Nhấp vào nút ***Execute*** để chạy chương trình.



Kết quả là một biểu đồ hiển thị . Dựa trên màn hình, bạn có thể kết luận rằng khi giá trị Na_to_K là lớn hơn hoặc bằng 15, thuốc Y là thuốc được lựa chọn.



CHƯƠNG 3: ÁP DỤNG CLEMENTINE VÀO BÀI TOÁN KHAI PHÁ DỮ LIỆU

Sản phẩm phần mềm SPSS Clementine là một phần mềm hữu ích trong việc thống kê dữ liệu và hỗ trợ trong khai phá dữ liệu. Như chúng ta đã biết để tìm kiếm các thông tin, tri thức ở những kho dữ liệu lớn là một việc khó khăn đối với những nhà thống kê học, đặc biệt đối với cơ sở dữ liệu về thống kê dân số có một tầm quan trọng lớn đối với nền kinh tế quốc dân, nên việc khai phá dữ liệu để tìm kiếm thông tin, tri thức đòi hỏi cần phải chính xác và nhanh chóng giúp cho Đảng và Nhà nước, các địa phương, các khu vực kinh tế tập thể, tư nhân…xây dựng kế hoạch phát triển kinh tế - xã hội nhằm nâng cao chất lượng cuộc sống của người dân. Có rất nhiều công cụ hỗ trợ trong lĩnh vự này nhưng SPSS Clementine là một phần mềm điển hình giúp người dùng xây dựng các mô hình khai phá dữ liệu cho toàn bộ quá trình của bài toán mà nó khác với các phương pháp truyền thống.

Trong phần này em đã áp dụng phần mềm SPSS Clementine để xây dựng các mô hình nhằm khai phá dữ liệu trong dữ liệu thống kê dân số của thành phố Hải Phòng năm 2009, từ đó rút ra một số tri thức có thể được áp dụng trong thực tế. Quy trình của bài toán như sau:

Bước 1 : Ban đầu có hai tập dữ liệu thô có tên là ***Nguoi.sav*** và ***ho.sav*** chứa đầy đủ thông tin như ; giới tính, tuổi, trình độ chuyên môn kĩ thuật, tổng số nam, nữ….

Bước 2 : Xử lý dữ liệu : sau quá trình tinh chỉnh dữ liệu được tập dữ liệu mới là :***Nguoi100_3.sa***v và ***ho_4.sav.***

Bước 3 : Biến đổi dữ liệu: làm mịn dữ liệu để đưa dữ liệu về dạng thuận lợi nhất.

Bước 4 : Khai phá dữ liệu : áp dụng những kỹ thuật phân tích, kỹ thuật thống kê để xử lý dữ liệu…từ đó tìm ra mối liên hệ giữa các thông tin.

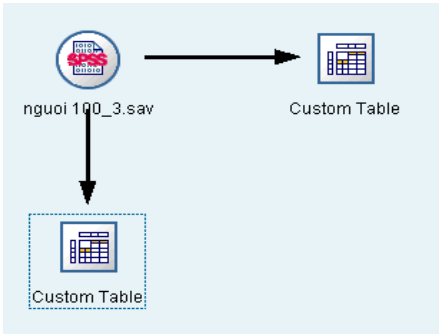
Bước 5 : Đánh giá kết quả và giải thích.

Ví dụ 1 :

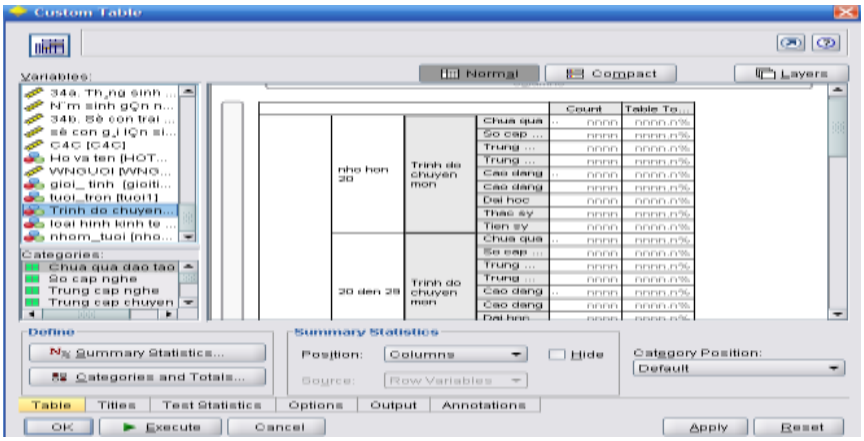
Từ File ***Nguoi100_3.sav,*** yêu cầu bài toán đặt ra là thống kê và so sánh tỷ lệ số lượng giữa nhóm độ tuổi và trình độ chuyên môn như thế nào. Sau đó vẽ biểu đồ để minh họa tỷ lệ đó.

Từ tab Source chọn SPSS File, kích đúp vào SPSS File và chọn tập dữ liệu ***Nguoi100_3.sav:***

Từ tập ***Nguoi100_3.sav***, chọn nút ***Custom Table***, màn hình xuất hiện như sau:



Nhấn đúp vào nút ***Custom Table***, trích chọn dữ liệu như hình sau:



Nhấn vào nút ***Execute*** để chạy.Kết quả là hình vẽ sau:

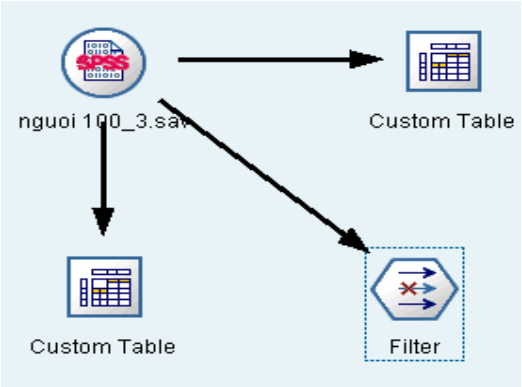
Custom Table #3

File Edit View Insert Pivot Format

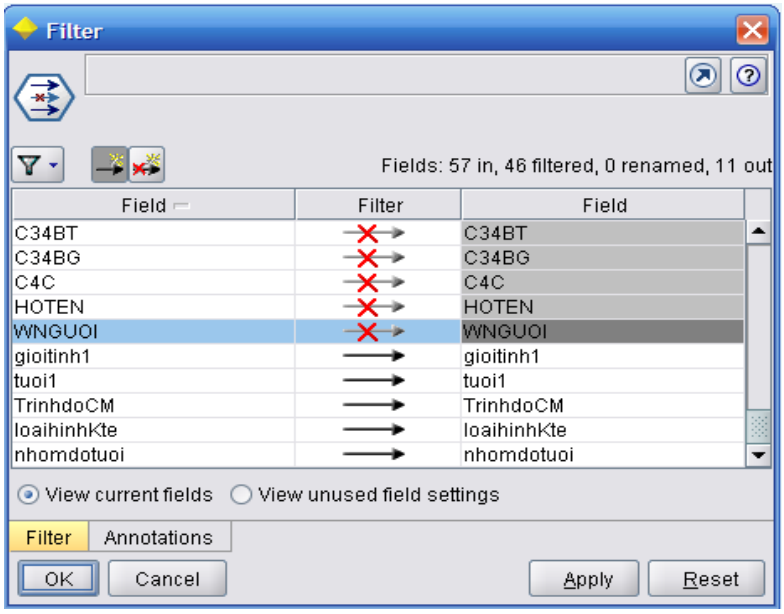
Table 1

				Count	Table Total N %
nhom_tuoi	nho hon 20	Trinh do chuyen mon	Chua qua dao tao CMKT	30208	10.6%
			So cap nghe	649	.2%
			Trung cap nghe	453	.2%
			Trung cap chuyen nghiep	170	.1%
			Cao dang nghe	85	.0%
			Cao dang	95	.0%
			Dai hoc	66	.0%
			Thac sy	0	.0%
			Tien sy	0	.0%
	20 den 29	Trinh do chuyen mon	Chua qua dao tao CMKT	31128	10.9%
			So cap nghe	2558	.9%
			Trung cap nghe	3842	1.3%
			Trung cap chuyen nghiep	2264	.8%
			Cao dang nghe	554	.2%
			Cao dang	1329	.5%
			Dai hoc	5111	1.8%
			Thac sy	80	.0%
			Tien sy	0	.0%
	30 den 39	Trinh do chuyen mon	Chua qua dao tao CMKT	31480	11.0%
			So cap nghe	1908	.7%
			Trung cap nghe	1659	.6%
			Trung cap chuyen nghiep	1243	.4%
			Cao dang nghe	197	.1%
			Cao dang	634	.2%
			Dai hoc	4074	1.4%
			Thac sy	131	0%

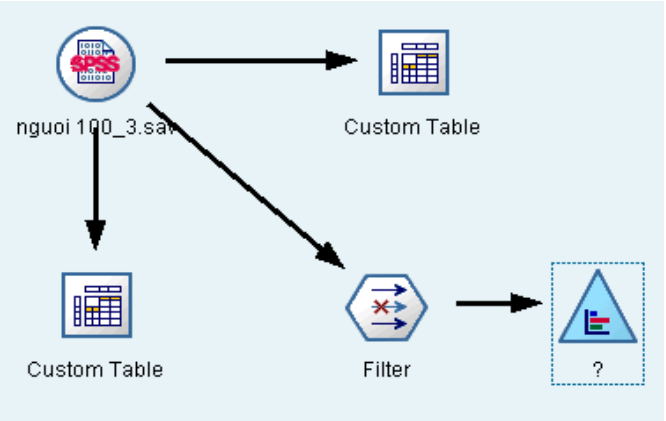
Chọn nút ***Filter***, màn hình xuất hiện như hình sau:



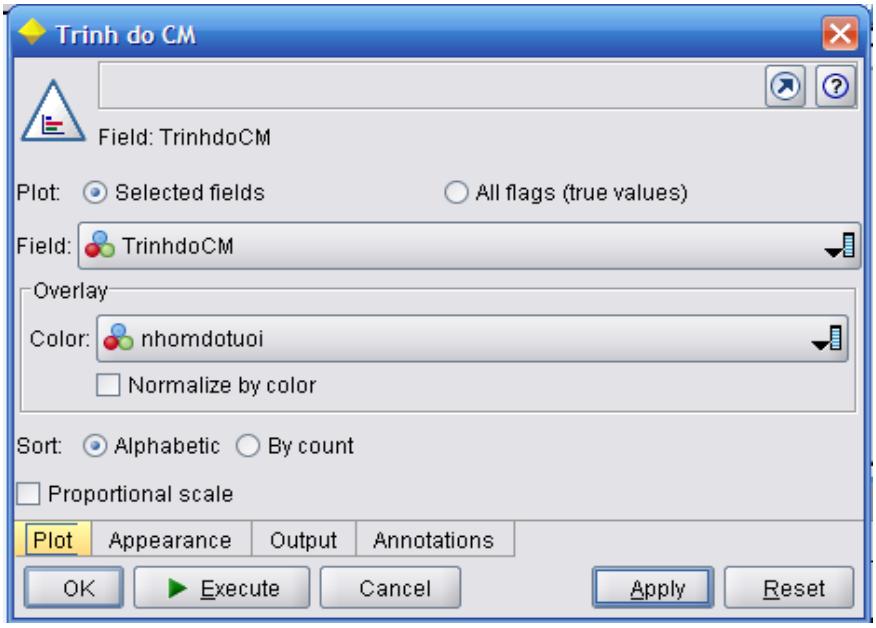
Nhấn đúp vào nút ***Filter***, loại bỏ một số trường dữ liệu không cần thiết.



