<https://drive.google.com/drive/folders/1KPBewqe2w9ukMtgrjfGKsKpsoOfaDqe-?fbclid=IwAR3pNEofdvXGCzDzllWVDKwj38CzfHB6msfLoAEyava95vBFW3BpPoNezF4>

tìm hiểu luật kết hợp - nhóm 9

323-A2

Link đăng ký nhóm

[Data Mining 63HTTT1](https://docs.google.com/spreadsheets/d/1GGj0W4wLSQ7XNmUrvQCyScLwVvTLLJwLGM-0W1kfyis/edit?usp=sharing)

0 - Unsatisfied

1 - Neutral

2 - Satisfied

<https://www.kaggle.com/datasets/uom190346a/e-commerce-customer-behavior-dataset>

ifelse(A<18,1,ifelse(A<27,3,ifelse(A<36,5,ifelse(A<45,7,ifelse(A<54,9,11)))))

[Toàn cảnh thị trường (cafef.vn)](https://s.cafef.vn/du-lieu.chn)

[General Statistics Office of Vietnam (gso.gov.vn)](https://www.gso.gov.vn/)

[data.world](https://data.world/datasets/ecommerce)

Các nghiên cứu liên quan bài toán  
Các quốc gia khác:

Hoa Kỳ - Cục Thống kê Lao động (Bureau of Labor Statistics): https://www.bls.gov/

Anh - Cục Thống kê Quốc gia (Office for National Statistics): https://www.ons.gov.uk/

Úc - Cơ quan Thống kê Úc (Australian Bureau of Statistics): https://www.abs.gov.au/

Canada - Cục Thống kê Canada (Statistics Canada): https://www.statcan.gc.ca/

Đức - Cục Thống kê Liên bang Đức (Federal Statistical Office of Germany): https://www.destatis.de/

Nhật Bản - Cục Thống kê Nhật Bản (Statistics Japan): https://www.stat.go.jp/

Mục tiêu mình để làm gì ?  
Đánh giá mức độ hài lòng: Dựa trên mối quan hệ của thuộc tính  
Dữ liệu mình có nhãn phân lớp để làm gì ?  
Khai phá dữ liệu bằng phương pháp phân lớp để phân loại hành vi khách hàng trên sàn thương mại điện tử

0 - Unsatisfied - không hài lòng

1 - Neutral - trung lập

2 - Satisfied - hài lòng  
Câu hỏi trao đổi:

1. Có nhãn là bài toán có giám sát
2. Mục tiêu:

* Phân khúc khách hàng: Phân tích và phân loại khách hàng dựa trên nhân khẩu học, thói quen chi tiêu và mức độ hài lòng.
* Phân tích sự hài lòng: Điều tra các yếu tố ảnh hưởng đến sự hài lòng của khách hàng và xác định các lĩnh vực cần cải thiện.
* Chiến lược quảng cáo, chiến lược duy trì.

1. Mục tiêu:

* Phân cụm: Mục tiêu của bài toán phân cụm là tìm ra cách chia dữ liệu thành các nhóm (clusters) tự nhiên mà không có thông tin trước về nhãn hay phân loại của từng nhóm. Trong phân cụm, chúng ta cố gắng tìm ra cấu trúc ẩn của dữ liệu và xác định các nhóm tương tự nhau.
* Phân lớp: Mục tiêu của bài toán phân lớp là xây dựng một mô hình dự đoán để gán nhãn (class label) cho các mẫu dữ liệu mới dựa trên các đặc trưng (features) đã biết trước. Trong phân lớp, chúng ta có các nhãn đã biết sẵn và huấn luyện mô hình để phân loại các mẫu vào các nhãn đã xác định trước đó.

Bài toán ID3:

* Dự đoán nhãn của mẫu mới chưa có nhãn

Các câu hỏi của các nhóm:  
Nhóm 19  
- Bài toán hồi quy để làm gì? Hồi quy tính có nhãn để ra 1 dữ liệu dự báo bạn phải ra 1 phương trình hồi quy y = f(x) khi tôi tính y = 1.7 thì anh làm tròn lên thì nó là 2 thuộc về nhãn 2 vì từ 1.7 lên 2 nó có sai số epsilon lên 0.3 khi anh hồi quy lên ẵm 2 cái sai số  
Nhóm 22  
- Làm thế nào đảm bảo chất liệu táo?  
- Phương pháp phân lớp là gì?  
Đa số các nhóm bị không phân biệt được phân lớp là phân cụm  
Đề bài các nhóm cũng chưa chọn chính xác   
Thầy cho từ h đến hết ngày mai nhóm nào cập nhật lại đề bài bản cuối cùng  
Vì đa số các nhóm không xác định được bài toán

