

1. Dữ liệu  
   + Dữ liệu phân lớp: <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>

* Dữ liệu phân cụm: <https://www.kaggle.com/datasets/yasserh/wine-quality-dataset>

1. Slide:<https://www.canva.com/design/DAGBqX1dvCU/_G_769d5PXN7He6AaITQmQ/edit?utm_content=DAGBqX1dvCU&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton>

Nội dung slide bao gồm:

* Giới thiệu dữ liệu
* Trình bày các thuật toán

1. Code, file run algorithm

—----------------------------------—-Bài làm sẽ thực hiện các file—-----------------------------------------

**ID3 - Tuyết + code (nếu có)**

**Model evaluation - Toàn**

**KNN - Huy**

**K-Mean - Huệ + code (nếu có)**

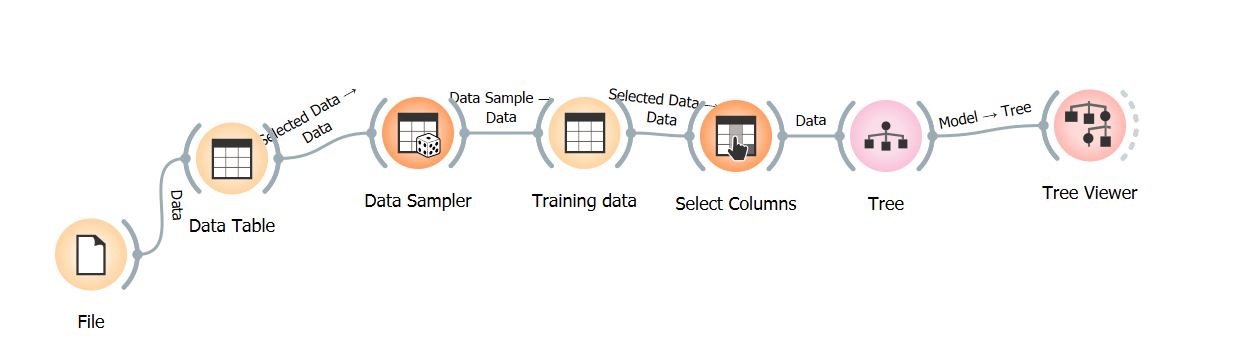
