# TRƯỜNG ĐẠI HỌC BÁCH KHOA
# ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

---


# BÁO CÁO BÀI TẬP NHÓM
# Đề tài: Phân tích tệp dữ liệu Las Vegas Strip

Môn học: Chuyên đề nghiên cứu khoa học dữ liệu ứng dụng

GVHD: PGS.TS Nguyễn Mạnh Tuân

Học viên:

| | |
|---|---|
| Đinh Thanh Phong | 2270243 |
| Nguyễn Thị Hạnh Thảo | 2270753 |
| Nguyễn Công Trực | 2270200 |
| Hồ Duy Đạt Phúc | 2051172 |

# Mục lục

# 1. Đánh giá chất lượng bộ dữ liệu Lasvegas Tripadvisor:

Bộ dữ liệu cung cấp các thông tin tổng kết từ những đánh giá trực tuyến của khách du lịch đã trải nghiệm thực tế và sử dụng dịch vụ tại các khách sạn nằm tại Lasvegas. Sự phát triển của ngành du lịch và các dịch vụ đặt phòng khiến cho sự cạnh tranh ngày càng gay gắt, ngành khách sạn phải đổi mới cách kinh doanh. Sự hài lòng và sự trung thành của khách hàng hết sức quan trọng đối với khách sạn. Do đó việc phân tích dữ liệu để hiểu được trải nghiệm của khách hàng với các dịch vụ được cung cấp trở thành một trong những xu hướng chính của lĩnh vực khách sạn. Từ đó đem đến góc nhìn tổng quát đa chiều cho người quản lý để đổi mới chất lượng các dịch vụ của khách sạn.

Nghiên cứu bộ dữ liệu có 504 dữ liệu đánh giá trực tuyến 21 khách sạn ở Las Vegas năm 2015 với các tiêu chí cụ thể bao gồm 20 thuộc tính: số trải nghiệm, mùa nghỉ dưỡng, điểm đánh giá, xếp hạng sao của khách sạn, cơ sở vật chất của khách sạn, các dịch vụ giải trí, thể thao của khách sạn, thông tin khách hàng như nhóm người thuộc quốc gia, khu vực, đối tượng khách hàng là doanh nhân, gia đình, bạn bè, độc thân hay những cặp đôi, số năm là thành viên, thời gian đánh giá,...Đối với các thuộc tính số phòng khách sạn, khách hàng tới từ khu vực, số năm là thành viên, ngày, tháng cung cấp đánh giá xuất hiện giá trị dữ liệu bị thiếu. Tuy nhiên vấn đề quan trọng của tập dữ liệu tập trung vào việc đánh giá chất lượng dịch vụ và khả năng đáp ứng dịch vụ của các khách

sạn thông qua các thuộc tính cụ thể có tính năng định lượng và phân loại từ các dữ liệu từ các nguồn dữ liệu chi tiết thu thập được từ người đánh giá như tiện ích hồi bơi, phòng tập, tennis, spa, casino và truy cập internet miễn phí.

## 2. Mục tiêu phân tích:

Mục tiêu phân tích bộ dữ liệu nhằm nỗ lực khai thác thông tin để trả lời các câu hỏi nghiên cứu sau: Có thể dự đoán điểm số của một bài đánh giá khách sạn trực tuyến chỉ bằng cách sử dụng dữ liệu định lượng đầu vào không? Các tính năng ảnh hưởng đến hầu hết các điểm đánh giá trong khách sạn là gì? Làm thế nào để mỗi tính năng đó ảnh hưởng đến điểm số và kiến thức này có thể hữu ích cho các nhà quản lý khách sạn không? Bộ dữ liêu đã đóng góp các thông tin chính với mục tiêu:

• Tạo mô hình dự đoán điểm đánh giá dựa trên các đặc điểm định lượng của người dùng/người đánh giá và khách sạn, cũng như khoảng thời gian lưu trú cụ thể;

• Góp phần nghiên cứu phản hồi của khách hàng và đánh giá trực tuyến bằng cách cung cấp một cách tiếp cận mới về dữ liệu được sử dụng, các tính năng định lượng, trái ngược với các phân tích phổ biến nhất về chính văn bản đánh giá;

• Hiểu cách người dùng bị ảnh hưởng vốn có bởi các tính năng của khách sạn khi gửi điểm số bên cạnh nhận xét bằng văn bản trên các nền tảng trực tuyến, chẳng hạn như TripAdvisor.

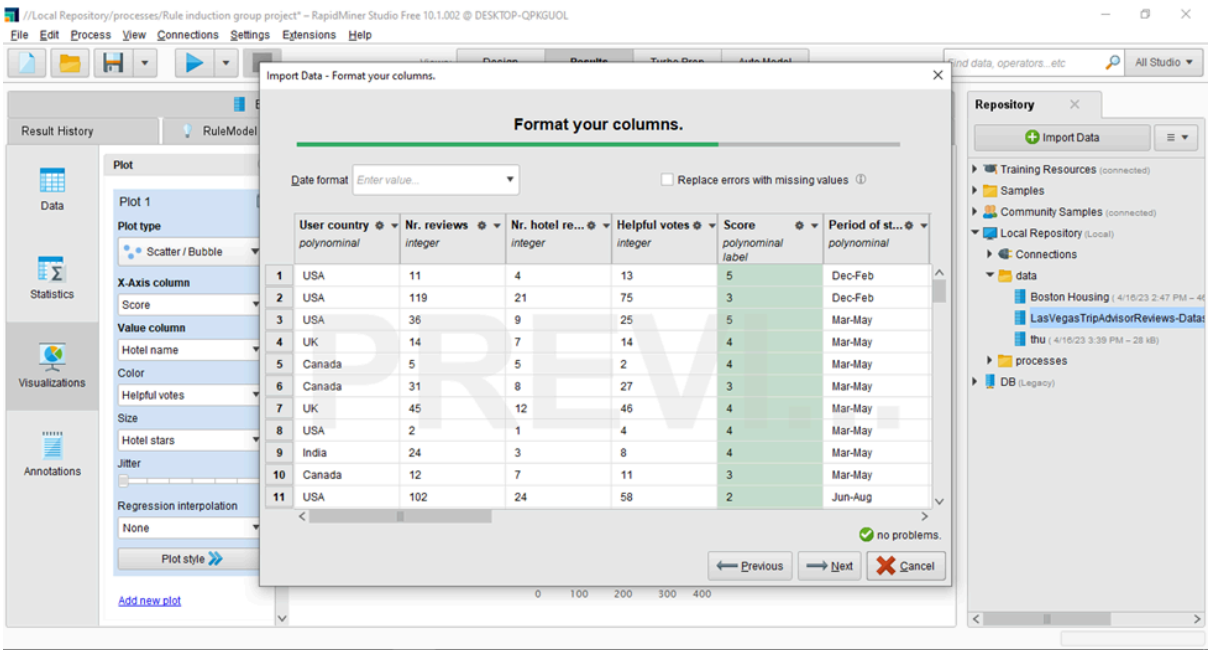
## 3. Xử lý bộ dữ liệu sử dụng 3 giải thuật Rule Induction, Decision tree và k-NN.

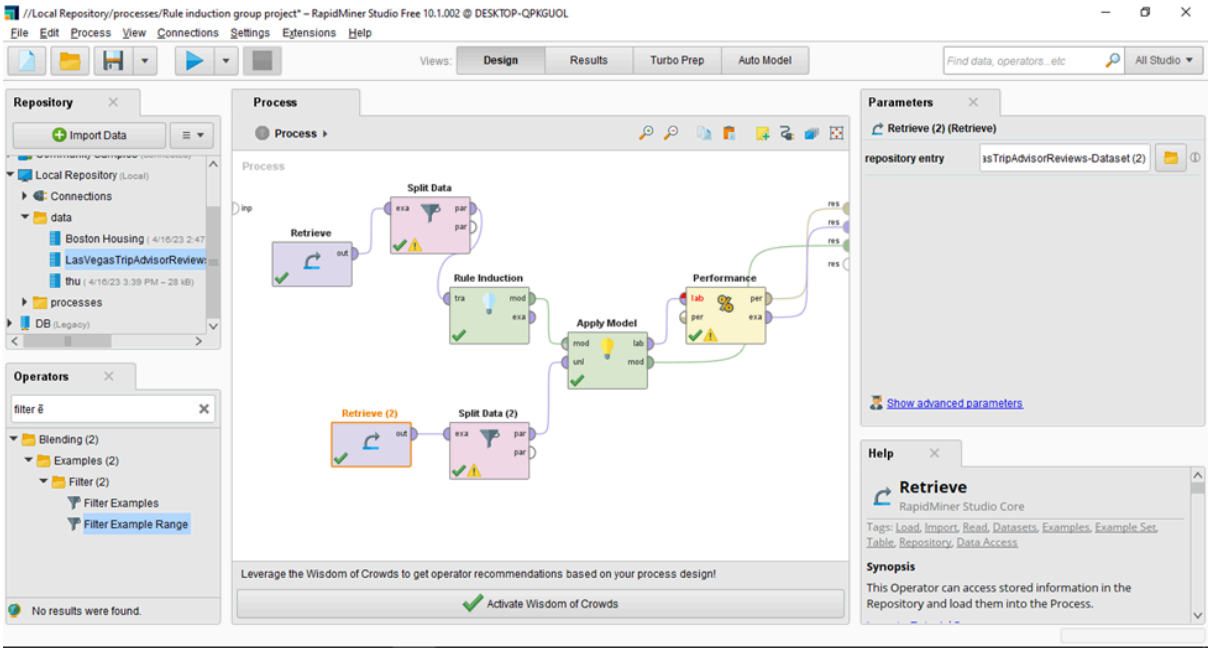
Nhóm dùng 3 nhiệm giải thuật Rule Induction, Decision tree và k-NN để thực hiện phân tích và so sánh kết quả (performance) của mô hình bằng cách sử dụng các công cụ sử dụng độ đo confusion matrix, precision, recall, accuracy.

