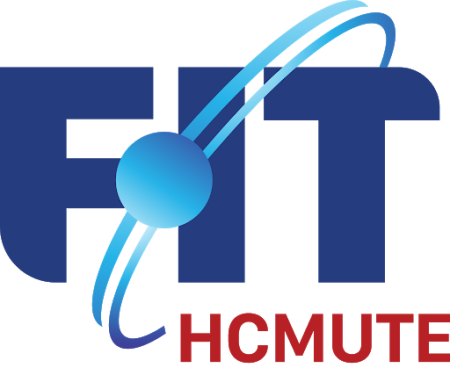
**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

**BỘ MÔN KHAI PHÁ DỮ LIỆU**

-----🙞🙜🕮🙞🙜-----

****

**Môn học: DataMinning**

**KHAI PHÁ DỮ LIỆU VỀ SỰ HÀI LÒNG CỦA HÀNH KHÁCH TRONG CHUYẾN BAY**

**GVDH:** **Nguyễn Văn Thành**

**SVTH:**

**Trần Văn Đức 20133037**

**Lê Huy Hoàng 20133044**

**Trần Văn Thắng 20133121**

**TP.HCM, tháng 05 năm 2023**

# **LỜI CẢM ƠN**

Lời đầu tiên nhóm xin phép được gửi lời cảm ơn chân thành và sâu sắc nhất đến với Khoa Công Nghệ Thông Tin – Trường Đại Học Sư Phạm Kỹ Thuật Thành Phố Hồ Chí Minh đã tạo điều kiện cho nhóm chúng em được học tập, phát triển nền tảng kiến thức sâu sắc và thực hiện đề tài này.

Với sự hướng dẫn nhiệt tình, giảng dạy tận tình đầy đủ kiến thức của thầy Nguyễn Văn Thành, chúng em đã học tập và hiểu được những kiến thức cơ bản về môn Khai phá dữ liệu. Qua đó chúng em biết cách cài đặt và sử dụng Weka để hỗ trợ cho việc học tập và khai phá dữ liệu.

Tuy nhiên lượng kiến thức là vô tận và với khả năng hạn hẹp chúng em đã rất cố gắng để hoàn thành một cách tốt nhất. Chính vì vậy việc xảy ra những thiếu sót là điều khó có thể tránh khỏi. Chúng em hy vọng nhận được sự góp ý tận tình của thầy để qua đó chúng em có thể rút ra được bài học kinh nghiệm và hoàn thiện và cải thiện nâng cấp lại sản phẩm của mình một cách tốt nhất có thể.Chúng em xin chân thành cảm ơn!

**PHẦN MỞ ĐẦU**

## **1. Tính cấp thiết của đề tài**

Dịch vụ hàng không hiện nay đang là một loại hình di chuyển được mợi người lựa chọn. Sở dĩ nó được ưa chuộng không chỉ bởi vì tốc độ di chuyển cùng độ an toàn của chuyến bay mà còn nằm ở các dịch vụ chăm sóc hành khách trên chuyến bay

Mức sống của người dân càng lên cao thì chất lượng cuộc sống cũng tăng theo. Do đó thì các dịch vụ trên chuyến bay cũng sẽ phải tốt lên theo hằng ngày.

Khi được học môn Khai phá dữ liệu, chúng em có cơ hội được tiếp xúc với các thuật toán phân nhóm,phân cụm,luật kết hợp, để phục vụ cho mục đích phân tích và làm rõ hơn về một tập dữ liệu lớn. Vậy nên, với mong muốn khai phá về dữ liệu các dịch vụ trong chuyến bay để từ đó dự đoán được khách hàng có hài lòng hay không về chất lượng chuyến bay thông qua các đánh giá ở những dịch vụ để đưa ra các giải pháp cải thiện chất lượng phục vụ trong chuyến bay đối với các hãng hàng không.

Vậy nên nhóm chúng em đã quyết định thực hiện đề tài: ***“Khai phá dữ liệu về phân tích sự hài lòng của hành khách trong chuyến bay.”.***

## **Đối tượng nghiên cứu**

Đối với đề tài này, chúng em tập trung nghiên cứu các đối tượng sau:

- Tìm hiểu về Weka.

- Phân tích tập dữ liệu và tiền xử lí về sự hài lòng của hành khách trong

chuyến bay.

- Thể hiện các biểu đồ để làm rõ mục tiêu phân tích.

## **Phạm vi nghiên cứu**

Chúng em sẽ nghiên cứu các vấn đề cơ bản về khai phá dữ liệu. Vẽ các biểu đồ để phân tích rõ ý nghĩa đề tài của nhóm. Xây dựng dashboard tổng quan về các biểu đồ đã được hoàn thành.

Ngoài ra nhóm sẽ sử dụng code python để kết hợp vào trong đề tài.

## **Bố cục đề tài**

Chương 1: Cơ sở lý thuyết.

Chương 2: Xây dựng dashboard và phân tích.

## **Kết quả dự kiến đạt được**

Thông qua việc tìm hiểu đề tài này chúng em dự kiến sẽ hoàn thành những nhiệm vụ như sau:

- Tìm hiểu về tổng quan, đặc điểm, lợi ích, tính năng của SuperSet.

* Xây dựng được một dashboard về tập dữ dữ liệu mà nhóm chọn.
* Phân tích rõ hơn về các biểu đồ đã vẽ.
* Tìm hiểu thêm về code python tích hợp trong đề tài.

# **PHẦN NỘI DUNG**

# **CHƯƠNG 1: CƠ SỞ LÝ THUYẾT**

## **Tổng quan về khai phá dữ liệu**

* 1. **Khái niệm**

**Mục tiêu của việc khai phá dữ liệu**

.

## **2. Sơ lược về tập dữ liệu**

* Dữ liệu này do một tổ chức hàng không đưa ra. Tên thực tế của công ty không được đưa ra do nhiều mục đích khác nhau, đó là lý do tại sao có tên hãng hàng không không bị tiết lộ.
* Tập dữ liệu bao gồm thông tin chi tiết về những khách hàng đã bay với họ. Phản hồi của khách hàng về các bối cảnh khác nhau và dữ liệu chuyến bay của họ đã được tổng hợp.
* Mục đích chính của tập dữ liệu này là dự đoán liệu một khách hàng trong tương lai có hài lòng với dịch vụ của họ hay không khi cung cấp thông tin chi tiết về các giá trị tham số khác.
* Ngoài ra, các hãng hàng không cũng cần biết khía cạnh nào của các dịch vụ do họ cung cấp phải được nhấn mạnh nhiều hơn để tạo ra nhiều khách hàng hài lòng hơn.

