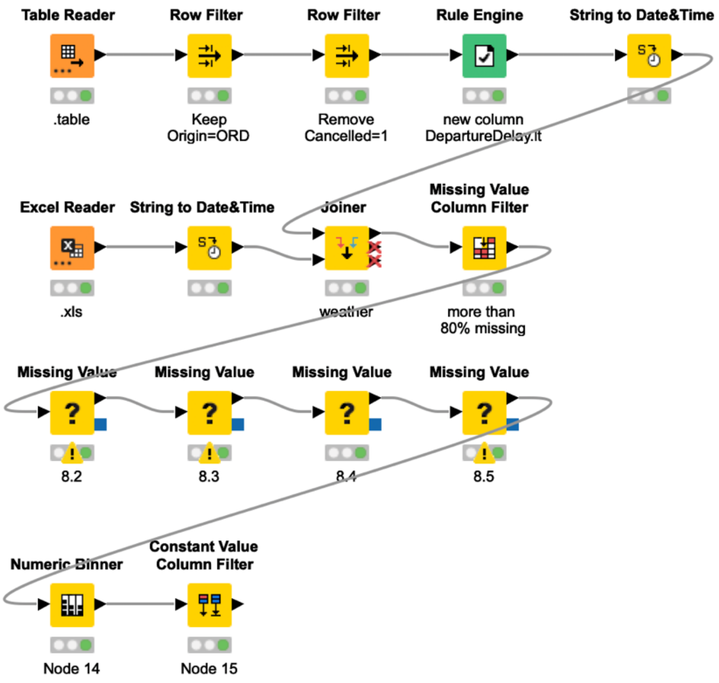
1. Exercise Data Import and Pre-Processing

Goal: Access, extend and prepare data.



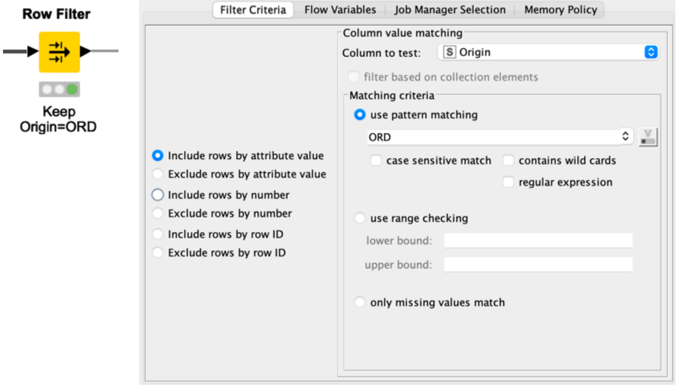
Download Datasets from Moodle:

1. *AirlineDataset.table*
2. *GHCN-Daily_source.xls contains daily weather information like precipitation, snowfall, snow depth, temperature, wind speed and wind direction measured at Chicago O'Hare International Airport.*

1. Read the data *AirlineDataset.table* (***Table Reader node)***

Advanced Settings → nichts angekreuzt

2. Keep only records with Origin=ORD (**Row Filter node)**



Column value matching

Column to test: S Origin

□ filter based on collection elements

Matching criteria

● use pattern matching

ORD

□ case sensitive match □ contains wild cards

□ regular expression

○ use range checking

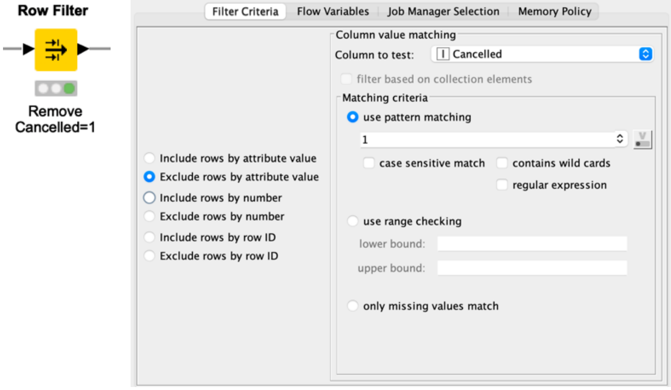
lower bound:

upper bound:

○ only missing values match

● 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

3. Remove canceled flights, i.e. Cancelled=1 (**Row Filter node)**



Column value matching

Column to test: I Cancelled

□ filter based on collection elements

Matching criteria

● use pattern matching

1

□ case sensitive match □ contains wild cards

□ regular expression

○ use range checking

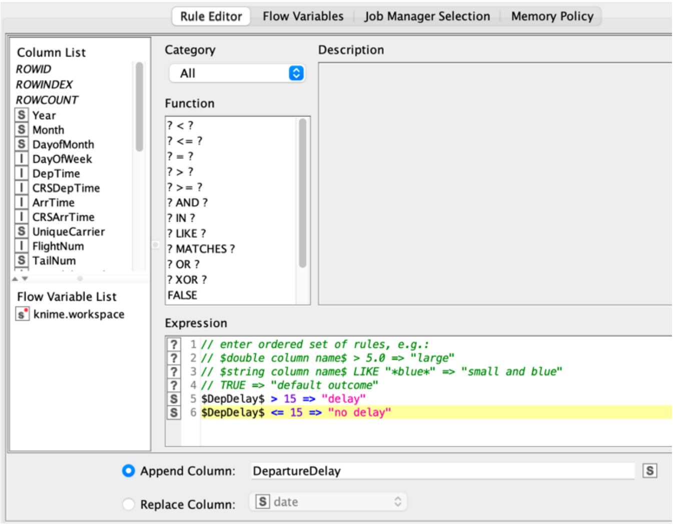
lower bound:

upper bound:

○ only missing values match

○ 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

4. Create a new column called DepartureDelay. It acquires the value "delay" if DepDelay > 15min and "no delay" otherwise (**Rule Engine node).**



Column List
ROWID
ROWINDEX
ROWCOUNT
S Year
S Month
S DayofMonth
I DayOfWeek
I DepTime
I CRSDepTime
I ArrTime
I CRSArrTime
S UniqueCarrier
I FlightNum
S TailNum

Flow Variable List
S knime.workspace

Category: All

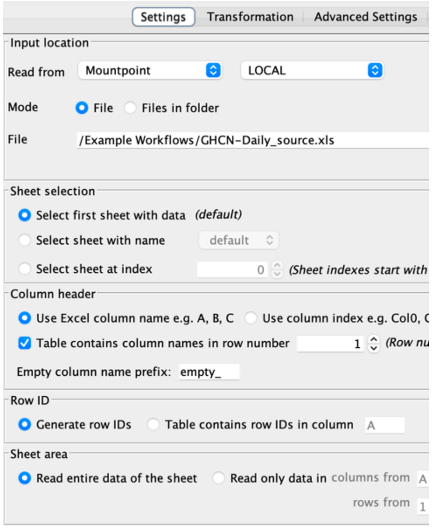
Function
? < ?
? <= ?
? = ?
? > ?
? >= ?
? AND ?
? IN ?
? LIKE ?
? MATCHES ?
? OR ?
? XOR ?
FALSE

Description

Expression
```
// enter ordered set of rules, e.g.:
// $double column name$ > 5.0 => "large"
// $string column name$ LIKE "*blue*" => "small and blue"
// TRUE => "default outcome"
$DepDelay$ > 15 => "delay"
$DepDelay$ <= 15 => "no delay"
```

