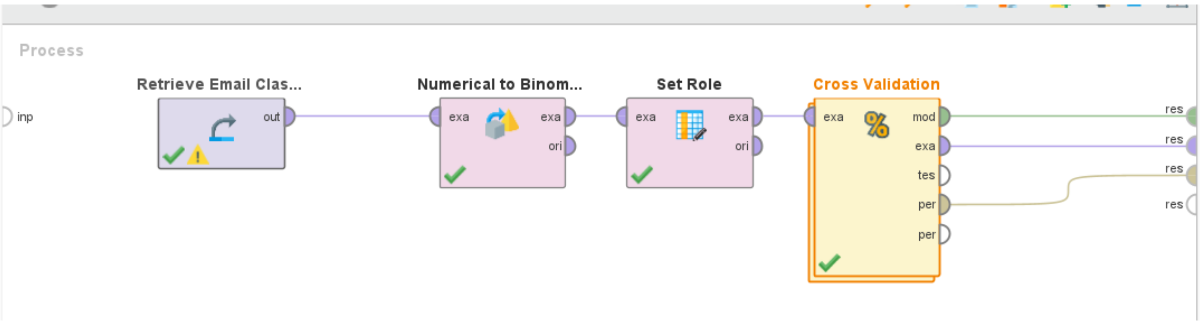
Classification Analysis

Name: Jahnvi Rameshbhai Patel

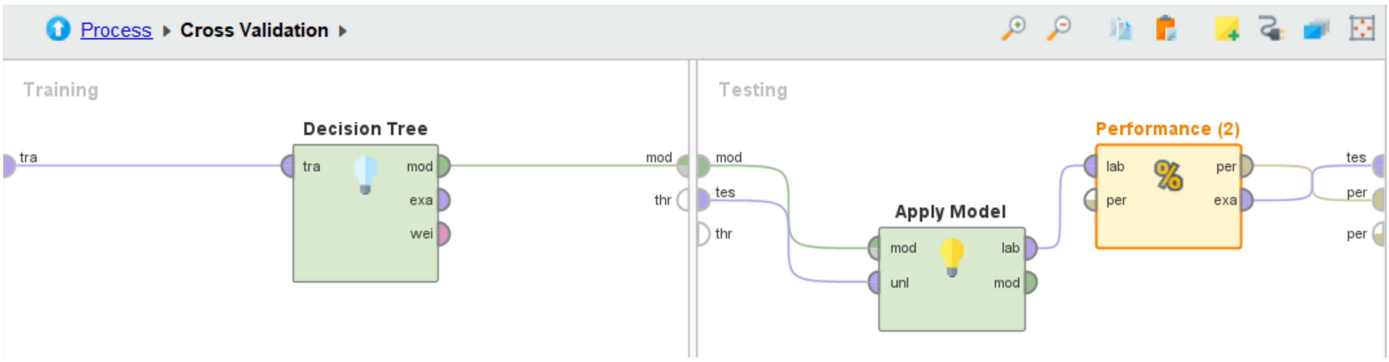
1: Email Classification Dataset

Model 1: Decision tree

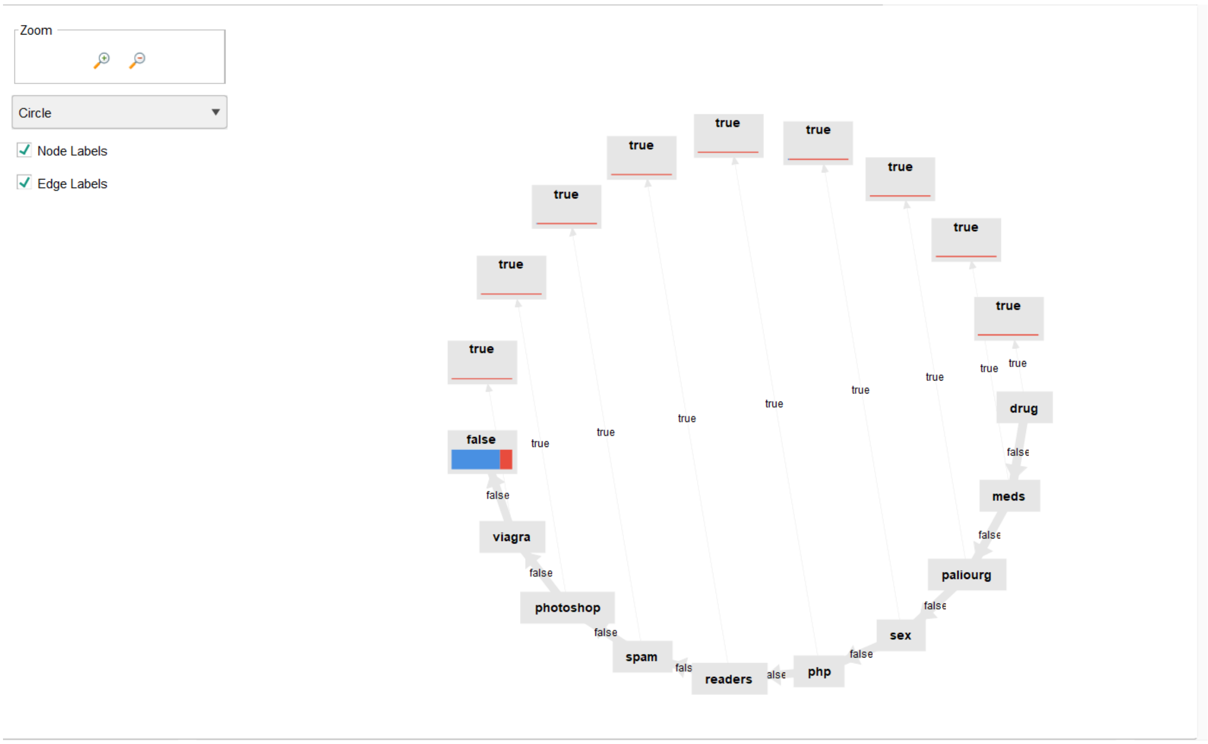
Process:



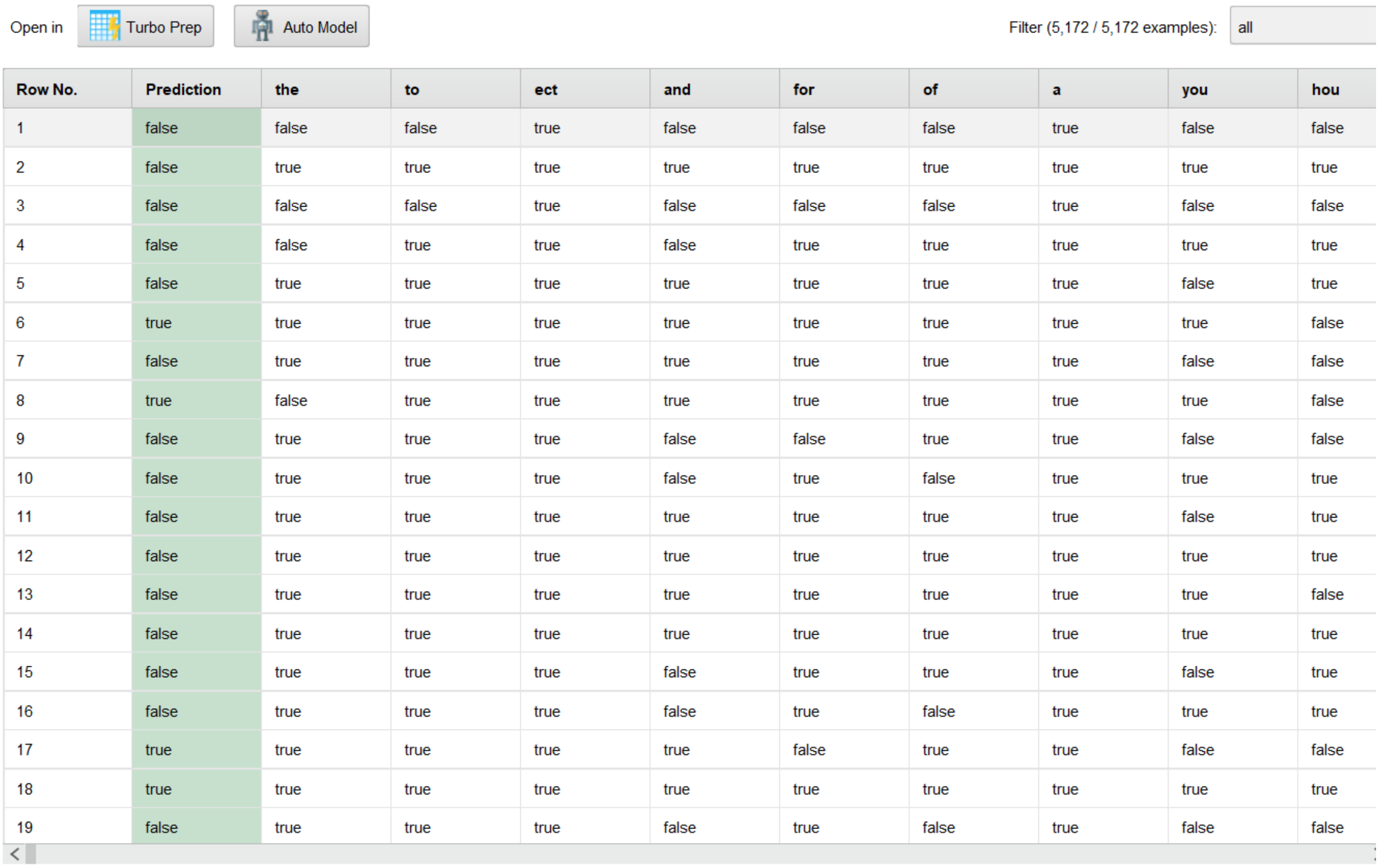
Cross validation:



Tree:



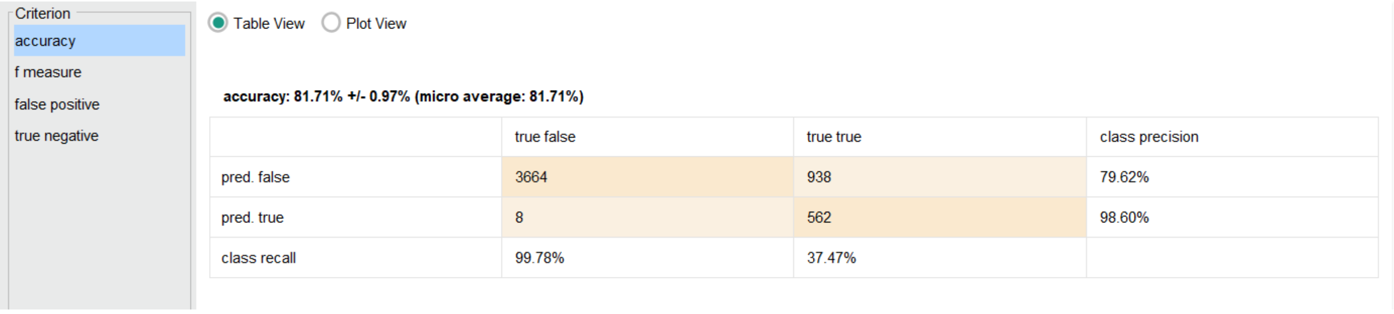
Example set:



Filter (5,172 / 5,172 examples): all

Row No.	Prediction	the	to	ect	and	for	of	a	you	hou
1	false	false	false	true	false	false	false	true	false	false
2	false	true	true	true	true	true	true	true	true	true
3	false	false	false	true	false	false	false	true	false	false
4	false	false	true	true	false	true	true	true	true	true
5	false	true	true	true	true	true	true	true	false	true
6	true	true	true	true	true	true	true	true	true	false
7	false	true	true	true	true	true	true	true	false	false
8	true	false	true	true	true	true	true	true	true	false
9	false	true	true	true	false	false	true	true	false	false
10	false	true	true	true	false	true	false	true	true	true
11	false	true	true	true	true	true	true	true	false	true
12	false	true	true	true	true	true	true	true	true	true
13	false	true	true	true	true	true	true	true	true	false
14	false	true	true	true	true	true	true	true	true	true
15	false	true	true	true	false	true	true	true	false	true
16	false	true	true	true	false	true	false	true	true	true
17	true	true	true	true	true	false	true	true	false	false
18	true	true	true	true	true	true	true	true	true	true
19	false	true	true	true	false	true	false	true	false	false

Confusion matrix:



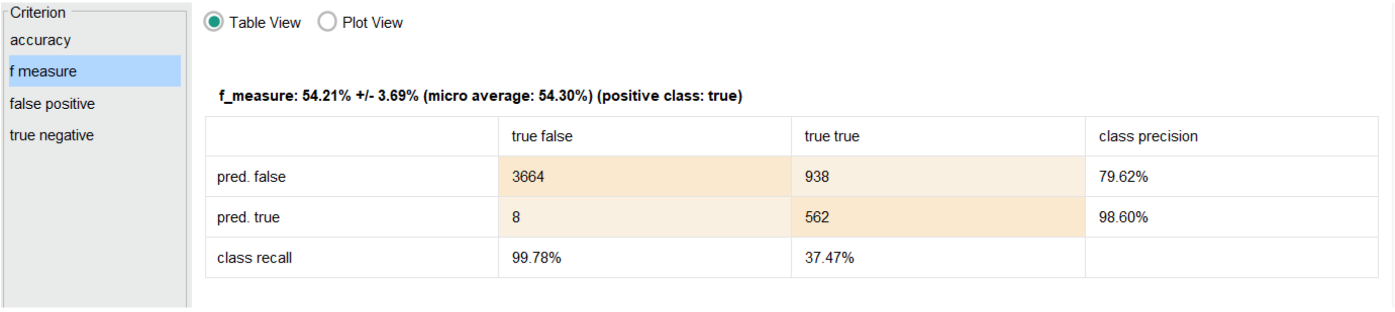
● Table View ○ Plot View

accuracy: 81.71% +/- 0.97% (micro average: 81.71%)

	true false	true true	class precision
pred. false	3664	938	79.62%
pred. true	8	562	98.60%
class recall	99.78%	37.47%	

In this case, an accuracy of 81.71% means that out of all the instances in the test set, the model correctly predicted the class label of 81.71% of them. While this accuracy may be considered high, it is important to note that the performance of the model can vary depending on the specific dataset and problem being tackled.

