IBA Institute of Business Administration Karachi

Leadership and Ideas for Tomorrow

INTRODUCTION TO DATA MINING
FINAL PROJECT
AIRLINE PASSENGER SATISFACTION - CLASSIFICATION

GROUP MEMBERS:

SAFA ABDUL HAI - 18625

MARIUM AFZAL -19756

MAHUM FATIMA KHAN - 19747

Introduction:

Recently, there has been a surge in interest in the field of data mining, where the goal is to predict correct and useful knowledge for users. We used a summary of our study and exploration of the Arline passenger satisfaction Data in this project to come up with useful, important, and intriguing data properties. The CRISP-DM data mining methodology was heavily used in the development of our project model to predict as accurate as possible statistics for our users.

Problem Description:

Passenger satisfaction is a high concern for airlines' management as their prime concern is whether the customer will choose their airlines for their next trip or not. Airline passenger satisfaction is affected by many factors, but at its root, this type of customer satisfaction is no different from that of any other business. Therefore, various data mining techniques are utilized in this project to build a classification model for Airlines management to determine if the customer is satisfied or not depending on various factors.

Data Collection:

The dataset, primarily known as "Airline Passenger Satisfaction" was found on Kaggle, https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction, while we were searching for a classification data set for our Data Mining project. This data set was collected by TJ Klein in 2019 through a survey.

Data Description:

The data set has two csv files named Test and Train. The test.csv file contains 26k records while the train.csv file contains 104k records. There are 25 attributes in total. The following are the given attributes:

Gender: Gender of the passengers (Female, Male)

Customer Type: The customer type (Loyal customer, disloyal customer)

Age: The actual age of the passengers

Type of Travel: Purpose of the flight of the passengers (Personal Travel, Business Travel)

Class: Travel class in the plane of the passengers (Business, Eco, Eco Plus)

Flight distance: The flight distance of this journey

Inflight wifi service: Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)

Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient

Ease of Online booking: Satisfaction level of online booking

Gate location: Satisfaction level of Gate location

Food and drink: Satisfaction level of Food and drink

Online boarding: Satisfaction level of online boarding

Seat comfort: Satisfaction level of Seat comfort

Inflight entertainment: Satisfaction level of inflight entertainment

On-board service: Satisfaction level of On-board service

Leg room service: Satisfaction level of Leg room service

Baggage handling: Satisfaction level of baggage handling

Check-in service: Satisfaction level of Check-in service

Inflight service: Satisfaction level of inflight service

Cleanliness: Satisfaction level of Cleanliness

Departure Delay in Minutes: Minutes delayed when departure

Arrival Delay in Minutes: Minutes delayed when Arrival

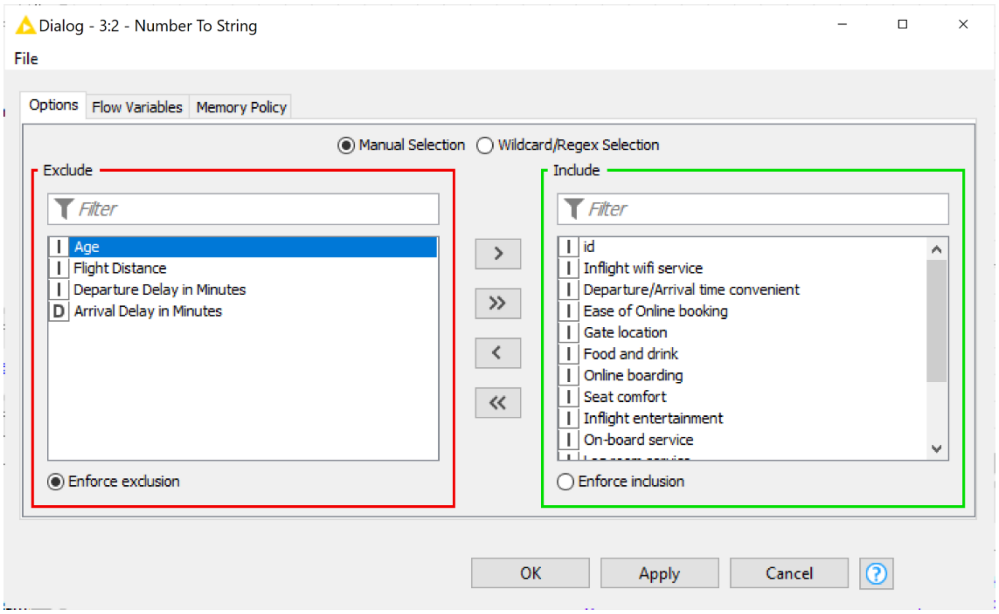
Satisfaction: Airline satisfaction level(Satisfaction, neutral or dissatisfaction)

Attempt 1:

Data Preprocessing:

The steps taken to prepare the data are as follow:

1. **Number to String:**
   All the attributes except the continuous attributes such as Age, Flight Distance, Departure Delay in Minutes, Arrival Delay in Minutes are converted to String data type.



Dialog - 3:2 - Number To String

File

Options | Flow Variables | Memory Policy

● Manual Selection ○ Wildcard/Regex Selection

Exclude

- I Age
- I Flight Distance
- I Departure Delay in Minutes
- D Arrival Delay in Minutes

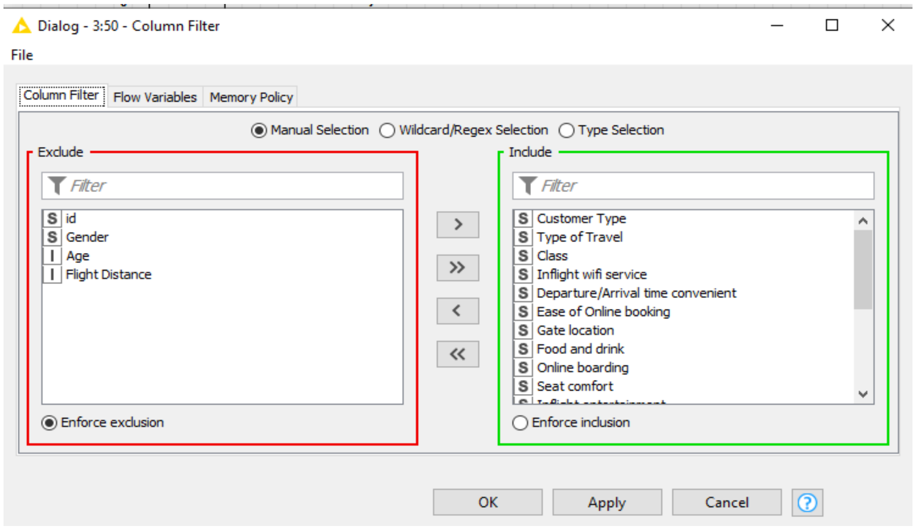
● Enforce exclusion

Include

- I id
- I Inflight wifi service
- I Departure/Arrival time convenient
- I Ease of Online booking
- I Gate location
- I Food and drink
- I Online boarding
- I Seat comfort
- I Inflight entertainment
- I On-board service

○ Enforce inclusion

2. **Column Filter:**
   The 4 columns ID, Gender, Age, and Flight Distance are filtered out using Column Filter as including these columns did not have much effect on our prediction/ accuracy.



Dialog - 3:50 - Column Filter

File

Column Filter | Flow Variables | Memory Policy

● Manual Selection ○ Wildcard/Regex Selection ○ Type Selection

Exclude

- S id
- S Gender
- I Age
- I Flight Distance

● Enforce exclusion

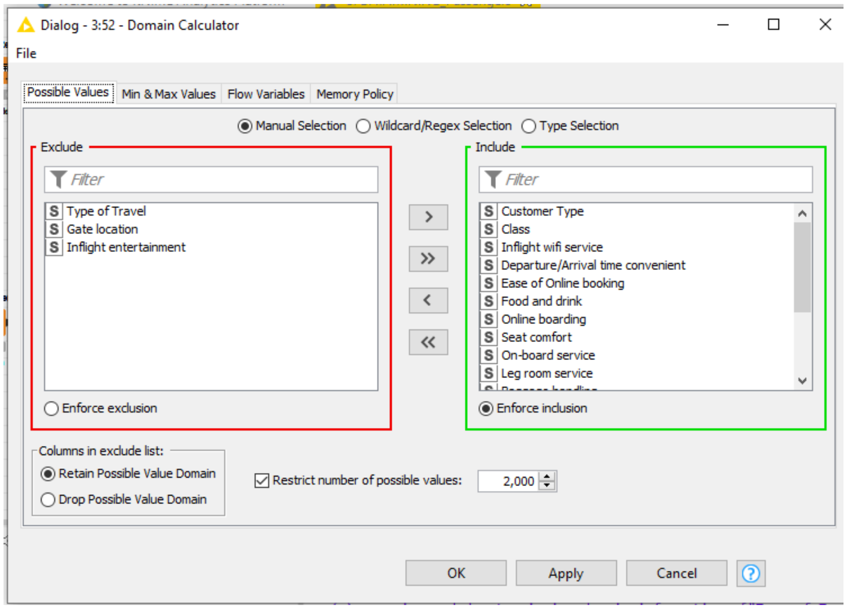
Include

- S Customer Type
- S Type of Travel
- S Class
- S Inflight wifi service
- S Departure/Arrival time convenient
- S Ease of Online booking
- S Gate location
- S Food and drink
- S Online boarding
- S Seat comfort

○ Enforce inclusion

3. **Domain Calculator:**

After applying the previous techniques on the dataset, the domain information of the data has changed. So we have used the Domain Calculator for this purpose.



Dialog - 3:52 - Domain Calculator

File

Possible Values | Min & Max Values | Flow Variables | Memory Policy

● Manual Selection ○ Wildcard/Regex Selection ○ Type Selection

Exclude

- S Type of Travel
- S Gate location
- S Inflight entertainment

○ Enforce exclusion

Include

- S Customer Type
- S Class
- S Inflight wifi service
- S Departure/Arrival time convenient
- S Ease of Online booking
- S Food and drink
- S Online boarding
- S Seat comfort
- S On-board service
- S Leg room service

● Enforce inclusion

Columns in exclude list:

● Retain Possible Value Domain

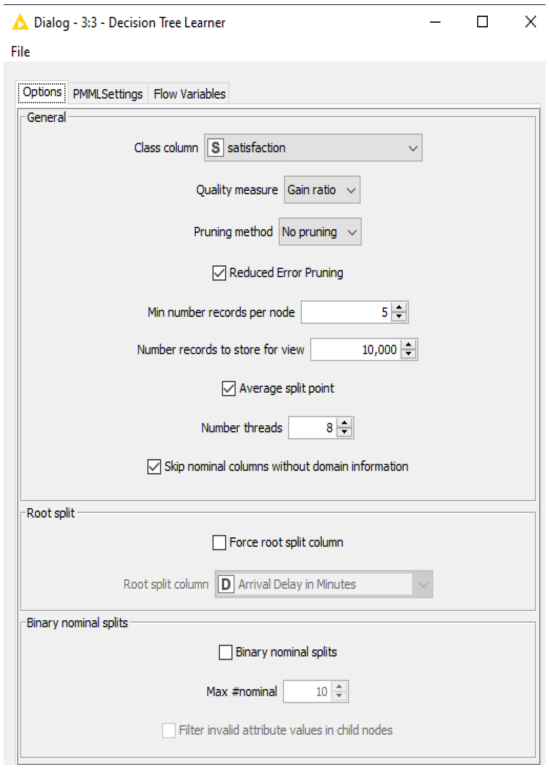
○ Drop Possible Value Domain

☑ Restrict number of possible values: 2,000

Model Building:

After preprocessing the entire data, we used four classifiers to build our model. They are as follow:

1. **Decision Tree:** We have used the following configuration on the Decision Tree Learner



General

Class column: S satisfaction

Quality measure: Gain ratio

Pruning method: No pruning

☑ Reduced Error Pruning

Min number records per node: 5

Number records to store for view: 10,000

☑ Average split point

Number threads: 8

☑ Skip nominal columns without domain information

Root split

☐ Force root split column

Root split column: D Arrival Delay in Minutes

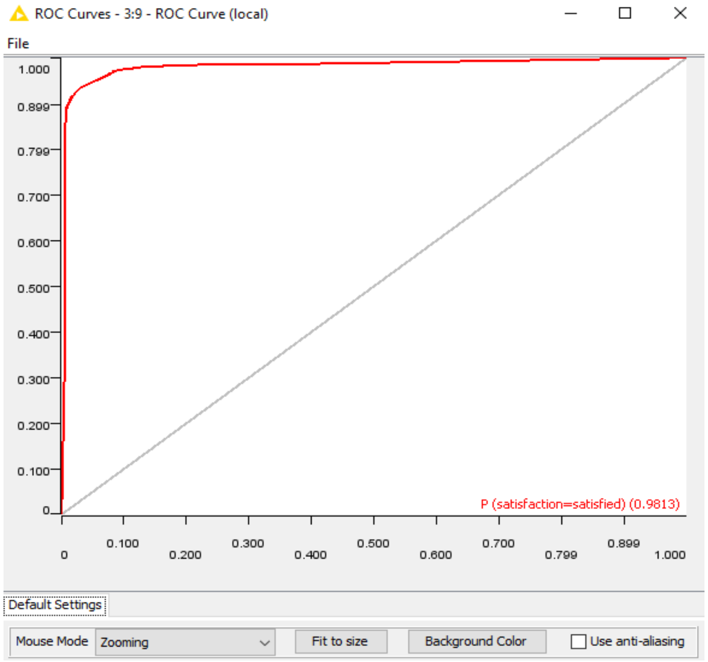
Binary nominal splits

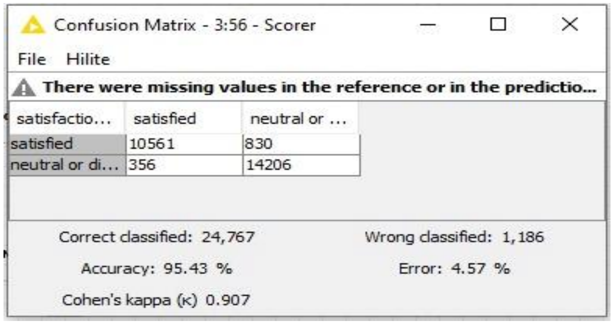
☐ Binary nominal splits

Max #nominal: 10

☐ Filter invalid attribute values in child nodes

ROC:





There were missing values in the reference or in the predictio...

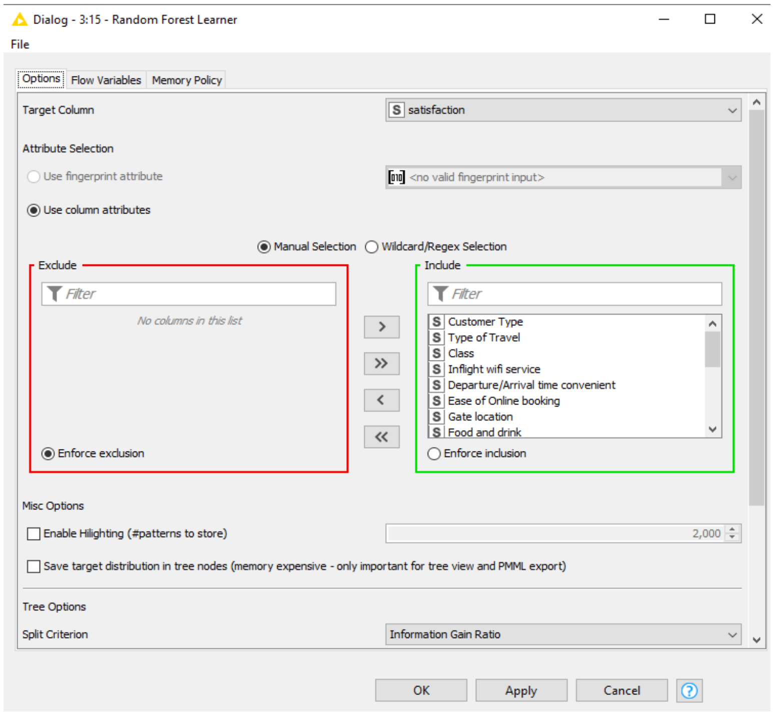
satisfactio...	satisfied	neutral or ...
satisfied	10561	830
neutral or di...	356	14206

Correct classified: 24,767 Wrong classified: 1,186

Accuracy: 95.43 % Error: 4.57 %

Cohen's kappa (κ) 0.907

2. **Random Forest:** We have used the following configuration on the Random Forest Learner



Target Column: S satisfaction

Attribute Selection

○ Use fingerprint attribute <no valid fingerprint input>

● Use column attributes

● Manual Selection ○ Wildcard/Regex Selection

Exclude
Filter
No columns in this list
● Enforce exclusion

Include
Filter
S Customer Type
S Type of Travel
S Class
S Inflight wifi service
S Departure/Arrival time convenient
S Ease of Online booking
S Gate location
S Food and drink
○ Enforce inclusion

Misc Options

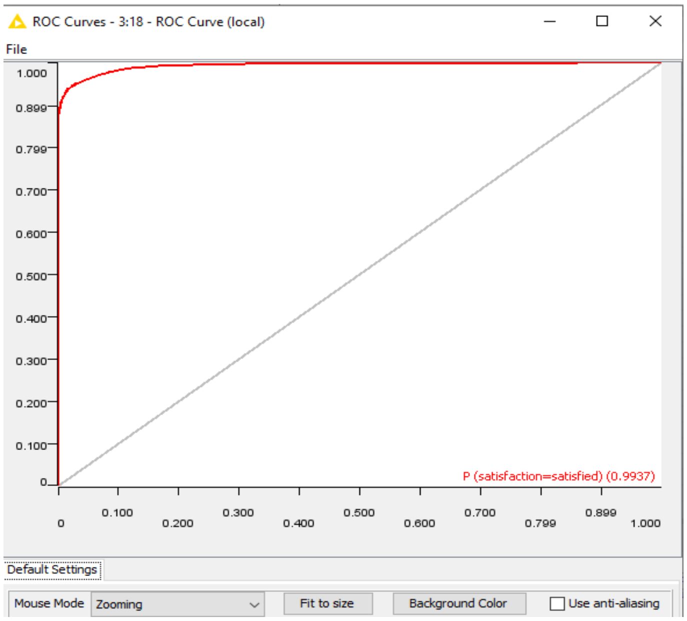
□ Enable Hiliting (#patterns to store) 2,000

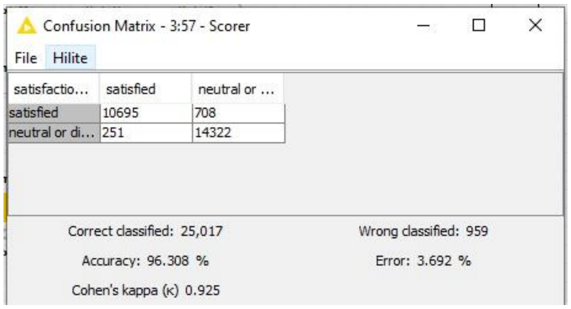
□ Save target distribution in tree nodes (memory expensive - only important for tree view and PMML export)

Tree Options

Split Criterion: Information Gain Ratio

ROC:





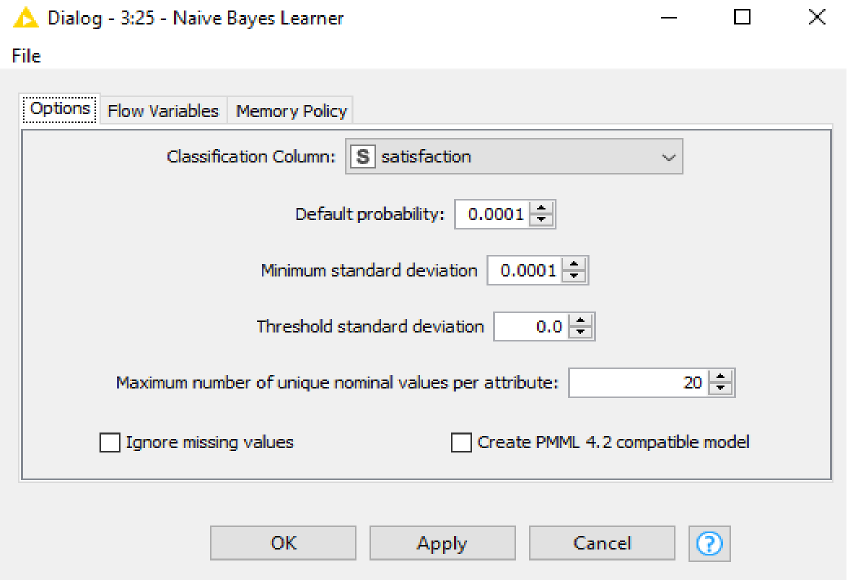
satisfactio...	satisfied	neutral or ...
satisfied	10695	708
neutral or di...	251	14322

