# Experiment 5

## Aim:

Experiment to explore Rapid Miner and implement classification models like Decision Tree and Naive Bayes etc.

## Theory:

In machine learning, a decision tree is a predictive model that maps observations about an item to conclusions about its target value. It is a tree-structured model where each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. Decision trees can be used for both regression and classification tasks, and they are widely used in various applications, such as finance, medicine, and engineering. The goal of building a decision tree is to create a model that predicts the value of a target variable based on several input variables. The decision tree algorithm uses a recursive process to split the data into smaller subsets based on the input variables until it reaches a point where it can make a prediction.

The basic steps involved in creating a decision tree are as follows:

1. Collect and prepare data: This involves collecting and organizing the data, and then preparing it for analysis. This may include data cleaning, data transformation, and feature selection.
2. Choose an algorithm: There are various algorithms available for building decision trees, including ID3, C4.5, CART, and CHAID. The choice of algorithm will depend on the specific problem and the characteristics of the data.
3. Build the tree: This involves applying the chosen algorithm to the data to create the decision tree. The tree is built by recursively partitioning the data into subsets based on the values of the input features, and selecting the feature that provides the most information gain at each step.
4. Evaluate the tree: Once the tree is built, it needs to be evaluated to determine its accuracy and effectiveness. This may involve using cross-validation or other techniques to estimate the performance of the tree on new data.
5. Use the tree: Finally, the decision tree can be used to make predictions on new data. This involves traversing the tree from the root to a leaf node based on the values of the input features, and outputting the corresponding class label or value.

1. **Gini Index-**

Gini index is a measure of impurity or inequality used in decision trees to determine the purity of a given split in the data. It ranges from 0 to 1, where 0 represents a completely pure split (all observations belong to the same class) and 1 represents a completely impure split (an equal number of observations belong to each class).

The formula for calculating the Gini index is:
Gini = 1 - (p_1)^2 - (p_2)^2 - ... - (p_k)^2
where k is the number of classes, and p_i is the proportion of observations belonging to class i in the split.

2. **Information Gain-**

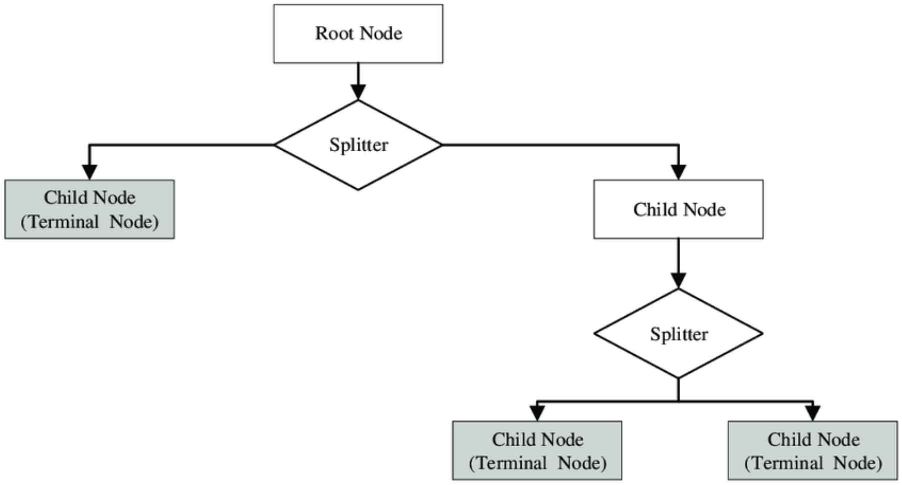
Information Gain is a measure used in decision tree algorithms to determine the relevance of a feature to a target variable. It measures the reduction in entropy or degree of disorder in the target variable when a feature is used to split the data into subsets. The formula for Information Gain is:

Information Gain = Entropy(parent) - [Weighted Avg. * Entropy(children)]
where,
Entropy(parent) is the entropy of the target variable for the entire dataset
Weighted Avg. is the weighted average of the entropy for each child node
Entropy(children) is the entropy of the target variable for each child node

3. **Entropy**-

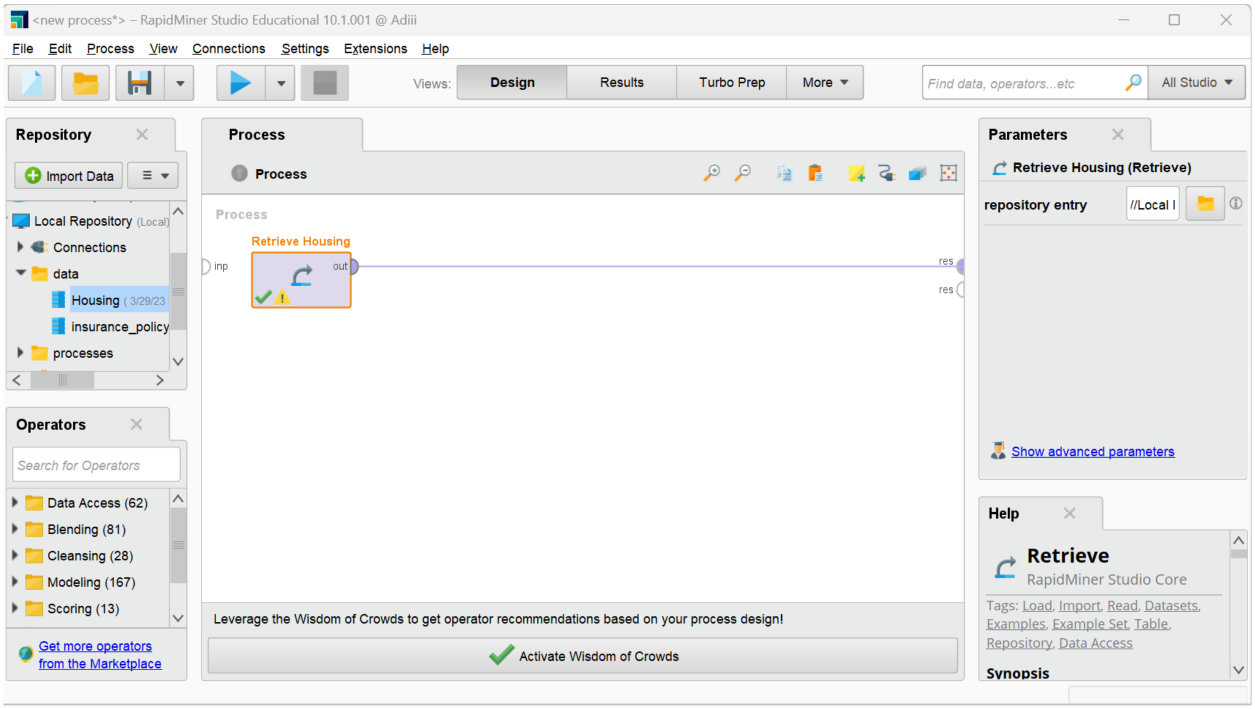
In the context of machine learning and decision trees, entropy is a measure of impurity or disorder within a set of examples. It is used to quantify the amount of uncertainty or randomness in a set of data. A dataset with low entropy is considered more uniform and has less randomness, whereas a dataset with high entropy is considered more disordered and has more randomness. The entropy of a dataset S with respect to a binary target variable Y can be calculated using the following formula:

Entropy(S) = -p_1 \log_2 p_1 - p_2 \log_2 p_2$
where p1 is the proportion of examples in S that belong to class 1, and p2 is the proportion of examples in S that belong to class 2. The logarithm base 2 is used to measure the entropy in bits.


Root Node
Splitter
Child Node (Terminal Node)
Child Node
Splitter
Child Node (Terminal Node)
Child Node (Terminal Node)

## Implementation:

Decision tree

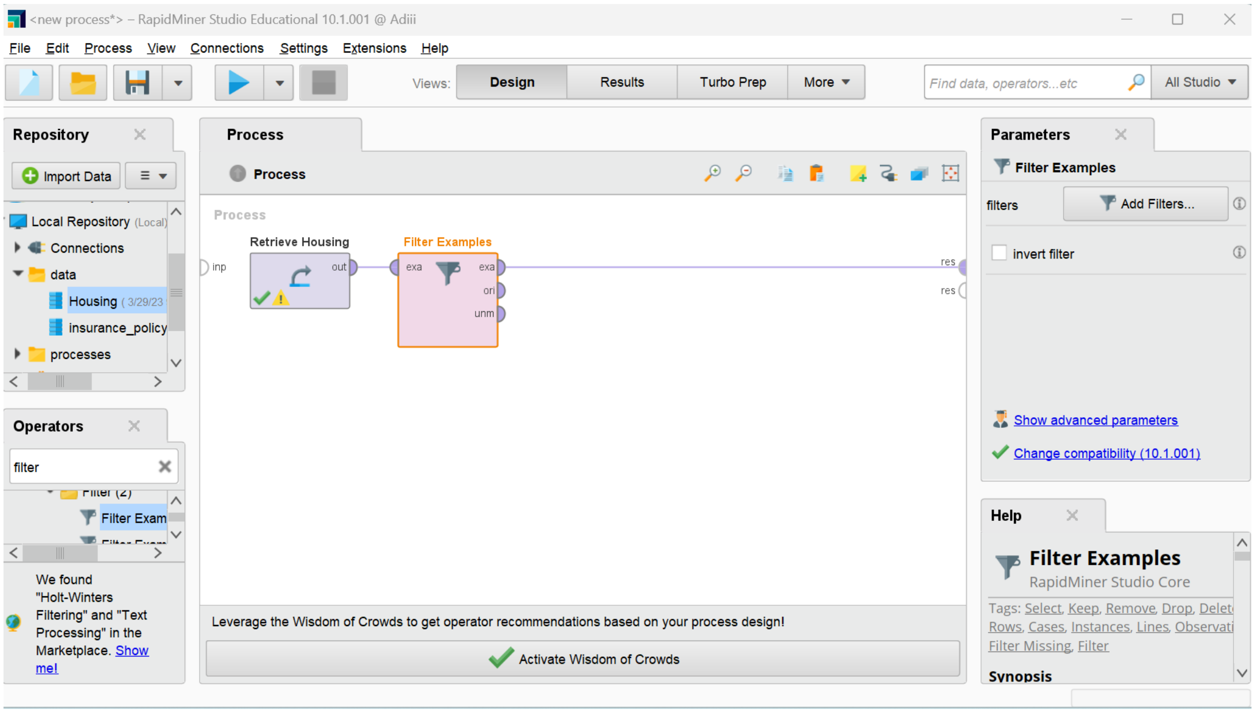


Retrieve Housing (Retrieve)

repository entry: //Local I

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Tags: Load, Import, Read, Datasets, Examples, Example Set, Table, Repository, Data Access