F measure:



● Table View ○ Plot View

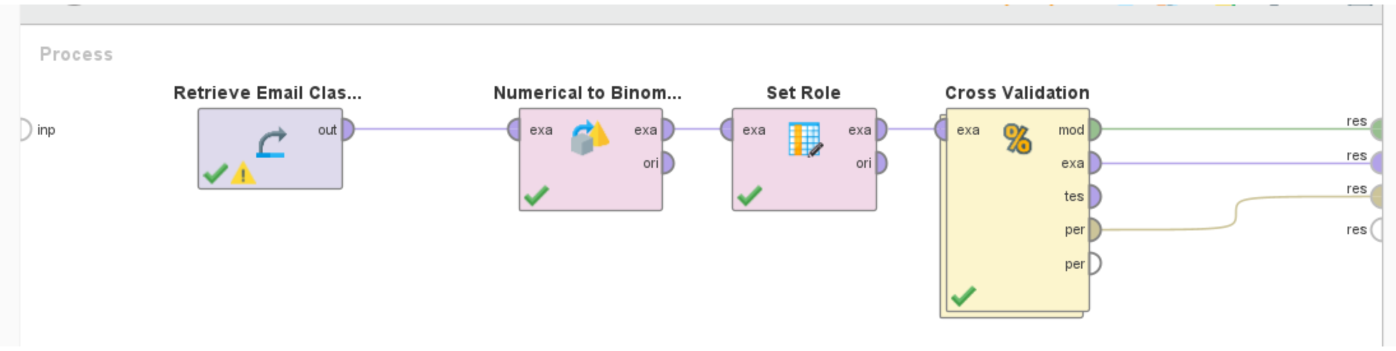
f_measure: 54.21% +/- 3.69% (micro average: 54.30%) (positive class: true)

	true false	true true	class precision
pred. false	3664	938	79.62%
pred. true	8	562	98.60%
class recall	99.78%	37.47%	

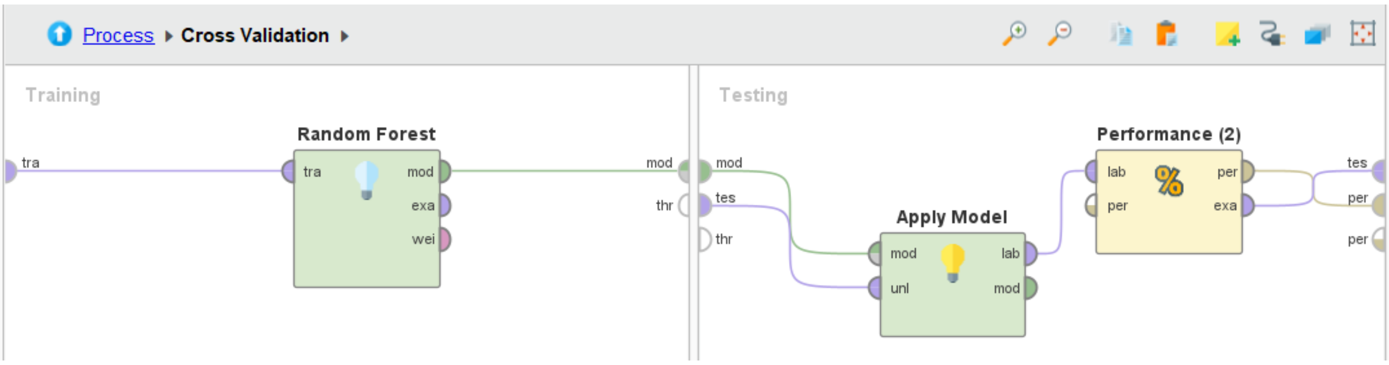
In this case, a value of 54.21% for the F-measure indicates that the model is not performing as well as it could be. This could be due to a variety of factors, such as imbalanced classes, noisy data, or poor feature selection.

Model 2: Random Forest

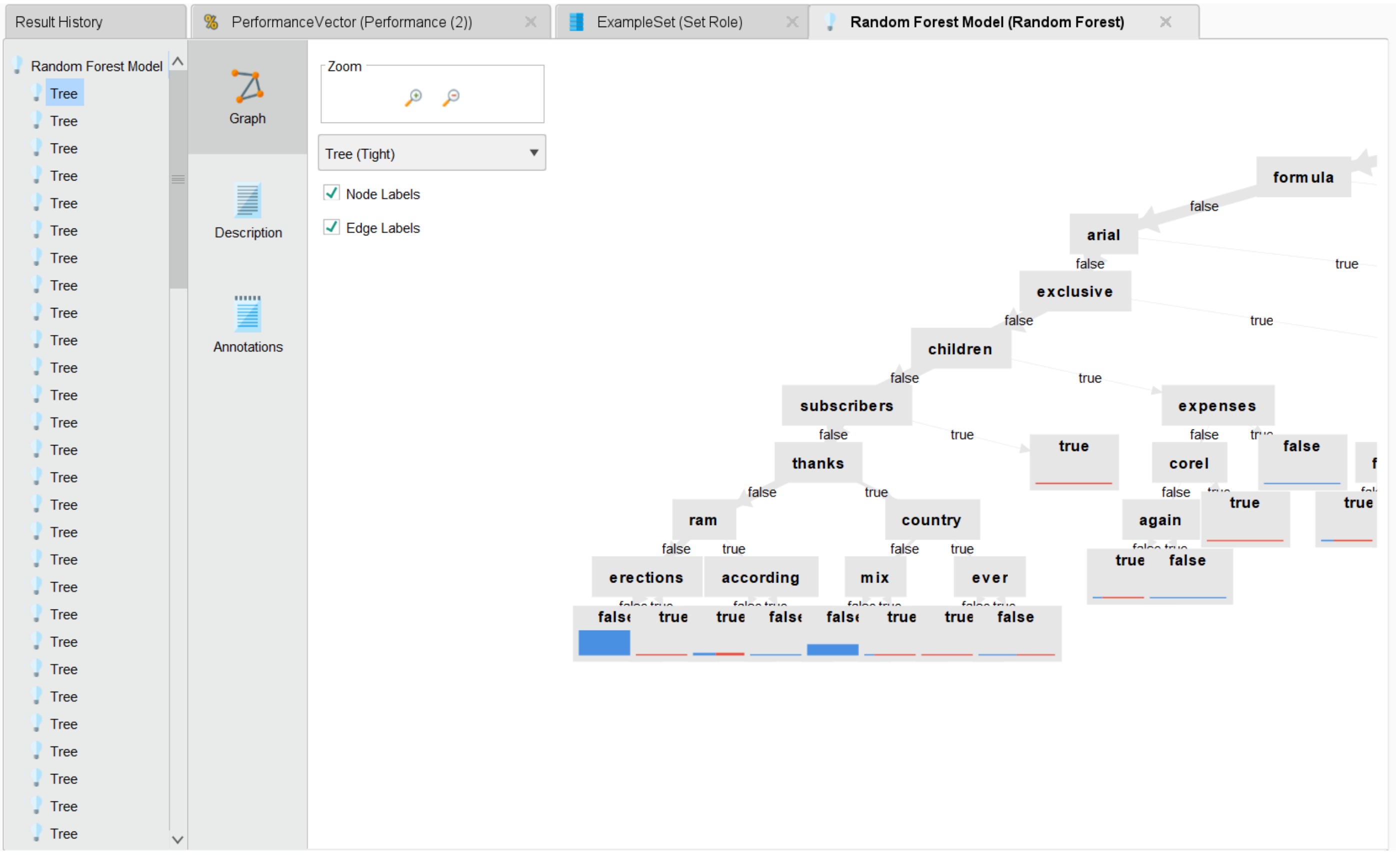
Process:



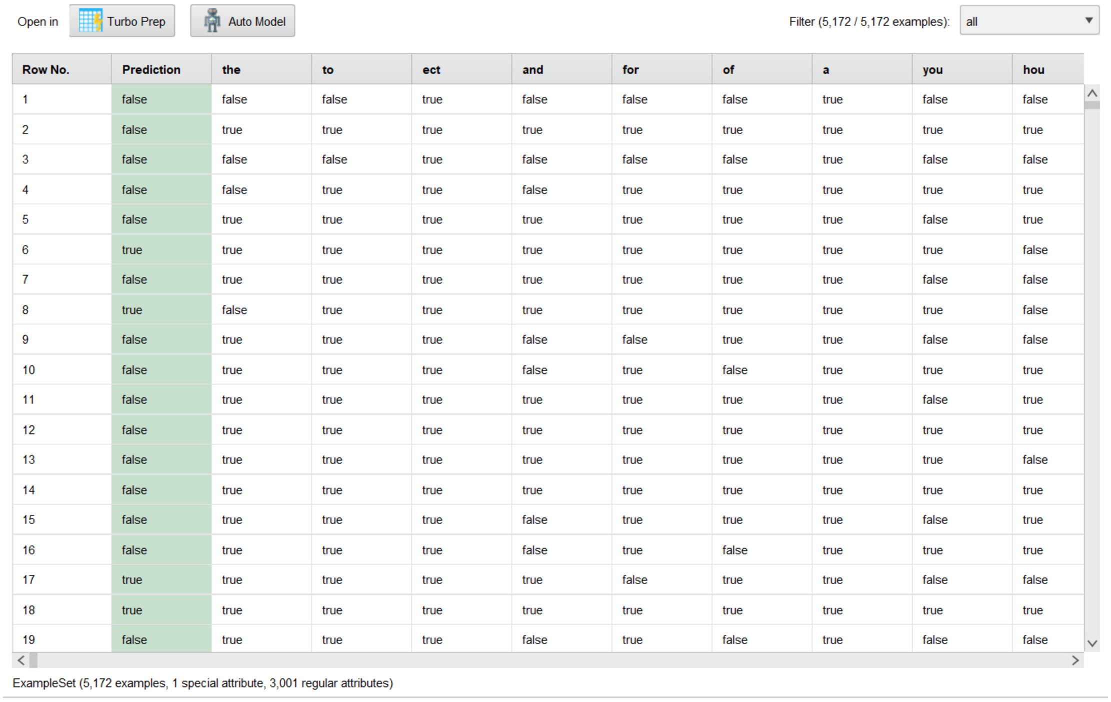
Cross validation:



Tree:



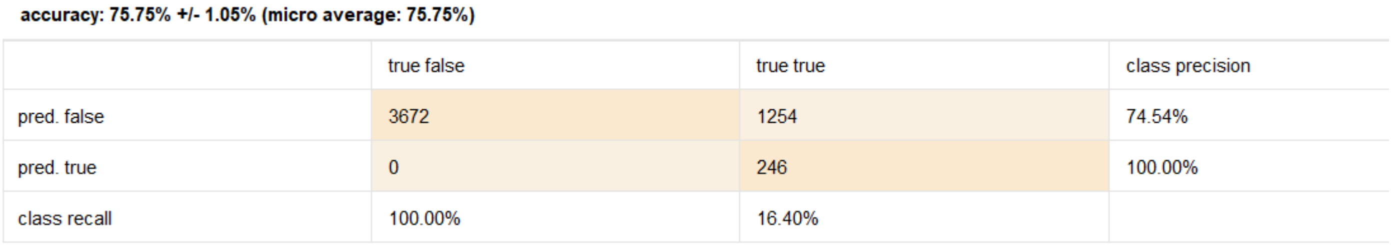
Example Set:



Row No.	Prediction	the	to	ect	and	for	of	a	you	hou
1	false	false	false	true	false	false	false	true	false	false
2	false	true	true	true	true	true	true	true	true	true
3	false	false	false	true	false	false	false	true	false	false
4	false	false	true	true	false	true	true	true	true	true
5	false	true	true	true	true	true	true	true	false	true
6	true	true	true	true	true	true	true	true	true	false
7	false	true	true	true	true	true	true	true	false	false
8	true	false	true	true	true	true	true	true	true	false
9	false	true	true	true	false	false	true	true	false	false
10	false	true	true	true	false	true	false	true	true	true
11	false	true	true	true	true	true	true	true	false	true
12	false	true	true	true	true	true	true	true	true	true
13	false	true	true	true	true	true	true	true	true	false
14	false	true	true	true	true	true	true	true	true	true
15	false	true	true	true	false	true	true	true	false	true
16	false	true	true	true	false	true	false	true	true	true
17	true	true	true	true	true	false	true	true	false	false
18	true	true	true	true	true	true	true	true	true	true
19	false	true	true	true	false	true	false	true	false	false

ExampleSet (5,172 examples, 1 special attribute, 3,001 regular attributes)

Confusion matrix:

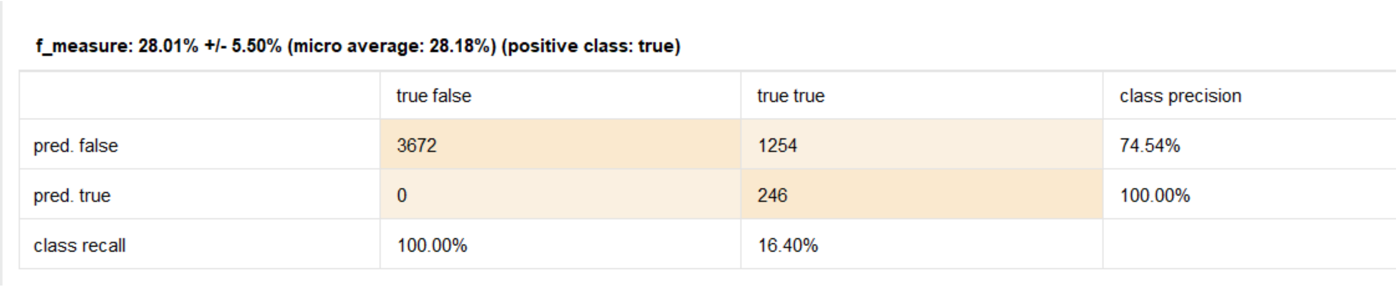


accuracy: 75.75% +/- 1.05% (micro average: 75.75%)

	true false	true true	class precision
pred. false	3672	1254	74.54%
pred. true	0	246	100.00%
class recall	100.00%	16.40%	

In a confusion matrix, the accuracy is calculated as the ratio of the number of correctly classified instances to the total number of instances in the test set. So, in this case, an accuracy of 75.75% means that out of all the instances in the test set, the model correctly predicted the class label of 75.75% of them.