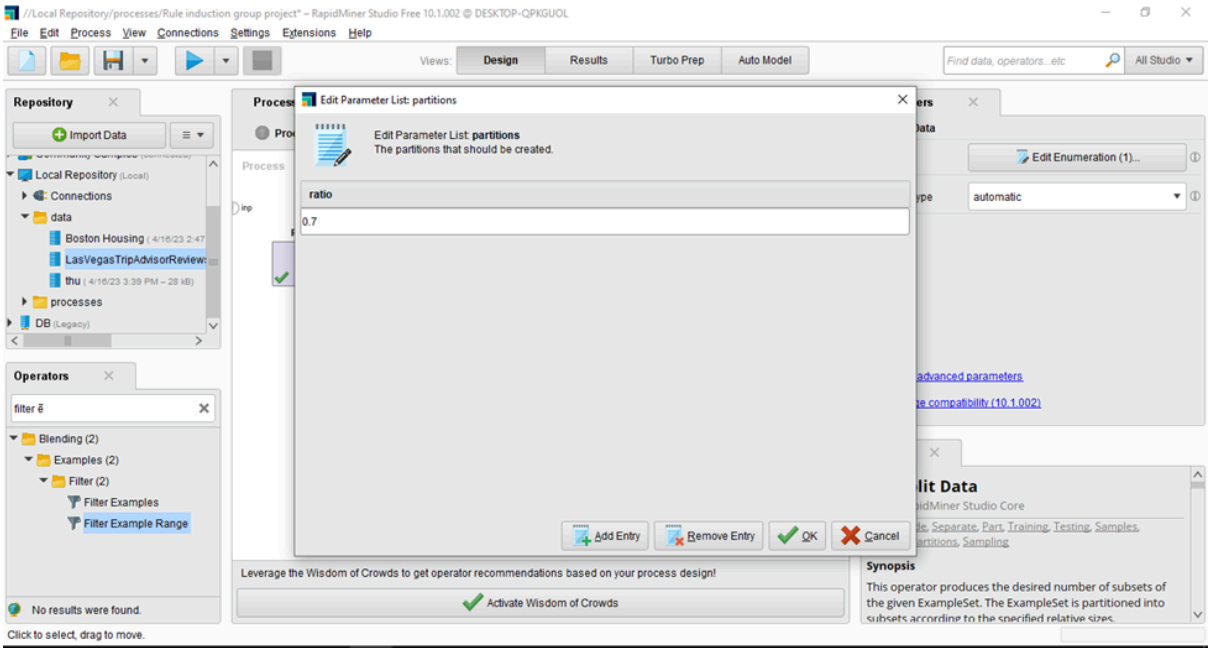
Thuộc tính được gắn nhãn label được chọn chung cho cả 3 giải thuật là thuộc tính score.
Để đạt được kết quả tốt nhất trong mỗi giải thuật, nhóm thực hiện lựa chọn các attributes khác nhau và chia tập train/test khác nhau ở mỗi giải thuật, nhằm đạt được kết quả tốt nhất cho bài toán.

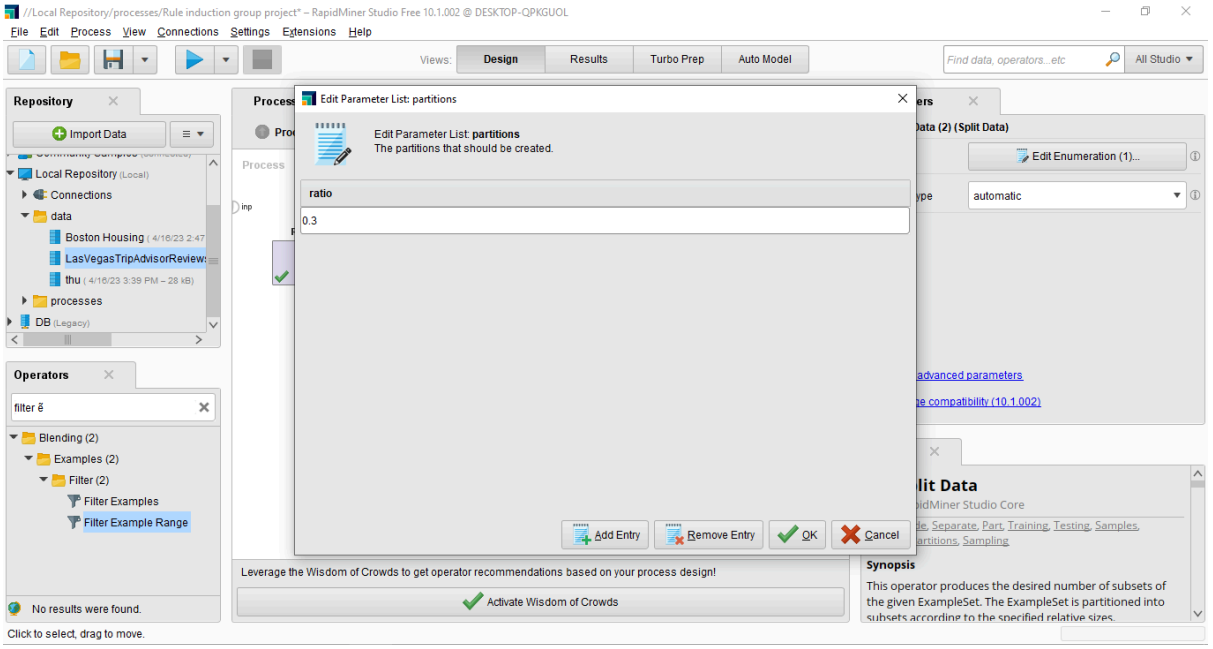
## 3.1) Giải thuật: Rule Induction.

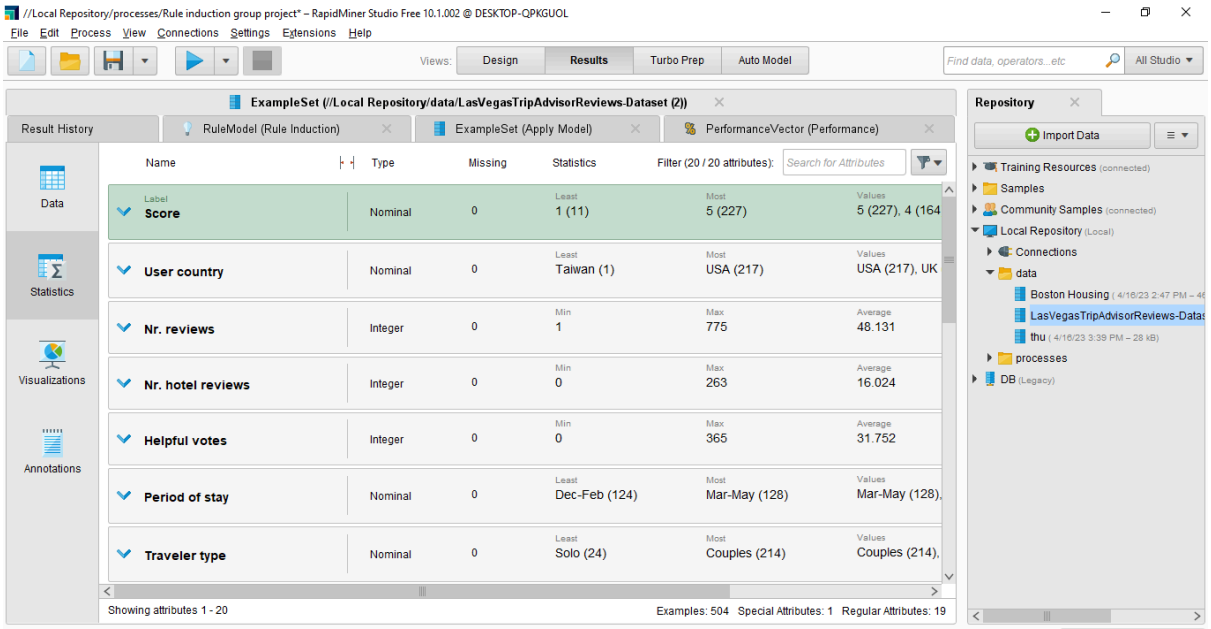
- Thuộc tính “score” được chọn gắn label.
- Sử dụng tất cả các thuộc tính khác làm attribute.
- Tỷ lệ phân chia Train/Test = 0.7/0.3