Correct classified: 25,017 Wrong classified: 959

Accuracy: 96.308 % Error: 3.692 %

Cohen's kappa (κ) 0.925

1. **Naive Bayes:** The following Configuration was used for Naïve Bayes



Classification Column: S satisfaction

Default probability: 0.0001

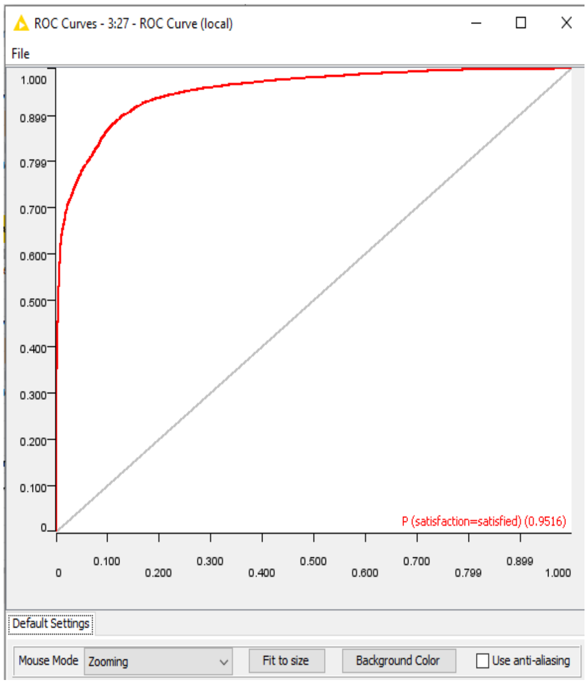
Minimum standard deviation 0.0001

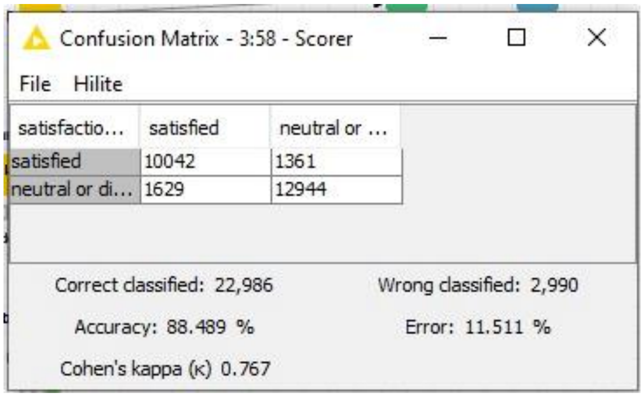
Threshold standard deviation 0.0

Maximum number of unique nominal values per attribute: 20

□ Ignore missing values □ Create PMML 4.2 compatible model

ROC:





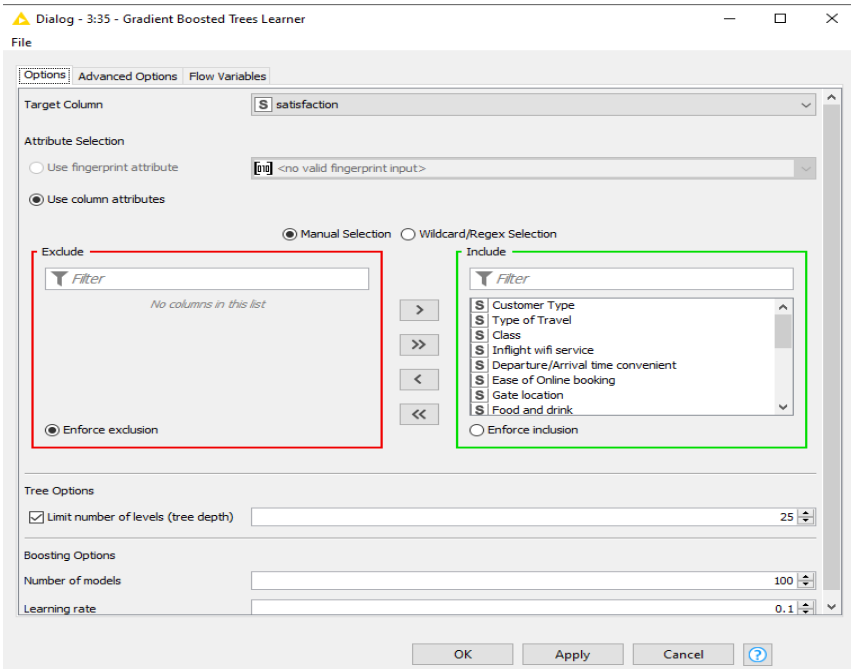
satisfactio...	satisfied	neutral or ...
satisfied	10042	1361
neutral or di...	1629	12944

Correct classified: 22,986 Wrong classified: 2,990

Accuracy: 88.489 % Error: 11.511 %

Cohen's kappa (κ) 0.767

2. **Gradient Boost:** The following configuration was used for gradient boosting



Target Column: S satisfaction

Attribute Selection

○ Use fingerprint attribute <no valid fingerprint input>

● Use column attributes

● Manual Selection ○ Wildcard/Regex Selection

Exclude: Filter — No columns in this list — ● Enforce exclusion

Include: Filter

- S Customer Type
- S Type of Travel
- S Class
- S Inflight wifi service
- S Departure/Arrival time convenient
- S Ease of Online booking
- S Gate location
- S Food and drink

○ Enforce inclusion

Tree Options

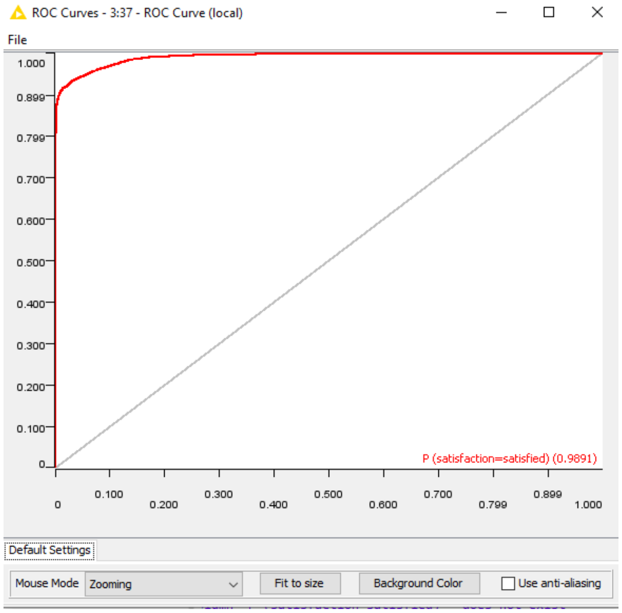
☑ Limit number of levels (tree depth): 25

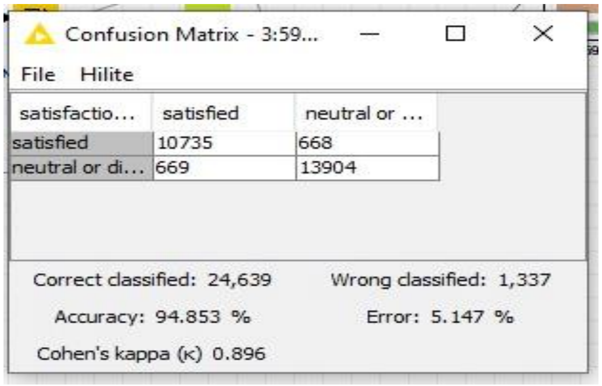
Boosting Options

Number of models: 100

Learning rate: 0.1

ROC:





satisfactio...	satisfied	neutral or ...
satisfied	10735	668
neutral or di...	669	13904

Correct classified: 24,639 Wrong classified: 1,337

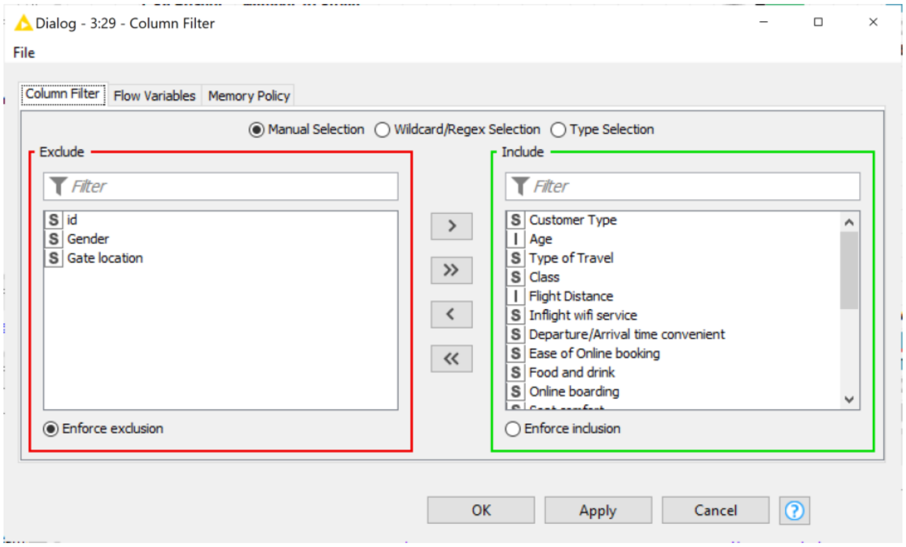
Accuracy: 94.853 % Error: 5.147 %

Cohen's kappa (κ) 0.896

<u>Attempt 2:</u>

Data Preprocessing:

1. **Column Filter:** From our previous attempt we found out that the flight distance and age are important factors therefore their columns should be included and gate location was excluded since it did not make any significant difference in our prediction.



Dialog - 3:29 - Column Filter

● Manual Selection ○ Wildcard/Regex Selection ○ Type Selection

Exclude

- S id
- S Gender
- S Gate location

● Enforce exclusion

Include

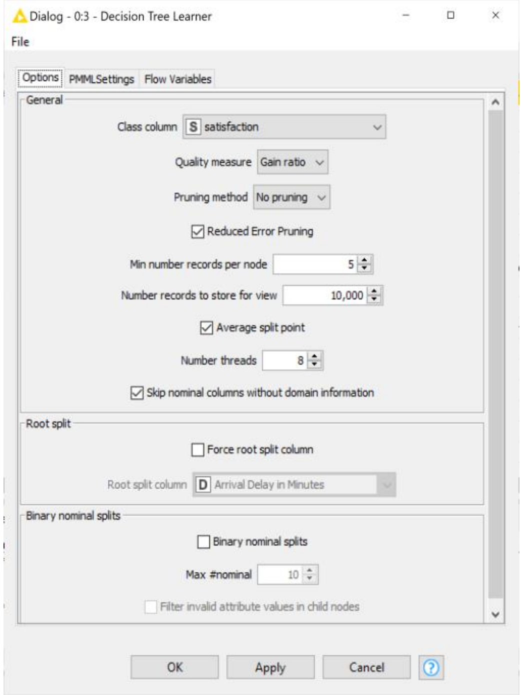
- S Customer Type
- I Age
- S Type of Travel
- S Class
- I Flight Distance
- S Inflight wifi service
- S Departure/Arrival time convenient
- S Ease of Online booking
- S Food and drink
- S Online boarding

○ Enforce inclusion

Model Building:

After changing the column filter's configuration, we again checked the accuracy using four classifiers. They are as follow:

1. **Decision Tree:**



Dialog - 0:3 - Decision Tree Learner

General

Class column: S satisfaction

Quality measure: Gain ratio

Pruning method: No pruning

☑ Reduced Error Pruning

Min number records per node: 5

Number records to store for view: 10,000

☑ Average split point

Number threads: 8

☑ Skip nominal columns without domain information

Root split

☐ Force root split column

Root split column: D Arrival Delay in Minutes

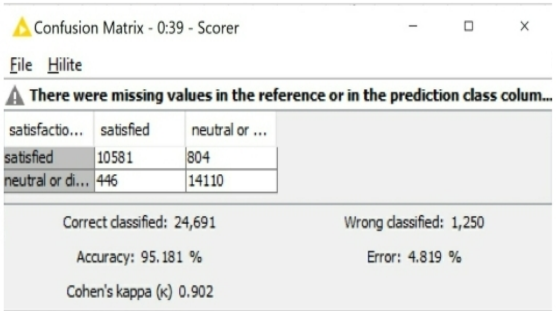
Binary nominal splits

☐ Binary nominal splits

Max #nominal: 10

☐ Filter invalid attribute values in child nodes

ROC: The accuracy decreased a little by 0.01 percent from approximately 98% to 97%.



Confusion Matrix - 0:39 - Scorer

⚠ There were missing values in the reference or in the prediction class colum...

satisfactio...	satisfied	neutral or ...
satisfied	10581	804
neutral or di...	446	14110

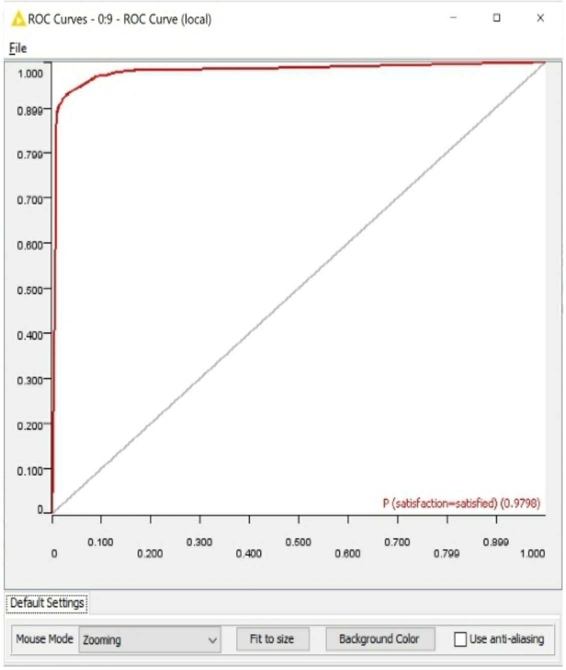
Correct classified: 24,691

Wrong classified: 1,250

Accuracy: 95.181 %

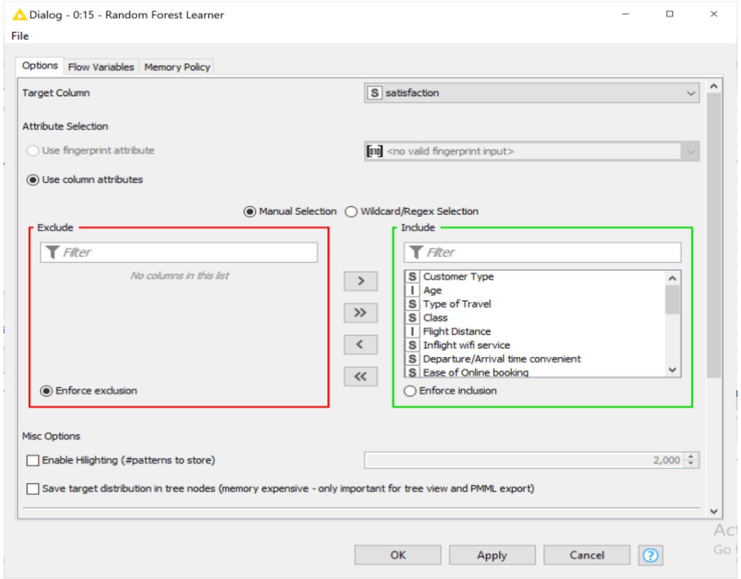
Error: 4.819 %

Cohen's kappa (κ) 0.902



ROC Curves - 0:9 - ROC Curve (local)

2. Random Forest:



Dialog - 0:15 - Random Forest Learner

File

Options | Flow Variables | Memory Policy

Target Column: S satisfaction

Attribute Selection

○ Use fingerprint attribute <no valid fingerprint input>

● Use column attributes

● Manual Selection ○ Wildcard/Regex Selection

Exclude: Filter — No columns in this list

● Enforce exclusion

Include: Filter

- S Customer Type
- I Age
- S Type of Travel
- S Class
- I Flight Distance
- S Inflight wifi service
- S Departure/Arrival time convenient
- S Ease of Online booking

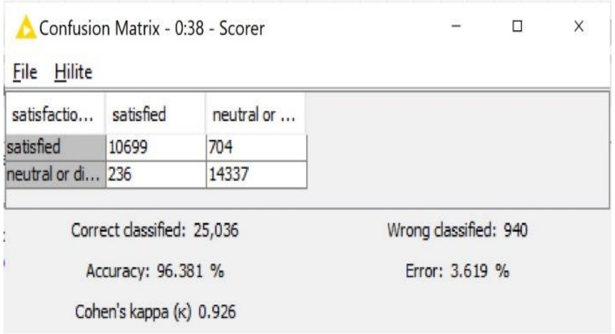
○ Enforce inclusion

Misc Options

□ Enable Hilighting (#patterns to store) 2,000

□ Save target distribution in tree nodes (memory expensive - only important for tree view and PMML export)

ROC & Scorer: The accuracy increased.