Filter Examples

filters

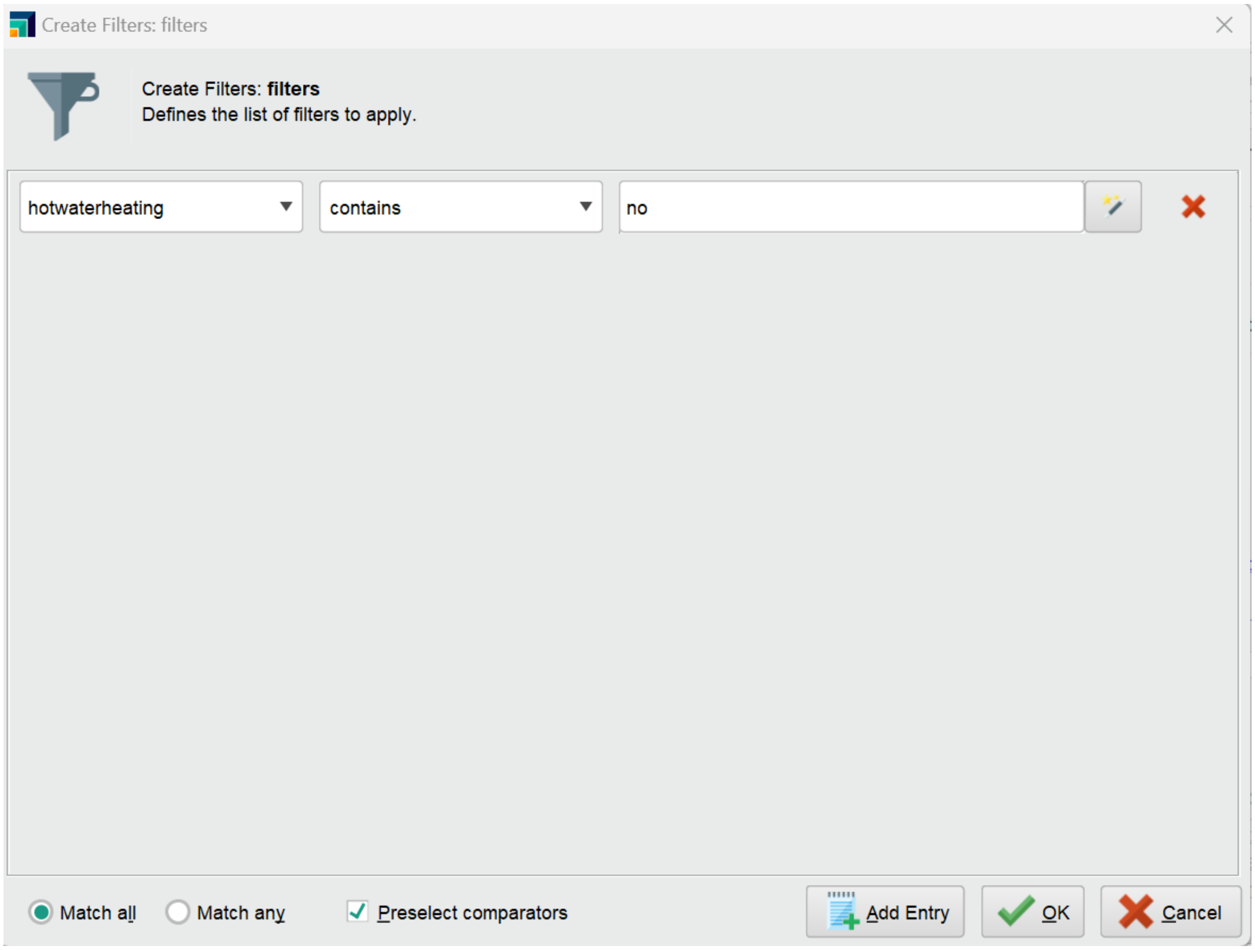
□ invert filter

Operators search: filter

We found "Holt-Winters Filtering" and "Text Processing" in the Marketplace. Show me!

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Tags: Select, Keep, Remove, Drop, Delet Rows, Cases, Instances, Lines, Observati Filter Missing, Filter

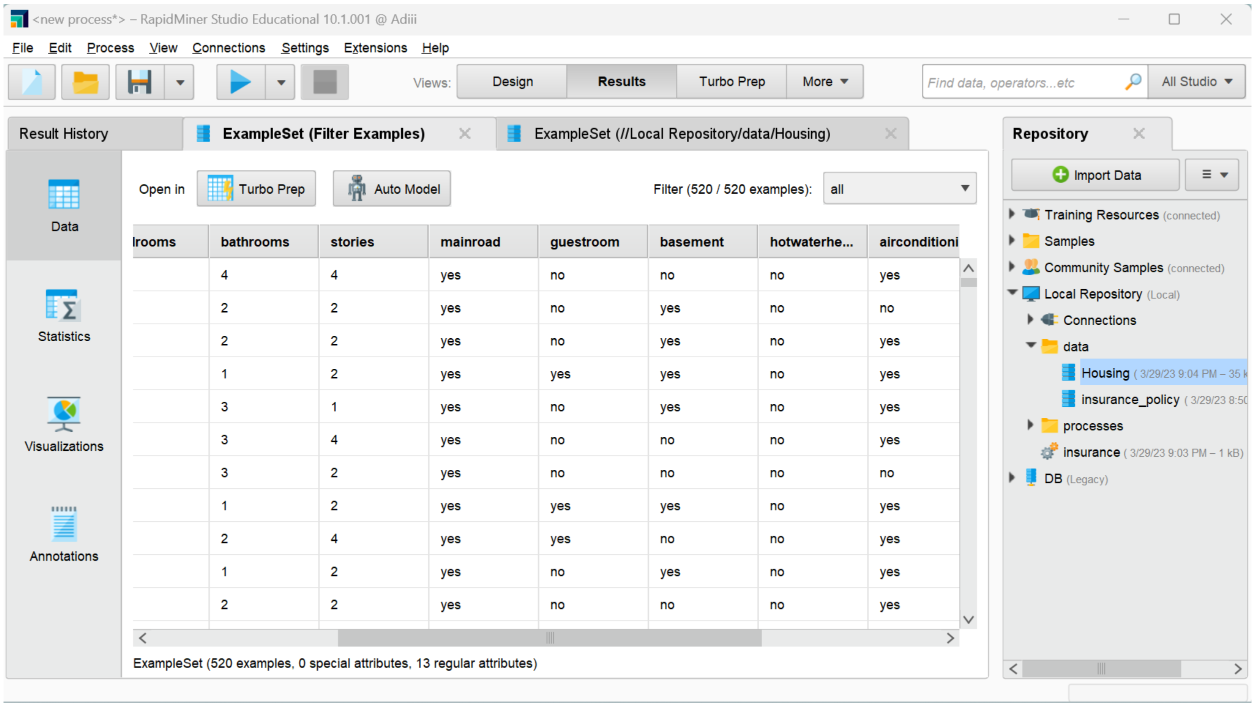


Create Filters: filters

Create Filters: **filters**
Defines the list of filters to apply.