**2.1. Dữ liệu**

Gồm 129880 dòng và 23 cột

- Gender: Giới tính của hành khách (Nữ, Nam)

- Customer\_type: Loại khách hàng (Khách hàng trung thành, khách hàng không trung thành)

- Age: Tuổi thực của hành khách

- Type\_of\_travel: Mục đích chuyến bay của hành khách (Đi du lịch cá nhân, Đi công tác)

- Customer\_class: Hạng du lịch trên máy bay của hành khách (hạng Thương gia, hạng Eco, hạng Eco Plus)

- Flight\_distance: Khoảng cách bay của hành trình này

- Inflight\_wifi\_service: Mức độ hài lòng đối với dịch vụ wifi trên chuyến bay (0: Không áp dụng; 1-5)

- Departure\_arrival\_time\_convenient: Mức độ hài lòng về Thời gian đi / đến thuận tiện

- Ease\_of\_online\_booking: Mức độ hài lòng khi đặt phòng trực tuyến

- Gate\_location: Mức độ hài lòng về vị trí cổng

- Food\_and\_drink: Mức độ hài lòng về Đồ ăn và thức uống

- Online\_boarding: Mức độ hài lòng của nội trú trực tuyến

- Seat\_comfort: Mức độ hài lòng về sự thoải mái của chỗ ngồi

- Inflight\_entertainment: Mức độ hài lòng của giải trí trên chuyến bay

- Onboard\_service: Mức độ hài lòng của dịch vụ trên chuyến bay

- Leg\_room\_service: Mức độ hài lòng của dịch vụ phòng chân

- Baggage\_handling: Mức độ hài lòng của việc xếp dỡ hành lý

- Checkin\_service: Mức độ hài lòng về dịch vụ nhận phòng

- Inflight\_service: Mức độ hài lòng của dịch vụ trên chuyến bay

- Cleanliness: Mức độ hài lòng về Sạch sẽ

- Departure\_delay\_in\_minutes: Bị hoãn vài phút khi khởi hành

- Arrival\_delay\_in\_minutes: Bị Chậm Phút Khi Đến

- Satisfaction: Mức độ hài lòng của hãng hàng không (Hài lòng, trung tính hoặc không hài lòng)

**2.2. Trực quan hóa dữ liệu**

Dashboard chung tổng quát:

Có sử dụng PowerBI để vẽ các chart.

Tổng quát về đánh giá chung của hành khách:

Graphical user interface

Description automatically generated

Nhìn chung vào biểu đồ trên ta thấy mức đánh giá hài lòng của hành khách trong chuyến bay thấp hơn đánh giá trung bình hoặc không hài lòng.

Các loại dịch vụ trong chuyến bay:

**A screenshot of a computer

Description automatically generated**

Có thể thấy, mỗi loại dịch vụ nó đều có ảnh hưởng riêng tới hành khách sử dụng dịch vụ đó. Không thể nói là loại dịch vụ nào là quan trọng nhất , mà tùy vào mỗi nhu cầu của người sử dụng thì sẽ có những dịch vụ tương ứng. Cho nên sẽ có những tiêu chuẩn đánh giá khác nhau. Vì vậy các hãng hàng không nên tối ưu nhất các dịch vụ trong suốt chuyến bay của khách hàng, để nhằm mang đến cho khách hàng những cảm giác thoải mái và trải nghiệm tốt trong chuyến bay dài.

Các yếu tố liên quan đến hành khách:

**Chart, histogram

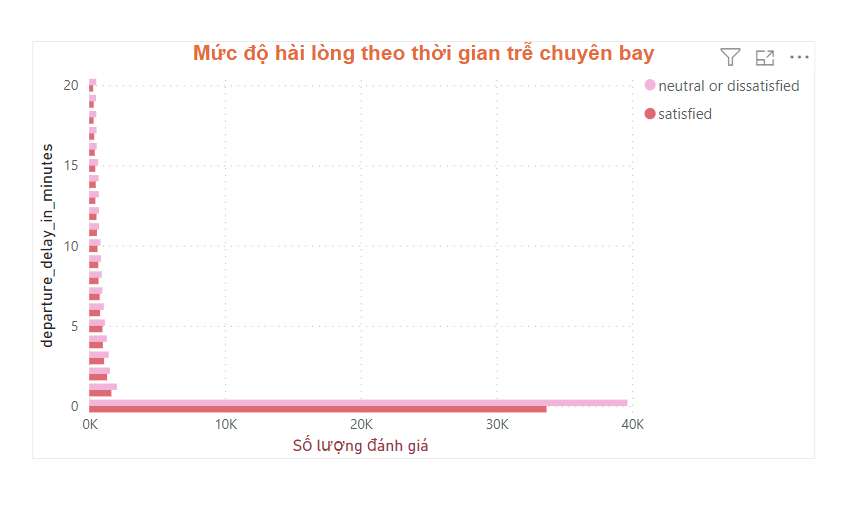
Description automatically generated**

**Chart, bar chart

Description automatically generated**

**Chart, funnel chart

Description automatically generated**

****

Ta thấy thời gian trễ chuyến bay không ảnh hưởng nhìu đến sự hài lòng của khách hàng

**Kết luận:** Qua các biểu đồ, chúng ta có thể thấy được phần nào sự hài lòng của các hàng khách qua các biểu đồ và cũng có thể hiểu được, điều gì có chất lượng chưa tốt để từ đó cải thiện.

