University of Minho
School of Engineering

ISLab

Machine Learning and Decision-Making

ADI @ LEI/3º, MiEI/4º - 2º Semestre

Filipe Gonçalves, Inês Alves, Cesar Analide

Part VIII – April 2022

Contents

- Model Validation Techniques
- Feature Selection
- Hands On

Validation Techniques

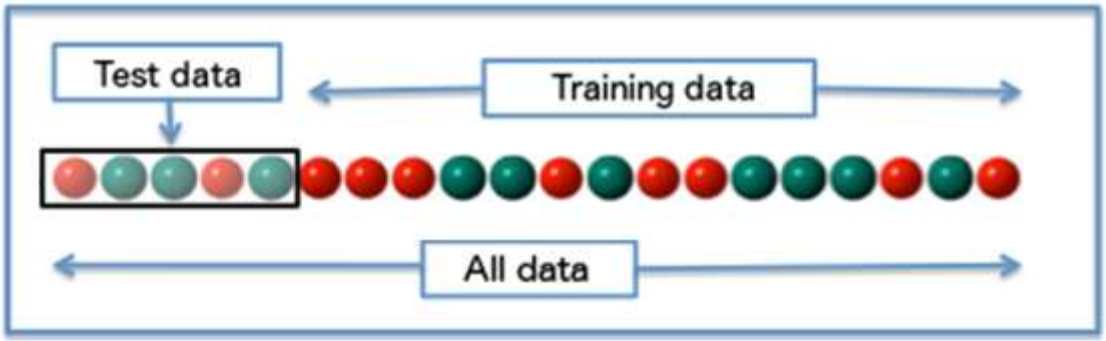
Cross Validation

Hold-out Validation

But before going into Cross Validation, a model validation technique, do you know you have already been using another model validation technique?

It is known as Hold-Out Validation!

In essence, it means we validate the model on unseen data, i.e., we use a "partitioning method" to split the learning and the testing data once. This means we hold-out a subset of data for testing (80/20; 75/25; 65;35…)!



Cross Validation

Cross validation is another model validation technique!

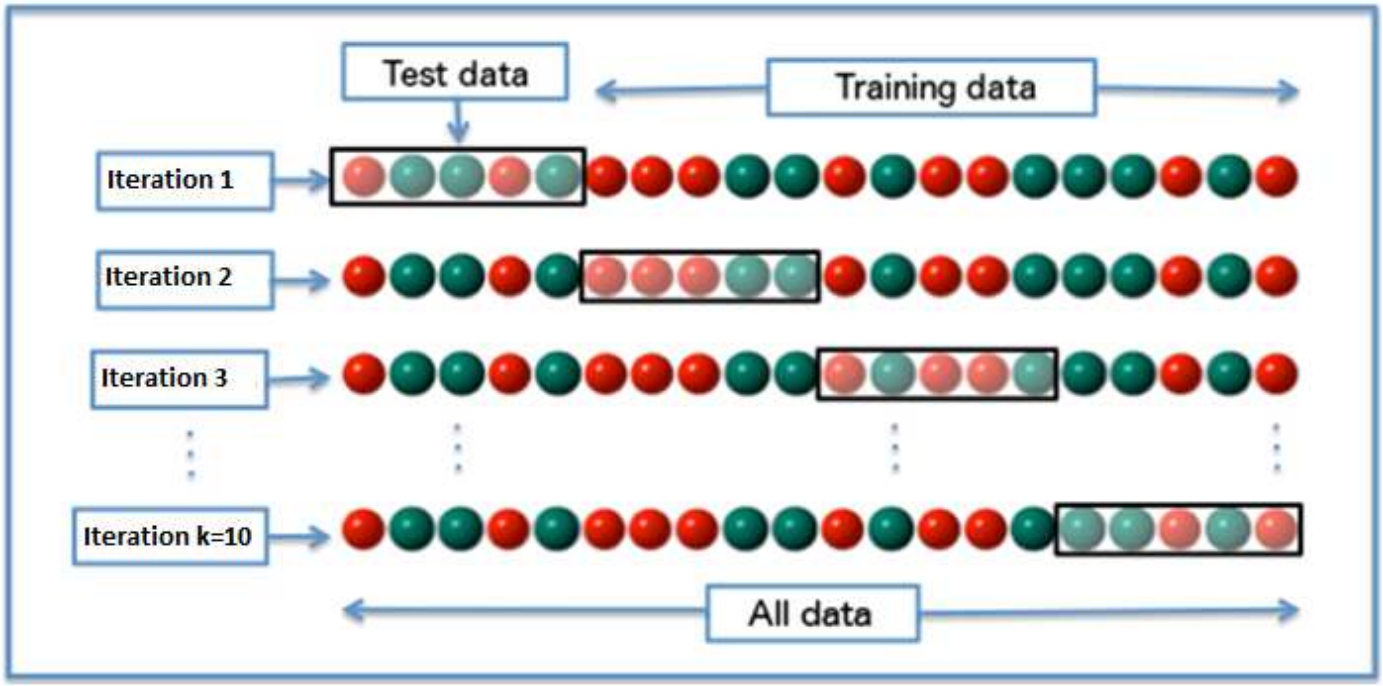
The goal is to have an accurate metric of how the model will perform in practice.