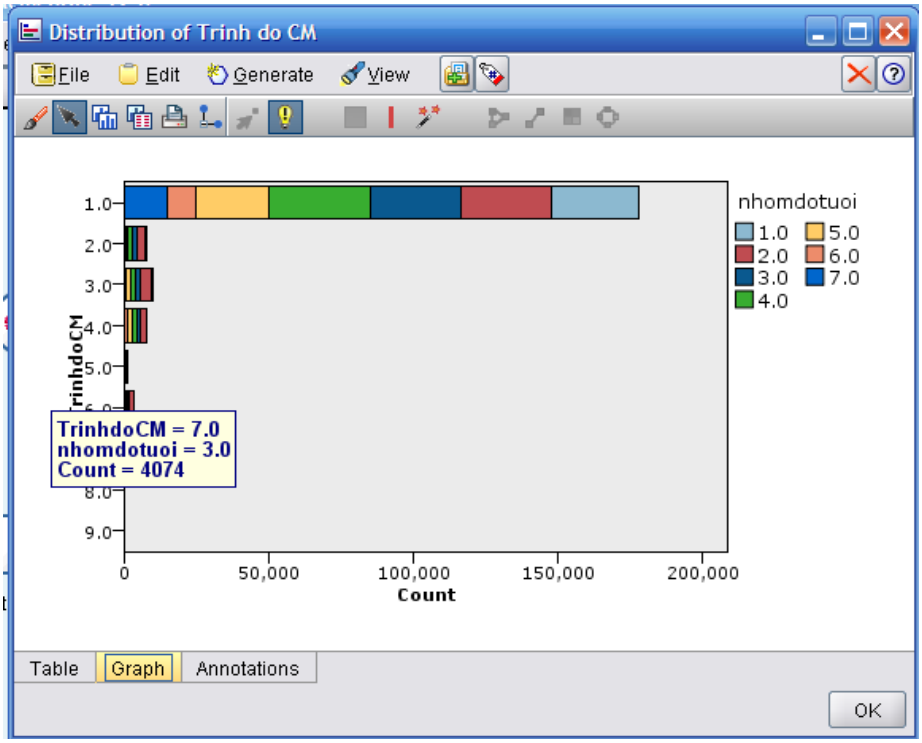
Để vẽ biểu đồ chọn nút ***Distribution***, màn hình xuất hiện như sau:



Nhấn đúp vào nút ***Distribution***, xử lý dữ liệu, nhấn ***Execute*** để chạy.



Kết quả là hình vẽ như sau:



Dựa vào bảng kết quả thống kê và biểu đồ cho ta thấy tỷ lệ số người chưa qua đào tạo chuyên môn kỹ thuật và sơ cấp nghề của các nhóm độ tuổi là cao.

Đồng thời kết quả cho thấy tỷ lệ số người ở trình độ trung cấp, cao đẳng, cao đẳng nghề là thấp, tỷ lệ số người ở trình độ đại học là bình thường.

Ví dụ 2:

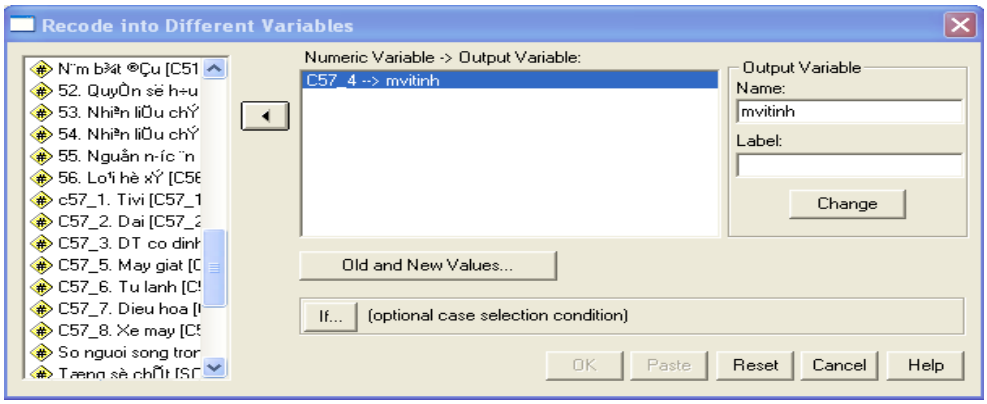
Từ file ***ho.sav*** gồm các thông tin như : tổng số người, tổng số nam, tổng số nữ, số hộ có tivi-máy vi tính-tủ lạnh-điều hòa-xe máy….

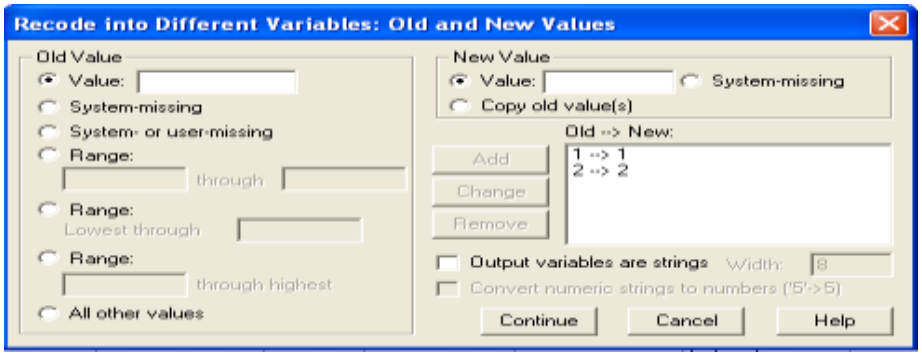
Yêu cầu bài toán đặt ra là : dựa vào những thông tin đó hãy thống kê và so sánh tỷ lệ có và không có các tiện nghi trong đời sống sinh hoạt, từ đó rút ra nhận xét.

Ban đầu có tập dữ liệu thô là ***ho.sav*** dưới dạng SPSS, ta đi tinh chỉnh dữ liệu.

Quá trình tinh chỉnh dữ liệu được thực hiện trong SPSS.

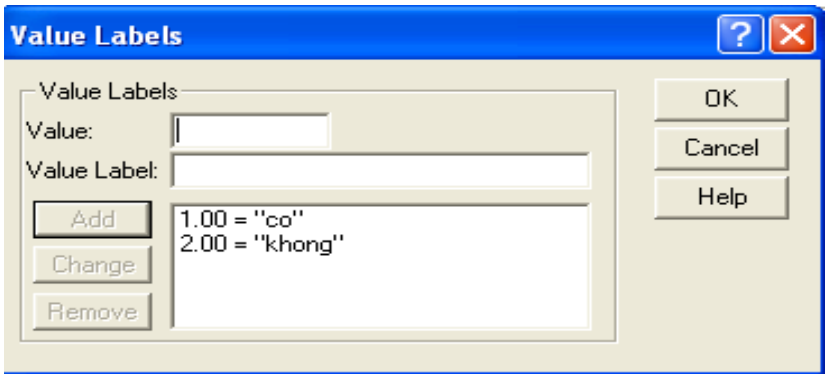
Chọn trường dữ liệu là ***C57_4*** : Máy vi tính





Gán lại trường dữ liệu : 1 – có

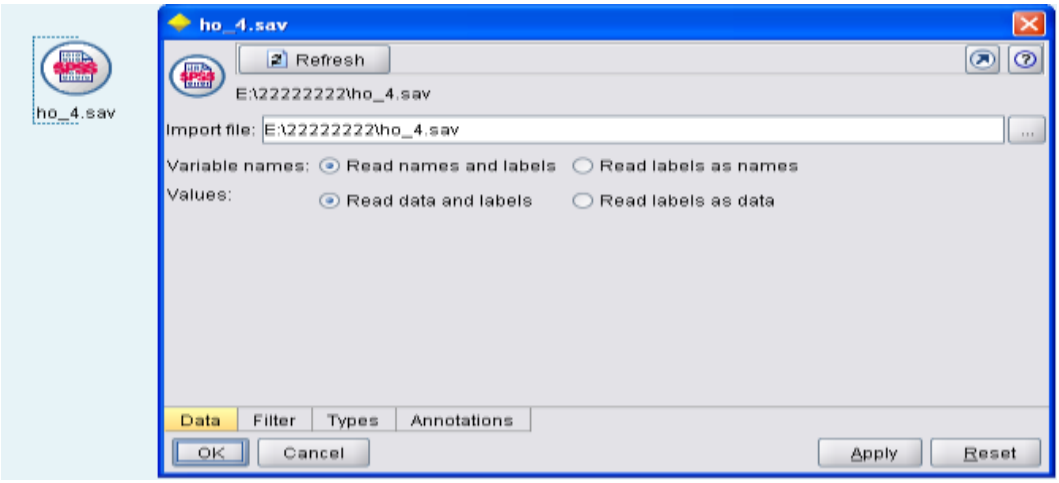
2 - không



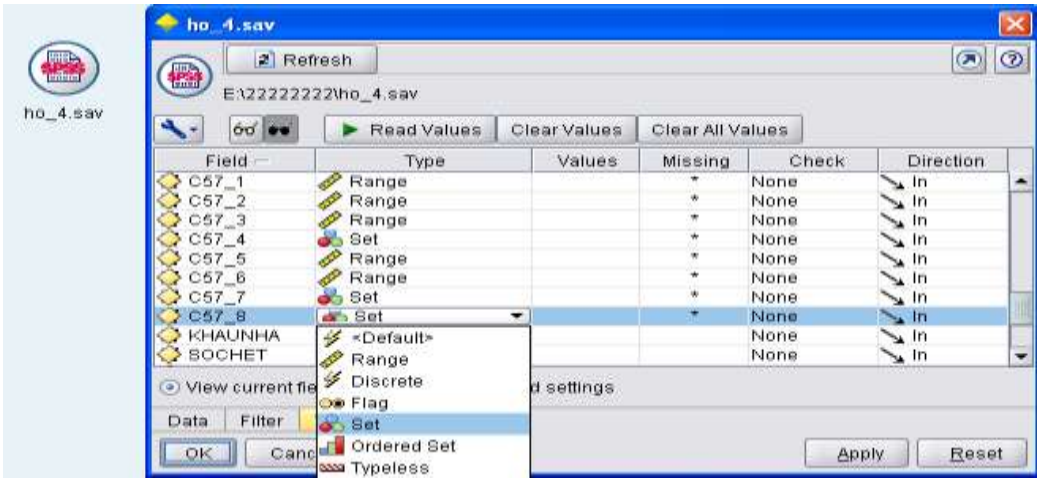
Sau quá trình tinh chỉnh :