**2.3. Tiền xử lí dữ liệu**

Đọc file csv

|  |
| --- |
| import pandas as pd  df = pd.read\_csv("airline\_passenger\_satisfaction.csv") |

Xóa các cột không sử dụng

|  |
| --- |
| df=df.drop(['flight\_distance','departure\_delay\_in\_minutes','arrival\_delay\_in\_minutes','age','customer\_class','type\_of\_travel'],axis=1) |

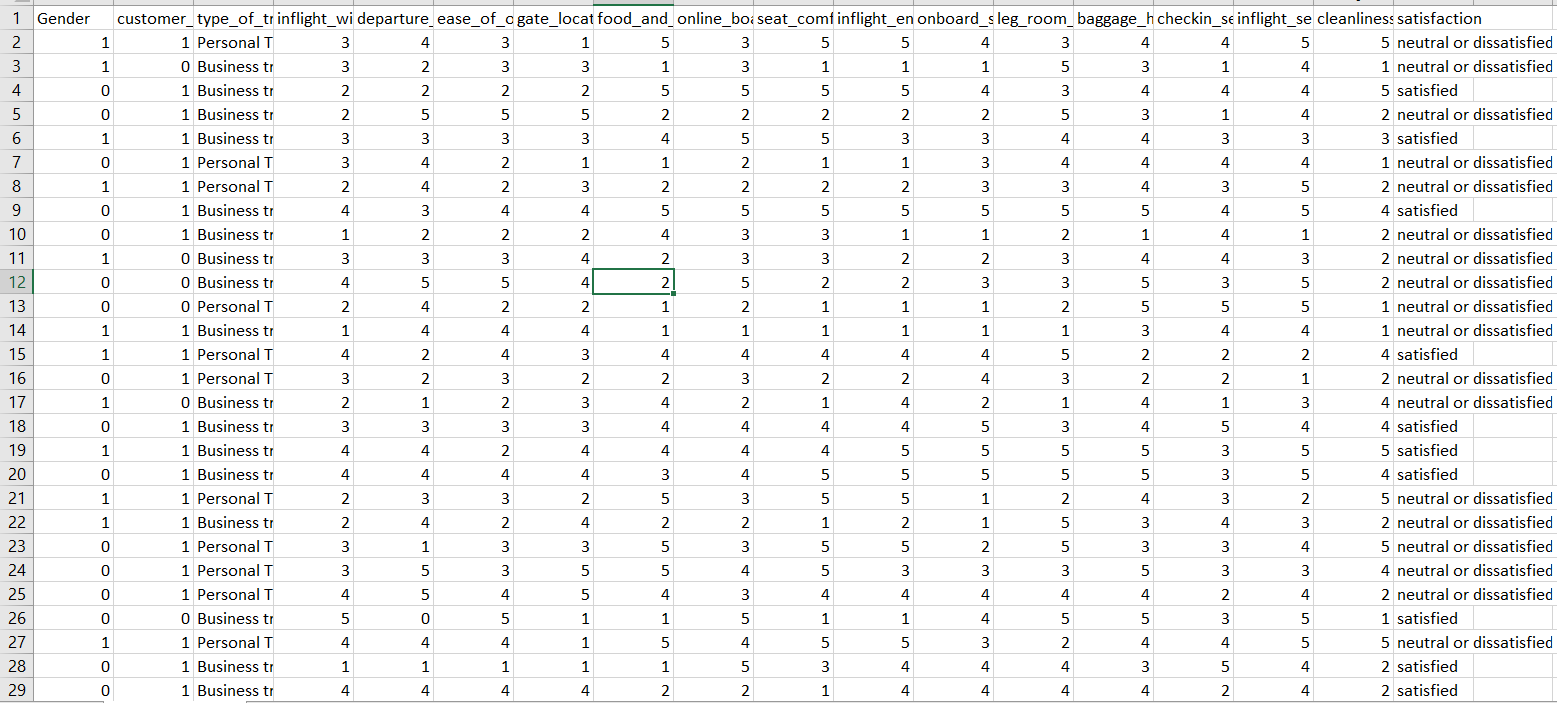
Chuyển đổi dữ liệu các cột customer\_type, Gender từ norminal thành numeric .

|  |
| --- |
| for index,value in enumerate(df['customer\_type']):      if value == "Loyal Customer":          df.at[index,'customer\_type']="1"      else:          df.at[index,'customer\_type']="0"  for index,value in enumerate(df['Gender']):      if value == "Male":          df.at[index,'Gender']="1"      else:          df.at[index,'Gender']="0"  // đổ dữ liệu  df['customer\_type']=df['customer\_type'].astype(int)  df['Gender']=df['Gender'].astype(int)  // show dữ liệu  df.info() |

Tạo ra file csv mới

|  |
| --- |
| df.to\_csv("Dataset\_ming.csv") |

Dữ liệu sau khi được xử lí:



**CHƯƠNG 2: KHAI PHÁ DỮ LIỆU BẰNG WEKA**

1. **Thêm dữ liệu vào Weka**

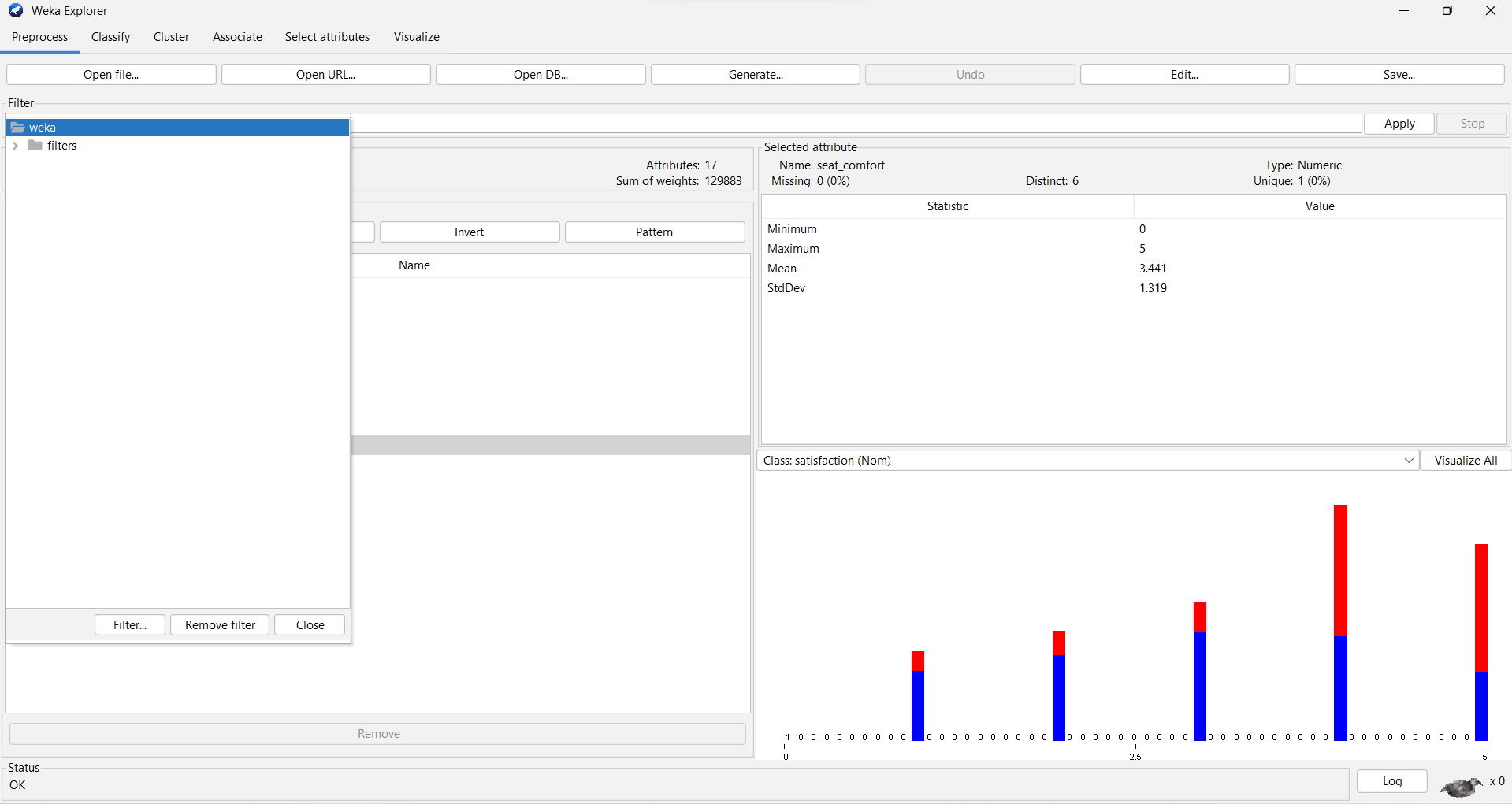
EXPLORER->OPEN FILE (Chọn tệp dữ liệu sau khi đã được xử lí)