F measure:



f_measure: 28.01% +/- 5.50% (micro average: 28.18%) (positive class: true)

	true false	true true	class precision
pred. false	3672	1254	74.54%
pred. true	0	246	100.00%
class recall	100.00%	16.40%	

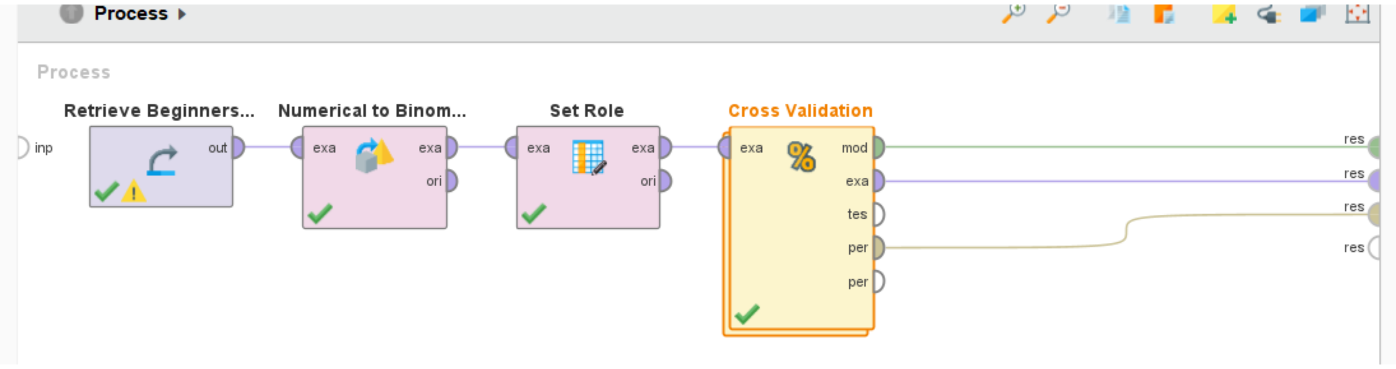
F-measure with a value of 28.01% in Random Forest using RapidMiner indicates that the model is not accurately classifying the instances in the test set, and its performance is poor. The model may require further tuning or optimization to improve its performance.

Considering both the models, a model with an accuracy of 81.71% is likely making more accurate predictions than a model with an accuracy of 75.75% and F-measure of 54.21% is relatively higher than an F-measure of 28.01%, which means that the model with an F-measure of 54.21% is likely performing better than the model with an F-measure of 28.01%. Hence, I will select model 1 based on the comparison.

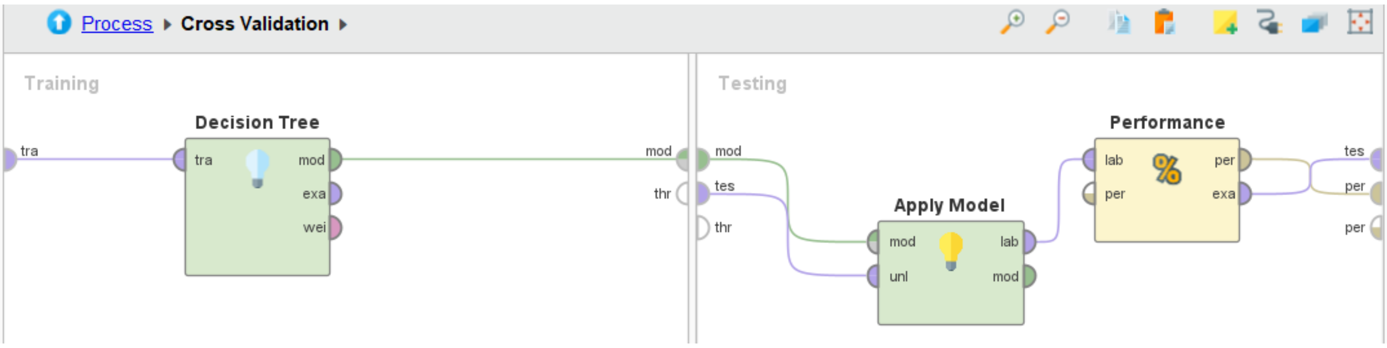
2: Beginners Classification Dataset

Model 1: Decision tree

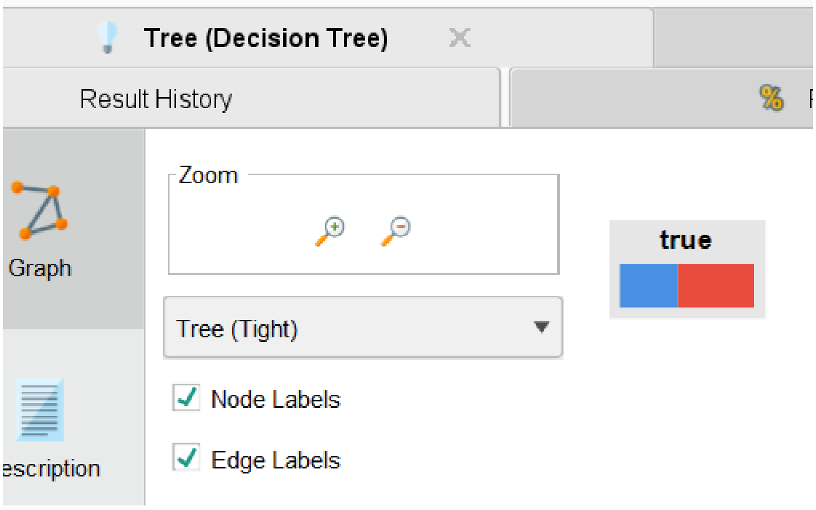
Process:



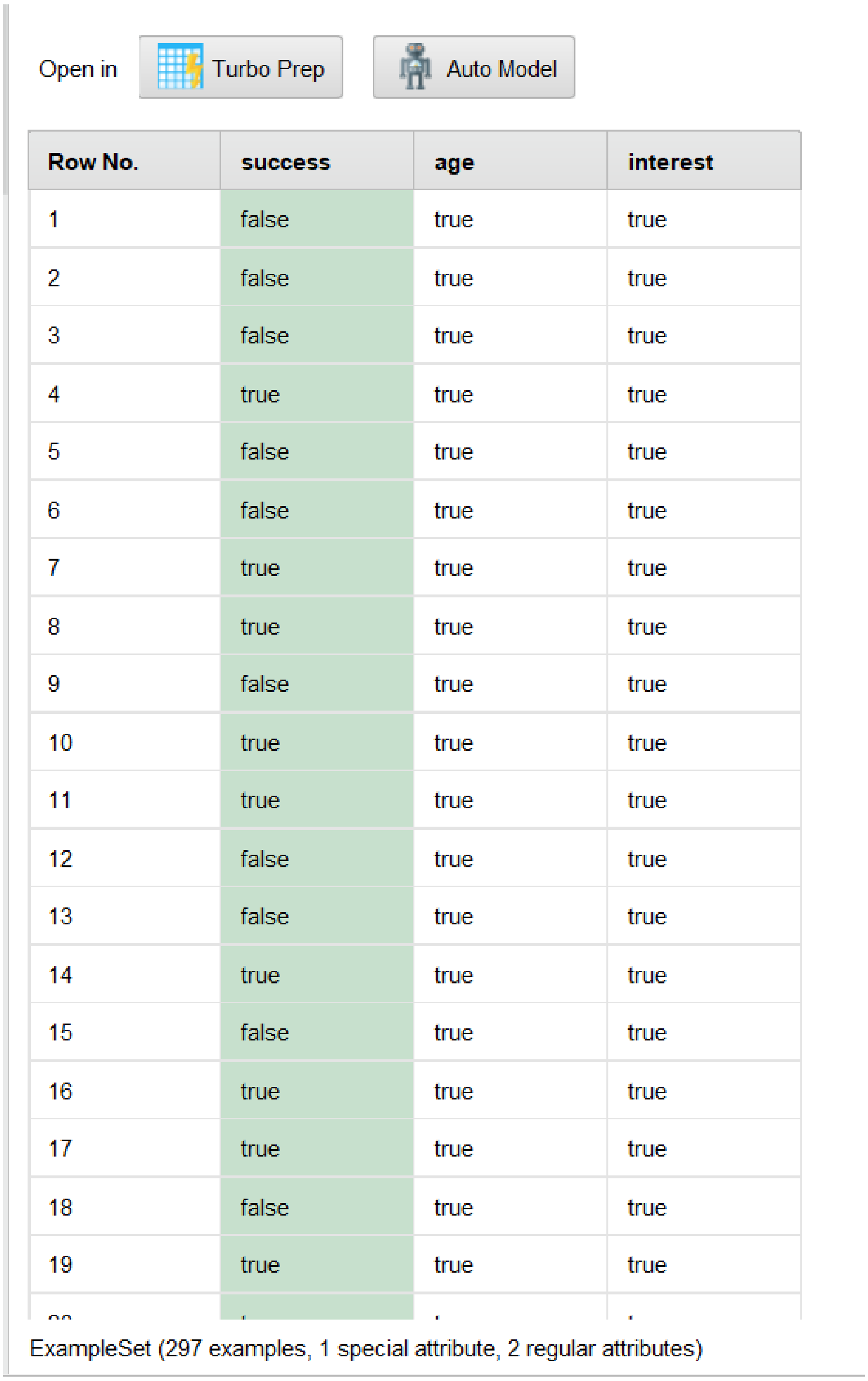
Cross validation:



Tree:



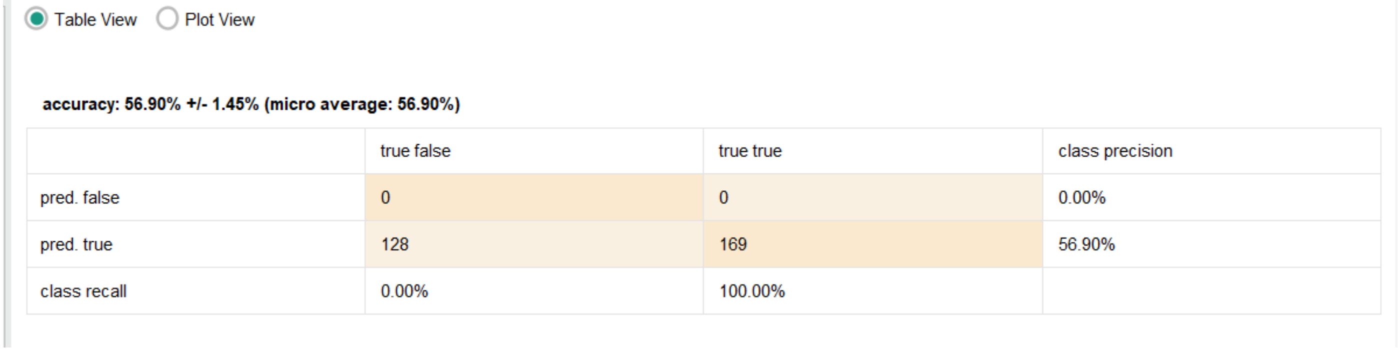
Example Set:



Row No.	success	age	interest
1	false	true	true
2	false	true	true
3	false	true	true
4	true	true	true
5	false	true	true
6	false	true	true
7	true	true	true
8	true	true	true
9	false	true	true
10	true	true	true
11	true	true	true
12	false	true	true
13	false	true	true
14	true	true	true
15	false	true	true
16	true	true	true
17	true	true	true
18	false	true	true
19	true	true	true

ExampleSet (297 examples, 1 special attribute, 2 regular attributes)

Confusion matrix:



● Table View ○ Plot View

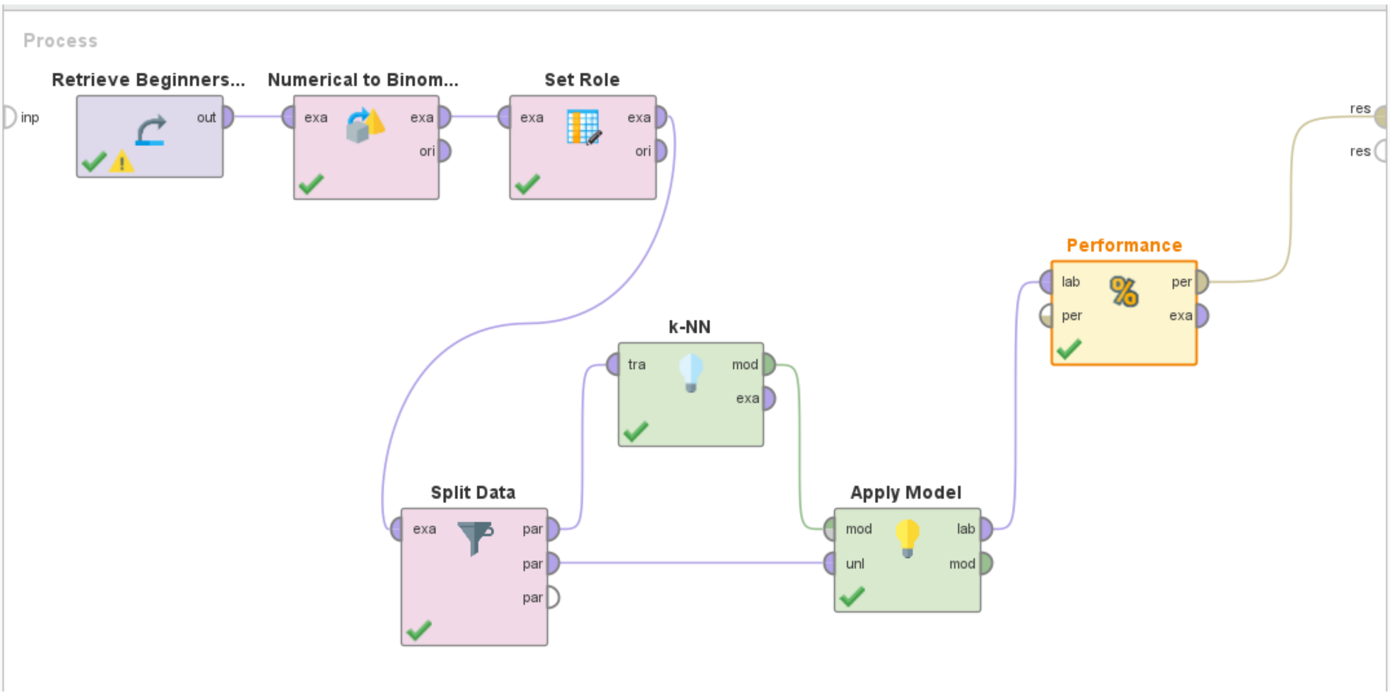
accuracy: 56.90% +/- 1.45% (micro average: 56.90%)

	true false	true true	class precision
pred. false	0	0	0.00%
pred. true	128	169	56.90%
class recall	0.00%	100.00%	