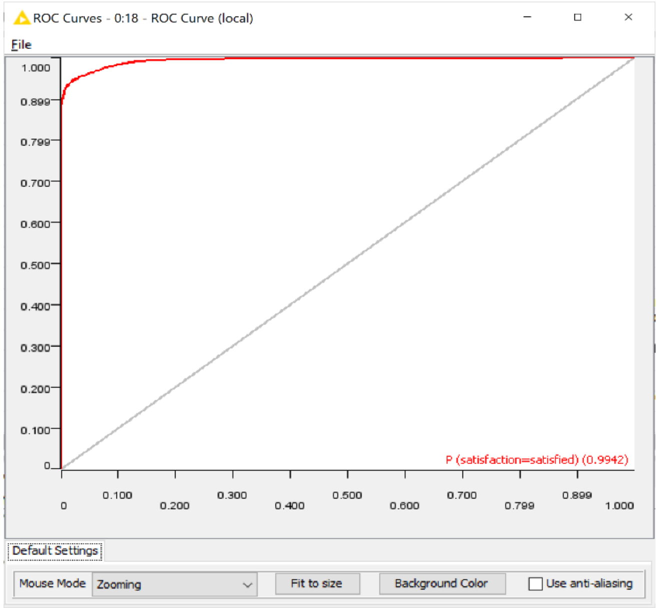
Confusion Matrix - 0:38 - Scorer

satisfactio...	satisfied	neutral or ...
satisfied	10699	704
neutral or di...	236	14337

Correct classified: 25,036 Wrong classified: 940

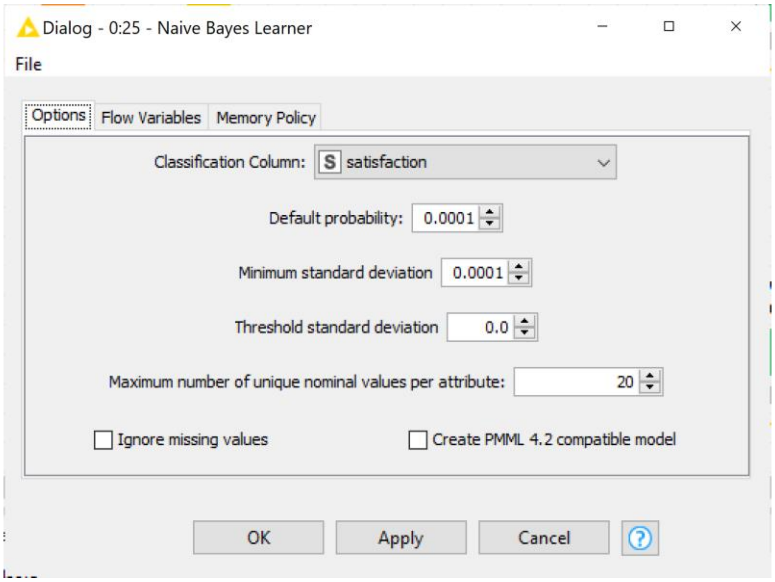
Accuracy: 96.381 % Error: 3.619 %

Cohen's kappa (κ) 0.926



ROC Curves - 0:18 - ROC Curve (local)

3. Naive Bayes:



Dialog - 0:25 - Naive Bayes Learner

Classification Column: S satisfaction

Default probability: 0.0001

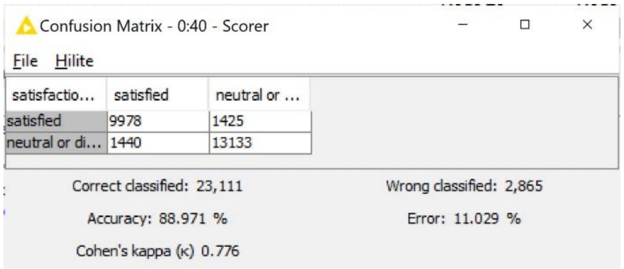
Minimum standard deviation 0.0001

Threshold standard deviation 0.0

Maximum number of unique nominal values per attribute: 20

☐ Ignore missing values ☐ Create PMML 4.2 compatible model

ROC & Scorer:



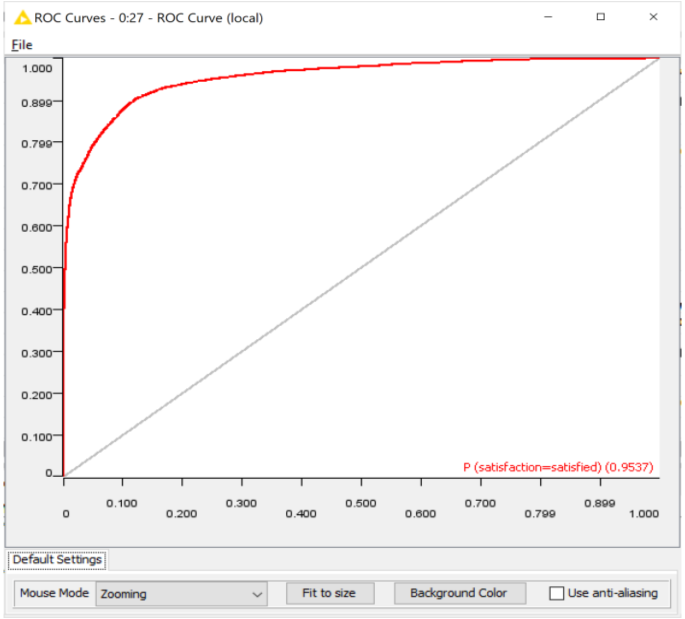
Confusion Matrix - 0:40 - Scorer

satisfactio...	satisfied	neutral or ...
satisfied	9978	1425
neutral or di...	1440	13133

Correct classified: 23,111 Wrong classified: 2,865

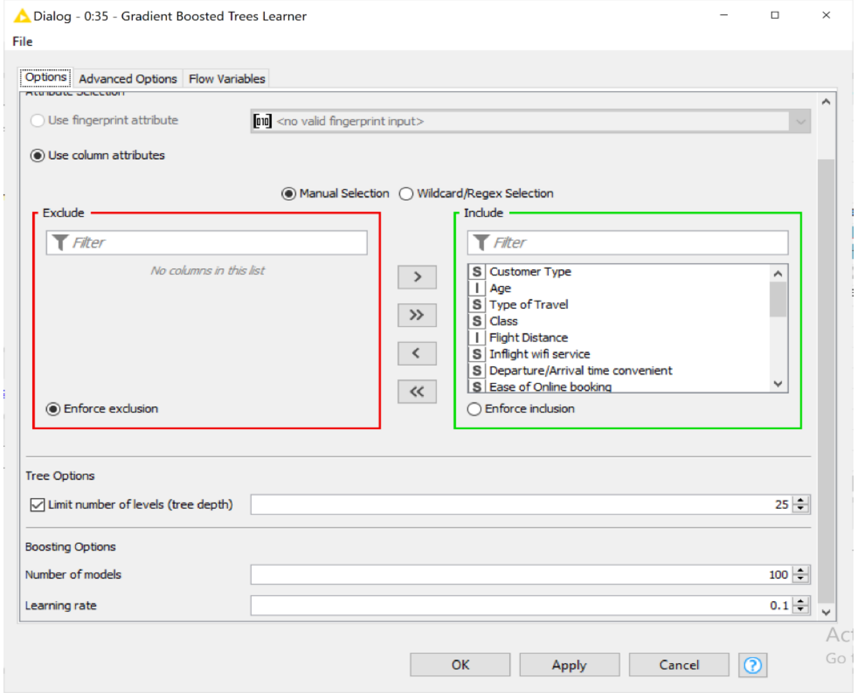
Accuracy: 88.971 % Error: 11.029 %

Cohen's kappa (κ) 0.776



ROC Curves - 0:27 - ROC Curve (local)

4. Gradient Boosting:



Dialog - 0:35 - Gradient Boosted Trees Learner

File

Options | Advanced Options | Flow Variables

Attribute Selection

○ Use fingerprint attribute <no valid fingerprint input>

● Use column attributes

● Manual Selection ○ Wildcard/Regex Selection

Exclude: No columns in this list

● Enforce exclusion

Include:
- S Customer Type
- I Age
- S Type of Travel
- S Class
- I Flight Distance
- S Inflight wifi service
- S Departure/Arrival time convenient
- S Ease of Online booking

○ Enforce inclusion

Tree Options

☑ Limit number of levels (tree depth) 25

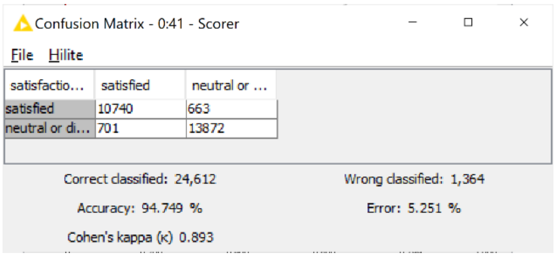
Boosting Options

Number of models 100

Learning rate 0.1

OK | Apply | Cancel

ROC & Scorer:



Confusion Matrix - 0:41 - Scorer

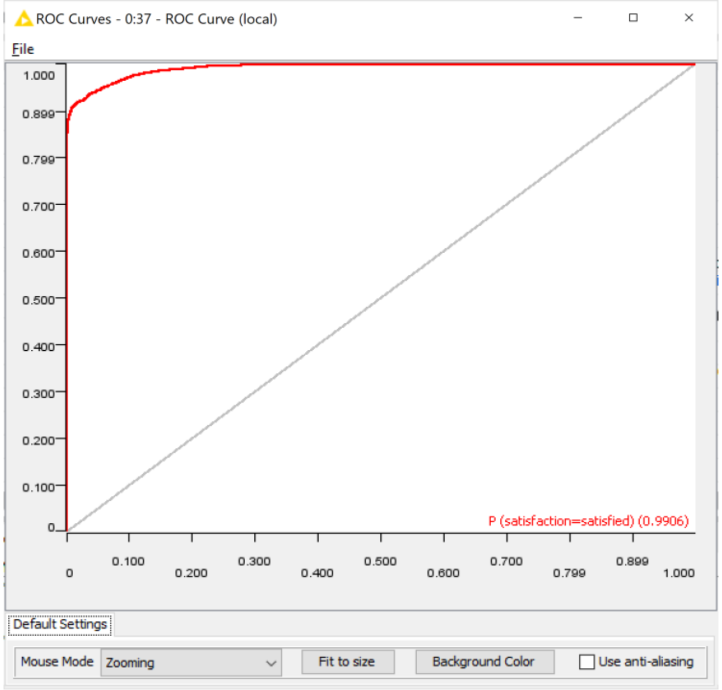
File Hilite

satisfactio…	satisfied	neutral or …
satisfied	10740	663
neutral or di…	701	13872

Correct classified: 24,612 Wrong classified: 1,364

Accuracy: 94.749 % Error: 5.251 %

Cohen's kappa (κ) 0.893



ROC Curves - 0:37 - ROC Curve (local)

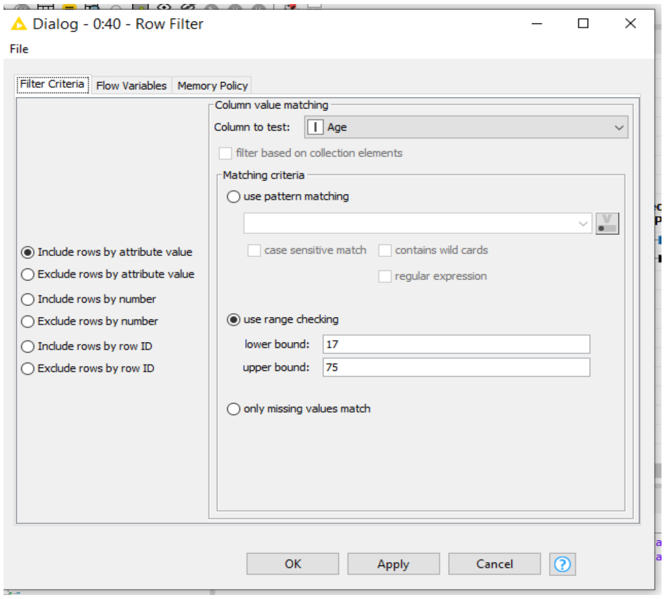
File

Attempt 3:

Data Preprocessing:

1. **Row Filter**

Keeping all the above conditions applied constant we applied a new filter -row filter- where we set the range of the age from 17 to 75 i.e. only the responses of people whose age lies in this range matter. People less than 18 are considered kids whose respond does not hold much value.



Dialog - 0:40 - Row Filter

File

Filter Criteria | Flow Variables | Memory Policy

● Include rows by attribute value
○ Exclude rows by attribute value
○ Include rows by number
○ Exclude rows by number
○ Include rows by row ID
○ Exclude rows by row ID

Column value matching

Column to test: Age

□ filter based on collection elements

Matching criteria

○ use pattern matching

□ case sensitive match □ contains wild cards

□ regular expression

● use range checking

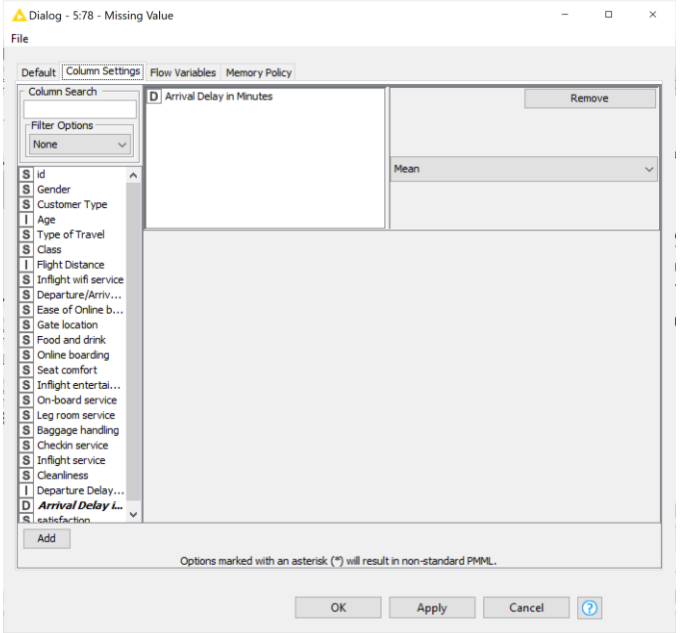
lower bound: 17

upper bound: 75

○ only missing values match

OK | Apply | Cancel

2. **Missing Value:** It was found that we had 310 missing data in arrival delay time so we used missing value filter.



Dialog - 5:78 - Missing Value

File

Default | Column Settings | Flow Variables | Memory Policy

Column Search

Filter Options: None

S id
S Gender
S Customer Type
I Age
S Type of Travel
S Class
I Flight Distance
S Inflight wifi service
S Departure/Arriv…
S Ease of Online b…
S Gate location
S Food and drink
S Online boarding
S Seat comfort
S Inflight entertai…
S On-board service
S Leg room service
S Baggage handling
S Checkin service
S Inflight service
S Cleanliness
I Departure Delay…
D *Arrival Delay i…*
S satisfaction

Add

D Arrival Delay in Minutes

Remove

Mean

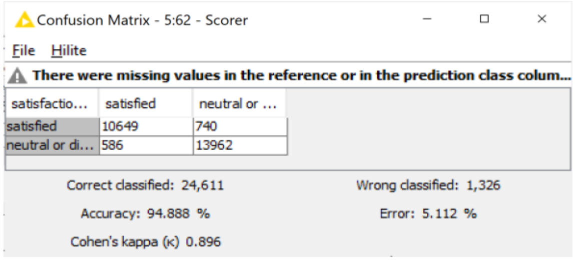
Options marked with an asterisk (*) will result in non-standard PMML.

OK | Apply | Cancel

<u>Models:</u> The configurations for all the models were kept constant, only a few filters mentioned above were added to our models. However the accuracy of all models were effected at most 0.01. Therefore it was concluded that applying the row filter did not have much of an effect.

1. Decision Tree

Scorer



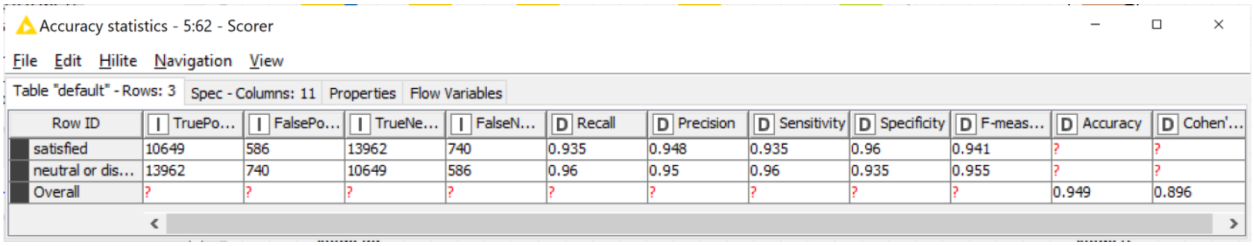
There were missing values in the reference or in the prediction class colum...

satisfactio...	satisfied	neutral or ...
satisfied	10649	740
neutral or di...	586	13962

Correct classified: 24,611 Wrong classified: 1,326

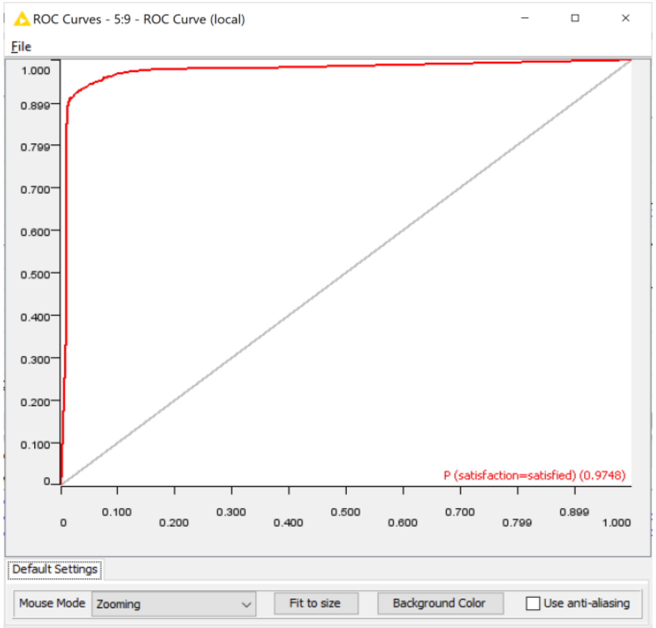
Accuracy: 94.888 % Error: 5.112 %

Cohen's kappa (κ) 0.896


Accuracy statistics - 5:62 - Scorer
File Edit Hilite Navigation View
Table "default" - Rows: 3
Spec - Columns: 11
Properties
Flow Variables

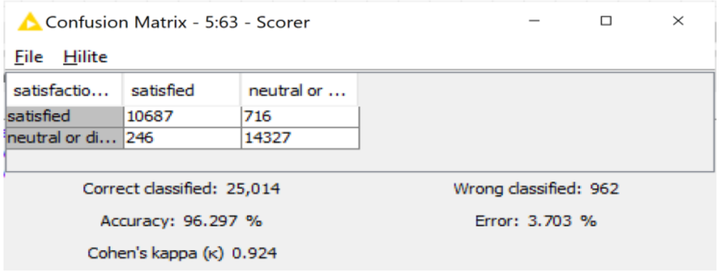
Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
satisfied	10649	586	13962	740	0.935	0.948	0.935	0.96	0.941	?	?
neutral or dis...	13962	740	10649	586	0.96	0.95	0.96	0.935	0.955	?	?
Overall	?	?	?	?	?	?	?	?	?	0.949	0.896

ROC:



2. Random Forest

ROC & Scorer:

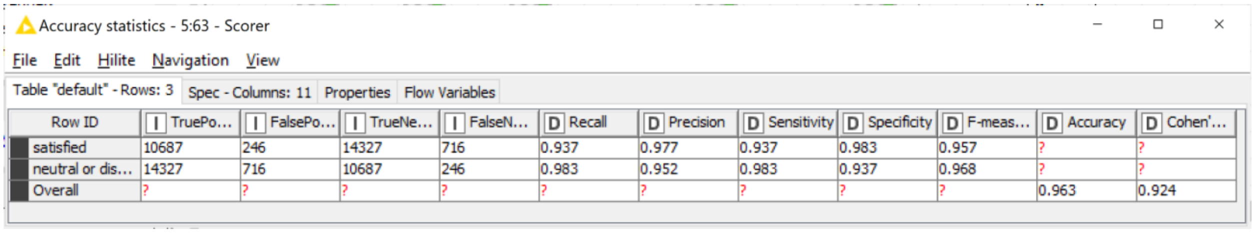


satisfactio...	satisfied	neutral or ...
satisfied	10687	716
neutral or di...	246	14327

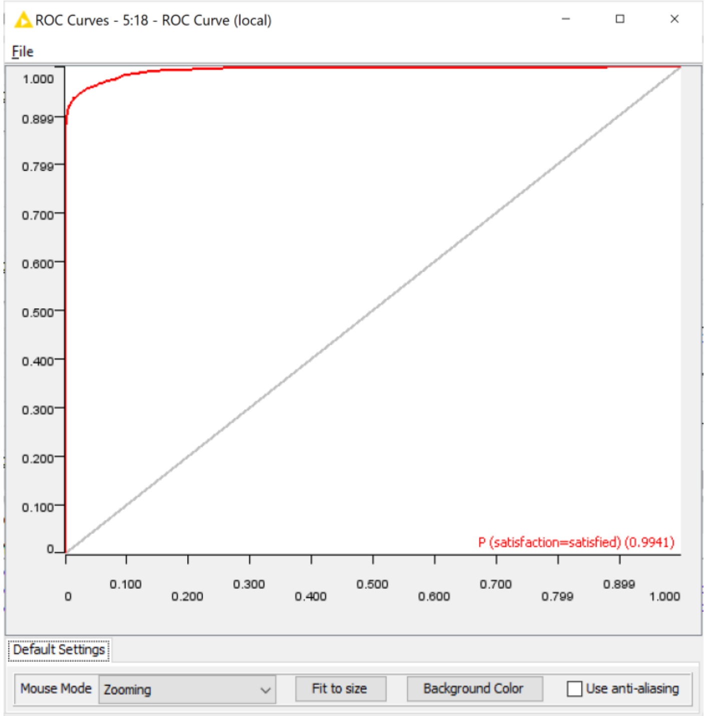
Correct classified: 25,014 Wrong classified: 962

