Mubashir Ali Data Mining

Muhammad Abdullah

Business Understanding

Background:

This dataset is the most infamous shipwrecks in the history of the ship titanic. The problem is to find the passengers survived and died during the ship wreck.

Business Objectives:

The objective is to apply the gain_ratio and gini_index for the survivors.

Data Understanding

Collect Initial Data

The data provided is consists of:

- dataset.xlsx: contains information about the passengers' id, class, name, sex, age and survival_result with 312 rows.

The data does not contain any personal information.

Description of the data:

The file containing the data required for data mining is dataset.xlsx. Below is the description of each of its fields and their data type.

Variable Name	Description	Type
passengerID	Passenger id used in the ship	Integer
PassengerClass	Class used by the passenger	Integer
Name	Name of the passenger	Nominal
Sex	Gender of the passenger	Nominal
Age	Defines the age of the passenger	Integer
Survival_result	Survival result of the passenger	Nominal

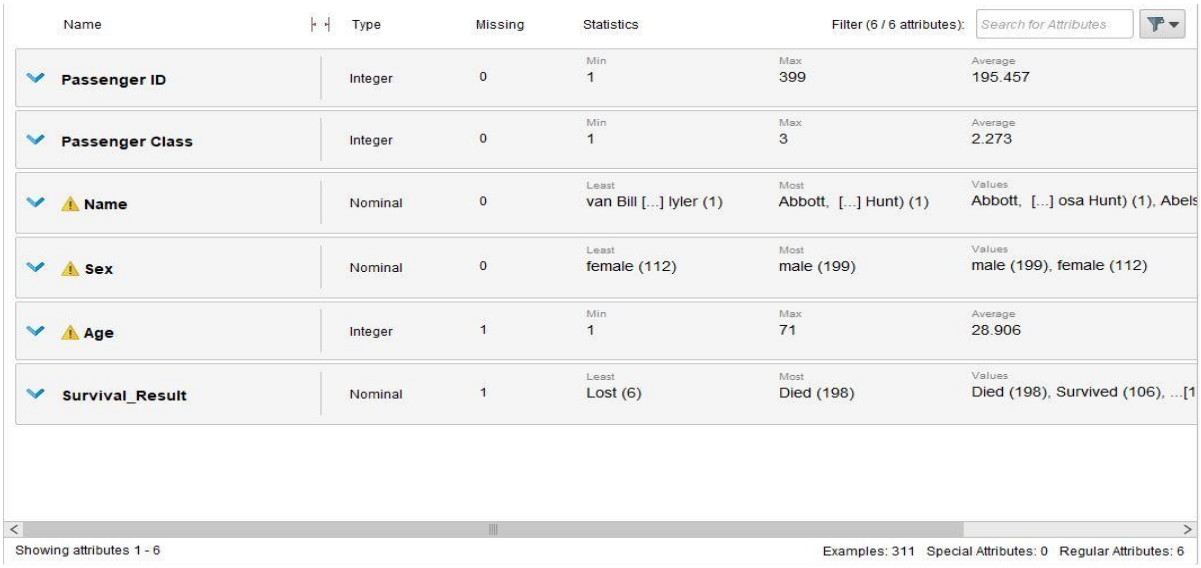
Data Preparation

Clean Data

This process was performed using the RapidMiner tool, and was executed as follows:

The dataset was imported first into a RapidMiner repository.

The dataset contains an attribute (**Age**) with missing value. And then a process was created and started. The missing values can be calculated with the operator (decision tree) that we use in this assignment.



Filter (6 / 6 attributes): Search for Attributes

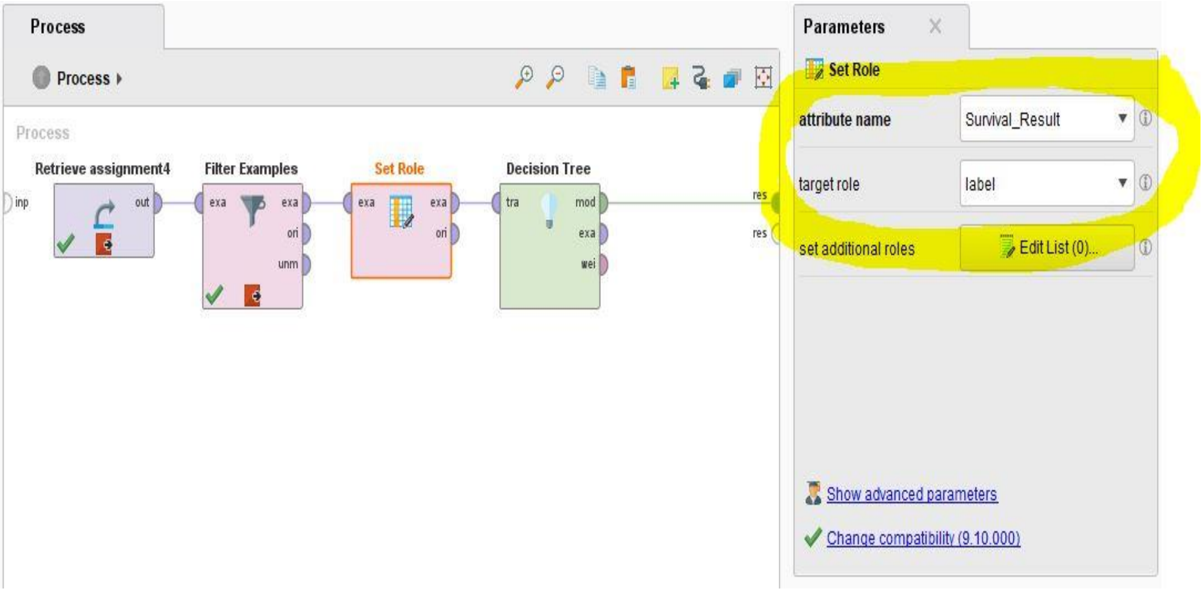
Name	Type	Missing	Statistics		
Passenger ID	Integer	0	Min 1	Max 399	Average 195.457
Passenger Class	Integer	0	Min 1	Max 3	Average 2.273
⚠ Name	Nominal	0	Least van Bill [...] lyler (1)	Most Abbott, [...] Hunt) (1)	Values Abbott, [...] osa Hunt) (1), Abels
⚠ Sex	Nominal	0	Least female (112)	Most male (199)	Values male (199), female (112)
⚠ Age	Integer	1	Min 1	Max 71	Average 28.906
Survival_Result	Nominal	1	Least Lost (6)	Most Died (198)	Values Died (198), Survived (106), ...[1