A screenshot of a computer

Description automatically generated

1. Sử dụng thuật toán cây quyết định J48

Chia tập để train

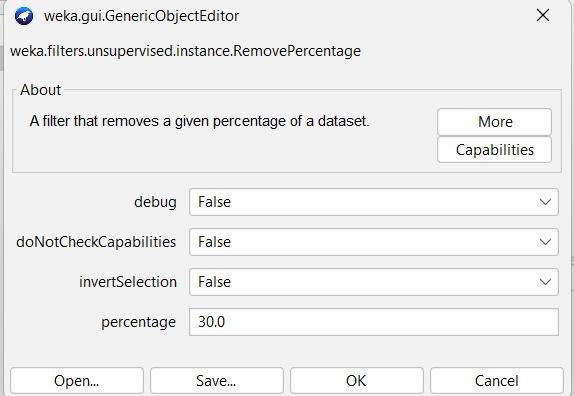


Sử dụng RemovePercentage để chia tập dữ liệu train

A screenshot of a computer

Description automatically generated with medium confidence

Chọn thuộc tính phù hợp



Và nhấn Apply

Graphical user interface, application

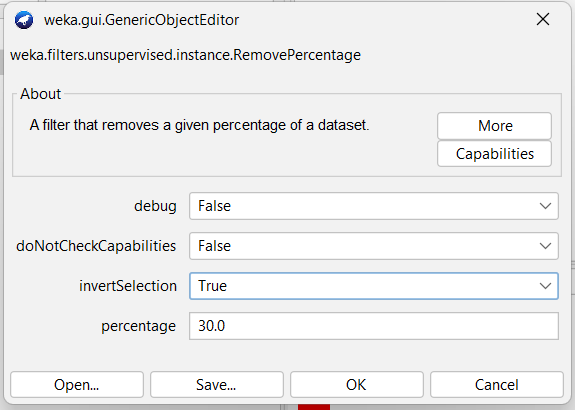
Description automatically generated

Lưu tập dữ liệu và để là train

**Graphical user interface, text, application

Description automatically generated**

Chia tập test



Sau khi nhấn apply

A screenshot of a computer

Description automatically generated with medium confidence

Lưu thành file test

Graphical user interface, text

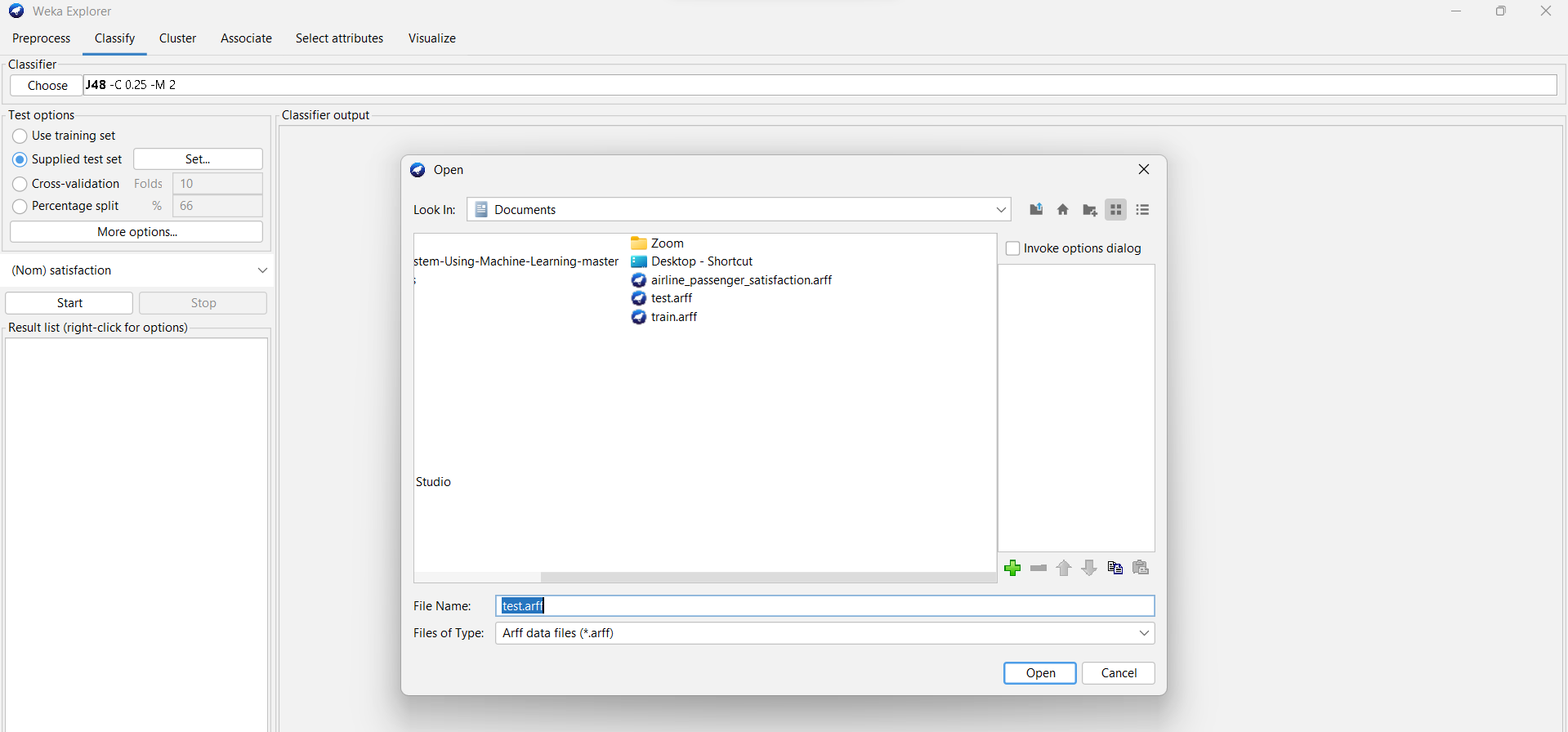
Description automatically generated

Ở tập train ta chọn vào Phân nhóm và chọn thuật toán Tree J48