In essence, it consists in dividing the dataset into k folds. In each run of the model, k-1 folds are used for training and 1 fold (the remaining) is used as test. Keep repeating the process until all folds have been used for testing.

The final error metric is based on the mean value of all error metrics.

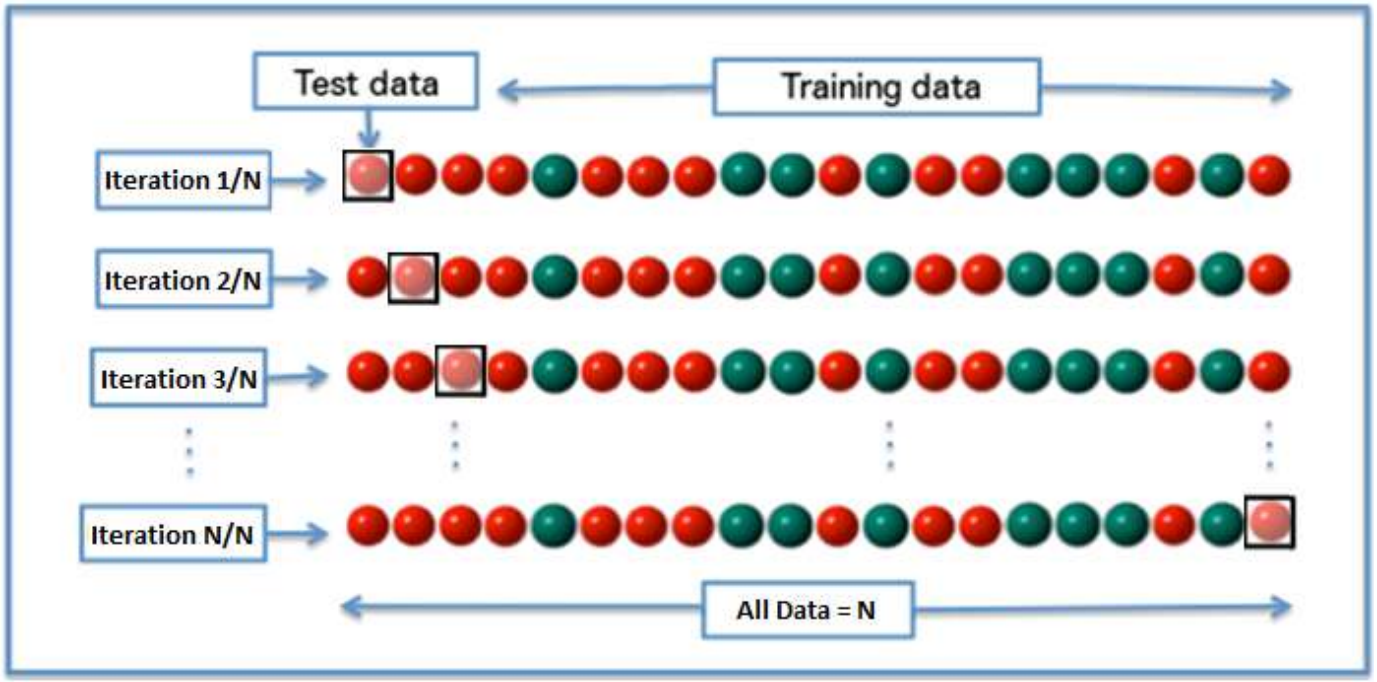
$$E = \frac{1}{k}\sum_{i=1}^{k} E_i$$

k-fold Cross Validation



Usually, k=10.

Leave-one-out Cross Validation (k=N)



The special case of having k=N. Expensive…
But a good approach when we have a small dataset.

Cross Validation - How many folds?

