Introduction to Data Analytics

Assignment 3

Kaung Khant Kyaw
14447837

Data Mining Problem

This assignment mainly focuses on data training and model optimization. This assignment required me to predict previous data by using and modifying classifiers as a data scientist in a company. I used KNIME software as a tool to see the dataset and use models in the tool to finish this assignment.

Input

The dataset given is about National Basketball Association players and their stats to train a model that can predict if they will be playing in the next 5 years future.

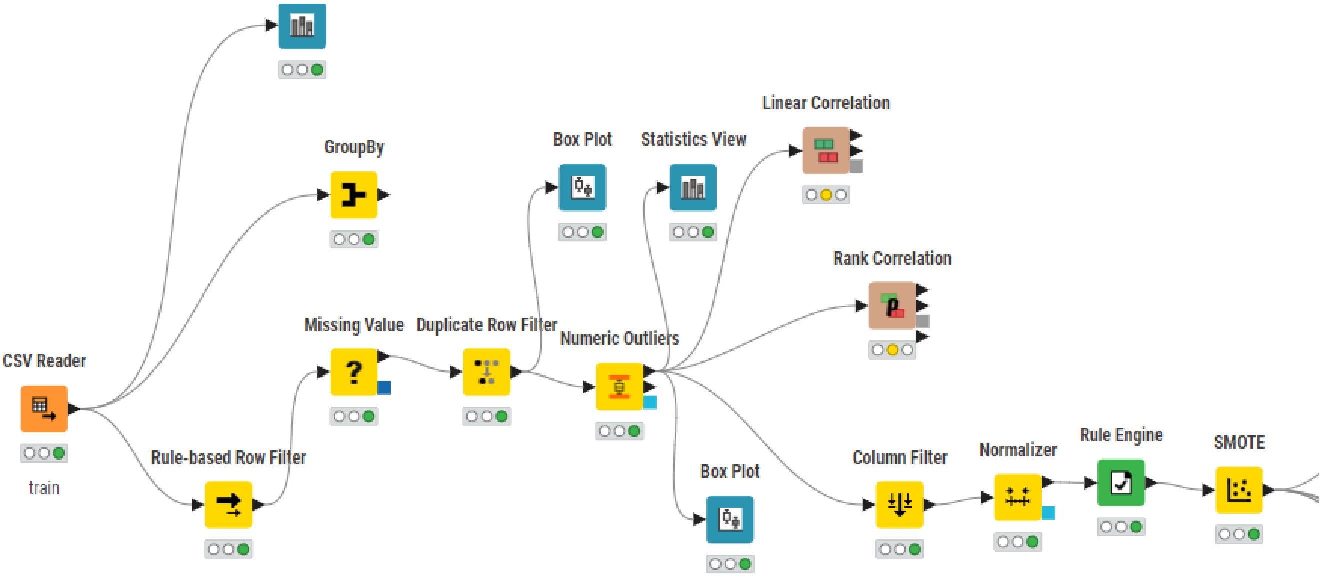
Output

The problem is that the data given includes some negative values as well as outliers, and since negative values are unstable to predict a model, we would have to pre-process this data first before training a model to predict. We also had to remove the outliers too.

Goal

The goal of this assessment is to demonstrate the ability to preprocess a dataset, train a predictive model, and optimize the classifier for predicting future outcomes. Identify and handle negative values in the dataset, ensuring the data is suitable for training a predictive model. Detect and remove outliers that could skew the model's performance. Utilize KNIME software to train a model on the preprocessed dataset. Experiment with various classifiers to determine the best fit for the data. Modify and fine-tune the selected classifier(s) to improve prediction accuracy. Evaluate the performance of the model using appropriate metrics. Use the trained and optimized model to predict whether NBA players will be playing in the next 5 years based on their stats.

Data Pre-Processing

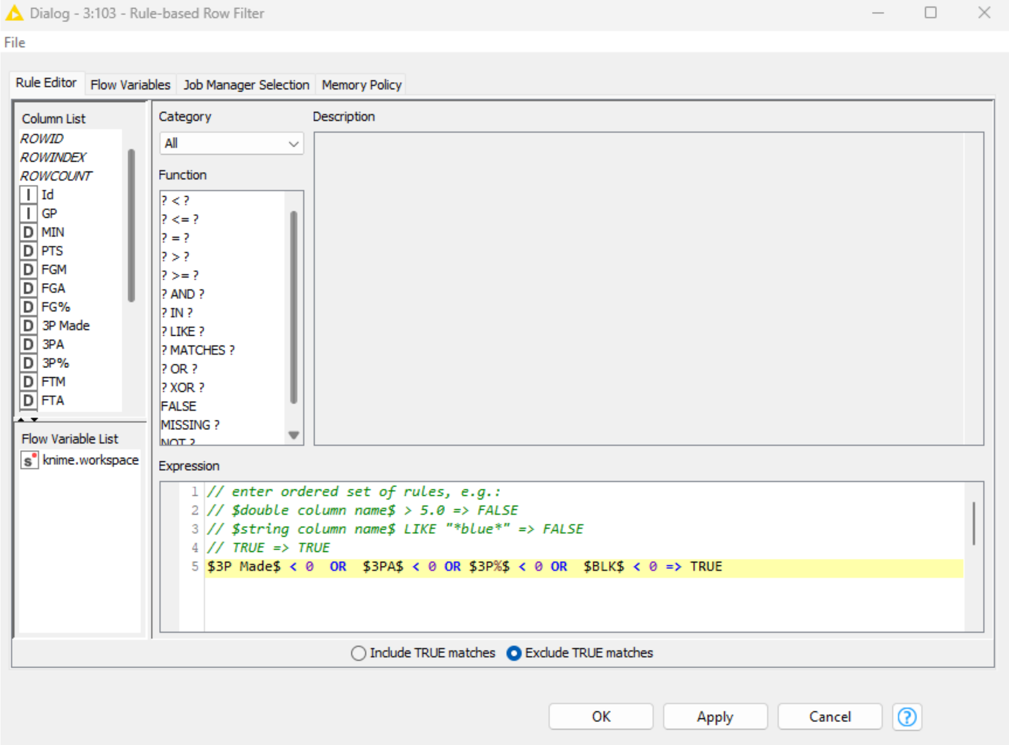


CSV reader

Input the dataset given as a csv file type which is train.csv.

Rule-based row Filter

Filter out the negative values since they are unstable when building a model.

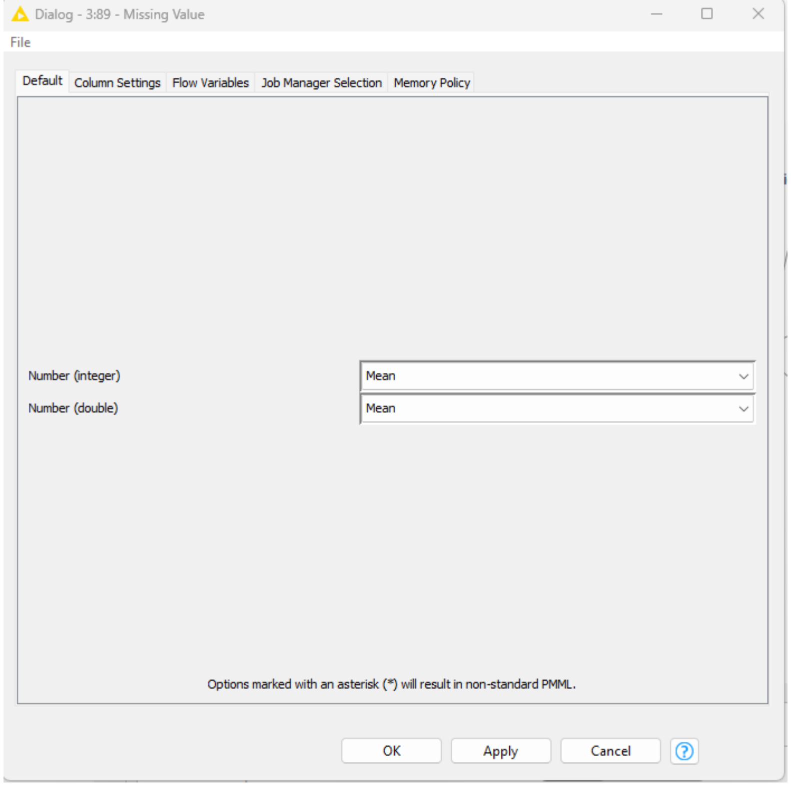


```
// enter ordered set of rules, e.g.:
// $double column name$ > 5.0 => FALSE
// $string column name$ LIKE "*blue*" => FALSE
// TRUE => TRUE
$3P Made$ < 0  OR  $3PA$ < 0 OR $3P%$ < 0 OR  $BLK$ < 0 => TRUE
```

○ Include TRUE matches ● Exclude TRUE matches

Missing Value

Even though, there were not any missing values, I set the missing values to the mean values just in case there is one and I could not find it.



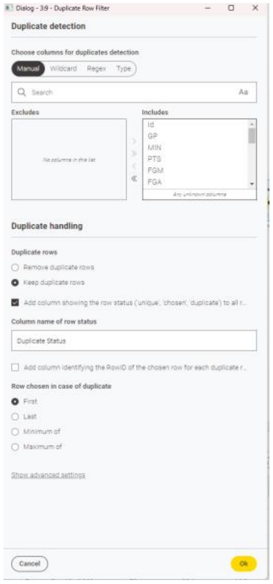
Number (integer): Mean

Number (double): Mean

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

Duplicate row filter

This node is for checking if there were any duplicate rows in the dataset, luckily there were not any, so I did not have to deal with it.



Duplicate detection

Choose columns for duplicates detection

Excludes: No columns in this list

Includes: Id, GP, MIN, PTS, FGM, FGA, Any unknown columns

Duplicate handling

Duplicate rows

○ Remove duplicate rows

● Keep duplicate rows

☑ Add column showing the row status ('unique', 'chosen', 'duplicate') to all r...

Column name of row status: Duplicate Status

□ Add column identifying the RowID of the chosen row for each duplicate r...

Row chosen in case of duplicate

● First

○ Last

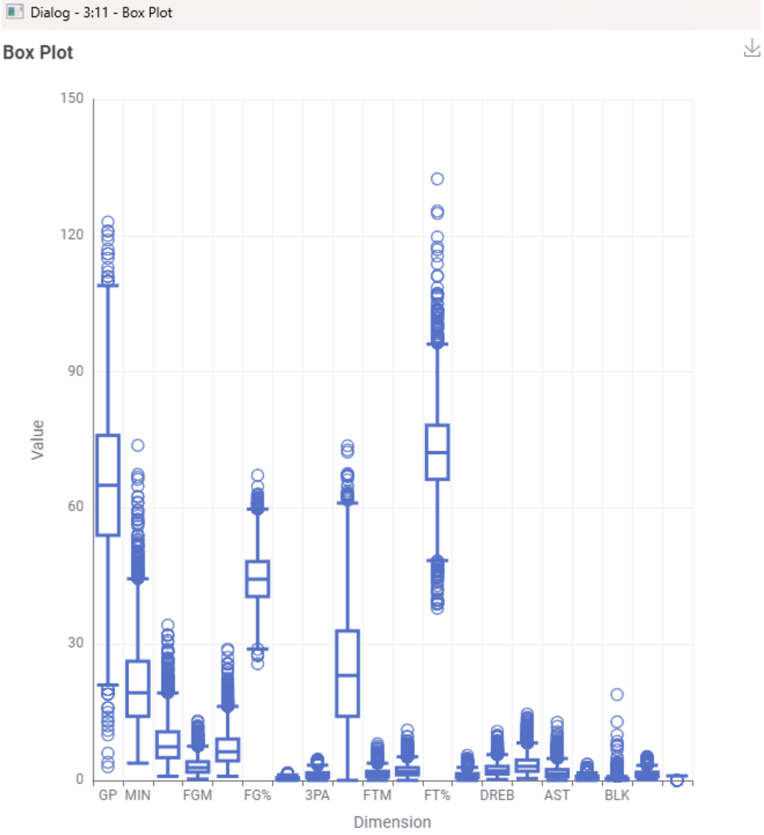
○ Minimum of

○ Maximum of

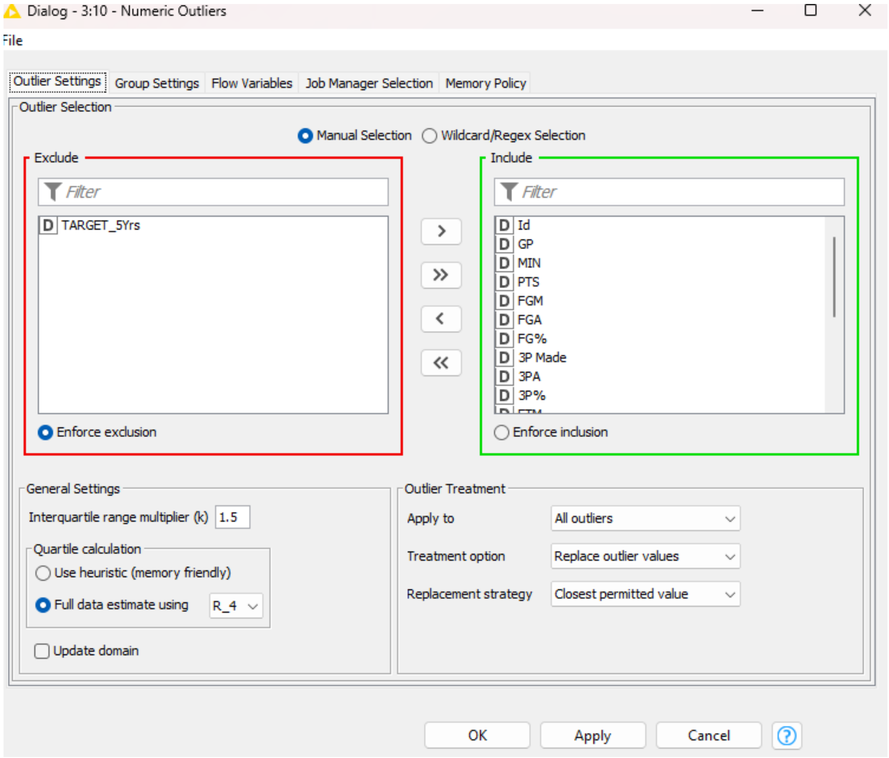
Numeric Outliers

I also used box plot to check the outliers before and after using numeric outliers which is used to replace the outlier values.