Graphical user interface, application, Word

Description automatically generated

Chọn tập test



Ở test option chọn Cross-validation và phân loại theo biến Satisfacction

Graphical user interface, text, application

Description automatically generated

Sau khi chạy thuật toán xong ta thu được kết quả

A picture containing table

Description automatically generated

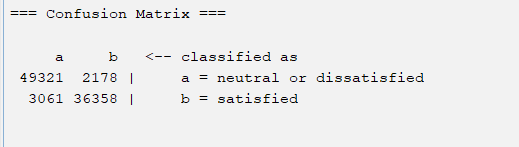
Ở đây ta thấy

Trường hợp phân nhóm đúng khá cao đạt 94.2327%

Trường hợp phân nhóm sai đạt 5.7623%



Nhận xét Confusion Matrix



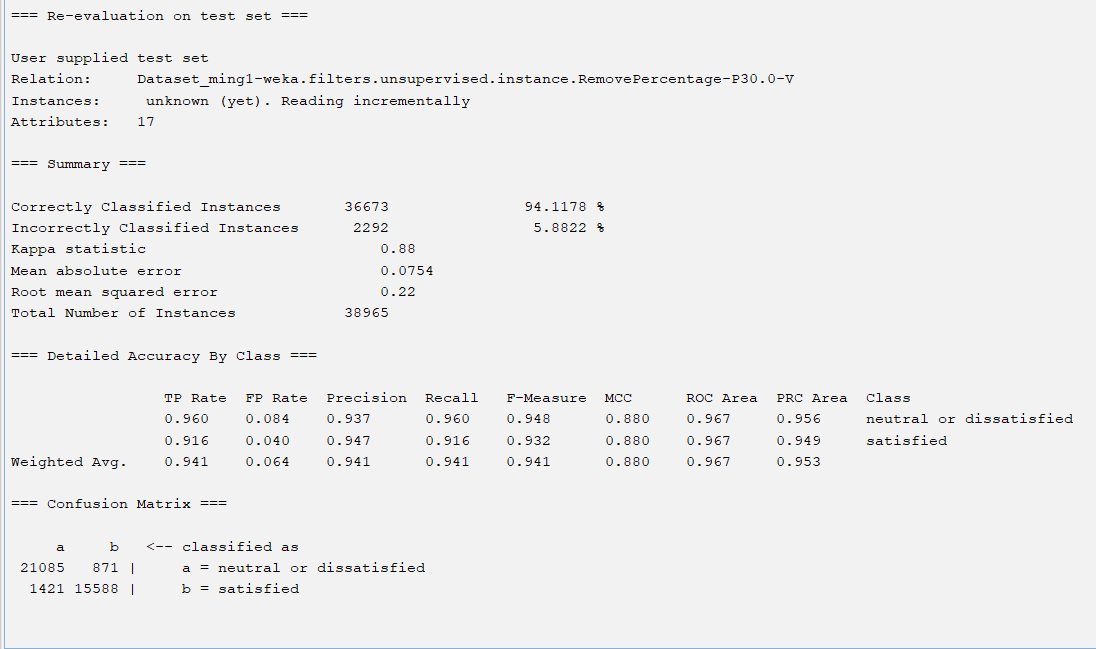
Thuật toán đã phân ra 2 nhóm với giá trị a là neutral or dissatisfiled và b là satisfiled. Ở ma trận này các giá trị ở vị trí đường chéo huyền là các giá trị phân loại đúng ví dụ như ở nhóm a có 49321 giá trị được phân loại đúng ở nhóm a và 36358 là giá trị phân loại đúng ở nhóm b. Các giá trị ngoài đường chéo này là các giá trị phân loại sai và đạt 5.7623 %