Accuracy: 96.297 % Error: 3.703 %

Cohen's kappa (κ) 0.924

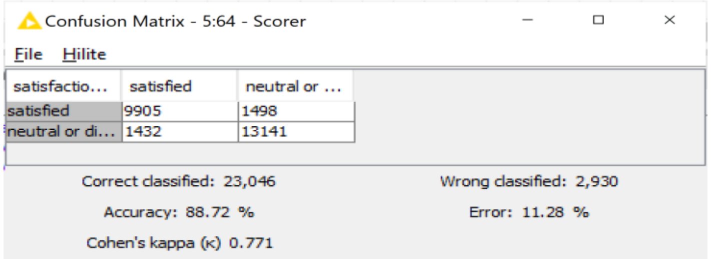

Accuracy statistics - 5:63 - Scorer
File Edit Hilite Navigation View
Table "default" - Rows: 3
Spec - Columns: 11
Properties
Flow Variables

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
satisfied	10687	246	14327	716	0.937	0.977	0.937	0.983	0.957	?	?
neutral or dis...	14327	716	10687	246	0.983	0.952	0.983	0.937	0.968	?	?
Overall	?	?	?	?	?	?	?	?	?	0.963	0.924



3. Naive Bayes

ROC & Scorer:



satisfactio…	satisfied	neutral or …
satisfied	9905	1498
neutral or di…	1432	13141

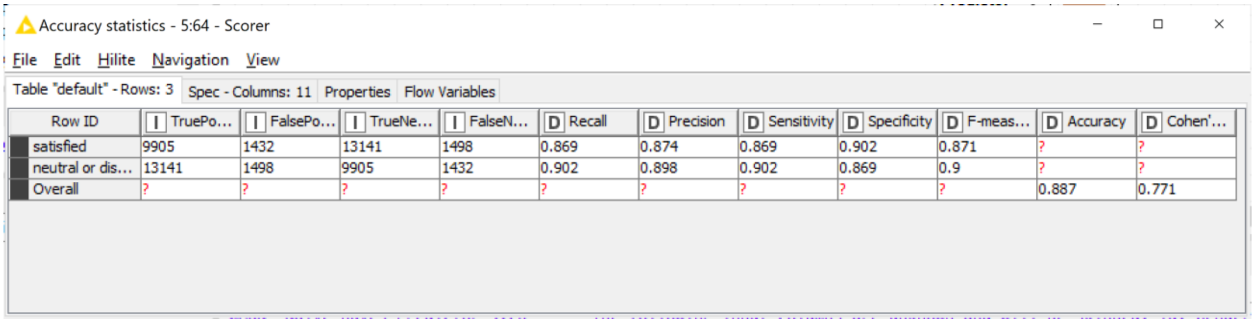
Correct classified: 23,046

Wrong classified: 2,930

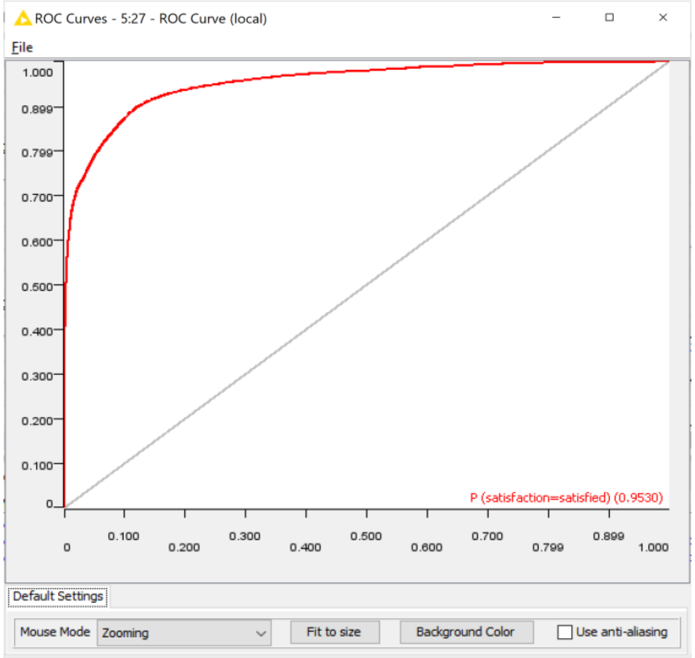
Accuracy: 88.72 %

Error: 11.28 %

Cohen's kappa (κ) 0.771

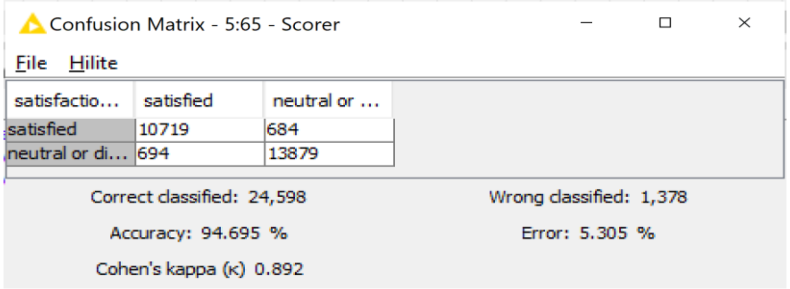

Accuracy statistics - 5:64 - Scorer
File Edit Hilite Navigation View
Table "default" - Rows: 3
Spec - Columns: 11
Properties
Flow Variables

Row ID	I TruePo…	I FalsePo…	I TrueNe…	I FalseN…	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas…	D Accuracy	D Cohen'…
satisfied	9905	1432	13141	1498	0.869	0.874	0.869	0.902	0.871	?	?
neutral or dis…	13141	1498	9905	1432	0.902	0.898	0.902	0.869	0.9	?	?
Overall	?	?	?	?	?	?	?	?	?	0.887	0.771



4. Gradient Boosting

ROC And Scorer:



Confusion Matrix - 5:65 - Scorer

File Hilite

satisfactio…	satisfied	neutral or …
satisfied	10719	684
neutral or di…	694	13879

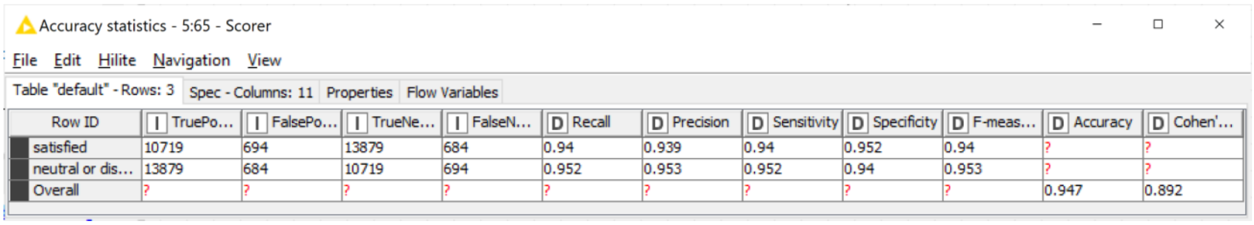
Correct classified: 24,598

Wrong classified: 1,378

Accuracy: 94.695 %

Error: 5.305 %

Cohen's kappa (κ) 0.892

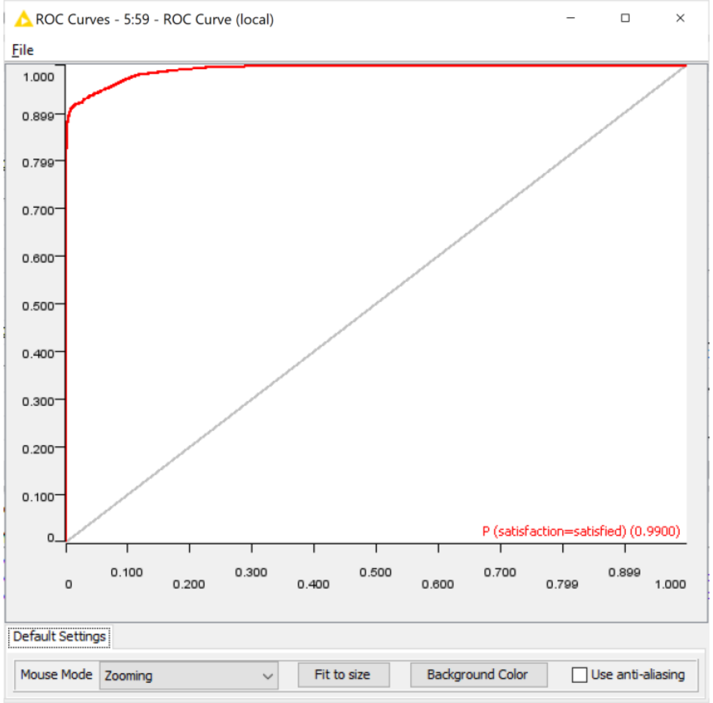


Accuracy statistics - 5:65 - Scorer

File Edit Hilite Navigation View

Table "default" - Rows: 3 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	I TruePo…	I FalsePo…	I TrueNe…	I FalseN…	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas…	D Accuracy	D Cohen'…
satisfied	10719	694	13879	684	0.94	0.939	0.94	0.952	0.94	?	?
neutral or dis…	13879	684	10719	694	0.952	0.953	0.952	0.94	0.953	?	?
Overall	?	?	?	?	?	?	?	?	?	0.947	0.892

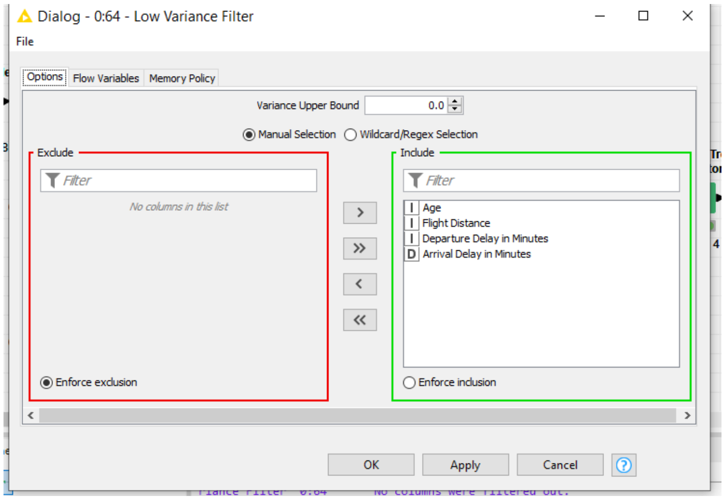


ROC Curves - 5:59 - ROC Curve (local)

Attempt 4:

Data Preprocessing

Low Variance Filter-In this attempt we applied low variance filter that ignores values that are below a certain range, however in our data set there were no columns that can be filtered out using low variance filter. hence , nodes' configurations were not changed a bit. nodes' configurations were not changed a bit.



Dialog - 0:64 - Low Variance Filter

File

Options | Flow Variables | Memory Policy

Variance Upper Bound 0.0

● Manual Selection ○ Wildcard/Regex Selection

Exclude

Filter

No columns in this list

● Enforce exclusion

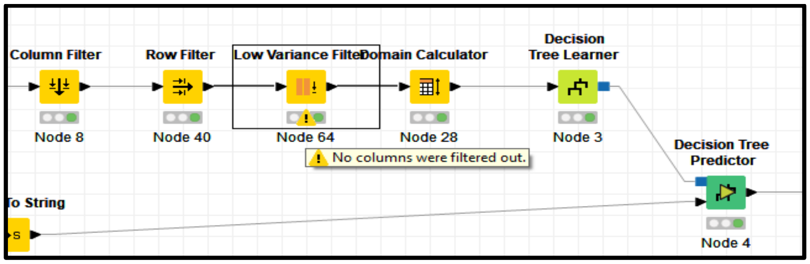
Include

Filter

I Age
I Flight Distance
I Departure Delay in Minutes
D Arrival Delay in Minutes

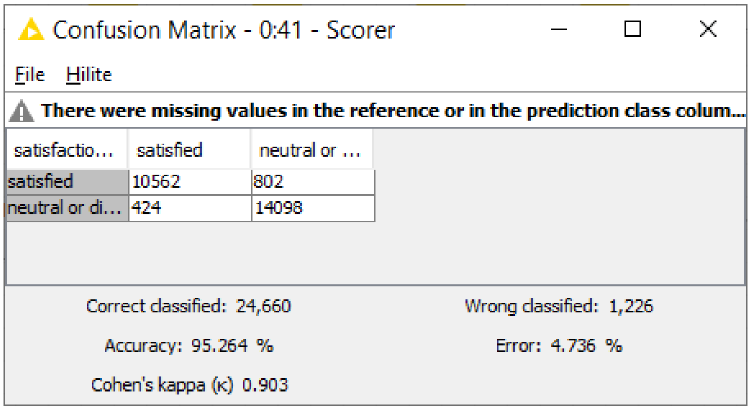
○ Enforce inclusion

OK | Apply | Cancel



Model Building

Since the low variance filter did not affect any of the columns therefore the accuracy was also not affected. Accuracy is the same as in attempt 3 for all the classification models.



Confusion Matrix - 0:41 - Scorer

File Hilite

⚠ There were missing values in the reference or in the prediction class colum...

satisfactio...	satisfied	neutral or ...
satisfied	10562	802
neutral or di...	424	14098

Correct classified: 24,660 Wrong classified: 1,226

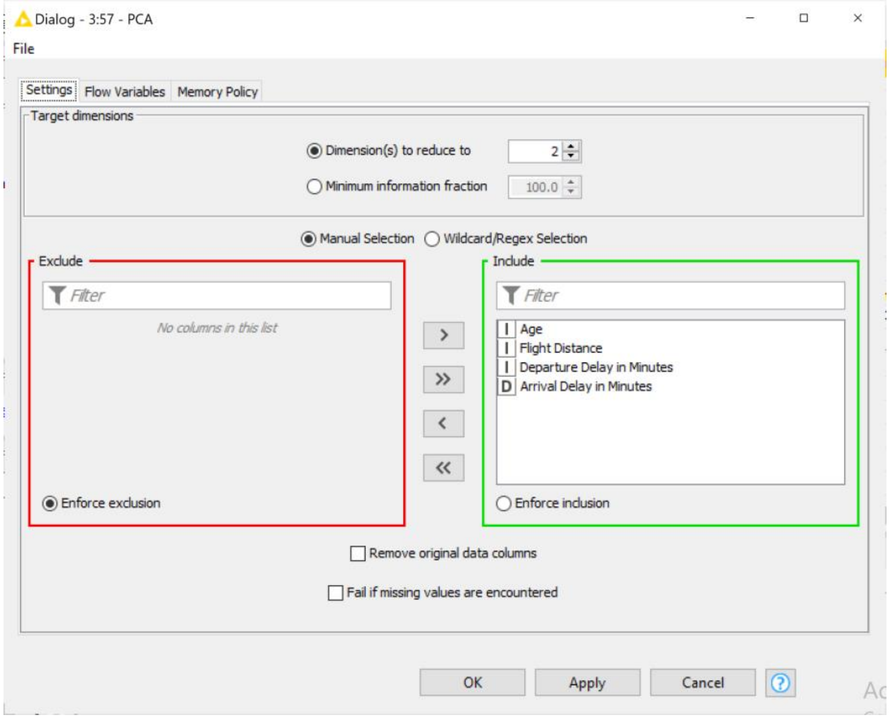
Accuracy: 95.264 % Error: 4.736 %

Cohen's kappa (κ) 0.903

<u>Attempt 5:</u>

Data Preprocessing:

1. **PCA:**
   Keeping the same configuration as Attempt 3, we added PCA to data preprocessing. Four numerical columns were reduced to two.



Target dimensions

● Dimension(s) to reduce to 2

○ Minimum information fraction 100.0

● Manual Selection ○ Wildcard/Regex Selection

Exclude

Filter

No columns in this list

● Enforce exclusion

Include

Filter

- I Age
- I Flight Distance
- I Departure Delay in Minutes
- D Arrival Delay in Minutes

○ Enforce inclusion

□ Remove original data columns

□ Fail if missing values are encountered

<u>Model Building</u>

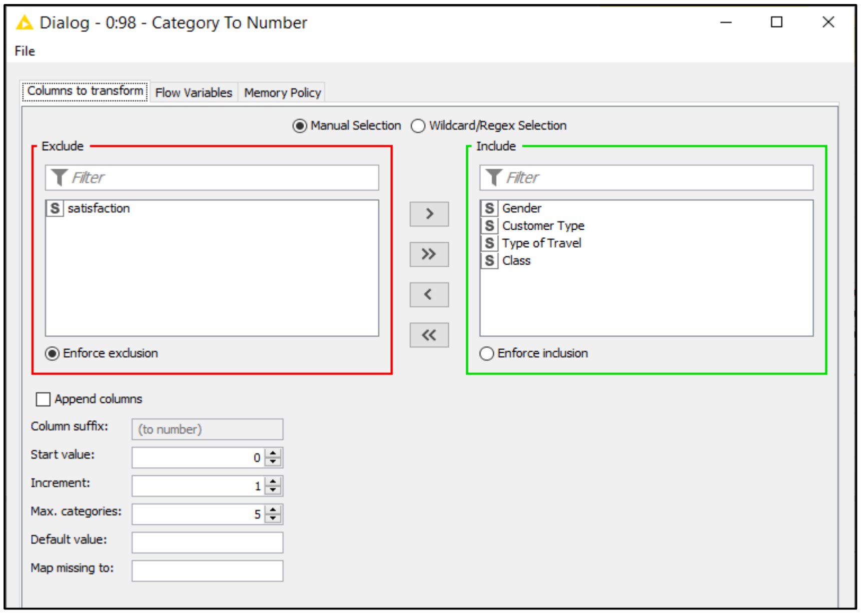
The configuration of all the classification models is kept the same as attempt 2 and 3. We found that the accuracy of each model declined after using PCA therefore we did not use it in the next attempts.

Attempt 6:

Data Preprocessing

1. **Category to Number:**

In this attempt we converted all the string columns to numerical values using category to number converter as the model used required all the values to be numerical. All the other filters used above were used with the same configuration.