Showing attributes 1 - 6

Examples: 311 Special Attributes: 0 Regular Attributes: 6

Format Data

We use '**Set Role**' operator and select the **survival_result** from the attribute name and set the target role to **label**.



Parameters

Set Role

attribute name: Survival_Result

target role: label

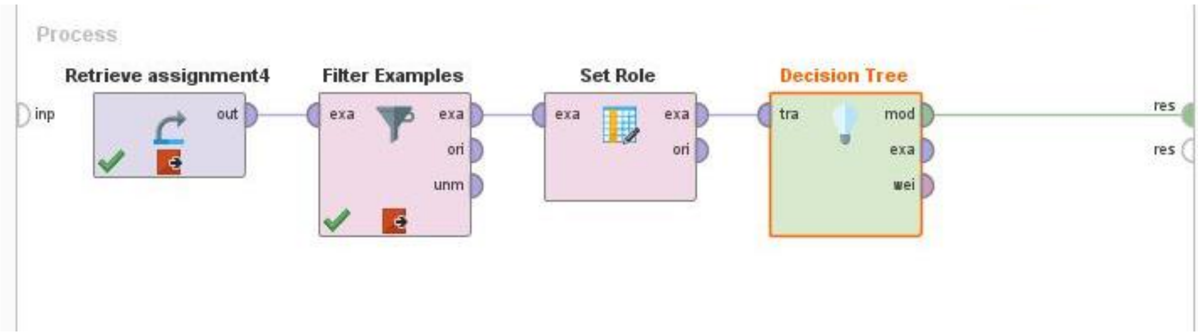
set additional roles: Edit List (0)...

Show advanced parameters

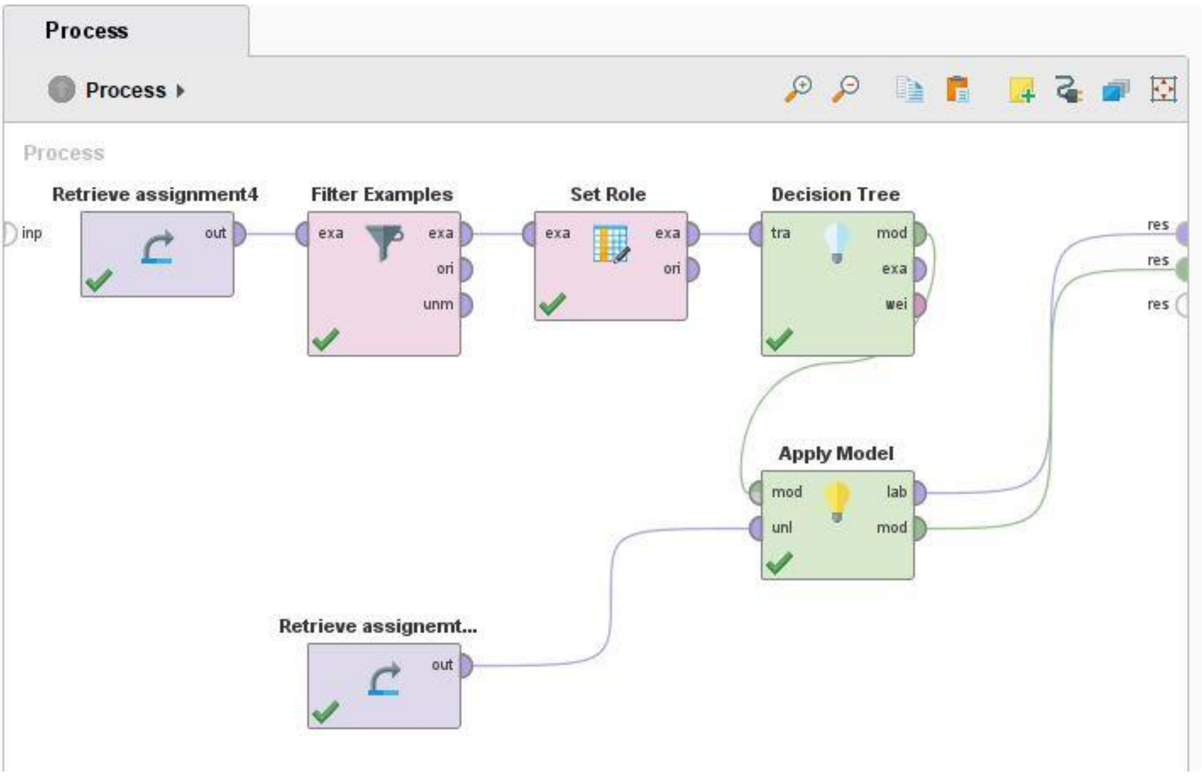
Change compatibility (9.10.000)