| | | |
|---|---|---|
| hotwaterheating | contains | no |

● Match all ○ Match any ☑ Preselect comparators

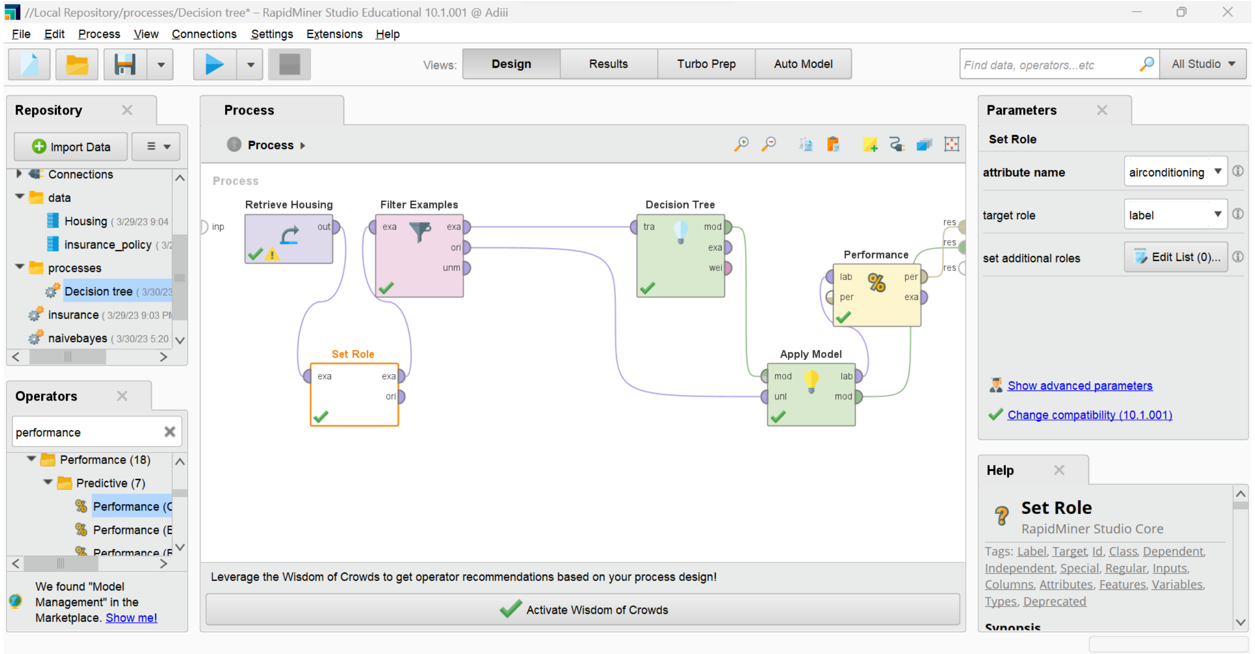


<new process*> – RapidMiner Studio Educational 10.1.001 @ Adiii

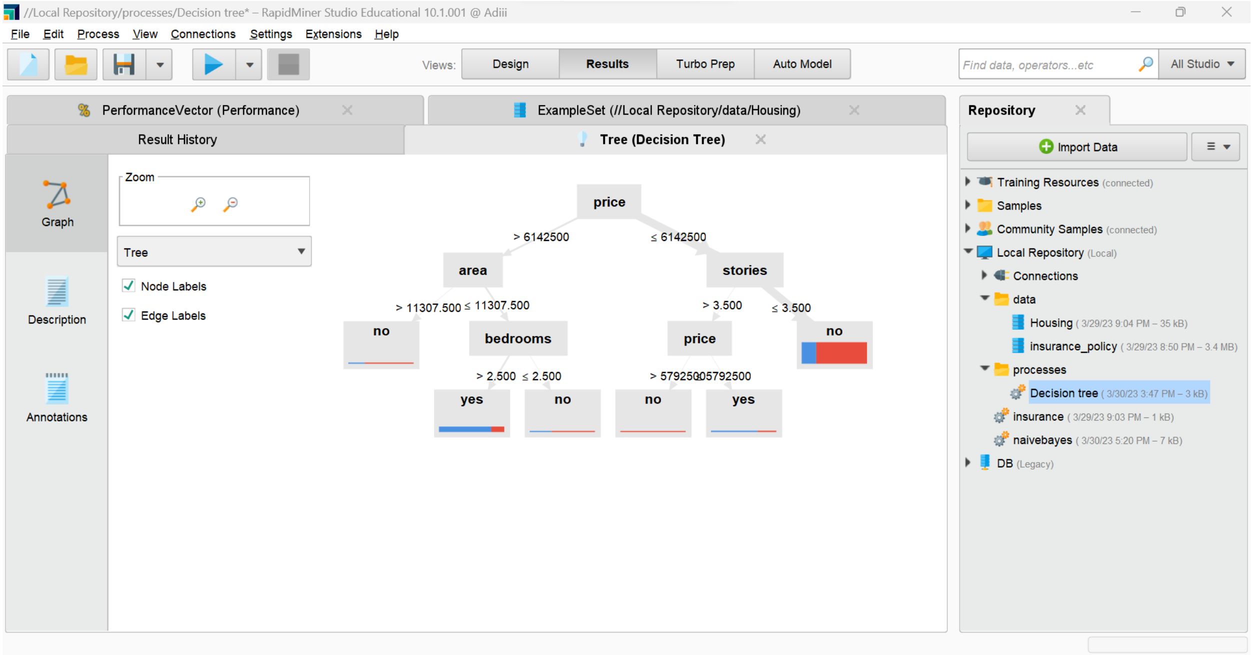
Filter (520 / 520 examples): all

| lrooms | bathrooms | stories | mainroad | guestroom | basement | hotwaterhe... | airconditioni |
|---|---|---|---|---|---|---|---|
| | 4 | 4 | yes | no | no | no | yes |
| | 2 | 2 | yes | no | yes | no | no |
| | 2 | 2 | yes | no | yes | no | yes |
| | 1 | 2 | yes | yes | yes | no | yes |
| | 3 | 1 | yes | no | yes | no | yes |
| | 3 | 4 | yes | no | no | no | yes |
| | 3 | 2 | yes | no | no | no | no |
| | 1 | 2 | yes | yes | yes | no | yes |
| | 2 | 4 | yes | yes | no | no | yes |
| | 1 | 2 | yes | no | yes | no | yes |
| | 2 | 2 | yes | no | no | no | yes |

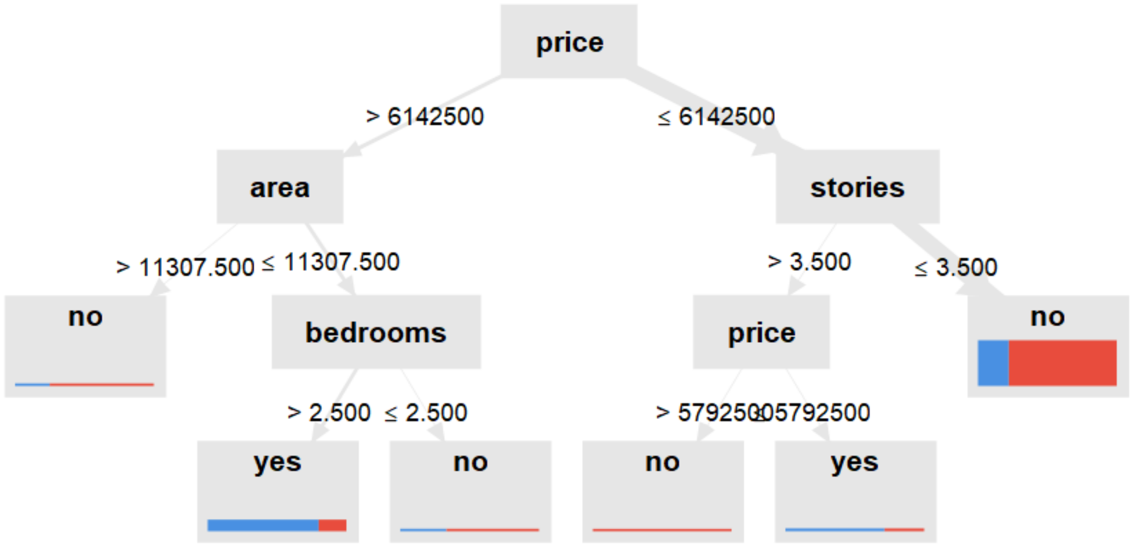
ExampleSet (520 examples, 0 special attributes, 13 regular attributes)

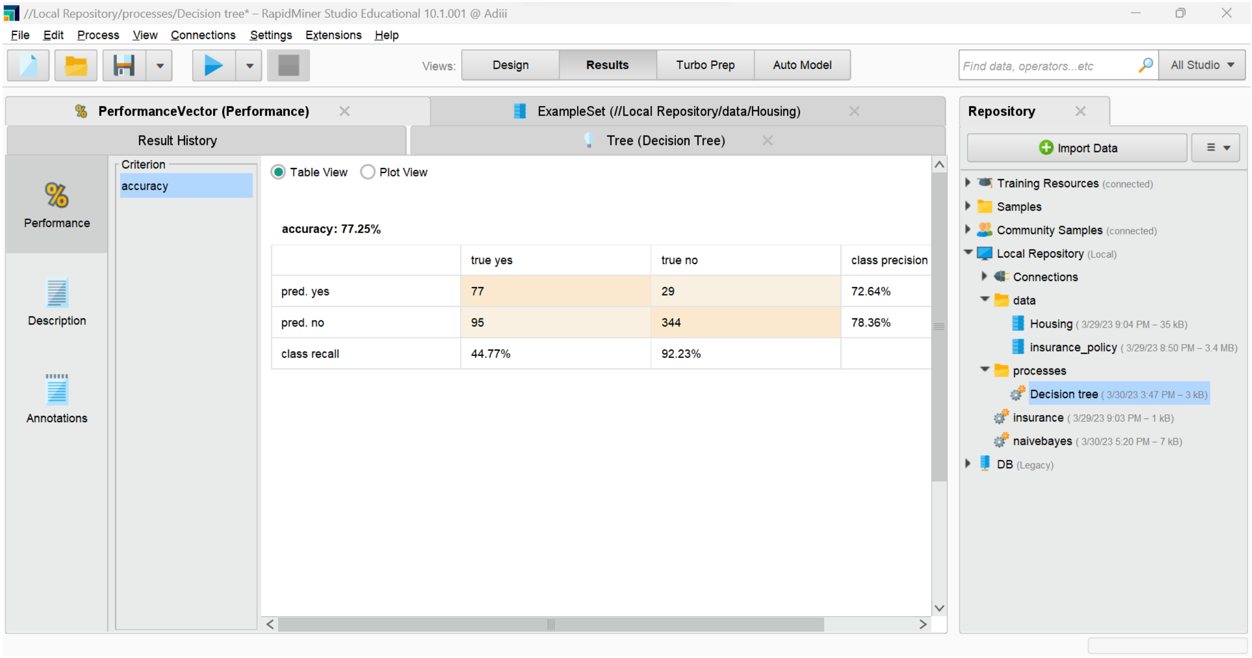

//Local Repository/processes/Decision tree* – RapidMiner Studio Educational 10.1.001 @ Adiii
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Repository
Import Data
Connections
data
Housing ( 3/29/23 9:04
insurance_policy ( 3/2
processes
Decision tree ( 3/30/23
insurance ( 3/29/23 9:03 PM
naivebayes ( 3/30/23 5:20
Operators
performance
Performance (18)
Predictive (7)
Performance (C
Performance (E
Performance (F
We found "Model Management" in the Marketplace. Show me!
Process
Process
Process
Retrieve Housing
inp
out
Filter Examples
exa
exa
ori
unm
Decision Tree
tra
mod
exa
wei
Performance
lab
per
per
exa
res
res
res
Set Role
exa
exa
ori
Apply Model
mod
unl
lab
mod
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Set Role
Show advanced parameters
Change compatibility (10.1.001)
Help
Set Role
RapidMiner Studio Core
Tags: Label, Target, Id, Class, Dependent, Independent, Special, Regular, Inputs, Columns, Attributes, Features, Variables, Types, Deprecated
Synopsis

| | |
|---|---|
| attribute name | airconditioning |
| target role | label |
| set additional roles | Edit List (0)... |