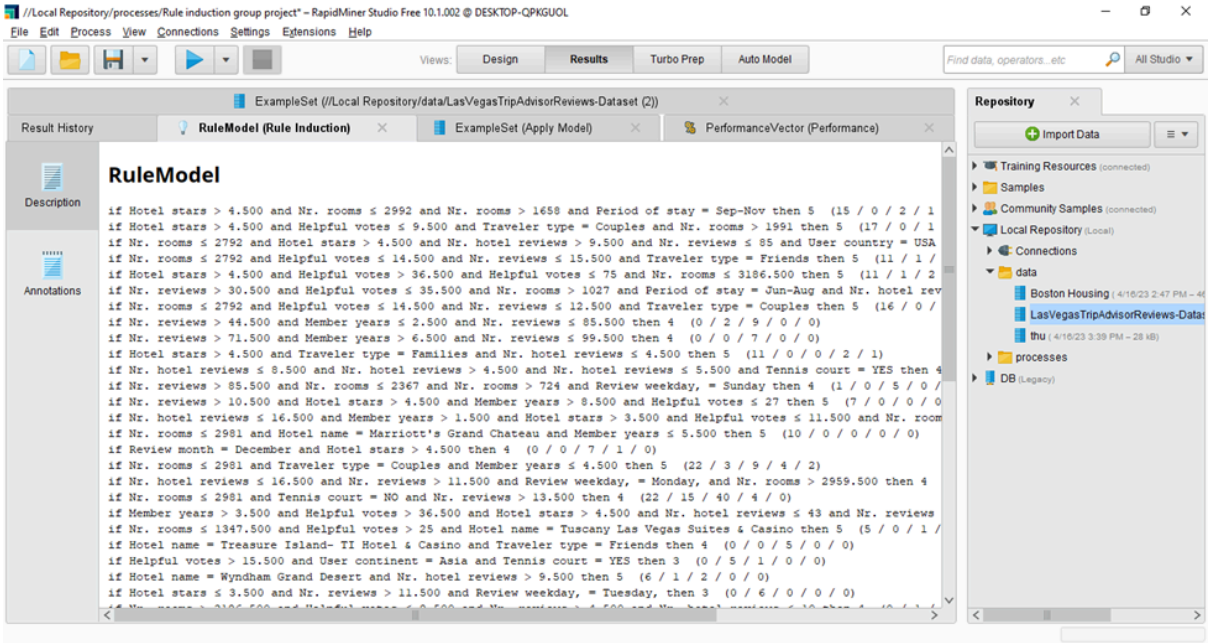


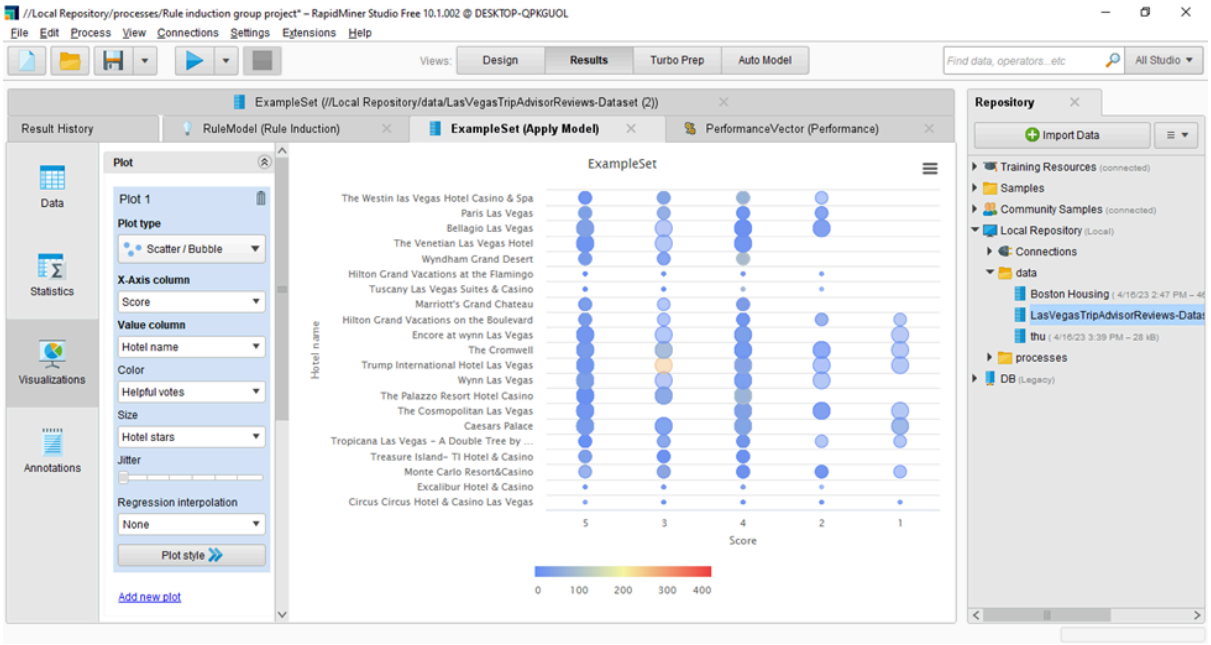

//Local Repository/processes/Rule induction group project* – RapidMiner Studio Free 10.1.002 @ DESKTOP-QPKGUOL
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Repository
Import Data
Local Repository (Local)
Connections
data
Boston Housing ( 4/16/23 2:47
LasVegasTripAdvisorReview
thu ( 4/16/23 3:39 PM – 28 kB)
processes
DB (Legacy)
Operators
filter ē
Blending (2)
Examples (2)
Filter (2)
Filter Examples
Filter Example Range
No results were found.
Process
Process
Split Data
Retrieve
Rule Induction
Performance
Apply Model
Retrieve (2)
Split Data (2)
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Retrieve (2) (Retrieve)
repository entry
sTripAdvisorReviews-Dataset (2)
Show advanced parameters
Help
Retrieve
RapidMiner Studio Core
Tags: Load, Import, Read, Datasets, Examples, Example Set, Table, Repository, Data Access
Synopsis
This Operator can access stored information in the Repository and load them into the Process.

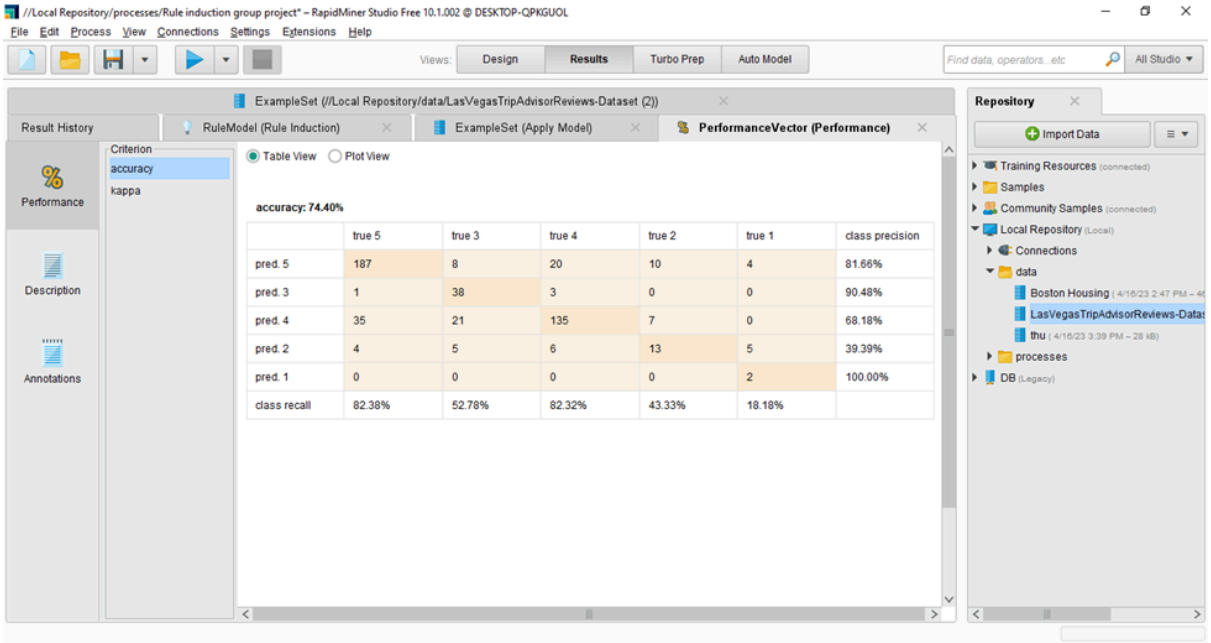

//Local Repository/processes/Rule induction group project* – RapidMiner Studio Free 10.1.002 @ DESKTOP-QPKGUOL
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Edit Parameter List: partitions
Edit Parameter List: partitions
The partitions that should be created.
ratio
0.7
Add Entry
Remove Entry
OK
Cancel
Edit Enumeration (1)...
automatic
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Click to select, drag to move.


//Local Repository/processes/Rule induction group project* – RapidMiner Studio Free 10.1.002 @ DESKTOP-QPKGUOL
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Repository
Import Data
Local Repository (Local)
Connections
data
Boston Housing ( 4/16/23 2:47
LasVegasTripAdvisorReviews
thu ( 4/16/23 3:39 PM – 28 kB)
processes
DB (Legacy)
Operators
filter
Blending (2)
Examples (2)
Filter (2)
Filter Examples
Filter Example Range
No results were found.
Click to select, drag to move.
Edit Parameter List: partitions
Edit Parameter List: partitions
The partitions that should be created.
ratio
0.3
Add Entry
Remove Entry
OK
Cancel
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Edit Enumeration (1)...
automatic
advanced parameters
compatibility (10.1.002)
Synopsis
This operator produces the desired number of subsets of the given ExampleSet. The ExampleSet is partitioned into


//Local Repository/processes/Rule induction group project* – RapidMiner Studio Free 10.1.002 @ DESKTOP-QPKGUOL
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
ExampleSet (//Local Repository/data/LasVegasTripAdvisorReviews-Dataset (2))
Result History
RuleModel (Rule Induction)
ExampleSet (Apply Model)
PerformanceVector (Performance)
Data
Statistics
Visualizations
Annotations
Name
Type
Missing
Statistics
Filter (20 / 20 attributes):
Search for Attributes
Score
Nominal
0
Least 1 (11)
Most 5 (227)
Values 5 (227), 4 (164
User country
Nominal
0
Least Taiwan (1)
Most USA (217)
Values USA (217), UK
Nr. reviews
Integer
0
Min 1
Max 775
Average 48.131
Nr. hotel reviews
Integer
0
Min 0
Max 263
Average 16.024
Helpful votes
Integer
0
Min 0
Max 365
Average 31.752
Period of stay
Nominal
0
Least Dec-Feb (124)
Most Mar-May (128)
Values Mar-May (128),
Traveler type
Nominal
0
Least Solo (24)
Most Couples (214)
Values Couples (214),
Showing attributes 1 - 20
Examples: 504 Special Attributes: 1 Regular Attributes: 19
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
Boston Housing
LasVegasTripAdvisorReviews-Datas
thu ( 4/16/23 3:39 PM – 28 kB)
processes
DB (Legacy)