**Đánh giá model thông qua tập test**

Graphical user interface, application

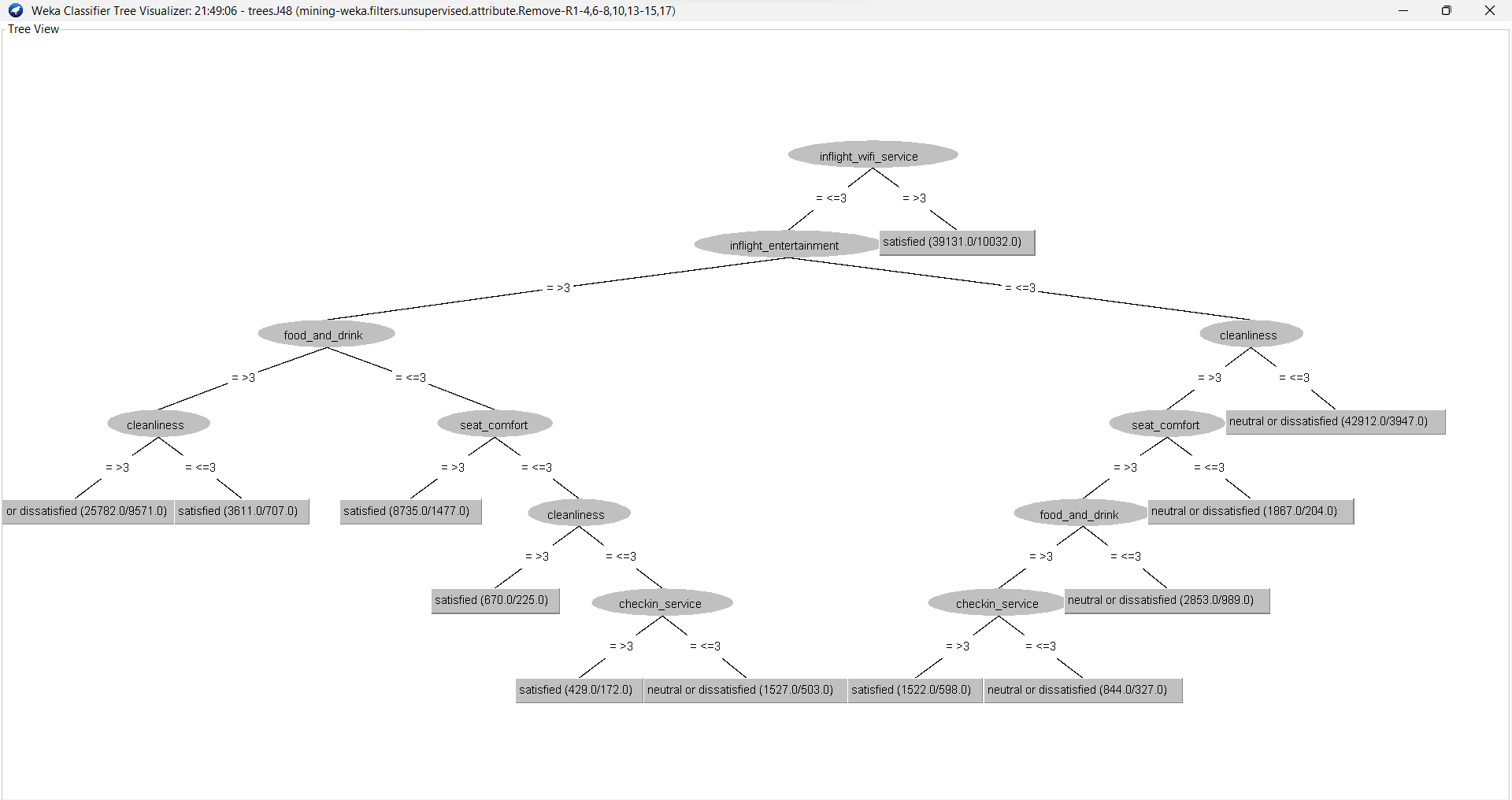
Description automatically generated

ở Model vừa xây dựng chọn chọn Re-evaluate model on current test set để đánh giá



Kết quả đánh giá tập test có 94,1178% dữ liệu được phân loại đúng và 5.8822 dữ

Để thấy được cây quyết định thì em sẽ tiền xữ lí lại dữ liệu và bỏ bớt các đánh giá chỉ để lại các đánh giá đặc trưng để xem cây được sinh ra



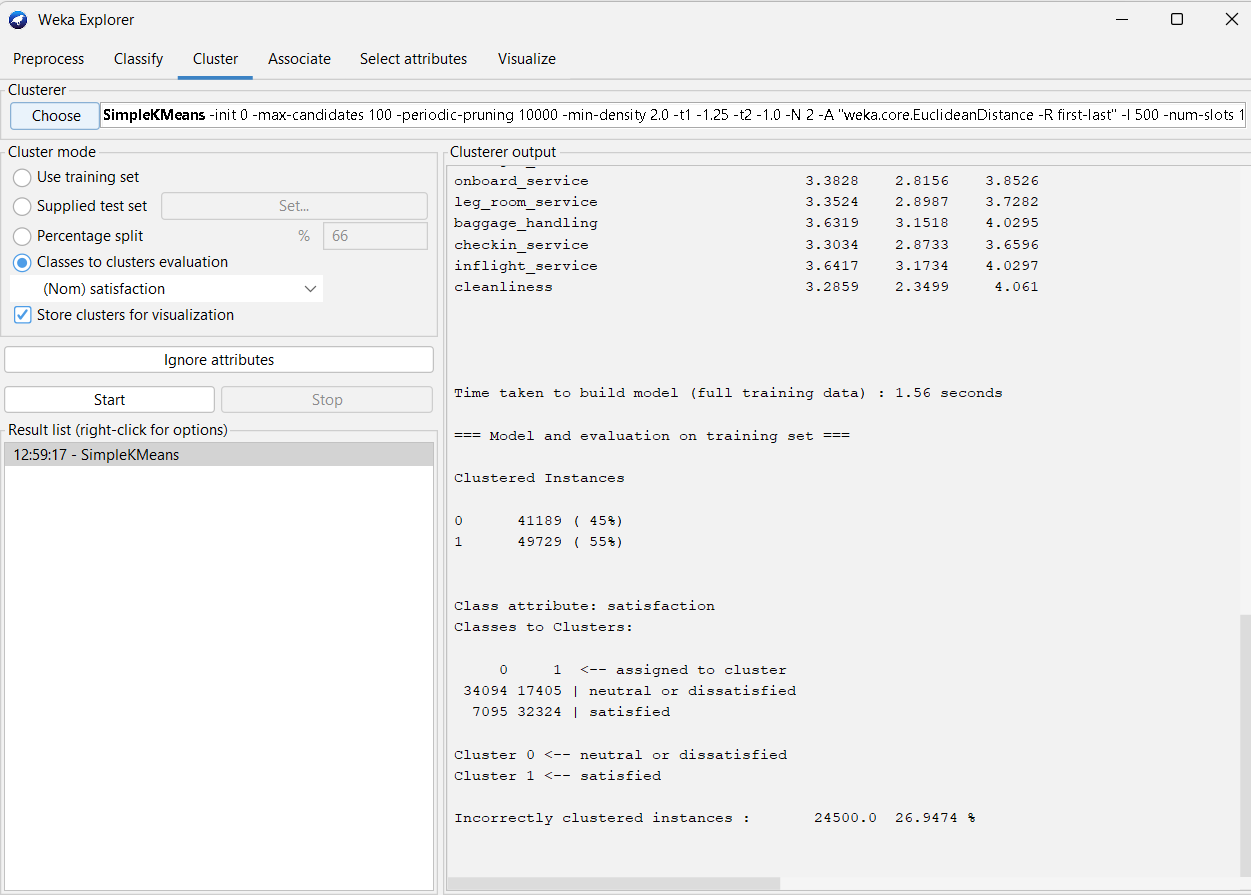
Đựa vào biểu đồ trên ta có thể rút qua các dự đoán khách hàng hài lòng như sau