Modeling:

In this model we use the decision tree operator to find the gain_ratio and gini_index for the survived passengers.



The same model was used for the both gain_ratio and gini_index.

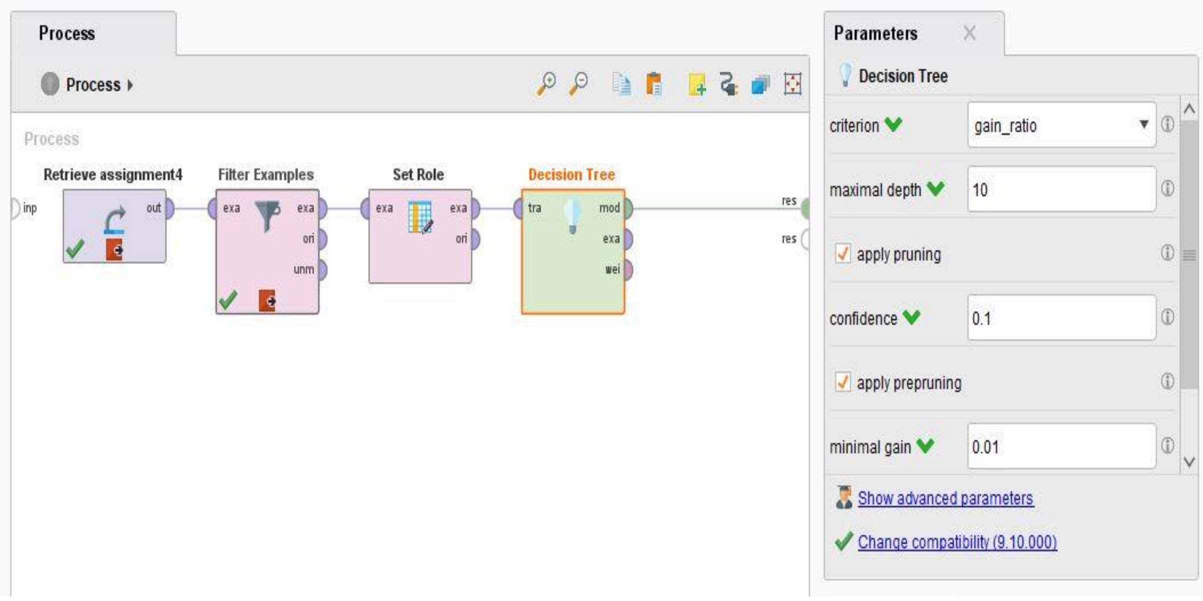


The above model was used as the second process after we removed the survival_result attribute from the first dataset. And in this new dataset we add 3 more names from our friend list.

Gain_Ratio:

Out of 312 there are 78 survived and the maximum strength of the surviving are the males.

The below figures shows the result we get from the decision tree after optioning the gain_ratio.