● Manual Selection ○ Wildcard/Regex Selection

Exclude: S satisfaction

● Enforce exclusion

Include: S Gender, S Customer Type, S Type of Travel, S Class

○ Enforce inclusion

□ Append columns

Column suffix: (to number)

Start value: 0

Increment: 1

Max. categories: 5

Default value:

Map missing to:

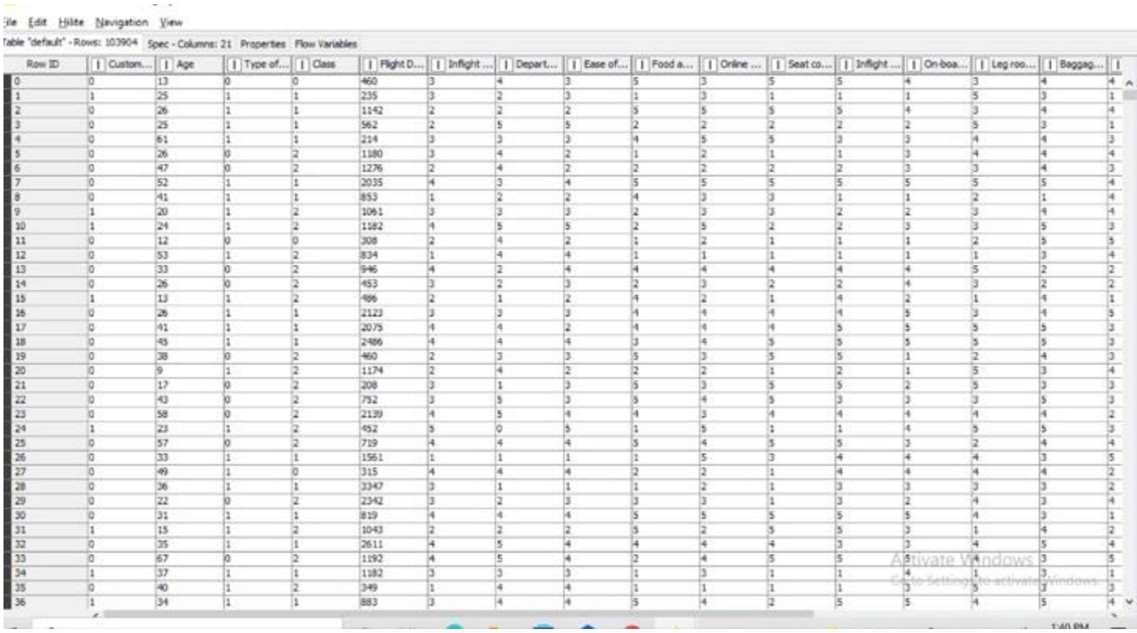


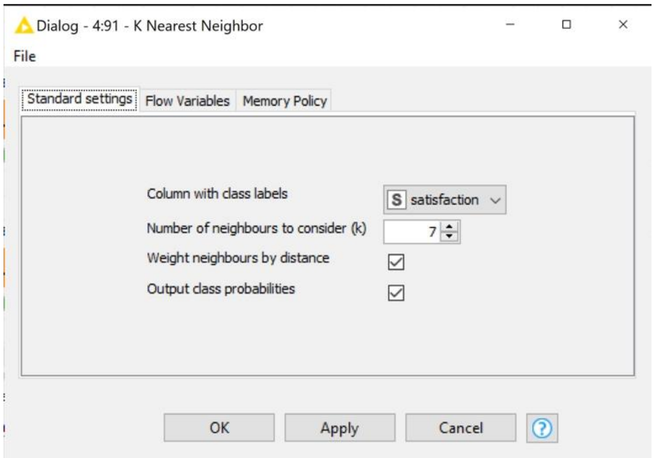
Table "default" - Rows: 103904 Spec - Columns: 21

Row ID	Custom...	Age	Type of...	Class	Flight D...	Inflight ...	Depart...	Ease of...	Food a...	Online ...	Seat co...	Inflight ...	On-boa...	Leg roo...	Baggag...	
0	0	13	0	0	460	3	4	3	5	3	5	5	4	3	4	4
1	1	25	1	1	235	3	2	3	1	3	1	1	1	5	3	1
2	0	26	1	1	1142	2	2	2	5	5	5	5	4	3	4	4
3	0	25	1	1	562	2	5	5	2	2	2	2	2	5	3	1
4	0	61	1	1	214	3	3	3	4	5	5	3	3	4	4	3
5	0	26	0	2	1180	3	4	2	1	2	1	1	3	4	4	4
6	0	47	0	2	1276	2	4	2	2	2	2	2	3	3	4	3
7	0	52	1	1	2035	4	3	4	5	5	5	5	5	5	5	4
8	0	41	1	1	853	1	2	2	4	3	3	1	1	2	1	4
9	1	20	1	2	1061	3	3	3	2	3	3	2	2	3	4	4
10	1	24	1	2	1182	4	5	5	2	5	2	2	3	3	5	3
11	0	12	0	0	308	2	4	2	1	2	1	1	1	2	5	5
12	0	53	1	2	834	1	4	4	1	1	1	1	1	1	3	4
13	0	33	0	2	946	4	2	4	4	4	4	4	4	5	2	2
14	0	26	0	2	453	3	2	3	2	3	2	2	4	3	2	2
15	1	13	1	2	486	2	1	2	4	2	1	4	2	1	4	1
16	0	26	1	1	2123	3	3	3	4	4	4	4	5	3	4	5
17	0	41	1	1	2075	4	4	2	4	4	4	5	5	5	5	3
18	0	45	1	1	2486	4	4	4	3	4	5	5	5	5	5	3
19	0	38	0	2	460	2	3	3	5	3	5	5	1	2	4	3
20	0	9	1	2	1174	2	4	2	2	2	1	2	1	5	3	4
21	0	17	0	2	208	3	1	3	5	3	5	5	2	5	3	3
22	0	43	0	2	752	3	5	3	5	4	5	3	3	3	5	3
23	0	58	0	2	2139	4	5	4	4	3	4	4	4	4	4	2
24	1	23	1	2	452	5	0	5	1	5	1	1	4	5	5	3
25	0	57	0	2	719	4	4	4	5	4	5	5	3	2	4	4
26	0	33	1	1	1561	1	1	1	1	5	3	4	4	4	3	5
27	0	49	1	0	315	4	4	4	2	2	1	4	4	4	4	2
28	0	36	1	1	3347	3	1	1	1	2	1	3	3	3	3	2
29	0	22	0	2	2342	3	2	3	3	3	1	3	2	4	3	4
30	0	31	1	1	819	4	4	4	5	5	5	5	5	4	3	1
31	1	15	1	2	1043	2	2	2	5	2	5	5	3	1	4	2
32	0	35	1	1	2611	4	5	4	4	4	4	3	3	4	5	4
33	0	67	0	2	1192	4	5	4	2	4	5	5	5	4	3	5
34	1	37	1	1	1182	3	3	3	1	3	1	1	4	1	3	1
35	0	40	1	2	349	1	4	4	1	1	1	1	3	5	3	3
36	1	34	1	1	883	3	4	4	5	4	2	5	5	4	5	4

Model Building:

K Nearest Neighbor

For this attempt we used KNN model, with the following configuration.



Dialog - 4:91 - K Nearest Neighbor

File

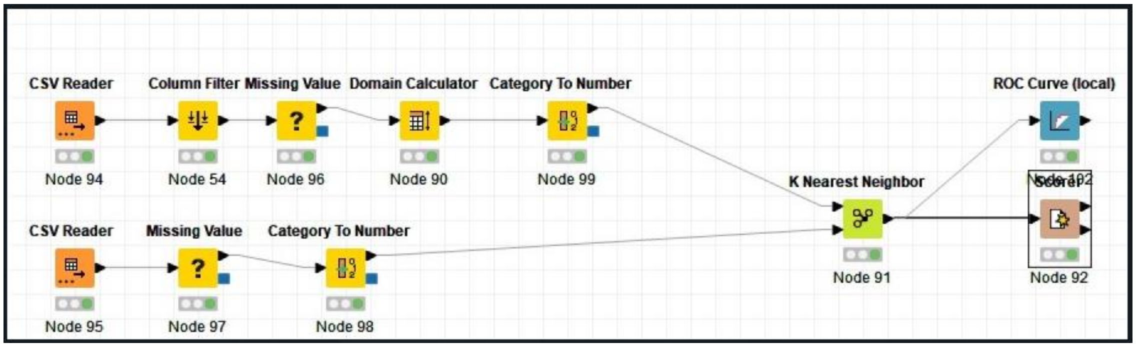
Standard settings | Flow Variables | Memory Policy

Column with class labels: S satisfaction

Number of neighbours to consider (k): 7

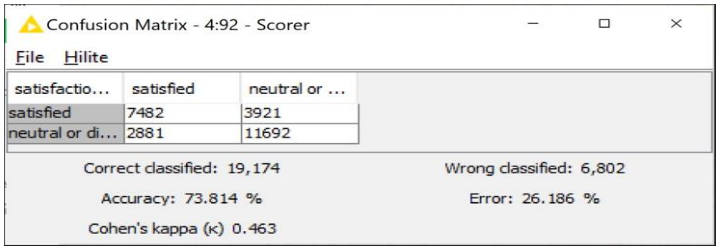
Weight neighbours by distance ☑

Output class probabilities ☑



Scorer

Using KNN the accuracy of our model decline from approximately 95% to 80%.



Confusion Matrix - 4:92 - Scorer

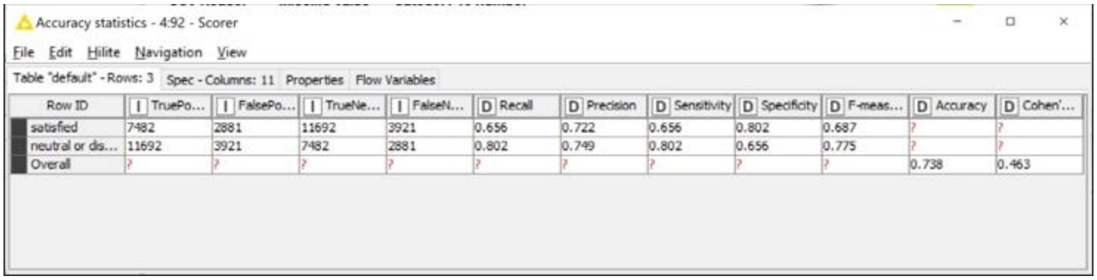
File Hilite

satisfactio...	satisfied	neutral or ...
satisfied	7482	3921
neutral or di...	2881	11692

Correct classified: 19,174 Wrong classified: 6,802

Accuracy: 73.814 % Error: 26.186 %

Cohen's kappa (κ) 0.463



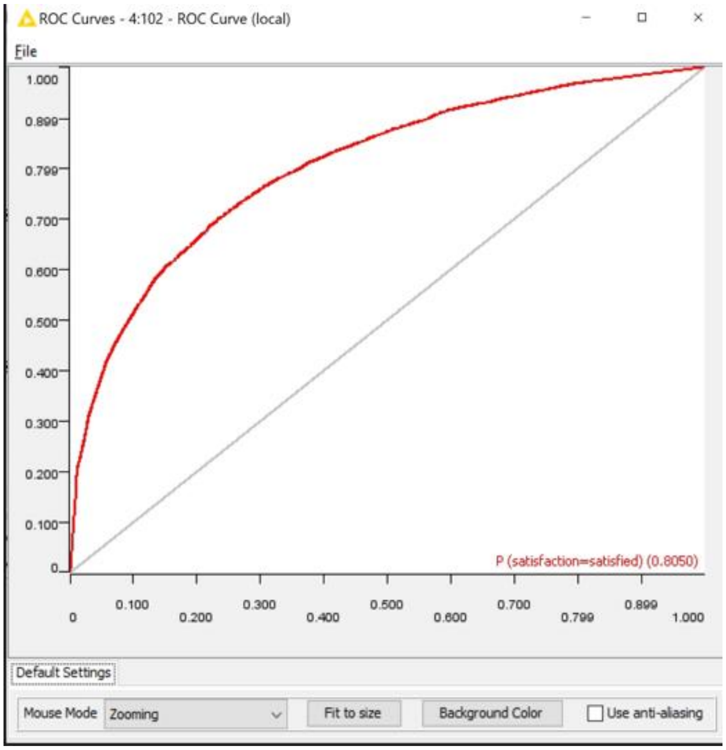
Accuracy statistics - 4:92 - Scorer

File Edit Hilite Navigation View

Table "default" - Rows: 3 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
satisfied	7482	2881	11692	3921	0.656	0.722	0.656	0.802	0.687	?	?
neutral or dis...	11692	3921	7482	2881	0.802	0.749	0.802	0.656	0.775	?	?
Overall	?	?	?	?	?	?	?	?	?	0.738	0.463

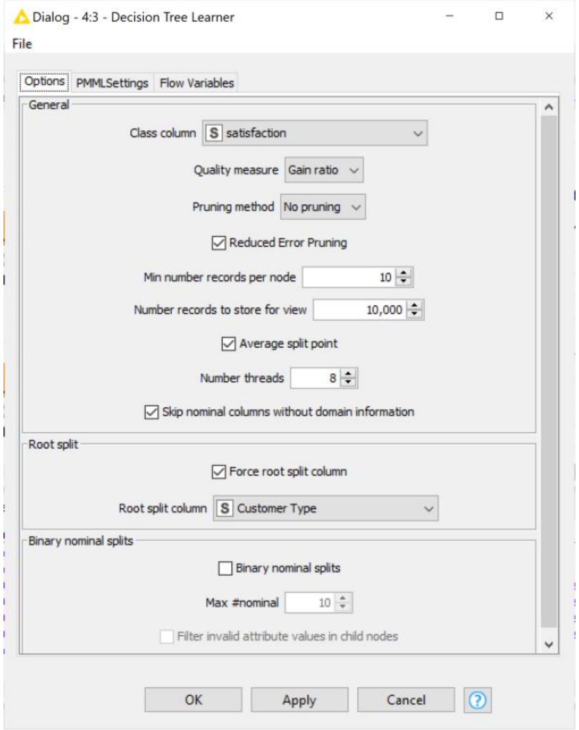
ROC:


ROC Curves - 4:102 - ROC Curve (local)
File
1.000
0.899
0.799
0.700
0.600
0.500
0.400
0.300
0.200
0.100
0
P (satisfaction=satisfied) (0.8050)
0
0.100
0.200
0.300
0.400
0.500
0.600
0.700
0.799
0.899
1.000
Default Settings
Mouse Mode
Zooming
Fit to size
Background Color
□ Use anti-aliasing

Attempt 7:

Changing configurations in Decision Tree:

7.1- For this attempt, we changed the minimum number of records per node from 5 to 10 and Root split the columns from Customer Type (Loyal or Disloyal Customers). We tried root split using other attributes as well but we got the highest accuracy using Customer Type.



General

Class column: S satisfaction

Quality measure: Gain ratio

Pruning method: No pruning

☑ Reduced Error Pruning

Min number records per node: 10

Number records to store for view: 10,000

☑ Average split point

Number threads: 8

☑ Skip nominal columns without domain information

Root split

☑ Force root split column

Root split column: S Customer Type

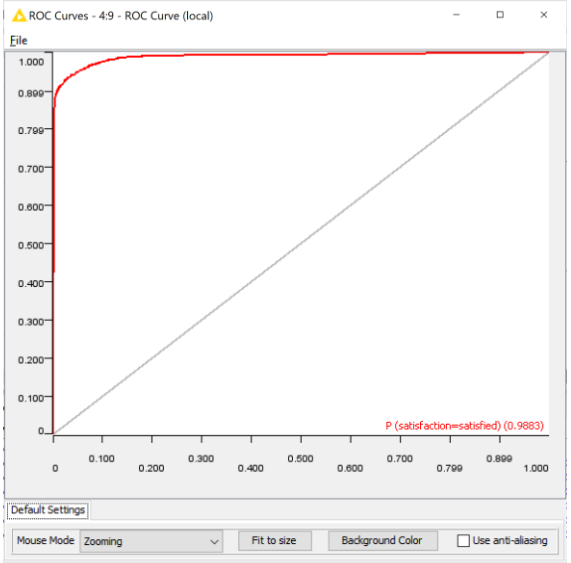
Binary nominal splits

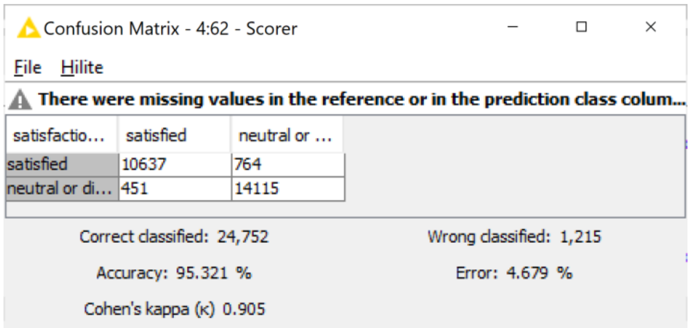
☐ Binary nominal splits

Max #nominal: 10

☐ Filter invalid attribute values in child nodes

ROC And Scorer:





⚠ There were missing values in the reference or in the prediction class colum...

satisfactio...	satisfied	neutral or ...
satisfied	10637	764
neutral or di...	451	14115

Correct classified: 24,752

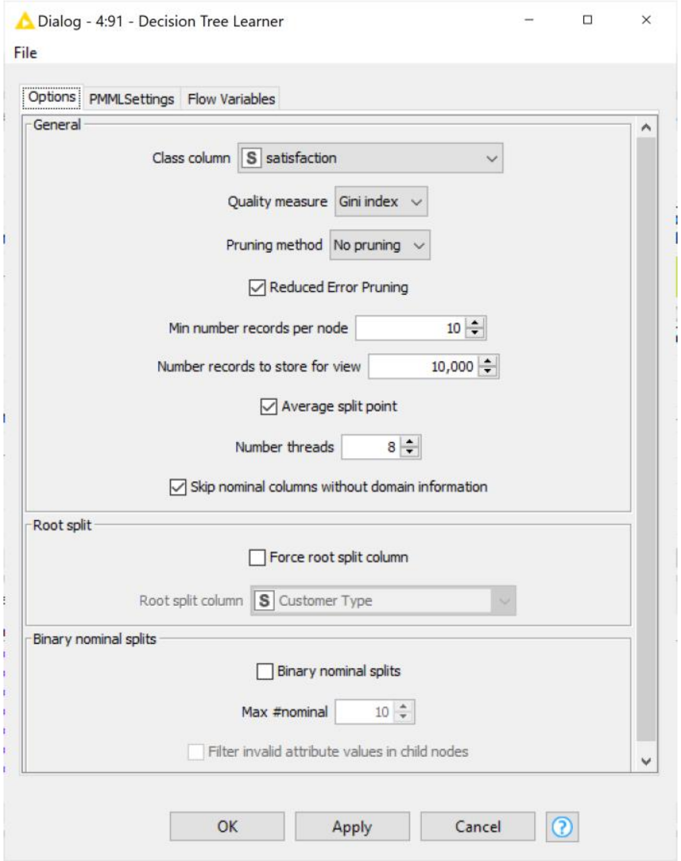
Wrong classified: 1,215

Accuracy: 95.321 %

Error: 4.679 %

Cohen's kappa (κ) 0.905

7.2- Changing the Quality measure from gain ratio to gini index reduced our accuracy.



Dialog - 4:91 - Decision Tree Learner

File

Options | PMMLSettings | Flow Variables

General

Class column: S satisfaction

Quality measure: Gini index

Pruning method: No pruning

☑ Reduced Error Pruning

Min number records per node: 10

Number records to store for view: 10,000

☑ Average split point

Number threads: 8

☑ Skip nominal columns without domain information

Root split

□ Force root split column

Root split column: S Customer Type

Binary nominal splits

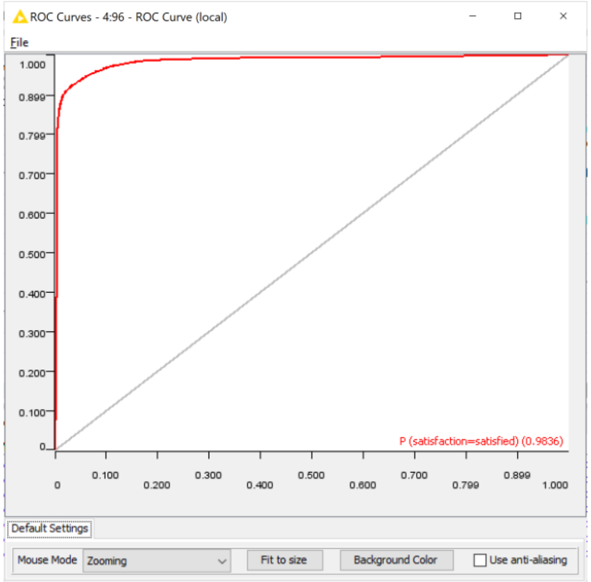
□ Binary nominal splits

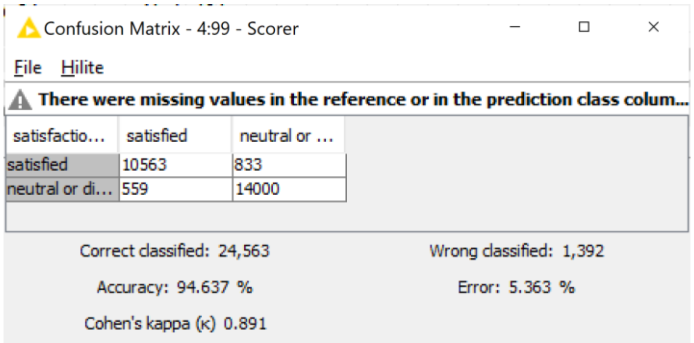
Max #nominal: 10

□ Filter invalid attribute values in child nodes

OK | Apply | Cancel

ROC And Scorer:





Confusion Matrix - 4:99 - Scorer

File Hilite

There were missing values in the reference or in the prediction class colum...