● Append Column: DepartureDelay
○ Replace Column: date

5. Read the *GHCN-Daily_source.xls file* (***Excel Reader (XLS) node)***



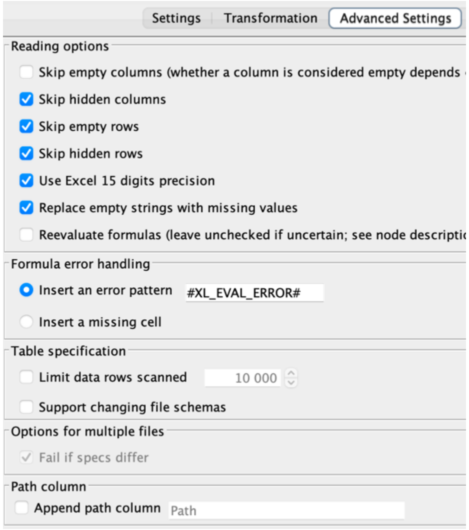
Input location
Read from: Mountpoint LOCAL
Mode: ● File ○ Files in folder
File: /Example Workflows/GHCN-Daily_source.xls

Sheet selection
● Select first sheet with data *(default)*
○ Select sheet with name default
○ Select sheet at index 0 *(Sheet indexes start with*

Column header
● Use Excel column name e.g. A, B, C ○ Use column index e.g. Col0, C
☑ Table contains column names in row number 1 *(Row nu*
Empty column name prefix: empty_

Row ID
● Generate row IDs ○ Table contains row IDs in column A

Sheet area
● Read entire data of the sheet ○ Read only data in columns from A
rows from 1



Reading options
□ Skip empty columns (whether a column is considered empty depends
☑ Skip hidden columns
☑ Skip empty rows
☑ Skip hidden rows
☑ Use Excel 15 digits precision
☑ Replace empty strings with missing values
□ Reevaluate formulas (leave unchecked if uncertain; see node descriptio

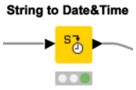
Formula error handling
● Insert an error pattern #XL_EVAL_ERROR#
○ Insert a missing cell

Table specification
□ Limit data rows scanned 10 000
□ Support changing file schemas

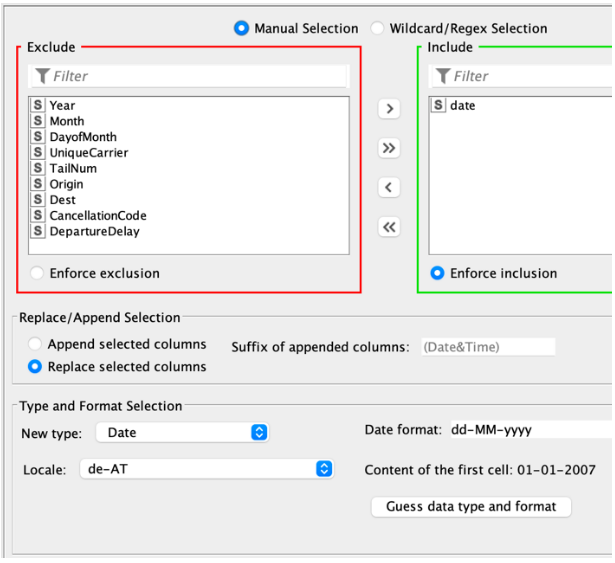
Options for multiple files
✓ Fail if specs differ