//Local Repository/processes/Rule induction group project* – RapidMiner Studio Free 10.1.002 @ DESKTOP-QPKGUOL
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
ExampleSet (//Local Repository/data/LasVegasTripAdvisorReviews-Dataset (2))
Result History
RuleModel (Rule Induction)
ExampleSet (Apply Model)
PerformanceVector (Performance)
Description
Annotations
RuleModel
if Hotel stars > 4.500 and Nr. rooms ≤ 2992 and Nr. rooms > 1658 and Period of stay = Sep-Nov then 5 (15 / 0 / 2 / 1
if Hotel stars > 4.500 and Helpful votes ≤ 9.500 and Traveler type = Couples and Nr. rooms > 1991 then 5 (17 / 0 / 1
if Nr. rooms ≤ 2792 and Hotel stars > 4.500 and Nr. hotel reviews > 9.500 and Nr. reviews ≤ 85 and User country = USA
if Nr. rooms ≤ 2792 and Helpful votes ≤ 14.500 and Nr. reviews ≤ 15.500 and Traveler type = Friends then 5 (11 / 1 /
if Hotel stars > 4.500 and Helpful votes > 36.500 and Helpful votes ≤ 75 and Nr. rooms ≤ 3186.500 then 5 (11 / 1 / 2
if Nr. reviews > 30.500 and Helpful votes ≤ 35.500 and Nr. rooms > 1027 and Period of stay = Jun-Aug and Nr. hotel rev
if Nr. rooms ≤ 2792 and Helpful votes ≤ 14.500 and Nr. reviews ≤ 12.500 and Traveler type = Couples then 5 (16 / 0 /
if Nr. reviews > 44.500 and Member years ≤ 2.500 and Nr. reviews ≤ 85.500 then 4 (0 / 2 / 9 / 0 / 0)
if Nr. reviews > 71.500 and Member years > 6.500 and Nr. reviews ≤ 99.500 then 4 (0 / 0 / 7 / 0 / 0)
if Hotel stars > 4.500 and Traveler type = Families and Nr. hotel reviews ≤ 4.500 then 5 (11 / 0 / 0 / 2 / 1)
if Nr. hotel reviews ≤ 8.500 and Nr. hotel reviews > 4.500 and Nr. hotel reviews ≤ 5.500 and Tennis court = YES then 4
if Nr. reviews > 85.500 and Nr. rooms ≤ 2367 and Nr. rooms > 724 and Review weekday, = Sunday then 4 (1 / 0 / 5 / 0 /
if Nr. reviews > 10.500 and Hotel stars > 4.500 and Member years > 8.500 and Helpful votes ≤ 27 then 5 (7 / 0 / 0 / 0
if Nr. hotel reviews ≤ 16.500 and Member years > 1.500 and Hotel stars > 3.500 and Helpful votes ≤ 11.500 and Nr. room
if Nr. rooms ≤ 2981 and Hotel name = Marriott's Grand Chateau and Member years ≤ 5.500 then 5 (10 / 0 / 0 / 0 / 0)
if Review month = December and Hotel stars > 4.500 then 4 (0 / 0 / 7 / 1 / 0)
if Nr. rooms ≤ 2981 and Traveler type = Couples and Member years ≤ 4.500 then 5 (22 / 3 / 9 / 4 / 2)
if Nr. hotel reviews ≤ 16.500 and Nr. reviews > 11.500 and Review weekday, = Monday, and Nr. rooms > 2959.500 then 4
if Nr. rooms ≤ 2981 and Tennis court = NO and Nr. reviews > 13.500 then 4 (22 / 15 / 40 / 4 / 0)
if Member years > 3.500 and Helpful votes > 36.500 and Hotel stars > 4.500 and Nr. hotel reviews ≤ 43 and Nr. reviews
if Nr. rooms ≤ 1347.500 and Helpful votes > 25 and Hotel name = Tuscany Las Vegas Suites & Casino then 5 (5 / 0 / 1 /
if Hotel name = Treasure Island- TI Hotel & Casino and Traveler type = Friends then 4 (0 / 0 / 5 / 0 / 0)
if Helpful votes > 15.500 and User continent = Asia and Tennis court = YES then 3 (0 / 5 / 1 / 0 / 0)
if Hotel name = Wyndham Grand Desert and Nr. hotel reviews > 9.500 then 5 (6 / 1 / 2 / 0 / 0)
if Hotel stars ≤ 3.500 and Nr. reviews > 11.500 and Review weekday, = Tuesday, then 3 (0 / 6 / 0 / 0 / 0)
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
Boston Housing
LasVegasTripAdvisorReviews-Datas
thu ( 4/16/23 3:39 PM – 28 kB)
processes
DB (Legacy)



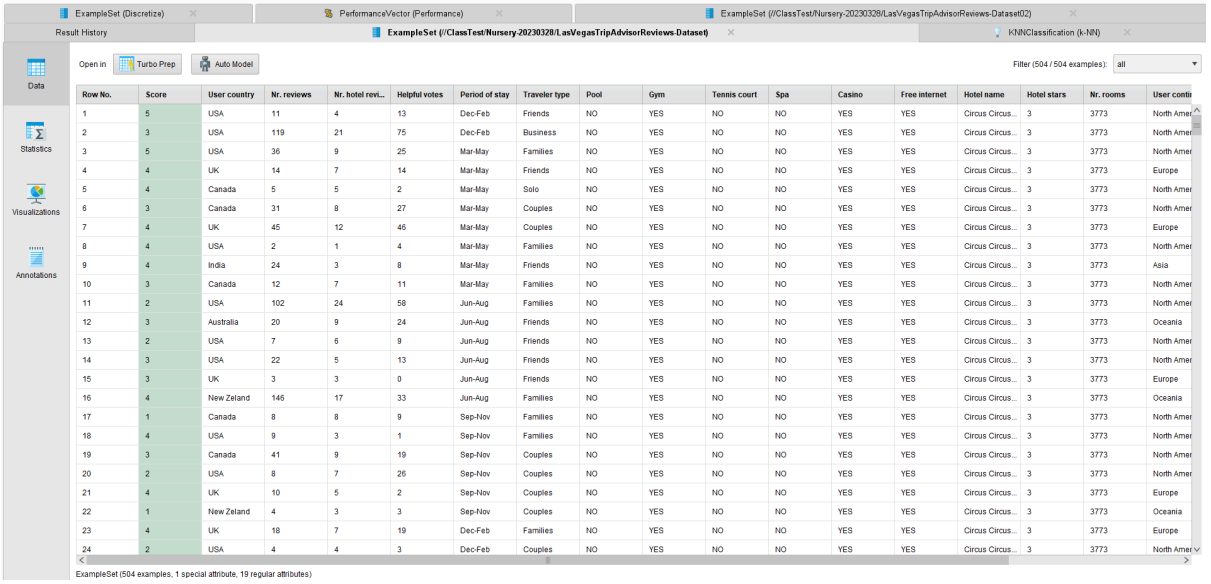

Kết quả giải thuật Rule Induction:

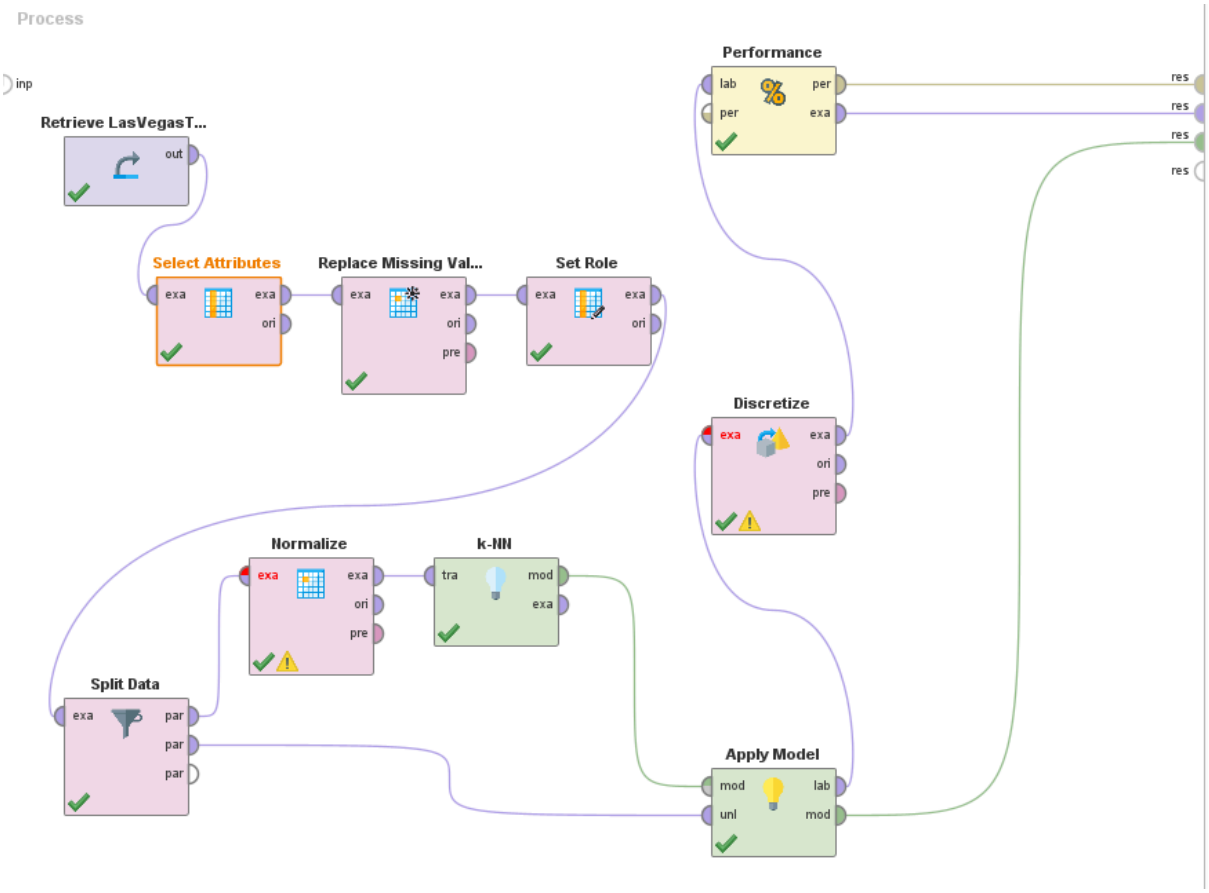
- Recall, Precision: như bảng đính kèm ở trên.
- Accuracy: 74,40%