//Local Repository/processes/Decision tree* – RapidMiner Studio Educational 10.1.001 @ Adiii
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
PerformanceVector (Performance)
ExampleSet (//Local Repository/data/Housing)
Result History
Tree (Decision Tree)
Graph
Description
Annotations
Zoom
Tree
Node Labels
Edge Labels
price
> 6142500
≤ 6142500
area
stories
> 11307.500
≤ 11307.500
> 3.500
≤ 3.500
no
bedrooms
price
no
> 2.500
≤ 2.500
> 5792500
≤ 5792500
yes
no
no
yes
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
Housing ( 3/29/23 9:04 PM – 35 kB)
insurance_policy ( 3/29/23 8:50 PM – 3.4 MB)
processes
Decision tree ( 3/30/23 3:47 PM – 3 kB)
insurance ( 3/29/23 9:03 PM – 1 kB)
naivebayes ( 3/30/23 5:20 PM – 7 kB)
DB (Legacy)

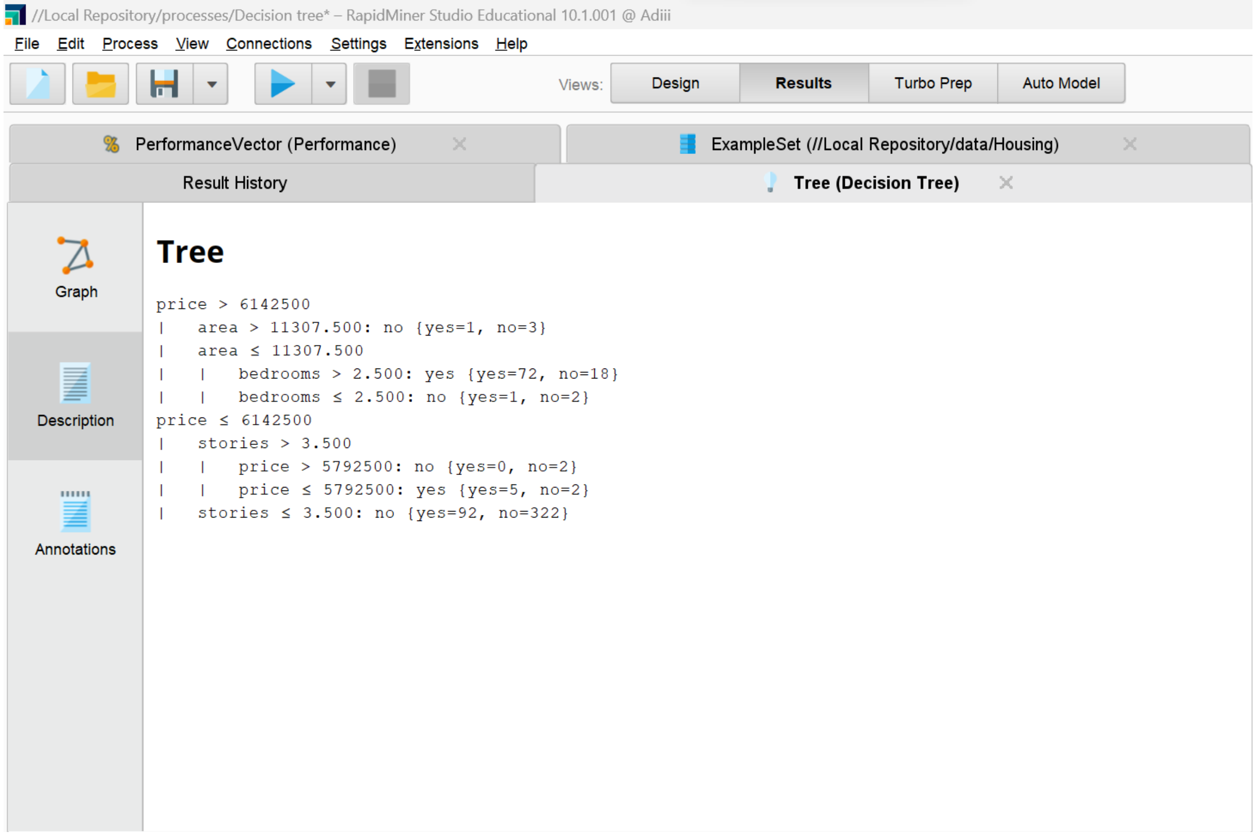
☑ Node Labels
☑ Edge Labels




//Local Repository/processes/Decision tree* – RapidMiner Studio Educational 10.1.001 @ Adiii
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
PerformanceVector (Performance)
ExampleSet (//Local Repository/data/Housing)
Result History
Tree (Decision Tree)
Performance
Description
Annotations
Criterion
accuracy
● Table View ○ Plot View
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
Housing ( 3/29/23 9:04 PM – 35 kB)
insurance_policy ( 3/29/23 8:50 PM – 3.4 MB)
processes
Decision tree ( 3/30/23 3:47 PM – 3 kB)
insurance ( 3/29/23 9:03 PM – 1 kB)
naivebayes ( 3/30/23 5:20 PM – 7 kB)
DB (Legacy)

accuracy: 77.25%

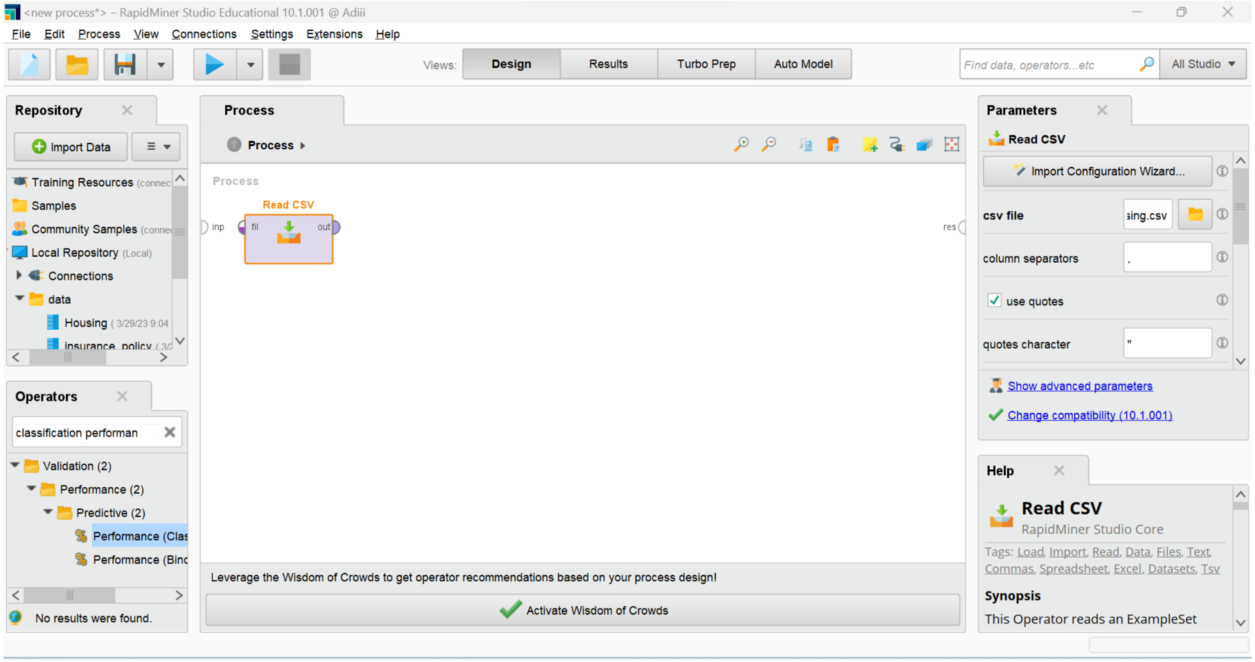
| | true yes | true no | class precision |
|---|---|---|---|
| pred. yes | 77 | 29 | 72.64% |
| pred. no | 95 | 344 | 78.36% |
| class recall | 44.77% | 92.23% | |


//Local Repository/processes/Decision tree* – RapidMiner Studio Educational 10.1.001 @ Adiii
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
PerformanceVector (Performance)
ExampleSet (//Local Repository/data/Housing)
Result History
Tree (Decision Tree)
Graph
Description
Annotations

### Tree

```
price > 6142500
|   area > 11307.500: no {yes=1, no=3}
|   area ≤ 11307.500
|   |   bedrooms > 2.500: yes {yes=72, no=18}
|   |   bedrooms ≤ 2.500: no {yes=1, no=2}
price ≤ 6142500
|   stories > 3.500
|   |   price > 5792500: no {yes=0, no=2}
|   |   price ≤ 5792500: yes {yes=5, no=2}
|   stories ≤ 3.500: no {yes=92, no=322}
```

## Naive Bayes


<new process*> – RapidMiner Studio Educational 10.1.001 @ Adiii
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 (connec
Samples
Community Samples (conne
Local Repository (Local)
Connections
data
Housing (3/29/23 9:04
insurance_policy (3/2
Operators
classification performan
Validation (2)
Performance (2)
Predictive (2)
Performance (Clas
Performance (Bino
No results were found.
Process
Process
Read CSV
inp
fil
out
res
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Read CSV
Import Configuration Wizard...
Show advanced parameters
Change compatibility (10.1.001)
Help
Read CSV
RapidMiner Studio Core