1.Customer ID: ID khách hàng

2. Gender: Giới tính

3.Age: Tuổi

4. City: Thành phố

5. Membership Type: Loại thành viên

6. Total Spend: Tổng chi tiêu

7. Items Purchased: Các mặt hàng đã mua

8. Average Rating: Đánh giá trung bình

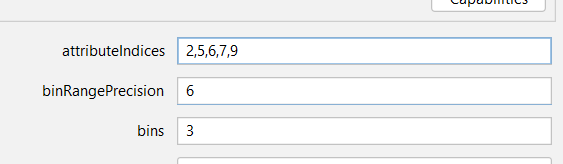
9. Discount Applied: Áp dụng giảm giá

10. Days Since Last Purchase: Số ngày kể từ lần mua cuối cùng

11. Satisfaction Level: Mức độ hài lòng

Cách dùng weka

1. Bổ sung dữ liệu thiếu cho cột nhãn
2. Xóa cột ID
3. Discretize



1. Chuyển Nominal

Sử dụng ID3 không phù hợp vì phân tích dữ liệu theo biên độ thời gian

Chia Age thành 2 nhóm

* 26-35: young adults.
* 36-43: middle-aged adults.

ifelse(A<36,0,1)

Nguồn:

<https://vi.wikipedia.org/wiki/Thanh_ni%C3%AAn>

Chia Total Spend thành 3 nhóm

* Low spending (410 - 600)
* Average spending (601 - 1000)
* High spending (1001 - 1520)

ifelse(A<601,0,ifelse(A<1001,1,2))

Chia Items Purchased thành 3 nhóm

* Little product (2 - 5)
* Average product (6 - 10)
* Many product (11 - 21)

ifelse(A<6,0,ifelse(A<11,1,2))

Chia Average Rating thành 3 nhóm

* Low rating (3 - 3.9)
* Average rating (4 - 4.4)
* High rating (4.5 - 4.9)

ifelse(A<4,0,ifelse(A<4.5,1,2))

Chia Days Since Last Purchase thành 3 nhóm

* Recent purchase (9 - 21 days)
* Regular purchase (22 - 42 days)
* Infrequent purchase (43 - 63 days)

ifelse(A<22,0,ifelse(A<43,1,2))

minsup <https://cempartner.com/vi/cem-partner-blog/do-luong-su-hai-long-cua-khach-hang.html>

L 5 :

["'Average product'", "'Low rating'", "'Low spending'", "'Middle-aged adults'", 'Bronze'], ["'Average product'", "'Low rating'", "'Low spending'", "'Middle-aged adults'", 'Female'], ["'Average product'", "'Low rating'", "'Low spending'", 'Bronze', 'Female'],

["'Average product'", "'Low rating'", "'Middle-aged adults'", 'Bronze', 'Female'],

["'Average product'", "'Low spending'", "'Middle-aged adults'", 'Bronze', 'Female'],

["'Average rating'", "'Average spending'", "'Young adults'", 'Male', 'Silver'],

["'High rating'", "'High spending'", "'Many product'", "'Recent purchase'", "'Young adults'"], ["'High rating'", "'High spending'", "'Many product'", "'Recent purchase'", '2'],

["'High rating'", "'High spending'", "'Many product'", "'Young adults'", '2'],

["'High rating'", "'High spending'", "'Many product'", "'Young adults'", 'Gold'],

["'High rating'", "'High spending'", "'Many product'", '2', 'Gold'],

["'High rating'", "'High spending'", "'Recent purchase'", "'Young adults'", '2'],

["'High rating'", "'High spending'", "'Young adults'", '2', 'Gold'],

["'High rating'", "'Many product'", "'Recent purchase'", "'Young adults'", '2'],

["'High rating'", "'Many product'", "'Young adults'", '2', 'Gold'],

["'High spending'", "'Many product'", "'Recent purchase'", "'Young adults'", '2'],

["'High spending'", "'Many product'", "'Young adults'", '2', 'Gold'],

["'Low rating'", "'Low spending'", "'Middle-aged adults'", "'Regular purchase'", 'Bronze'],

["'Low rating'", "'Low spending'", "'Middle-aged adults'", "'Regular purchase'", 'Female'], ["'Low rating'", "'Low spending'", "'Middle-aged adults'", 'Bronze', 'Female'],

["'Low rating'", "'Low spending'", "'Regular purchase'", 'Bronze', 'Female'],

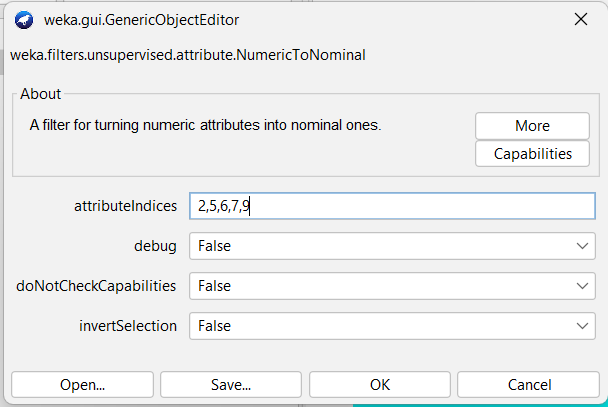
["'Low rating'", "'Middle-aged adults'", "'Regular purchase'", 'Bronze', 'Female'],