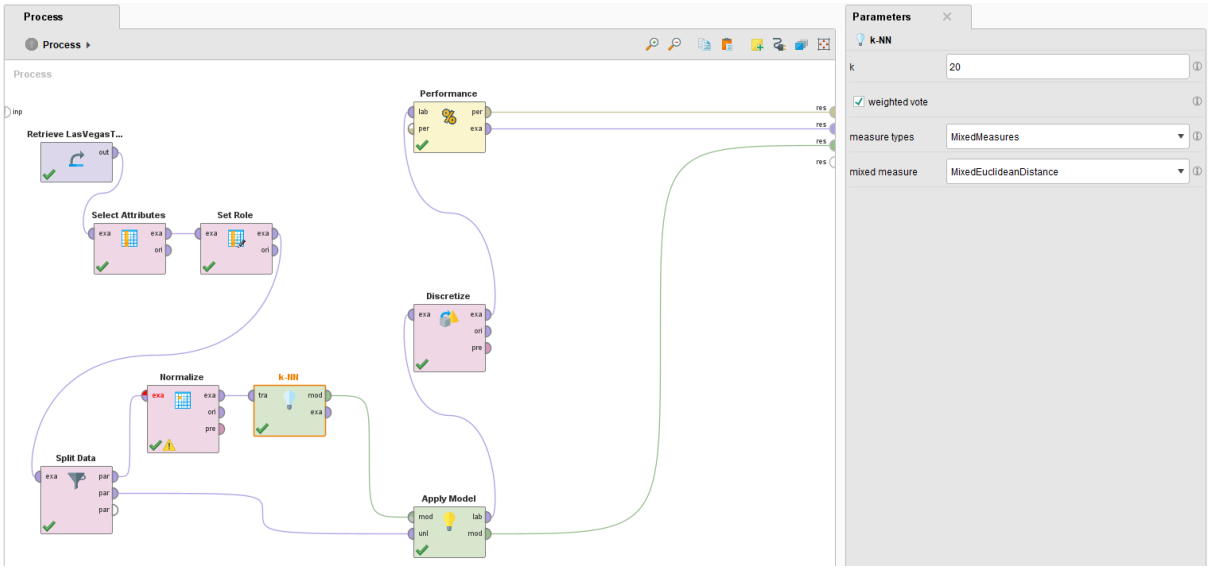
## 3.2) Giải thuật: k-NN.

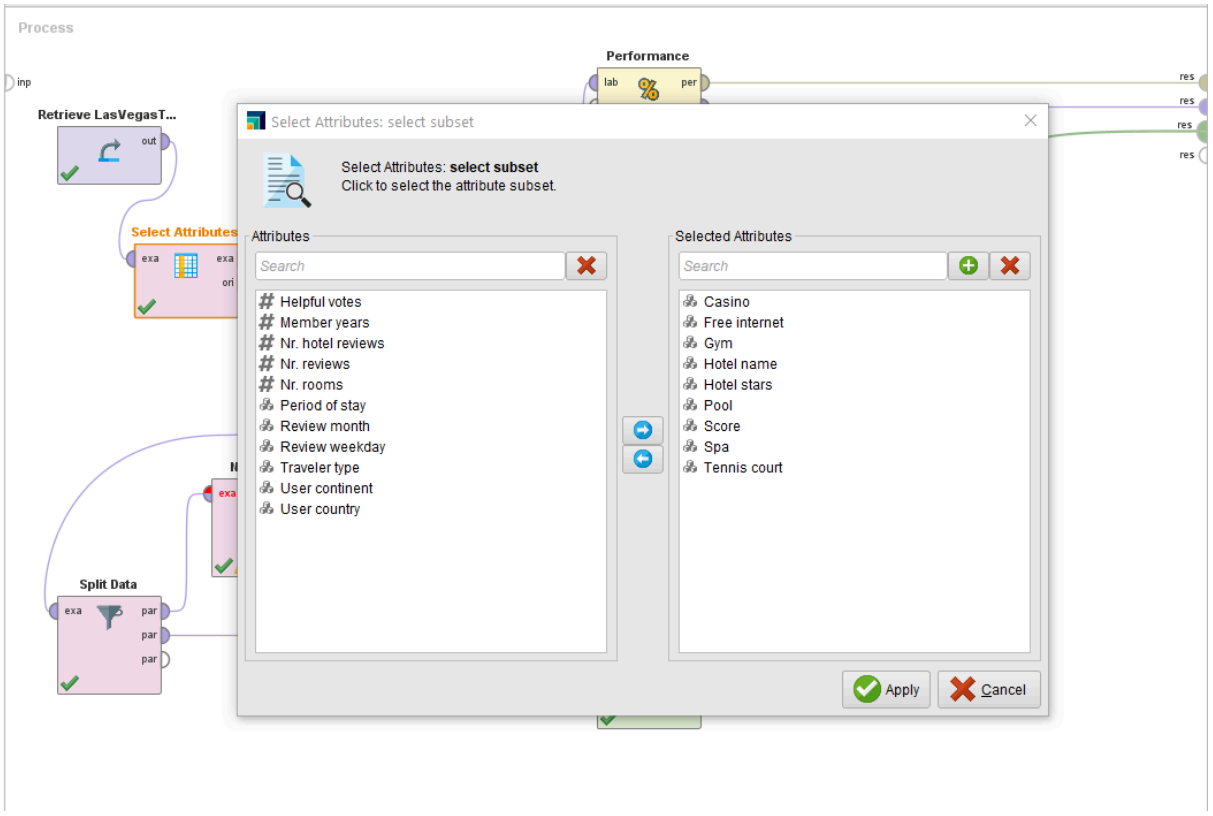
- Thuộc tính “score” được chọn gắn label.

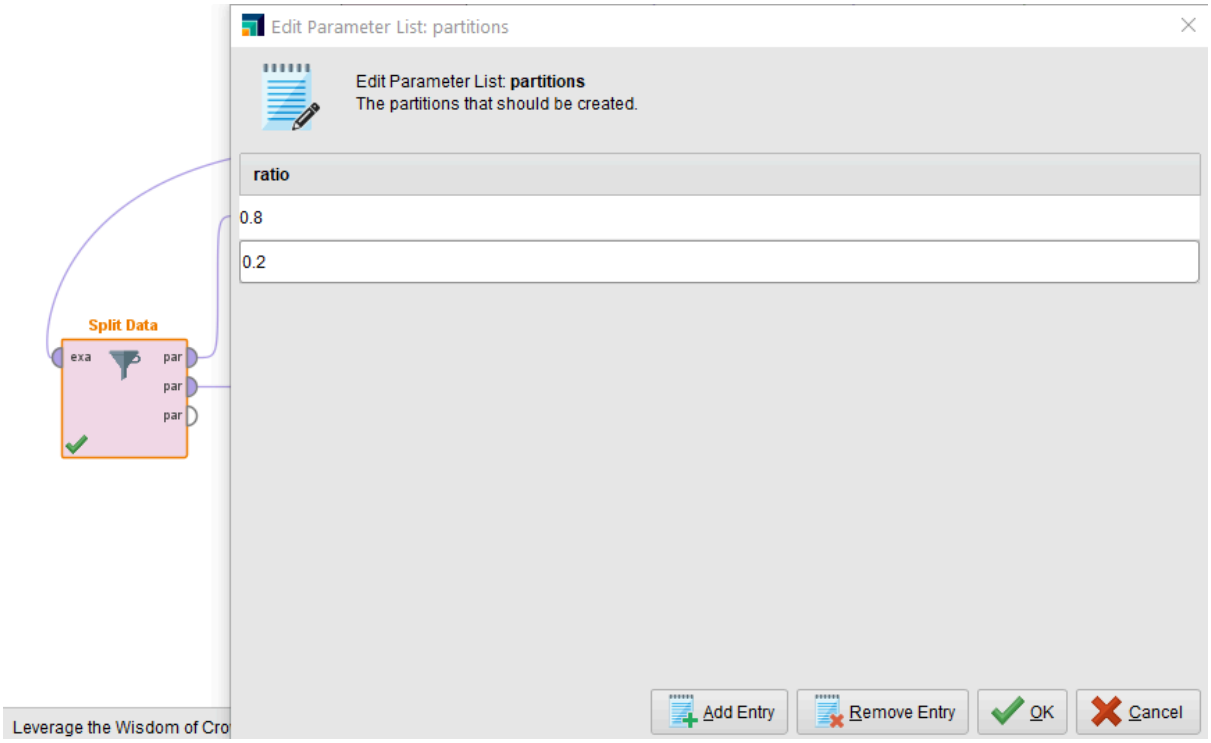
- Các thuộc tính được chọn làm attribute: Casino, Free Internet, Gym, Hotel name, Hotel star, Pool, Spa, Tennis court.
- Chọn k = 20
- Tỷ lệ phân chia Train/Test = 0.8/0.2