csv file: sing.csv

column separators: ,

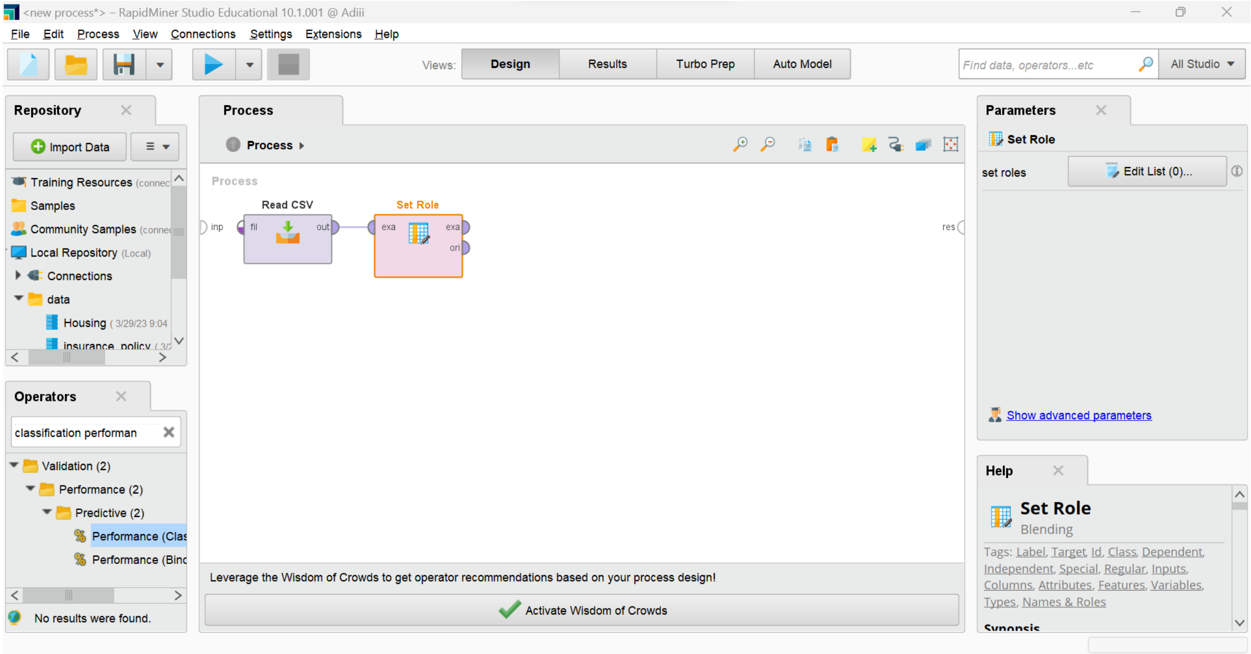
☑ use quotes

quotes character: "

Tags: Load, Import, Read, Data, Files, Text, Commas, Spreadsheet, Excel, Datasets, Tsv

Synopsis

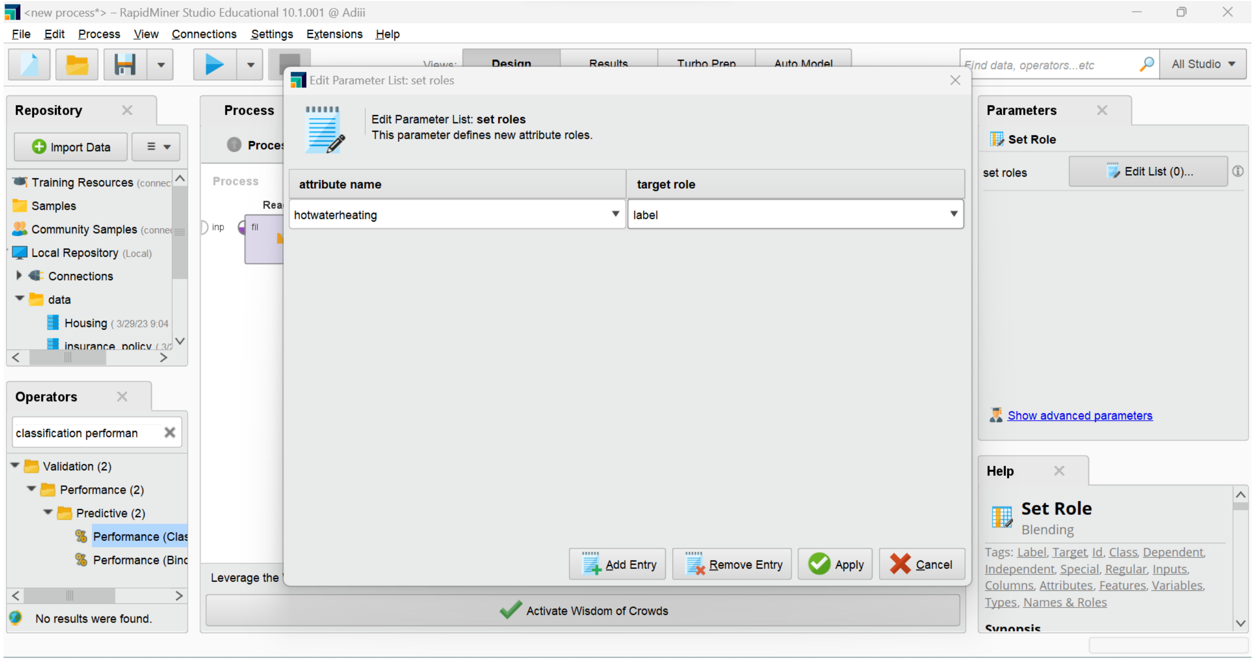
This Operator reads an ExampleSet



<new process*> – RapidMiner Studio Educational 10.1.001 @ Adiii
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 (connec
Samples
Community Samples (conne
Local Repository (Local)
Connections
data
Housing (3/29/23 9:04
insurance_policy (3/
Operators
Validation (2)
Performance (2)
Predictive (2)
Performance (Clas
Performance (Binc
No results were found.
Process
Read CSV
Set Role
inp
fil
out
exa
exa
ori
res
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Set Role
Show advanced parameters
Help
Set Role
Blending
Tags: Label, Target, Id, Class, Dependent, Independent, Special, Regular, Inputs, Columns, Attributes, Features, Variables, Types, Names & Roles
Synopsis

classification performan

set roles: Edit List (0)...

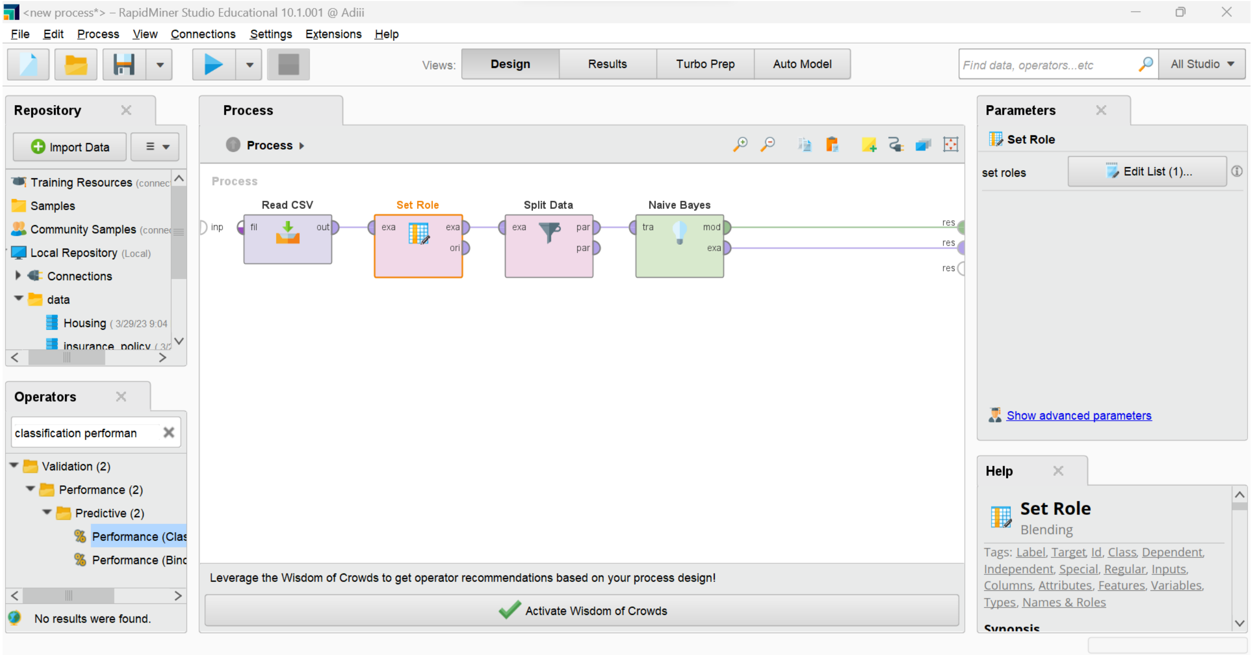


<new process*> – RapidMiner Studio Educational 10.1.001 @ Adiii
File Edit Process View Connections Settings Extensions Help
Edit Parameter List: set roles
Add Entry
Remove Entry
Apply
Cancel
Repository
Operators
Parameters
Set Role
Show advanced parameters
Help
Set Role
Blending
Tags: Label, Target, Id, Class, Dependent, Independent, Special, Regular, Inputs, Columns, Attributes, Features, Variables, Types, Names & Roles
Synopsis
Activate Wisdom of Crowds

Edit Parameter List: **set roles**
This parameter defines new attribute roles.

| attribute name | target role |
|---|---|
| hotwaterheating | label |

set roles: Edit List (0)...



<new process*> – RapidMiner Studio Educational 10.1.001 @ Adiii

Operators search: classification performan

No results were found.

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Parameters: Set Role

set roles: Edit List (1)...