Từ tab ***Source***, chọn nút ***SPSS File***

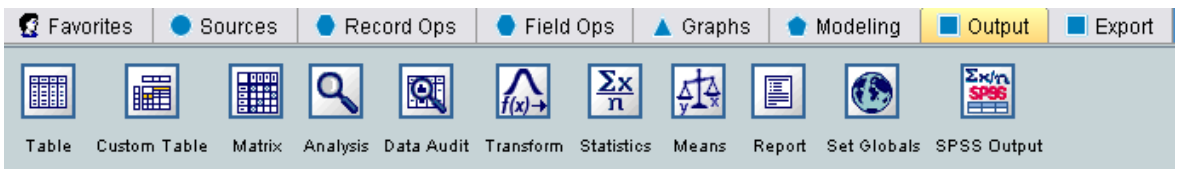
Kích đúp vào ***SPSS File*** và chọn tập dữ liệu ***ho_4.sav***



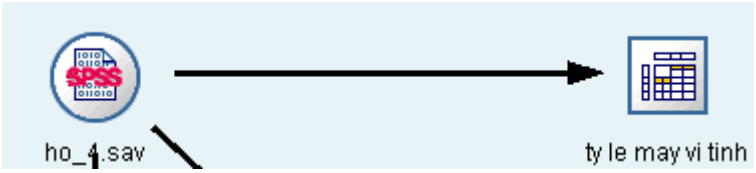
Nhấn vào tab ***Type*** để thay đổi loại dữ liệu



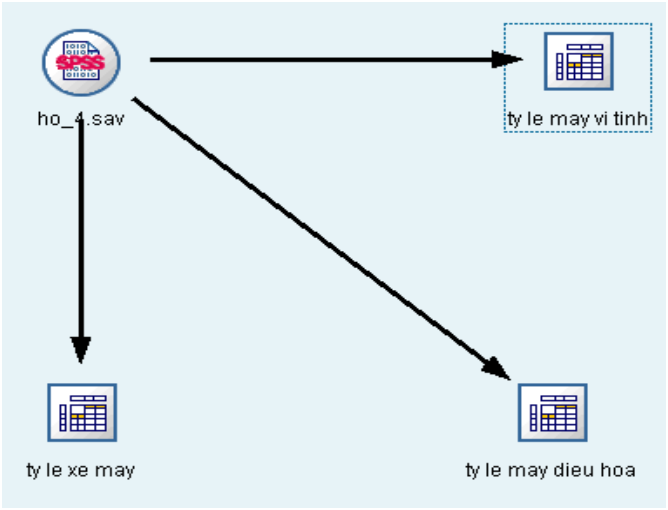
Từ tab ***Output*** chọn ***Custom Table*** để xuất dữ liệu:



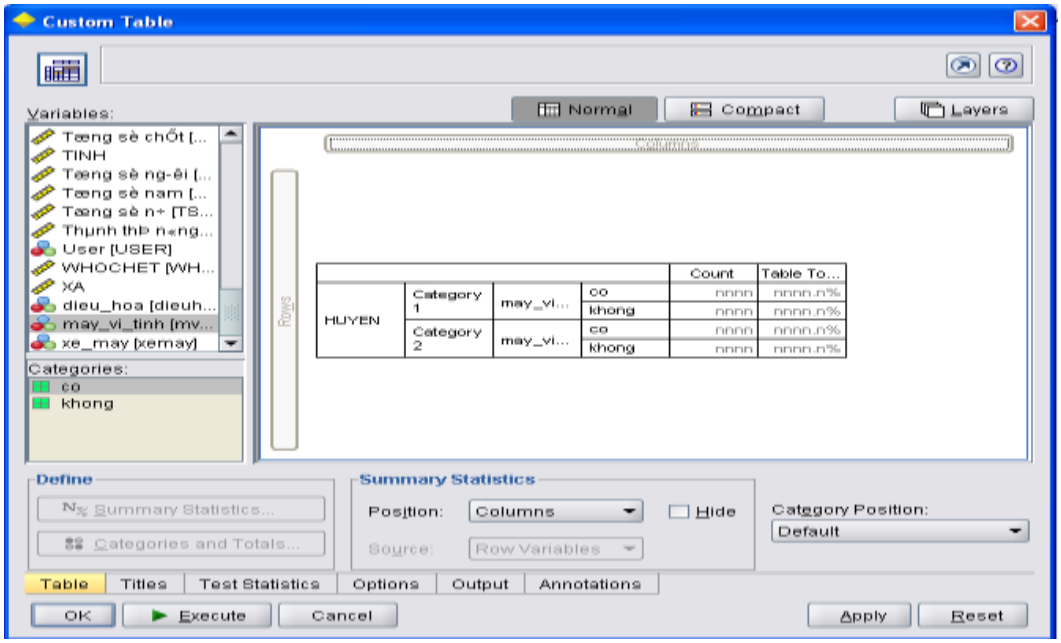
Kích đúp vào ***Custom Table*** và đổi tên ***Custom Table*** thành ***tỷ lệ máy tính*** như hình vẽ sau:



Làm tương tự đối với ***tỷ lệ máy điều hòa*** và ***tỷ lệ xe máy***.



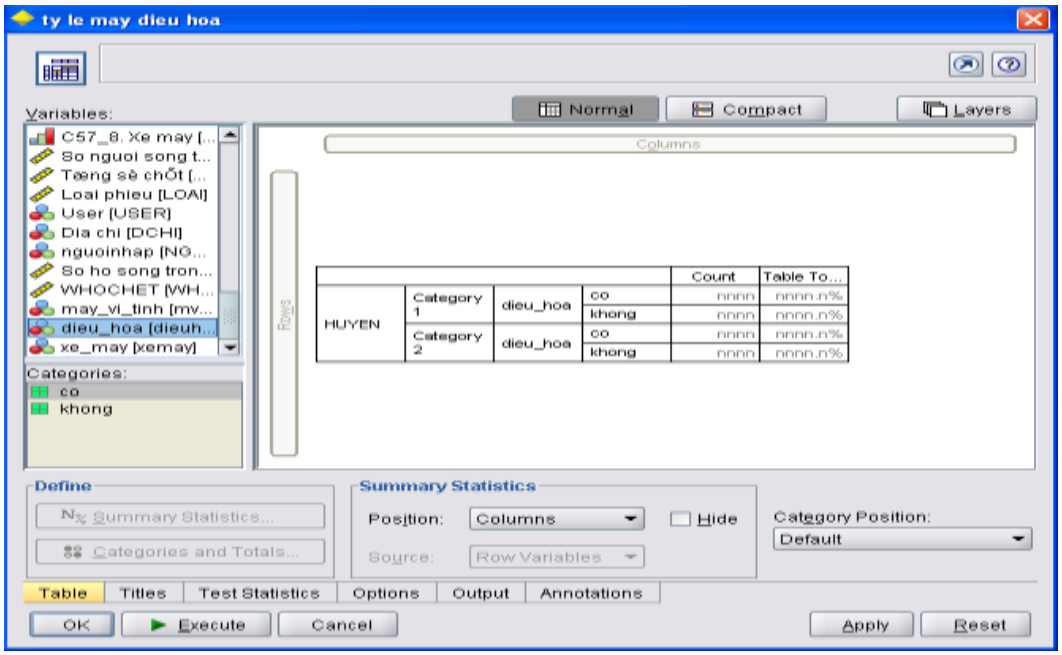
Kích đúp vào ***Custom Table*** và chọn trường dữ liệu ***may_vi_tinh*** và gắp thả chúng vào màn hình bên trong như hình vẽ sau:



Nhấn ***Execute*** để chạy, kết quả là bảng như sau:

				Count	Table Total N %
HUYEN	303	may_vi_tinh	co	2040	2.4%
			khong	4327	5.2%
	304	may_vi_tinh	co	2780	3.3%
			khong	3640	4.3%
	305	may_vi_tinh	co	2388	2.9%
			khong	4310	5.1%
	306	may_vi_tinh	co	1663	2.0%
			khong	4160	5.0%
	307	may_vi_tinh	co	1436	1.7%
			khong	5027	6.0%
	308	may_vi_tinh	co	646	.8%
			khong	4143	4.9%
	309	may_vi_tinh	co	642	.8%
			khong	4098	4.9%
	311	may_vi_tinh	co	807	1.0%
			khong	6255	7.5%
	312	may_vi_tinh	co	913	1.1%
			khong	5674	6.8%
	313	may_vi_tinh	co	315	.4%
			khong	5628	6.7%
	314	may_vi_tinh	co	372	.4%
			khong	5677	6.8%
	315	may_vi_tinh	co	265	.3%
			khong	5670	6.8%
	316	may_vi_tinh	co	212	.3%
			khong	6309	7.5%
	317	may_vi_tinh	co	237	.3%
			khong	3867	4.6%