criterion: gain_ratio

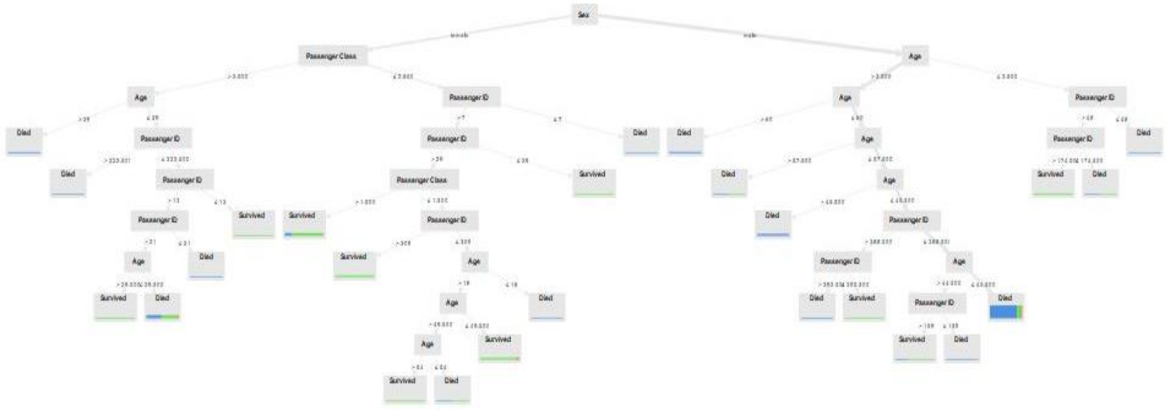
maximal depth: 10

☑ apply pruning

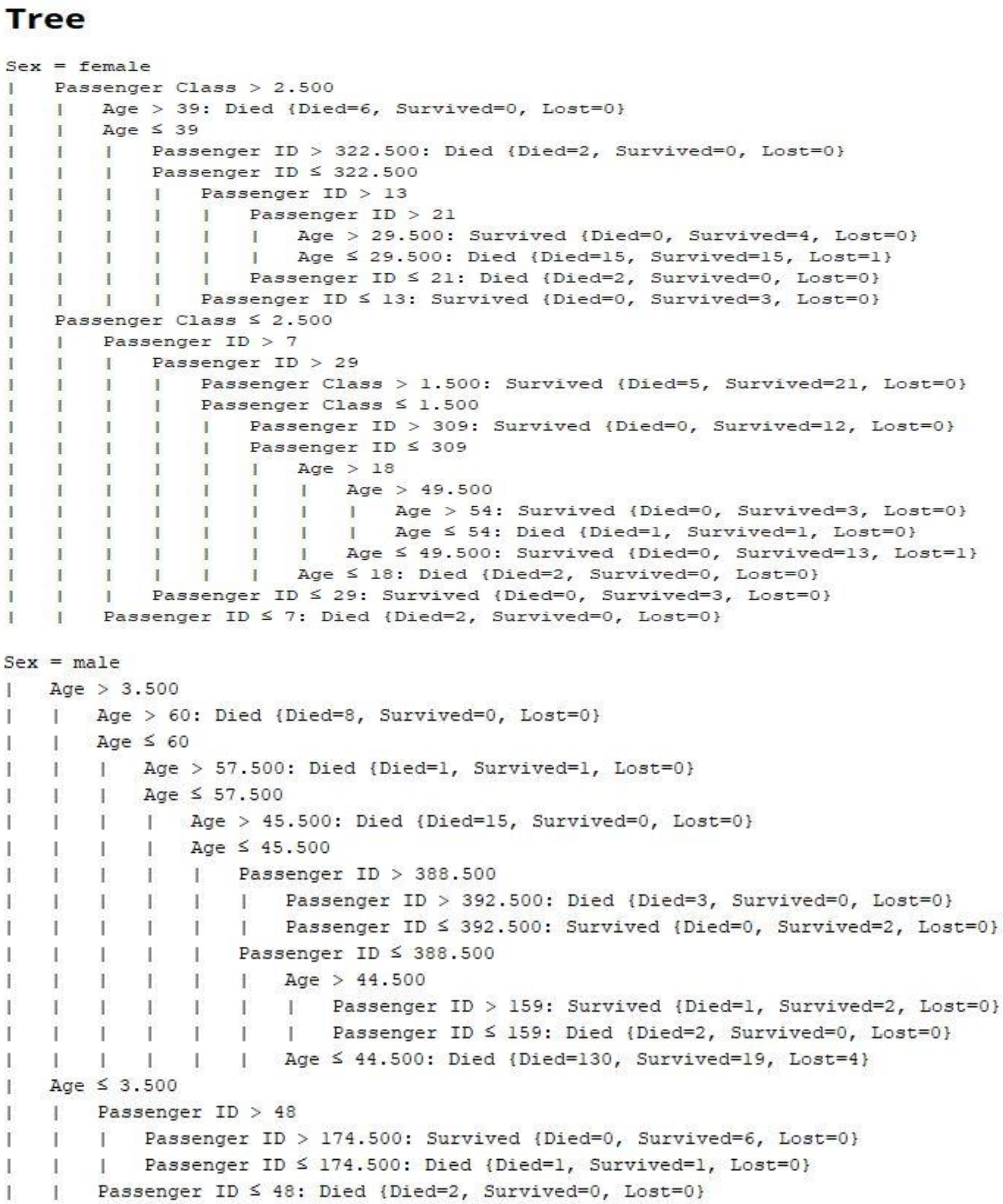
confidence: 0.1

☑ apply prepruning

minimal gain: 0.01



This is the decision tree after we run the process.