Path column
□ Append path column Path

6. Convert the date columns in both datasets from the data type string to date&time (**String to Date&Time node)**



AirlineDataset.table



● Manual Selection ○ Wildcard/Regex Selection

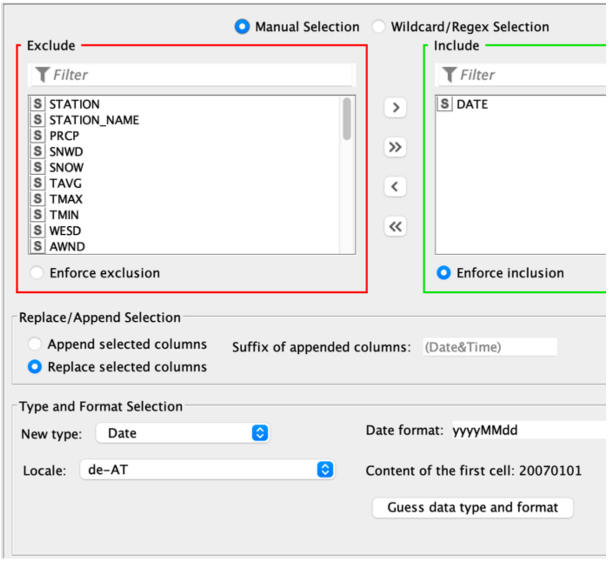
Exclude
S Year
S Month
S DayofMonth
S UniqueCarrier
S TailNum
S Origin
S Dest
S CancellationCode
S DepartureDelay
○ Enforce exclusion

Include
S date
● Enforce inclusion

Replace/Append Selection
○ Append selected columns Suffix of appended columns: (Date&Time)
● Replace selected columns

Type and Format Selection
New type: Date Date format: dd-MM-yyyy
Locale: de-AT Content of the first cell: 01-01-2007
Guess data type and format

GHCN-Daily_source.xls



● Manual Selection ○ Wildcard/Regex Selection

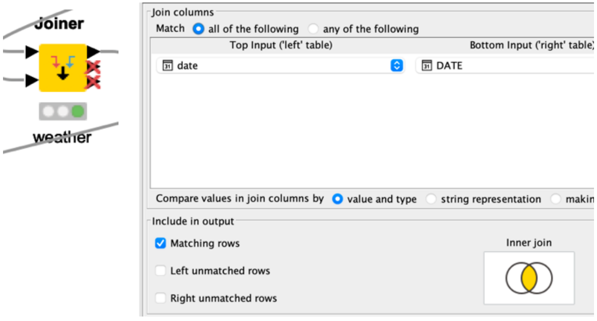
Exclude
S STATION
S STATION_NAME
S PRCP
S SNWD
S SNOW
S TAVG
S TMAX
S TMIN
S WESD
S AWND
○ Enforce exclusion

Include
S DATE
● Enforce inclusion

Replace/Append Selection
○ Append selected columns Suffix of appended columns: (Date&Time)
● Replace selected columns

Type and Format Selection
New type: Date Date format: yyyyMMdd
Locale: de-AT Content of the first cell: 20070101
Guess data type and format

7. Join the weather data with the airline data using the date columns as the joining columns (**Joiner node). Use inner join.**



Join columns

Match ● all of the following ○ any of the following

Top Input ('left' table)	Bottom Input ('right' table)
date	DATE

Compare values in join columns by ● value and type ○ string representation ○ makin

Include in output

☑ Matching rows

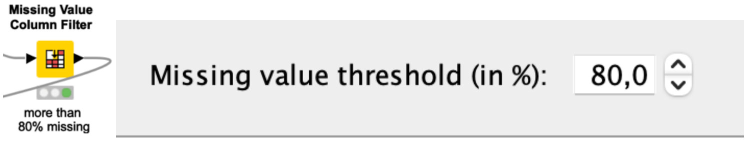
□ Left unmatched rows

□ Right unmatched rows

Inner join

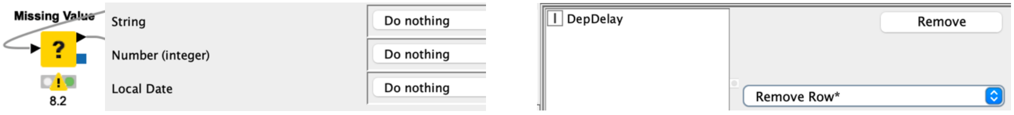
8. Handle missing values:

8.1 Remove columns that contain more than 80% missing values (**Missing Value Column Filter node)**



Missing value threshold (in %): 80,0

8.2 If the data contains rows where the value of DepDelay is missing, remove them (**Row Filter node OR Missing Value node)**