**Hierarchical - Hằng + code (Nếu có)**

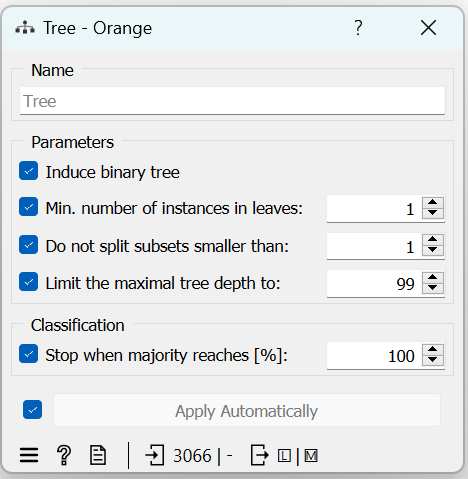
**Association Rule - Toàn**

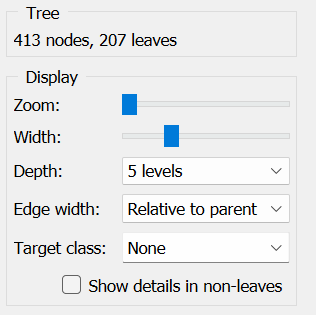
**ID 3:**

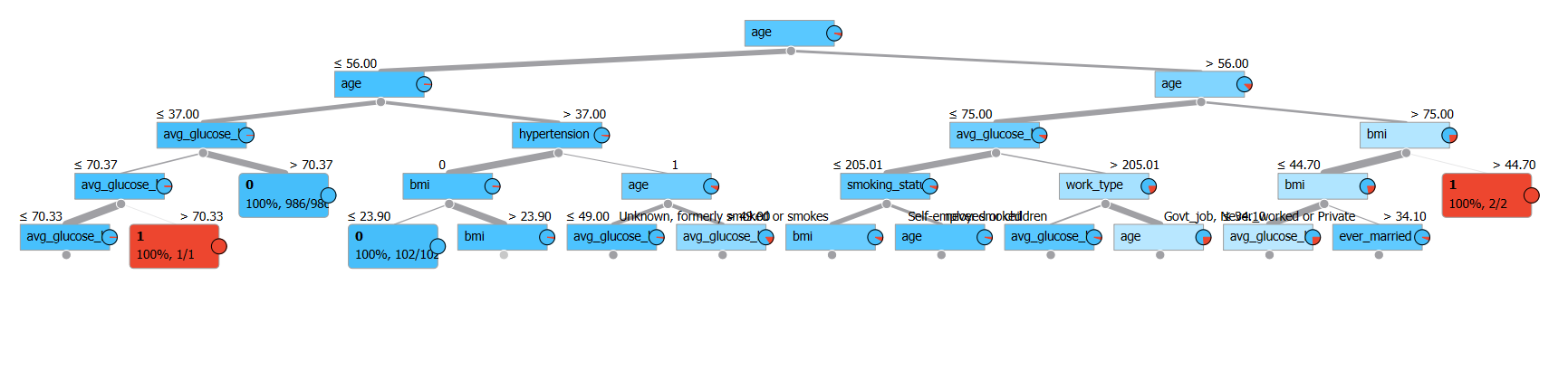
****

****

****

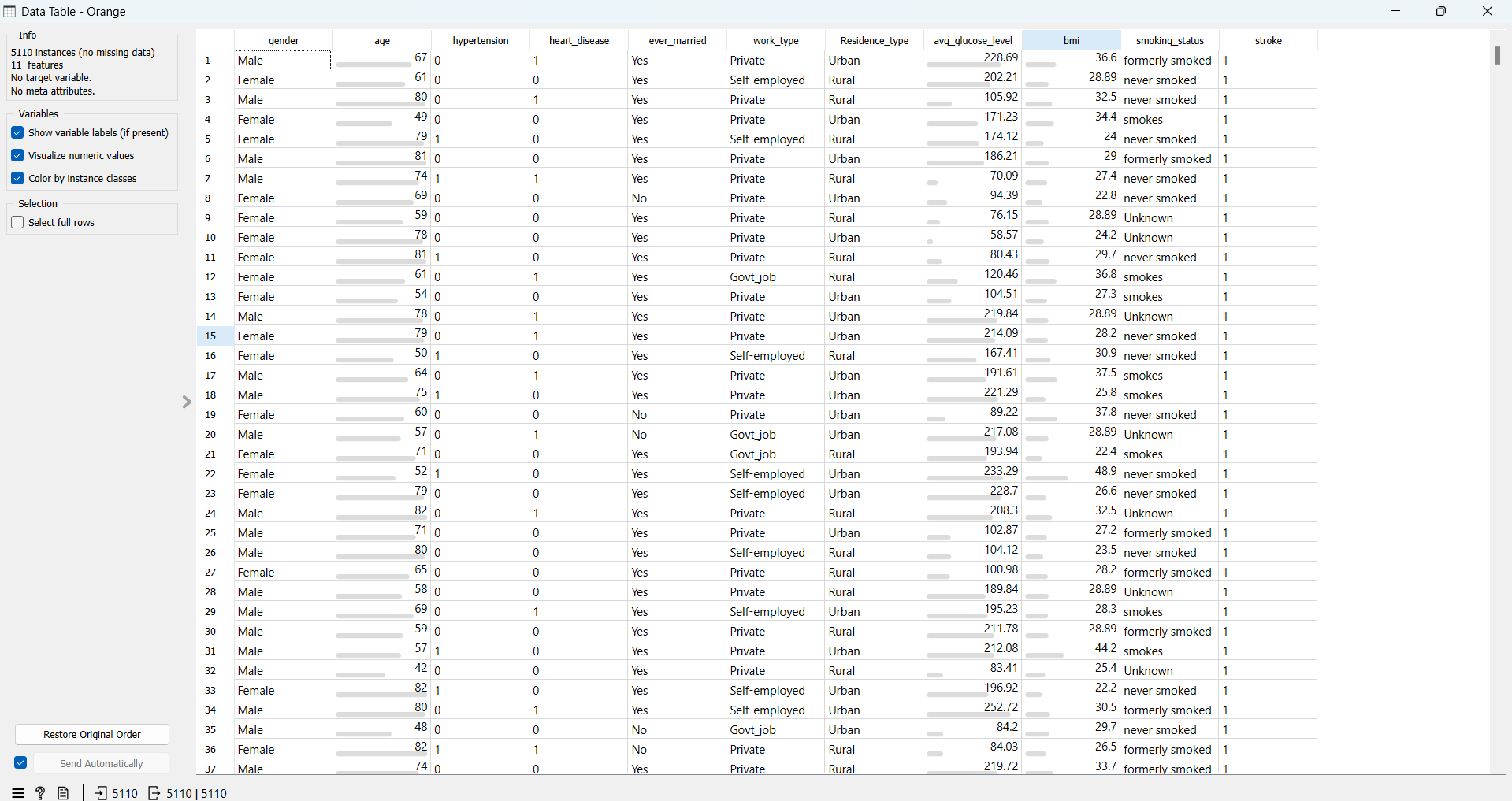
****

****

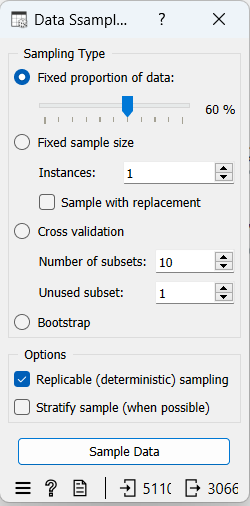
****

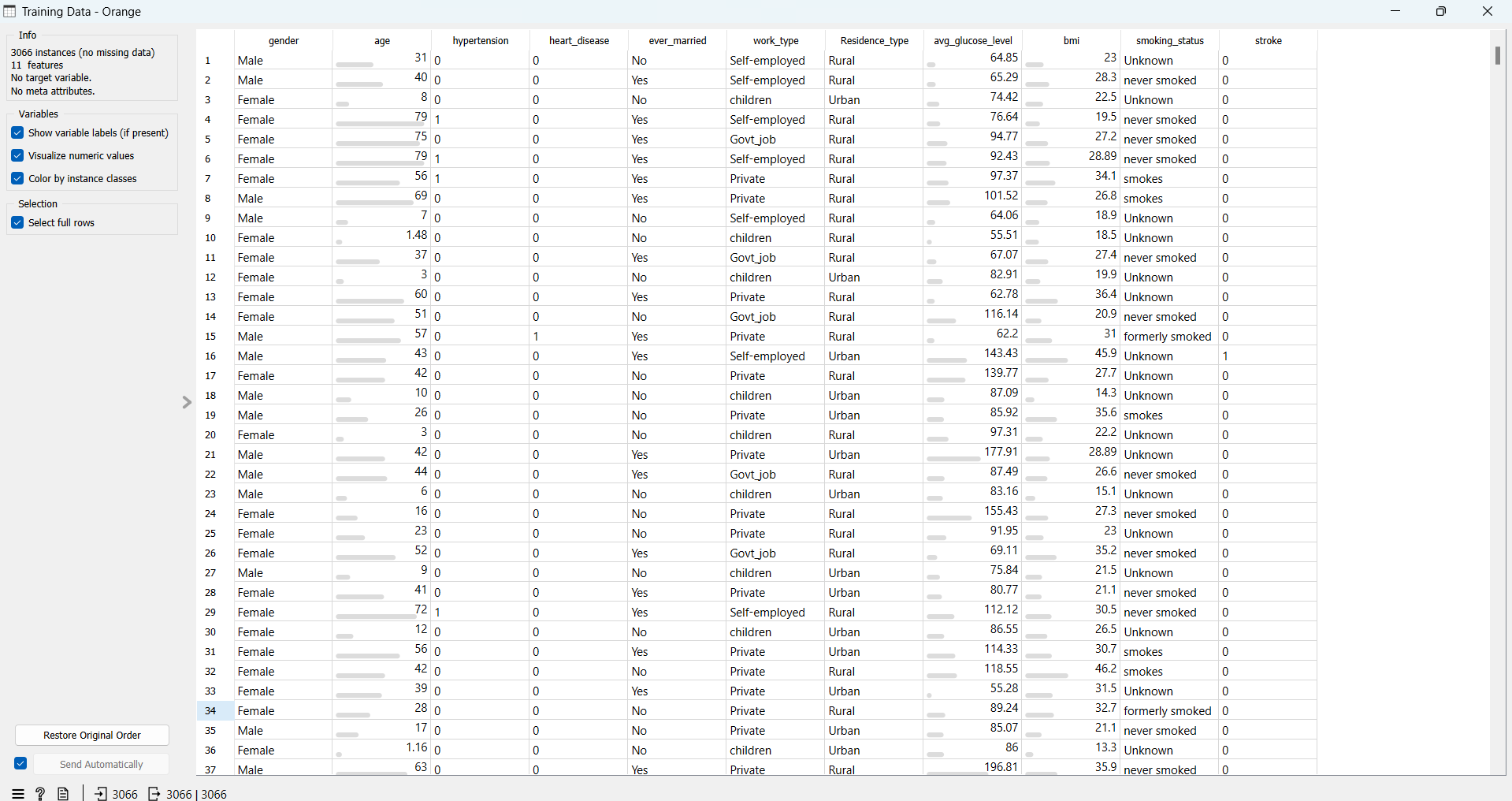


* Dataset:



* Training Data:



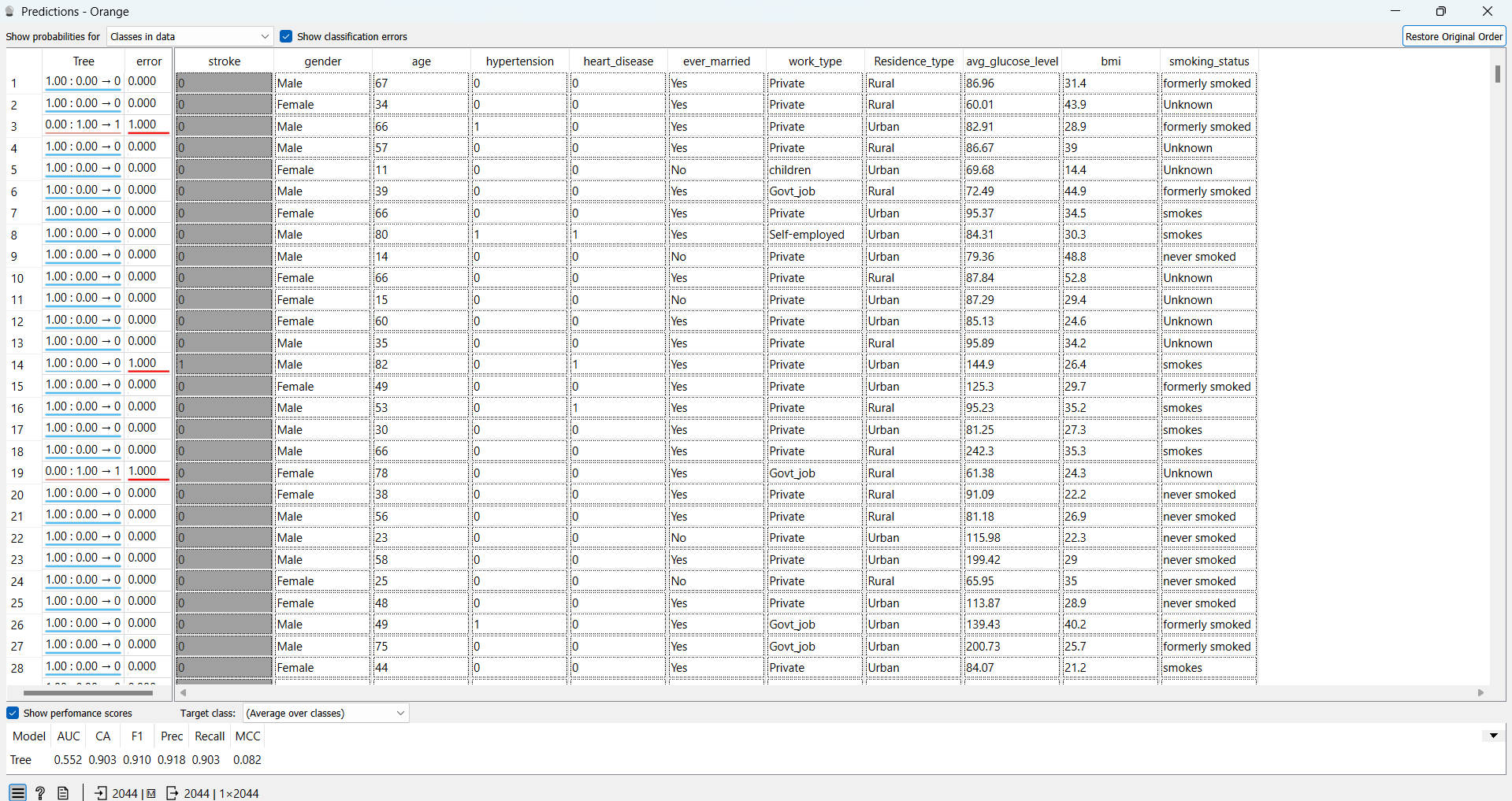


* Testing data

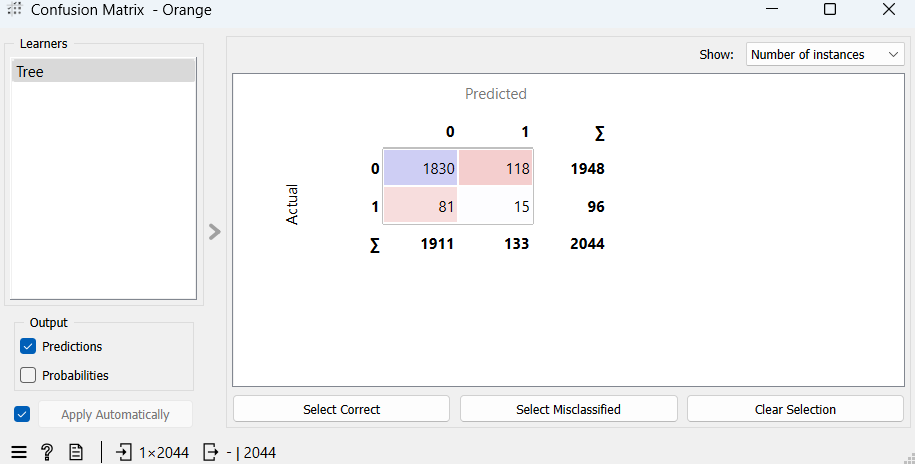
Dữ liệu còn lại trong dataset



* Các chỉ số:

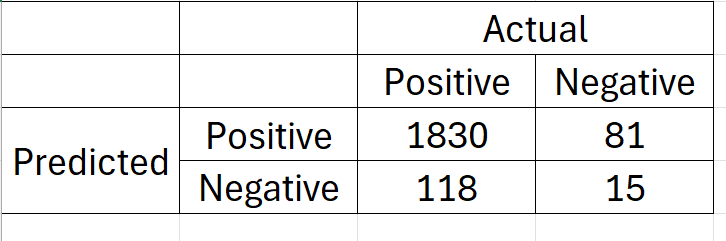


* Confusion matrix



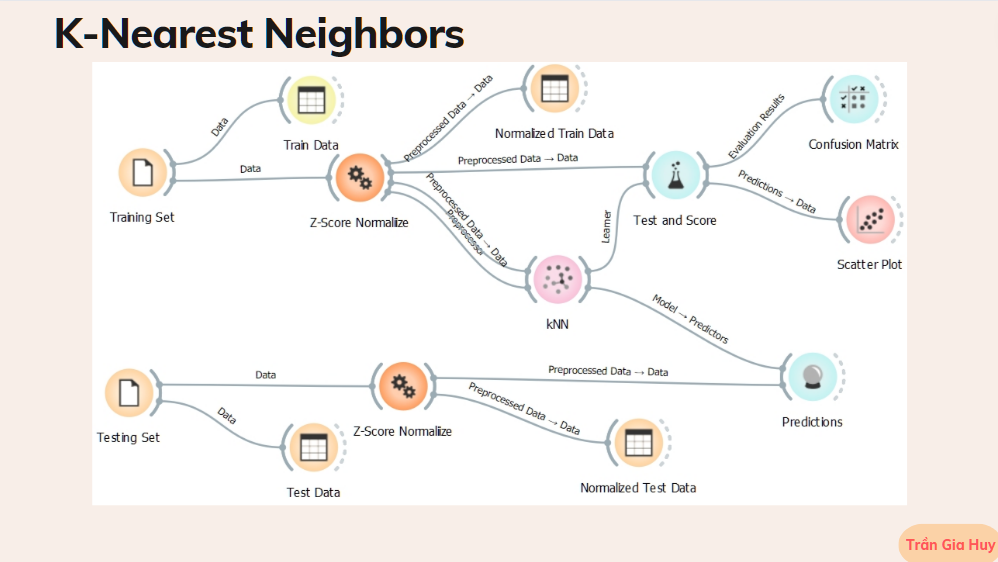
* Thực tế: có 1948 người ko bị bệnh, có 96 người bị đột quỵ
* Dự đoán: Có 1911 người ko bị bệnh, có 133 người bị đột quỵ.