Box plot before using this node



Numeric Outliers



Outlier Selection

● Manual Selection ○ Wildcard/Regex Selection

Exclude: D TARGET_5Yrs

● Enforce exclusion

Include: D Id, D GP, D MIN, D PTS, D FGM, D FGA, D FG%, D 3P Made, D 3PA, D 3P%

○ Enforce inclusion

General Settings

Interquartile range multiplier (k) 1.5

Quartile calculation

○ Use heuristic (memory friendly)

● Full data estimate using R_4

□ Update domain

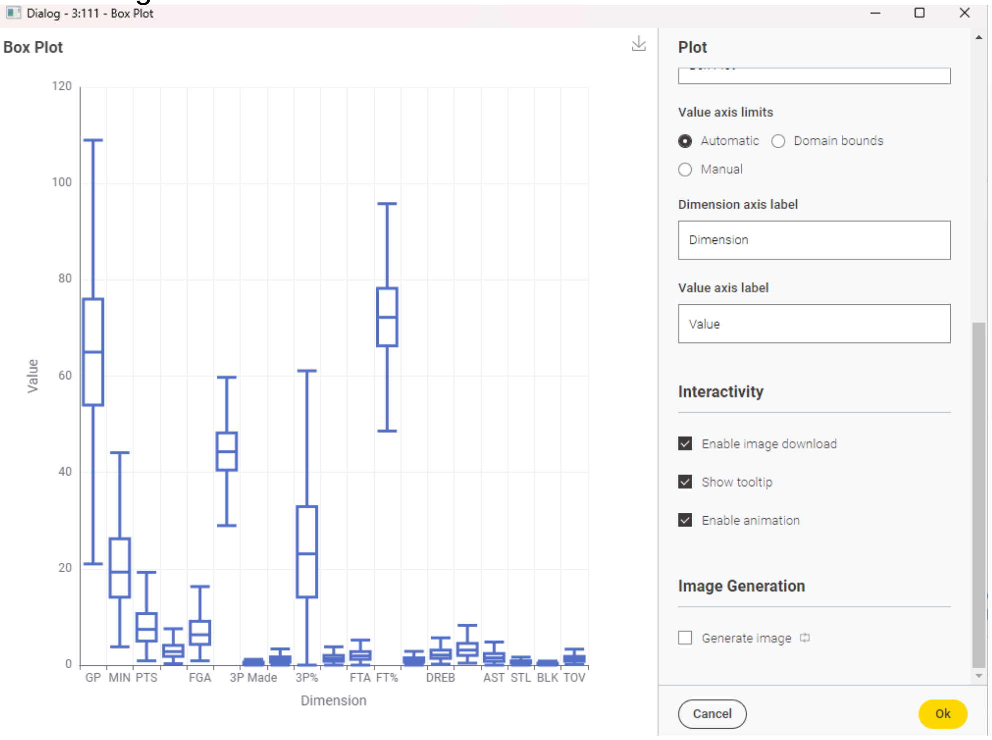
Outlier Treatment

Apply to: All outliers

Treatment option: Replace outlier values

Replacement strategy: Closest permitted value

After using Numeric Outliers



Dialog - 3:111 - Box Plot

Box Plot

Plot

Value axis limits

● Automatic ○ Domain bounds

○ Manual

Dimension axis label

Dimension

Value axis label

Value

Interactivity

☑ Enable image download

☑ Show tooltip

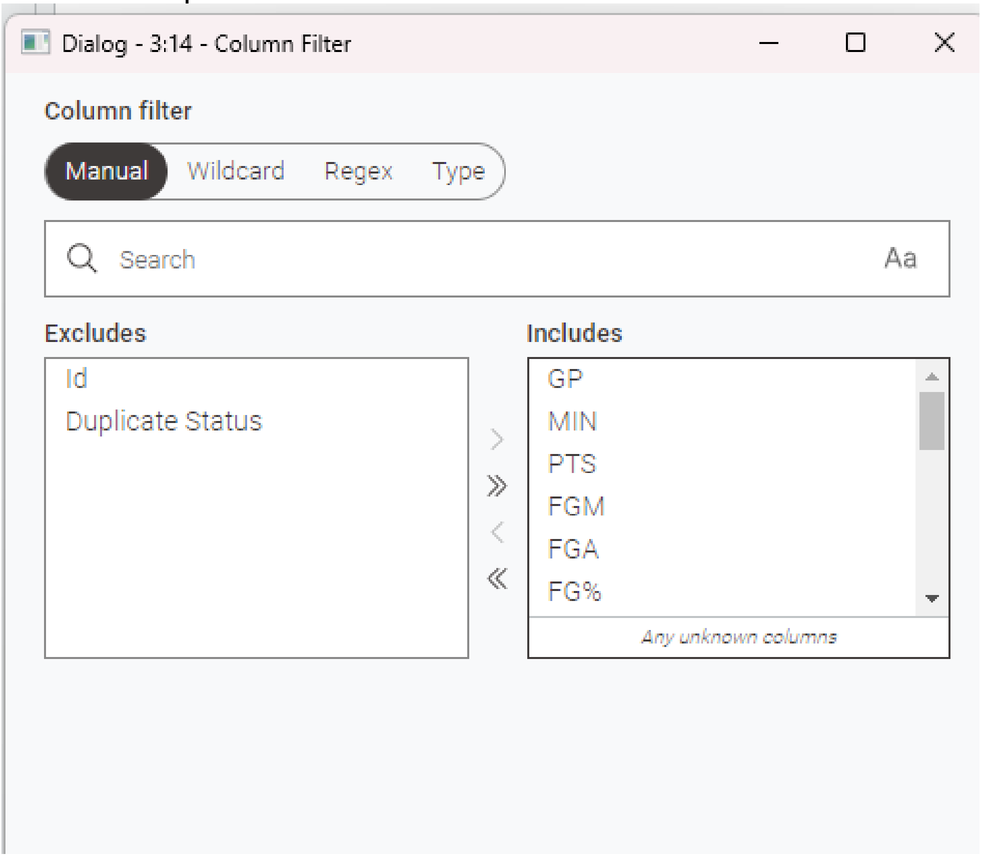
☑ Enable animation

Image Generation

□ Generate image

Column Filter

In the column filter, I filtered out the player's ID. And the new duplicate status that is used to check if the row is duplicated or not.



Dialog - 3:14 - Column Filter

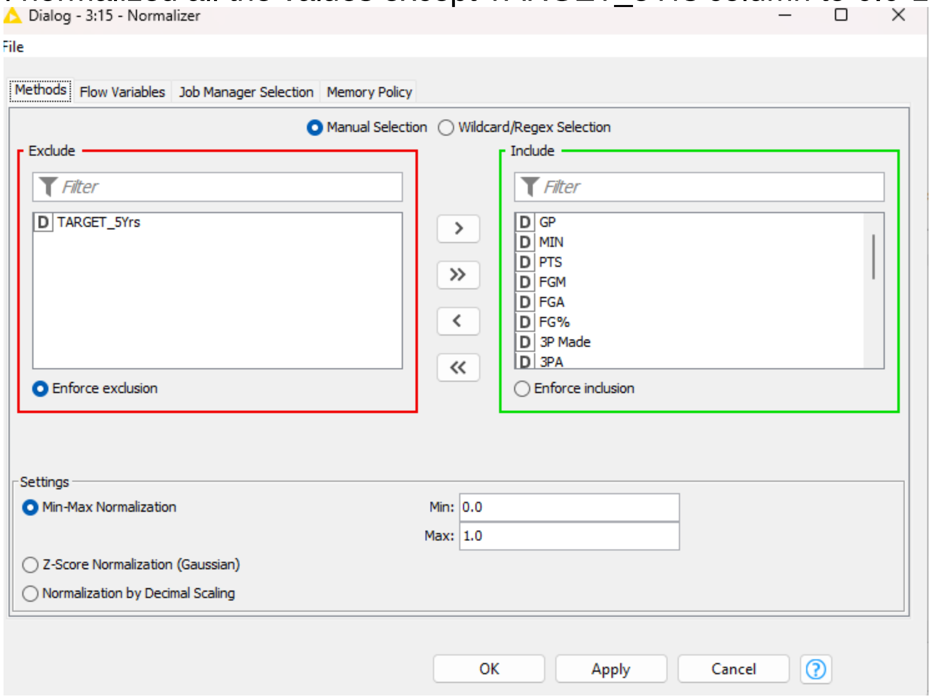
Column filter

Excludes	Includes
Id	GP
Duplicate Status	MIN
	PTS
	FGM
	FGA
	FG%

Any unknown columns

Normalizer

I normalized all the values except TARGET_5Yrs column to 0.0-1.0 ratio.



● Manual Selection ○ Wildcard/Regex Selection

Exclude: D TARGET_5Yrs

● Enforce exclusion

Include: D GP, D MIN, D PTS, D FGM, D FGA, D FG%, D 3P Made, D 3PA

○ Enforce inclusion

Settings

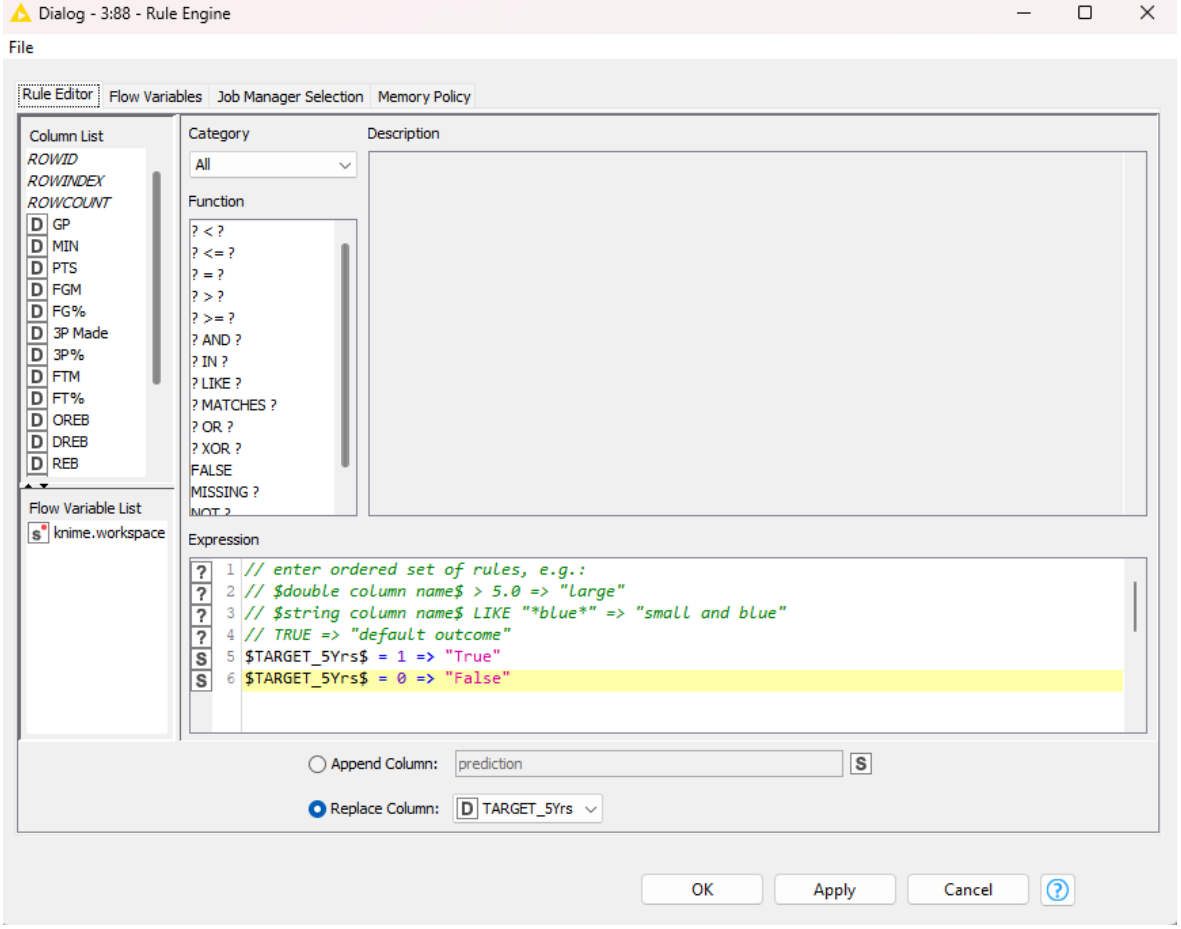
● Min-Max Normalization Min: 0.0 Max: 1.0

○ Z-Score Normalization (Gaussian)

○ Normalization by Decimal Scaling

Rule engine

I used rule-engine to change my TARGET_5Yrs values (1,0) to (True,False) and nominal data type for the client to understand clearly and nominal for future modelling.



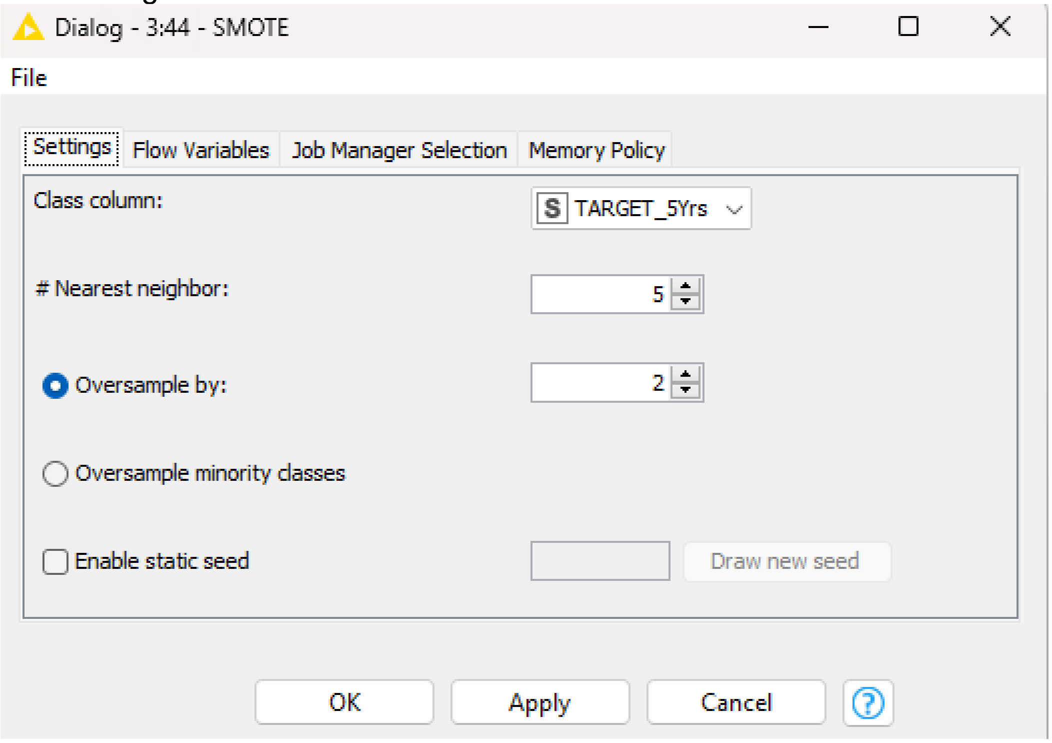
```
// enter ordered set of rules, e.g.:
// $double column name$ > 5.0 => "large"
// $string column name$ LIKE "*blue*" => "small and blue"
// TRUE => "default outcome"
$TARGET_5Yrs$ = 1 => "True"
$TARGET_5Yrs$ = 0 => "False"
```

○ Append Column: prediction

● Replace Column: D TARGET_5Yrs

Smote

SMOTE (Synthetic Minority Over-sampling Technique) is used to oversample the input data to enrich the training data.



Class column: S TARGET_5Yrs

Nearest neighbor: 5

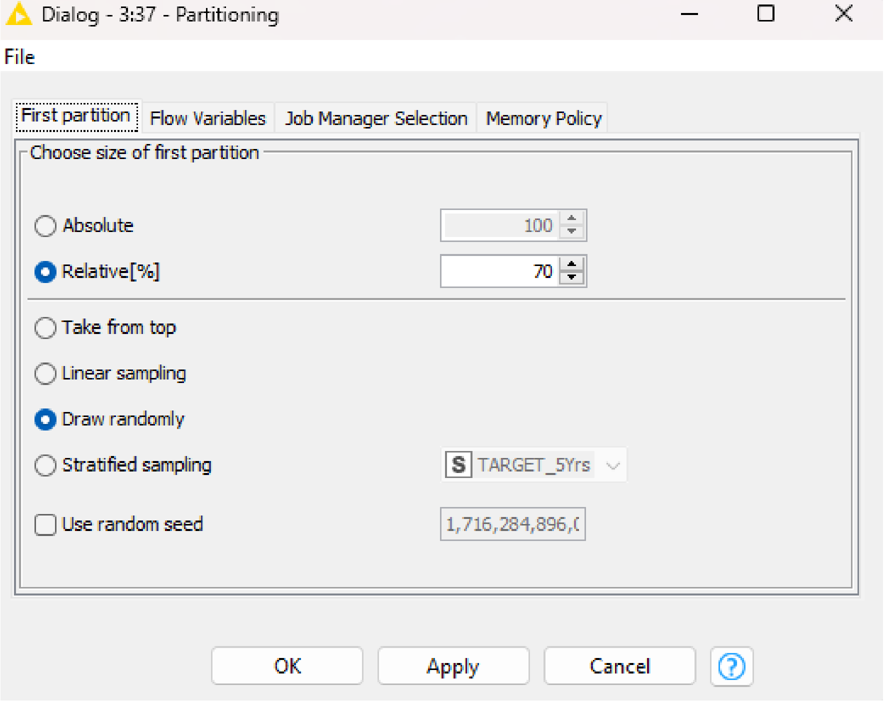
● Oversample by: 2

○ Oversample minority classes

□ Enable static seed

Partitioning

All the partitioning done in this model are 70 percent relative.