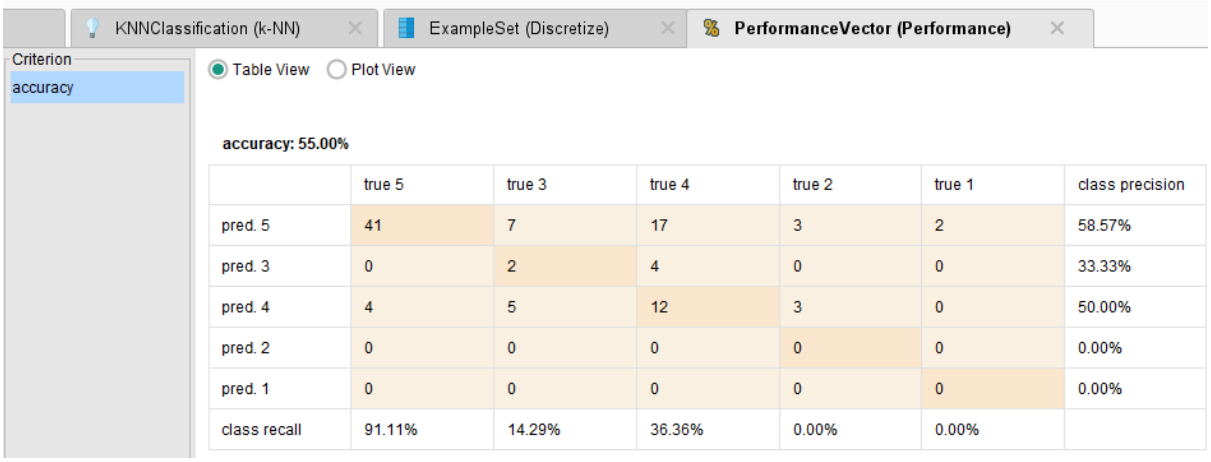



Process
Process
Retrieve LasVegasT...
Performance
Select Attributes
Set Role
Discretize
Normalize
k-NN
Split Data
Apply Model
Parameters
k-NN
k
20
weighted vote
measure types
MixedMeasures
mixed measure
MixedEuclideanDistance


Process
Performance
Retrieve LasVegasT...
Select Attributes
Split Data
Select Attributes: select subset
Select Attributes: select subset
Click to select the attribute subset.
Attributes
Search
Helpful votes
Member years
Nr. hotel reviews
Nr. reviews
Nr. rooms
Period of stay
Review month
Review weekday
Traveler type
User continent
User country
Selected Attributes
Search
Casino
Free internet
Gym
Hotel name
Hotel stars
Pool
Score
Spa
Tennis court
Apply
Cancel




| | true 5 | true 3 | true 4 | true 2 | true 1 | class precision |
|---|---|---|---|---|---|---|
| pred. 5 | 41 | 7 | 17 | 3 | 2 | 58.57% |
| pred. 3 | 0 | 2 | 4 | 0 | 0 | 33.33% |
| pred. 4 | 4 | 5 | 12 | 3 | 0 | 50.00% |
| pred. 2 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. 1 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| class recall | 91.11% | 14.29% | 36.36% | 0.00% | 0.00% | |

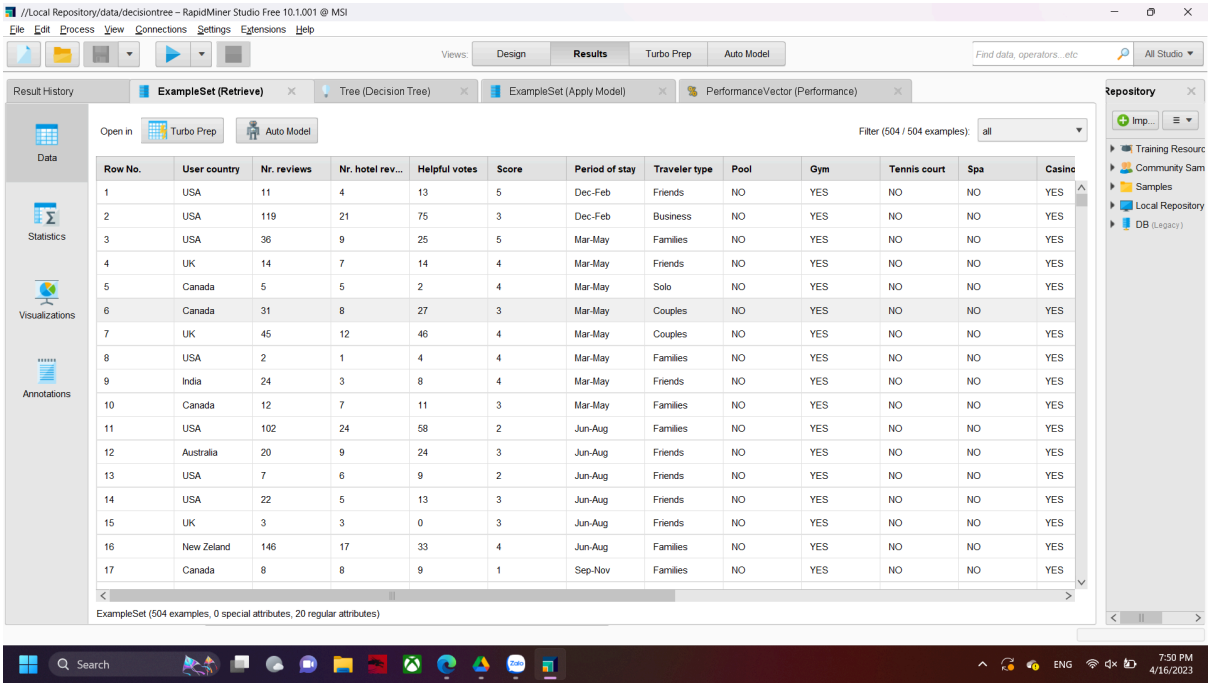
Kết quả giải thuật k-NN:

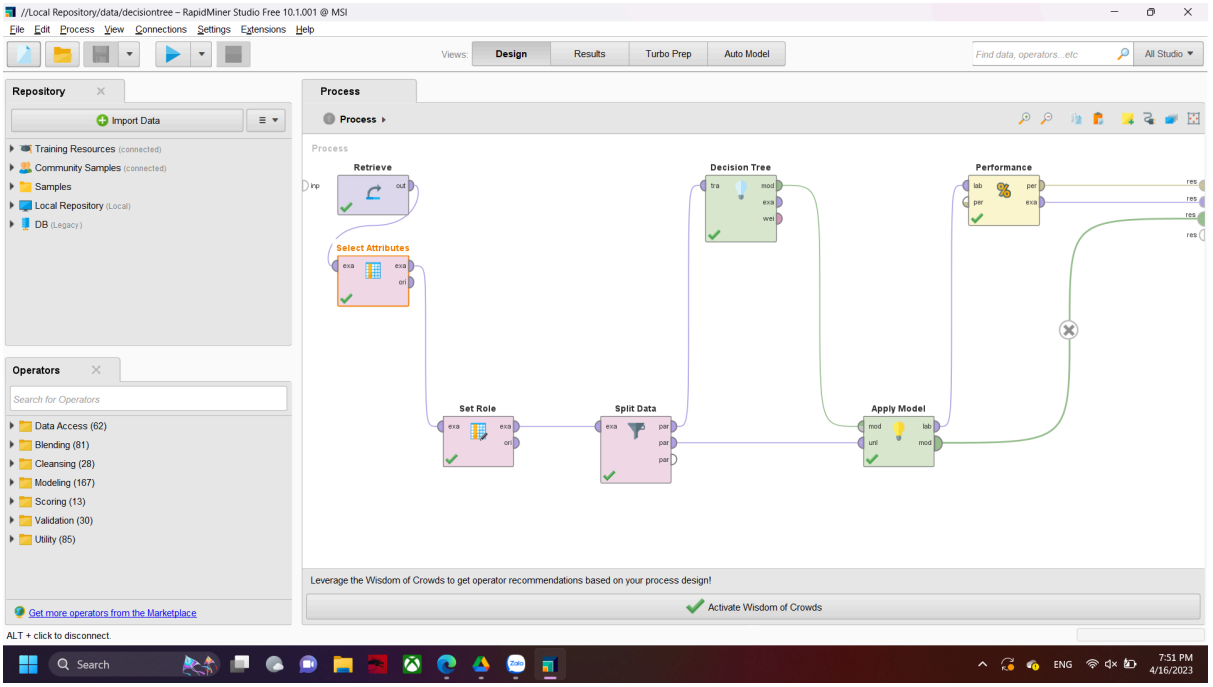
- Recall, Precision: như bảng đính kèm ở trên.
- Accuracy: 55,50%