["'Low spending'", "'Middle-aged adults'", "'Regular purchase'", 'Bronze', 'Female'],

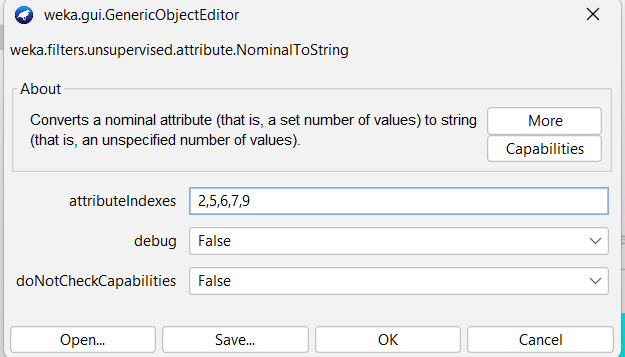
["'Many product'", "'Recent purchase'", "'Young adults'", 'False', 'Male']]

Cách tiền xử lý:

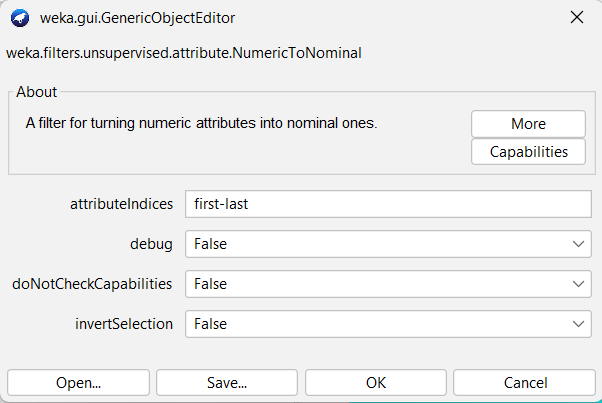
* Xóa cột Customer ID
* Bổ sung dữ liệu thiếu Satisfaction Level = 0.0
* Nhận xét thấy cột 2,5,6,7,9 dữ liệu liên tục => rời rạc hóa
* MathExpression 5 thuộc tính
* Chia Age thành 2 nhóm
* 26-35: 0
* 36-43: 1
  + ifelse(A<36,0,1)
  + Nguồn:
  + https://vi.wikipedia.org/wiki/Thanh\_ni%C3%AAn
* Chia Total Spend thành 3 nhóm
* 0 (410 - 600)
* 1 (601 - 1000)
* 2 (1001 - 1520)
  + ifelse(A<601,0,ifelse(A<1001,1,2))
* Chia Items Purchased thành 3 nhóm
* 0 (2 - 5)
* 1 (6 - 10)
* 2 (11 - 21)
  + ifelse(A<6,0,ifelse(A<11,1,2))
* Chia Average Rating thành 3 nhóm
* 0 (3 - 3.9)
* 1 (4 - 4.4)
* 2 (4.5 - 4.9)
  + ifelse(A<4,0,ifelse(A<4.5,1,2))
* Chia Days Since Last Purchase thành 3 nhóm
* 0 (9 - 21 days)
* 1 (22 - 42 days)
* 2 (43 - 63 days)
  + ifelse(A<22,0,ifelse(A<43,1,2))
* Chuyển về chữ có nghĩa, NumbericToNominal 2,5,6,7,9

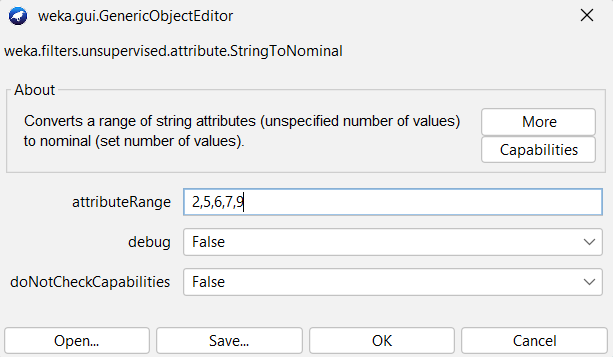


* NominalToString 2,5,6,7,9



* Chọn edit và thay đổi từng giá trị trong thuộc tính
* Chia Age thành 2 nhóm
* 0 - Young adults.
* 1 - Middle-aged adults.
* Chia Total Spend thành 3 nhóm
* 0 - Low spending
* 1 - Average spending
* 2 - High spending
* Chia Items Purchased thành 3 nhóm
* 0 - Little product
* 1 - Average product
* 2 - Many product
* Chia Average Rating thành 3 nhóm
* 0 - Low rating
* 1 - Average rating
* 2 - High rating
* Chia Days Since Last Purchase thành 3 nhóm
* 0 - Recent purchase
* 1 - Regular purchase
* 2 - Infrequent purchase
* Chuẩn hóa NumbericToNominal và StringToNominal