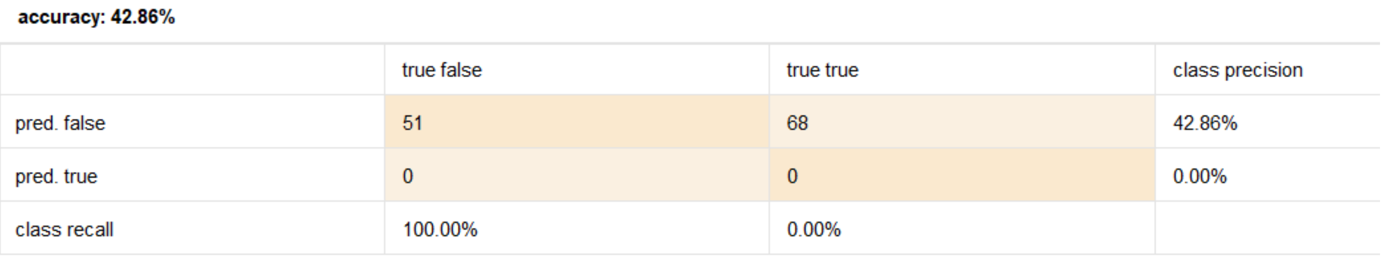
When the confusion matrix has an accuracy of 56.90% in decision tree using RapidMiner, it means that the model is making correct predictions for only 56.90% of the instances in the test set.

Model 2: KNN

Process:



Confusion matrix:



accuracy: 42.86%

	true false	true true	class precision
pred. false	51	68	42.86%
pred. true	0	0	0.00%
class recall	100.00%	0.00%	

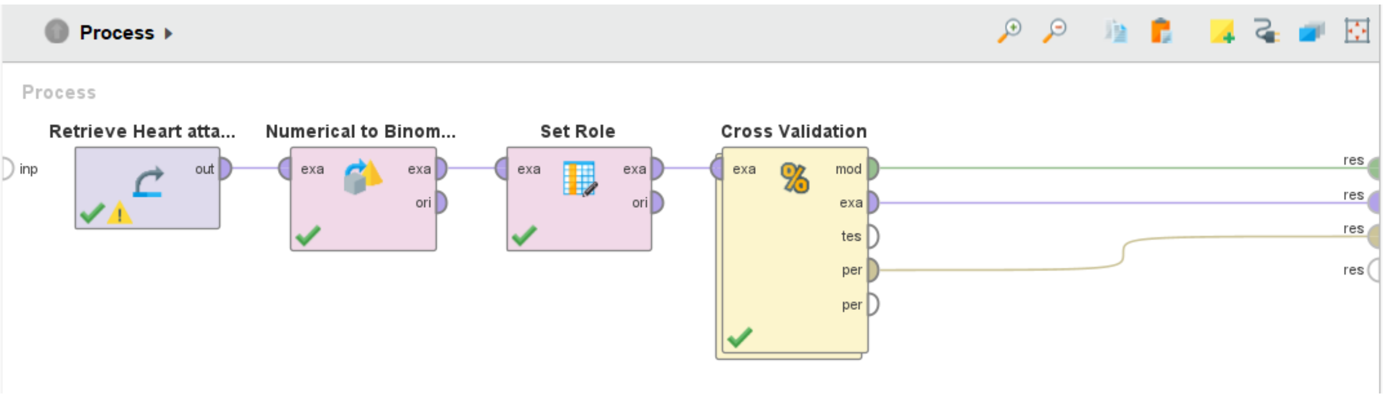
A 42.86% accuracy is not very high and suggests that the model is not performing well on the dataset.

In any case, an accuracy of 42.86% or 56.90% suggests that the model is not performing well on the dataset. However, if I have to choose one, I will go with model 1 in this case.

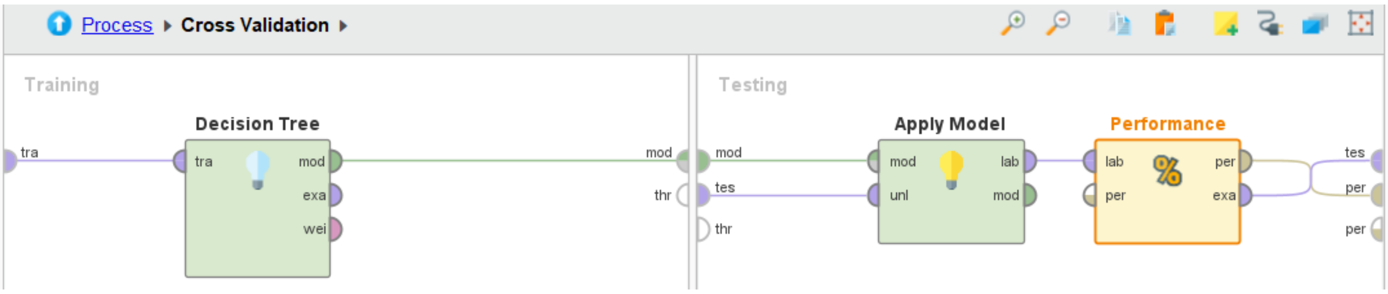
3: Heart attack Analysis

Model 1: Decision tree

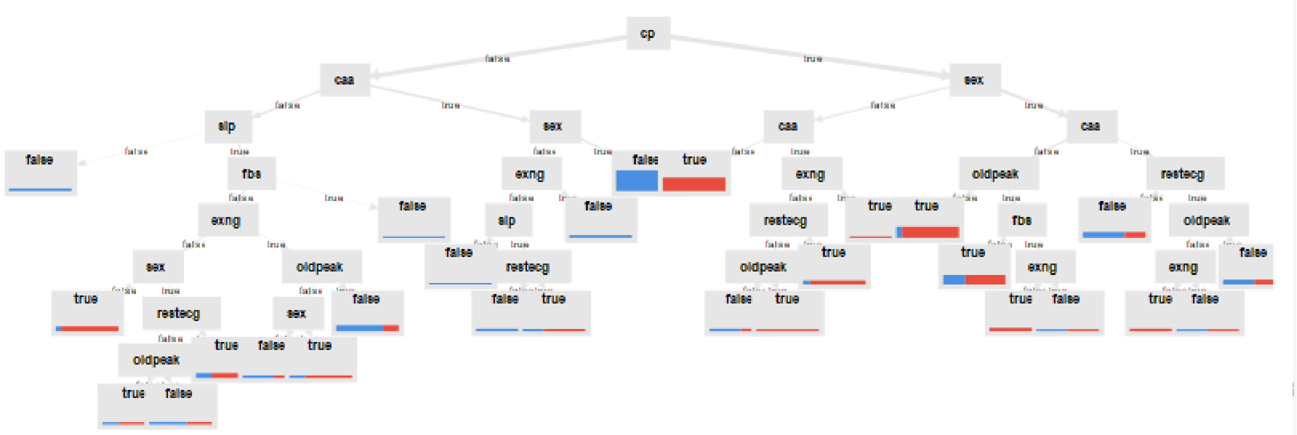
Process:



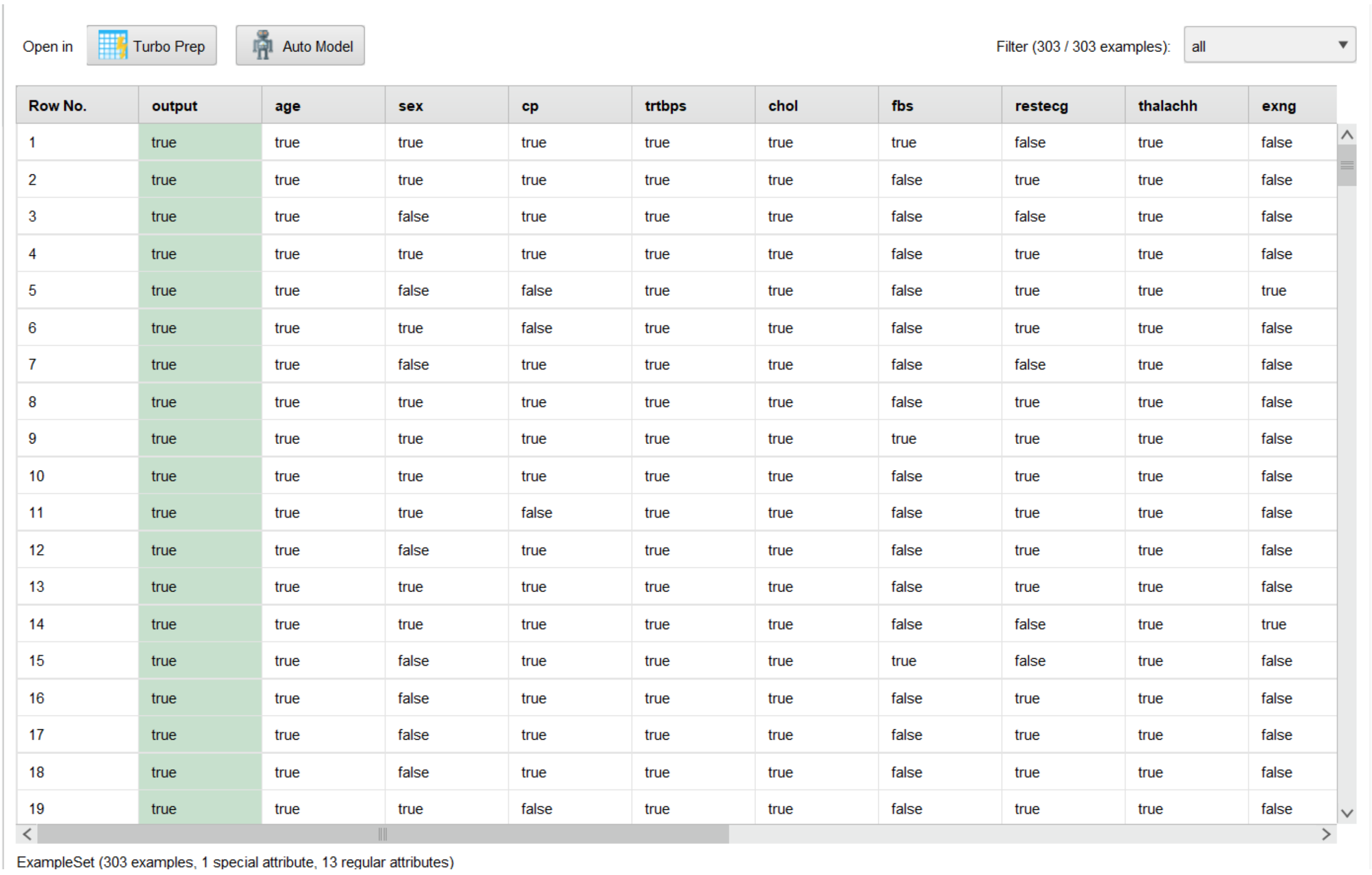
Cross validation:



Tree:



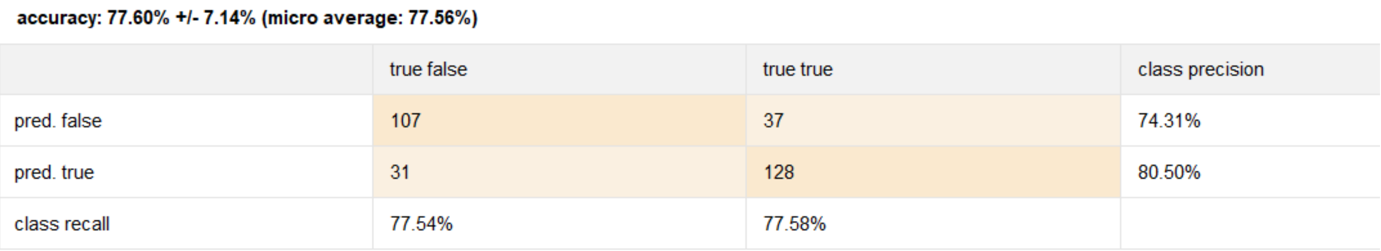
Example Set:



Row No.	output	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng
1	true	true	true	true	true	true	true	false	true	false
2	true	true	true	true	true	true	false	true	true	false
3	true	true	false	true	true	true	false	false	true	false
4	true	true	true	true	true	true	false	true	true	false
5	true	true	false	false	true	true	false	true	true	true
6	true	true	true	false	true	true	false	true	true	false
7	true	true	false	true	true	true	false	false	true	false
8	true	true	true	true	true	true	false	true	true	false
9	true	true	true	true	true	true	true	true	true	false
10	true	true	true	true	true	true	false	true	true	false
11	true	true	true	false	true	true	false	true	true	false
12	true	true	false	true	true	true	false	true	true	false
13	true	true	true	true	true	true	false	true	true	false
14	true	true	true	true	true	true	false	false	true	true
15	true	true	false	true	true	true	true	false	true	false
16	true	true	false	true	true	true	false	true	true	false
17	true	true	false	true	true	true	false	true	true	false
18	true	true	false	true	true	true	false	true	true	false
19	true	true	true	false	true	true	false	true	true	false

ExampleSet (303 examples, 1 special attribute, 13 regular attributes)

Confusion matrix:

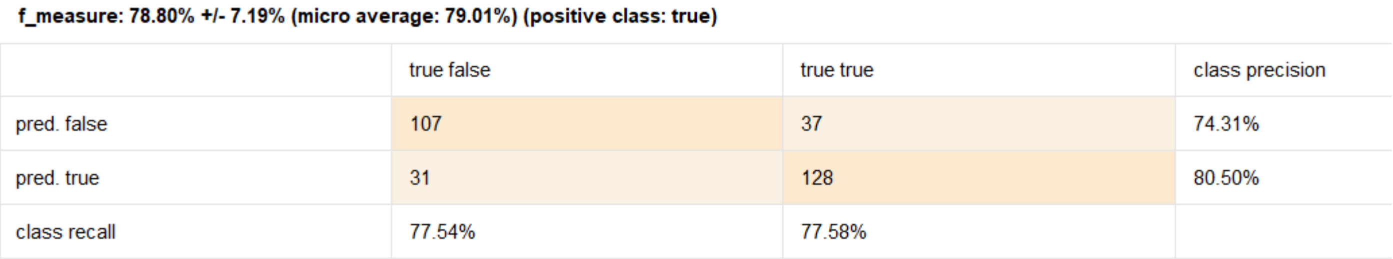


accuracy: 77.60% +/- 7.14% (micro average: 77.56%)

	true false	true true	class precision
pred. false	107	37	74.31%
pred. true	31	128	80.50%
class recall	77.54%	77.58%	

This means that the model correctly classified 77.60% of the instances in the dataset, while misclassifying the remaining 22.40%.