String: Do nothing

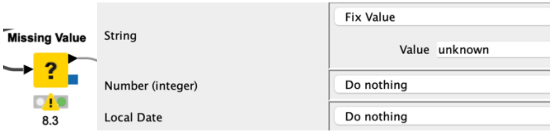
Number (integer): Do nothing

Local Date: Do nothing

DepDelay: Remove Row*

Remove

8.3 Set missing values in string columns to a fixed value "unknown" (**Missing Value node)**

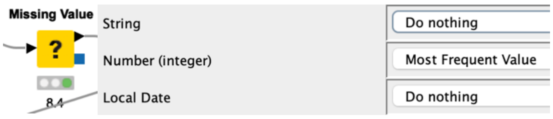


String: Fix Value, Value unknown

Number (integer): Do nothing

Local Date: Do nothing

8.4 Set missing values in integer columns to the most frequent value in the column (**Missing Value node)**

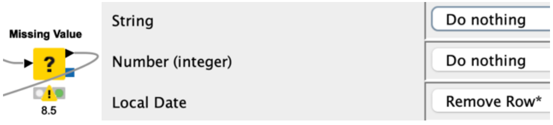


String: Do nothing

Number (integer): Most Frequent Value

Local Date: Do nothing

8.5 Remove rows that have missing values in a column of type date&time (**Missing Value node)**

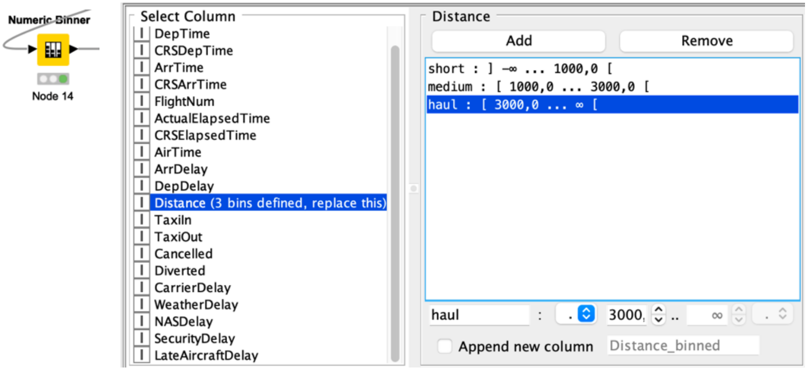


String: Do nothing

Number (integer): Do nothing

Local Date: Remove Row*

9. Bin flight distance to three bins: long, short and medium haul (**Numeric Binner node)**



Select Column: DepTime, CRSDepTime, ArrTime, CRSArrTime, FlightNum, ActualElapsedTime, CRSElapsedTime, AirTime, ArrDelay, DepDelay, Distance (3 bins defined, replace this), TaxiIn, TaxiOut, Cancelled, Diverted, CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay

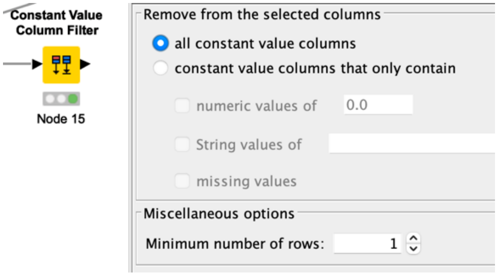
Distance

```
short : ] -∞ ... 1000,0 [
medium : [ 1000,0 ... 3000,0 [
haul : [ 3000,0 ... ∞ [
```

haul : . 3000, .. ∞ .