Choose size of first partition

○ Absolute 100

● Relative[%] 70

○ Take from top

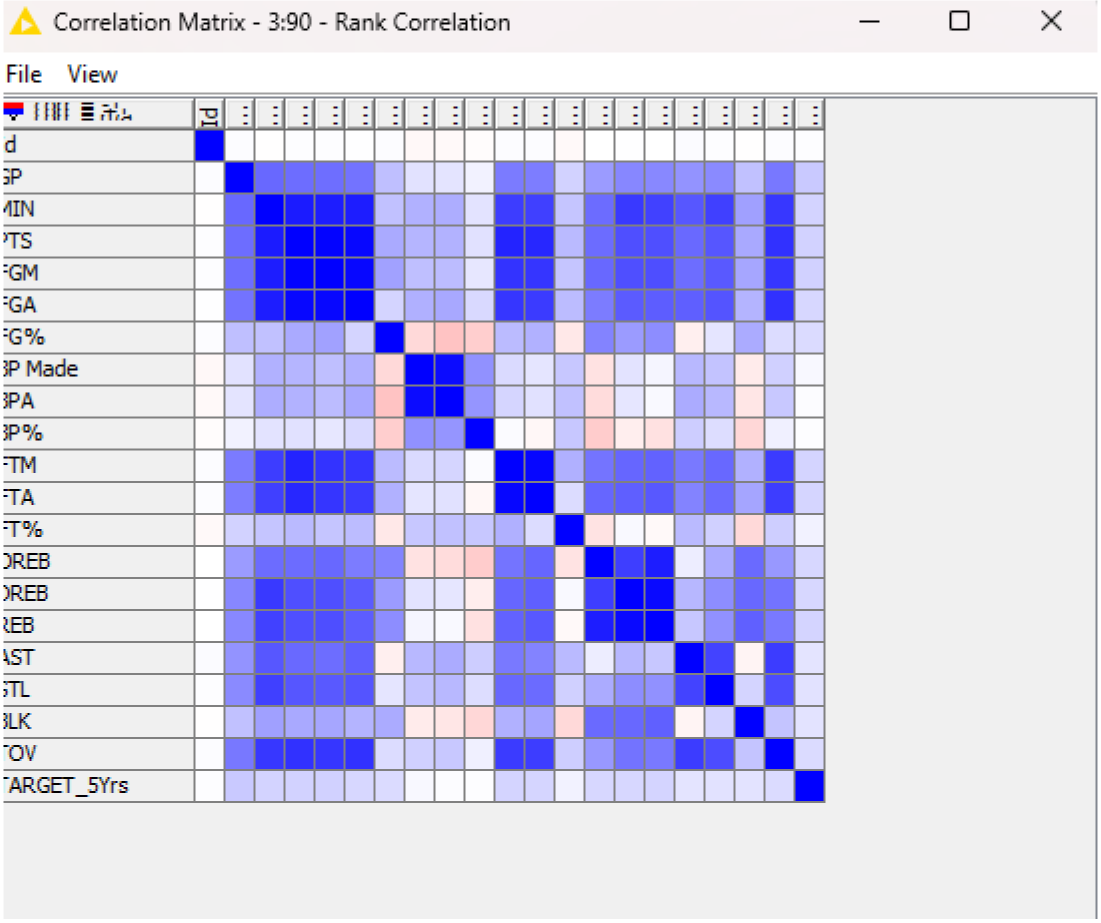
○ Linear sampling

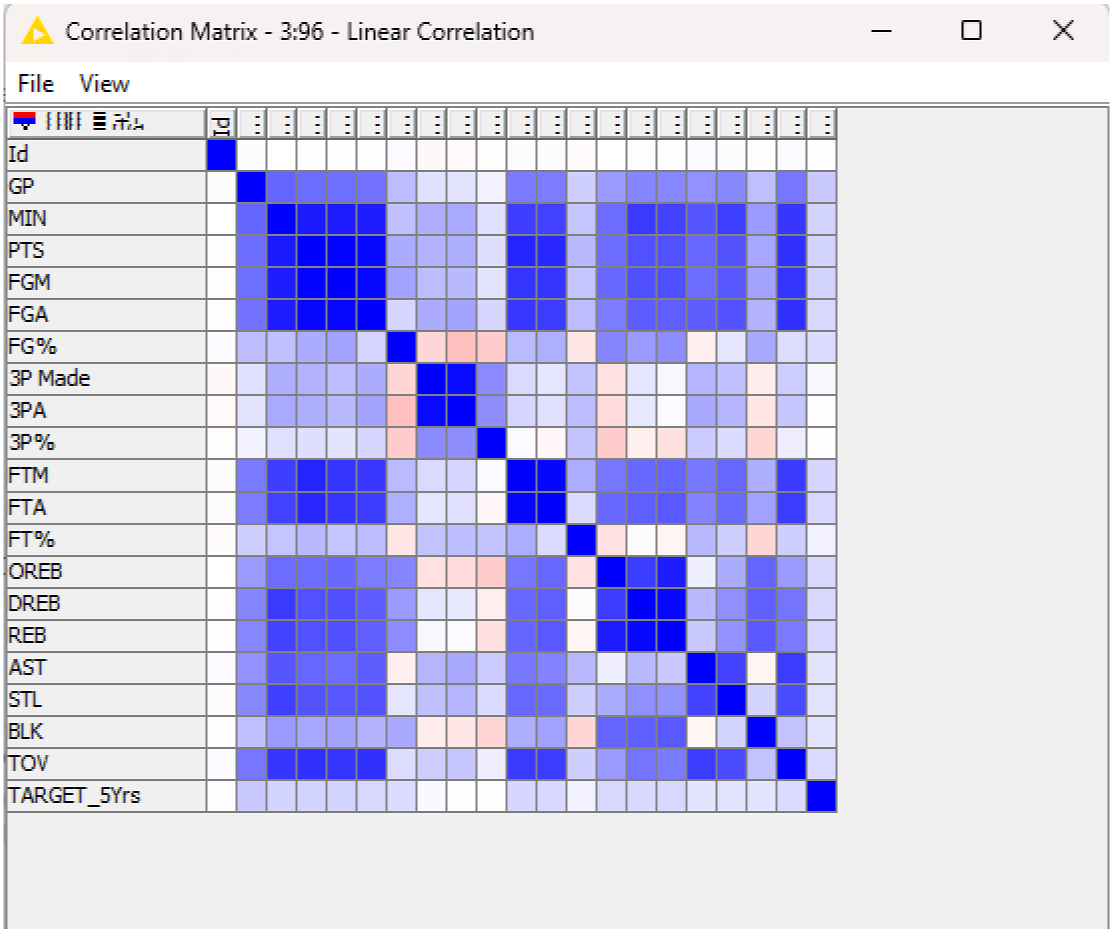
● Draw randomly

○ Stratified sampling S TARGET_5Yrs

□ Use random seed 1,716,284,896,(

Linear and Rank Correlation



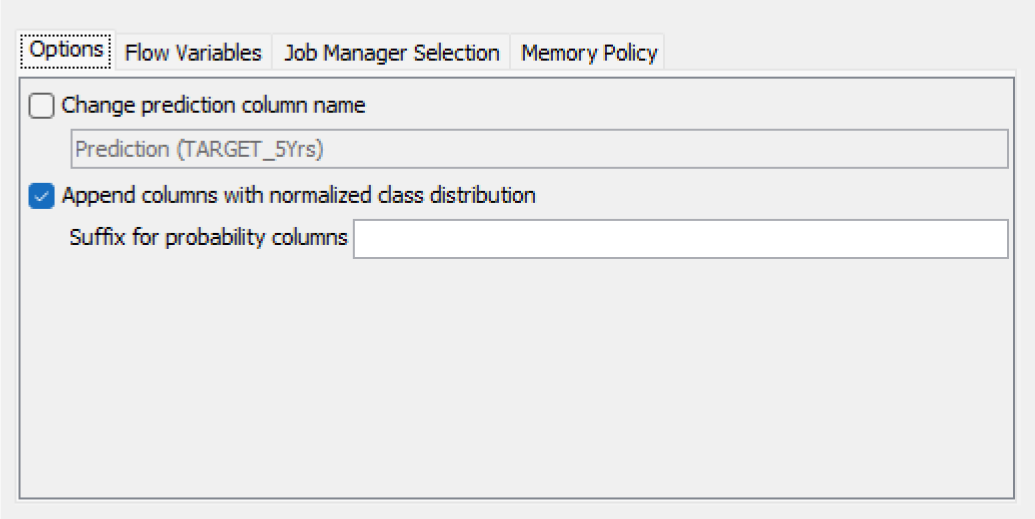


This matrix provides a quick overview of how pairs of variables are related, helping identify important relationships that might warrant further investigation or consideration in predictive models

Classifiers & Explanations

All predictors

I append a new column with a normalized class distribution to be able to see newly predicted columns in the ROC curve. This applies to all predictors used in this assessment.



□ Change prediction column name

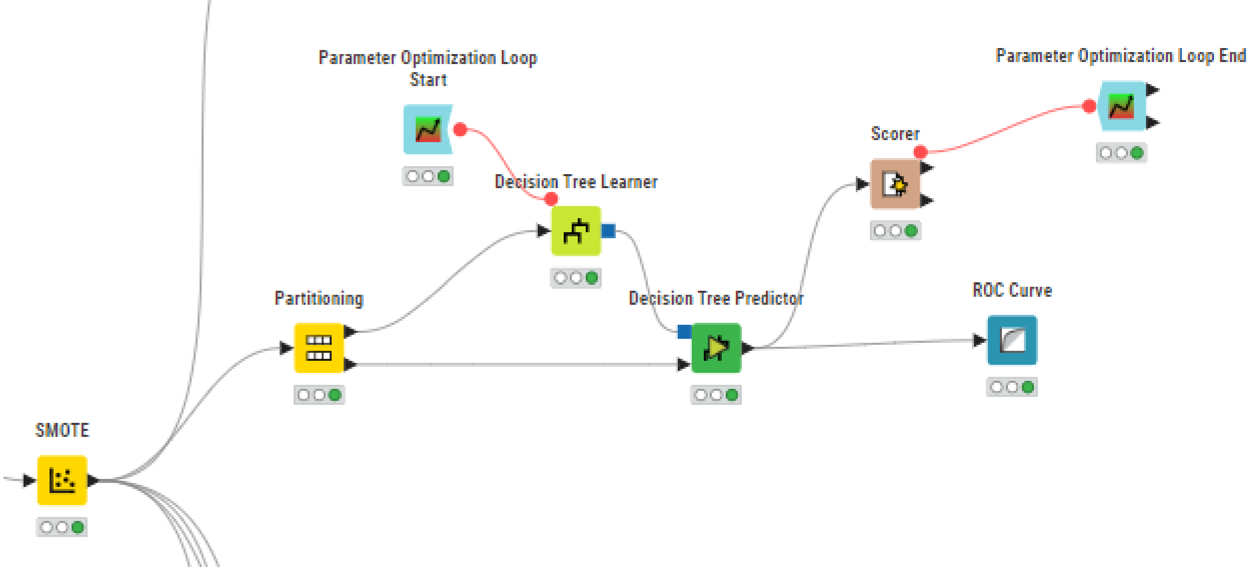
Prediction (TARGET_5Yrs)

☑ Append columns with normalized class distribution

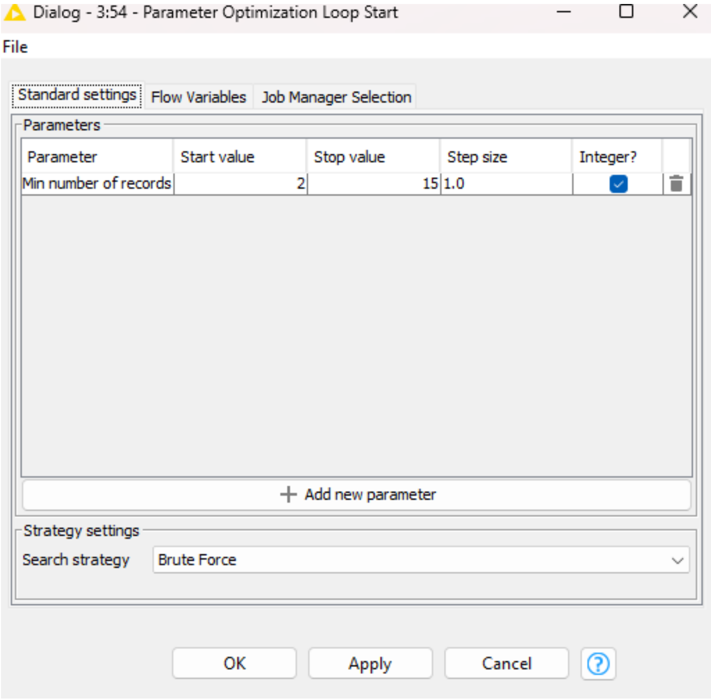
Suffix for probability columns

Decision Tree

This classifier worked and the data is predicted into structure trees. This is effective for noisy data and can handle both numerical and categorical data making it versatile.



I used parameter optimization loop to adjust the parameter for the learner to have the highest accuracy for the score.



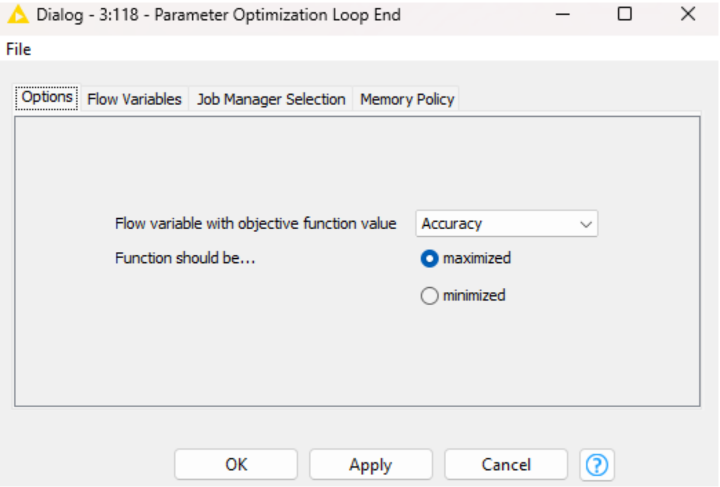
Dialog - 3:54 - Parameter Optimization Loop Start

Parameters

Parameter	Start value	Stop value	Step size	Integer?
Min number of records	2	15	1.0	☑

Strategy settings

Search strategy: Brute Force

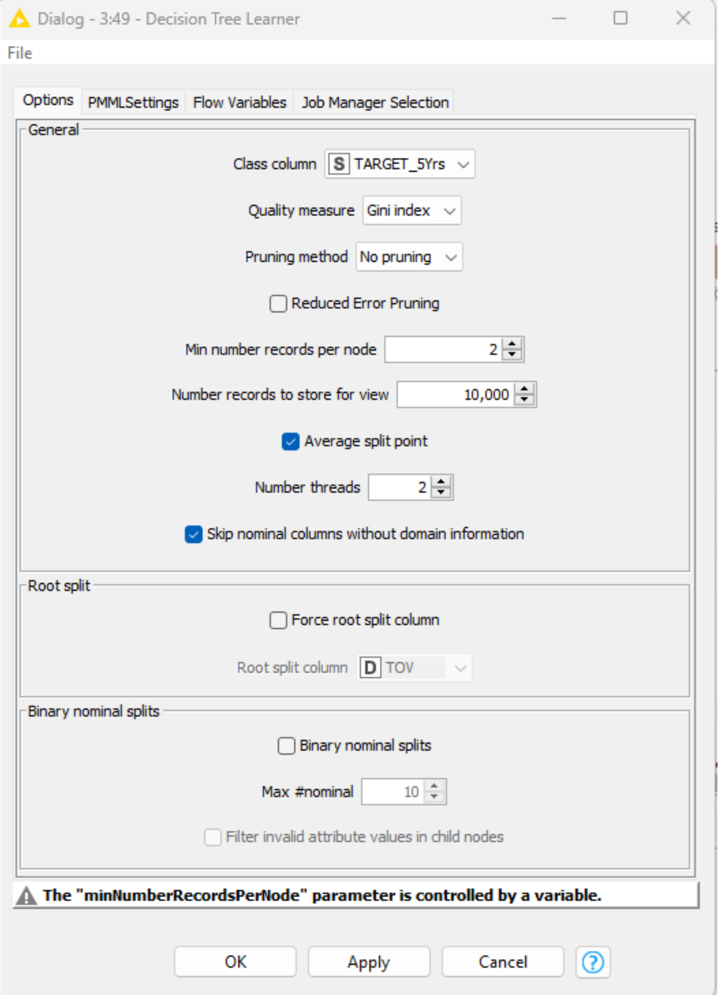


Dialog - 3:118 - Parameter Optimization Loop End

Flow variable with objective function value: Accuracy

Function should be… ● maximized ○ minimized

The loop shows that the accuracy would be the highest if the minimum number of records per node is 2, so I adjusted the node to 2 and resulted in the best accuracy possible for this classifier.