Có thể dùng cái này:

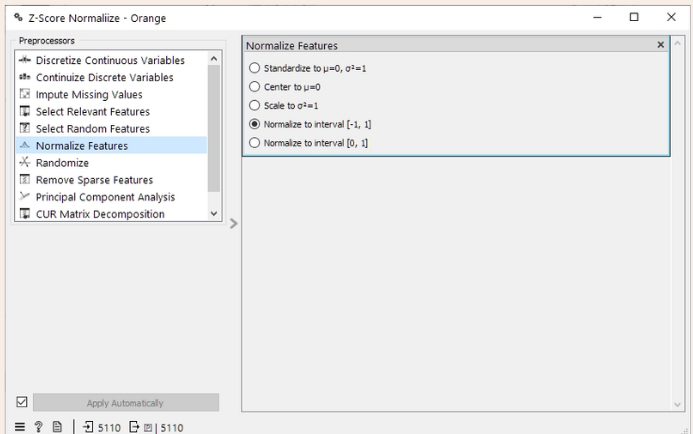


**KNN - Huy**

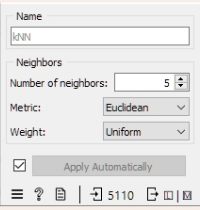
**Orange**

****

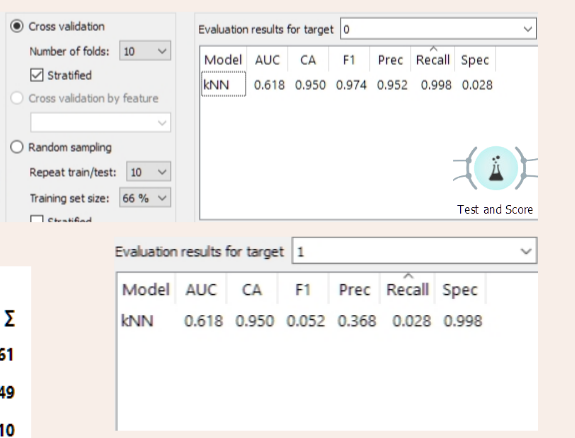
* Training set: Tệp dữ liệu huấn luyện
* Testing set: Tệp dữ liệu kiểm thử
* Preprocessing (Z-score Normalize): Chuẩn hoá dữ
* liệu để đưa về cùng đơn vị đo
* Knn: thuật toán
* Test and score: Tính và đánh giá các độ đo của thuật toán
* Confusion Matrix: Biểu diễn phân loại theo confusion matrix
* Scatter Plot: Biểu diễn giá trị dữ liệu target trên biểu đồ phân tán
* Prediction: Dự đoán cho dữ liệu target ở tệp dữ liệu kiểm thử

****

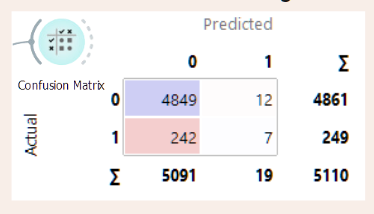
* Normalize to interval [-1, 1]: Chuẩn hoá Z-Score
* Normalize to interval [0, 1]: Chuẩn hoá Min-Max



Trong thuật toán, chọn K = 5 (Vì qua nhiều lần chạy thử với các giá trị 3,5,7,9,11 thì k = 5 cho ra các độ đo AUC, F1 và Accuracy tương đối ổn hơn so với k = các giá trị khác).

****

* Fold: Chia dữ liệu train ra 10 gói để đánh giá chéo
* Stratified: trộn dữ liệu trước khi chia fold và chia sao cho tỷ lệ phân bố các lớp trong các fold là cân xứng và tương đối với nhau.
* AUC (Area Under Curve) = 0.618 => Mức ý nghĩa của model ở mức trung bình hay Model không mang ý nghĩa quá lớn.
* Các giá trị F1, Precision, Recall ở target 0 cao còn ở target 1 thì thấp => Model thiên về phân loại đối tượng không/ít có khả năng bị đột quỵ

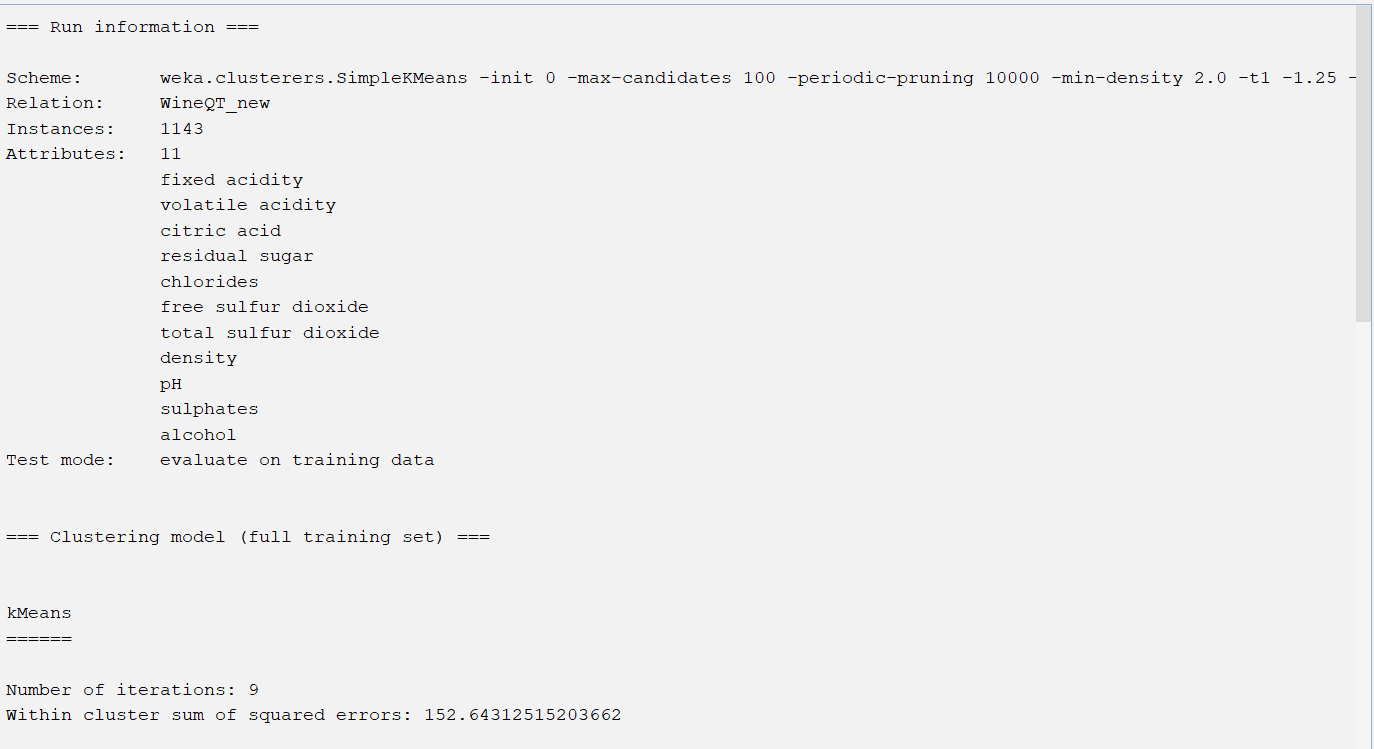


* Các số in đậm ko phải là kết quả tính độ đo mà là tổng số các đối tượng tính theo từng cột/hàng.

Từ matrix, ta thấy FN = 242 quá lớn với TP = 7 mà => Model ko có nhiều ý nghĩa khi chuẩn đoán đối tượng có khả năng bị đột quỵ.