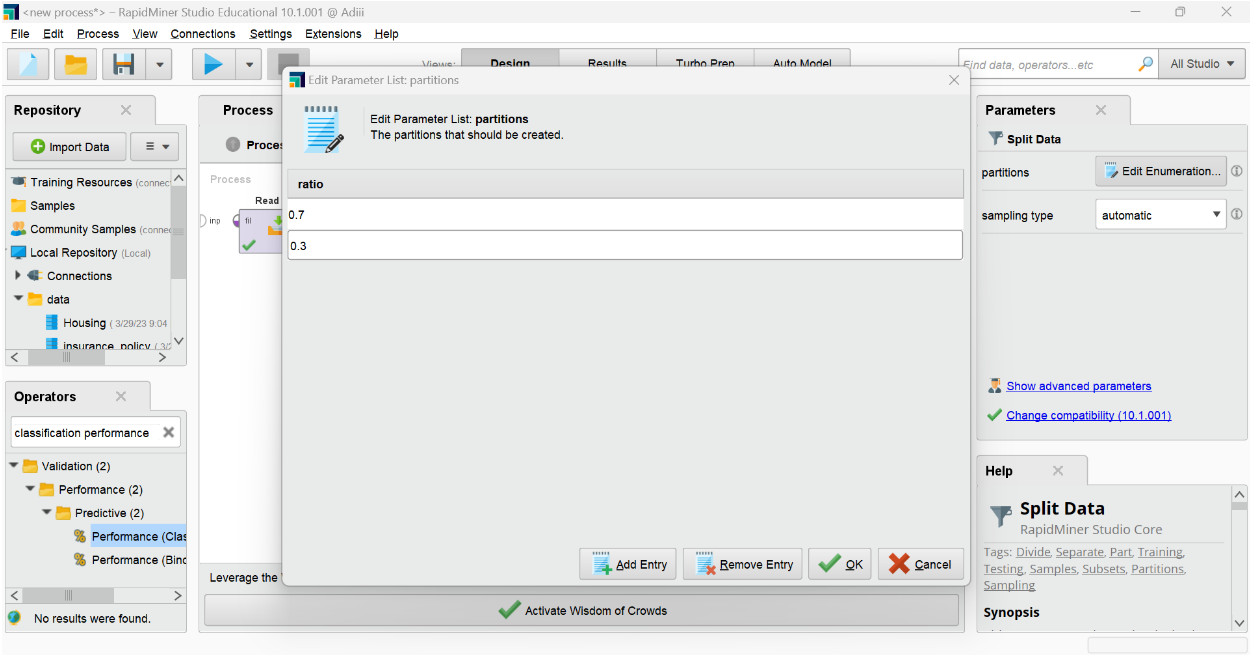
Show advanced parameters

Help: **Set Role**

Blending

Tags: Label, Target, Id, Class, Dependent, Independent, Special, Regular, Inputs, Columns, Attributes, Features, Variables, Types, Names & Roles

Synopsis



<new process*> – RapidMiner Studio Educational 10.1.001 @ Adiii

Edit Parameter List: partitions

Edit Parameter List: **partitions**
The partitions that should be created.

| ratio |
|---|
| 0.7 |
| 0.3 |

Operators search: classification performance

No results were found.

Leverage the

Parameters: Split Data

partitions: Edit Enumeration...

sampling type: automatic

Show advanced parameters

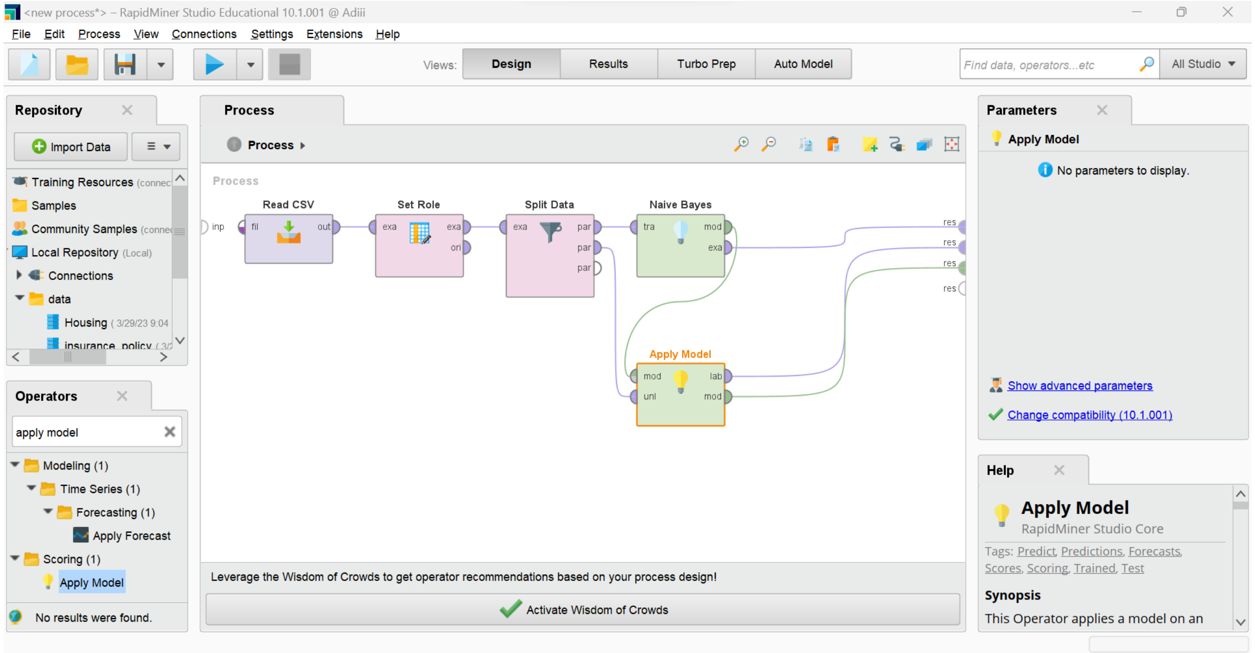
Change compatibility (10.1.001)

Help: **Split Data**

RapidMiner Studio Core

Tags: Divide, Separate, Part, Training, Testing, Samples, Subsets, Partitions, Sampling

Synopsis



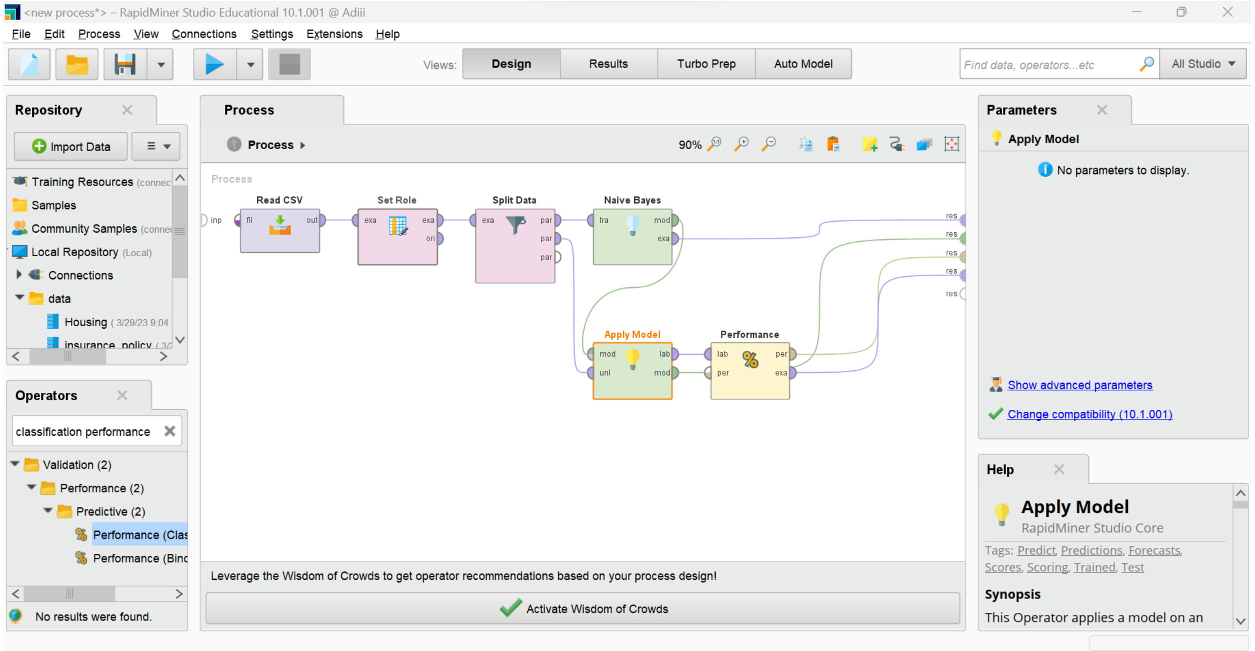
apply model

No results were found.

No parameters to display.

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

This Operator applies a model on an



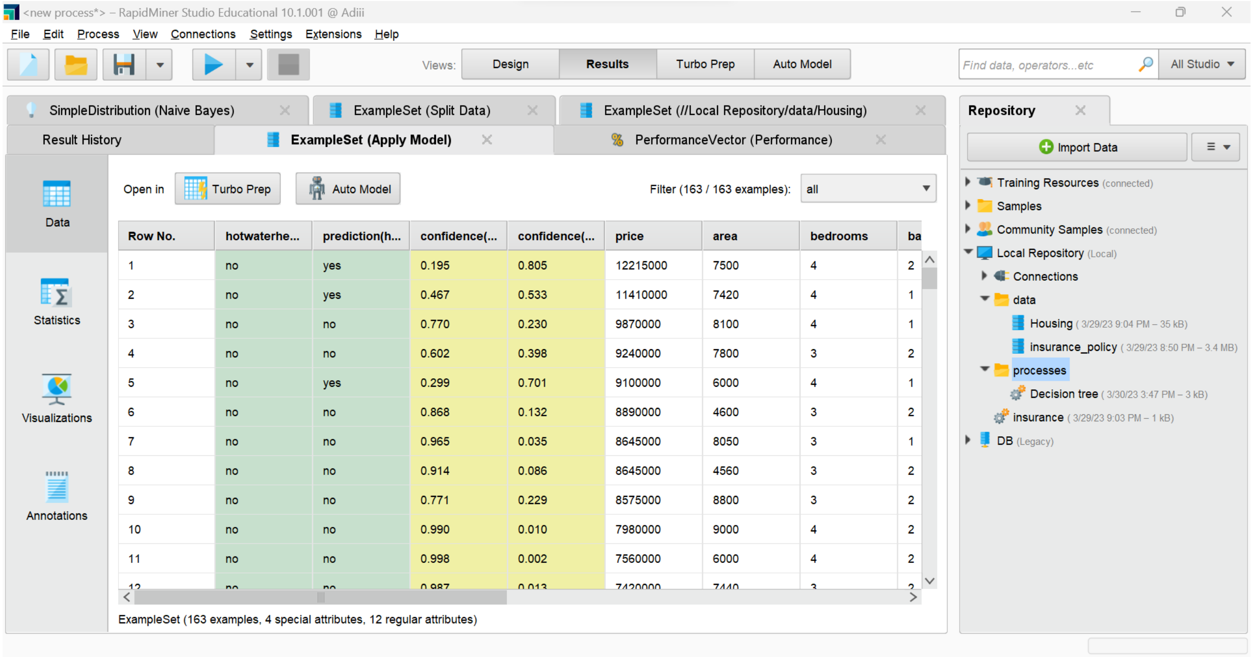
classification performance

No results were found.