□ Append new column Distance_binned

10. Remove columns containing only constant values (**Constant Value Column Filter node)**



Remove from the selected columns

● all constant value columns

○ constant value columns that only contain

□ numeric values of 0.0

□ String values of

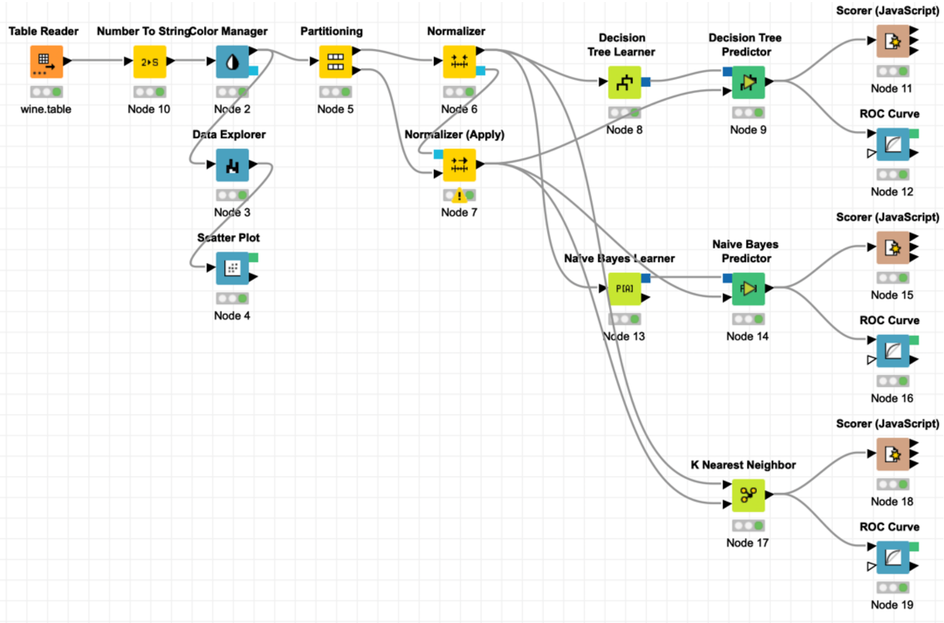
□ missing values

Miscellaneous options

Minimum number of rows: 1

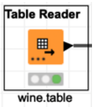
2. Wine data classification exercise

-Chemical properties of 178 wines are examined, resulting in 13 numerical features.
-There are 3 different types of wines in this data set, described by the column Type.
-Goal of this analysis to classify these wines based on their features.



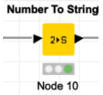
1. Reading the data set

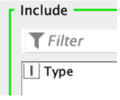
-Read the file "*wine.table*" with the **Table Reader node**



Advanced Settings → nichts angekreuzt

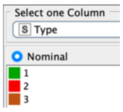
-**Number to String node**




Include
Filter
Type

2. Explore the data

-Use the **Color Manager node** to assign colors to different classes of the target variable Type

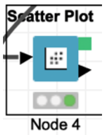

Select one Column
S Type

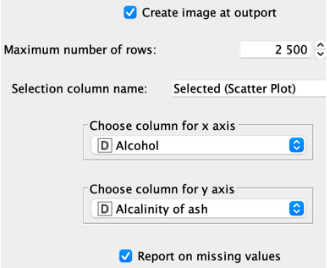
● Nominal

🟩 1
🟥 2
🟧 3

-Use the **Data Explorer node** to examine statsitics and distributions of the features

-Use the **Scatterplot node** to plot various attributes




☑ Create image at outport

Maximum number of rows: 2 500

Selection column name: Selected (Scatter Plot)

Choose column for x axis: D Alcohol

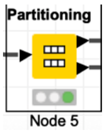
Choose column for y axis: D Alcalinity of ash

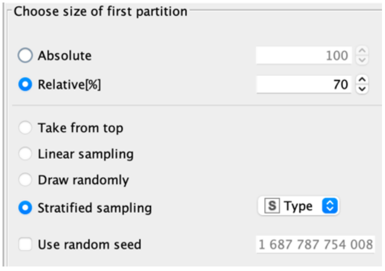
☑ Report on missing values

3. Partitioning

-Use the **Partitioning node** to split the data set into the training (70%) & testing (30%) data sets