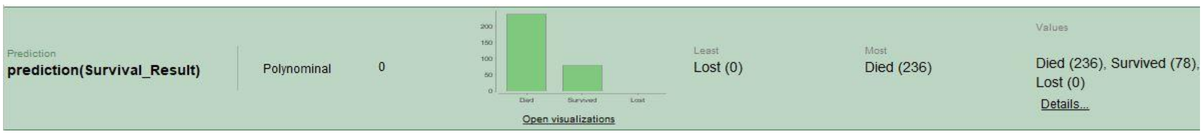


Tree

```
Sex = female
|   Passenger Class > 2.500
|   |   Age > 39: Died {Died=6, Survived=0, Lost=0}
|   |   Age ≤ 39
|   |   |   Passenger ID > 322.500: Died {Died=2, Survived=0, Lost=0}
|   |   |   Passenger ID ≤ 322.500
|   |   |   |   Passenger ID > 13
|   |   |   |   |   Passenger ID > 21
|   |   |   |   |   |   Age > 29.500: Survived {Died=0, Survived=4, Lost=0}
|   |   |   |   |   |   Age ≤ 29.500: Died {Died=15, Survived=15, Lost=1}
|   |   |   |   |   Passenger ID ≤ 21: Died {Died=2, Survived=0, Lost=0}
|   |   |   |   Passenger ID ≤ 13: Survived {Died=0, Survived=3, Lost=0}
|   Passenger Class ≤ 2.500
|   |   Passenger ID > 7
|   |   |   Passenger ID > 29
|   |   |   |   Passenger Class > 1.500: Survived {Died=5, Survived=21, Lost=0}
|   |   |   |   Passenger Class ≤ 1.500
|   |   |   |   |   Passenger ID > 309: Survived {Died=0, Survived=12, Lost=0}
|   |   |   |   |   Passenger ID ≤ 309
|   |   |   |   |   |   Age > 18
|   |   |   |   |   |   |   Age > 49.500
|   |   |   |   |   |   |   |   Age > 54: Survived {Died=0, Survived=3, Lost=0}
|   |   |   |   |   |   |   |   Age ≤ 54: Died {Died=1, Survived=1, Lost=0}
|   |   |   |   |   |   |   Age ≤ 49.500: Survived {Died=0, Survived=13, Lost=1}
|   |   |   |   |   |   Age ≤ 18: Died {Died=2, Survived=0, Lost=0}
|   |   |   Passenger ID ≤ 29: Survived {Died=0, Survived=3, Lost=0}
|   |   Passenger ID ≤ 7: Died {Died=2, Survived=0, Lost=0}

Sex = male
|   Age > 3.500
|   |   Age > 60: Died {Died=8, Survived=0, Lost=0}
|   |   Age ≤ 60
|   |   |   Age > 57.500: Died {Died=1, Survived=1, Lost=0}
|   |   |   Age ≤ 57.500
|   |   |   |   Age > 45.500: Died {Died=15, Survived=0, Lost=0}
|   |   |   |   Age ≤ 45.500
|   |   |   |   |   Passenger ID > 388.500
|   |   |   |   |   |   Passenger ID > 392.500: Died {Died=3, Survived=0, Lost=0}
|   |   |   |   |   |   Passenger ID ≤ 392.500: Survived {Died=0, Survived=2, Lost=0}
|   |   |   |   |   Passenger ID ≤ 388.500
|   |   |   |   |   |   Age > 44.500
|   |   |   |   |   |   |   Passenger ID > 159: Survived {Died=1, Survived=2, Lost=0}
|   |   |   |   |   |   |   Passenger ID ≤ 159: Died {Died=2, Survived=0, Lost=0}
|   |   |   |   |   |   Age ≤ 44.500: Died {Died=130, Survived=19, Lost=4}
|   Age ≤ 3.500
|   |   Passenger ID > 48
|   |   |   Passenger ID > 174.500: Survived {Died=0, Survived=6, Lost=0}
|   |   |   Passenger ID ≤ 174.500: Died {Died=1, Survived=1, Lost=0}
|   |   Passenger ID ≤ 48: Died {Died=2, Survived=0, Lost=0}
```

The above 2 figures shows the details of decision tree for the gain_ratio factor.



Prediction
prediction(Survival_Result)
Polynominal
0

Least
Lost (0)

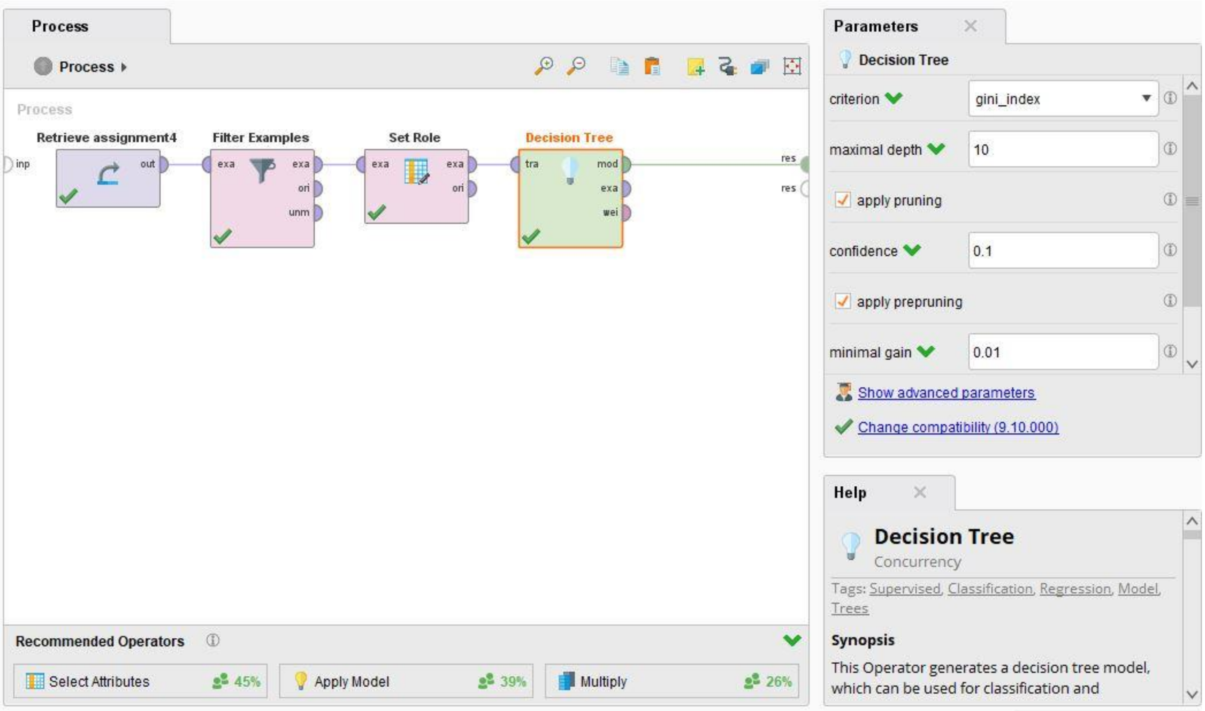
Most
Died (236)

Values
Died (236), Survived (78), Lost (0)
Details...