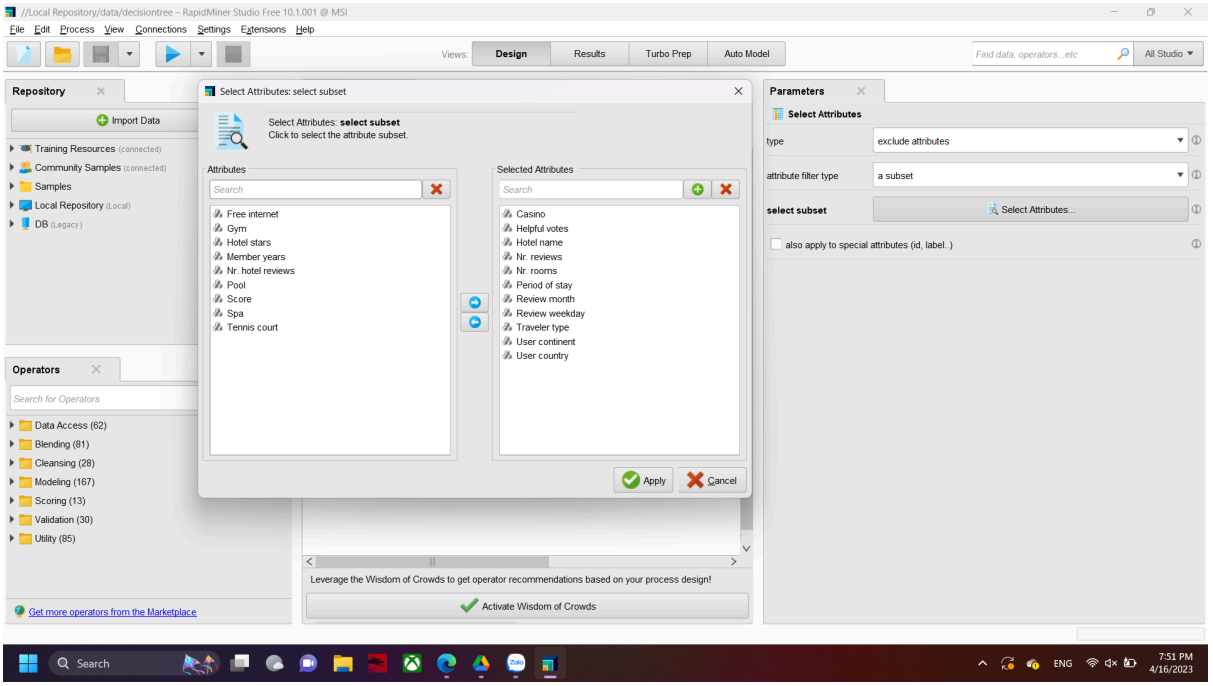
## 3.3) Giải thuật: Decision tree.

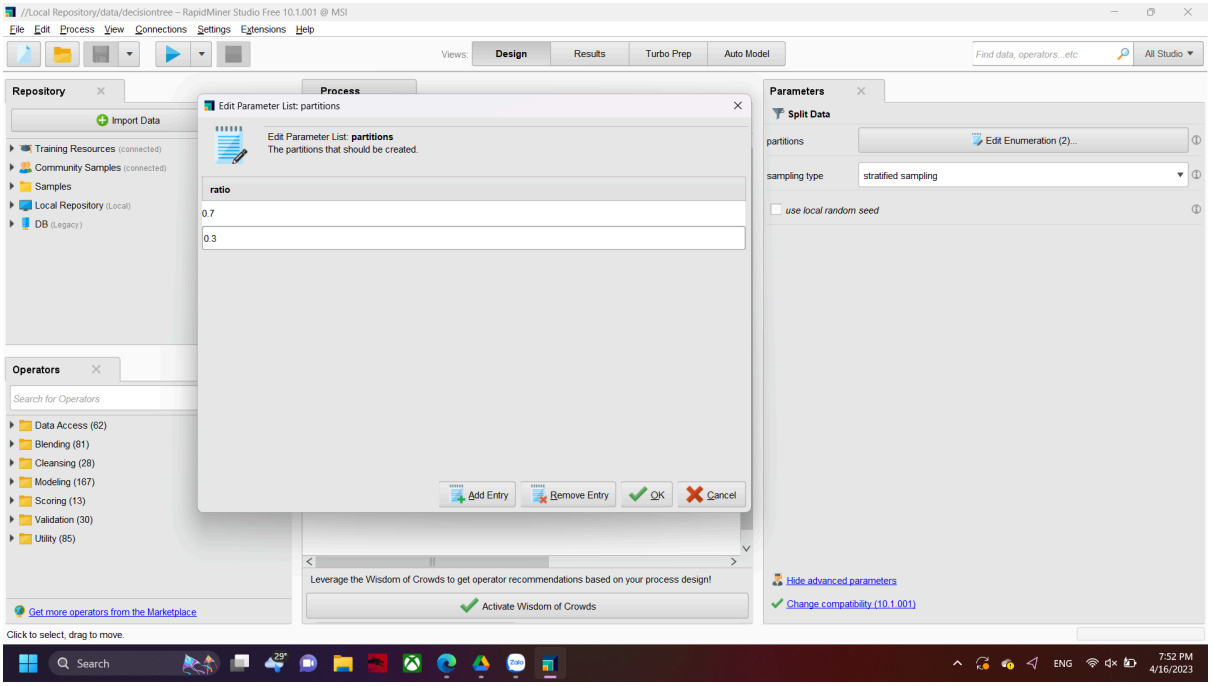
- Thuộc tính “score” được chọn gắn label.
- Các thuộc tính được chọn làm attribute: Casino, Helpful votes, Nr. reviews, Nr. rooms, Period of stay, Review months, Review weekday, Travel type, User continent, User country.