NavBayes

class NaiveBayesClassifier:

def \_\_init\_\_(self):

# Khởi tạo biến thành viên class\_probabilities và feature\_probabilities

self.class\_probabilities = None

self.feature\_probabilities = None

def fit(self, X, y):

"""

Huấn luyện mô hình Naive Bayes Classifier dựa trên dữ liệu huấn luyện.

Parameters:

X (numpy.ndarray): Ma trận dữ liệu đầu vào.

y (numpy.ndarray): Mảng nhãn của dữ liệu huấn luyện.

Returns:

None

"""

# ...

def predict(self, X):

"""

Dự đoán nhãn cho các mẫu dữ liệu đầu vào sử dụng mô hình đã được huấn luyện.

Parameters:

X (numpy.ndarray): Ma trận dữ liệu đầu vào.

Returns:

list: Danh sách các nhãn dự đoán.

"""

# ...

def get\_feature\_probabilities(self):

"""

Trả về xác suất của các đặc trưng cho từng lớp sau khi mô hình đã được huấn luyện.

Returns:

defaultdict: Từ điển lưu trữ xác suất của các đặc trưng cho từng lớp.

"""

# ...

def evaluate(self, X, y):

"""

Đánh giá hiệu suất của mô hình dự đoán.

Parameters:

X (numpy.ndarray): Ma trận dữ liệu đầu vào.

y (numpy.ndarray): Mảng nhãn thực tế.

Returns:

tuple: Bốn giá trị - độ chính xác, số lượng dự đoán sai, kappa statistic, ma trận nhầm lẫn.

"""

# ...

def read\_data(file\_path):

"""

Đọc dữ liệu từ file CSV.

Parameters:

file\_path (str): Đường dẫn đến file CSV.

Returns:

tuple: Ba giá trị - ma trận dữ liệu đầu vào, mảng nhãn và tên đặc trưng.

"""

# ...

def main():

"""

Chương trình chính.

Returns:

None

"""

# ...

**Từ ma trận tương quan, bạn có thể suy ra các thông tin và thực hiện các hoạt động sau:**

Đánh giá mối quan hệ tương quan: Ma trận tương quan cho bạn cái nhìn tổng quan về mối quan hệ tương quan giữa các biến trong tập dữ liệu. Bằng cách xem xét các giá trị tương quan, bạn có thể đánh giá độ mạnh yếu và hướng của mối quan hệ giữa các biến. Điều này có thể giúp bạn hiểu các mẫu hoặc liên kết tồn tại giữa các biến.

Phát hiện biến độc lập: Nếu giữa các biến không có mối quan hệ tương quan đáng kể, bạn có thể kết luận rằng chúng là độc lập và không ảnh hưởng lẫn nhau. Điều này có thể hữu ích trong việc lựa chọn các biến quan trọng và loại bỏ những biến dư thừa trong mô hình hóa dữ liệu.

Xác định biến quan trọng: Nếu một biến có mối quan hệ tương quan mạnh với nhiều biến khác, nó có thể được coi là biến quan trọng trong việc giải thích biến đổi của các biến khác. Bằng cách xem xét ma trận tương quan, bạn có thể xác định các biến quan trọng và tập trung nghiên cứu hoặc sử dụng chúng trong các mô hình hoặc phân tích dữ liệu.

Xây dựng mô hình dự đoán: Ma trận tương quan có thể cung cấp cho bạn thông tin quan trọng để xây dựng mô hình dự đoán. Bằng cách chọn các biến có mối quan hệ tương quan mạnh với biến mục tiêu, bạn có thể xây dựng mô hình dự đoán hiệu quả hơn và đạt được độ chính xác cao hơn.

Phân tích thành phần chính: Ma trận tương quan thường được sử dụng trong phân tích thành phần chính (PCA), một phương pháp giảm chiều dữ liệu. PCA giúp giảm số chiều của dữ liệu bằng cách tìm các thành phần chính (principal components) có ảnh hưởng lớn nhất đến sự biến thiên của dữ liệu. Ma trận tương quan được sử dụng để tính toán các thành phần chính và phân tích cấu trúc của dữ liệu.

Tóm lại, ma trận tương quan cung cấp thông tin về mối quan hệ giữa các biến trong tập dữ liệu và có thể được sử dụng để đánh giá, xác định và sử dụng thông tin đó trong các hoạt động phân tích dữ liệu và xây dựng mô hình.