Dialog - 3:49 - Decision Tree Learner

General

Class column: S TARGET_5Yrs

Quality measure: Gini index

Pruning method: No pruning

□ Reduced Error Pruning

Min number records per node: 2

Number records to store for view: 10,000

☑ Average split point

Number threads: 2

☑ Skip nominal columns without domain information

Root split

□ Force root split column

Root split column: D TOV

Binary nominal splits

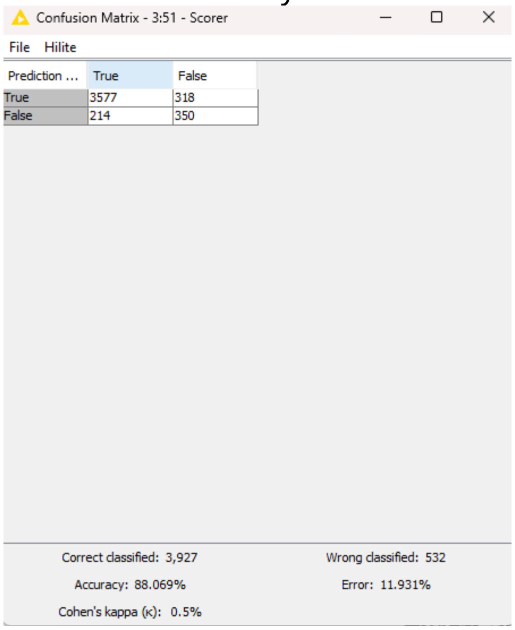
□ Binary nominal splits

Max #nominal: 10

□ Filter invalid attribute values in child nodes

The "minNumberRecordsPerNode" parameter is controlled by a variable.

Here is the accuracy of the classifier



Prediction ...	True	False
True	3577	318
False	214	350

Correct classified: 3,927

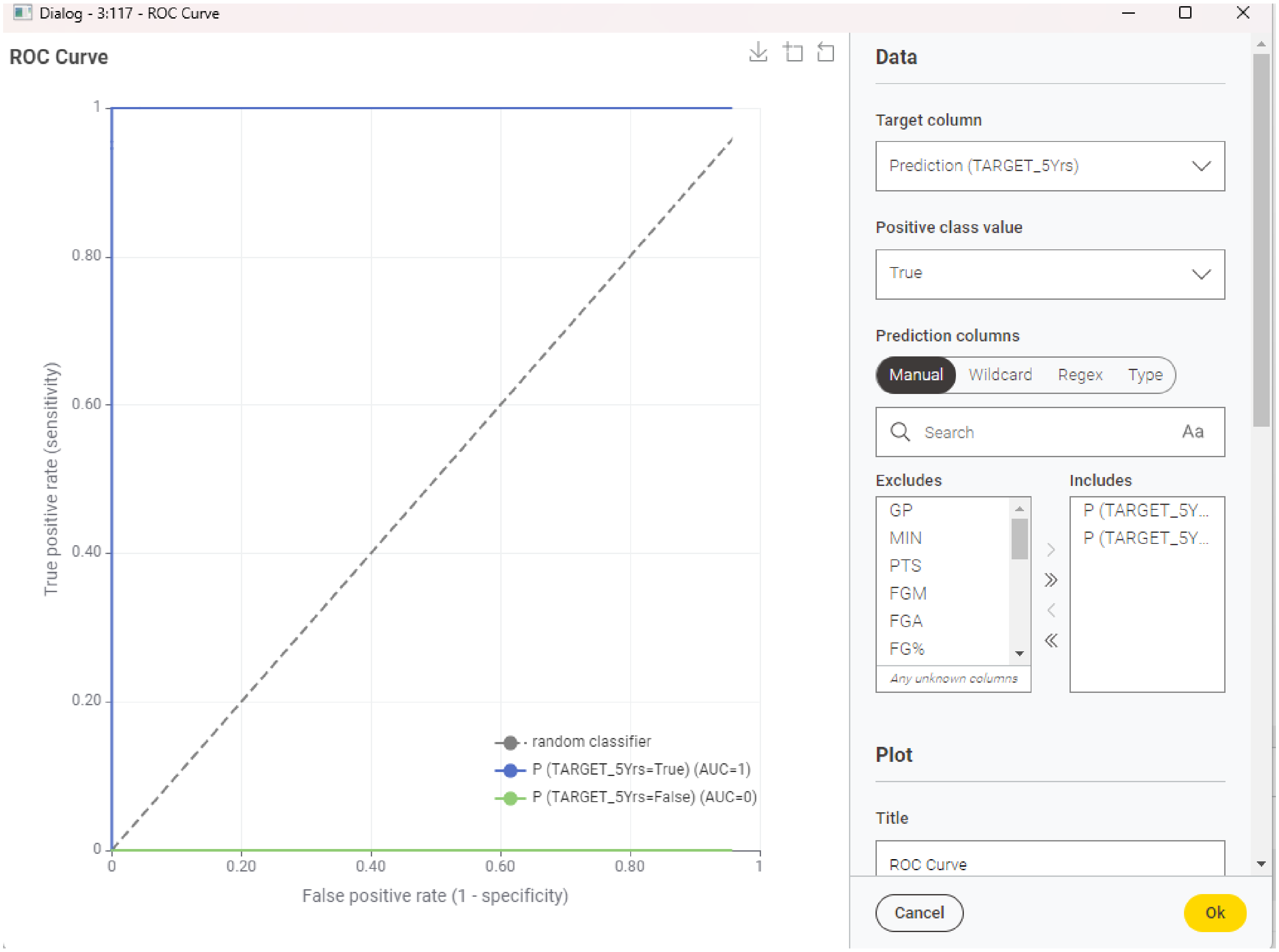
Wrong classified: 532

Accuracy: 88.069%

Error: 11.931%

Cohen's kappa (κ): 0.5%

ROC Curve


Dialog - 3:117 - ROC Curve
ROC Curve
1
0.80
0.60
0.40
0.20
0
True positive rate (sensitivity)
0
0.20
0.40
0.60
0.80
1
False positive rate (1 - specificity)
random classifier
P (TARGET_5Yrs=True) (AUC=1)
P (TARGET_5Yrs=False) (AUC=0)
Data
Manual
Wildcard
Regex
Type
Search
Aa
Plot
Cancel
Ok

Target column: Prediction (TARGET_5Yrs)

Positive class value: True

Prediction columns

Excludes: GP, MIN, PTS, FGM, FGA, FG%

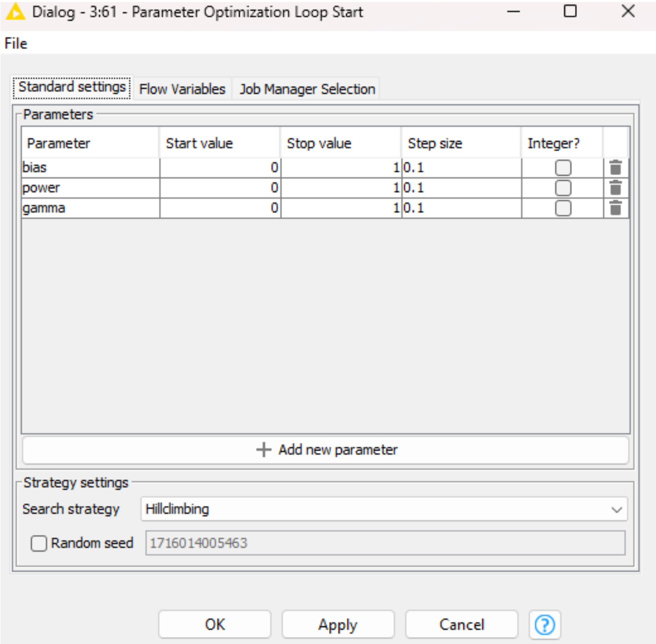
Any unknown columns

Includes: P (TARGET_5Y..., P (TARGET_5Y...

Title: ROC Curve

SVM Learner

Just as Decision Tree Learner, I used parameter optimization loop to adjust the polynomial variables to get the most accurate result possible.



Dialog - 3:61 - Parameter Optimization Loop Start

File

Standard settings | Flow Variables | Job Manager Selection

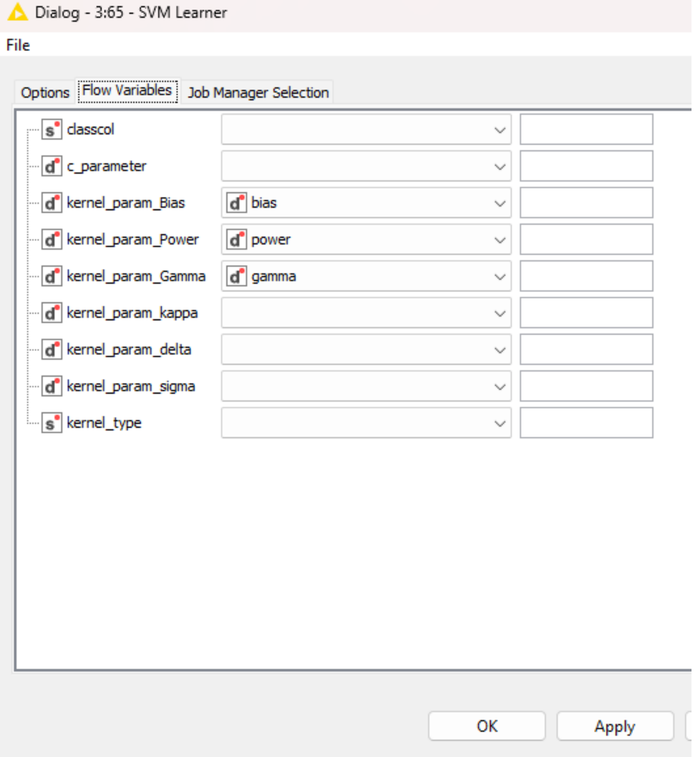
Parameters

Parameter	Start value	Stop value	Step size	Integer?
bias	0	1	0.1	□
power	0	1	0.1	□
gamma	0	1	0.1	□

Strategy settings

Search strategy: Hillclimbing

□ Random seed: 1716014005463

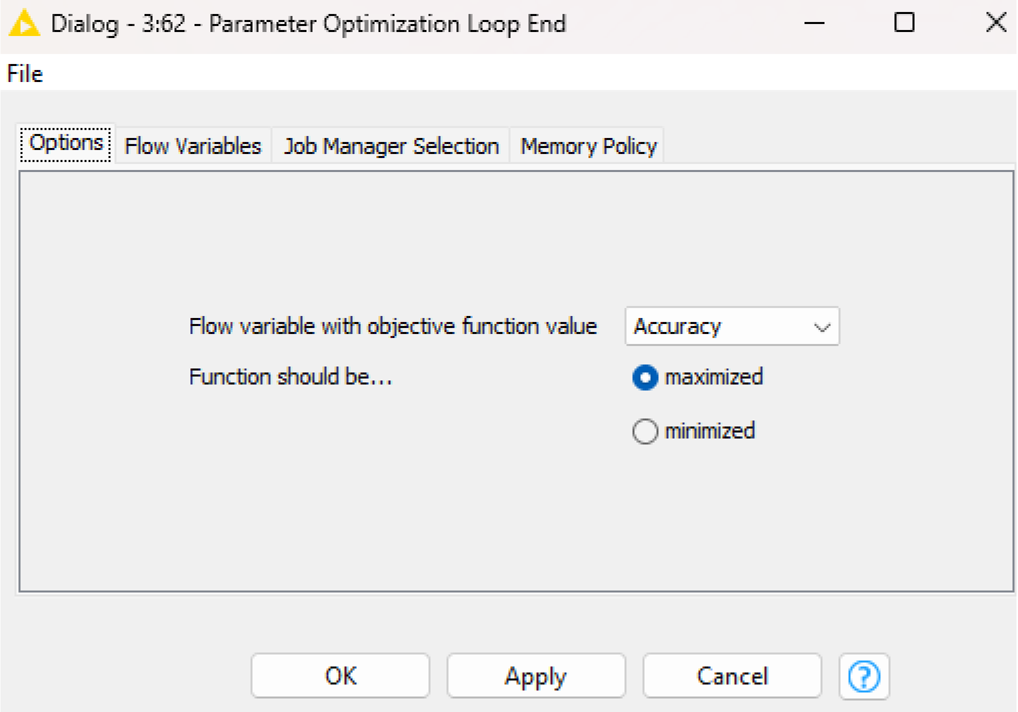


Dialog - 3:65 - SVM Learner

File

Options | Flow Variables | Job Manager Selection

classcol		
c_parameter		
kernel_param_Bias	bias	
kernel_param_Power	power	
kernel_param_Gamma	gamma	
kernel_param_kappa		
kernel_param_delta		
kernel_param_sigma		
kernel_type		



Dialog - 3:62 - Parameter Optimization Loop End

File

Options | Flow Variables | Job Manager Selection | Memory Policy

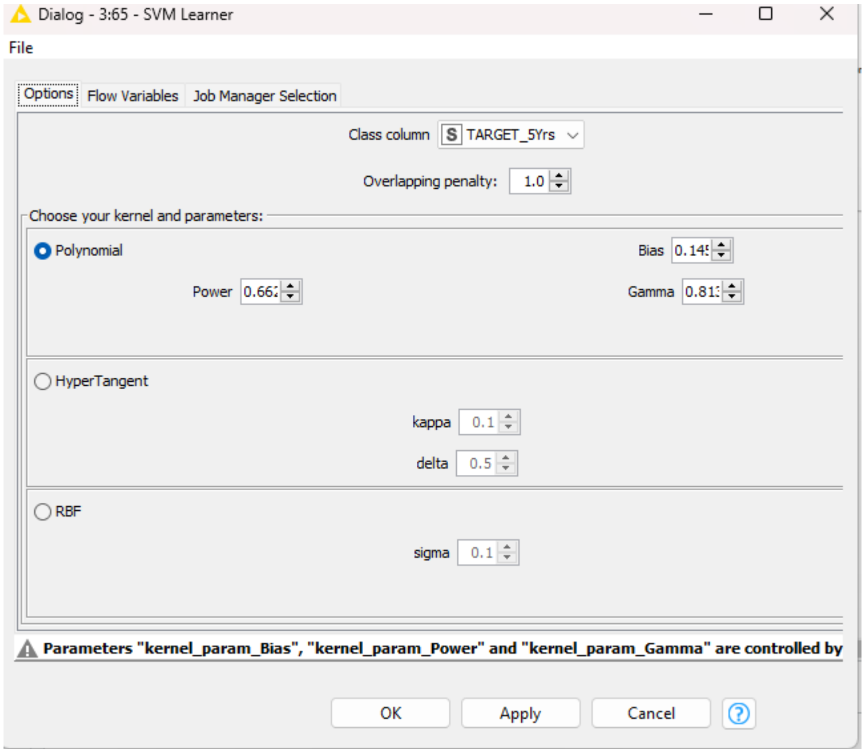
Flow variable with objective function value: Accuracy

Function should be…

● maximized

○ minimized

Using this finds the variables input for the best accuracy and, the settings of SVM learner are



Class column S TARGET_5Yrs

Overlapping penalty: 1.0

Choose your kernel and parameters:

● Polynomial

Bias 0.14

Power 0.66

Gamma 0.81

○ HyperTangent

kappa 0.1

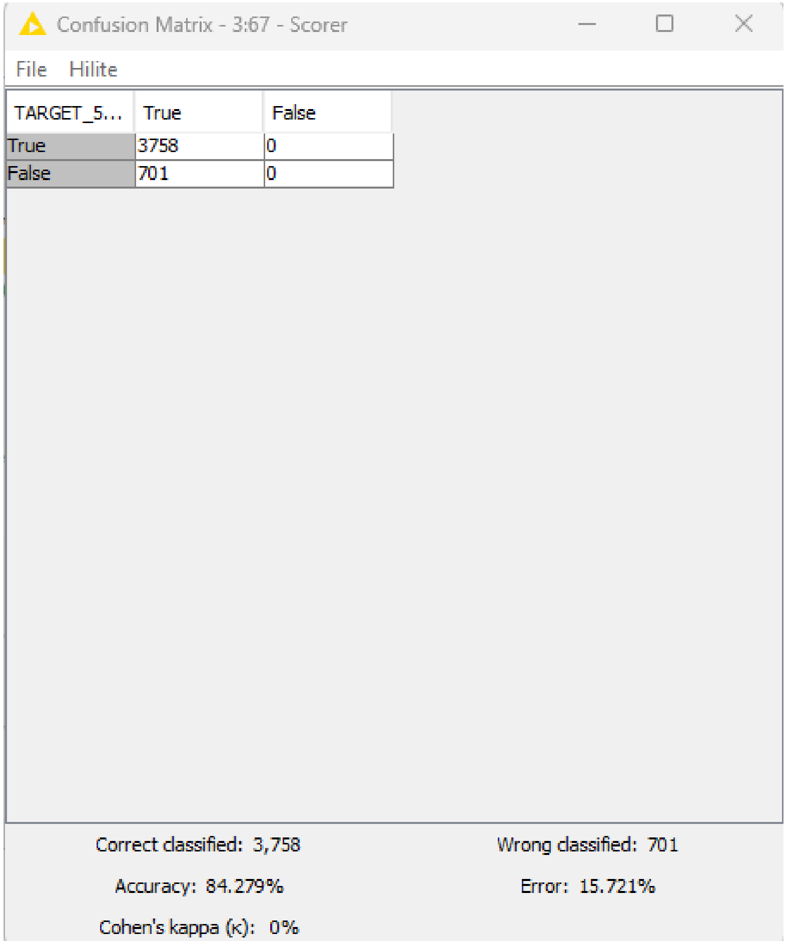
delta 0.5

○ RBF

sigma 0.1

Parameters "kernel_param_Bias", "kernel_param_Power" and "kernel_param_Gamma" are controlled by

The score of this classifier is



TARGET_5...	True	False
True	3758	0
False	701	0

Correct classified: 3,758

Wrong classified: 701

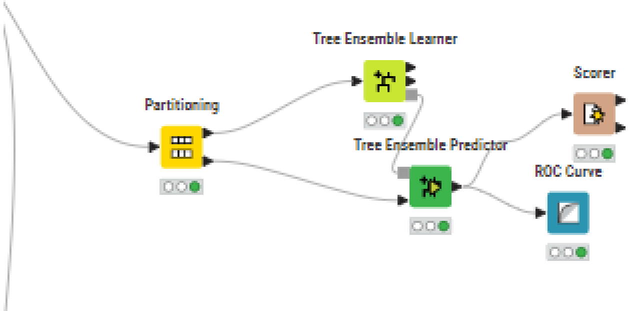
Accuracy: 84.279%

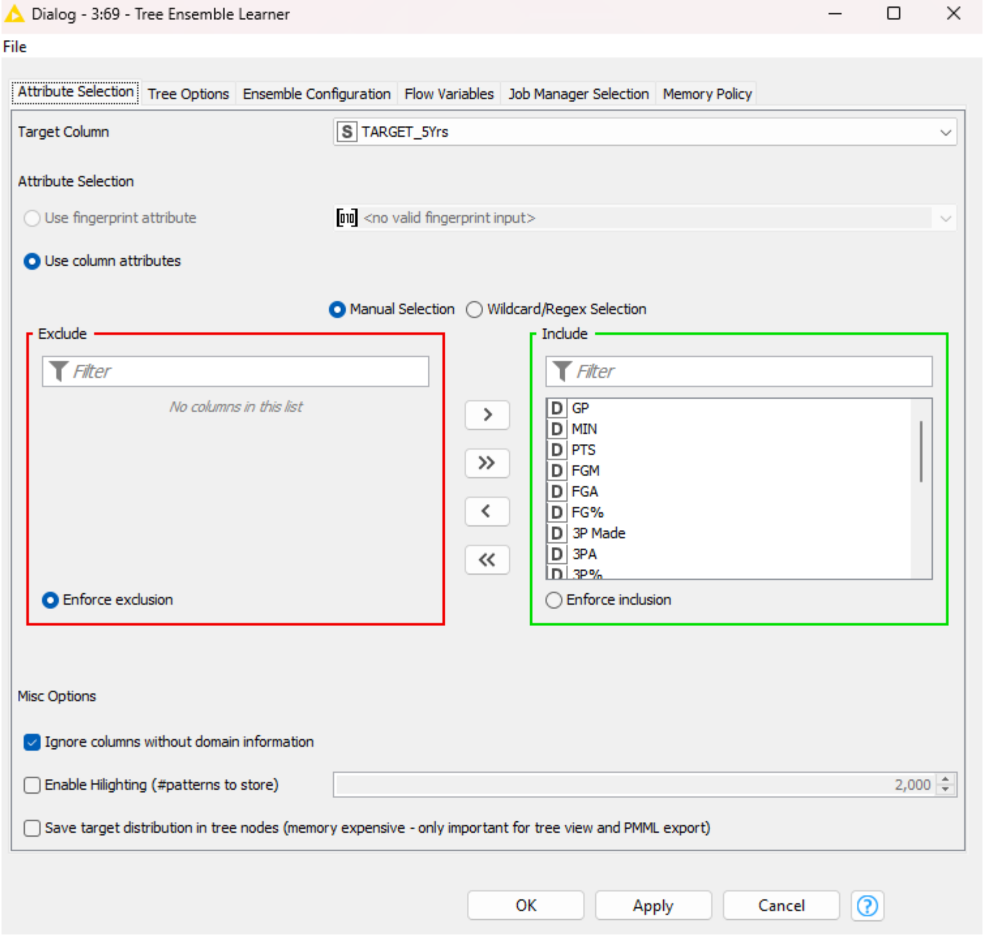
Error: 15.721%

Cohen's kappa (κ): 0%

Tree ensemble Learner

As there are no changeable parameters in Tree ensemble learner, I used the default settings for the learner.




Dialog - 3:69 - Tree Ensemble Learner

File

Attribute Selection | Tree Options | Ensemble Configuration | Flow Variables | Job Manager Selection | Memory Policy

Target Column: S TARGET_5Yrs

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

D GP
D MIN
D PTS
D FGM
D FGA
D FG%
D 3P Made
D 3PA
D 3P%

○ Enforce inclusion

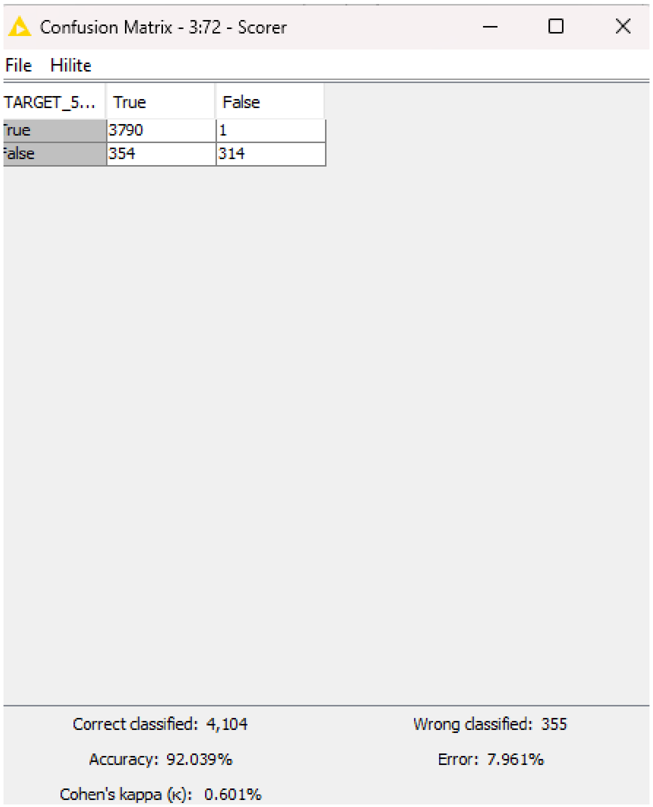
Misc Options

☑ Ignore columns without domain information

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

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

The score of this classifier is



TARGET_5...	True	False
True	3790	1
False	354	314

Correct classified: 4,104

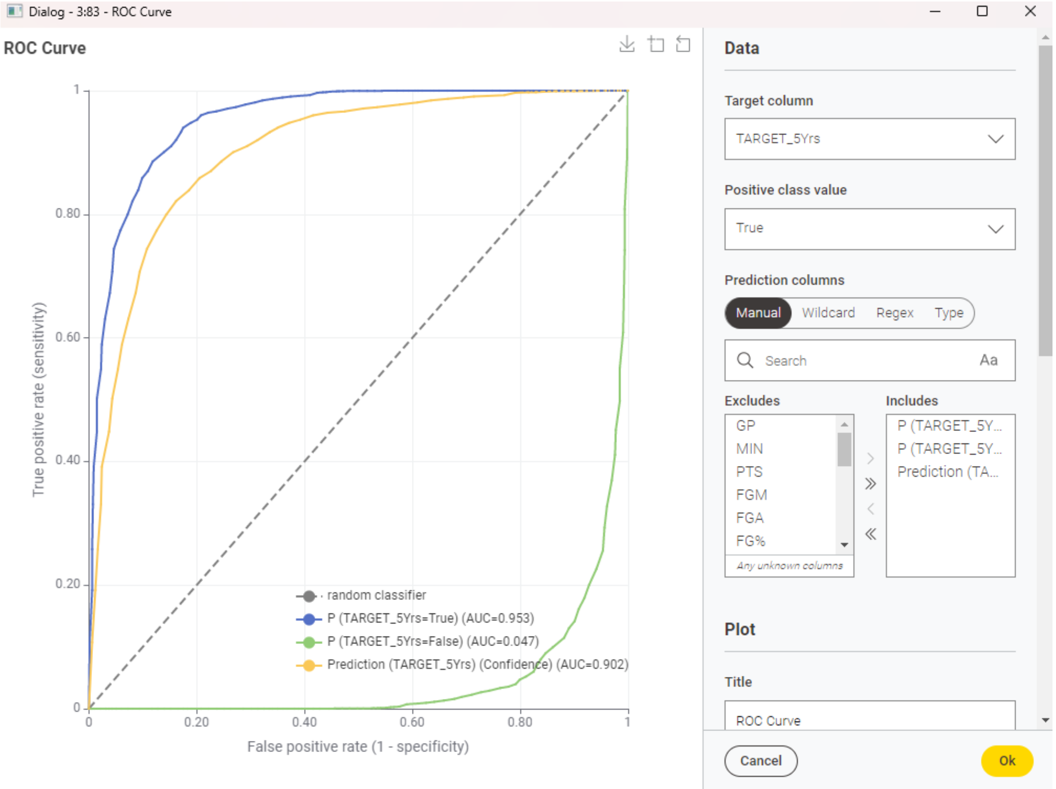
Wrong classified: 355

Accuracy: 92.039%

Error: 7.961%

Cohen's kappa (κ): 0.601%

ROC curve


Dialog - 3:83 - ROC Curve
ROC Curve
1
0.80
0.60
0.40
0.20
0
True positive rate (sensitivity)
0
0.20
0.40
0.60
0.80
1
False positive rate (1 - specificity)
random classifier
P (TARGET_5Yrs=True) (AUC=0.953)
P (TARGET_5Yrs=False) (AUC=0.047)
Prediction (TARGET_5Yrs) (Confidence) (AUC=0.902)
Data
Manual
Wildcard
Regex
Type
Search
Aa
Any unknown columns
Plot
Cancel
Ok

Target column: TARGET_5Yrs

Positive class value: True

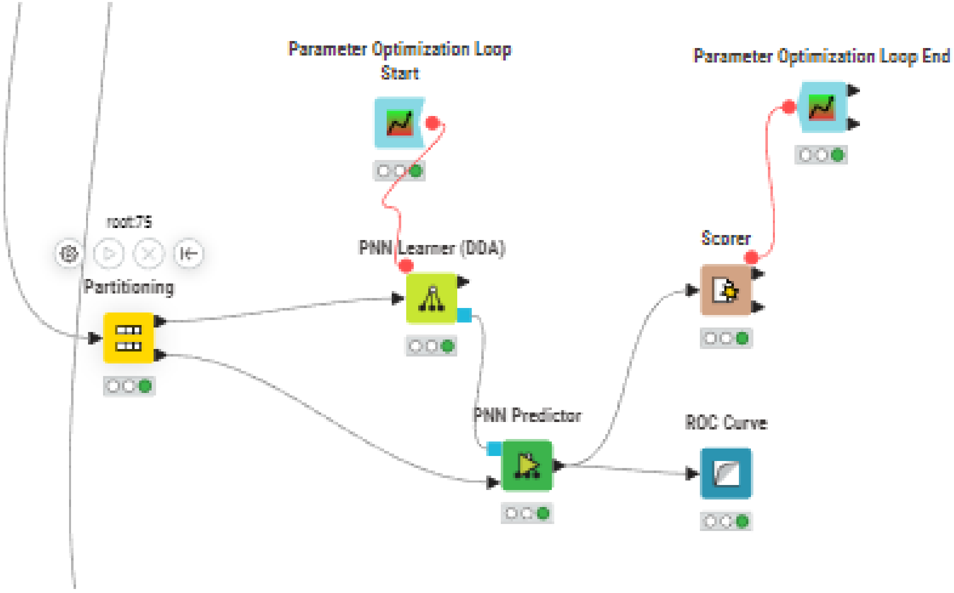
Prediction columns

Excludes: GP, MIN, PTS, FGM, FGA, FG%

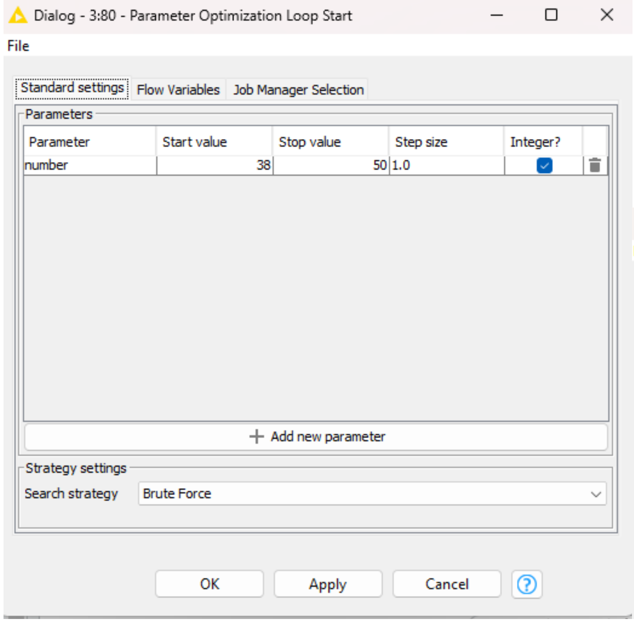
Includes: P (TARGET_5Y..., P (TARGET_5Y..., Prediction (TA...