**K-Mean - Huệ**

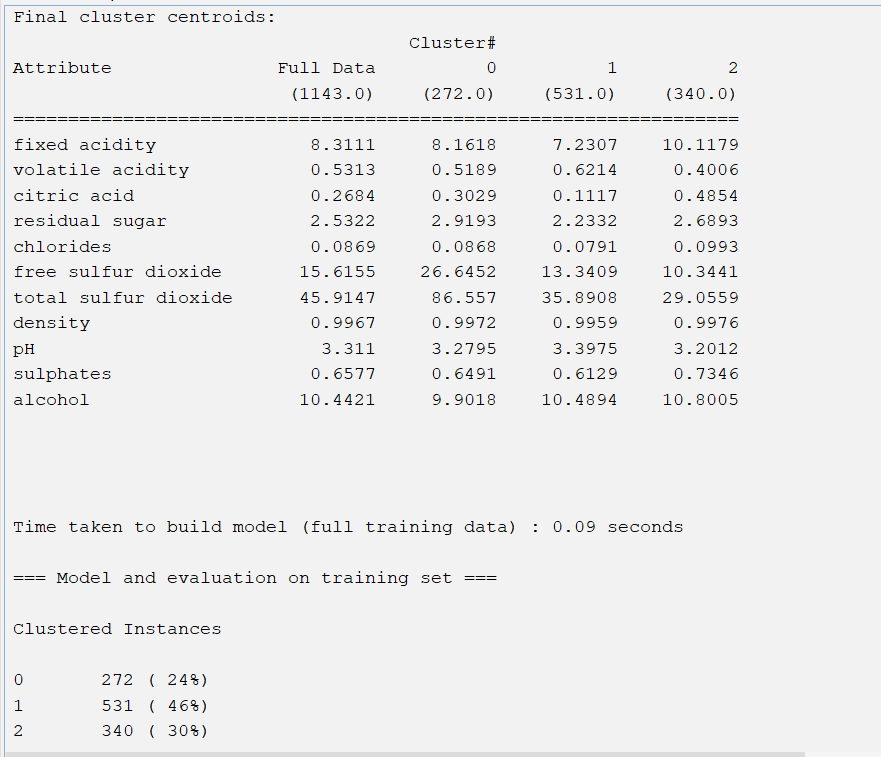
Weka

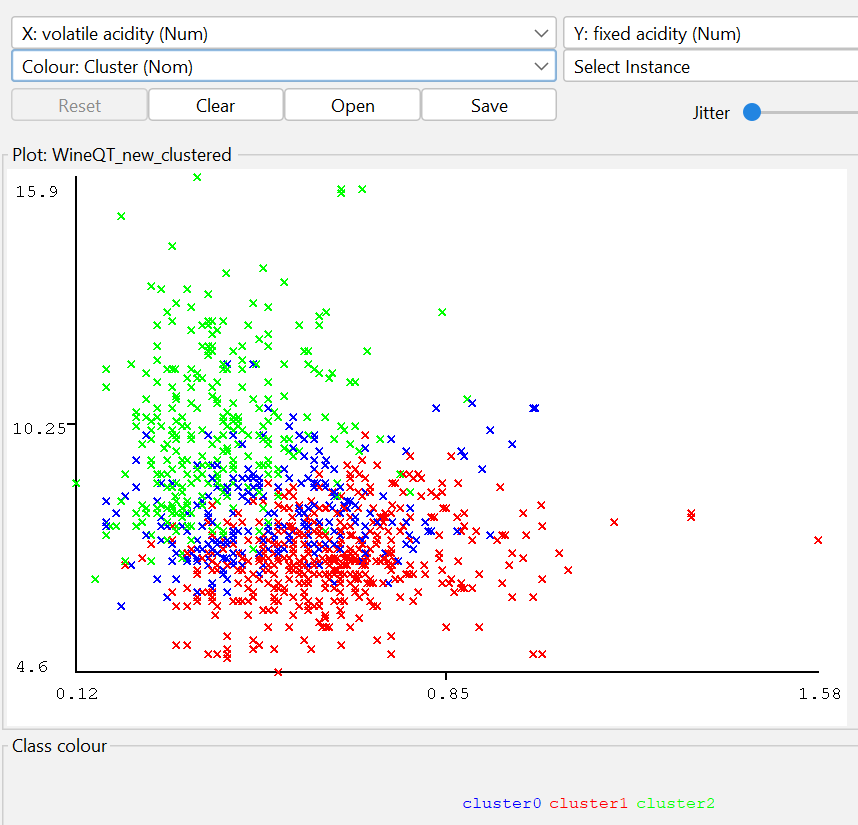




number of iterations: số vòng lặp lại để tính cụm center

Within cluster sum of squared errors: (WSS) là một phép đo độ biến động của các điểm dữ liệu bên trong các cụm (cluster) trong một thuật toán gom cụm như K-means. Độ phân tán giữa các điểm dữ liệu





**Code:**

**Tính khoảng cách bằng pp Elbow để chọn k cụm**

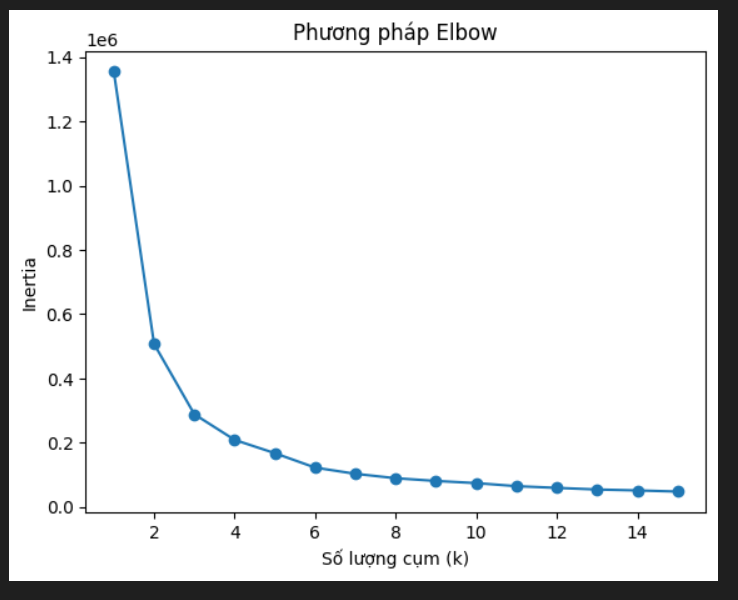
****

**kmeans=Kmeans(n\_cluster=k) ⇒ tạo các đối tượng với số lượng cụm mong muốn là k**

**kmeans.fit(X) ⇒ huấn luyện mô hình Kmeans trên dữ liệu X (xác định các tâm cụm và gán điểm dữ liệu vào cụm gần nhất)**

**kmeans.inertia\_ ⇒inertia (trọng số) là tổng bình phương khoảng cách từ mỗi điểm dữ liệu đến tâm cụm gần nhất của nó. Trong ngữ cảnh của K-means, inertia là một phép đo của sự "tập trung" của các điểm dữ liệu trong các cụm.**

**Cụm với inertia thấp hơn có xu hướng có các điểm gần nhau hơn về mặt không gian, do đó, làm cho cụm trở nên "tập trung" hơn.**

****

**Hiển thị kết quả**



#plt.line2D mô tả màu sắc và chú thích của biểu đồ thông quan hàm plt.legend

#handles=[]: Tham số này là một danh sách các đối tượng mà chúng ta muốn thêm vào chú thích. Trong trường hợp này, danh sách này chỉ chứa một đối tượng.

#[0],[0]:> không cần thiết nên để 0

#plt.Line2D(xdata, ydata, linestyle='-', color=None, marker=None, markersize=None, label=None)

# Hiển thị kết quả phân cụm trên biểu đồ

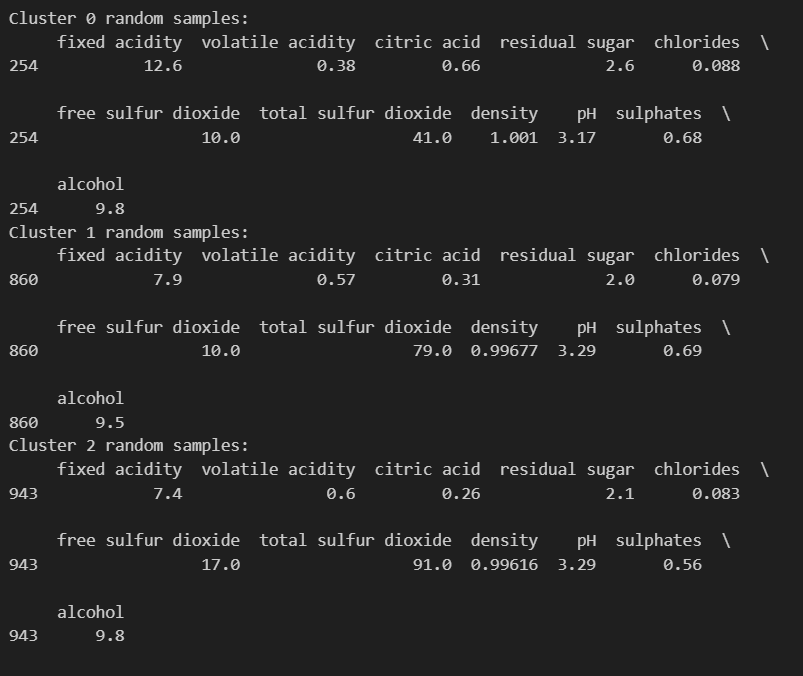
for i in range(3):: Vòng lặp này sẽ lặp qua các giá trị từ 0 đến 2 (tức là các cụm từ 0 đến 2)

# cluster\_data = data[labels == i]: Dòng này tạo ra một DataFrame mới cluster\_data từ DataFrame data, chỉ chứa các dòng mà nhãn của chúng tương ứng với giá trị i. Điều này có nghĩa là nó chỉ lấy các điểm dữ liệu thuộc về cụm i.

# random\_samples = cluster\_data.sample(n=1, random\_state=42): Dòng này chọn một mẫu ngẫu nhiên từ cluster\_data. Trong trường hợp này, chỉ có một mẫu được chọn (n=1). Tham số random\_state=42 đảm bảo rằng việc chọn mẫu ngẫu nhiên này sẽ tạo ra kết quả nhất quán trong mỗi lần chạy chương trình.

# print(f"Cluster {i} random samples:"): Dòng này in ra tiêu đề cho mỗi cụm, cho biết rằng các số liệu được in sau đó là mẫu ngẫu nhiên từ cụm đó. Biến i được sử dụng để hiển thị số cụm tương ứng.

****

****

**5. Hierarchical Clustering - Phân cụm thứ bậc (Hằng)**

**5.1 Giải thích dữ liệu**

Dữ liệu này chứa các thông tin về các loại rượu, với mỗi loại được mô tả bằng các biến đầu vào dựa trên các kiểm tra vật lý-hóa học và một biến đầu ra dựa trên dữ liệu cảm quan về chất lượng.

Các biến đầu vào

1. Fixed Acidity : Nồng độ axit cố định trong rượu (g/L).

2. Volatile Acidity : Nồng độ axit bay hơi trong rượu (g/L).

3. Citric Acid : Nồng độ axit citric trong rượu (g/L).

4. Residual Sugar : Lượng đường còn lại sau khi quá trình lên men hoàn tất (g/L).

5. Chlorides : Lượng muối trong rượu (g/L).

6. Free Sulfur Dioxide : Lượng lưu huỳnh dạng tự do trong rượu (mg/L).

7. Total Sulfur Dioxide : Tổng lượng lưu huỳnh trong rượu (mg/L).

8. Density : Tỷ trọng của rượu (g/cm^3).

9. pH : Độ axit hoặc kiềm của rượu.