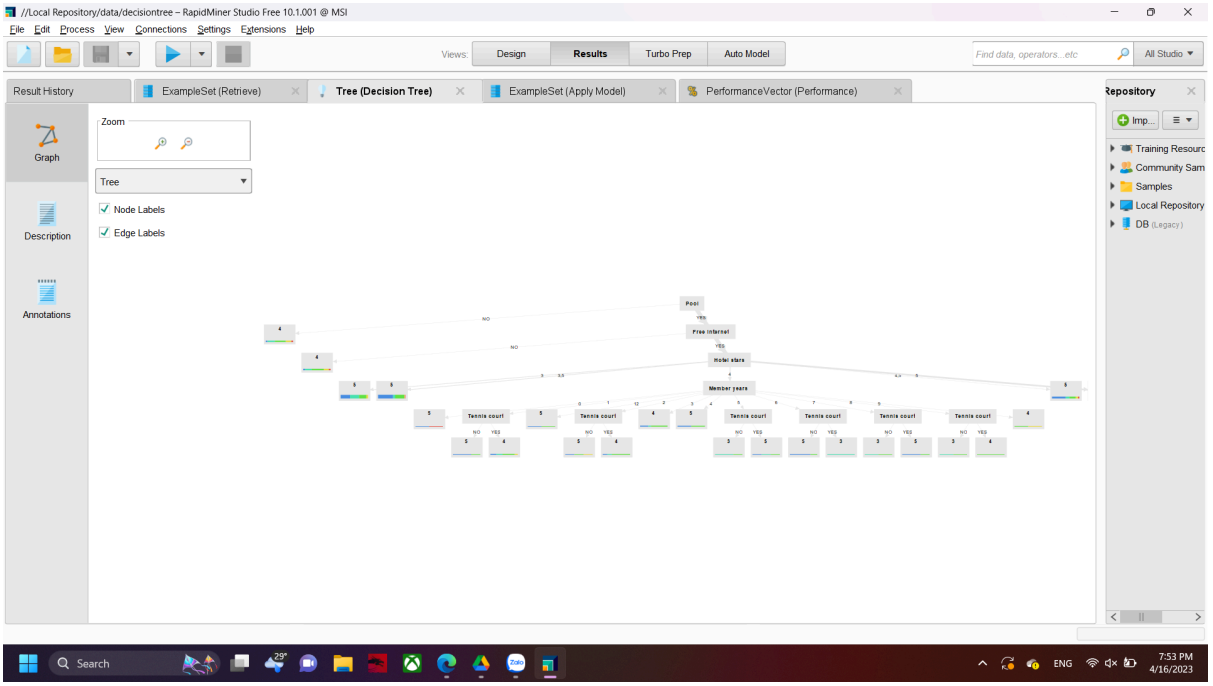
- Tỷ lệ phân chia Train/Test = 0.7/0.3

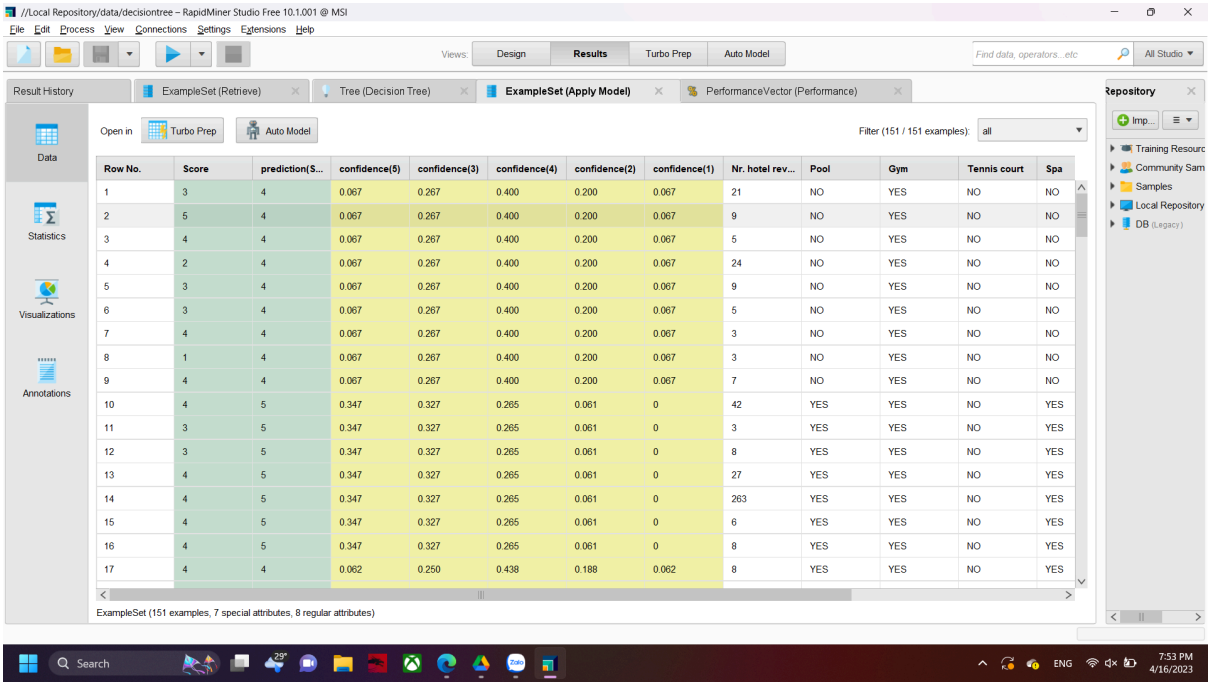


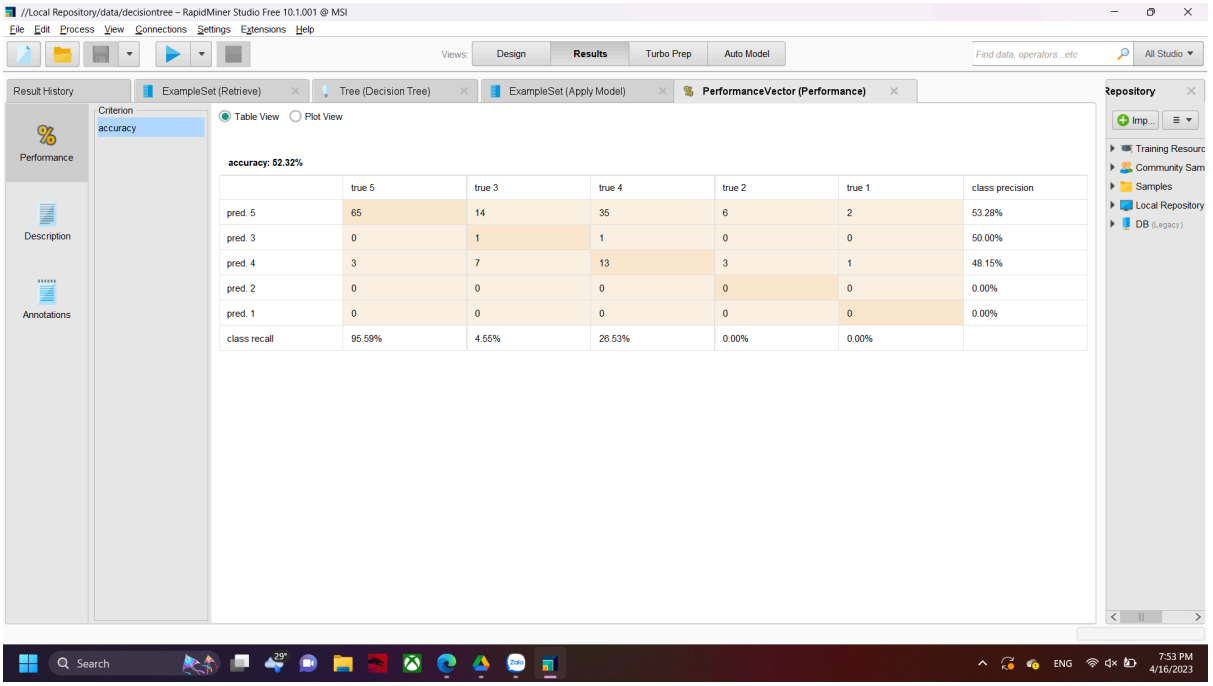




//Local Repository/data/decisiontree – RapidMiner Studio Free 10.1.001 @ MSI
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Repository
Import Data
Training Resources (connected)
Community Samples (connected)
Samples
Local Repository (Local)
DB (Legacy)
Select Attributes: select subset
Select Attributes: select subset
Click to select the attribute subset.
Attributes
Search
Free internet
Gym
Hotel stars
Member years
Nr. hotel reviews
Pool
Score
Spa
Tennis court
Selected Attributes
Search
Casino
Helpful votes
Hotel name
Nr. reviews
Nr. rooms
Period of stay
Review month
Review weekday
Traveler type
User continent
User country
Apply
Cancel
Parameters
Select Attributes
type
exclude attributes
attribute filter type
a subset
select subset
Select Attributes...
also apply to special attributes (id, label..)
Operators
Search for Operators
Data Access (62)
Blending (81)
Cleansing (28)
Modeling (167)
Scoring (13)
Validation (30)
Utility (85)
Get more operators from the Marketplace
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Search
ENG
7:51 PM
4/16/2023


//Local Repository/data/decisiontree – RapidMiner Studio Free 10.1.001 @ MSI
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Repository
Import Data
Training Resources (connected)
Community Samples (connected)
Samples
Local Repository (Local)
DB (Legacy)
Process
Edit Parameter List: partitions
Edit Parameter List: partitions
The partitions that should be created.
ratio
0.7
0.3
Add Entry
Remove Entry
OK
Cancel
Parameters
Split Data
partitions
Edit Enumeration (2)...
sampling type
stratified sampling
use local random seed
Operators
Search for Operators
Data Access (62)
Blending (81)
Cleansing (28)
Modeling (167)
Scoring (13)
Validation (30)
Utility (85)
Get more operators from the Marketplace
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Hide advanced parameters
Change compatibility (10.1.001)
Click to select, drag to move.
Search
ENG
7:52 PM
4/16/2023


//Local Repository/data/decisiontree – RapidMiner Studio Free 10.1.001 @ MSI
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Result History
ExampleSet (Retrieve)
Tree (Decision Tree)
ExampleSet (Apply Model)
PerformanceVector (Performance)
Graph
Description
Annotations
Zoom
Tree
Node Labels
Edge Labels
Repository
Training Resourc
Community Sam
Samples
Local Repository
DB (Legacy)
Search
ENG
7:53 PM
4/16/2023


//Local Repository/data/decisiontree – RapidMiner Studio Free 10.1.001 @ MSI
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Result History
ExampleSet (Retrieve)
Tree (Decision Tree)
ExampleSet (Apply Model)
PerformanceVector (Performance)
Data
Statistics
Visualizations
Annotations
Open in Turbo Prep Auto Model
Filter (151 / 151 examples): all
Row No. Score prediction(S... confidence(5) confidence(3) confidence(4) confidence(2) confidence(1) Nr. hotel rev... Pool Gym Tennis court Spa
1 3 4 0.067 0.267 0.400 0.200 0.067 21 NO YES NO NO
2 5 4 0.067 0.267 0.400 0.200 0.067 9 NO YES NO NO
3 4 4 0.067 0.267 0.400 0.200 0.067 5 NO YES NO NO
4 2 4 0.067 0.267 0.400 0.200 0.067 24 NO YES NO NO
5 3 4 0.067 0.267 0.400 0.200 0.067 9 NO YES NO NO
6 3 4 0.067 0.267 0.400 0.200 0.067 5 NO YES NO NO
7 4 4 0.067 0.267 0.400 0.200 0.067 3 NO YES NO NO
8 1 4 0.067 0.267 0.400 0.200 0.067 3 NO YES NO NO
9 4 4 0.067 0.267 0.400 0.200 0.067 7 NO YES NO NO
10 4 5 0.347 0.327 0.265 0.061 0 42 YES YES NO YES
11 3 5 0.347 0.327 0.265 0.061 0 3 YES YES NO YES
12 3 5 0.347 0.327 0.265 0.061 0 8 YES YES NO YES
13 4 5 0.347 0.327 0.265 0.061 0 27 YES YES NO YES
14 4 5 0.347 0.327 0.265 0.061 0 263 YES YES NO YES
15 4 5 0.347 0.327 0.265 0.061 0 6 YES YES NO YES
16 4 5 0.347 0.327 0.265 0.061 0 8 YES YES NO YES
17 4 4 0.062 0.250 0.438 0.188 0.062 8 YES YES NO YES
ExampleSet (151 examples, 7 special attributes, 8 regular attributes)
Repository
Training Resourc
Community Sam
Samples
Local Repository
DB (Legacy)
Search
ENG
7:53 PM
4/16/2023




Kết quả giải thuật Decision tree:

- Recall, Precision: như bảng đính kèm ở trên.
- Accuracy: 55,50%

# 4. Kết luận, đưa ra đánh giá về kết quả và đề xuất áp dụng mô hình cho nghiệp vụ khách sạn:

Thông qua việc sử dụng 3 giải thuật: Rule induction, Decision tree và k-NN nhận thấy giải thuật Rule induction cho độ chính xác (*accurancy*) cao nhất 74,4% so với giải thuật decision tree có độ chính xác 52,32% và giải thuật k-NN có độ chính xác 55%.

Từ model xây dựng được qua các giải thuật ở trên, người dùng có thể dự đoán được “score” cho các hotel mà khách đánh giá. Người dùng có một cái nhìn tổng quan về mức quan trọng và các yếu tố ảnh hưởng đến

“score”. Từ đó, người dùng có thể áp dụng các nghiệp vụ chuyên ngành để cải thiện dịch vụ và nâng cao chất lượng dịch vụ hơn.