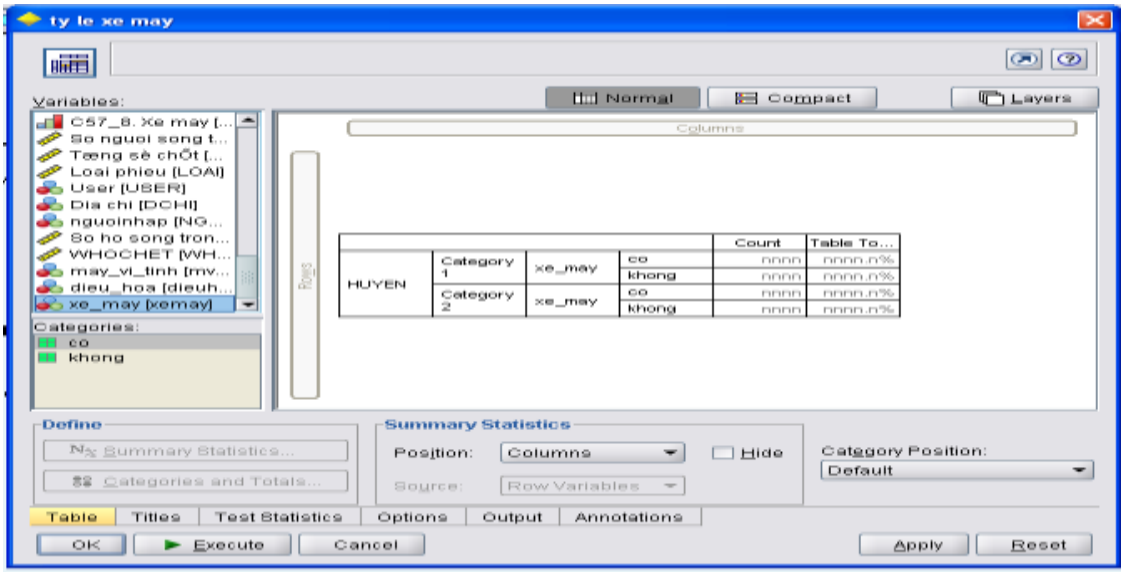
Tương tự đối với ***tỷ lệ máy điều hòa***.



Nhấn ***Execute*** để chạy, kết quả như bảng sau:

				Count	Table Total N %
HUYEN	303	dieu_hoa	co	1315	1.6%
			khong	5050	6.0%
	304	dieu_hoa	co	1522	1.8%
			khong	4899	5.9%
	305	dieu_hoa	co	1476	1.8%
			khong	5230	6.2%
	306	dieu_hoa	co	736	.9%
			khong	5084	6.1%
	307	dieu_hoa	co	480	.6%
			khong	5978	7.1%
	308	dieu_hoa	co	306	.4%
			khong	4484	5.4%
	309	dieu_hoa	co	150	.2%
			khong	4589	5.5%
	311	dieu_hoa	co	345	.4%
			khong	6719	8.0%
	312	dieu_hoa	co	343	.4%
			khong	6243	7.5%
	313	dieu_hoa	co	43	.1%
			khong	5900	7.0%
	314	dieu_hoa	co	48	.1%
			khong	6002	7.2%
	315	dieu_hoa	co	41	.0%
			khong	5890	7.0%
	316	dieu_hoa	co	61	.1%
			khong	6462	7.7%
	317	dieu_hoa	co	195	.2%
			khong	3911	4.7%

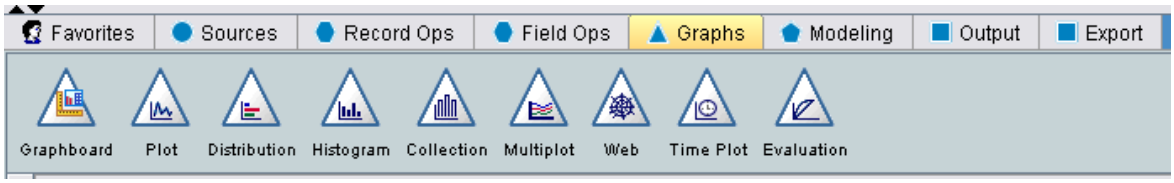
Tương tự đối với ***xe máy***:



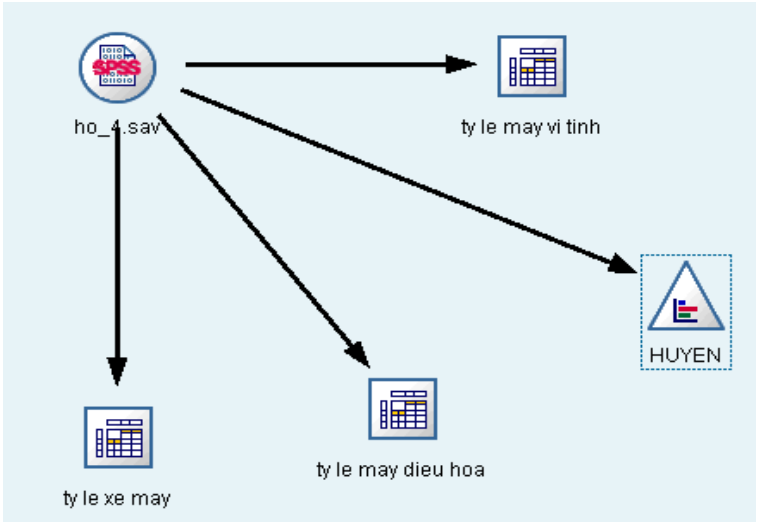
Nhấn ***Execute*** để chạy :

				Count	Table Total N %
HUYEN	303	xe_may	co	5177	6.2%
			khong	1200	1.4%
	304	xe_may	co	5361	6.4%
			khong	1077	1.3%
	305	xe_may	co	5588	6.7%
			khong	1133	1.4%
	306	xe_may	co	4739	5.7%
			khong	1092	1.3%
	307	xe_may	co	4829	5.8%
			khong	1652	2.0%
	308	xe_may	co	3266	3.9%
			khong	1528	1.8%
	309	xe_may	co	3533	4.2%
			khong	1216	1.5%
	311	xe_may	co	4665	5.6%
			khong	2399	2.9%
	312	xe_may	co	4911	5.9%
			khong	1686	2.0%
	313	xe_may	co	4134	4.9%
			khong	1814	2.2%
	314	xe_may	co	3915	4.7%
			khong	2145	2.6%
	315	xe_may	co	3404	4.1%
			khong	2535	3.0%
	316	xe_may	co	3234	3.9%
			khong	3294	3.9%
	317	xe_may	co	1823	2.2%
			khong	2283	2.7%

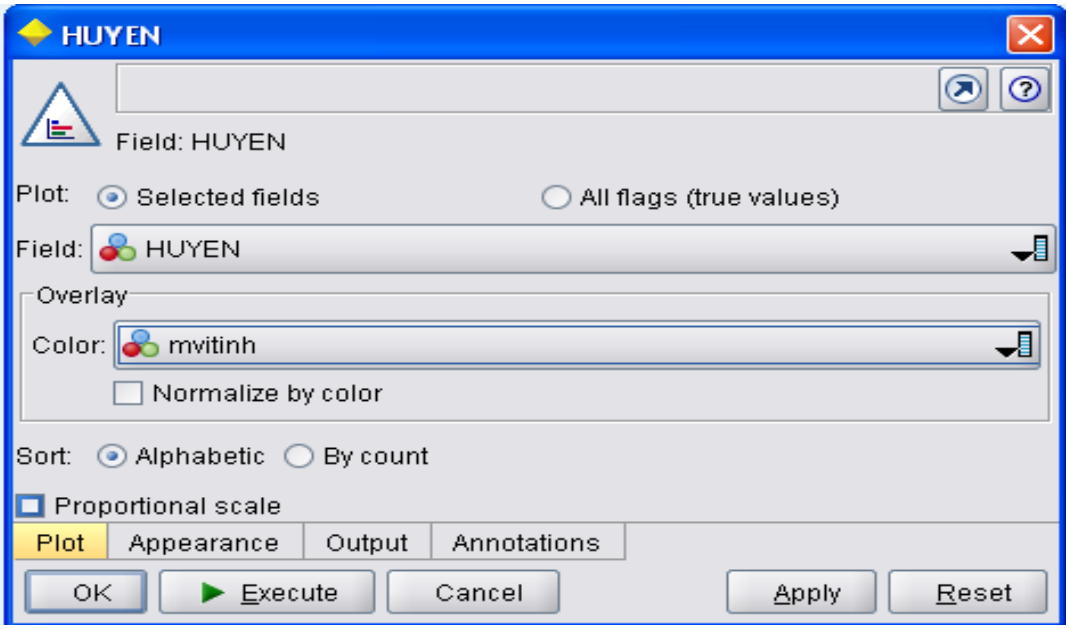
Để có cái nhìn trực quan hơn đối với kết quả thu được ta dùng đồ thị để biểu diễn kết quả đó. Để chọn loại biểu đồ nhấn vào tab ***Graphs*** như hình sau:



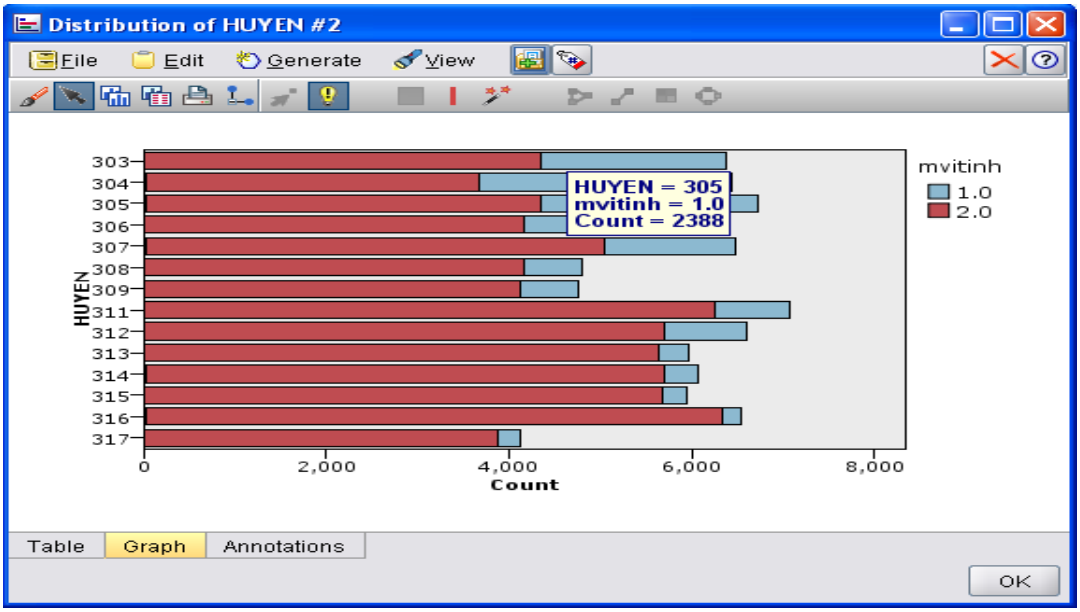
Kích đúp vào ***Distribution*** hoặc gắp thả chúng vào màn hình bên trong.