1. Nếu Dịch vụ Wifi được chấm điểm > 3 thì sẽ thuộc nhóm hài lòng(39131,10032)
2. Nếu Wifi < =3, Giải trí > 3, Thức Ăn > 3 , Sạch sẽ <= 3 thì sẽ thuộc nhóm hài lòng(3611/707)
3. Nếu Wifi < =3, Giải trí > 3, Thức Ăn <=3, Chỗ ngồi > 3 thì sẽ thuộc nhóm hài lòng(8735/1477)
4. Nếu Wifi < =3, Giải trí > 3, Thức Ăn <=3, Chỗ ngồi <= 3, Sạch sẽ > 3 thì sẽ thuộc nhóm hài lòng(670/225)
5. Wifi < =3, Giải trí > 3, Thức Ăn <=3, Chỗ ngồi <= 3, Sạch sẽ <= 3, Check-in > 3 thì sẽ thuộc nhóm hài lòng(429/172)
6. Wifi < =3, Giải trí <= 3, Thức Ăn > 3, Chỗ ngồi > 3, Sạch sẽ > 3, Check-in > 3 thì sẽ thuộc nhóm hài lòng (1522/598)

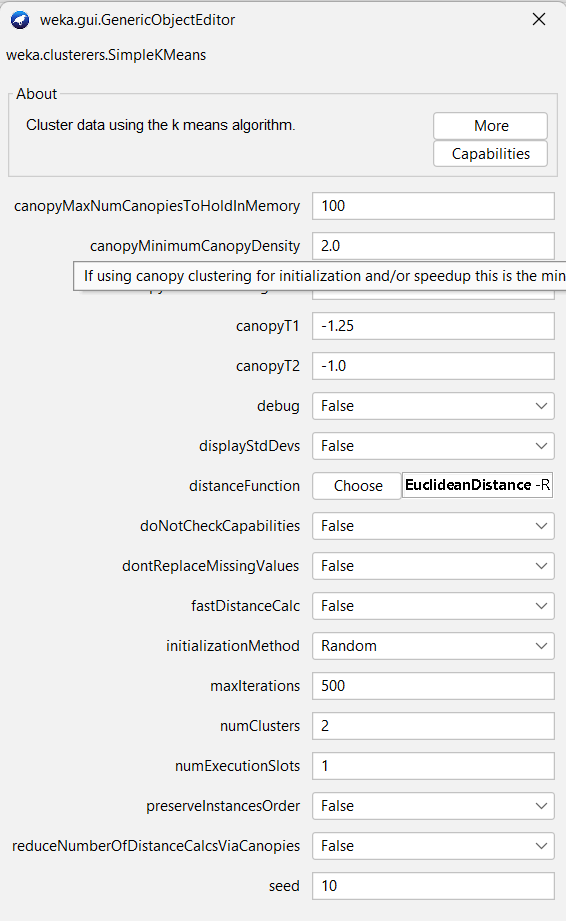
Kết luận:

Dịch vụ wifi là một dịch khá quan trọng chiếm được phần lớn sự hài lòng có thể là do việc lắp đặt wifi ở trên máy bay là điều khó khăn và rất tốn kém chi phí nên nếu dịch vụ wifi tốt được khách hàng hàng đánh giá cao thì rất dễ dẫn đến khách hàng đánh giá hài lòng. Bên cạnh đó ta thấy số lượng khách hàng hài lòng cao cũng nằm ở 2 dịch vụ là giải trí, chổ ngồi và sự sạch sẽ đây 3 yếu tố khá chủ chốt đánh vào trải nghiệm trực tiếp của khách hàng trong chuyến bay