Title: ROC Curve

PNN learner


Parameter Optimization Loop Start
Parameter Optimization Loop End
root:75
Partitioning
PNN Learner (DDA)
Scorer
PNN Predictor
ROC Curve

In the PNN learner, I used parameter optimization loop to find out the best possible accuracy, I made a variable for the maximum no of epochs. And the other settings are default



Dialog - 3:80 - Parameter Optimization Loop Start

File

Standard settings | Flow Variables | Job Manager Selection

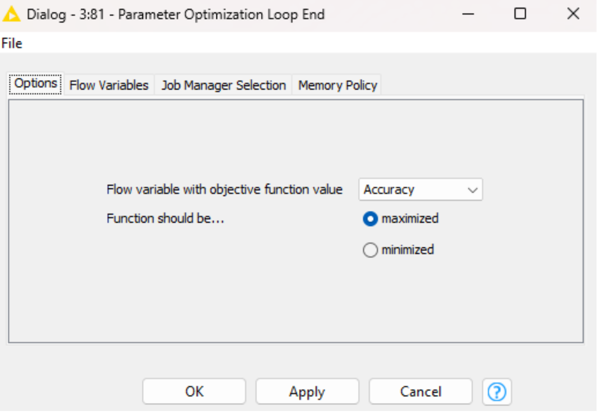
Parameters

Parameter	Start value	Stop value	Step size	Integer?
number	38	50	1.0	☑

+ Add new parameter

Strategy settings

Search strategy: Brute Force



Dialog - 3:81 - Parameter Optimization Loop End

File

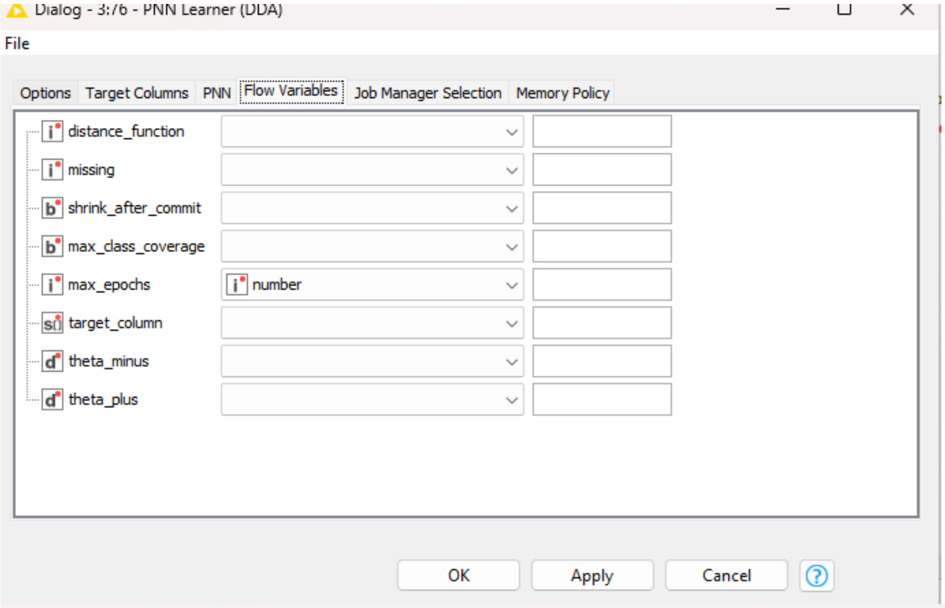
Options | Flow Variables | Job Manager Selection | Memory Policy

Flow variable with objective function value: Accuracy

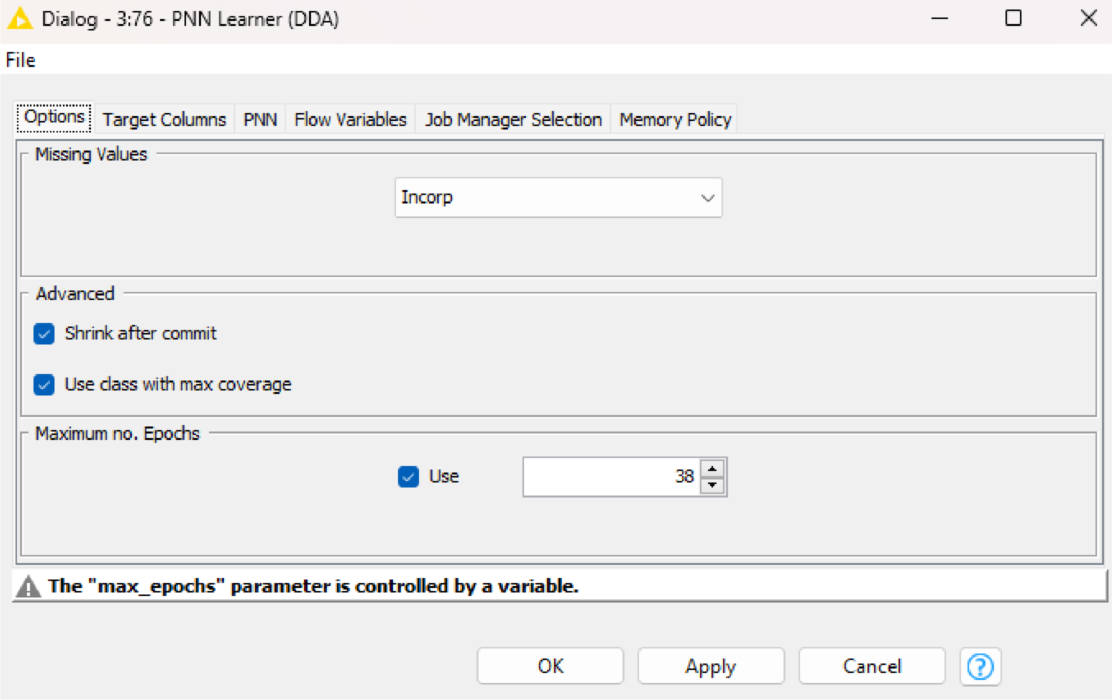
Function should be...

● maximized

○ minimized



distance_function
missing
shrink_after_commit
max_class_coverage
max_epochs: number
target_column
theta_minus
theta_plus



Missing Values: Incorp

Advanced

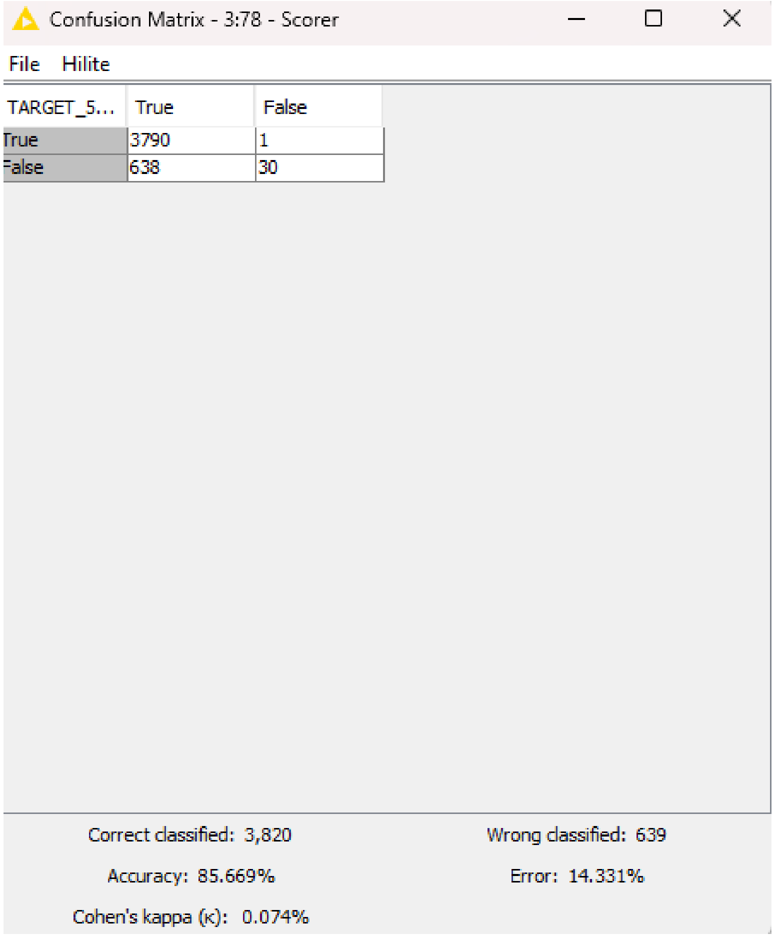
☑ Shrink after commit

☑ Use class with max coverage

Maximum no. Epochs: ☑ Use 38

The "max_epochs" parameter is controlled by a variable.

The score of the classifier is as follows



Confusion Matrix - 3:78 - Scorer

TARGET_5...	True	False
True	3790	1
False	638	30

Correct classified: 3,820

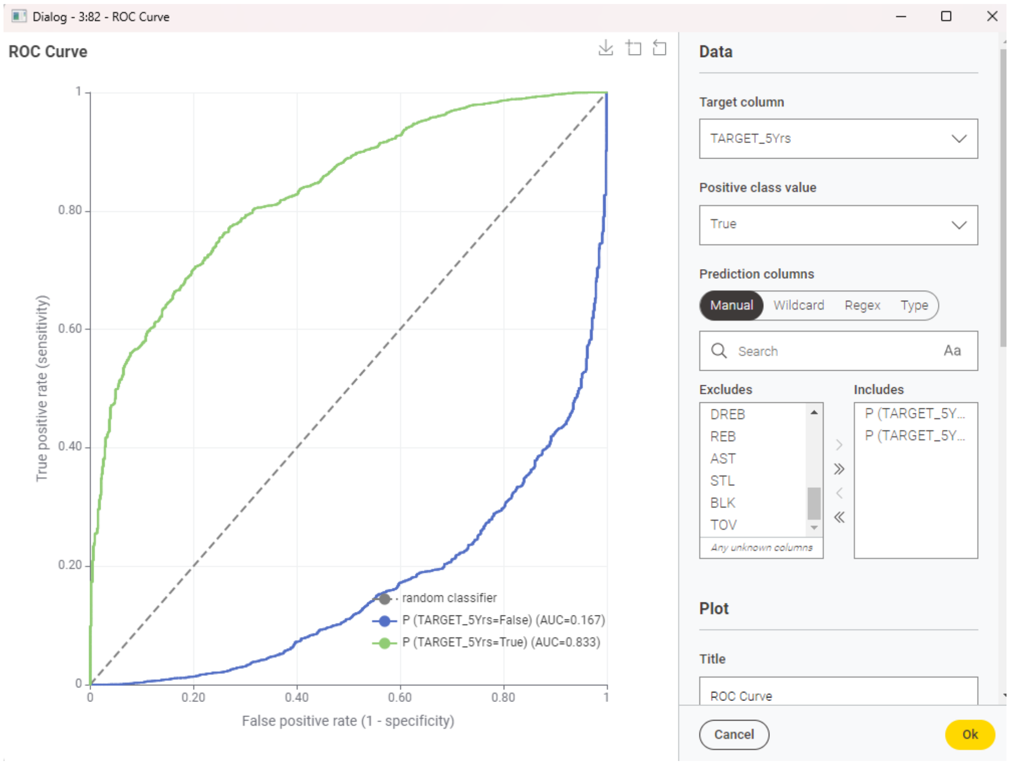
Wrong classified: 639

Accuracy: 85.669%

Error: 14.331%

Cohen's kappa (κ): 0.074%

ROC curve



Dialog - 3:82 - ROC Curve

ROC Curve
1
0.80
0.60
0.40
0.20
0
True positive rate (sensitivity)
0
0.20
0.40
0.60
0.80
1
False positive rate (1 - specificity)
random classifier
P (TARGET_5Yrs=False) (AUC=0.167)
P (TARGET_5Yrs=True) (AUC=0.833)

Data

Target column: TARGET_5Yrs

Positive class value: True

Prediction columns: Manual Wildcard Regex Type

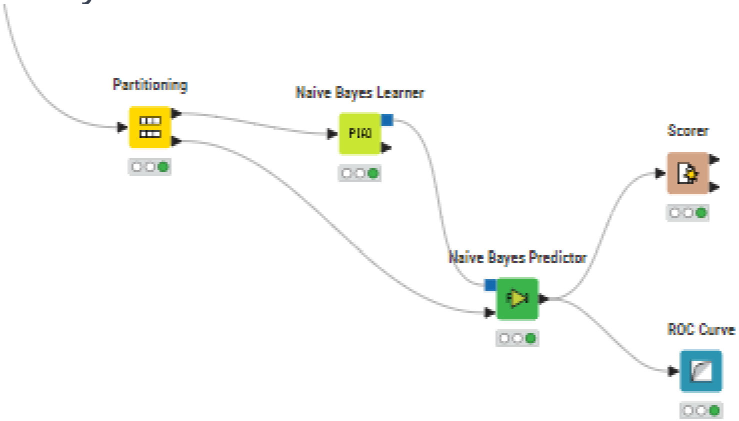
Excludes: DREB, REB, AST, STL, BLK, TOV, *Any unknown columns*

Includes: P (TARGET_5Y..., P (TARGET_5Y...