This above figure shows the results for the survival passengers out of our dataset.

<u>Gini_Index:</u>

From the dataset of 312 passengers the survived passengers are 105 with the males passengers are in the majority.



Process

Process ▸

Process

Retrieve assignment4 → Filter Examples → Set Role → Decision Tree

Parameters

Decision Tree

criterion ✓ gini_index

maximal depth ✓ 10

☑ apply pruning

confidence ✓ 0.1

☑ apply prepruning

minimal gain ✓ 0.01

Show advanced parameters

Change compatibility (9.10.000)

Help

Decision Tree

Concurrency

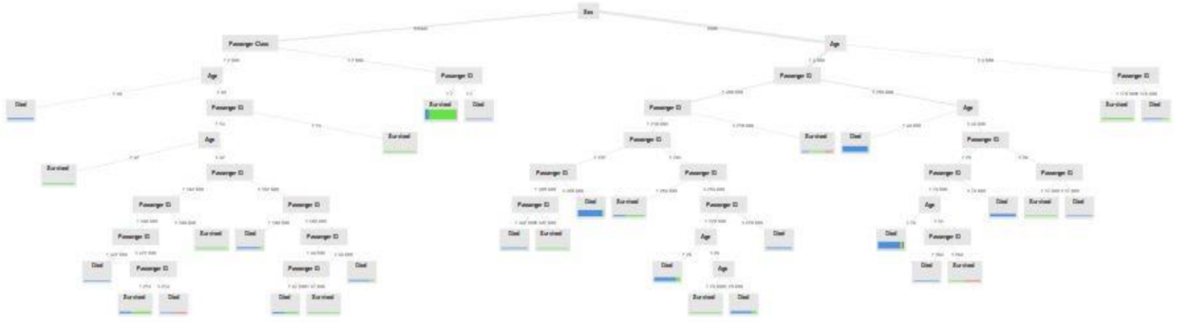
Tags: Supervised, Classification, Regression, Model, Trees

Synopsis

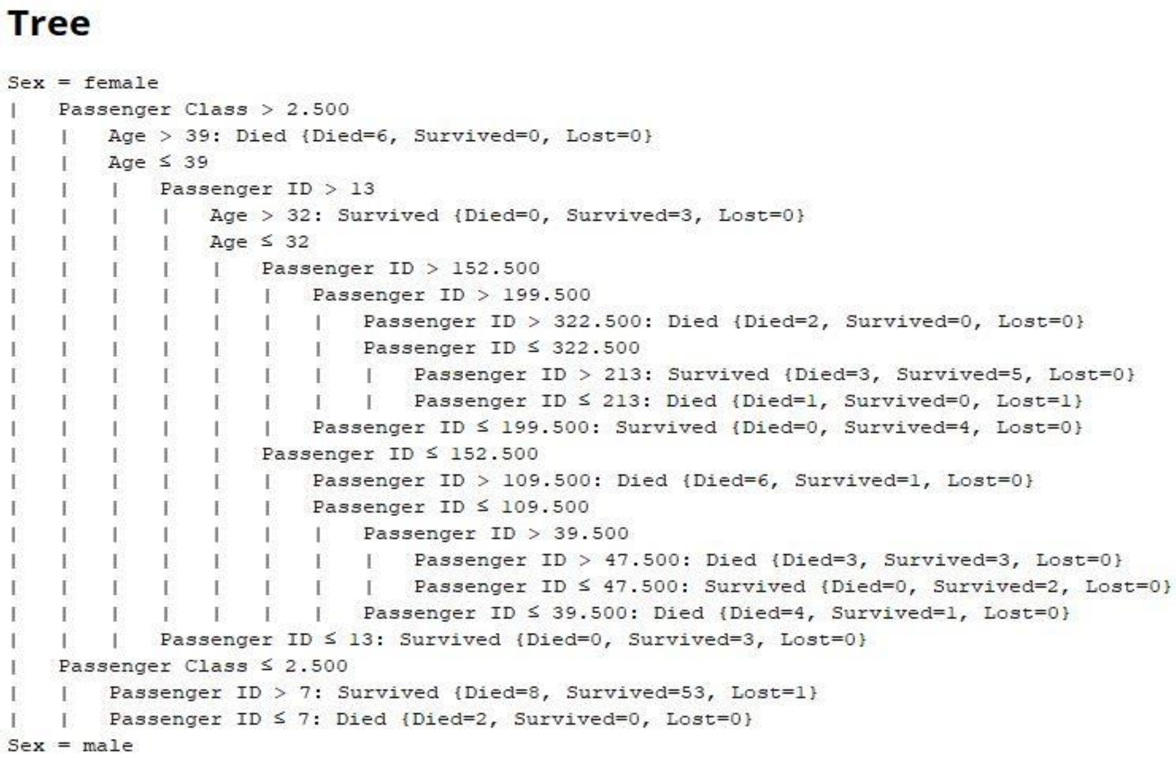
This Operator generates a decision tree model, which can be used for classification and