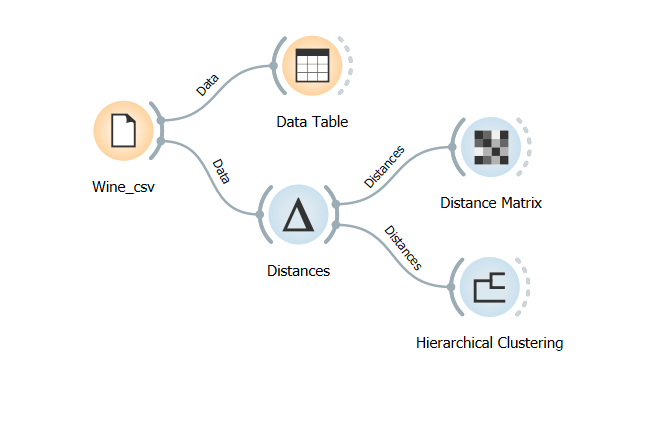
10. Sulphates : Nồng độ các hợp chất sulfat trong rượu (g/L).

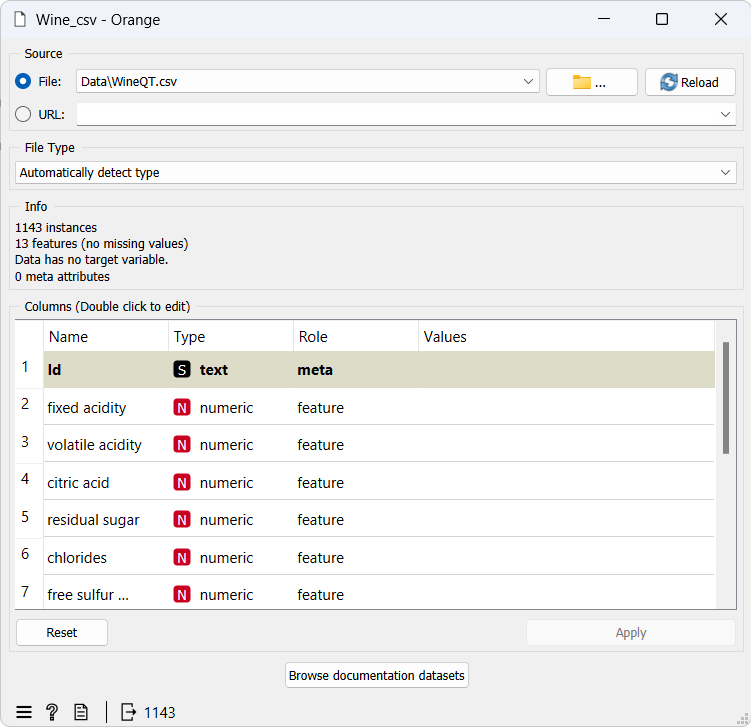
11. Alcohol : Phần trăm cồn trong rượu (% vol).

Biến đầu ra

12. Quality : Điểm chất lượng của rượu, được đánh giá dựa trên các yếu tố cảm quan, thường nằm trong khoảng từ 0 đến 10, với 10 đại diện cho chất lượng tốt nhất.

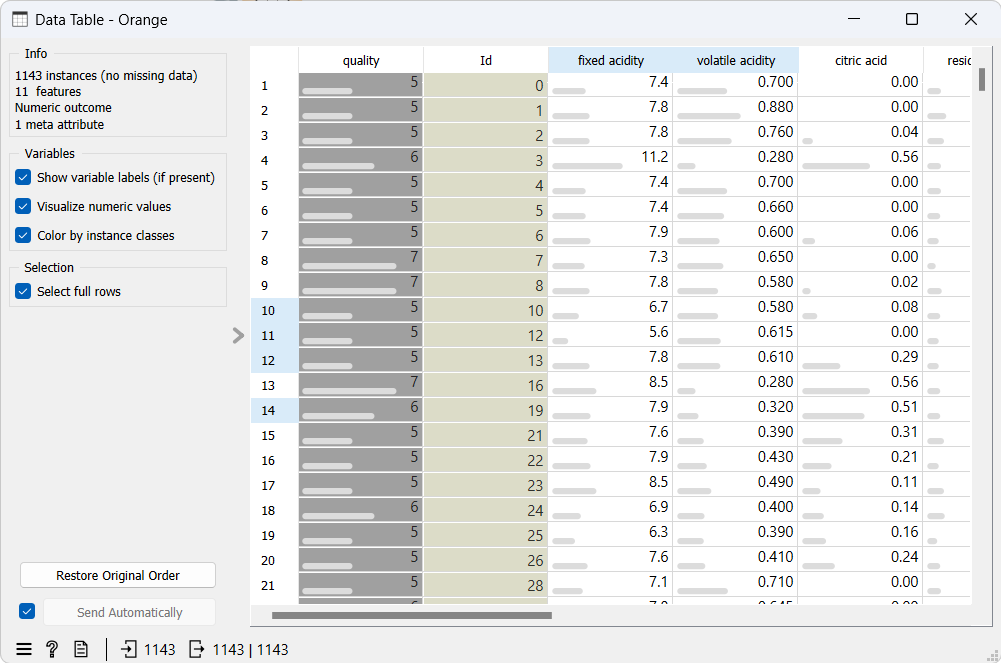
**5.2 Thực hiện trên Orange**

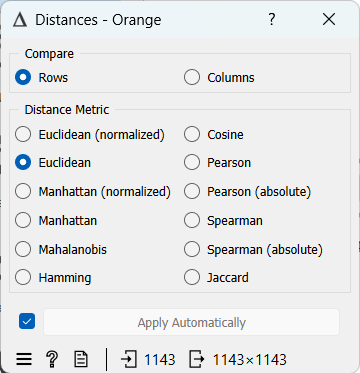
****

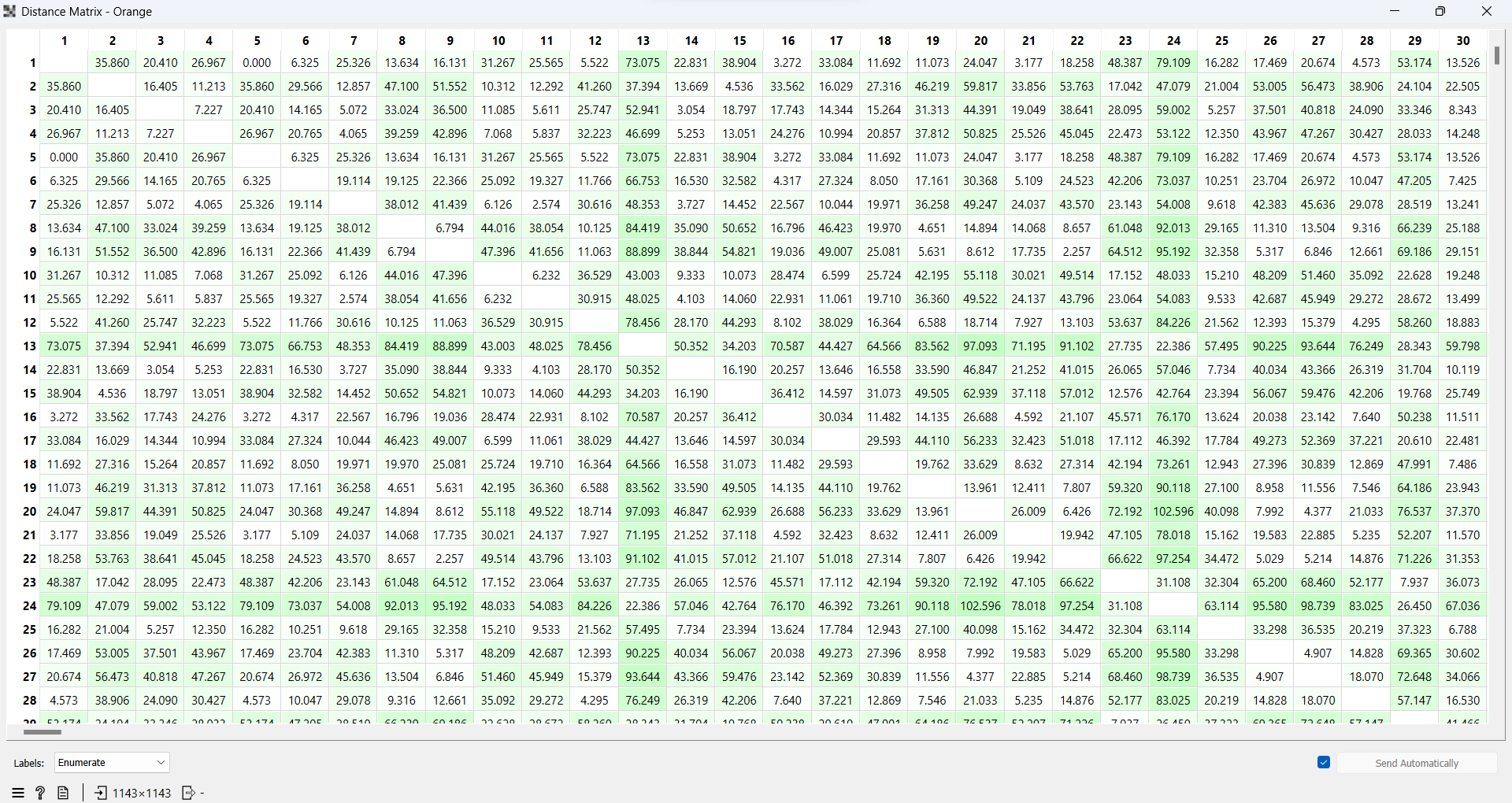
****

* **Thay đổi dữ liệu**
* **Id**: Type = **numeric → text** và chọn Role **= feature → meta** : biến cột **Id** thành tiêu đề
* **quality**: Role = **feature → target** : biến cột quality thành biến phụ thuộc
* 11 thuộc tính còn lại giữ nguyên: dùng đánh giá mức độ tương đồng để phân cụm rượu

⇒ Bảng ma trận khoảng cách sẽ tính theo 11 thuộc tính này ( không sử dụng 2 thuộc tính **Id** và **quality** vào để tính )

****

****

****

****

* **Giải thích các lựa chọn**

**Linkage:**

* Single: Đây là phương pháp liên kết đơn giản nhất, trong đó các cụm được hình thành bằng cách kết hợp hai cụm gần nhau nhất nhất.
* Complete: Đây là phương pháp liên kết phức tạp hơn, trong đó các cụm được hình thành bằng cách kết hợp hai cụm có khoảng cách trung bình ngắn nhất.
* Average: Phương pháp này tính trung bình khoảng cách giữa tất cả các điểm dữ liệu trong hai cụm trước khi kết hợp chúng.