No parameters to display.

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

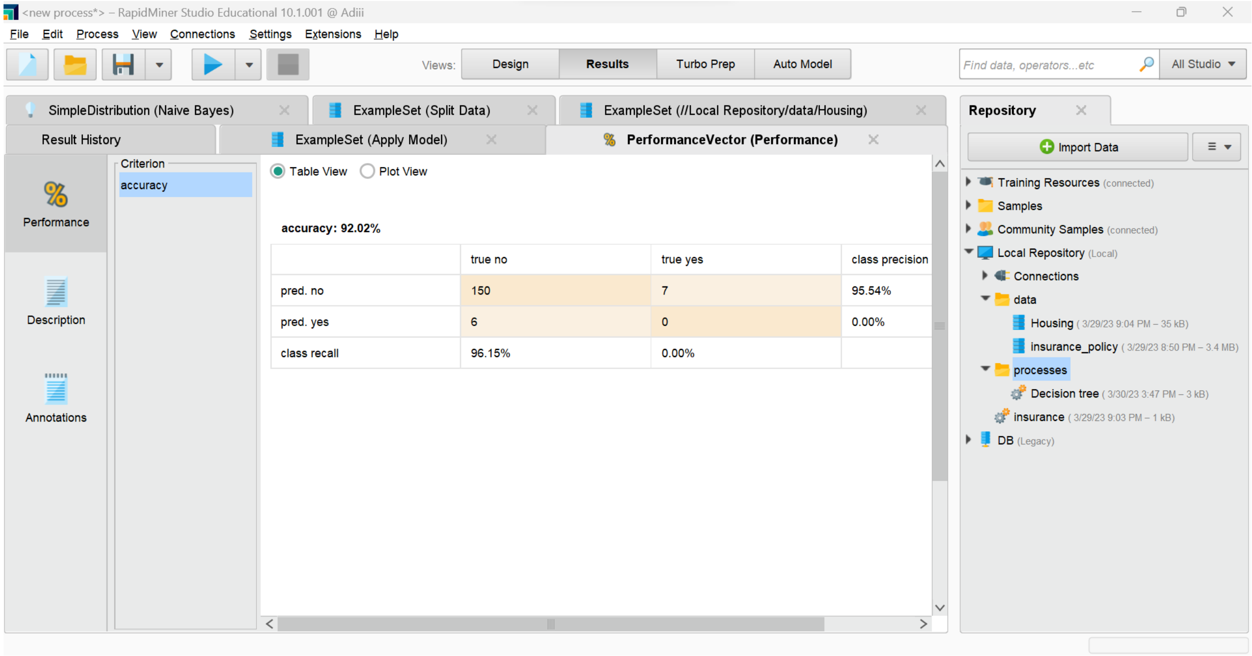
This Operator applies a model on an


<new process*> – RapidMiner Studio Educational 10.1.001 @ Adiii
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
SimpleDistribution (Naive Bayes)
ExampleSet (Split Data)
ExampleSet (//Local Repository/data/Housing)
Result History
ExampleSet (Apply Model)
PerformanceVector (Performance)
Data
Statistics
Visualizations
Annotations
Open in Turbo Prep Auto Model
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
Housing ( 3/29/23 9:04 PM – 35 kB)
insurance_policy ( 3/29/23 8:50 PM – 3.4 MB)
processes
Decision tree ( 3/30/23 3:47 PM – 3 kB)
insurance ( 3/29/23 9:03 PM – 1 kB)
DB (Legacy)

Filter (163 / 163 examples): all

| Row No. | hotwaterhe... | prediction(h... | confidence(... | confidence(... | price | area | bedrooms | ba |
|---|---|---|---|---|---|---|---|---|
| 1 | no | yes | 0.195 | 0.805 | 12215000 | 7500 | 4 | 2 |
| 2 | no | yes | 0.467 | 0.533 | 11410000 | 7420 | 4 | 1 |
| 3 | no | no | 0.770 | 0.230 | 9870000 | 8100 | 4 | 1 |
| 4 | no | no | 0.602 | 0.398 | 9240000 | 7800 | 3 | 2 |
| 5 | no | yes | 0.299 | 0.701 | 9100000 | 6000 | 4 | 1 |
| 6 | no | no | 0.868 | 0.132 | 8890000 | 4600 | 3 | 2 |
| 7 | no | no | 0.965 | 0.035 | 8645000 | 8050 | 3 | 1 |
| 8 | no | no | 0.914 | 0.086 | 8645000 | 4560 | 3 | 2 |
| 9 | no | no | 0.771 | 0.229 | 8575000 | 8800 | 3 | 2 |
| 10 | no | no | 0.990 | 0.010 | 7980000 | 9000 | 4 | 2 |
| 11 | no | no | 0.998 | 0.002 | 7560000 | 6000 | 4 | 2 |
| 12 | no | no | 0.987 | 0.013 | 7420000 | 7440 | 3 | 2 |

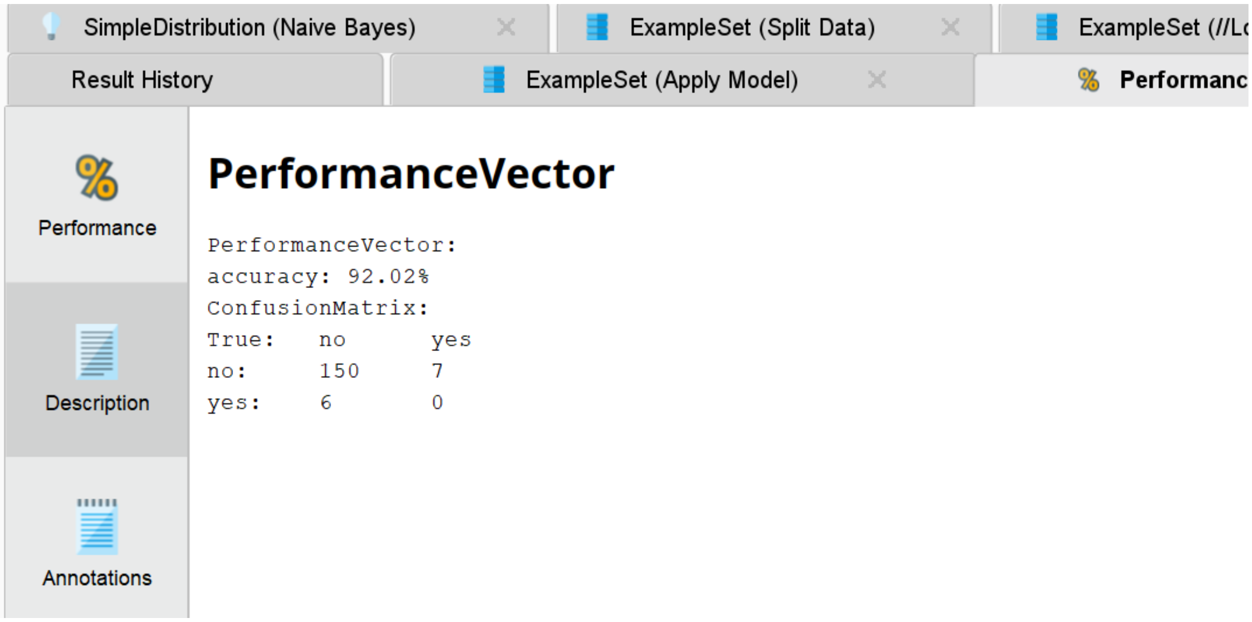
ExampleSet (163 examples, 4 special attributes, 12 regular attributes)


<new process*> – RapidMiner Studio Educational 10.1.001 @ Adiii
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
SimpleDistribution (Naive Bayes)
ExampleSet (Split Data)
ExampleSet (//Local Repository/data/Housing)
Result History
ExampleSet (Apply Model)
PerformanceVector (Performance)
Performance
Description
Annotations
Criterion
accuracy
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
Housing ( 3/29/23 9:04 PM – 35 kB)
insurance_policy ( 3/29/23 8:50 PM – 3.4 MB)
processes
Decision tree ( 3/30/23 3:47 PM – 3 kB)
insurance ( 3/29/23 9:03 PM – 1 kB)
DB (Legacy)

● Table View ○ Plot View

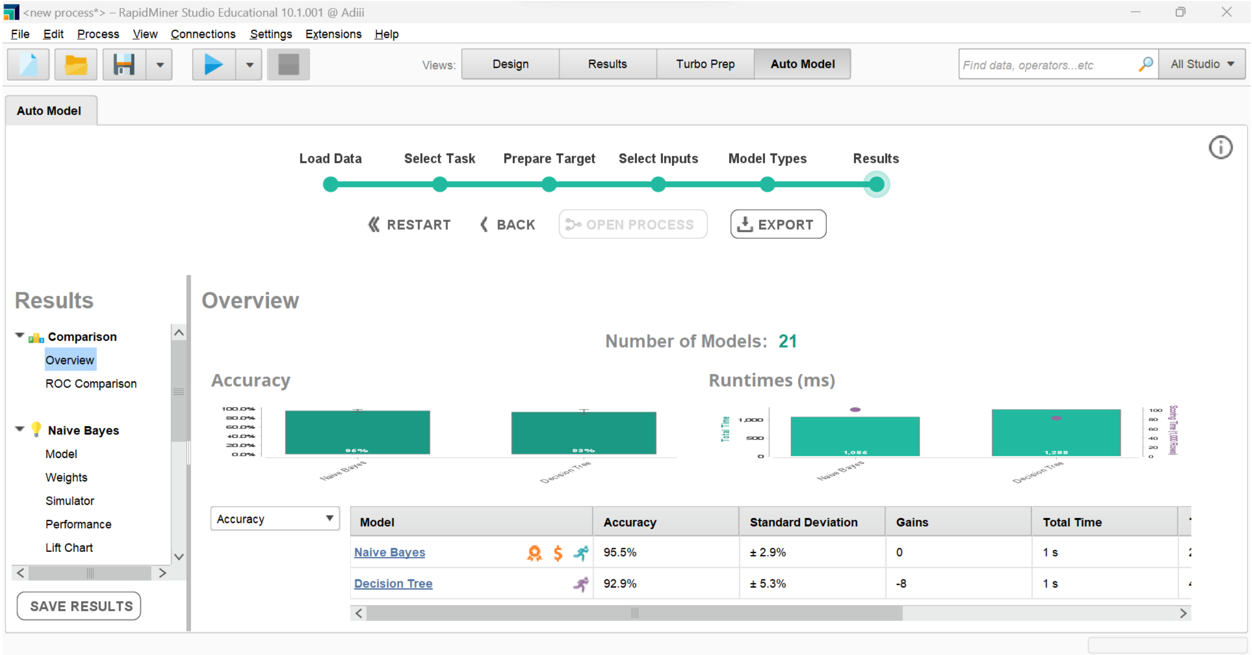
**accuracy: 92.02%**

| | true no | true yes | class precision |
|---|---|---|---|
| pred. no | 150 | 7 | 95.54% |
| pred. yes | 6 | 0 | 0.00% |
| class recall | 96.15% | 0.00% | |