Recommended Operators

Select Attributes 45% | Apply Model 39% | Multiply 26%

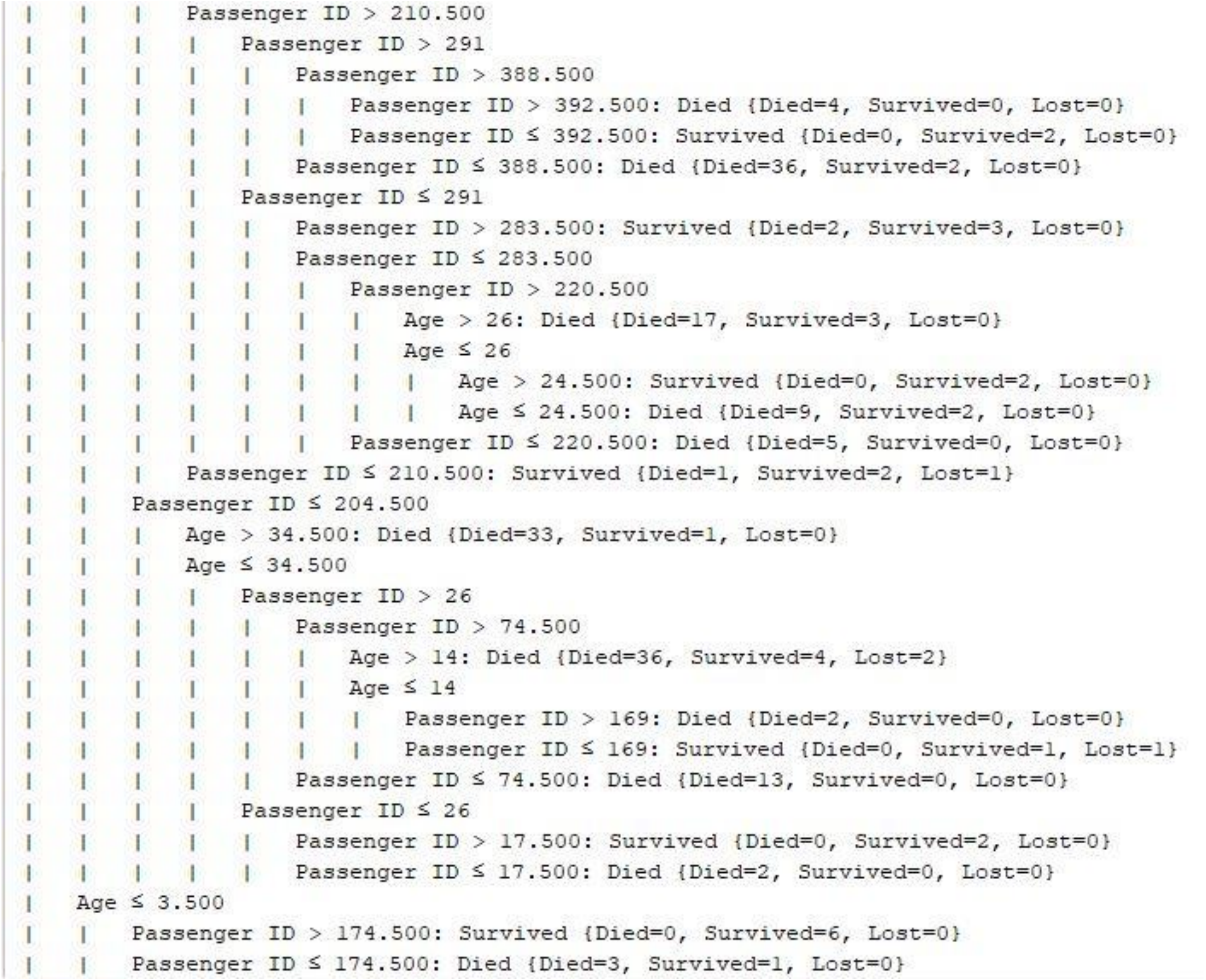


This is the decision tree after we run the process.