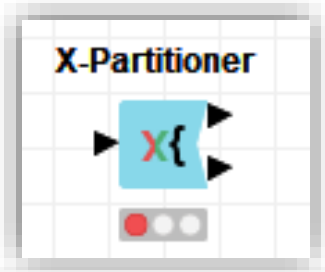
Well, …

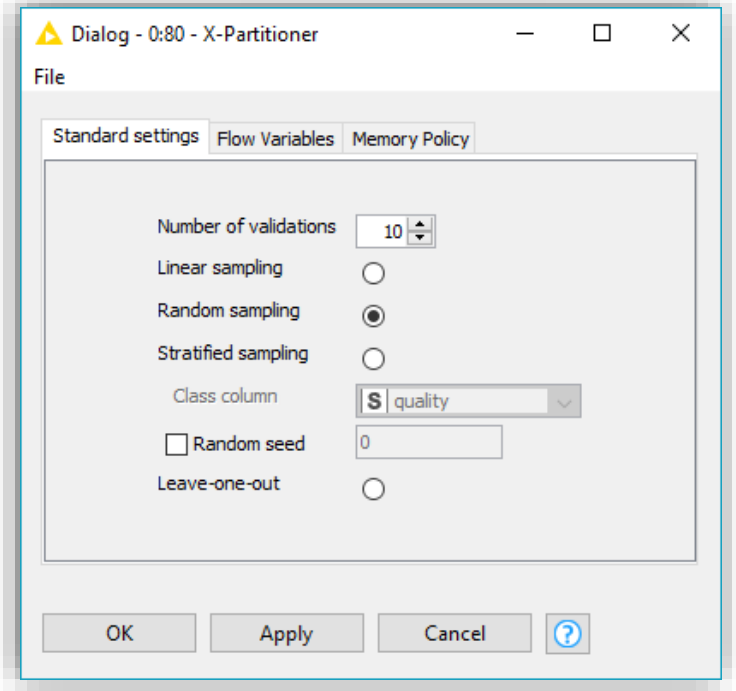
A greater number of folds will lead to a better error estimate of the model, a lower bias and less overfitting! However, it comes with a higher computational cost!

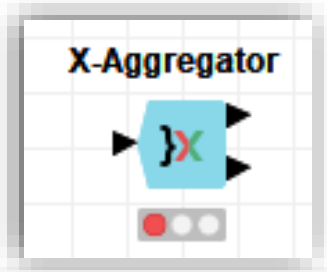
If we have a large dataset, a smaller k may be enough since we will have a larger amount of data for training. If we have a small dataset, we may want to use leave-one-out cross validation to maximize the amount of data for training…

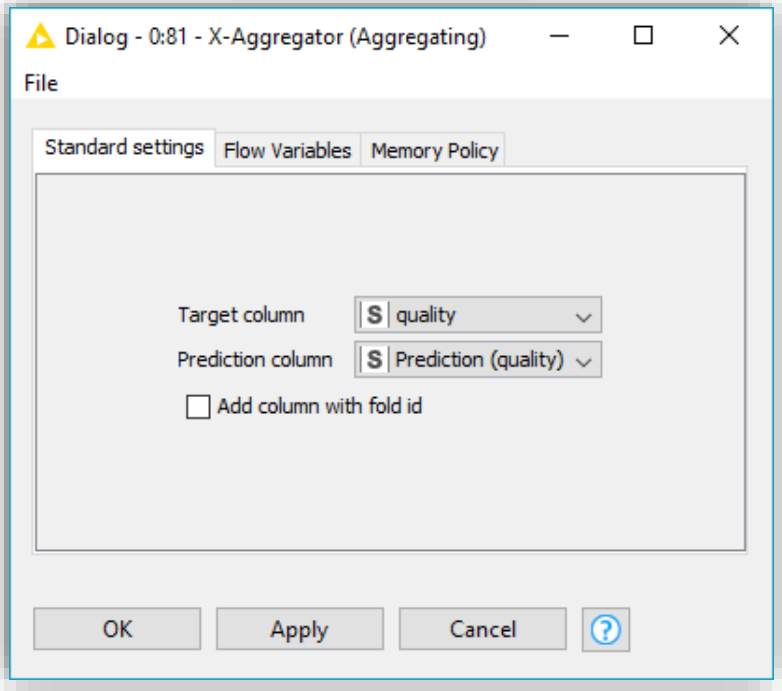
In reality, k depends on N!!
Rule of thumb –> k=10!