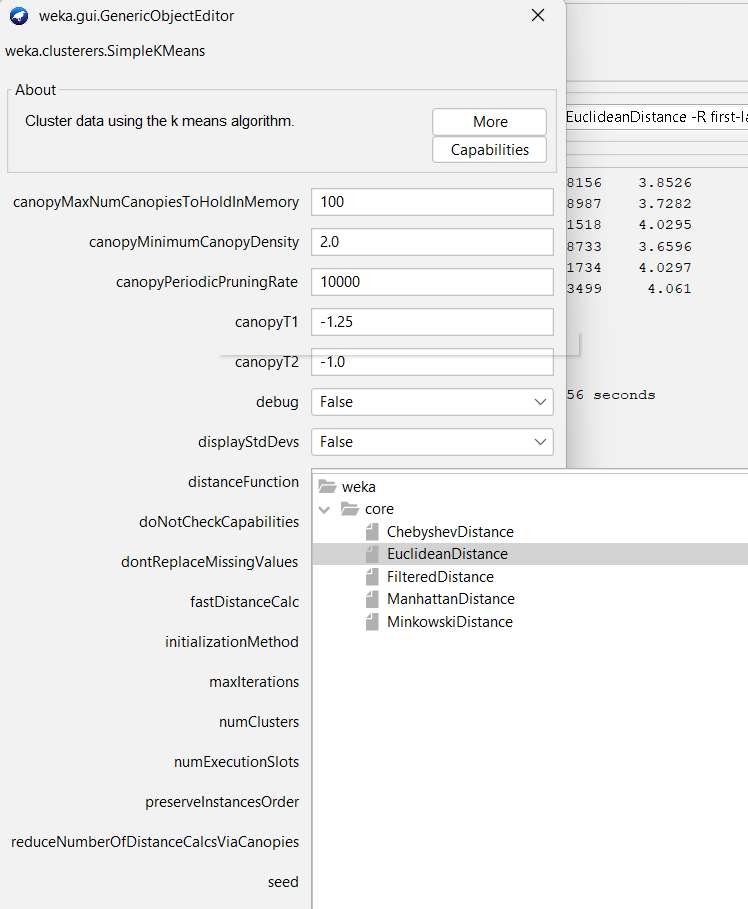
## Sử dụng thuật toán phân cụm K-mean



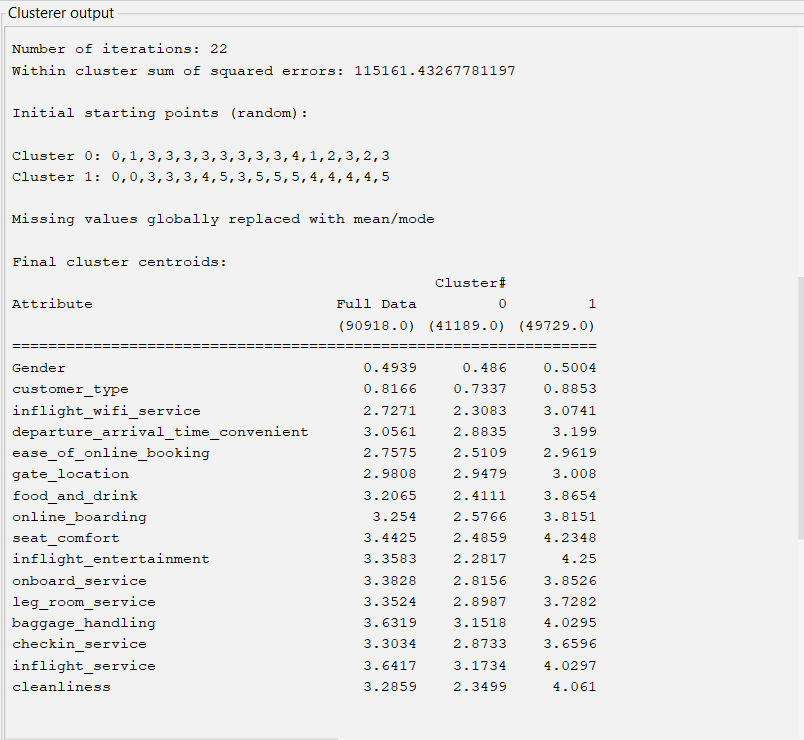
Set các thuộc tính cho thuật toán



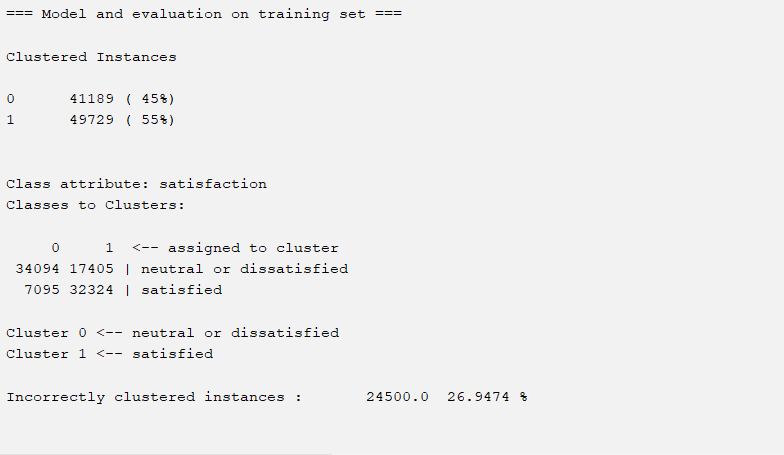
Chọn cách tính khoảng cách là EuclideanDistance



Sau đó khởi chạy thuật toán ta được các thông tin

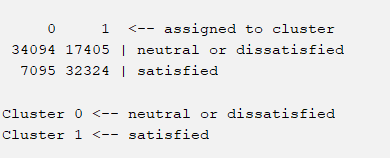


So sánh đánh giá



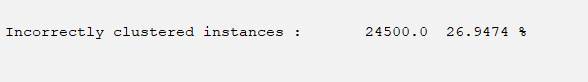
Ở trên model đã phân ra chúng ta thành 2 cụm 0 và 1 với Cluster 0 là neutral or dissatisfiled chiếm 45% và Cluster 1 là satisfiled với 55%

Trong đó dựa vào Confusion matrix ta thấy



ở Cụm có 34094 dữ liệu đúng với Cụm 0 (neutral or dissatisfied )và có 32324 là đúng với cụm(satisfied)

các dữ liệu còn lại được phân cụm nhưng bị sai trường hợp so với đánh giá chiếm tỉ lệ 26.9474%



Kết luận:

Thuật toán phân cụm không phù hợp cho bộ dữ liệu này bởi vì các trường hợp bị phân cụm sai còn khá cao khoảng 27%

## Association Rule

Sau khi chạy thuật toán Apriori trên tập dữ liệu ta thu được các luật kết hợp sau:

A picture containing text

Description automatically generated

Trong đó chọn lọc các luật đáng chú ý như:

1. Luật số 4 cho thấy Đồ ăn <= 3, Chổ ngồi <=3 thì có 94% sẽ đánh giá sạch sẽ <=3
2. Luật số 10 cho thấy Sạch sẽ <=3 , Khách hàng không hài lòng thì khả năng tới 92% khách hàng đã đánh giá <= 3 ở dịch vụ giải trí
3. Luật số 7 cho thấy nếu chổ ngồi <= 3, giải trí <=3 thì có tới 93% khách hàng sẽ đánh giá đồ ăn <=3
4. Luật số 5 cho thấy nếu khách hàng đánh giá thức ăn <=3, sạch sẽ <=3 và khách hàng không hài lòng thì khả năng 94% khách hàng đã đánh giá giải trí <=3

Kết luận:

Khi thức ăn và chỗ ngồi bị đánh giá thấp có thể không phải do thức ăn không ngon hay chổ ngồi không thoải mái mà do khâu vệ sinh tệ dẫn đến khách hàng đánh giá thấp ở cả 2 dịch vụ này. Khi sự sạch sẽ bị đánh giá thấp thì có khả năng cao bị đánh giá thấp ở các dịch vụ còn lại. Cho nên Hãng hàng không nên chú ý ở khâu vệ sinh trước khi đón tiếp khách hàng để họ có một trải nghiệm về chuyến bay thực sự tuyệt vời nhất.