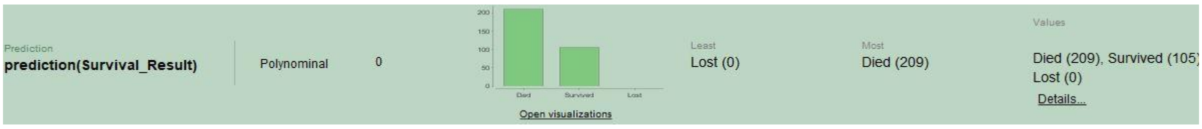
Tree

```
Sex = female
|   Passenger Class > 2.500
|   |   Age > 39: Died {Died=6, Survived=0, Lost=0}
|   |   Age ≤ 39
|   |   |   Passenger ID > 13
|   |   |   |   Age > 32: Survived {Died=0, Survived=3, Lost=0}
|   |   |   |   Age ≤ 32
|   |   |   |   |   Passenger ID > 152.500
|   |   |   |   |   |   Passenger ID > 199.500
|   |   |   |   |   |   |   Passenger ID > 322.500: Died {Died=2, Survived=0, Lost=0}
|   |   |   |   |   |   |   Passenger ID ≤ 322.500
|   |   |   |   |   |   |   |   Passenger ID > 213: Survived {Died=3, Survived=5, Lost=0}
|   |   |   |   |   |   |   |   Passenger ID ≤ 213: Died {Died=1, Survived=0, Lost=1}
|   |   |   |   |   |   Passenger ID ≤ 199.500: Survived {Died=0, Survived=4, Lost=0}
|   |   |   |   |   Passenger ID ≤ 152.500
|   |   |   |   |   |   Passenger ID > 109.500: Died {Died=6, Survived=1, Lost=0}
|   |   |   |   |   |   Passenger ID ≤ 109.500
|   |   |   |   |   |   |   Passenger ID > 39.500
|   |   |   |   |   |   |   |   Passenger ID > 47.500: Died {Died=3, Survived=3, Lost=0}
|   |   |   |   |   |   |   |   Passenger ID ≤ 47.500: Survived {Died=0, Survived=2, Lost=0}
|   |   |   |   |   |   |   Passenger ID ≤ 39.500: Died {Died=4, Survived=1, Lost=0}
|   |   |   Passenger ID ≤ 13: Survived {Died=0, Survived=3, Lost=0}
|   Passenger Class ≤ 2.500
|   |   Passenger ID > 7: Survived {Died=8, Survived=53, Lost=1}
|   |   Passenger ID ≤ 7: Died {Died=2, Survived=0, Lost=0}
Sex = male
```



```
|   |   |   Passenger ID > 210.500
|   |   |   |   Passenger ID > 291
|   |   |   |   |   Passenger ID > 388.500
|   |   |   |   |   |   Passenger ID > 392.500: Died {Died=4, Survived=0, Lost=0}
|   |   |   |   |   |   Passenger ID ≤ 392.500: Survived {Died=0, Survived=2, Lost=0}
|   |   |   |   |   Passenger ID ≤ 388.500: Died {Died=36, Survived=2, Lost=0}
|   |   |   |   Passenger ID ≤ 291
|   |   |   |   |   Passenger ID > 283.500: Survived {Died=2, Survived=3, Lost=0}
|   |   |   |   |   Passenger ID ≤ 283.500
|   |   |   |   |   |   Passenger ID > 220.500
|   |   |   |   |   |   |   Age > 26: Died {Died=17, Survived=3, Lost=0}
|   |   |   |   |   |   |   Age ≤ 26
|   |   |   |   |   |   |   |   Age > 24.500: Survived {Died=0, Survived=2, Lost=0}
|   |   |   |   |   |   |   |   Age ≤ 24.500: Died {Died=9, Survived=2, Lost=0}
|   |   |   |   |   |   Passenger ID ≤ 220.500: Died {Died=5, Survived=0, Lost=0}
|   |   |   Passenger ID ≤ 210.500: Survived {Died=1, Survived=2, Lost=1}
|   |   Passenger ID ≤ 204.500
|   |   |   Age > 34.500: Died {Died=33, Survived=1, Lost=0}
|   |   |   Age ≤ 34.500
|   |   |   |   Passenger ID > 26
|   |   |   |   |   Passenger ID > 74.500
|   |   |   |   |   |   Age > 14: Died {Died=36, Survived=4, Lost=2}
|   |   |   |   |   |   Age ≤ 14
|   |   |   |   |   |   |   Passenger ID > 169: Died {Died=2, Survived=0, Lost=0}
|   |   |   |   |   |   |   Passenger ID ≤ 169: Survived {Died=0, Survived=1, Lost=1}
|   |   |   |   |   Passenger ID ≤ 74.500: Died {Died=13, Survived=0, Lost=0}
|   |   |   |   Passenger ID ≤ 26
|   |   |   |   |   Passenger ID > 17.500: Survived {Died=0, Survived=2, Lost=0}
|   |   |   |   |   Passenger ID ≤ 17.500: Died {Died=2, Survived=0, Lost=0}
|   Age ≤ 3.500
|   |   Passenger ID > 174.500: Survived {Died=0, Survived=6, Lost=0}
|   |   Passenger ID ≤ 174.500: Died {Died=3, Survived=1, Lost=0}
```

The above 2 figures are the description of the decision tree.



Prediction
prediction(Survival_Result)

Polynominal 0

Least
Lost (0)

Most
Died (209)

Values
Died (209), Survived (105
Lost (0)
Details...

This above figure shows the results for the survival passengers out of our dataset.

Report:

Run your model using gain_ratio. Report your tree nodes, and discuss whether you and the people you know would have lived, died or been lost.

We added 3 people from our side and after running the model we predict that all are died.

Re-run your model using gini_index. Report differences in your tree structures. Discuss whether your chances for survival increase under Gini or Gain.

After running the both the models we get that the differences among the survived passengers are increased in gini_index rather than the gain_ratio model. The survival chances was increased in the gini_index.

References:

Dataset

https://gist.github.com/aficionado/7743748

https://en.wikipedia.org/wiki/Passengers_of_the_Titanic#cite_ref-nasserbaby_91-0

Tutorial for Orange

https://www.youtube.com/watch?v=D6zd7m2aYqU

https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/data/selectcolumns.html