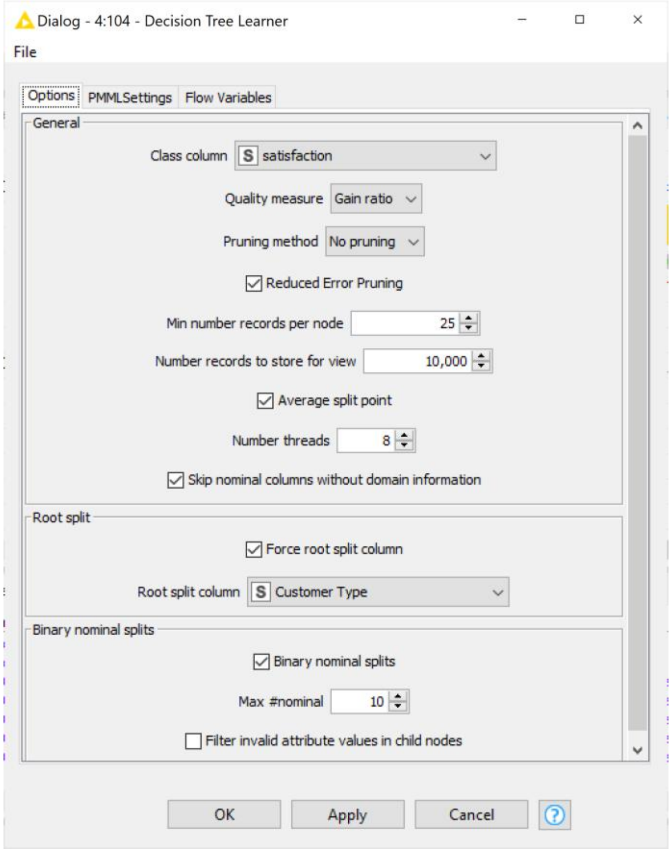
satisfactio...	satisfied	neutral or ...
satisfied	10563	833
neutral or di...	559	14000

Correct classified: 24,563 Wrong classified: 1,392

Accuracy: 94.637 % Error: 5.363 %

Cohen's kappa (κ) 0.891

7.3- Changed the Quality Measure back to Gain ratio with 25 as the minimum number of records per node, force splitting the column at customer type and binary nominal splits.



Dialog - 4:104 - Decision Tree Learner

File

Options | PMMLSettings | Flow Variables

General

Class column: S satisfaction

Quality measure: Gain ratio

Pruning method: No pruning

☑ Reduced Error Pruning

Min number records per node: 25

Number records to store for view: 10,000

☑ Average split point

Number threads: 8

☑ Skip nominal columns without domain information

Root split

☑ Force root split column

Root split column: S Customer Type

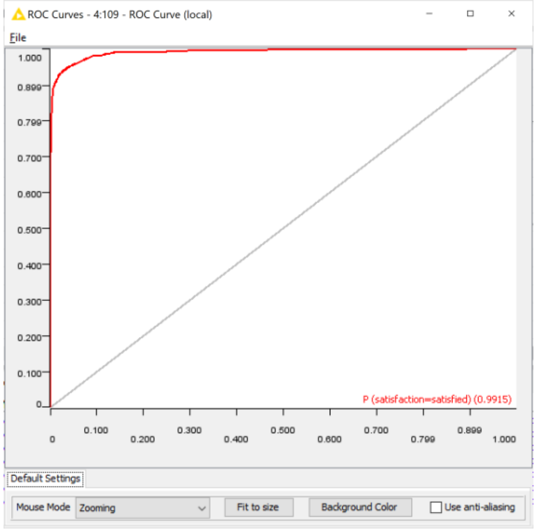
Binary nominal splits

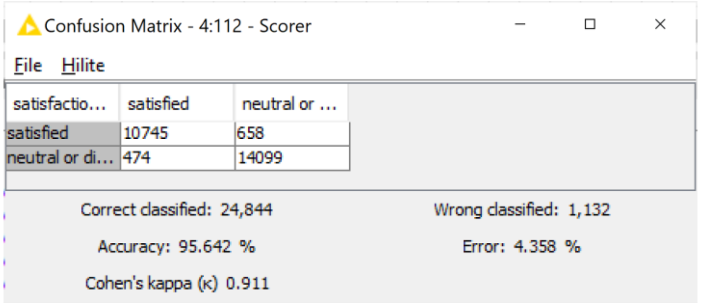
☑ Binary nominal splits

Max #nominal: 10

☐ Filter invalid attribute values in child nodes

ROC and Scorer:





Confusion Matrix - 4:112 - Scorer

File Hilite

satisfactio...	satisfied	neutral or ...
satisfied	10745	658
neutral or di...	474	14099

Correct classified: 24,844

Wrong classified: 1,132

Accuracy: 95.642 %

Error: 4.358 %

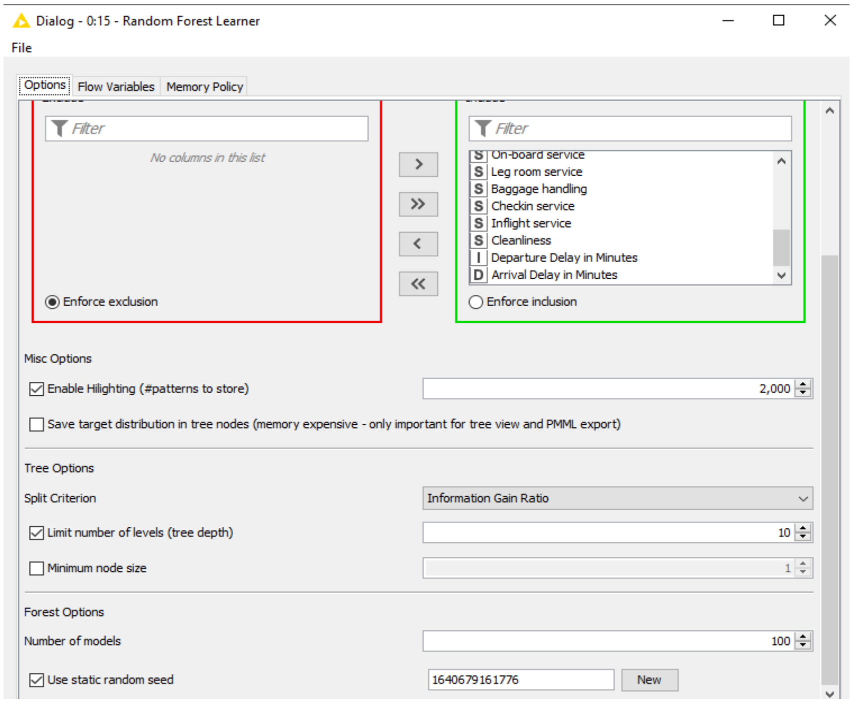
Cohen's kappa (κ) 0.911

For attempts 8-10 we did not made changes in our data preprocessing but we made changes in our model building to check if it had any effect on our accuracy or not.

Attempt 8:

Changing configurations in Random Forest:

8.1- I limited the number of levels (tree depth) to 10.


Dialog - 0:15 - Random Forest Learner

File

Options | Flow Variables | Memory Policy

Filter

No columns in this list

● Enforce exclusion

Filter

S	On-board service
S	Leg room service
S	Baggage handling
S	Checkin service
S	Inflight service
S	Cleanliness
I	Departure Delay in Minutes
D	Arrival Delay in Minutes

○ Enforce inclusion

Misc Options

☑ Enable Hilighting (#patterns to store) 2,000

□ Save target distribution in tree nodes (memory expensive - only important for tree view and PMML export)

Tree Options

Split Criterion: Information Gain Ratio

☑ Limit number of levels (tree depth) 10

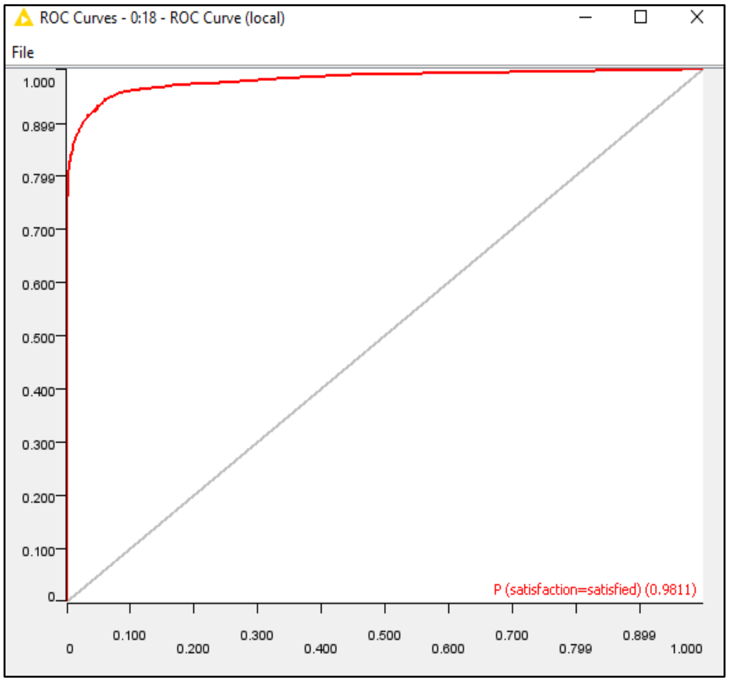
□ Minimum node size 1

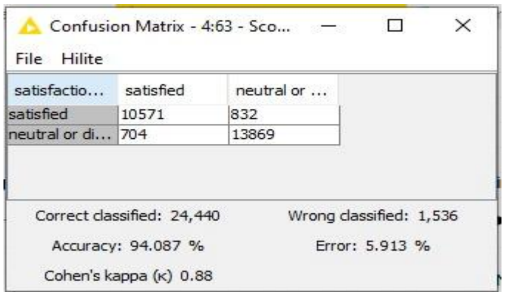
Forest Options

Number of models 100

☑ Use static random seed 1640679161776 New

ROC And Scorer:




Confusion Matrix - 4:63 - Sco...

File Hilite

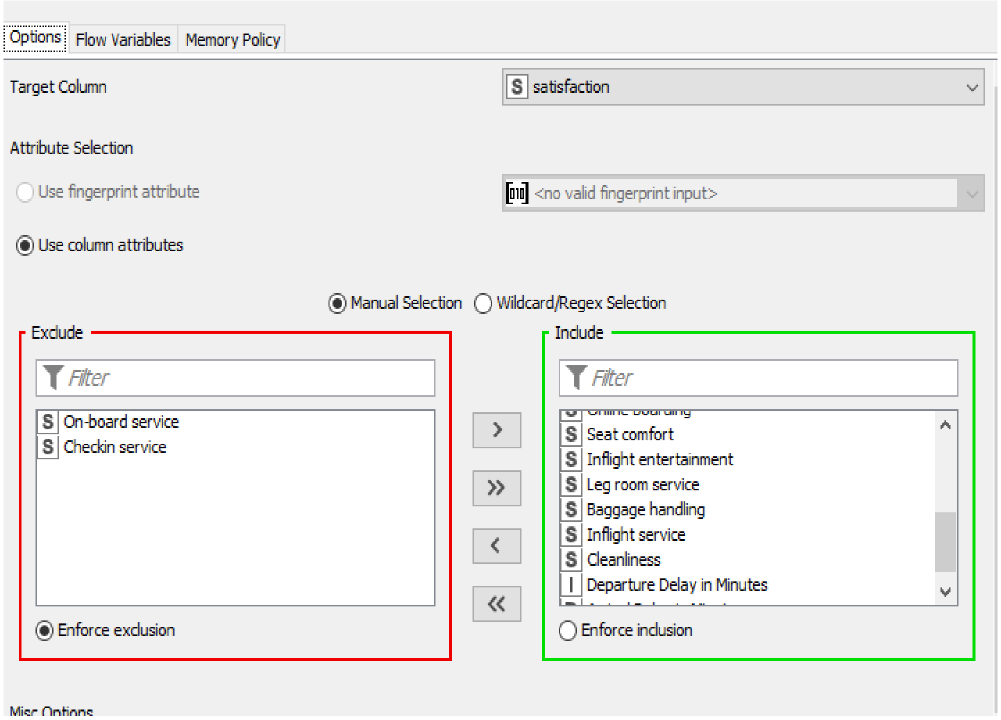
satisfactio...	satisfied	neutral or ...
satisfied	10571	832
neutral or di...	704	13869

Correct classified: 24,440 Wrong classified: 1,536

Accuracy: 94.087 % Error: 5.913 %

Cohen's kappa (κ) 0.88

8.2- Reset the changes made in 8.1. Then tried excluding on-board service and checkin service. Just to see if it will affect the results.



Target Column: S satisfaction

Attribute Selection

○ Use fingerprint attribute <no valid fingerprint input>

● Use column attributes

● Manual Selection ○ Wildcard/Regex Selection

Exclude
- S On-board service
- S Checkin service

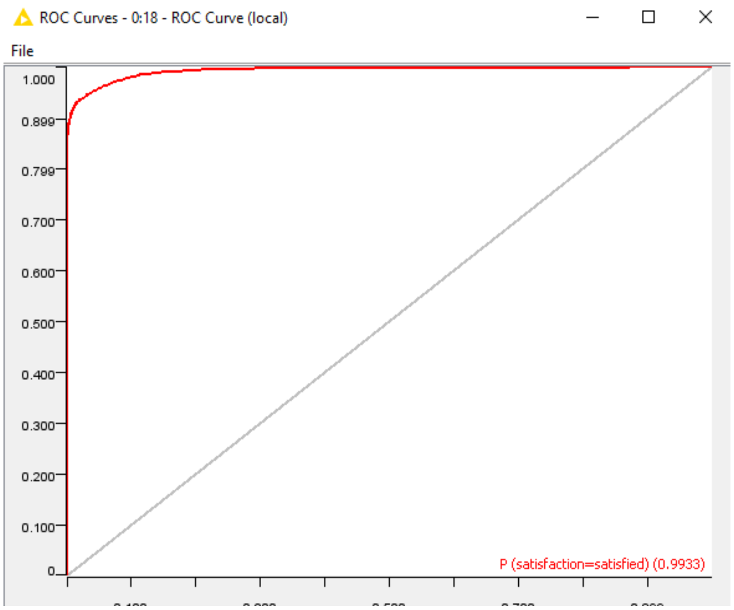
● Enforce exclusion

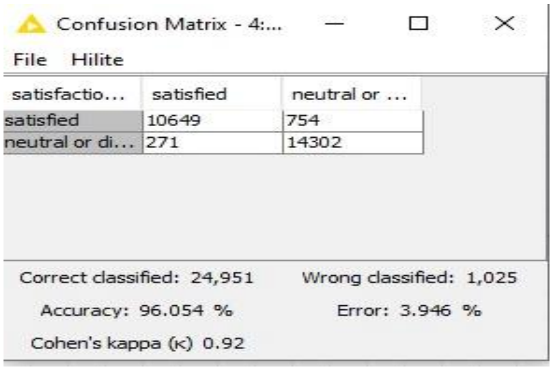
Include
- S Online Boarding
- S Seat comfort
- S Inflight entertainment
- S Leg room service
- S Baggage handling
- S Inflight service
- S Cleanliness
- I Departure Delay in Minutes

○ Enforce inclusion

Misc Options

ROC And Scorer:





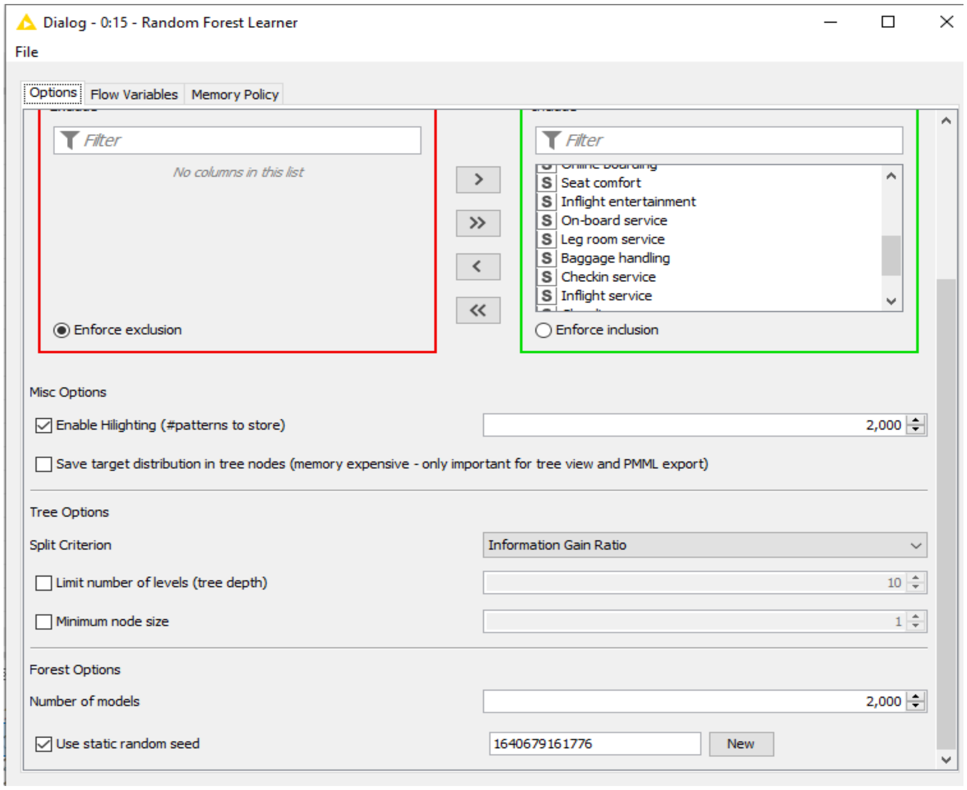
satisfactio...	satisfied	neutral or ...
satisfied	10649	754
neutral or di...	271	14302

Correct classified: 24,951 Wrong classified: 1,025

Accuracy: 96.054 % Error: 3.946 %

Cohen's kappa (κ) 0.92

8.3- Reset the changes made in 8.2. Increased the number of models from 100 to 2000.



Dialog - 0:15 - Random Forest Learner

File

Options | Flow Variables | Memory Policy

Filter

No columns in this list

● Enforce exclusion

Filter

S Seat comfort
S Inflight entertainment
S On-board service
S Leg room service
S Baggage handling
S Checkin service
S Inflight service

○ Enforce inclusion

Misc Options

☑ Enable Hilighting (#patterns to store) 2,000

□ Save target distribution in tree nodes (memory expensive - only important for tree view and PMML export)

Tree Options

Split Criterion: Information Gain Ratio

□ Limit number of levels (tree depth) 10

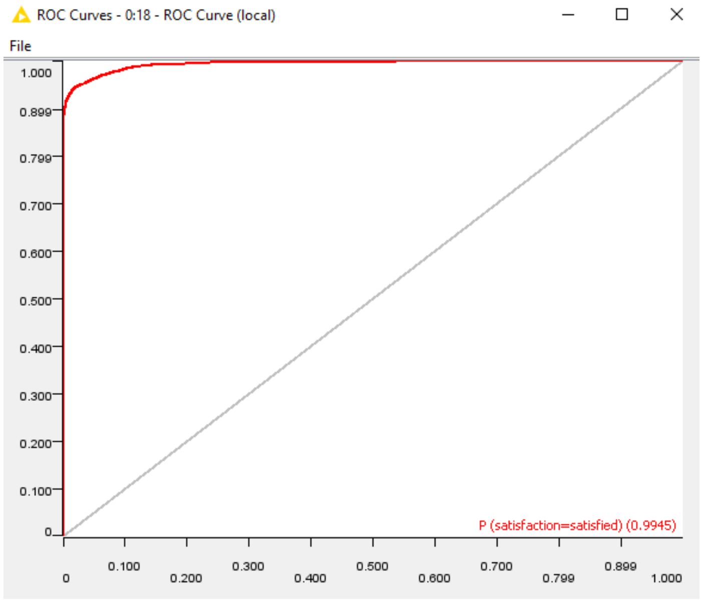
□ Minimum node size 1

Forest Options

Number of models: 2,000

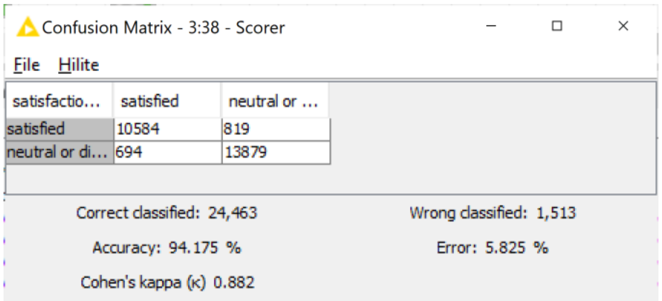
☑ Use static random seed 1640679161776 New

ROC And Scorer:



ROC Curves - 0:18 - ROC Curve (local)

File



Confusion Matrix - 3:38 - Scorer

File Hilite

satisfactio...	satisfied	neutral or ...
satisfied	10584	819
neutral or di...	694	13879

Correct classified: 24,463 Wrong classified: 1,513

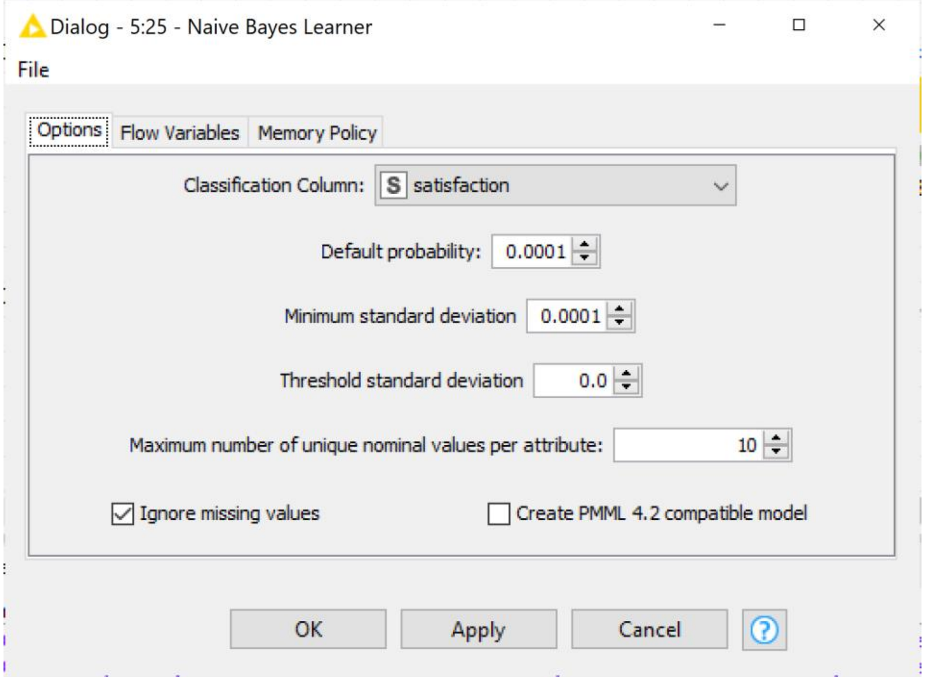
Accuracy: 94.175 % Error: 5.825 %

Cohen's kappa (κ) 0.882

Attempt 9:

Changing configurations in Naive Bayes Learner:

9.1- I changed the maximum number of unique nominal values per attribute from 20 to 10, to check whether it affects the accuracy or not.