F measure:



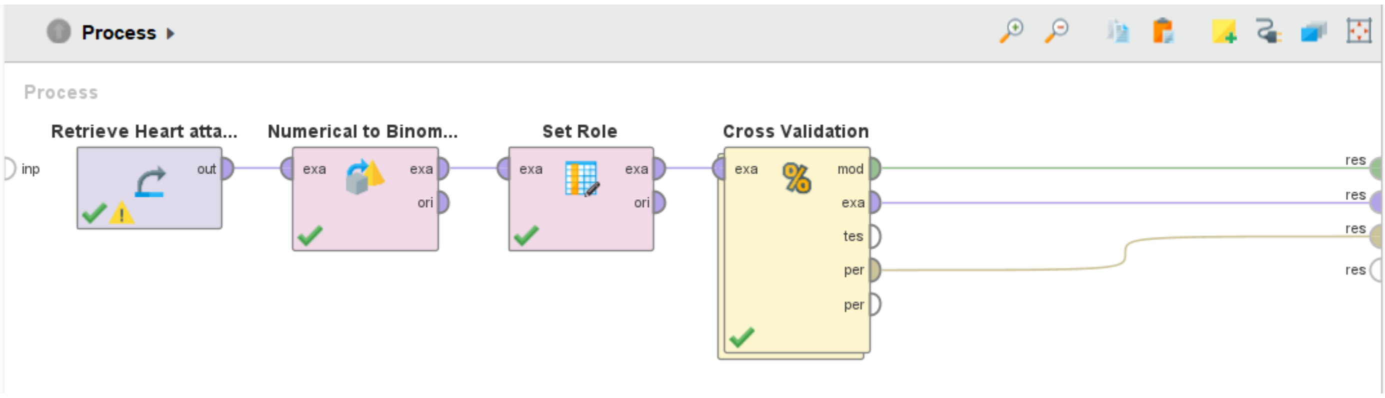
f_measure: 78.80% +/- 7.19% (micro average: 79.01%) (positive class: true)

	true false	true true	class precision
pred. false	107	37	74.31%
pred. true	31	128	80.50%
class recall	77.54%	77.58%	

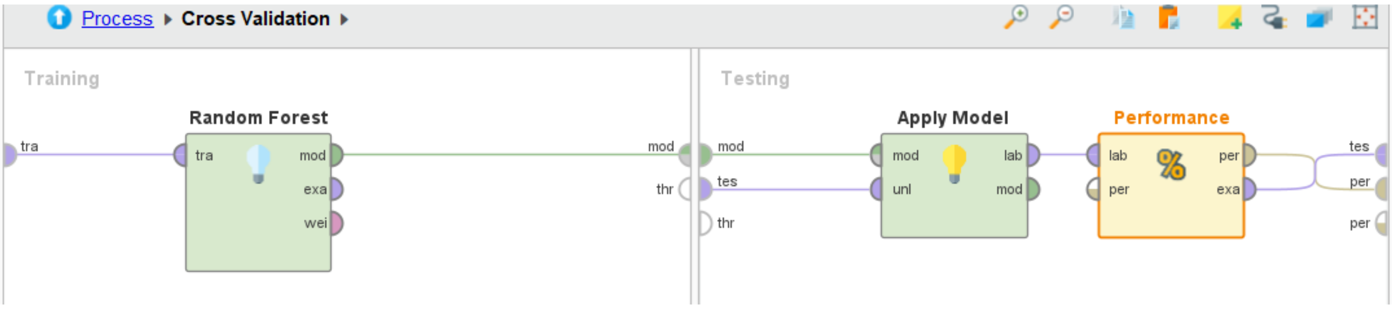
the F-measure has a value of 78.80%. This means that the model has a good balance between precision and recall, and is performing well on the dataset.

Model 2: Random Forest

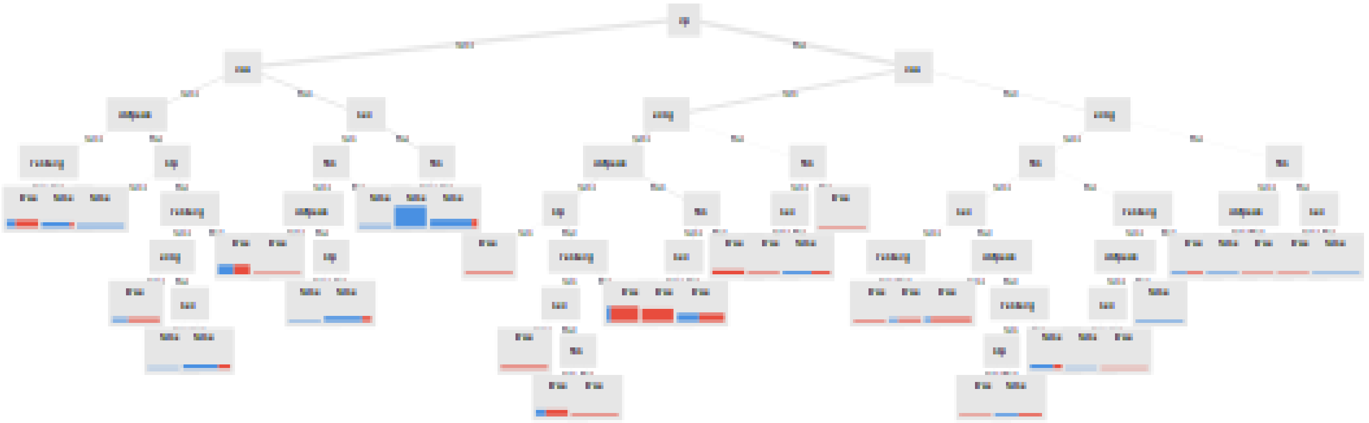
Process:



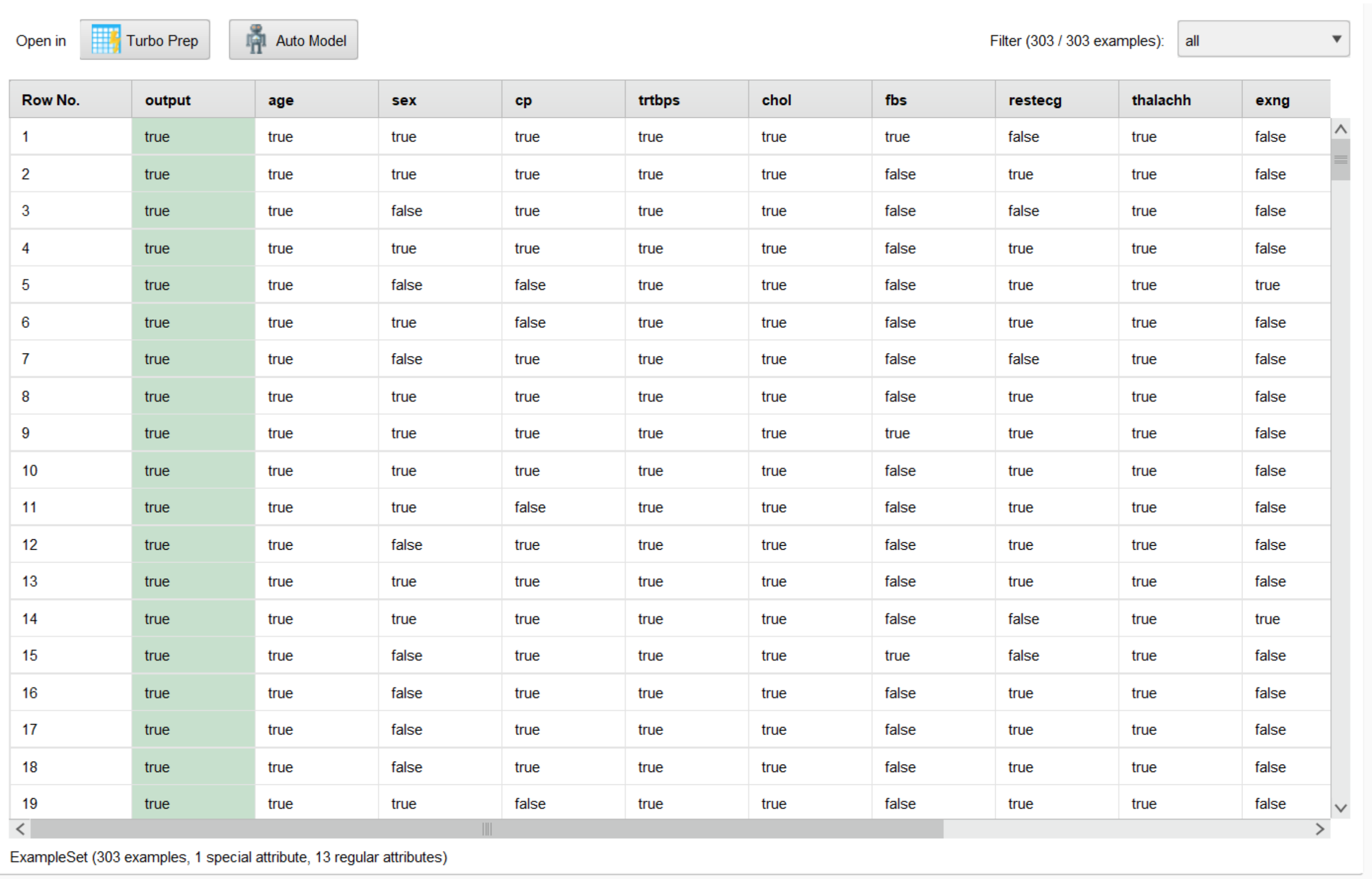
Cross validation:



Tree:



Example Set:



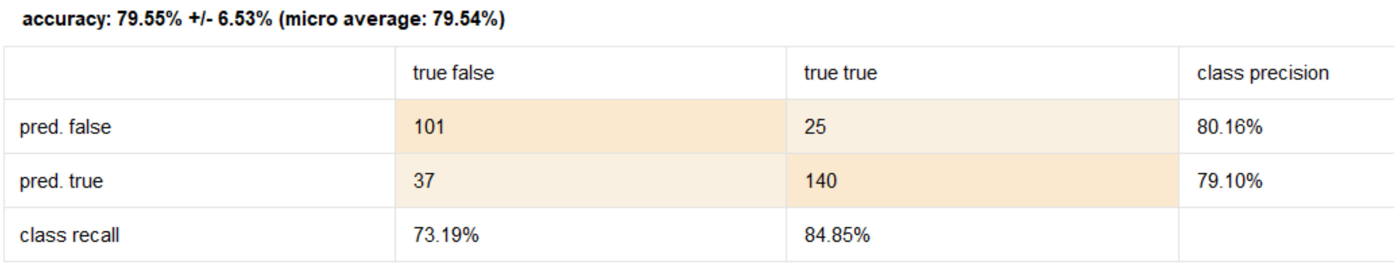
Open in

Filter (303 / 303 examples): all

Row No.	output	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng
1	true	true	true	true	true	true	true	false	true	false
2	true	true	true	true	true	true	false	true	true	false
3	true	true	false	true	true	true	false	false	true	false
4	true	true	true	true	true	true	false	true	true	false
5	true	true	false	false	true	true	false	true	true	true
6	true	true	true	false	true	true	false	true	true	false
7	true	true	false	true	true	true	false	false	true	false
8	true	true	true	true	true	true	false	true	true	false
9	true	true	true	true	true	true	true	true	true	false
10	true	true	true	true	true	true	false	true	true	false
11	true	true	true	false	true	true	false	true	true	false
12	true	true	false	true	true	true	false	true	true	false
13	true	true	true	true	true	true	false	true	true	false
14	true	true	true	true	true	true	false	false	true	true
15	true	true	false	true	true	true	true	false	true	false
16	true	true	false	true	true	true	false	true	true	false
17	true	true	false	true	true	true	false	true	true	false
18	true	true	false	true	true	true	false	true	true	false
19	true	true	true	false	true	true	false	true	true	false

ExampleSet (303 examples, 1 special attribute, 13 regular attributes)

Confusion matrix:

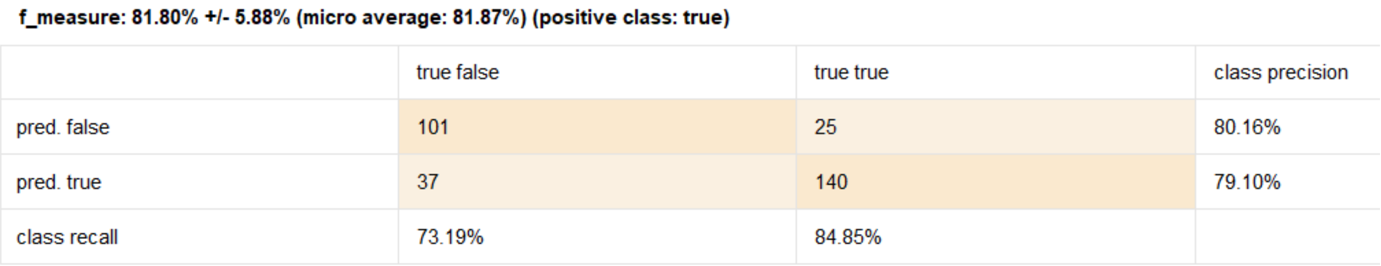


accuracy: 79.55% +/- 6.53% (micro average: 79.54%)

	true false	true true	class precision
pred. false	101	25	80.16%
pred. true	37	140	79.10%
class recall	73.19%	84.85%	

A 79.55% accuracy is a relatively high level of accuracy and suggests that the random forest model is performing well on the dataset.