-Stratified sampling




Choose size of first partition

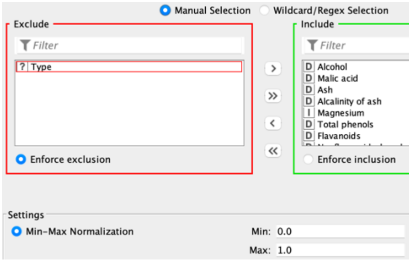
○ Absolute 100
● Relative[%] 70

○ Take from top
○ Linear sampling
○ Draw randomly
● Stratified sampling S Type

□ Use random seed 1 687 787 754 008

4. Normalization

-For the training data, normalize numerical features to the range of [0,1] with the **Normalizer node**



● Manual Selection ○ Wildcard/Regex Selection

Exclude

Filter

? Type

● Enforce exclusion

Include

Filter

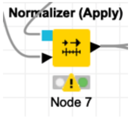
D Alcohol
D Malic acid
D Ash
D Alcalinity of ash
I Magnesium
D Total phenols
D Flavanoids

○ Enforce inclusion

Settings

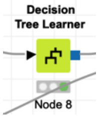
● Min-Max Normalization Min: 0.0 Max: 1.0

-Apply the normalization from the training data to the testing data with the **Normalizer (Apply) node**

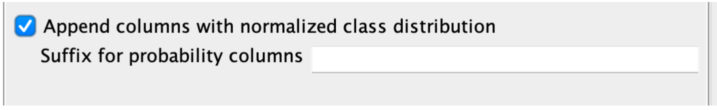


5. Train and apply a decision tree classification model

-Train a decision tree model with the **Decision Tree Learner node**.



-Apply the trained model to the testing data with **Decision Tree Predictor**

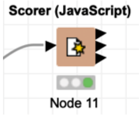


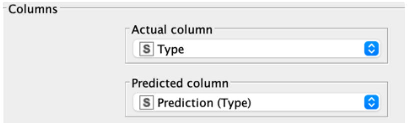
☑ Append columns with normalized class distribution

Suffix for probability columns

-Make sure to output class probabilities

-Evaluate the model performance with the **Scorer (JavaScript) node**



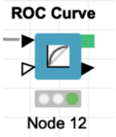


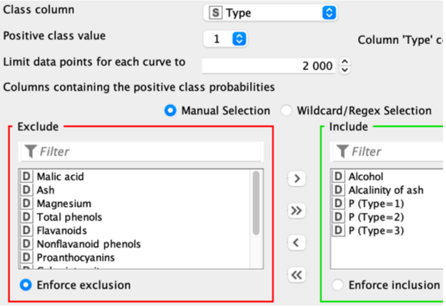
Columns

Actual column: S Type

Predicted column: S Prediction (Type)

-Plot the ROC curve





Class column: S Type

Positive class value: 1

Column 'Type' c

Limit data points for each curve to: 2 000

Columns containing the positive class probabilities

● Manual Selection ○ Wildcard/Regex Selection

Exclude

Filter

D Malic acid
D Ash
D Magnesium
D Total phenols
D Flavanoids
D Nonflavanoid phenols
D Proanthocyanins

● Enforce exclusion

Include

Filter

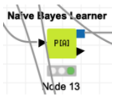
D Alcohol
D Alcalinity of ash
D P (Type=1)
D P (Type=2)
D P (Type=3)

○ Enforce inclusion

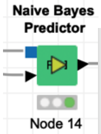
-Adjust parameters of the Decision Tree Learner to improve the classifier performance

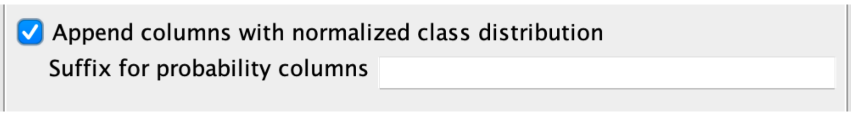
6. Train and apply a Naive Bayes classification model

-Train a naive Bayes model with the **Naive Bayes Learner node**.