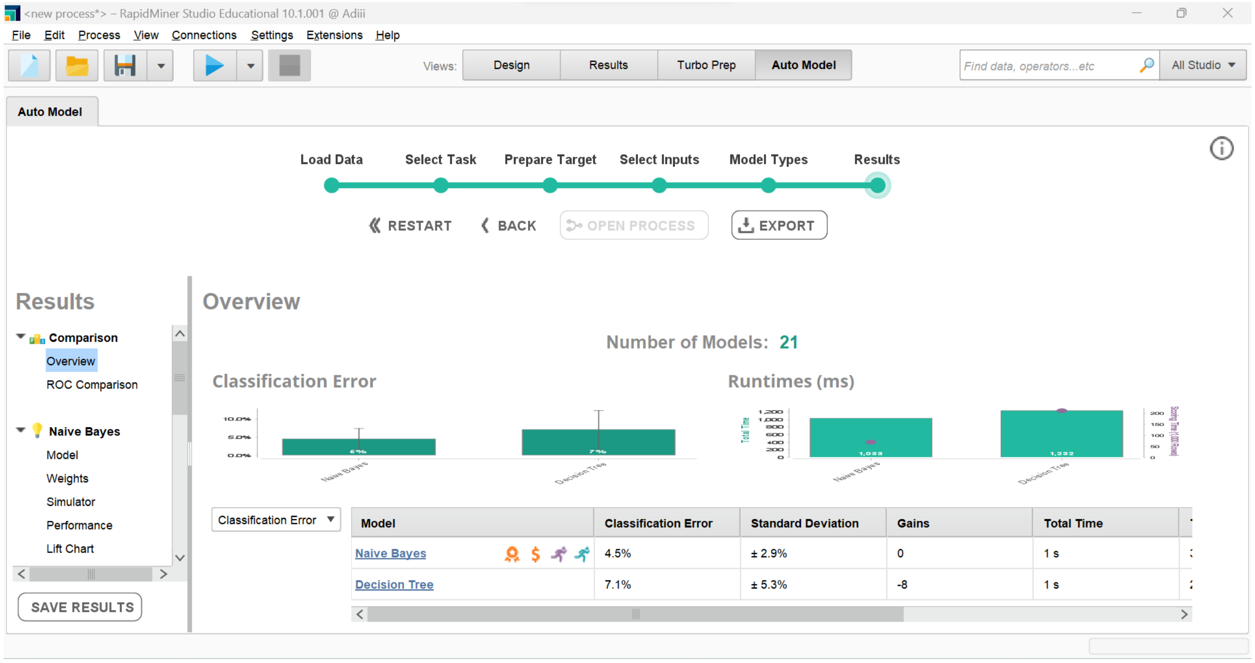
## PerformanceVector

```
PerformanceVector:
accuracy: 92.02%
ConfusionMatrix:
True:   no      yes
no:     150     7
yes:    6       0
```



Number of Models: 21

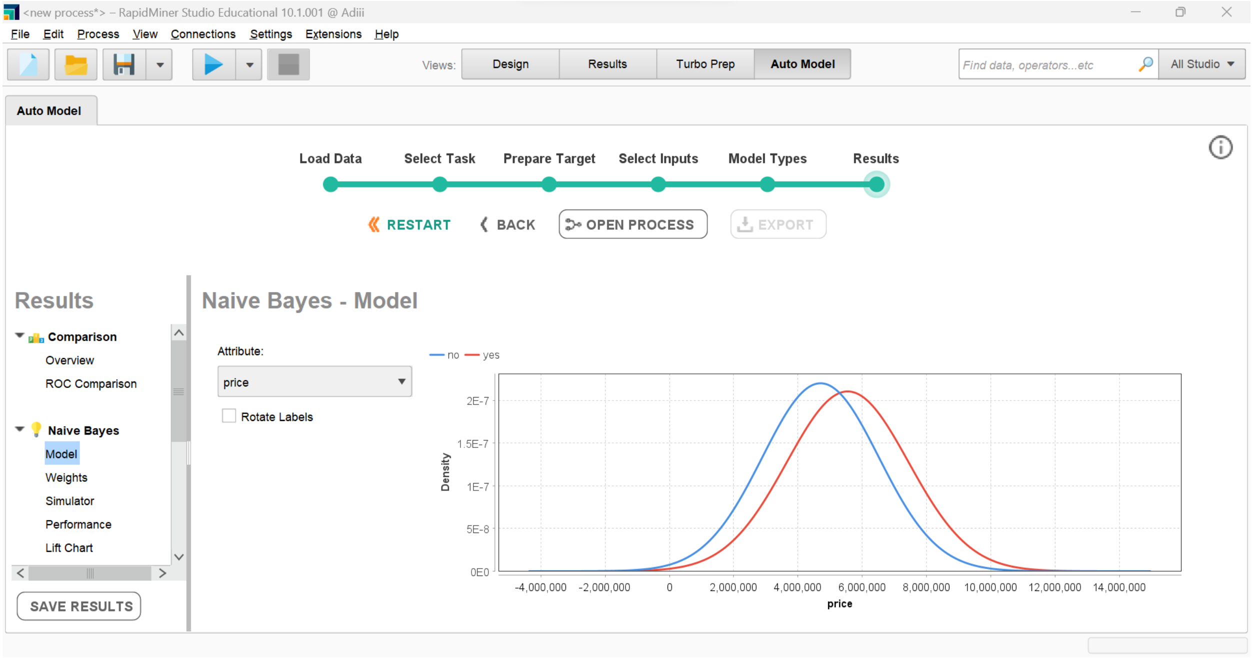
| Model | Accuracy | Standard Deviation | Gains | Total Time |
|---|---|---|---|---|
| Naive Bayes | 95.5% | ± 2.9% | 0 | 1 s |
| Decision Tree | 92.9% | ± 5.3% | -8 | 1 s |



Number of Models: 21

| Model | Classification Error | Standard Deviation | Gains | Total Time |
|---|---|---|---|---|
| Naive Bayes | 4.5% | ± 2.9% | 0 | 1 s |
| Decision Tree | 7.1% | ± 5.3% | -8 | 1 s |

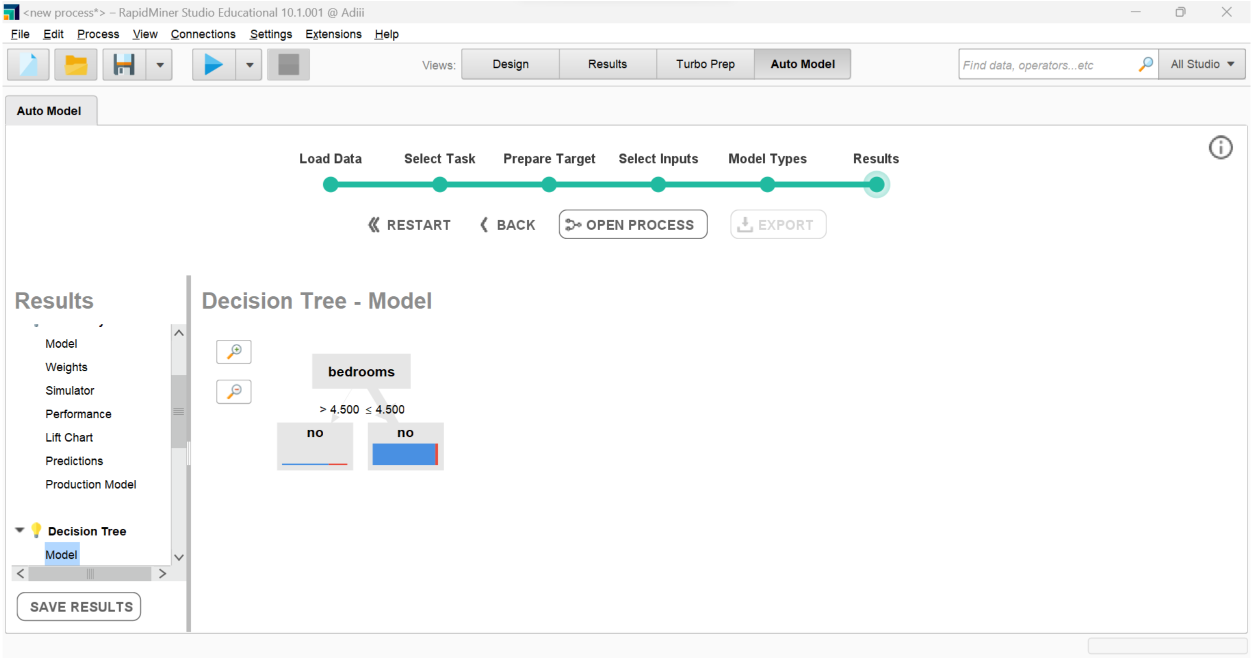
Naive bayes



Attribute: price

□ Rotate Labels

Decision tree



# Conclusion-

We have successfully explored Rapid Miner and implemented classification models like Decision Tree and Naive Bayes.