F measure:



f_measure: 81.80% +/- 5.88% (micro average: 81.87%) (positive class: true)

	true false	true true	class precision
pred. false	101	25	80.16%
pred. true	37	140	79.10%
class recall	73.19%	84.85%	

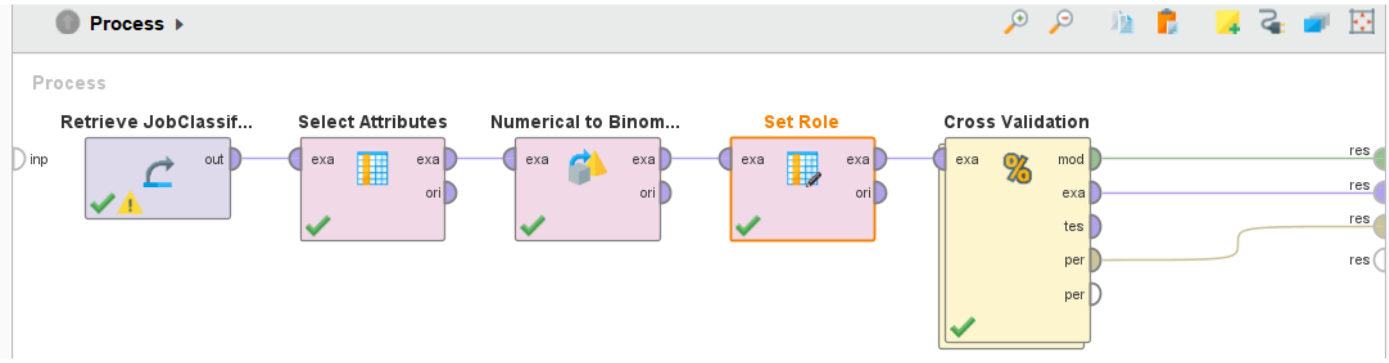
the F-measure has a value of 81.80%. This means that the model has a good balance between precision and recall, and is performing well on the dataset.

Comparing both the models I will choose model 2 in this case.

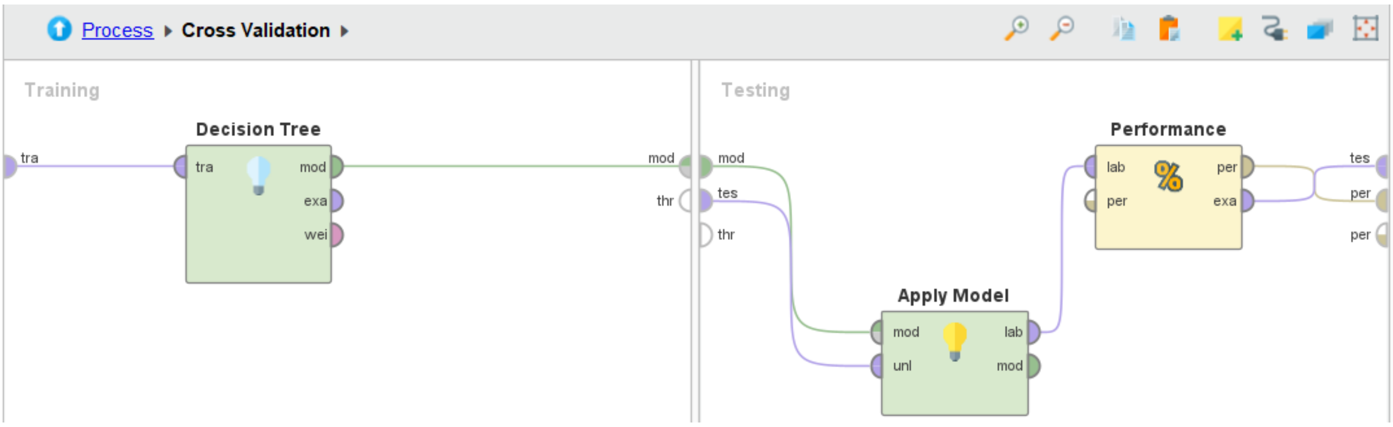
4: JobClassification

Model 1: Decision tree

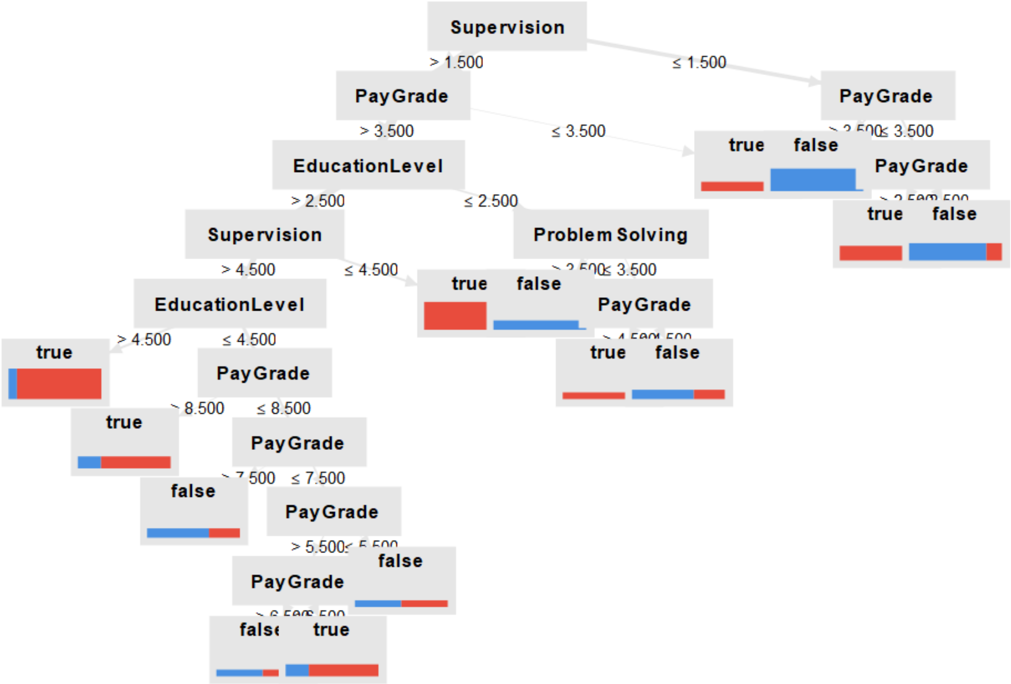
Process:



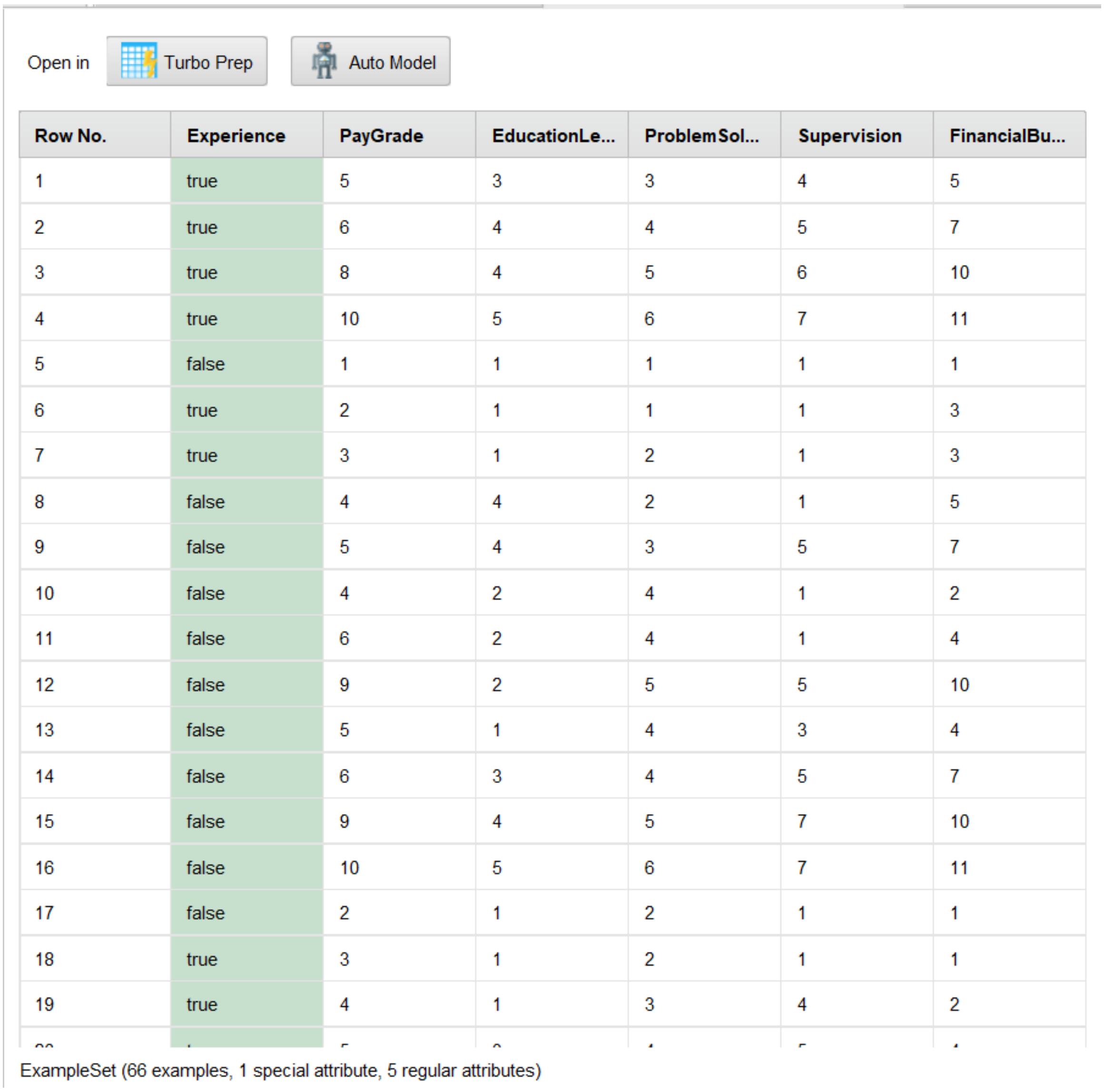
Cross validation:



Tree:



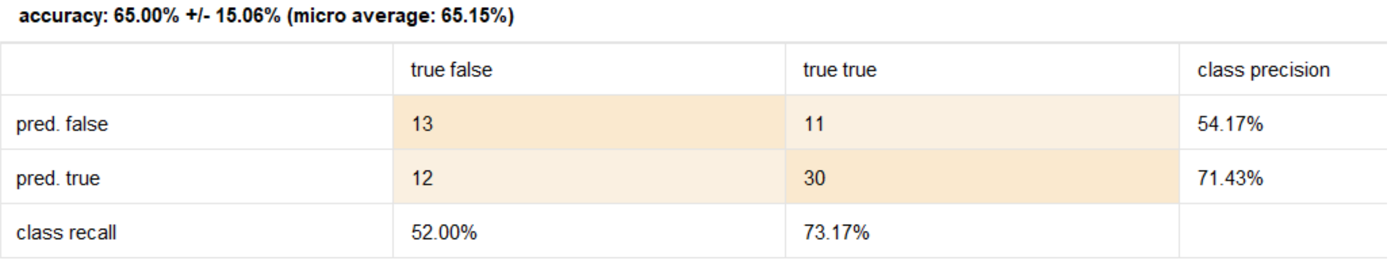
Example set:



Row No.	Experience	PayGrade	EducationLe...	ProblemSol...	Supervision	FinancialBu...
1	true	5	3	3	4	5
2	true	6	4	4	5	7
3	true	8	4	5	6	10
4	true	10	5	6	7	11
5	false	1	1	1	1	1
6	true	2	1	1	1	3
7	true	3	1	2	1	3
8	false	4	4	2	1	5
9	false	5	4	3	5	7
10	false	4	2	4	1	2
11	false	6	2	4	1	4
12	false	9	2	5	5	10
13	false	5	1	4	3	4
14	false	6	3	4	5	7
15	false	9	4	5	7	10
16	false	10	5	6	7	11
17	false	2	1	2	1	1
18	true	3	1	2	1	1
19	true	4	1	3	4	2

ExampleSet (66 examples, 1 special attribute, 5 regular attributes)

Confusion matrix:

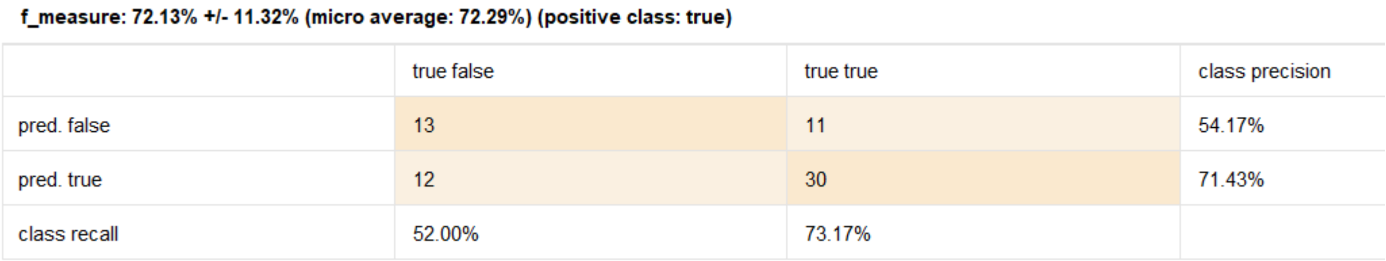


accuracy: 65.00% +/- 15.06% (micro average: 65.15%)

	true false	true true	class precision
pred. false	13	11	54.17%
pred. true	12	30	71.43%
class recall	52.00%	73.17%	

In the context of a decision tree model built using RapidMiner, a 65% accuracy in the confusion matrix means that the model correctly classified 65% of the observations in the dataset.