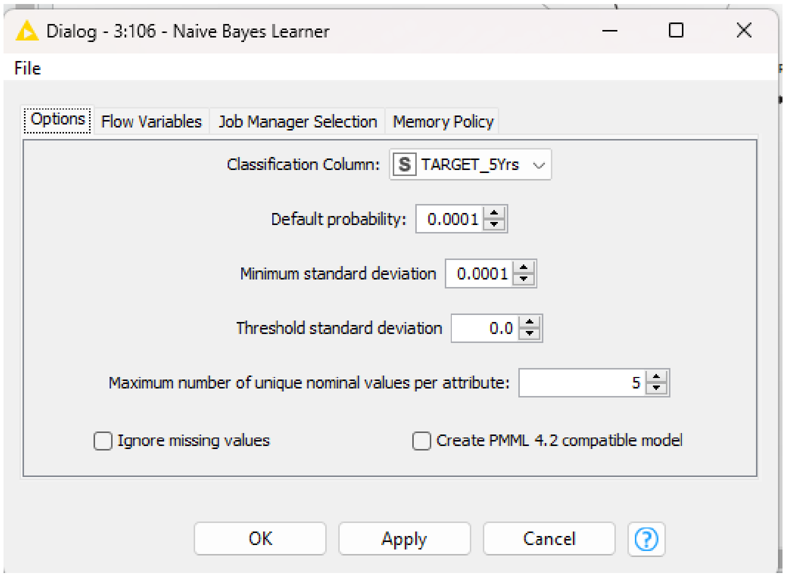
Plot

Title: ROC Curve

Naive Bayes Learner



Naive Bayes Learner with the default settings



Classification Column: S TARGET_5Yrs

Default probability: 0.0001

Minimum standard deviation 0.0001

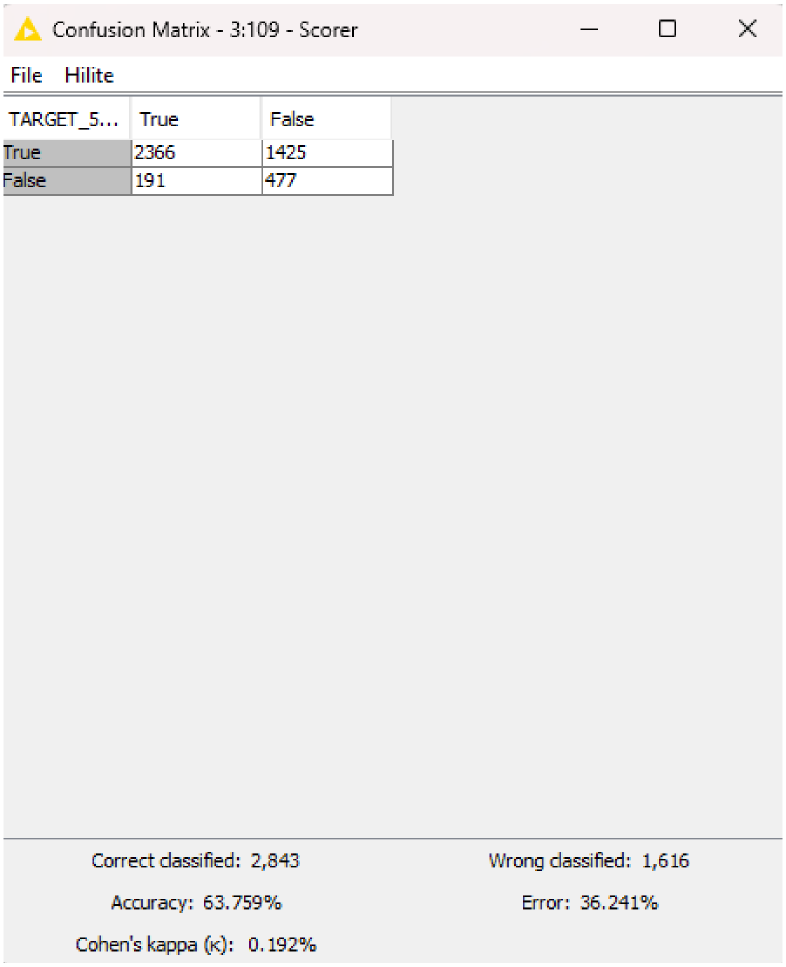
Threshold standard deviation 0.0

Maximum number of unique nominal values per attribute: 5

□ Ignore missing values

□ Create PMML 4.2 compatible model

The score of Naive Bayes Learner



TARGET_5...	True	False
True	2366	1425
False	191	477

Correct classified: 2,843

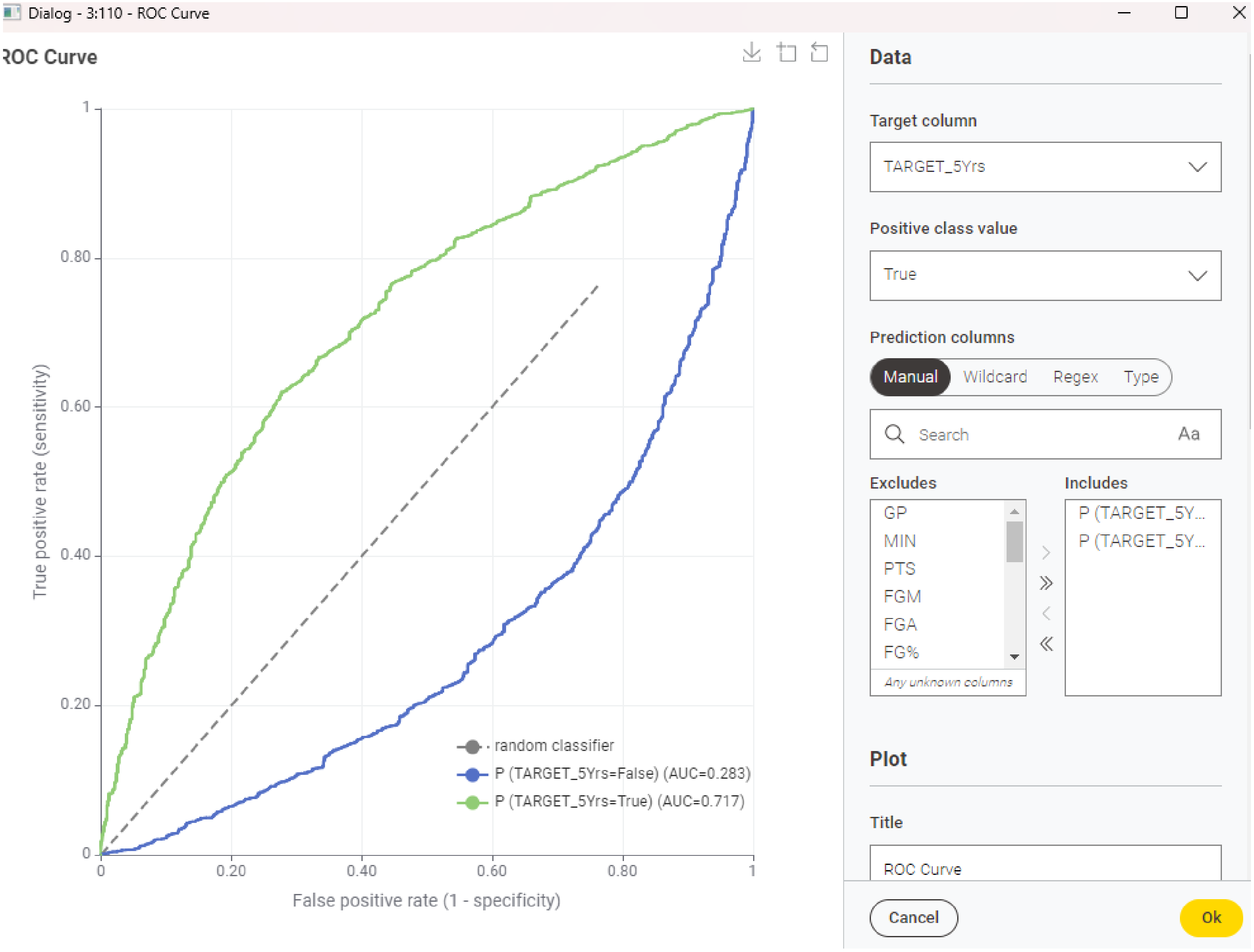
Wrong classified: 1,616

Accuracy: 63.759%

Error: 36.241%

Cohen's kappa (κ): 0.192%

ROC curve


ROC Curve
1
0.80
0.60
0.40
0.20
0
True positive rate (sensitivity)
0
0.20
0.40
0.60
0.80
1
False positive rate (1 - specificity)
random classifier
P (TARGET_5Yrs=False) (AUC=0.283)
P (TARGET_5Yrs=True) (AUC=0.717)

Dialog - 3:110 - ROC Curve

Data

Target column: TARGET_5Yrs

Positive class value: True

Prediction columns: Manual | Wildcard | Regex | Type

Search Aa

Excludes:
GP
MIN
PTS
FGM
FGA
FG%
Any unknown columns

Includes:
P (TARGET_5Y...
P (TARGET_5Y...

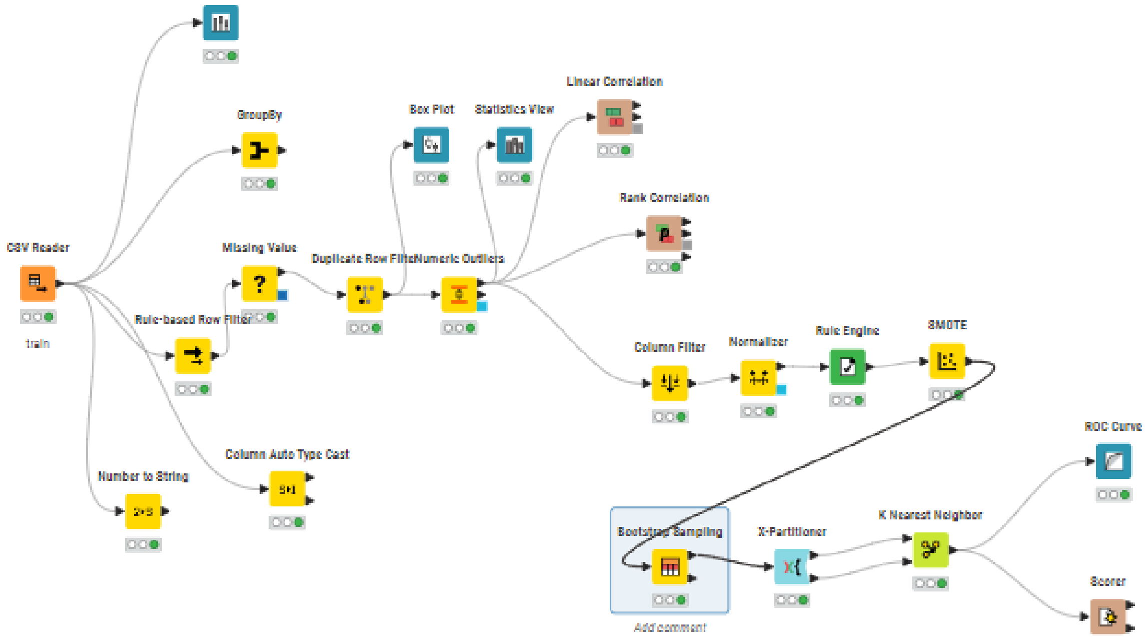
Plot

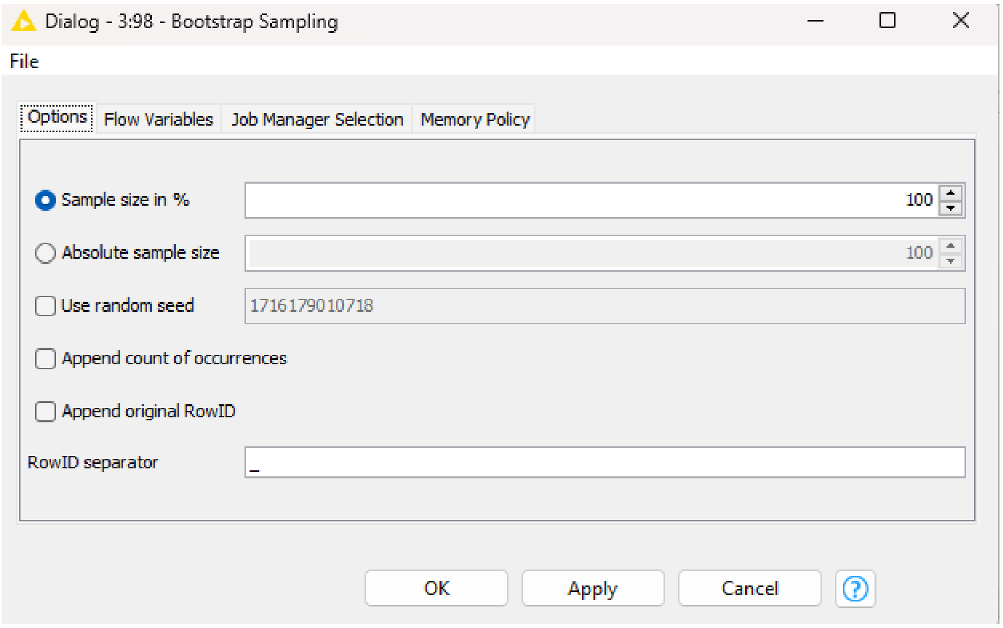
Title: ROC Curve

Cancel Ok

KNN

In the KNN classifier, I used Bootstrap Sampling node before partitioning, and X- partitioner for the partitioning node.




Dialog - 3:98 - Bootstrap Sampling

● Sample size in % 100

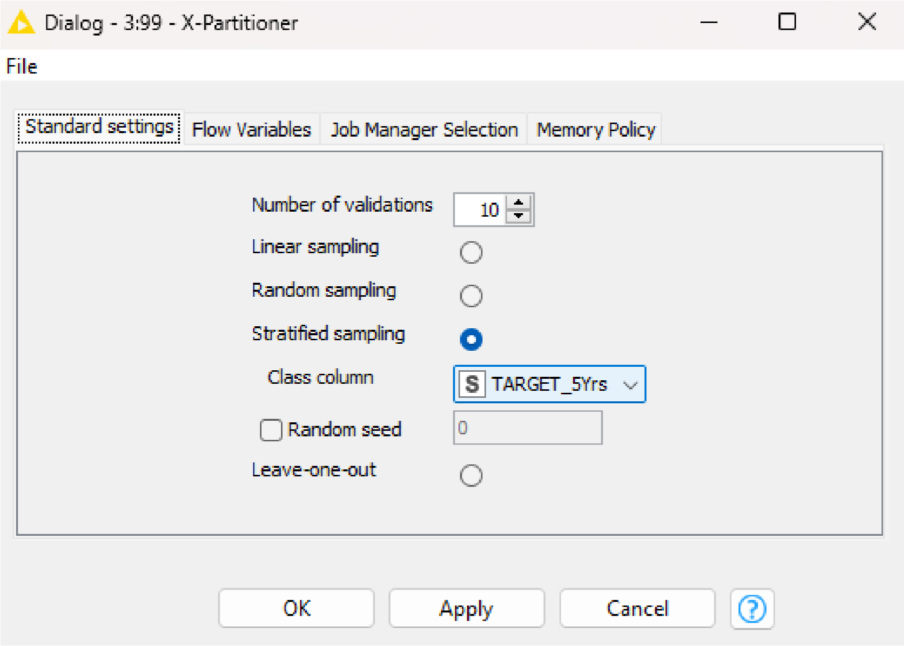
○ Absolute sample size 100

□ Use random seed 1716179010718

□ Append count of occurrences

□ Append original RowID

RowID separator _



Dialog - 3:99 - X-Partitioner

Number of validations 10

Linear sampling ○

Random sampling ○

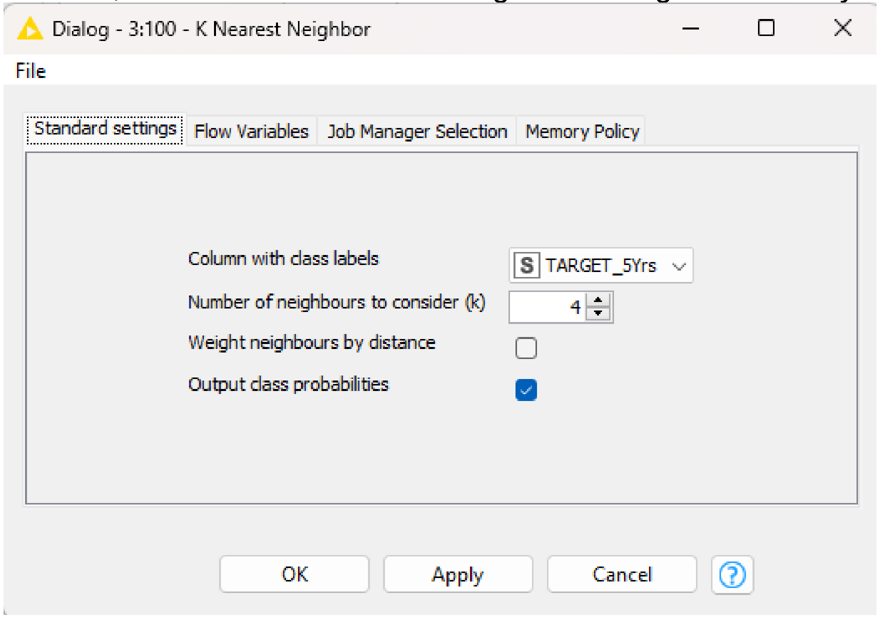
Stratified sampling ●

Class column S TARGET_5Yrs

□ Random seed 0

Leave-one-out ○

IN KNN, I used k value 4 so that it gives the higher accuracy than the default settings.