-Apply the trained model to the testing data with **Naive Bayes Predictor**



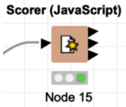


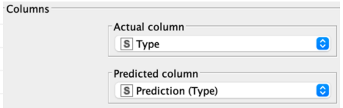
☑ Append columns with normalized class distribution

Suffix for probability columns

-Make sure to output class probabilities

-Evaluate the model performance with the **Scorer (JavaScript) node**



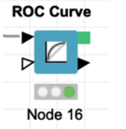


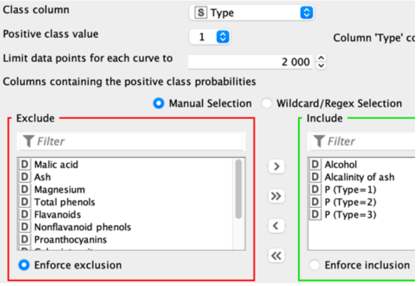
Columns

Actual column: S Type

Predicted column: S Prediction (Type)

-Plot the **ROC curve**





Class column: S Type

Positive class value: 1

Column 'Type' c

Limit data points for each curve to: 2 000

Columns containing the positive class probabilities

● Manual Selection ○ Wildcard/Regex Selection

Exclude
- D Malic acid
- D Ash
- D Magnesium
- D Total phenols
- D Flavanoids
- D Nonflavanoid phenols
- D Proanthocyanins

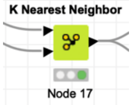
● Enforce exclusion

Include
- D Alcohol
- D Alcalinity of ash
- D P (Type=1)
- D P (Type=2)
- D P (Type=3)

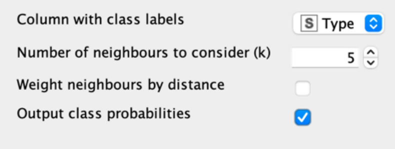
○ Enforce inclusion

7. kNN classification model

-Apply the kNN classification model with the **K Nearest Neighbor node**



-Set k=5. Use the training data set as the model.



Column with class labels: S Type

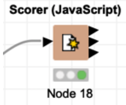
Number of neighbours to consider (k): 5

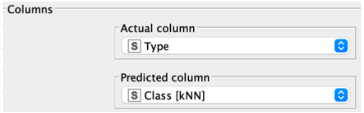
Weight neighbours by distance: □

Output class probabilities: ☑

-Make sure to output class probabilities

-Evaluate the model performance with the **Scorer (JavaScript) node**



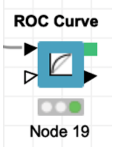


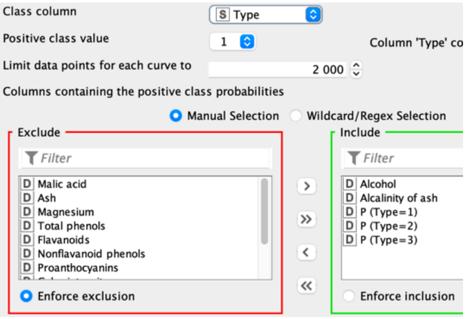
Columns

Actual column: S Type

Predicted column: S Class [kNN]

-Plot the ROC curve





Class column: S Type

Positive class value: 1

Column 'Type' co

Limit data points for each curve to: 2 000

Columns containing the positive class probabilities

● Manual Selection ○ Wildcard/Regex Selection

Exclude
- D Malic acid
- D Ash
- D Magnesium
- D Total phenols
- D Flavanoids
- D Nonflavanoid phenols
- D Proanthocyanins

● Enforce exclusion

Include
- D Alcohol
- D Alcalinity of ash
- D P (Type=1)
- D P (Type=2)
- D P (Type=3)

○ Enforce inclusion

-Adjust parameters of the K Nearest Neighbor node to improve the classifier performance