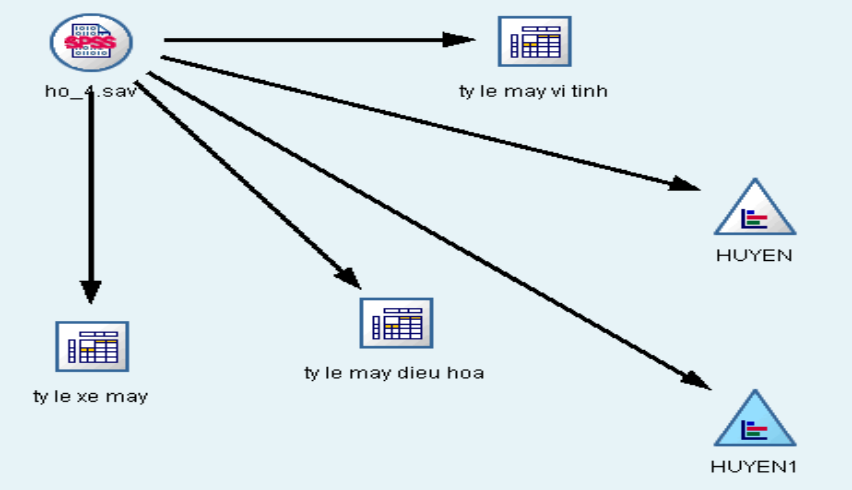
Kích đúp vào ***Distribution*** và chon trường dữ liệu như hình vẽ:

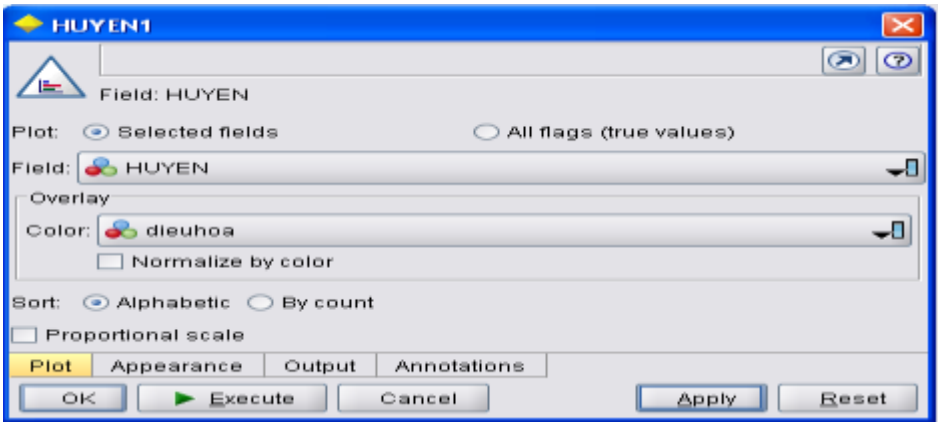


Nhấn ***Execute*** để chạy:

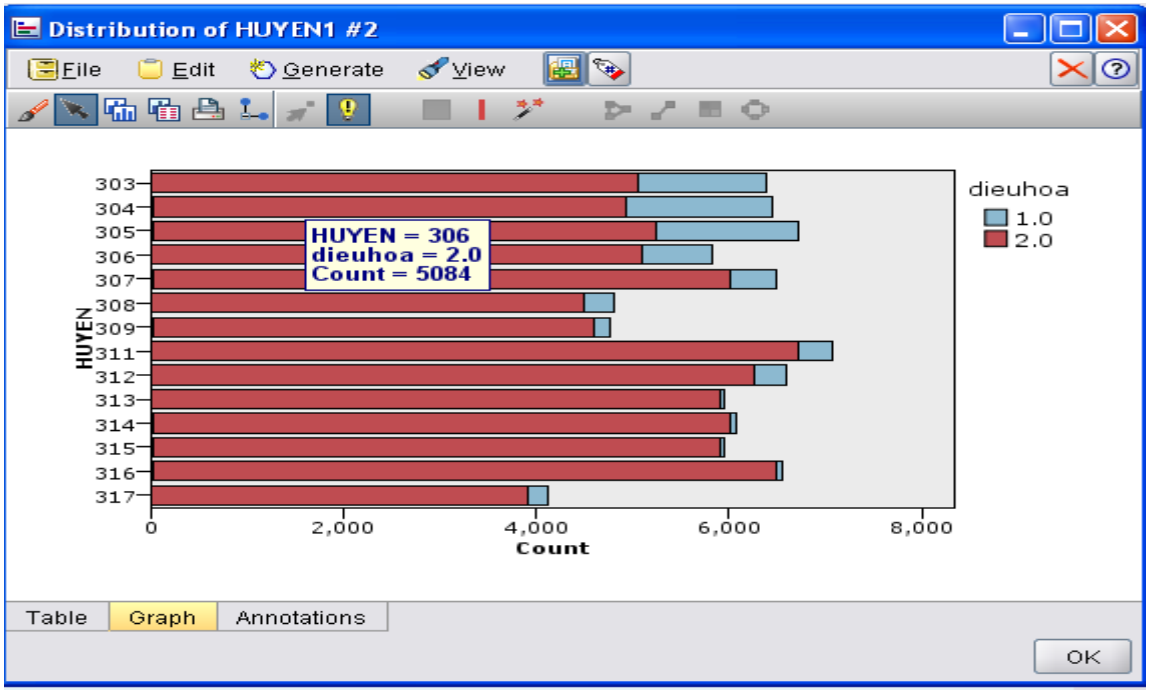


Làm tương tự đối với ***tỷ lệ máy điều hòa***.





Nhấn ***Execute*** để chạy:



Dựa vào bảng thống kê và biểu đồ cho thấy tỷ lệ người dân có sử dụng các thiết bị đồ dân dụng như: máy vi tính, điều hòa, xe máy…là chưa cao và đặc biệt đối với các huyện ở ngoại thành và vùng sâu xa thì tỷ lệ đó là thấp.

Điều đó có thể cho thấy nhu cầu sử dụng các thiết bị đó của người dân là cao, và đây là nhân tố có vai trò quan trọng trong việc đề ra tiêu chí kinh doanh đối với các nhà sản xuất và các nhà cung cấp các loại thiết bị đồ dân dụng đó.

KẾT LUẬN

Lĩnh vực khai phá dữ liệu là một lĩnh vực còn khá mới ở Việt Nam. Trong đồ án này em đã trình bày tổng quan về khai phá dữ liệu, các thuật toán trong phân cụm dữ liệu…, Đồng thời tìm hiểu về phần mềm chuyên dụng trong thống kê SPSS Clementine và ứng dụng vào việc xây dựng các mô hình trong việc khai phá dữ liệu thống kê dân số năm 2009 của thành phố Hải Phòng. Đây cũng là mục tiêu chính của đồ án.

Như chúng ta đã biết Khai phá dữ liệu là quá trình tìm kiếm và phát hiện ra tri thức mới. Để phát hiện ra tri thức mới đó cần xây dựng mô hình tổng quát hóa để khai phá dữ liệu đó. Phần mềm Clementine giúp giải quyết vấn đề đó.

Trong đồ án này, đã áp dụng phần mềm Clementine xây dựng mô hình để khai phá dữ liệu trong tập dữ liệu thống kê dân số của thành phố Hải Phòng năm 2009 trong một số tiêu chí. Từ đó phát hiện ra các tri thức mới hỗ trợ cho việc ra quyết định.

TÀI LIỆU THAM KHẢO

[1] . M. Kantardzic: Data Mining: Concepts, Models, Method, and Algorithms, John Wiley & Sons, New York, NY, 2003.

[2] . J. Grabmeier, and A. Rudolph: Techniques of Clustreing Algorithms in Data Mining, Data Mining and Knowledge Discovery, Kluwer Academic Publishers, Netherlands, pp.303-360, 2002.

[3] . R. Mattision: Data Warehousing and Data Mining for Telecommunications, Norwood, MA, 1997.

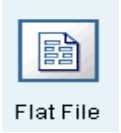
[4] . Sổ tay điều tra viên địa bàn mẫu, Ban chỉ đạo tổng điều tra dân số và nhà ở trung ương, Tổng cục thống kê – Cục thống kê thành phố Hải Phòng, 11/2008.

[5] . Hoàng Hải Xanh : Các kỹ thuật phân cụm trong Datamining, luận văn – Đại học Công nghệ, Đại học quốc gia Hà Nội.