F measure:



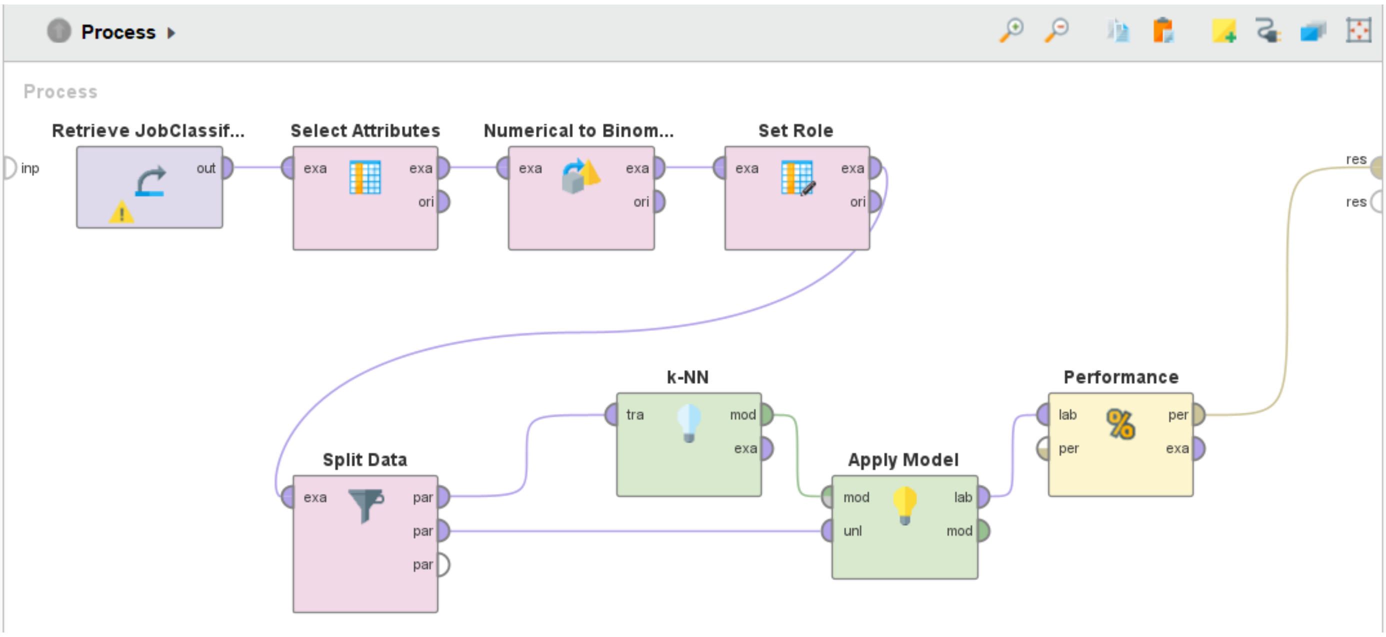
f_measure: 72.13% +/- 11.32% (micro average: 72.29%) (positive class: true)

	true false	true true	class precision
pred. false	13	11	54.17%
pred. true	12	30	71.43%
class recall	52.00%	73.17%	

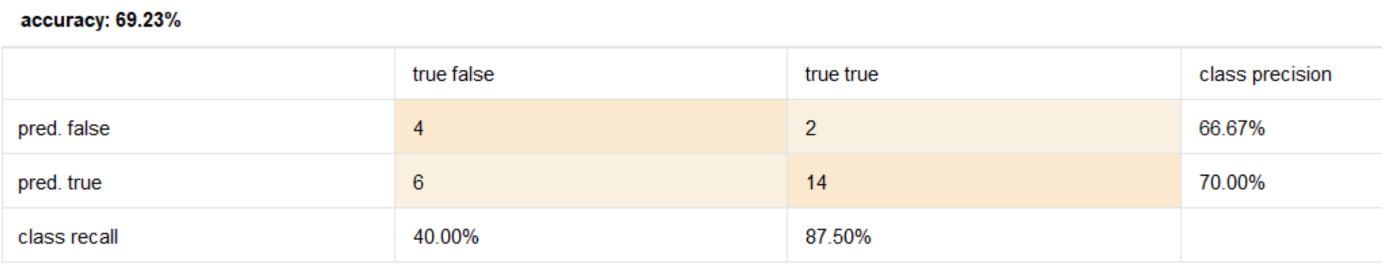
A 72.13% F-measure means that the model's precision and recall performance combined is relatively good.

Model 2: KNN

Process:



Confusion matrix:

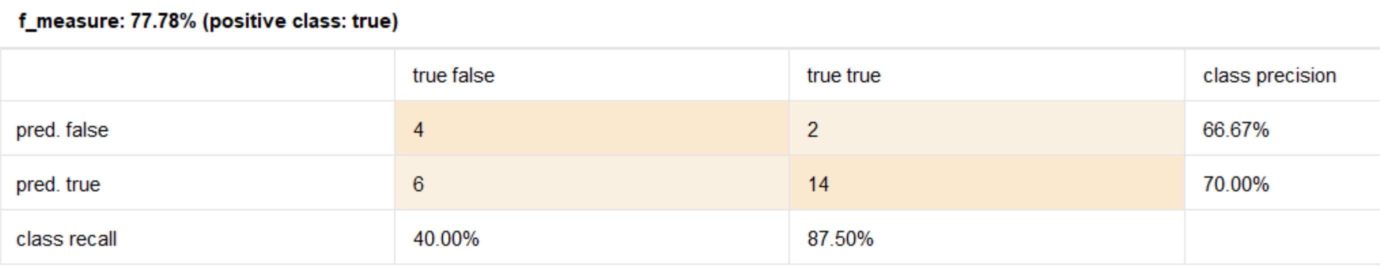


accuracy: 69.23%

	true false	true true	class precision
pred. false	4	2	66.67%
pred. true	6	14	70.00%
class recall	40.00%	87.50%	

A 69.23% accuracy in the confusion matrix means that the model correctly classified 69.23% of the observations in the dataset.

F measure:



f_measure: 77.78% (positive class: true)

	true false	true true	class precision
pred. false	4	2	66.67%
pred. true	6	14	70.00%
class recall	40.00%	87.50%	

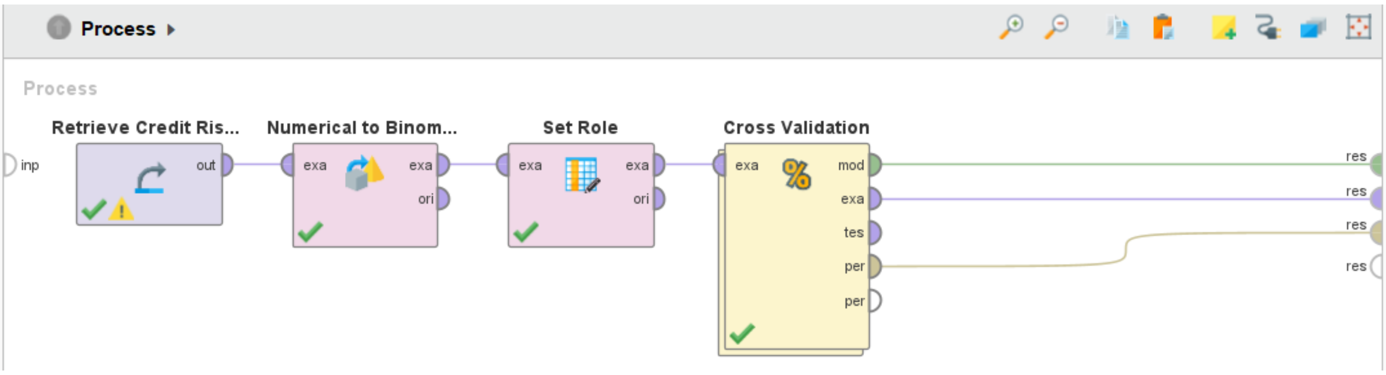
A 77.78% F-measure means that the model's precision and recall performance combined is relatively good.

Comparing both the models I would like to follow the KNN model for this data set.

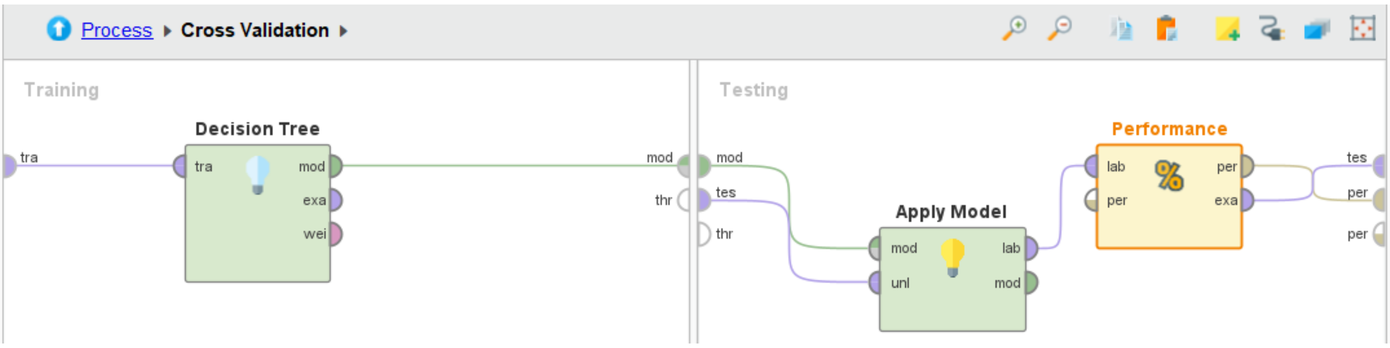
5: Credit Risk Classification

Model 1: Decision tree

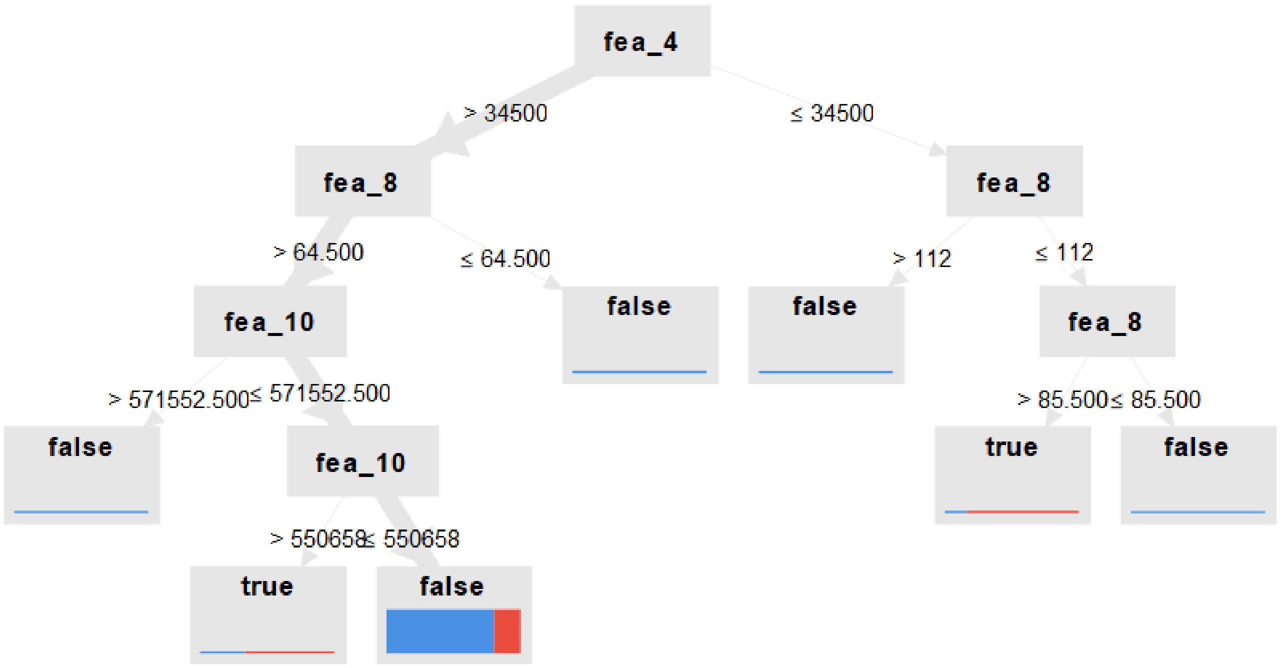
Process:



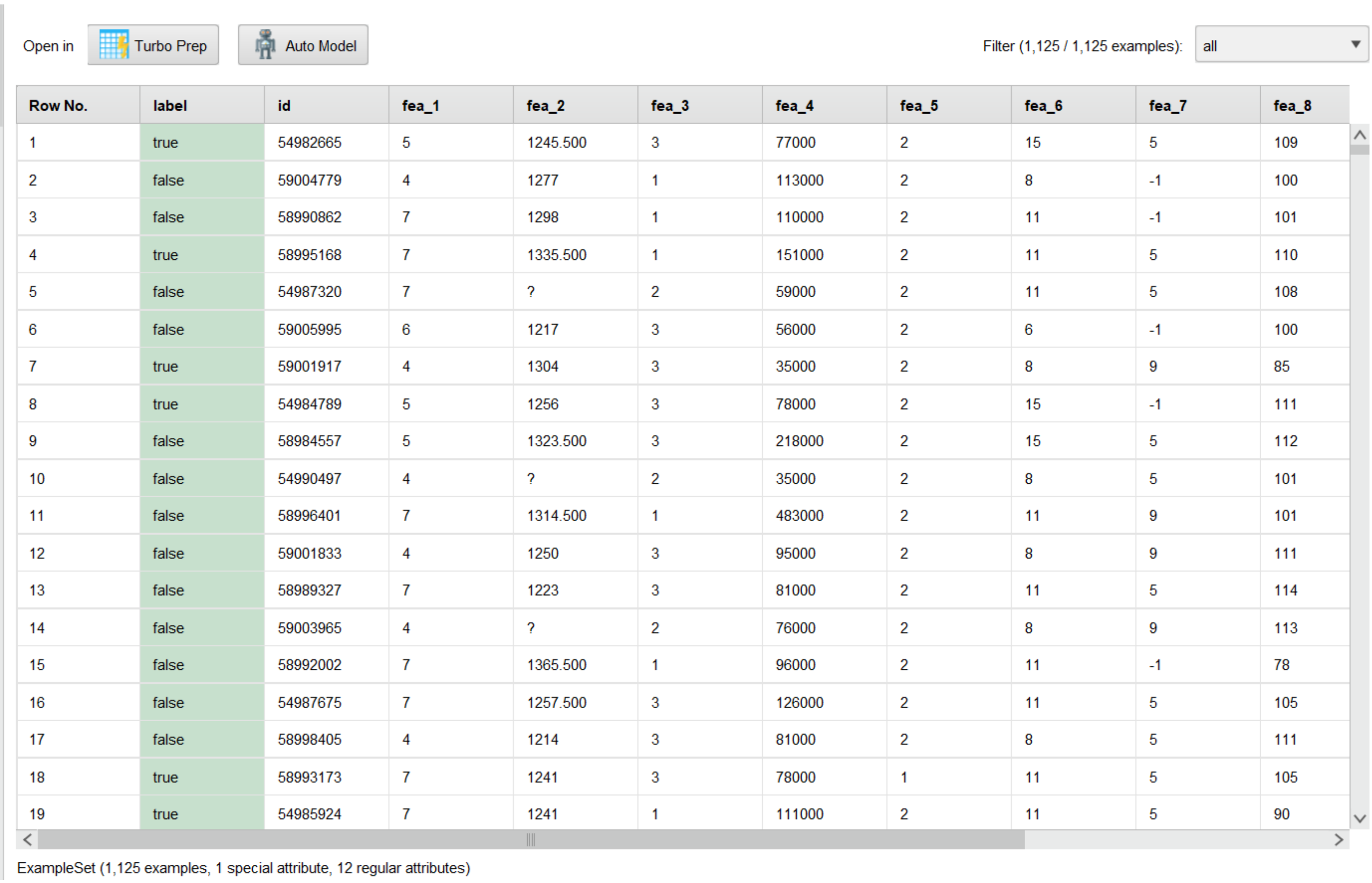
Cross validation:



Tree:



Example Set:

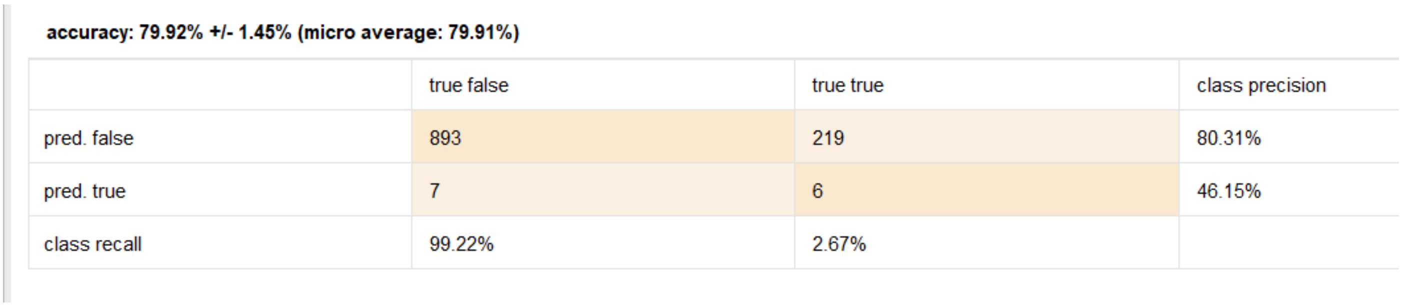


Filter (1,125 / 1,125 examples): all

Row No.	label	id	fea_1	fea_2	fea_3	fea_4	fea_5	fea_6	fea_7	fea_8
1	true	54982665	5	1245.500	3	77000	2	15	5	109
2	false	59004779	4	1277	1	113000	2	8	-1	100
3	false	58990862	7	1298	1	110000	2	11	-1	101
4	true	58995168	7	1335.500	1	151000	2	11	5	110
5	false	54987320	7	?	2	59000	2	11	5	108
6	false	59005995	6	1217	3	56000	2	6	-1	100
7	true	59001917	4	1304	3	35000	2	8	9	85
8	true	54984789	5	1256	3	78000	2	15	-1	111
9	false	58984557	5	1323.500	3	218000	2	15	5	112
10	false	54990497	4	?	2	35000	2	8	5	101
11	false	58996401	7	1314.500	1	483000	2	11	9	101
12	false	59001833	4	1250	3	95000	2	8	9	111
13	false	58989327	7	1223	3	81000	2	11	5	114
14	false	59003965	4	?	2	76000	2	8	9	113
15	false	58992002	7	1365.500	1	96000	2	11	-1	78
16	false	54987675	7	1257.500	3	126000	2	11	5	105
17	false	58998405	4	1214	3	81000	2	8	5	111
18	true	58993173	7	1241	3	78000	1	11	5	105
19	true	54985924	7	1241	1	111000	2	11	5	90

ExampleSet (1,125 examples, 1 special attribute, 12 regular attributes)

Confusion matrix:

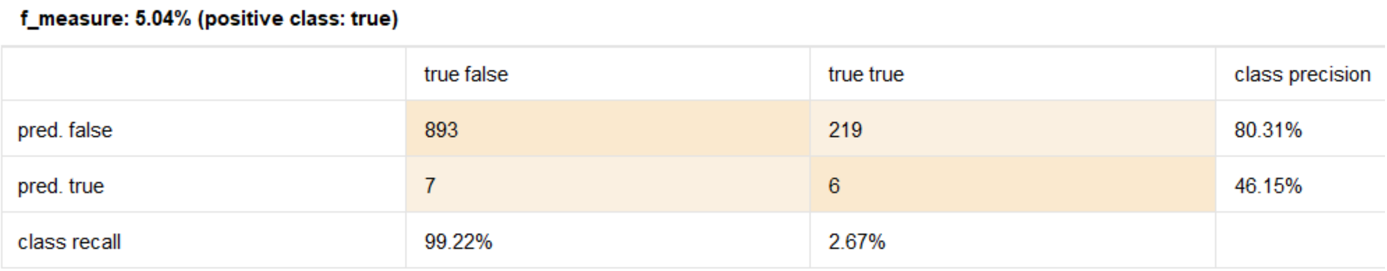


accuracy: 79.92% +/- 1.45% (micro average: 79.91%)

	true false	true true	class precision
pred. false	893	219	80.31%
pred. true	7	6	46.15%
class recall	99.22%	2.67%	

A 79.92% accuracy in the confusion matrix suggests that the decision tree model built using RapidMiner is performing relatively well in terms of overall accuracy.

F measure:



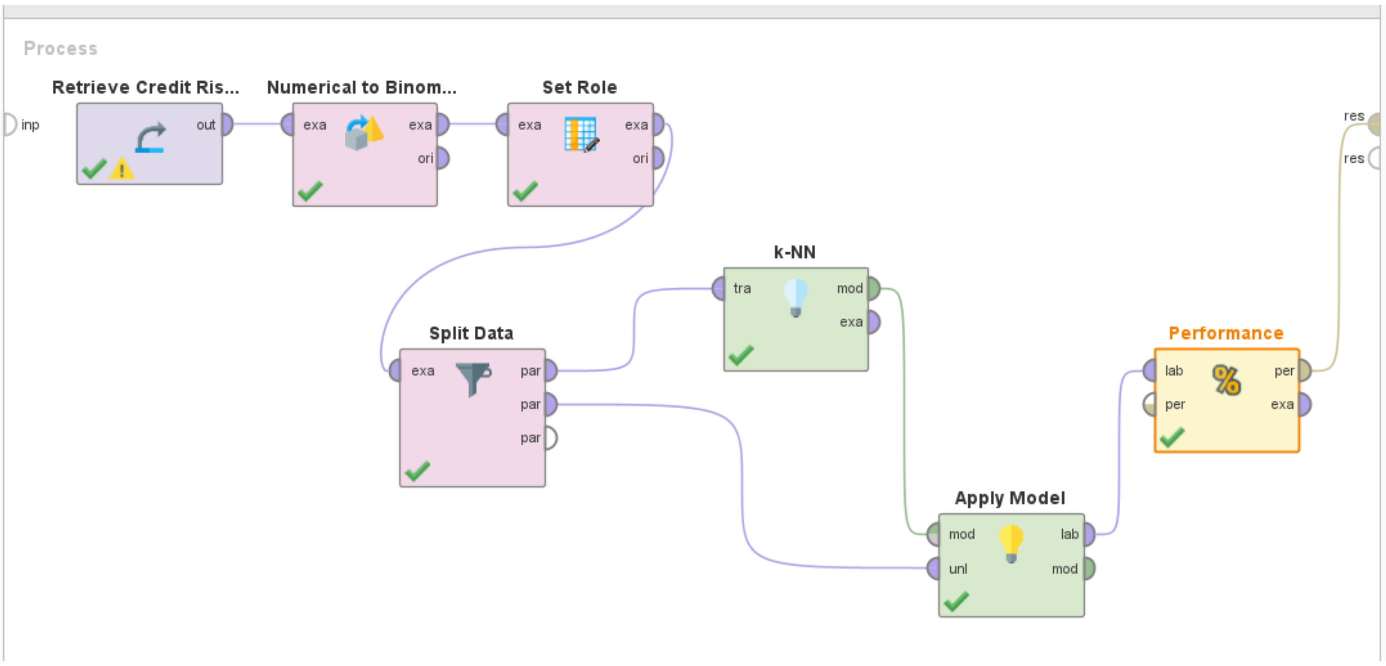
f_measure: 5.04% (positive class: true)

	true false	true true	class precision
pred. false	893	219	80.31%
pred. true	7	6	46.15%
class recall	99.22%	2.67%	

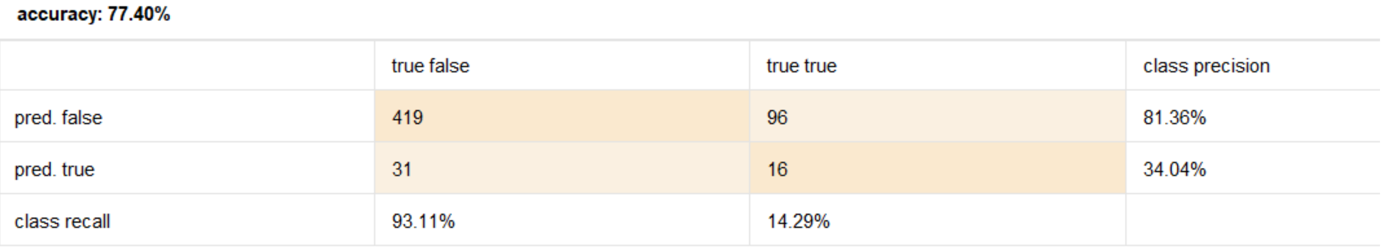
A 5.04% F-measure in a decision tree model built using RapidMiner indicates that the model is performing poorly in terms of precision and recall, and further analysis and improvement techniques are needed to enhance the model's performance.

Model 2: KNN

Process:



Confusion matrix:

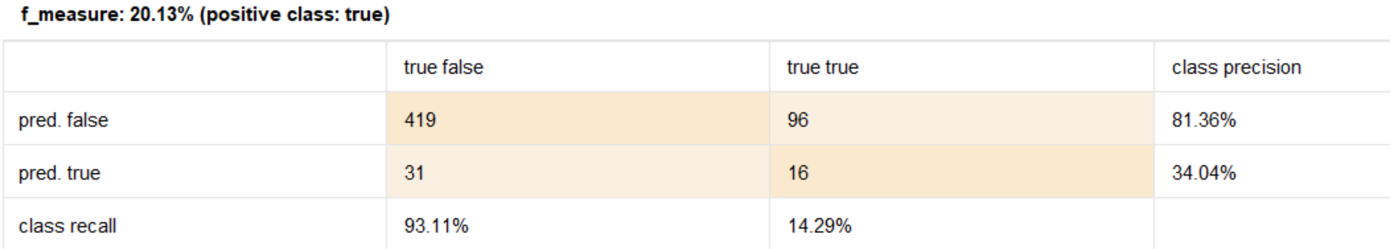


accuracy: 77.40%

	true false	true true	class precision
pred. false	419	96	81.36%
pred. true	31	16	34.04%
class recall	93.11%	14.29%	

A 77.40% accuracy in the confusion matrix means that the model correctly classified 77.40% of the observations in the dataset.

F measure:



f_measure: 20.13% (positive class: true)

	true false	true true	class precision
pred. false	419	96	81.36%
pred. true	31	16	34.04%
class recall	93.11%	14.29%	

A low F-measure value of 20.13% in a KNN model built using RapidMiner suggests that the model is not performing well.

Comparing the two models, we can see that the first model has a higher accuracy of 79.92% compared to the second model's accuracy of 77.40%. However, the second model has a higher F-measure of 20.13% compared to the first model's F-measure of 5.04%.

It is important to note that accuracy alone may not be the best metric to evaluate the performance of a classification model. In some cases, a model with a lower accuracy may still have a better performance if it has a higher F-measure, which considers both precision and recall.

In the first model with higher accuracy and lower F-measure, the model may be classifying most instances correctly, but at the same time, it may be incorrectly classifying many instances. This would result in a low F-measure, indicating poor precision and recall.

In the second model with lower accuracy and higher F-measure, the model may be missing some instances but is performing well in terms of precision and recall. This could be due to a smaller number of false positive predictions.

Hence, for this model I will consider overall model performance that is given by KNN model.