**Annotation:** thêm các chú thích cho các thuộc tính

**Color by:**

* None: Lựa chọn này sẽ không tô màu các điểm dữ liệu.
* Attribute: Lựa chọn này sẽ tô màu các điểm dữ liệu theo một thuộc tính đã chọn.

**Pruning:**

* None: không muốn loại bỏ bất kỳ cụm nào khỏi kết quả.
* Max depth: Lựa chọn này cho phép bạn đặt giới hạn tối đa cho số lượng cấp bậc trong cây phân cấp.

**Selection:**

* Manual: Lựa chọn này cho phép bạn chọn thủ công các điểm dữ liệu mà bạn muốn bao gồm trong phân tích.
* Height ratio: phân tích theo khoảng cách
* Top: Lựa chọn này sẽ chọn một phần trăm nhất định các điểm dữ liệu có giá trị cao nhất trong thuộc tính đã chọn.

**Zoom:** phóng to/ thu nhỏ để xem dữ liệu

* **Phân tích kết quả:**
* Đặt vấn đề: Sử dụng thuật toán Hierarchical cho dữ liệu này nhằm mục đích gì?
* Giả sử ta có 1 căn phòng trống muốn trưng bày các loại rượu theo từng cụm có mức độ tương đồng với nhau. Nhưng chưa biết nên phân thành mấy cụm để bố trí vào căn phòng thì Hierarchical là phương pháp giúp giải quyết vấn đề này.
* Phân tích kết quả:
* Cụm1 và Cụm2: phân cụm rõ ràng và cố định
* Cụm 1: 2 loại rượu
* Cụm 2: 3 loại rượu
* Cụm 3 với nhiều loại rượu với mức độ tương đồng khác nhau, cần phân tích sâu hơn để phân cụm hợp lý bằng cách tăng độ sâu của cây

**6. Hồi quy tuyến tính đa biến (Hằng)**

**6.1 Thực hiện trên Orange**

* **Đặt vấn đề:** Để dự đoán chất lượng của rượu dựa vào các thuộc tính của nó
* **Hồi quy tuyến tính là gì?**

- Là phương pháp học máy có giám sát đơn giản, được sử dụng để dự đoán giá trị đầu ra, nó dựa trên thống kê để thiết lập mối quan hệ giữa một biến phụ thuộc và một nhóm tập hợp các biến độc lập.

* **Biến phụ thuộc**: biến mà giá trị được dự đoán hoặc ước lượng dựa trên các biến độc lập. (**quality)**
* **Biến độc lập:** gọi là biến giải thích, là các biến mà giá trị của chúng không phụ thuộc vào bất kỳ biến nào khác trong mô hình. (**11 thuộc tính còn lại)**
* Phương trình hồi quy tuyến tính có dạng



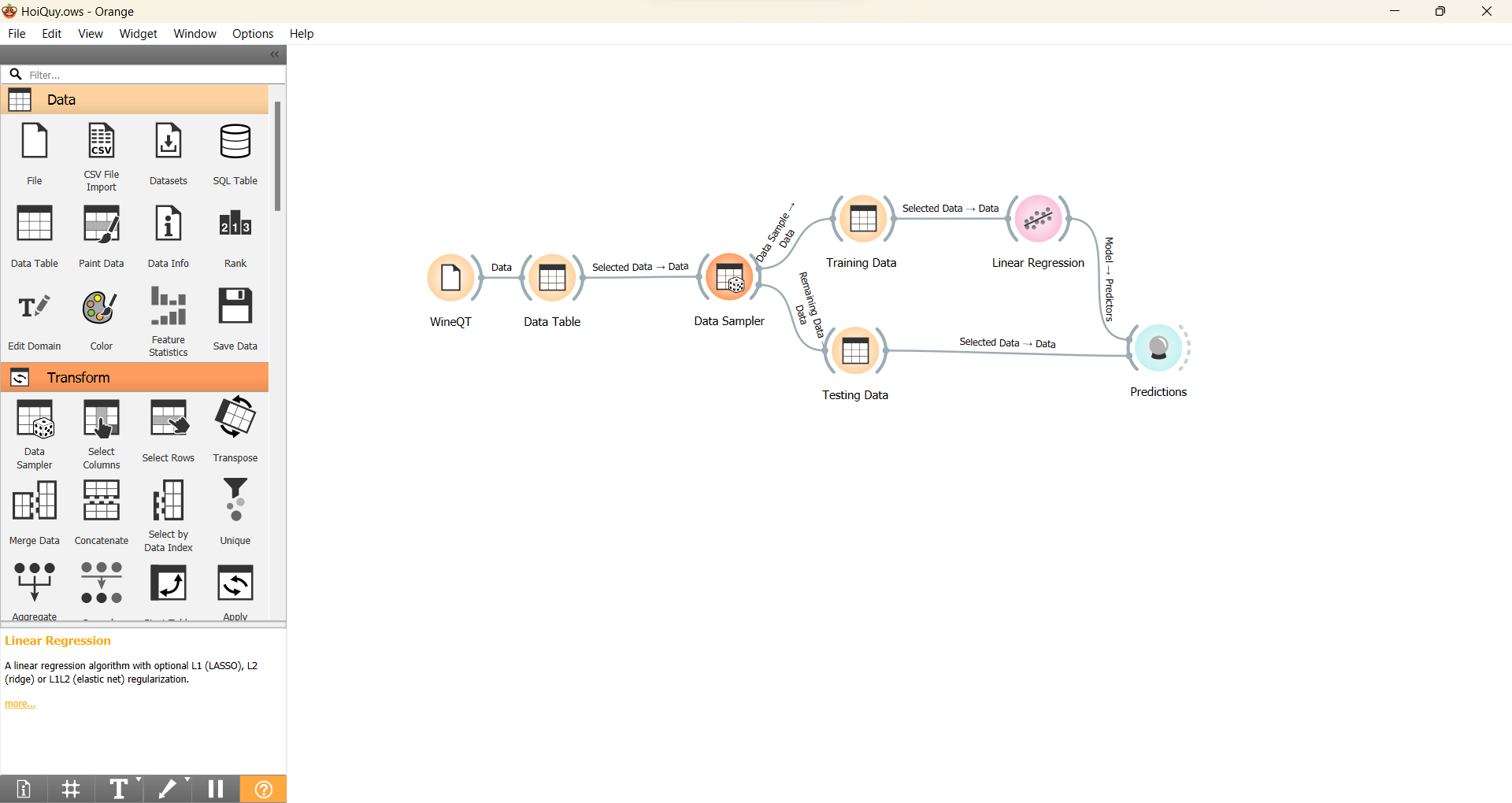
Trong đó:

Y: Biến phụ thuộc

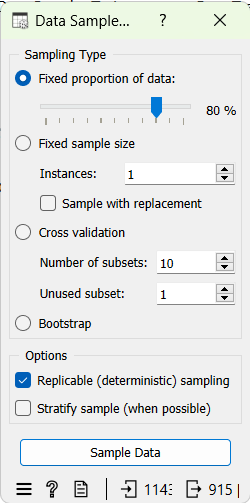
β0: Hệ số chặn

β1, β2, ..., βn: Hệ số hồi quy của các biến độc lập

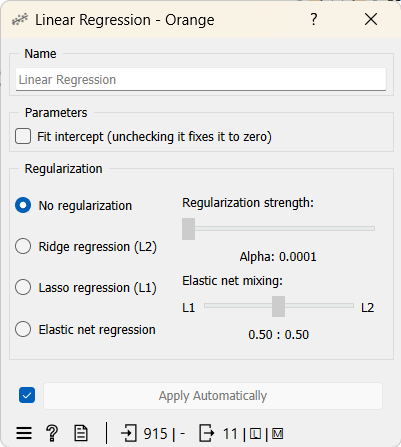
X1, X2, ..., Xn: Các biến độc lập

****

**Chia dữ liệu train (80%) và test (phần còn lại=20%)**

****

**Mô hình hồi quy**

****

* **Giải thích hình ảnh**
* **Fit intercept:** Chọn có lấy hệ số chặn vào dự đoán không. Thuật ngữ chặn là một hằng số được thêm vào tất cả các dự đoán.
* **Regularization:**
* **No regularization:** Tùy chọn này cho phép sử dụng chính quy hóa trong thuật toán hồi quy tuyến tính. Chính quy hóa là một kỹ thuật được sử dụng để ngăn chặn quá phù hợp, xảy ra khi thuật toán học máy học quá tốt trên dữ liệu đào tạo nhưng không thể tổng quát hóa tốt cho dữ liệu mới. ⇒ Chặn overfiting
* **Ridge regression (L2): l**à một loại chính quy hóa phổ biến được sử dụng để ngăn chặn overfiting . Nó hoạt động bằng cách thêm bình phương các trọng số mô hình vào hàm mất mát. Điều này có xu hướng đẩy các trọng số về phía 0, có thể giúp giảm phức tạp của mô hình và cải thiện khả năng tổng quát hóa của nó.
* **Lasso regression (L1):** là một loại chính quy hóa khác cũng có thể được sử dụng để ngăn chặn overfiting . Nó hoạt động bằng cách thêm giá trị tuyệt đối của các trọng số mô hình vào hàm mất mát. Điều này có xu hướng đẩy các trọng số về phía 0 hoặc hoàn toàn bằng 0, có thể giúp giảm phức tạp của mô hình và cải thiện khả năng tổng quát hóa của nó.
* **Regularization strength:** (dùng điều chỉnh khi chọn L2, L1)