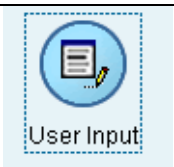
PHỤ LỤC A: CÁC NÚT ĐỂ XÂY DỰNG MÔ HÌNH

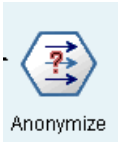
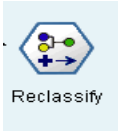
BIỂU TƯỢNG	MÔ TẢ
1. Database 	Các nút Database (cơ sở dữ liệu) có thể được sử dụng để nhập dữ liệu từ một loạt các gói khác bằng cách sử dụng ODBC (Open Database Connectivity), bao gồm Microsoft SQL Server, DB2, Oracle, và những người khác.
2. Var.file 	Các nút File Variable (biến) đọc dữ liệu từ các tập tin văn bản-có nghĩa là, các tập tin có chứa một số bản ghi của trường dữ liệu. Nút này cũng hữu ích cho các tập tin với các văn bản lớn và một số loại khác.
3. Select 	Nút Select : lựa chọn hoặc loại bỏ một tập hợp các bản ghi từ các luồng dữ liệu dựa trên một điều kiện cụ thể. Ví dụ, bạn có thể chọn các bản ghi mà liên quan đến một khu vực bán hàng cụ thể.
4. Sample 	Nút Sample (mẫu) chọn một tập hợp con của bản ghi. Một loạt các loại mẫu được hỗ trợ, bao gồm phân tầng, cụm, và nonrandom (cấu trúc) mẫu. Lấy mẫu có thể hữu ích để cải thiện hiệu suất, và để chọn nhóm các hồ sơ có liên quan hoặc các giao dịch để phân tích.
5. Aggregate 	Nút Aggregate (tổng hợp) thay thế một chuỗi các bản ghi đầu vào với tổng số lượng bản ghi .
6. Derive 	Nút Derive (thêm) đổi giá trị dữ liệu hoặc tạo ra các trường dữ liệu mới từ một hoặc nhiều trường dữ liệu hiện có. Nó tạo ra các trường dữ liệu loại: cờ, thiết lập, số và có điều kiện.
7. Type	Nút Type: xác định loại dữ liệu và giá trị của dữ liệu. Ví dụ, bạn có thể chỉ định một loại sử dụng (phạm vi, thiết lập, ra lệnh thiết

BIỂU TƯỢNG	MÔ TẢ
	lập, hay flag) cho mỗi trường dữ liệu, thiết lập các tùy chọn để xử lý các giá trị null mất tích và hệ thống, thiết lập vai trò của một trường dữ liệu cho các mục đích làm mẫu, xác định trường dữ liệu và các nhãn giá trị, và xác định giá trị cho một trường..
8. Filter 	Nút Filter lọc (loại bỏ) các lĩnh vực, đặt lại tên các trường, và bản đồ các trường từ nút một trong những nguồn khác.
9. Plot 	Nút Plot cho thấy mối quan hệ giữa các trường số. Bạn có thể tạo ra một bản đồ bằng cách sử dụng các điểm (a scatterplot) hoặc dòng.
10.Distribution 	Nút Distribution(phân phối) cho thấy sự xuất hiện của giá trị, chẳng hạn như loại thế chấp, giới tính. Thông thường, bạn có thể sử dụng các nút Distribution cho thấy sự mất cân bằng trong các dữ liệu, mà bạn sau đó có thể khắc phục bằng cách sử dụng một nút cân bằng trước khi tạo ra một mô hình.
11. Histogram 	Nút Histogram cho thấy sự xuất hiện của các giá trị cho các trường số. Nó thường được dùng để khám phá những dữ liệu trước khi thao tác và xây dựng mô hình. Tương tự như nút Distribution, các nút Histogram thường xuyên cho thấy sự mất cân bằng trong các dữ liệu.
12. Neural Net 	Nút Neural Net (nút mạng nơ-ron) sử dụng một mô hình đơn giản giông cách làm việc của bộ não con người xử lý thông tin. Nó hoạt động bằng cách mô phỏng một số lượng lớn các đơn vị xử lý kết nối đơn giản, giống như các phiên bản trừu tượng của tế bào thần kinh. Mạng lưới thần kinh có rất nhiều chức năng lập dự toán nói chung và yêu cầu tối thiểu thống kê hoặc kiến thức toán học để đào tạo hoặc áp dụng.
13. Kohonen	Nút Kohonen tạo ra một loại mạng nơron có thể được sử dụng để tập hợp các nhóm dữ liệu vào các nhóm riêng biệt. Khi mạng được thiết lập đầy đủ, bản ghi sẽ xuất hiện gần nhau trên bản đồ

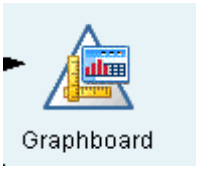
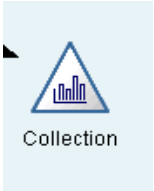
BIỂU TƯỢNG	MÔ TẢ
	số lượng, trong khi bản ghi sẽ xuất hiện khác nhau xa.
14. C5.0 	Nút C5.0 hoặc xây dựng một cây quyết định hoặc quy định một bộ. Mô hình này hoạt động bằng cách chia tách các mẫu dựa trên trường dữ liệu cung cấp các thông tin thu được tối đa ở từng cấp. Các mục tiêu phải rõ ràng. Được chia thành nhiều hơn hai nhóm con .
15. C&R Tree 	Nút C&R Tree (phân loại và hồi qui) là một cây dựa trên phân loại và phương pháp dự báo. Tương tự như C5.0, phương pháp này sử dụng phân vùng đệ quy chia các bản ghi thành các đoạn có giá trị tương tự như đầu ra. Bắt đầu bằng cách kiểm tra các trường dữ liệu đầu vào để tìm chia tốt nhất, chia thành hai nhóm con, sau đó chia thành hai nhóm con nhiều hơn, và như vậy, cho đến khi một trong những tiêu chí dừng được kích hoạt. Tất cả chia tách là nhị phân.
16. K-Means 	Các nút K-Means nhóm các tập dữ liệu thành các nhóm khác nhau (hoặc cụm). Xác định một số phương pháp cố định của các cụm, lặp đi lặp lại giao bản ghi cho các cụm, và điều chỉnh các trung tâm cụm sàng lọc hơn nữa cho đến khi không còn có thể cải thiện mô hình. Thay vì cố gắng để dự đoán một kết quả, có nghĩa là k-mean sử dụng một quá trình được gọi là học không có giám sát để phát hiện ra trong tập hợp các trường dữ liệu đầu vào.
17. Table 	Nút Table (bảng) hiển thị dữ liệu ở định dạng bảng, mà cũng có thể được ghi vào một tập tin. Điều này rất hữu ích bất cứ lúc nào mà ngươi sư dụng cần phải kiểm tra các giá trị dữ liệu của bạn hay xuất chúng để có thể dễ dàng đọc được.
18. Flat File 	Nút Flat File ghi dữ liệu vào một tập tin văn bản. Nó rất hữu dụng để xuất dữ liệu có thể được đọc bởi các phân tích khác hoặc bảng tính.

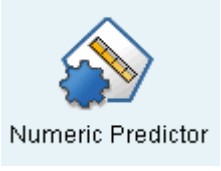
BIỂU TƯỢNG	MÔ TẢ
19. Enterprise View 	Nút Enterprise View tạo ra một kết nối đến một doanh nghiệp, cho phép bạn đọc Enterprise View dữ liệu vào dòng một và đóng gói một mô hình trong một sự kiện có thể được truy cập từ các kho lưu trữ của người dùng khác.
20. Fixed File 	Nút Fixed File : nhập dữ liệu từ trường dữ liệu không đổi từ các tập tin văn bản-có nghĩa là, các tập tin mà các trường là không giới hạn, nhưng bắt đầu từ vị trí tương tự và đang có chiều dài cố định. Kho dữ liệu được tạo ra hoặc di chuyển thường được lưu trữ ở dạng cố định.
21. SPSS File 	Nút SPSS File xuất dữ liệu đầu ra trong SPSS. Định dạng SAV. Các. File SAV có thể được đọc bởi SPSS Base và các sản phẩm khác. Đây cũng là định dạng được sử dụng cho các tập tin bộ nhớ cache trong Clementine.
22. Dimensions 	Nút Dimensions (Kích thước) xuất dữ liệu đầu ra trong các định dạng được sử dụng bởi phần mềm SPSS như nghiên cứu thị trường, dữ liệu Thư viện phải được cài đặt để sử dụng nút này.
23. SAS File 	Các nút SAS nhập dữ liệu vào Clementine.
24. Excel 	Nút Excel xuất dữ liệu đầu ra ở định dạng Microsoft Excel (xls).. Tùy chọn, bạn có thể chọn để khởi động Excel tự động và mở tập tin xuất dữ liệu khi nút được thực hiện.
25. User Input	Nút User Input : tạo hoặc tổng hợp dữ liệu từ đầu hoặc bằng cách thay đổi dữ liệu hiện có. Điều này rất hữu ích, ví dụ, khi bạn muốn tạo ra một tập dữ liệu kiểm tra cho người mẫu

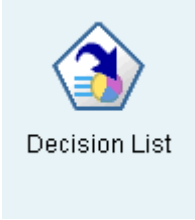
BIỂU TƯỢNG	MÔ TẢ
	
26. Balance 	Nút Balance (cân bằng) sửa chữa sự mất cân bằng trong một tập dữ liệu, vì vậy nó phù hợp với một điều kiện quy định. Các chỉ thị điều chỉnh cân bằng tỷ lệ của hồ sơ, nơi mà một điều kiện là đúng bởi các yếu tố quy định.
27. RFM Aggregate 	Các Recency, Tần số, tiền tệ (RFM) cho phép bạn có được nhật ký dữ liệu của khách hàng giao dịch, truyền đi bất kỳ dữ liệu không sử dụng, và kết hợp tất cả dữ liệu giao dịch còn lại của chúng vào một hàng duy nhất mà các danh sách khi chúng mới bị xử lý, có bao nhiêu giao dịch đã đưa ra, và tổng giá trị tiền tệ của những người giao dịch.
28. Sort 	Nút Sort : Phân loại các loại bản ghi vào tăng hay giảm dựa trên các giá trị của một hoặc nhiều trường dữ liệu.
29. Merge 	Nút Merge: chứa nhiều bản ghi đầu vào và tạo ra một bản ghi đầu ra có chứa một số hoặc tất cả các trường dữ liệu đầu vào. Nó rất hữu ích cho việc hợp nhất dữ liệu từ các nguồn khác nhau.
30. Append 	Nút Append : Hữu dụng cho việc kết hợp với cấu trúc tương tự tập hợp dữ liệu nhưng dữ liệu khác nhau.
31. Distinct 	Nút Distinct : loại bỏ các bản ghi, hoặc bằng cách loại bỏ bản ghi đầu tiên và chuyển qua bất kỳ bản sao để đến các dòng dữ liệu thay thế.