Classification Column: S satisfaction

Default probability: 0.0001

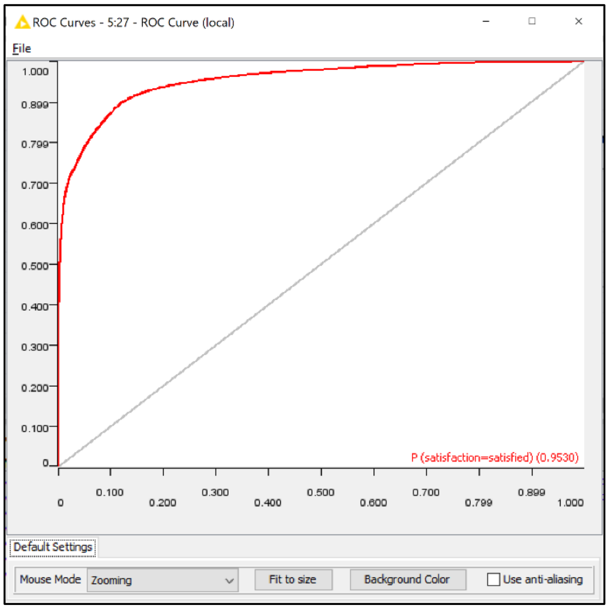
Minimum standard deviation 0.0001

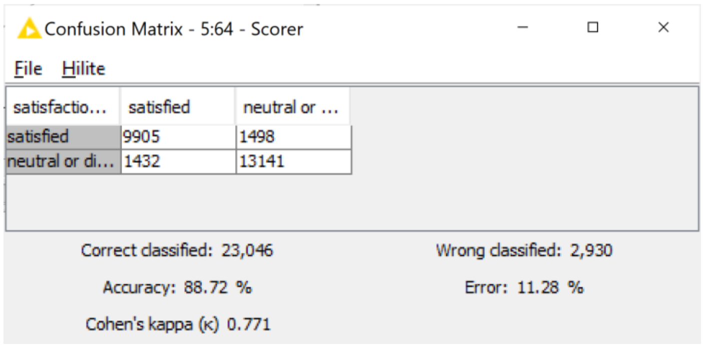
Threshold standard deviation 0.0

Maximum number of unique nominal values per attribute: 10

☑ Ignore missing values ☐ Create PMML 4.2 compatible model

ROC And Scorer:





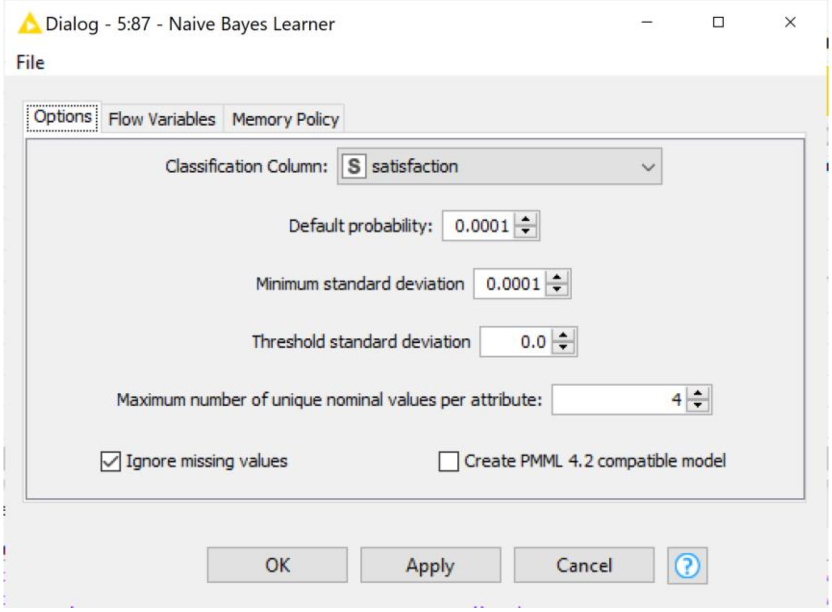
satisfactio...	satisfied	neutral or ...
satisfied	9905	1498
neutral or di...	1432	13141

Correct classified: 23,046 Wrong classified: 2,930

Accuracy: 88.72 % Error: 11.28 %

Cohen's kappa (κ) 0.771

9.2- Reset the changes made in 9.1, realized if we go below 6 in the maximum number of unique nominal values per attribute the accuracy drops sharply.



Dialog - 5:87 - Naive Bayes Learner

File

Options | Flow Variables | Memory Policy

Classification Column: S satisfaction

Default probability: 0.0001

Minimum standard deviation 0.0001

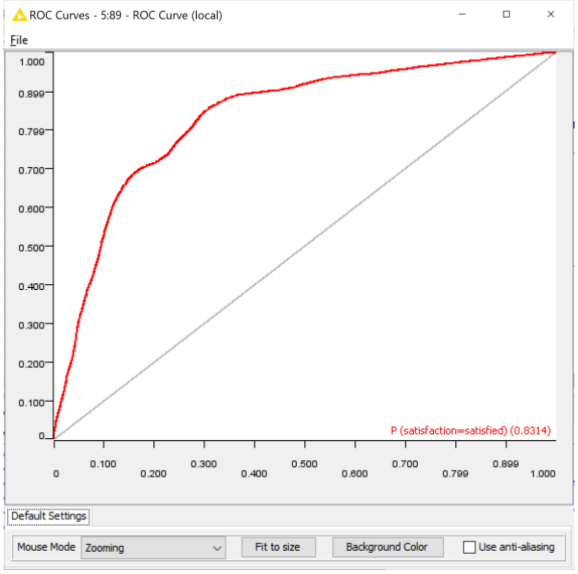
Threshold standard deviation 0.0

Maximum number of unique nominal values per attribute: 4

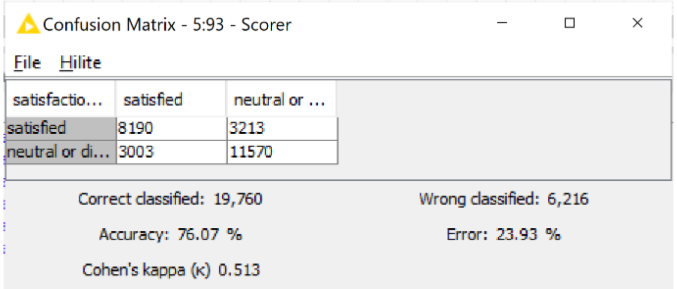
☑ Ignore missing values

□ Create PMML 4.2 compatible model

ROC And Scorer:



ROC Curves - 5:89 - ROC Curve (local)



Confusion Matrix - 5:93 - Scorer

satisfactio...	satisfied	neutral or ...
satisfied	8190	3213
neutral or di...	3003	11570

Correct classified: 19,760

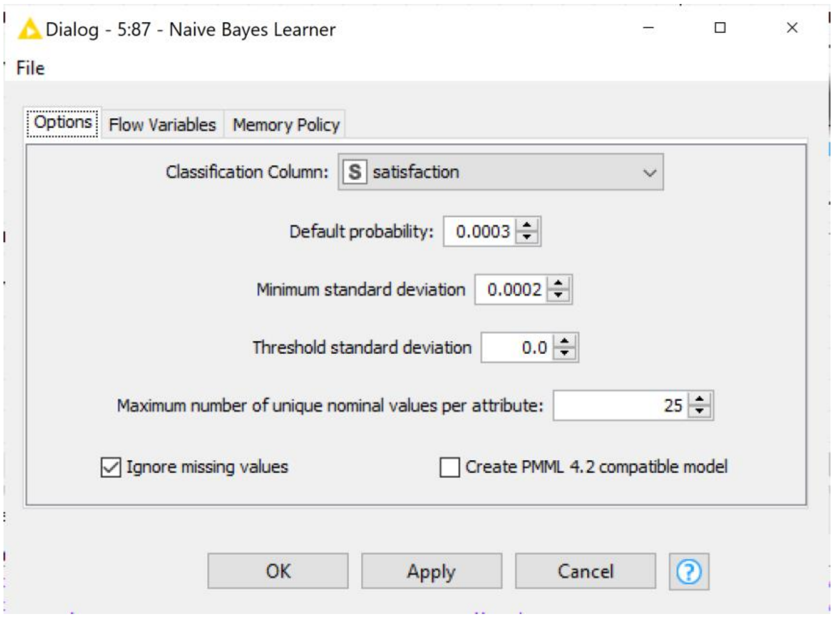
Wrong classified: 6,216

Accuracy: 76.07 %

Error: 23.93 %

Cohen's kappa (κ) 0.513

9.3: Changed the probability to 0.0003, minimum S.D to 0.0002 and increased the nominal values per attribute to 25. Even if we increased it more the accuracy did not change.



Dialog - 5:87 - Naive Bayes Learner

File

Options | Flow Variables | Memory Policy

Classification Column: S satisfaction

Default probability: 0.0003

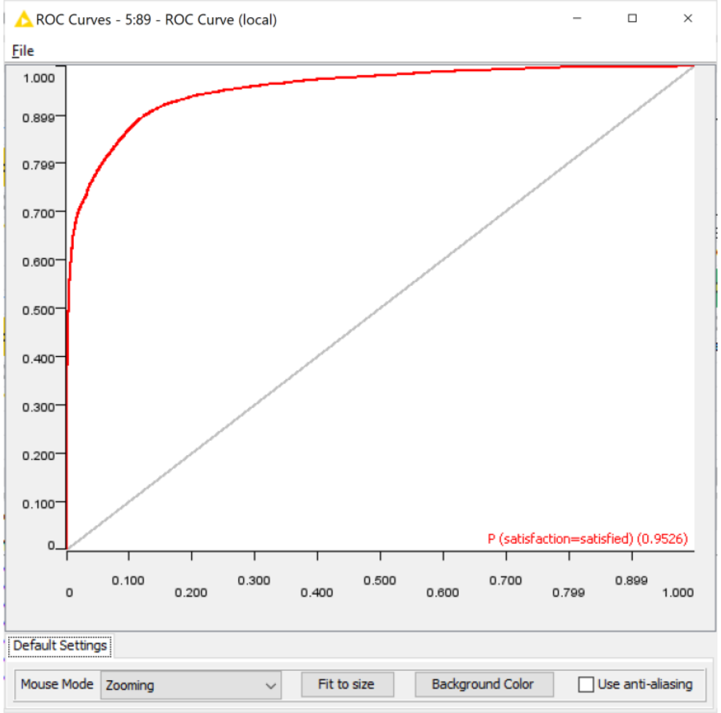
Minimum standard deviation 0.0002

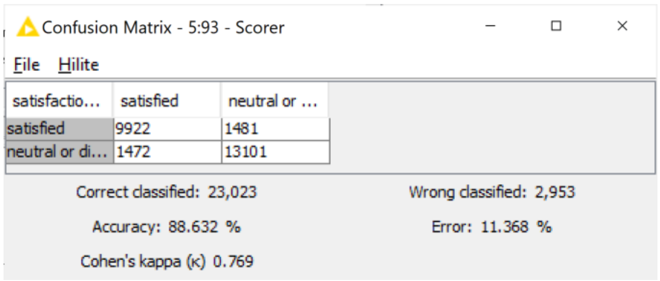
Threshold standard deviation 0.0

Maximum number of unique nominal values per attribute: 25

☑ Ignore missing values ☐ Create PMML 4.2 compatible model

ROC And Scorer:





Confusion Matrix - 5:93 - Scorer

File Hilite

satisfactio...	satisfied	neutral or ...
satisfied	9922	1481
neutral or di...	1472	13101

Correct classified: 23,023 Wrong classified: 2,953

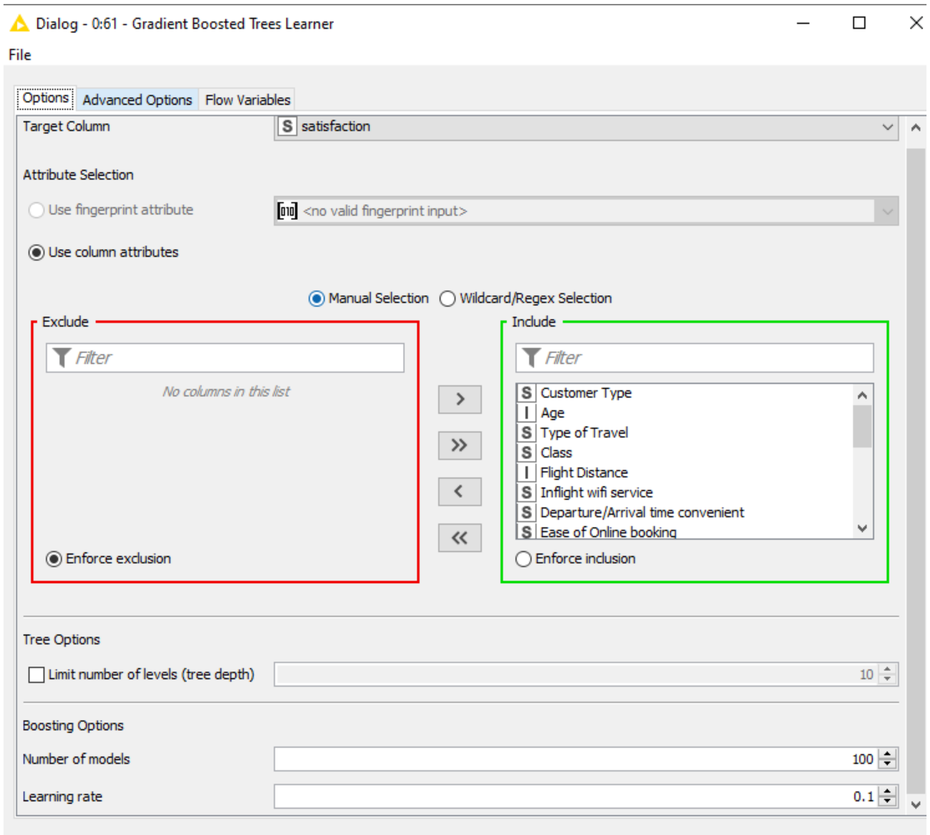
Accuracy: 88.632 % Error: 11.368 %

Cohen's kappa (κ) 0.769

Attempt 10:

Changing configurations in Gradient Boosting Learner:

10.1: Unticked the option of limiting the number of levels(tree depth)



Target Column: S satisfaction

Attribute Selection

○ Use fingerprint attribute <no valid fingerprint input>

● Use column attributes

● Manual Selection ○ Wildcard/Regex Selection

Exclude

Filter

No columns in this list

● Enforce exclusion

Include

Filter

S Customer Type
I Age
S Type of Travel
S Class
I Flight Distance
S Inflight wifi service
S Departure/Arrival time convenient
S Ease of Online booking

○ Enforce inclusion

Tree Options

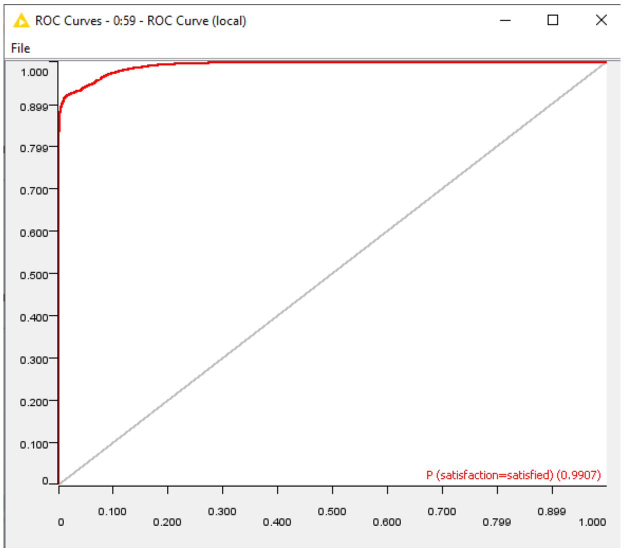
□ Limit number of levels (tree depth) 10

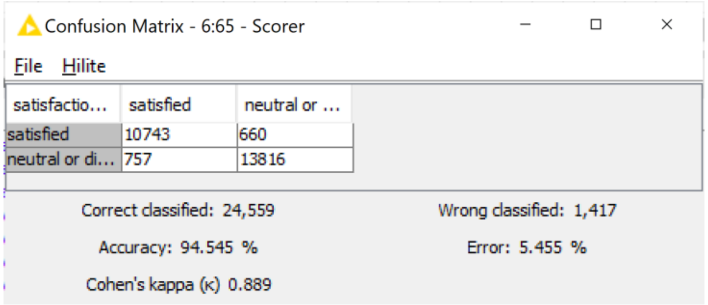
Boosting Options

Number of models 100

Learning rate 0.1

ROC And Scorer:





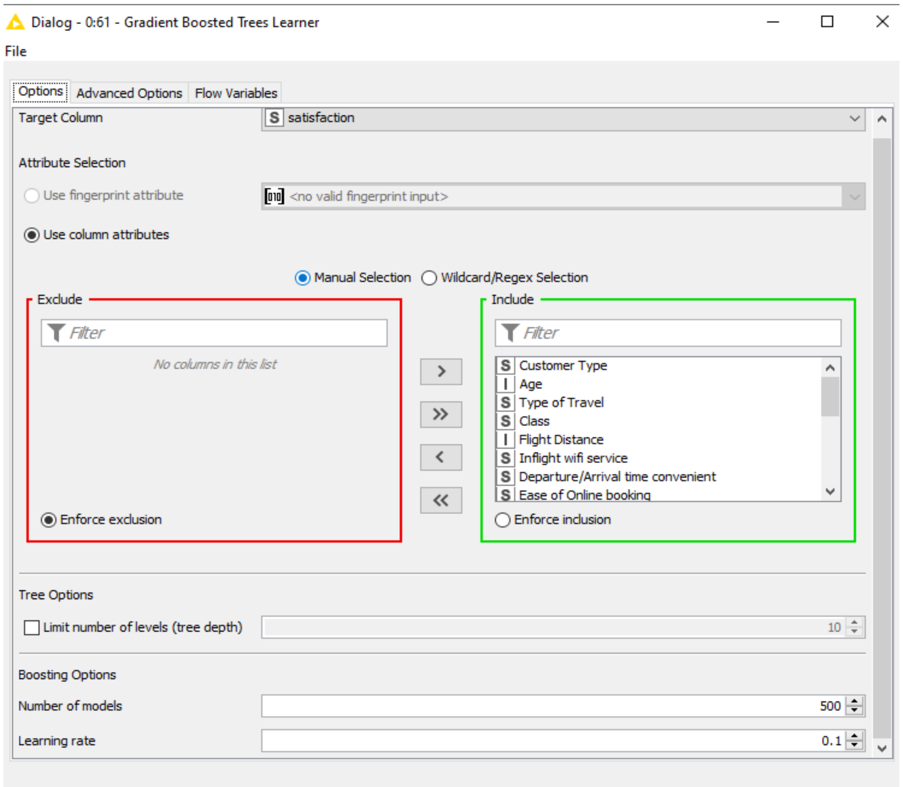
satisfactio...	satisfied	neutral or ...
satisfied	10743	660
neutral or di...	757	13816

Correct classified: 24,559 Wrong classified: 1,417

Accuracy: 94.545 % Error: 5.455 %

Cohen's kappa (κ) 0.889

10.2: Increased the number of models to 500.