**alpha: 0.0001**

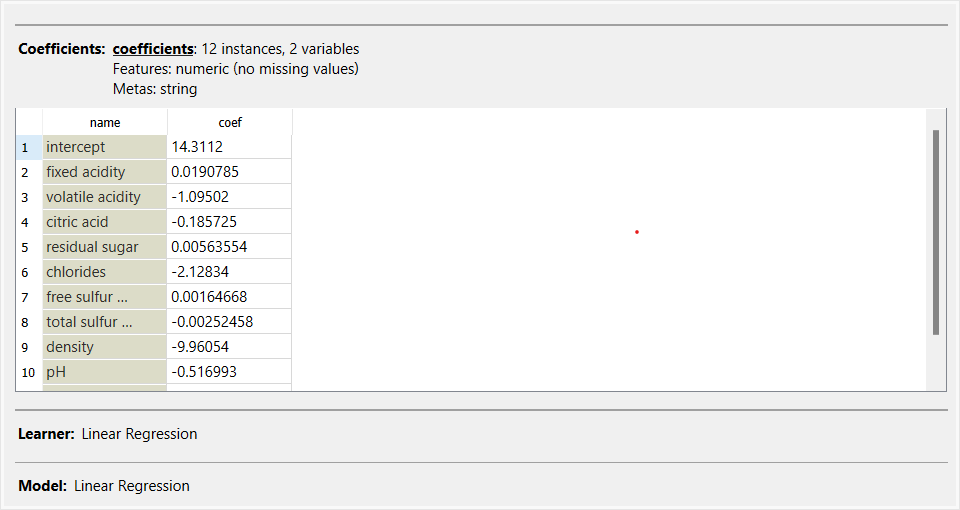
Đây là giá trị của sức mạnh chính quy hóa. Sức mạnh chính quy hóa kiểm soát mức độ ảnh hưởng của chính quy hóa đến mô hình. Giá trị cao hơn dẫn đến chính quy hóa mạnh hơn, có thể giúp ngăn ngừa quá phù hợp nhưng cũng có thể làm giảm hiệu suất của mô hình.

* **Elastic net regression:** Tùy chọn này cho phép sử dụng hồi quy mạng đàn hồi, là một loại chính quy hóa kết hợp hồi quy Ridge và hồi quy Lasso.
* Elastic net mixing

**0.50: 0.50**

Đây là tỷ lệ trộn trong hồi quy mạng đàn hồi. Tỷ lệ trộn kiểm soát sự cân bằng giữa hồi quy Ridge và hồi quy Lasso. Giá trị bằng 0 tương đương với hồi quy Ridge và giá trị bằng 1 tương đương với hồi quy Lasso

**Kết quả xây dựng mô hình hồi quy**

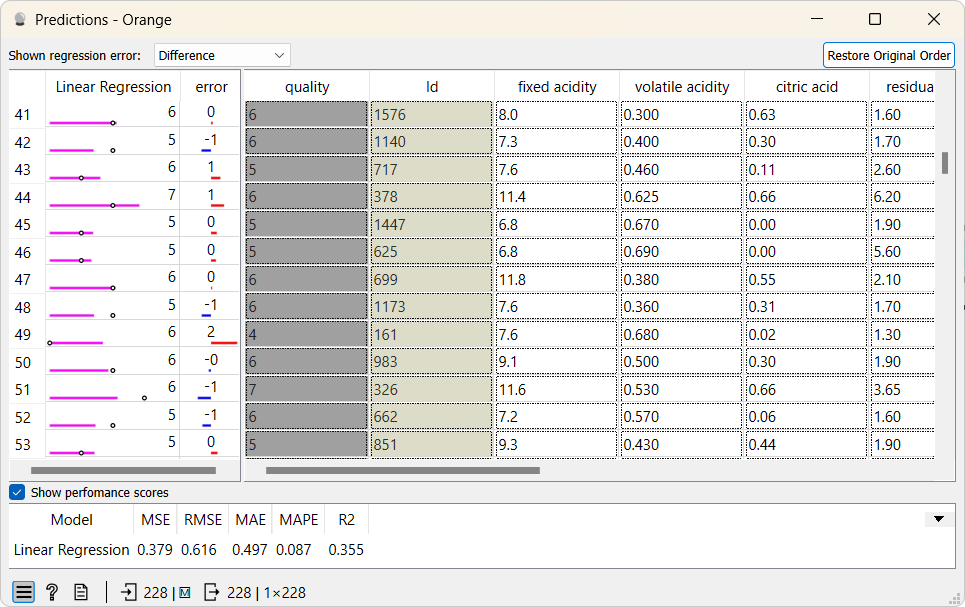
****

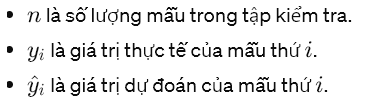
* **Giải thích:**
* intercept (β0: Hệ số chặn): khi không tính tới sự ảnh hưởng của các biến độc lập) thì Y = β0
* Các dòng còn lại:

| Tên biến độc lập  (11 biến đã chọn) | Hệ số hồi quy (β1, β2, ..., βn)  (mức độ và chiều hướng ảnh hưởng của biến độc lập đến biến phụ thuộc)   * Mức độ: Hệ số càng lớn ảnh hưởng còn lớn * Ảnh hưởng: dấu (-) thì biến độc lập nghịch với biến phụ thuộc, dấu (+) thì thuận |
| --- | --- |

* Đọc ý nghĩa: Mô hình hồi tuyến tính quy xây dựng được có
* Hệ số chặn = 14.3112 ⇒
* Hệ số hồi quy của **fixed acidity** = 0.0190785 ⇒ thuộc tính fixed acidity có **chiều hướng** ảnh hưởng **tốt** đến chất lượng của rượu tuy nhiên **mức độ** ảnh hưởng **không lớn**. Nghĩa là khi fixed acidity tăng thì chất lượng rượu cũng sẽ tăng nhưng không đáng kể
* ….
* Hệ số hồi quy của **PH**: -0.516993 ⇒ thuộc tính PH có **chiều hướng** ảnh hưởng **xấu** do mang dấu âm) đến chất lượng của rượu tuy nhiên **mức độ** ảnh hưởng **không quá lớn.**  Nghĩa là khi fixed acidity tăng thì chất lượng rượu sẽ giảm nhưng không đáng kể

**Kết quả dự đoán**

****

* **Giải thích hình ảnh:**
* **Đánh giá mô hình:** Dựa vào sự chênh lệch giữa kết quả dự đoán và kết quả thực tế
* **Sai số bình phương trung bình (MSE): **
* Giá trị MSE là 0.364.
* **Đánh giá:** MSE là thước đo độ lệch trung bình giữa các giá trị dự đoán và giá trị thực tế. Một giá trị MSE thấp hơn cho thấy mô hình dự đoán chính xác hơn.. Giá trị MSE càng nhỏ, mô hình càng có độ chính xác cao. Trong trường hợp này, MSE = 0.364 là một giá trị tương đối nhỏ, cho thấy mô hình có độ chính xác cao.
* **Sai số gốc bình phương trung bình (RMSE):**

****

* Giá trị RMSE là 0.603.
* **Đánh giá:** RMSE là căn bậc hai của MSE, ***dễ dàng giải thích hơn MSE.*** Giá trị RMSE càng nhỏ, mô hình càng có độ chính xác cao. Trong trường hợp này, RMSE = 0.603 là một giá trị tương đối nhỏ, cho thấy mô hình có độ chính xác cao.
* **Sai số trung bình tuyệt đối (MAE):**

****

* Giá trị MAE là 0.465.
* **Đánh giá:** MAE là một chỉ số đo lường sai số trung bình tuyệt đối giữa giá trị dự đoán của mô hình và giá trị thực. ***MAE ít nhạy cảm với các giá trị sai lệch lớn hơn so với MSE***.. Giá trị MAE càng nhỏ, mô hình càng có độ chính xác cao. Trong trường hợp này, MAE = 0.465 là một giá trị tương đối nhỏ, cho thấy mô hình có độ chính xác cao.
* **Sai số phần trăm trung bình tuyệt đối (MAPE):**

****

* Giá trị MAPE là 0.084.
* **Đánh giá:** MAPE là một chỉ số đo lường sai số phần trăm trung bình tuyệt đối giữa giá trị dự đoán của mô hình và giá trị thực tế. ***MAPE hữu ích khi các giá trị của biến phụ thuộc có các đơn vị đo khác nhau***. Giá trị MAPE càng nhỏ, mô hình càng có độ chính xác cao. Trong trường hợp này, MAPE = 0.084 là một giá trị tương đối nhỏ, cho thấy mô hình có độ chính xác cao.
* **Hệ số xác định (R^2):**

****

****

* Giá trị R^2 là 0.346.
* **Đánh giá:** R^2 là một chỉ số đo lường tỷ lệ biến đổi của giá trị thực tế được giải thích bởi mô hình. Giá trị R^2 càng cao, mô hình càng có khả năng giải thích tốt biến đổi của giá trị thực tế. Trong trường hợp này, R^2 = 0.346 là một giá trị tương đối nhỏ, cho thấy mô hình chỉ giải thích được 34,6% biến đổi của giá trị thực tế.