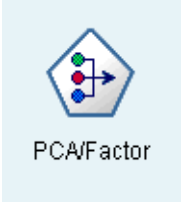
BIỂU TƯỢNG	MÔ TẢ
32. Ensemble 	Nút Ensemble : kết hợp hai hoặc nhiều mô hình nuggets để có được những dự đoán chính xác hơn có thể thu được từ bất kỳ mô hình một.
33. Filler 	Nút Filler : điền các giá trị thay thế trường dữ liệu và lưu trữ các thay đổi. Bạn có thể chọn để thay thế các giá trị dựa trên một điều kiện Clem, chẳng hạn như @ BLANK (@ lĩnh vực). Ngoài ra, bạn có thể chọn để thay thế tất cả các khoảng trống hoặc giá trị null với một giá trị cụ thể. Một nút Filler thường được sử dụng cùng với một nút Type để thay thế các giá trị mất tích.
34. Anonymize 	Nút Anonymize : biến đổi tên trường và các giá trị được đại diện ở cuối cùng, do đó làm dữ liệu gốc. Điều này có thể hữu ích nếu bạn muốn cho phép người khác xây dựng mô hình bằng cách sử dụng dữ liệu nhạy cảm, chẳng hạn như tên khách hàng hoặc các chi tiết khác.
35. Reclassify 	Nút Reclassify : biến đổi một tập các giá trị rời rạc khác nhau. Reclassify rất hữu dụng cho loại dữ liêu bị mất hoặc tập hợp dữ liệu để phân tích.
36. Binning 	Nút Binning (tự động) : tạo ra các trường dữ liệu được thiết lập mới dựa trên các giá trị của một hoặc nhiều trường dữ liệu .
37. RFM Analysis 	Nút RFM Analysis : xác định số lượng mà khách hàng có thể là tốt nhất bằng cách kiểm tra gần đây họ mới mua cái gì của cửa hàng bạn (recency), mức độ thường xuyên mua (tần số), và có bao nhiêu giao dịch (tiền tệ)
38. Partition	Nút Partition (phân vùng) tạo ra một phân vùng trường dữ liệu, tách các dữ liệu vào các tập con riêng cho việc đào tạo, thử

BIỂU TƯỢNG	MÔ TẢ
	nghiệm, và giai đoạn xây dựng mô hình .
39. SetToFlag 	Nút SeToFlag : Tập hợp nhiều trường dữ liệu gốc dựa trên các giá trị cụ thể được xác định cho một hoặc nhiều lĩnh vực đã được thiết lập.
40. Restructure 	Restructure (Cơ cấu lại) chuyển đổi một nhóm hoặc trường dữ liệu vào một nhóm các trường có thể được dân cư với các giá trị của trường dữ liệu nào khác. Ví dụ, cho một trường dữ liệu được đặt tên loại thanh toán, với giá trị của tín dụng, tiền mặt, và ghi nợ, ba trường dữ liệu mới sẽ được tạo ra (tín dụng, tiền mặt, thẻ ghi nợ), mỗi người mà có thể chứa giá trị thực tế thanh toán được thực hiện.
41. Transpose 	Nút Transpose : hoán đổi dữ liệu trong hàng và cột để ghi trở thành trường dữ liệu và các trường dữ liệu trở thành các bản ghi
42. Time Intervals 	Nút Time Intervals (khoảng thời gian) : quy định và tạo ra các khoảng nhãn (nếu cần) cho dữ liệu mô hình chuỗi thời gian.
43. History 	History tạo ra các trường dữ liệu mới có chứa các dữ liệu từ các trường dữ liệu trong bản ghi trước đó. Được sử dụng cho dữ liệu tuần tự, chẳng hạn như dữ liệu chuỗi thời gian. Trước khi sử dụng một nút History, bạn có thể muốn sắp xếp dữ liệu bằng cách sử dụng một nút Sort.
44. SPSS	Nút SPSS Transform : thực hiện biến đổi dữ liệu bằng cách sử

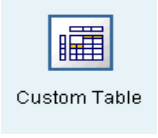
BIỂU TƯỢNG	MÔ TẢ
Transform 	dụng cú pháp lệnh SPSS.
45. Field Reorder 	Nút Field Reorder : xác định trình tự xuất hiện các trường dữ liệu trung bình. Chúng ảnh hưởng đến màn hình hiển thị của các trường trong một loạt các địa điểm, chẳng hạn như các bảng, danh sách, và Field Chooser. Chúng rất hữu ích khi làm việc với tập hợp các dữ liệu rộng để làm cho các trường dữ liệu được rõ ràng hơn.
46. Graphboard 	Nút Graphboard : cung cấp nhiều loại khác nhau của đồ thị . Sử dụng nút này, bạn có thể chọn các trường dữ liệu mà bạn muốn khai thác và sau đó chọn một đồ thị từ những người có sẵn cho các dữ liệu được chọn. nút sẽ tự động lọc ra bất cứ loại đồ thị mà sẽ không làm việc với những lựa chọn của trường.
47. Collection 	Nút Collection (sưu tập) cho thấy sự phân bố giá trị của trường số mà liên quan đến các giá trị khác. (Nó tạo ra các đồ thị tương tự như histograms) Đó là hữu ích cho việc minh họa một biến hoặc trường dữ liệu có giá trị thay đổi theo thời gian.. Sử dụng đồ họa 3-D, bạn cũng có thể bao gồm một trục biểu tượng hiển thị phân phối theo thể loại.
48. Multiplot 	Nút Multiplot tạo ra một bản đồ có hiển thị nhiều trường dữ liệu Y trên một trường dữ liệu X duy nhất. Multiplots rất hữu ích khi bạn muốn khám phá những biến động của các biến số theo thời gian.
49. Web	Nút Web : minh họa sự phát triển mối quan hệ giữa các giá trị của hai hoặc nhiều biểu tượng . Biểu đồ này sử dụng đường nối có độ rộng khác nhau để chỉ sức mạnh kết nối. Bạn có thể sử dụng một nút Web ví dụ, để khám phá mối quan hệ giữa việc mua một

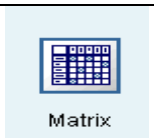
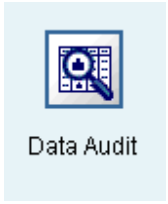
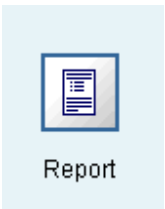
BIỂU TƯỢNG	MÔ TẢ
	nhóm các hạng mục tại một trang web thương mại điện tử.
50. Time Plot 	Nút Time Plot hiển thị một hoặc nhiều tập hợp dữ liệu chuỗi thời gian. Thông thường, trước tiên bạn sẽ sử dụng một khoảng thời gian để tạo ra một TimeLabel(bảng thời gian).
51. Evaluation 	Nút Evaluation : đánh giá và so sánh các mô hình được dự báo. Biểu đồ cho thấy đánh giá như thế nào dự đoán kết quả các mô hình cụ thể. Nó loại các bản ghi dựa trên giá trị dự đoán và độ tin cậy của các dự đoán. Nó chia tách các bản ghi thành các nhóm có kích thước bằng nhau (quantiles) và sau đó phát họa giá trị của các tiêu chí kinh doanh cho mỗi quantile từ cao nhất đến thấp nhất.
52. Binary Classifier 	Nút Binary Classifier : tạo ra và so sánh một số mô hình khác nhau cho kết quả nhị phân (có hoặc không có), cho phép bạn lựa chọn cách tiếp cận tốt nhất cho một phân tích đã cho. Một số thuật toán mô hình được hỗ trợ, làm cho nó có thể chọn các phương pháp bạn muốn sử dụng,và các tiêu chuẩn để so sánh các kết quả. Các nút tạo ra một tập các mô hình dựa trên các tùy chọn chỉ định và xếp hạng các ứng viên tốt nhất theo tiêu chuẩn mà bạn chỉ định.
53. Numeric Predictor 	Nút Numeric : dự đoán số ước lượng và so sánh các mô hình cho kết quả nhiều số liên tục bằng cách sử dụng một số phương pháp khác nhau. nút này hoạt động theo cách thức tương tự như nút Binary Classifier, cho phép bạn chọn các thuật toán để sử dụng và thử nghiệm để kết hợp với nhiều tùy chọn trong một mô hình duy nhất. các thuật toán hỗ trợ bao gồm mạng lưới thần kinh, C & R Tree, CHAID, hồi quy tuyến tính, hồi quy tuyến tính tổng quát, và máy vector hỗ trợ (SVM). Mô hình có thể được so sánh dựa trên sự tương quan lỗi, hoặc số của các biến được sử dụng.
54. Time	Nút Time Series ước lượng số mũ làm mịn, tích hợp Đường trung bình (Arima), và đa biến Arima (hoặc chuyển giao chức năng) các

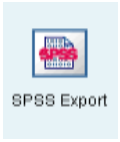
BIỂU TƯỢNG	MÔ TẢ
Series 	mô hình dữ liệu chuỗi thời gian và tạo ra các dự báo về hiệu quả trong tương lai. Một nút Time Series luôn luôn phải được đi trước bởi một nút khoảng thời gian.
55. QUEST 	Nút QUEST cung cấp một phương pháp phân loại nhị phân để xây dựng cây quyết định, được thiết kế để giảm thời gian xử lý cần thiết cho C & R trong khi cũng làm giảm xu hướng tìm thấy trong các phương pháp phân loại cây để ưu tiên cho phép nhiều hơn dự đoán chia tách.
56. CHAID 	Nút CHAID tạo ra cây quyết định sử dụng số liệu thống kê để xác định tối ưu chia tách. Không giống như C & R Tree và QUEST các nút CHAID có thể tạo ra cây nonbinary, có nghĩa là một số chia tách có nhiều hơn hai chi nhánh. Mục tiêu và các trường dữ liệu dự đoán có thể được nhiều .
57. Decision List 	Nút Decision List xác định các nhóm con, hoặc các phân đoạn, cho thấy một khả năng cao hơn hoặc thấp của một kết quả nhị phân tương đối so với tổng dân số. Bạn có thể kết hợp kiến thức kinh doanh của bạn thành mô hình bằng cách thêm các phân đoạn tuỳ chỉnh của riêng bạn và xem trước các mô hình thay thế cạnh nhau để so sánh kết quả. bao gồm một danh sách các quy tắc, trong đó quy định mỗi một điều kiện và một kết quả .
58. Regression 	Hồi quy tuyến tính là một kỹ thuật phổ biến cho các thống kê dữ liệu và làm cho dự đoán chính xác hơn hoặc làm giảm thiểu các sai lệch giữa dự đoán và số lượng các giá tri thực tế.
59. Factor /	Các Factor / PCA cung cấp dữ liệu giảm mạnh các kỹ thuật để giảm bớt sự phức tạp của dữ liệu . phân tích các thành phần chính