Dialog - 0:61 - Gradient Boosted Trees Learner

File

Options | Advanced Options | Flow Variables

Target Column: S satisfaction

Attribute Selection

○ Use fingerprint attribute: <no valid fingerprint input>

● Use column attributes

● Manual Selection ○ Wildcard/Regex Selection

Exclude: Filter

No columns in this list

● Enforce exclusion

Include: Filter

- S Customer Type
- I Age
- S Type of Travel
- S Class
- I Flight Distance
- S Inflight wifi service
- S Departure/Arrival time convenient
- S Ease of Online booking

○ Enforce inclusion

Tree Options

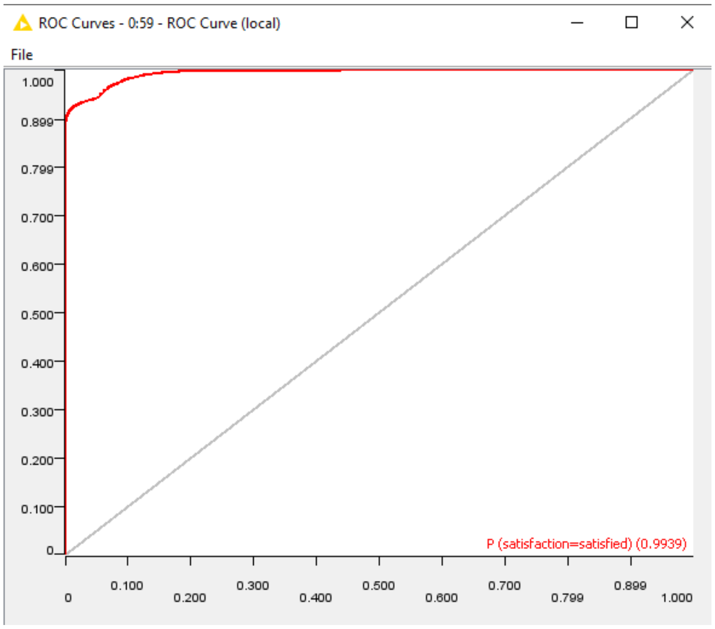
□ Limit number of levels (tree depth): 10

Boosting Options

Number of models: 500

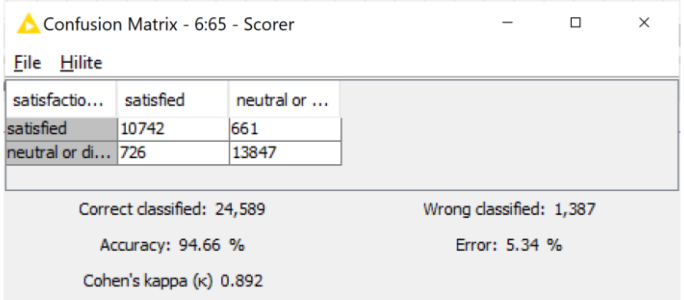
Learning rate: 0.1

ROC And Scorer:



ROC Curves - 0:59 - ROC Curve (local)

File



Confusion Matrix - 6:65 - Scorer

File Hilite

satisfactio...	satisfied	neutral or ...
satisfied	10742	661
neutral or di...	726	13847

Correct classified: 24,589

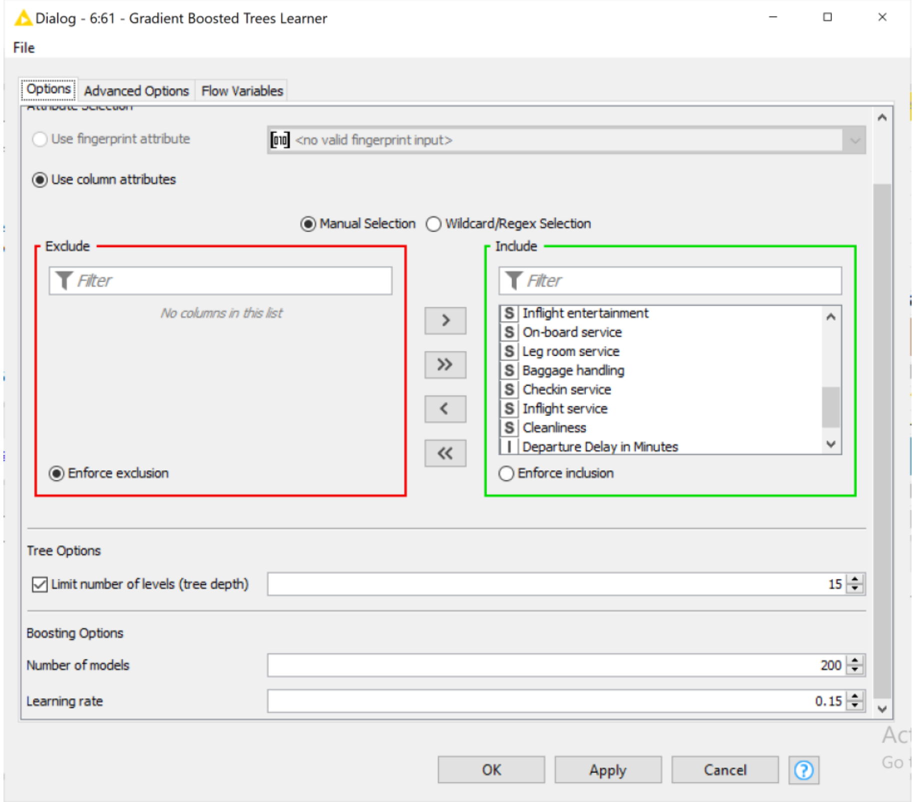
Wrong classified: 1,387

Accuracy: 94.66 %

Error: 5.34 %

Cohen's kappa (κ) 0.892

10.3: Limited the number of levels to 15 with 200 number of models and a learning rate of 0.15



Dialog - 6:61 - Gradient Boosted Trees Learner

File

Options | Advanced Options | Flow Variables

Attribute Selection

○ Use fingerprint attribute <no valid fingerprint input>

● Use column attributes

● Manual Selection ○ Wildcard/Regex Selection

Exclude

No columns in this list

● Enforce exclusion

Include

- S Inflight entertainment
- S On-board service
- S Leg room service
- S Baggage handling
- S Checkin service
- S Inflight service
- S Cleanliness
- I Departure Delay in Minutes

○ Enforce inclusion

Tree Options

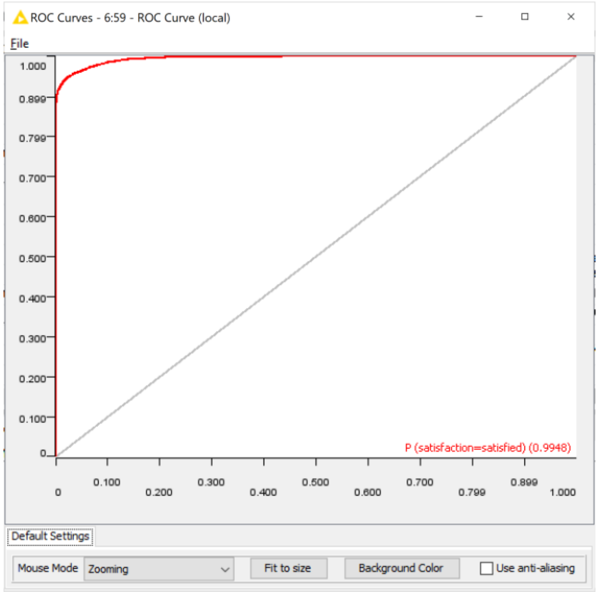
☑ Limit number of levels (tree depth) 15

Boosting Options

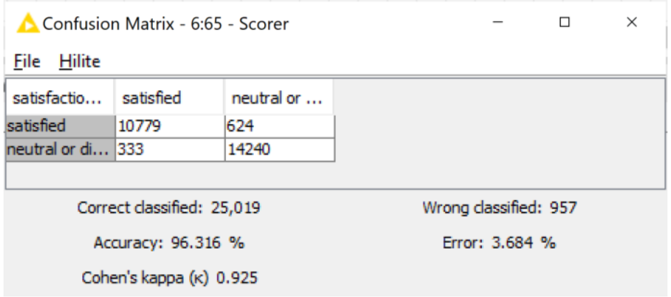
Number of models 200

Learning rate 0.15

ROC And Scorer:



ROC Curves - 6:59 - ROC Curve (local)



Confusion Matrix - 6:65 - Scorer

satisfactio...	satisfied	neutral or ...
satisfied	10779	624
neutral or di...	333	14240

Correct classified: 25,019 Wrong classified: 957

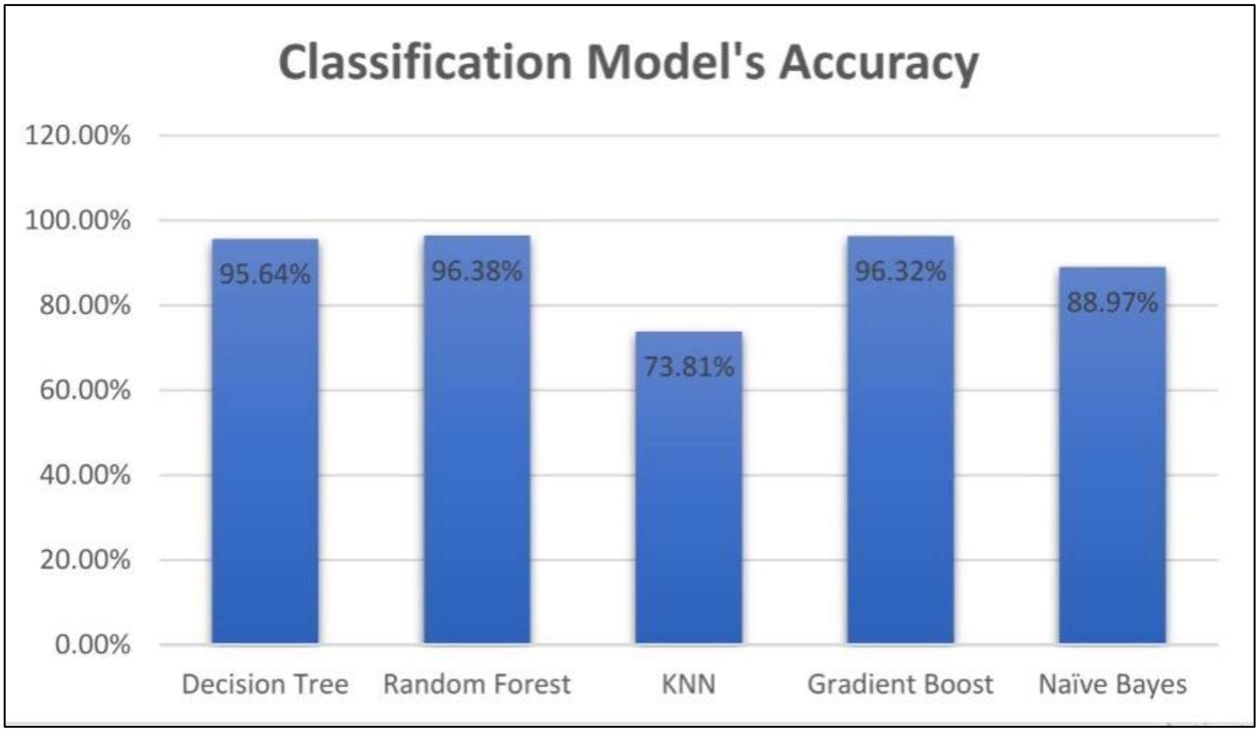
Accuracy: 96.316 % Error: 3.684 %

Cohen's kappa (κ) 0.925

Findings

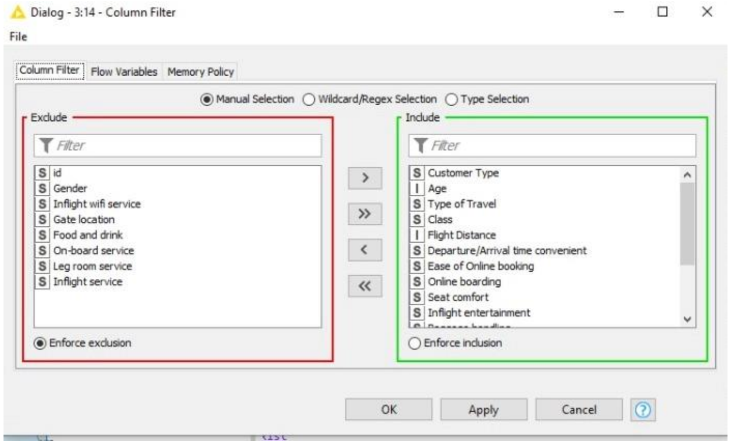
Finding Best Algorithm

We used various classification model for airlines to identify critical bottleneck to raise passenger satisfaction. As seen in the bar chart below, random forest gave us the best accuracy in attempt 2.



Factors Effecting Passenger Satisfaction

To figure out which features effect our accuracy the most, we took our best model and tried excluding certain features using column filter.



Dialog - 3:14 - Column Filter

● Manual Selection ○ Wildcard/Regex Selection ○ Type Selection

Exclude

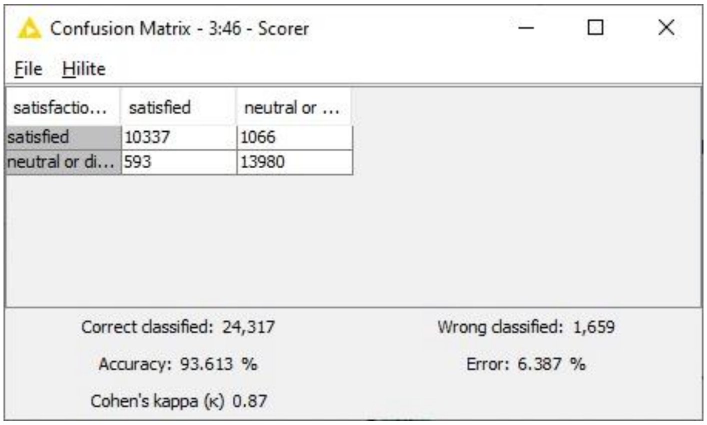
- S id
- S Gender
- S Inflight wifi service
- S Gate location
- S Food and drink
- S On-board service
- S Leg room service
- S Inflight service

● Enforce exclusion

Include

- S Customer Type
- I Age
- S Type of Travel
- S Class
- I Flight Distance
- S Departure/Arrival time convenient
- S Ease of Online booking
- S Online boarding
- S Seat comfort
- S Inflight entertainment

○ Enforce inclusion



Confusion Matrix - 3:46 - Scorer

satisfactio...	satisfied	neutral or ...
satisfied	10337	1066
neutral or di...	593	13980

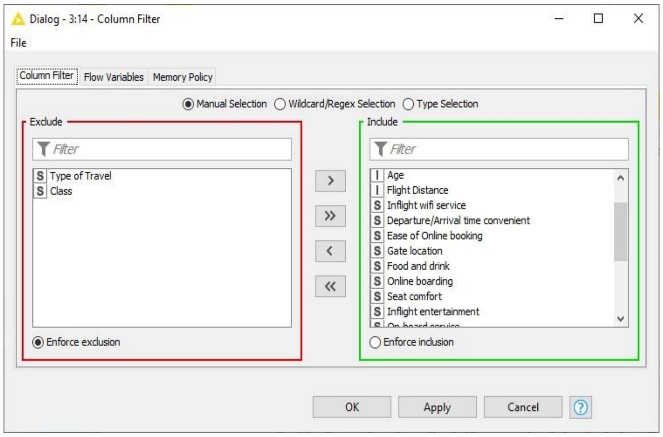
Correct classified: 24,317

Wrong classified: 1,659

Accuracy: 93.613 %

Error: 6.387 %

Cohen's kappa (κ) 0.87



Dialog - 3:14 - Column Filter

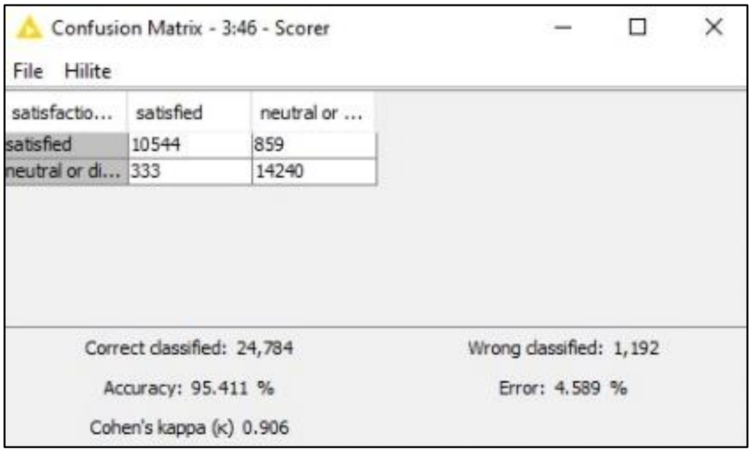
● Manual Selection ○ Wildcard/Regex Selection ○ Type Selection

Exclude
- S Type of Travel
- S Class

● Enforce exclusion

Include
- I Age
- I Flight Distance
- S Inflight wifi service
- S Departure/Arrival time convenient
- S Ease of Online booking
- S Gate location
- S Food and drink
- S Online boarding
- S Seat comfort
- S Inflight entertainment
- S On-board service

○ Enforce inclusion



Confusion Matrix - 3:46 - Scorer

satisfactio...	satisfied	neutral or ...
satisfied	10544	859
neutral or di...	333	14240

Correct classified: 24,784 Wrong classified: 1,192

Accuracy: 95.411 % Error: 4.589 %

Cohen's kappa (κ) 0.906

From some of the simulations, it was seen that airlines should focus on improving the Services provided in the flight. For instance, airlines could develop better software to allow easier access to inflight wi-fi, food & etc. Excluding the “class” feature we observed how it effected the overall satisfaction of the passenger. Business class passenger are more satisfied with the flight so the airlines should initiate cheaper packages for business class.

Concluding remarks

The dramatic drop in demand for passenger air transport due to the COVID-19 pandemic has caused tremendous decline in the viability in the aviation industry. As a result, in order to resurrect the industry in the face of the crisis, it is critical to understand client pain points and increase their satisfaction with the services supplied. The industry should involve a third party market research company to provide them with unique, fresh, and unbiased recommendations to continuously improve their services attracting more people to choose their airlines for their trip.

There may be various options for gathering client feedback. The most straightforward and traditional method is to use the customer feedback form that is provided throughout the trip. However, the majority of passengers are uninterested in completing feedback forms. Other methods for collecting client input include the airline's internet website or online mobile applications. After the travel, the passenger can be sent an email with a link asking for feedback.

Finally, we hope that the model will provide a reference and be utilized for business value to the airlines, and will help them improve their services to allure more passengers towards them.