**Association Rule**

Dataset thu thập: Gồm các cột:

Apple: Táo

Bread: Bánh mì

Butter: Bơ

Cheese: Phô mai

Corn: Bắp

Dill: Rau thì là

Eggs: Trứng

Ice cream: Kem

Kidney Beans: Đậu tây

Milk: Sữa

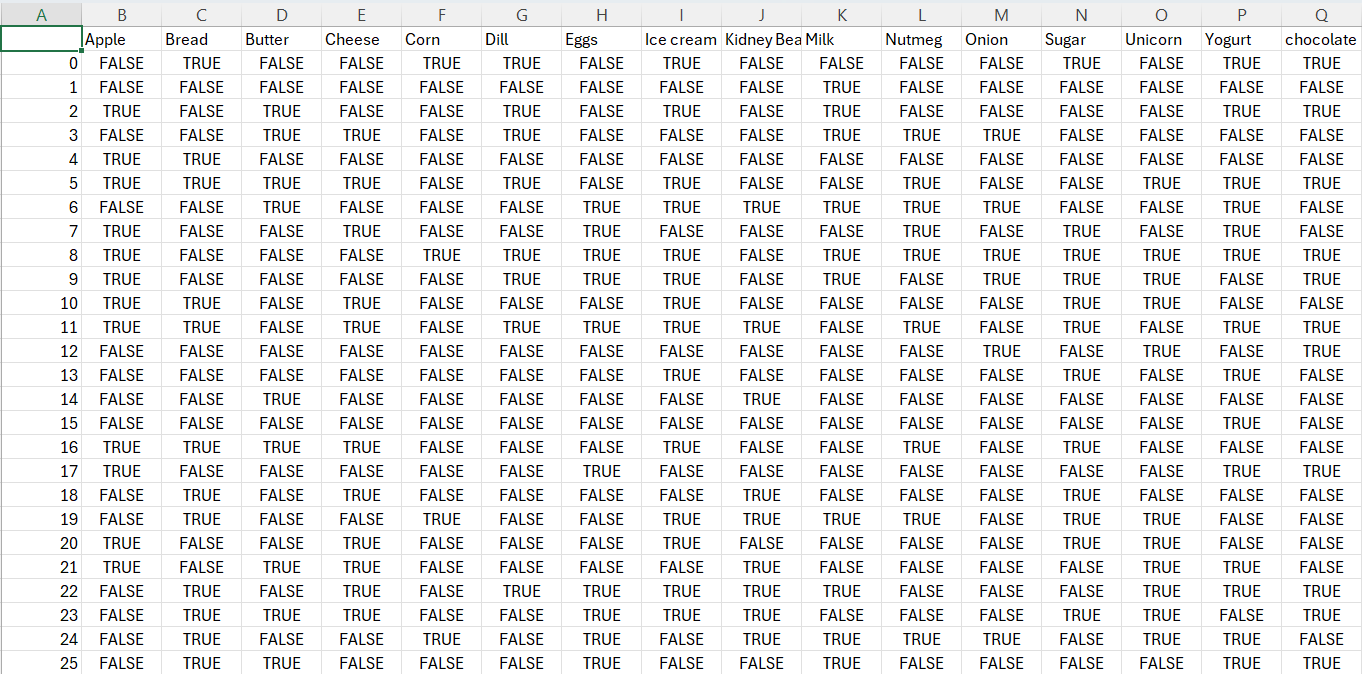
Onion: Hành

Sugar: Đường

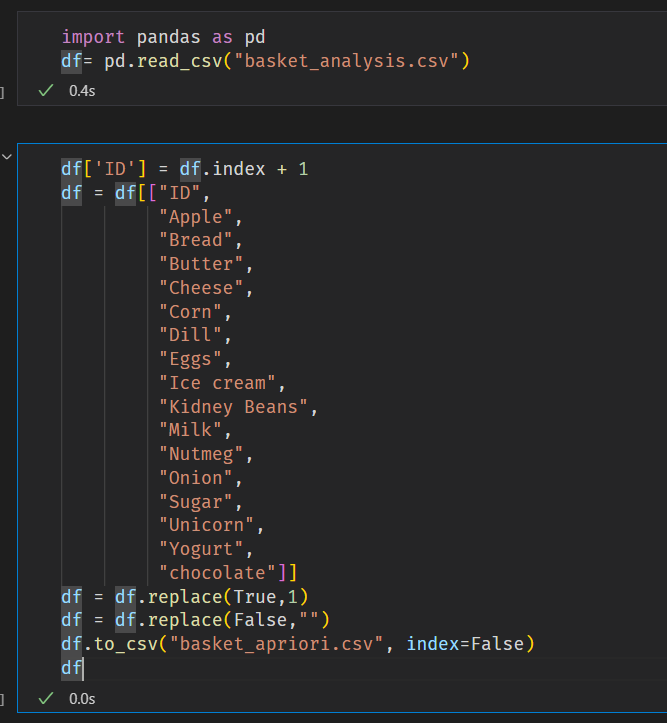
Unicorn: Phẩm màu

Yogurt: Sữa chua

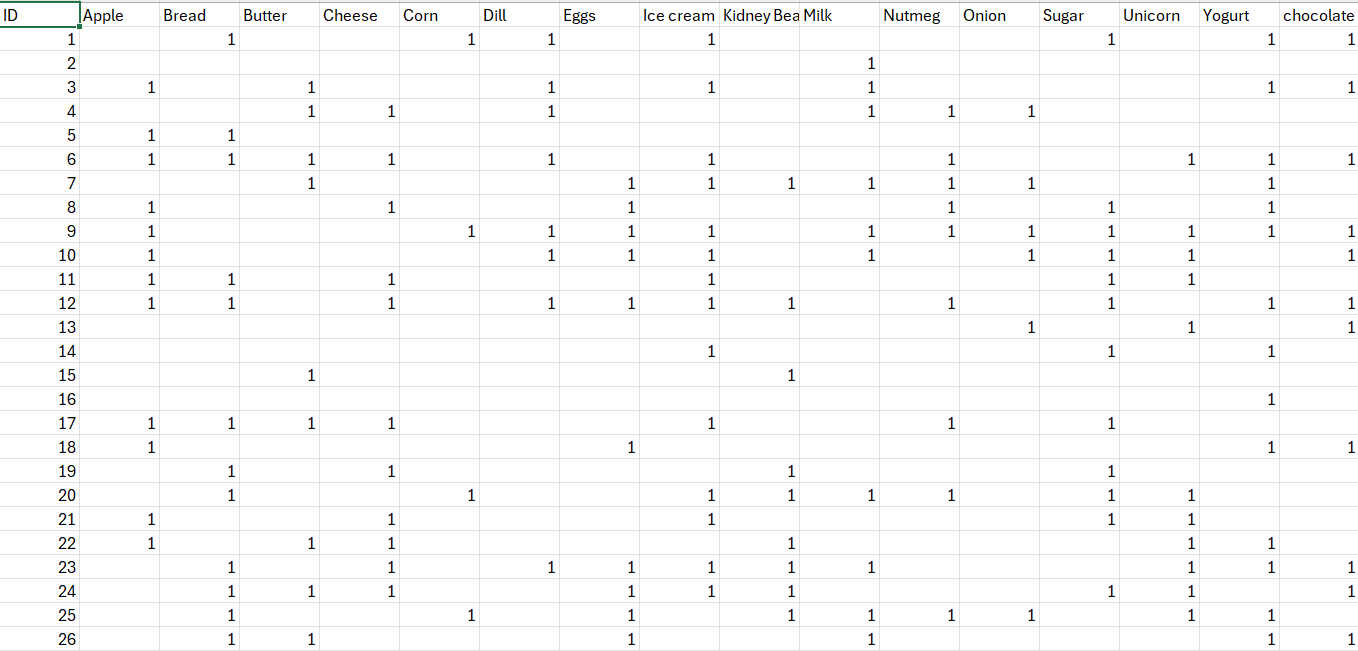
chocolate: Sô cô la

****

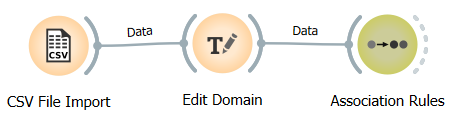
**Qua quá trình xử lý:**

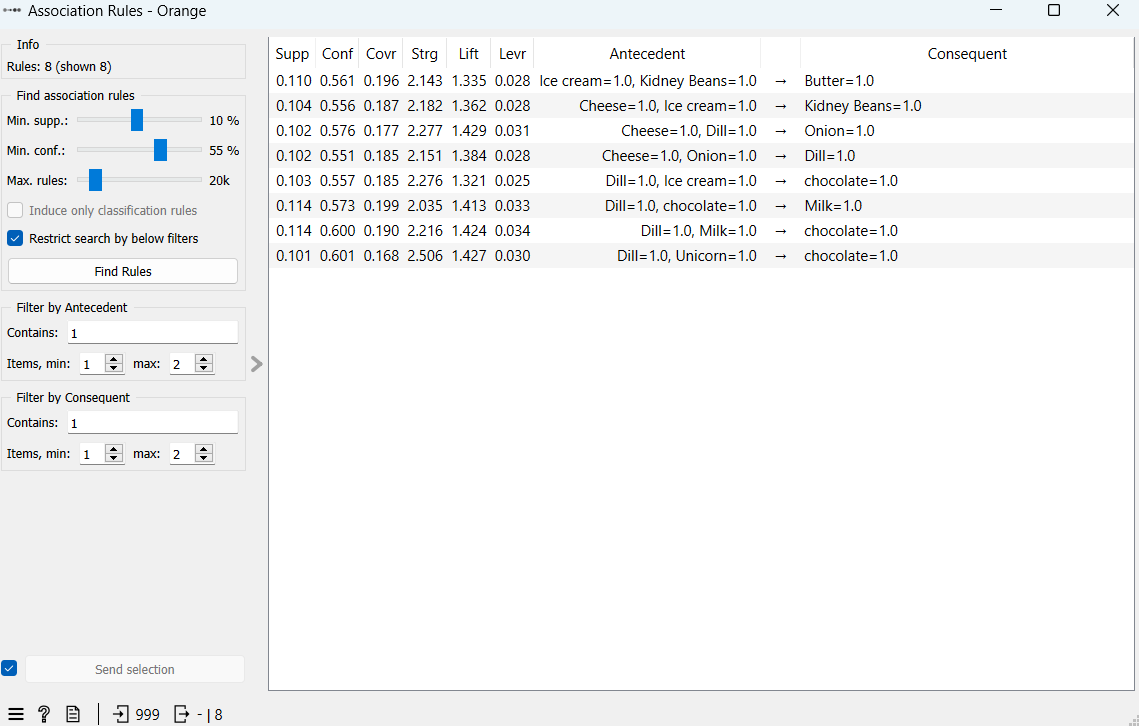
****

**File sau khi xử lý:**

****

**Chạy appriori trên orange với minsup=10% và mincof = 55%**

****

****

**⇒ Tạo được 8 luật kết hợp**

* **Contains: số lượng item tối thiểu trong luật (có thể là bên Atencendent hoặc Consequent)**
* **Supp: độ phổ biến**
* **Conf: độ tin cậy**
* **Covr: độ bao phủ**
* **Strg: độ phổ biến của Atencedent / độ phổ biến của Consequent**
* **Lift: data \* độ tin cậy / độ phổ biến**
* **Leverage: cái này ko cần quan tâm**