BIỂU TƯỢNG	MÔ TẢ
PCA 	(PCA) tìm thấy sự kết hợp tuyến tính của trường nhập dữ liệu .Yếu tố phân tích cố gắng xác định các yếu tố cơ bản để giải thích các mô hình của tương quan trong một bộ các lĩnh vực quan sát. Đối với cả hai phương pháp tiếp cận, mục tiêu là tìm thấy một số lượng nhỏ các lĩnh vực mà hiệu quả thu được tóm tắt các thông tin trong các thiết lập ban đầu của lĩnh vực
60. Feature Selection 	Nút Feature Selection : loại bỏ các trường dữ liệu dự báo dựa trên một bộ tiêu chuẩn (chẳng hạn như tỷ lệ phần trăm của giá trị mất tích); sau đó nó xếp tầm quan trọng của yếu tố dự đoán còn lại liên quan đến một mục tiêu . Ví dụ, với một tập dữ liệu với hàng trăm dự đoán tiềm năng, trong đó có nhiều khả năng có ích trong các kết quả mô hình bệnh nhân.
61. Discriminant 	Nút Discriminant làm cho các giả định chặt chẽ hơn hồi quy, nhưng có thể thay thế một giá trị, bổ sung một phân tích hồi quy khi đáp ứng được những giả định đặt ra.
62. Logistic 	Logistic là một kỹ thuật thống kê để phân loại hồ sơ dựa trên giá trị của trường dữ liệu đầu vào. Nó tương tự như hồi quy tuyến tính nhưng có một kết quả cụ thể rõ ràng thay vì một dải số.
63. Genlin 	Mô hình generalized linear (tuyến tính tổng quát) mở rộng mô hình tuyến tính tổng quát để các biến phụ thuộc là tuyến tính liên quan đến các yếu tố và thông qua một chức năng liên kết nhất định. Hơn nữa, mô hình này cho phép các biến phụ thuộc để có một phân phối không bình thường. Nó bao gồm các chức năng của một số các mô hình thống kê, bao gồm cả hồi quy tuyến tính, thống kê logic cho dữ liệu .

BIỂU TƯỢNG	MÔ TẢ
64. Cox 	Nút Cox (hồi quy) cho phép bạn xây dựng một mô hình cho sự tồn tai thời gian . Mô hình này tạo ra một chức năng để dự đoán xác suất các sự kiện quan tâm đã xảy ra tại một thời gian nhất định (t) cho các giá trị của các biến có yếu tố dự báo.
65. SVM 	Nút SVM (Support Vector Machine) cho phép bạn phân loại dữ liệu . SVM hoạt động tốt với các tập dữ liệu lớn.
66. Bayes Net 	Các nút mạng Bayes cho phép bạn xây dựng một mô hình xác suất bằng cách kết hợp quan sát và ghi lại dữ liệu với kiến thức thực tế để thiết lập khả năng xảy ra.
67. SLRM 	Nút SLRM :cho phép bạn xây dựng một mô hình trong đó có một trường hợp đơn lẻ, hoặc số lượng nhỏ các trường hợp mới.
68. GRI 	Nút phát hiện ra nguyên tắc liên kết trong dữ liệu. Ví dụ, khách hàng mua dao cạo râu và sữa tắm cũng có khả năng sẽ mua kem cạo râu. GRI có thể xử lý đầu vào rõ ràng, nhưng mục tiêu cũng phải được rõ ràng.
69. Apriori 	Nút Apriori lấy một bộ quy tắc từ dữ liệu, tạo ra các quy tắc với nội dung thông tin cao nhất. Apriori cung cấp năm phương pháp khác nhau của các quy tắc lựa chọn và sử dụng một chương trình để xử lý tập dữ liệu lớn hiệu quả. Đối với những dữ liệu lớn, Apriori thường phát triển nhanh hơn GRI, nó không giới hạn vào số lượng các quy tắc có thể được giữ lại, và nó có thể xử lý các quy tắc lên đến 32 điều kiện . Apriori yêu cầu đầu vào và đầu ra tất cả các trường dữ liệu phải rõ ràng, nhưng mang lại hiệu quả tốt hơn bởi vì nó được tối ưu hóa cho các loại dữ liệu.

BIỂU TƯỢNG	MÔ TẢ
70. CARMA 	Mô hình CARMA chép một bộ quy tắc từ các dữ liệu mà không yêu cầu bạn chỉ định.Trái ngược với Apriori và GRI, nút CARMA cung cấp các thiết lập để hỗ trợ xây dựng các quy tắc. Điều này có nghĩa rằng các quy tắc được tạo ra có thể được sử dụng đa dạng hơn các ứng dụng, ví dụ, để tìm một danh mục sản phẩm, dịch vụ (tiền đề) mà kết quả là các mục mà bạn muốn quảng cáo mùa lễ này.
71. Sequence 	Các nút Sequence(trình tự) phát hiện ra nguyên tắc liên kết trong dữ liệu tuần tự hoặc theo định hướng thời gian. Sequence là một danh sách các mục mà có xu hướng xảy ra theo một thứ tự dự đoán được. Các nút Sequence dựa trên thuật toán CARMA về nguyên tắc liên kết, trong đó sử dụng một phương pháp two-pass hiệu quả cho việc tìm kiếm các chuỗi.
72. TwoStep 	Các nút TwoStep sử dụng một phương pháp two-step clustering . Bước đầu tiên tạo một pass duy nhất thông qua các dữ liệu để nén dữ liệu đầu vào của một bộ quản lý của subclusters. Bước thứ hai sử dụng một phương pháp phân nhóm theo thứ bậc để dần dần hợp nhất subclusters thành các cụm lớn hơn và lớn hơn. TwoStep có lợi thế là tự động ước tính số lượng tối ưu các cụm cho các dữ liệu . Nó có thể xử lý các loại dữ liệu lớn một cách hiệu quả.
73. Anomaly 	Nút Anomaly phát hiện sự bất thường của các trường hợp bất thường, mà không phù hợp với mô hình dữ liệu "normal".
74. Custom Table 	Nút Custom Table hỗ trợ một loạt các lựa chọn, bao gồm cả khả năng làm chồng dữ liệu, hoặc biến lớp, để hiển thị tóm lược cho các thống kê lớn, và để hiển thị nhiều tập hợp .
75. Matrix	Nút Matrix tạo ra một bảng cho thấy mối quan hệ giữa các trường dữ liệu. Nó thường được sử dụng để hiển thị các mối quan hệ giữa hai trường dữ liệu mang tính tượng trưng, nhưng nó cũng có thể hiển thị các mối quan hệ giữa các trường flag hoặc các trường

BIỂU TƯỢNG	MÔ TẢ
	number.
76. Analysis 	Nút Analysis (Phân tích) đánh giá các khả năng để tạo ra các mô hình chính xác. Phân tích các nút khác nhau thực hiện so sánh giữa giá trị dự báo và giá trị thực tế cho một hoặc nhiều mô hình. Chúng cũng có thể so sánh các mô hình dự đoán với nhau.
77. Data Audit 	Nút Data Audit (kiểm toán dữ liệu) cung cấp một cái nhìn toàn diện đầu tiên các dữ liệu, bao gồm cả số liệu thống kê tóm tắt, lược đồ histograms và phân phối cho từng trường dữ liêu, cũng như thông tin về đầu ra, thiếu giá trị. Kết quả được hiển thị trong một ma trận có thể được sắp xếp và sử dụng để tạo kích thước đồ thị và các nút chuẩn bị dữ liệu.
78. Transform 	Nút Transform(chuyển đổi): cho phép bạn chọn và trực quan xem trước kết quả của biến đổi trước khi áp dụng chúng vào các trường dữ liệu được lựa chọn.
79. Statistics 	Nút Statistics (Thống kê) cung cấp thông tin tóm tắt cơ bản về trường dữ liệu số. Nó tính toán số liệu thống kê tóm tắt cho các trường dữ liệu riêng biệt và tương quan giữa các trường dữ liệu.
80. Means 	Ví dụ, bạn có thể so sánh: doanh thu trước và sau khi thực hiện chương trình khuyến mãi hoặc so sánh các khoản thu từ khách hàng không nhận được khuyến mãi với những người đã làm.
81. Report 	Nút Report(Báo cáo) : tạo ra các báo cáo có chứa các định dạng văn bản cố định cũng như dữ liệu và biểu thức khác xuất phát từ dữ liệu. Bạn chỉ rõ định dạng của báo cáo bằng cách sử dụng văn bản mẫu để xác định văn bản và số lượng công việc. Bạn có thể cung cấp văn bản bằng cách sử dụng các thẻ HTML trong bản mẫu và bằng cách thiết lập các tùy chọn trên tab đầu ra. Chúng có thể bao gồm các giá trị dữ liệu và đầu ra có điều kiện khác bằng

BIỂU TƯỢNG	MÔ TẢ
	cách sử dụng Clem .
82. Set Globals 	Nút Set Globals quét dữ liệu và tính giá trị, được sử dụng trong Clem. Ví dụ, bạn có thể sử dụng nút này để tính toán thống kê độ tuổi và sau đó sử dụng tổng số tuổi trong Clem bằng cách chèn chức năng @ GLOBAL_MEAN (tuổi).
83. SPSS Output 	Nút Output SPSS cho phép gọi một thủ tục SPSS để phân tích dữ liệu Clementine . Một loạt các thủ tục có sẵn để phân tích SPSS .
84. SPSS Export 	Nút SPSS Export (xuất dữ liệu đầu ra trong SPSS). Định dạng SAV. Các. File SAV có thể được đọc bởi SPSS và các sản phẩm khác. Đây cũng là định dạng được sử dụng cho các tập tin bộ nhớ cache trong Clementine
85. Dimensions 	Nút Dimensions (Kích thước) xuất dữ liệu đầu ra .
86. SAS Export 	Nút SAS Export xuất dữ liệu đầu ra dưới dạng SAS. SAS định dạng tập tin có sẵn: SAS cho Windows/OS2, SAS cho UNIX, hoặc SAS Phiên bản 7 / 8.
87. Publisher	Nút Publisher là một "phiên bản" đóng gói của một stream có thể được thực hiện bởi một engine . Thời gian chạy bên ngoài hoặc

BIỂU TƯỢNG	MÔ TẢ
	nhúng trong một ứng dụng bên ngoài để sử dụng trong một môi trường sản xuất.