Dialog - 3:100 - K Nearest Neighbor

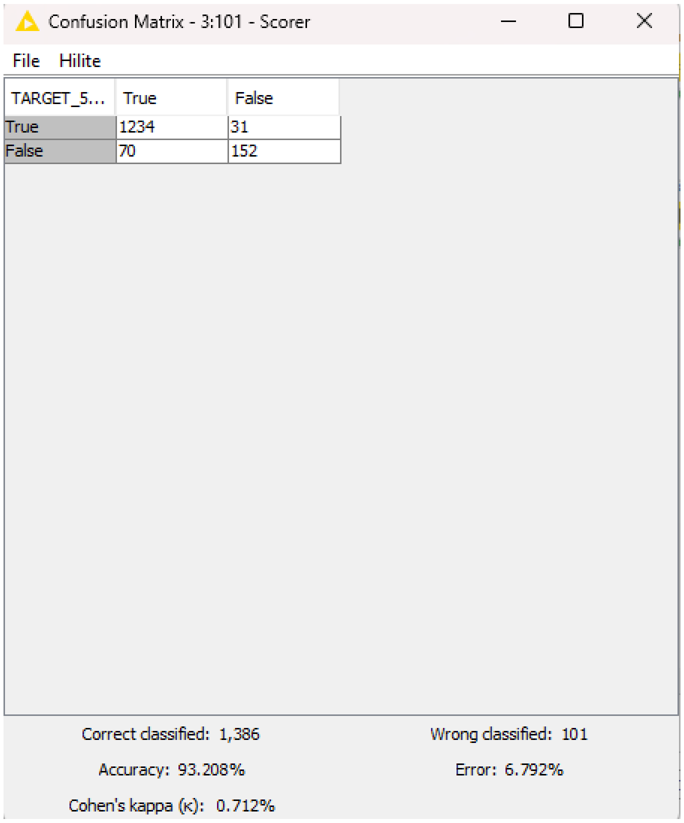
Column with class labels S TARGET_5Yrs

Number of neighbours to consider (k) 4

Weight neighbours by distance □

Output class probabilities ☑

Here is the score of the classifier



TARGET_5...	True	False
True	1234	31
False	70	152

Correct classified: 1,386

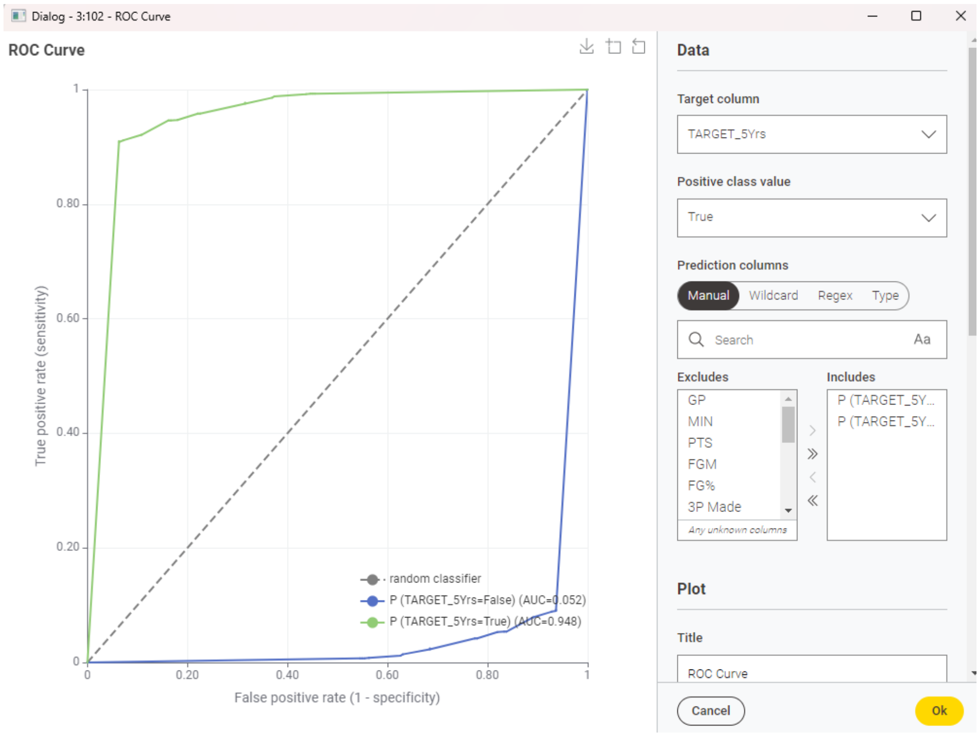
Wrong classified: 101

Accuracy: 93.208%

Error: 6.792%

Cohen's kappa (κ): 0.712%

ROC curve


Dialog - 3:102 - ROC Curve
ROC Curve
1
0.80
0.60
0.40
0.20
0
True positive rate (sensitivity)
0
0.20
0.40
0.60
0.80
1
False positive rate (1 - specificity)
random classifier
P (TARGET_5Yrs=False) (AUC=0.052)
P (TARGET_5Yrs=True) (AUC=0.948)
Data
Manual
Wildcard
Regex
Type
Search
Aa
Any unknown columns
Plot
Cancel
Ok

Target column: TARGET_5Yrs

Positive class value: True

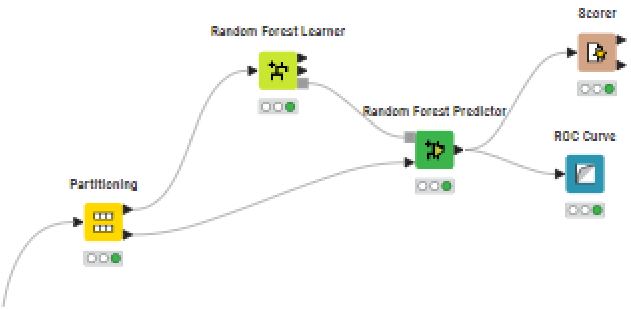
Prediction columns

Excludes: GP, MIN, PTS, FGM, FG%, 3P Made

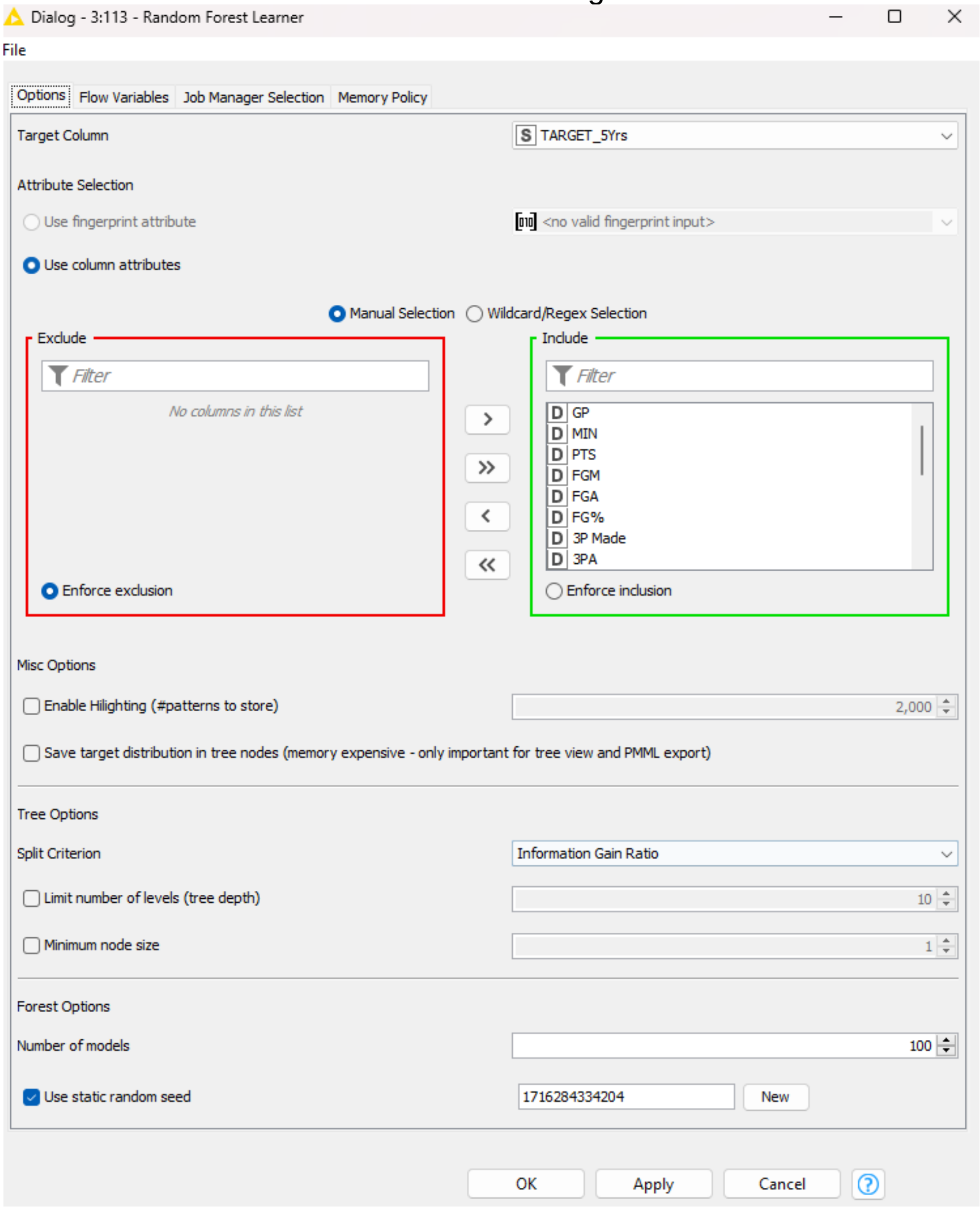
Includes: P (TARGET_5Y..., P (TARGET_5Y...

Title: ROC Curve

Random Forest Learner



I used random forest learner default settings



Dialog - 3:113 - Random Forest Learner

File

Options | Flow Variables | Job Manager Selection | Memory Policy

Target Column: S TARGET_5Yrs

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

D GP
D MIN
D PTS
D FGM
D FGA
D FG%
D 3P Made
D 3PA

○ 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

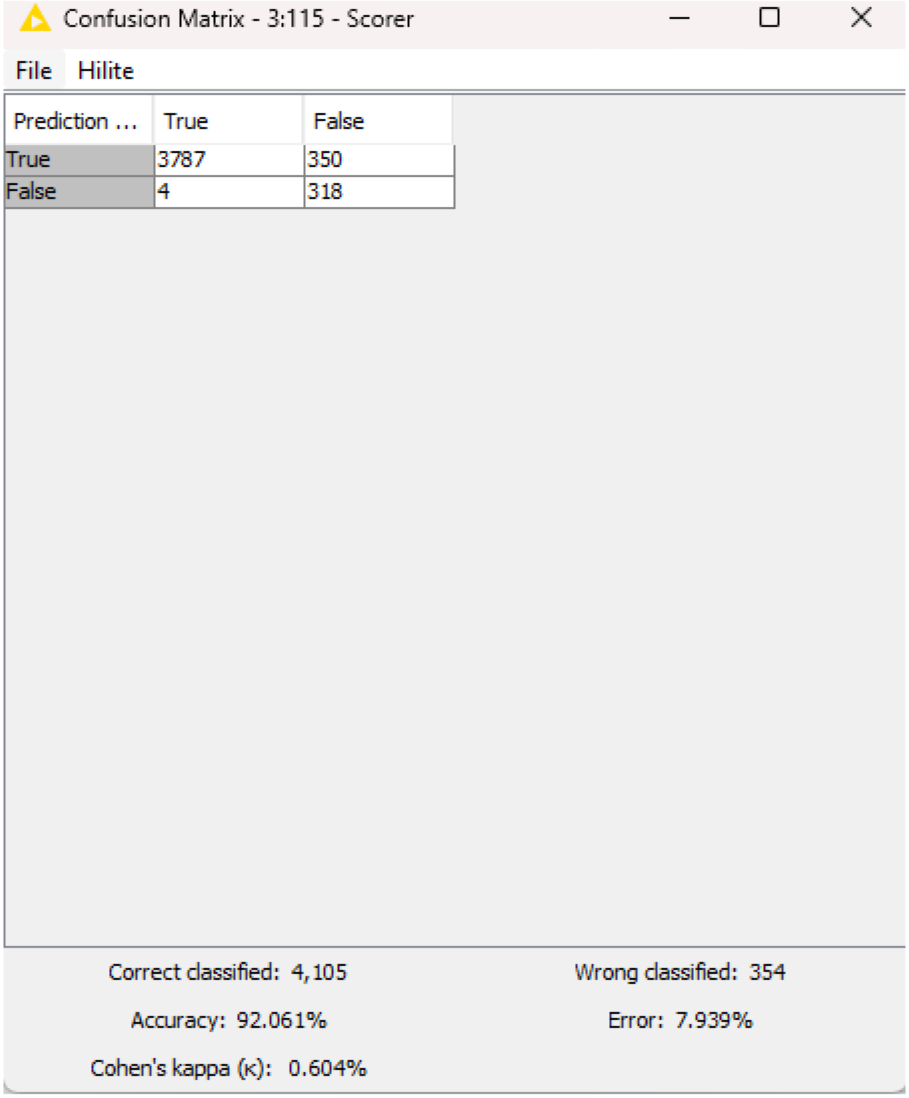
Forest Options

Number of models: 100

☑ Use static random seed — 1716284334204 — New

OK | Apply | Cancel | ?

The score of this classifier is



Prediction ...	True	False
True	3787	350
False	4	318

Correct classified: 4,105

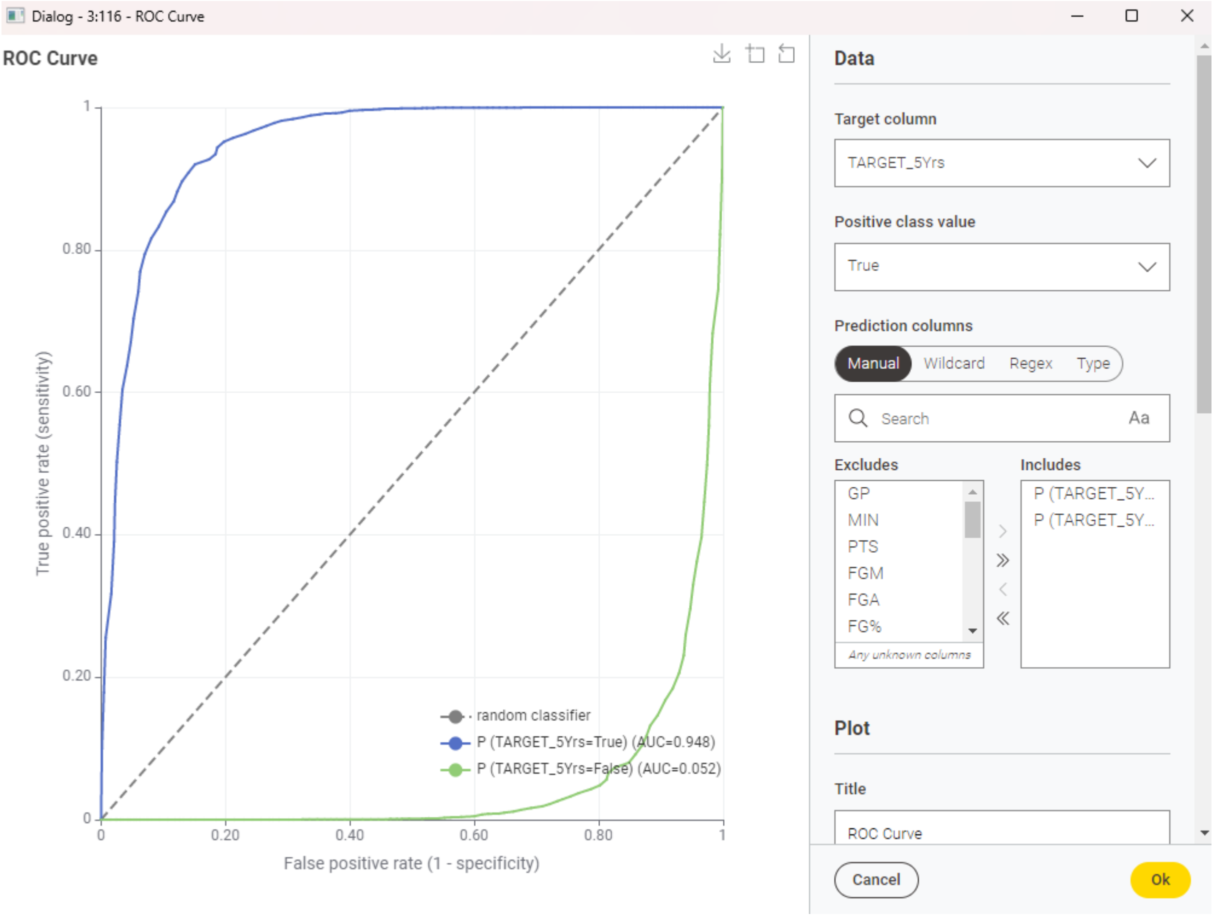
Wrong classified: 354

Accuracy: 92.061%

Error: 7.939%

Cohen's kappa (κ): 0.604%

ROC Curve



Dialog - 3:116 - ROC Curve

ROC Curve

1
0.80
0.60
0.40
0.20
0
True positive rate (sensitivity)
0
0.20
0.40
0.60
0.80
1
False positive rate (1 - specificity)
random classifier
P (TARGET_5Yrs=True) (AUC=0.948)
P (TARGET_5Yrs=False) (AUC=0.052)

Data

Target column: TARGET_5Yrs

Positive class value: True

Prediction columns

Manual
Wildcard
Regex
Type
Search
Aa

Excludes: GP, MIN, PTS, FGM, FGA, FG%

Includes: P (TARGET_5Y... P (TARGET_5Y...

Plot

Title: ROC Curve

Best Classifier

	Accuracy	Recall	Precision	F1 score

SVM	84.279	1	1	
Decision Tree	88.069	0.9183	0.9435	0.9307
PNN	85.669	0.8559	0.9997	0.9222
Naive Bayes	63.759	0.9253	0.6241	0.7454
KNN	93.208	0.9463	0.9754	0.9606
Tree Ensemble	92.039	0.9145	0.9997	0.9952
Random Forest	92.061	0.9989	0.9153	0.9552

With the metrics resulted, its accuracy, recall and precision is high on Tree Ensemble, also F1 score of Tree Ensemble is high.

Random Forest builds multiple decision trees (bagging) and combines their predictions. Each tree is trained on a random subset of features and instances.

Random Forest is robust, handles noisy data, and avoids overfitting. Its high precision suggests it’s good at minimizing false positives.