k-fold Cross Validation









k-fold Cross Validation

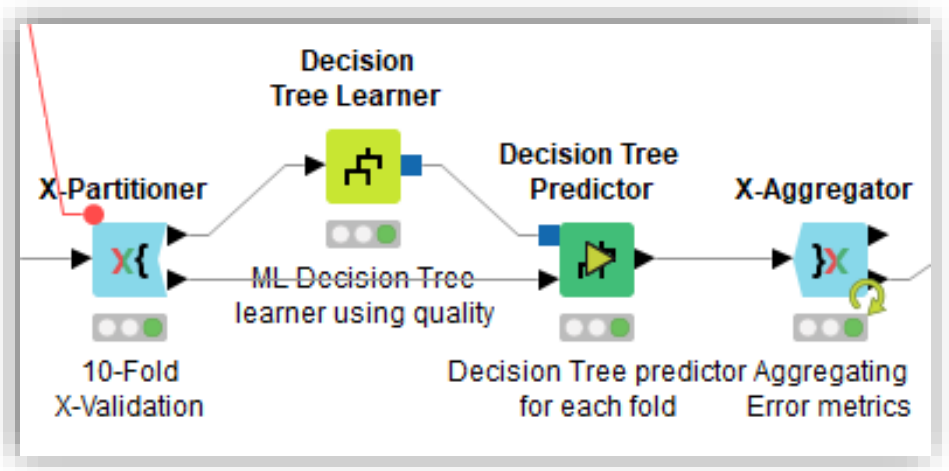
Error rates - 0:81 - X-Aggregator (Aggregating)

File Hilite Navigation View

Table "default" - Rows: 10 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	D Error in %	I Size of Test Set	I Error Count
fold 0	38.281	128	49
fold 1	35.156	128	45
fold 2	44.531	128	57
fold 3	42.969	128	55
fold 4	42.969	128	55
fold 5	41.406	128	53
fold 6	41.406	128	53
fold 7	40.625	128	52
fold 8	41.406	128	53
fold 9	42.52	127	54

k-fold Cross Validation



Feature Selection

(dimensionality reduction)

Feature Selection

Feature Selection (or dimensionality reduction)

Rationale:

Which features should we use to create a predictive model? Select a sub-set of the most important features to reduce dimensionality.

The removal of unimportant features:

- May affect significantly the performance of a model
- Reduces overfitting (less opportunity to make decisions based on noise)
- Improves accuracy
- Helps reducing the complexity of a model (reduces training time)

Feature Selection

Feature Selection (or dimensionality reduction)

Rationale:

Which features should we use to create a predictive model? Select a sub-set of the most important features to reduce dimensionality.

The removal of unimportant features:

- May affect significantly the performance of a model
- Reduces overfitting (less opportunity to make decisions based on noise)
- Improves accuracy
- Helps reducing the complexity of a model (reduces training time)

What can we remove:

- Redundant features (duplicate)
- Irrelevant and unneeded features (non-useful)

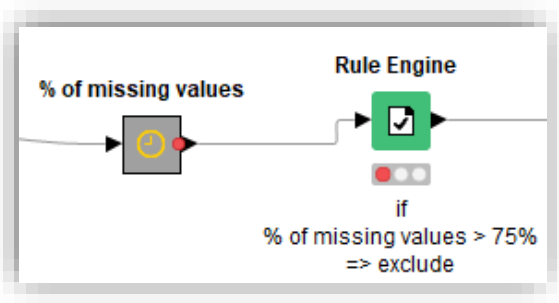
Feature Selection Methods:

- Filter methods
- Wrapper methods
- Embedded methods

Feature Selection

Filter Methods:

i. Remove a feature if the percentage of missing values is higher than a threshold



ii. Use the chi-square test to measure the degree of dependency between a feature and the target class
   - For each feature calculate X^2
   - Normalize X^2 and sort in descending order
   - Select *n* features with the highest importance (or those that are above the threshold)

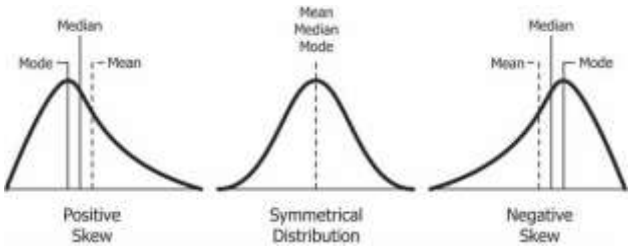
Feature Selection

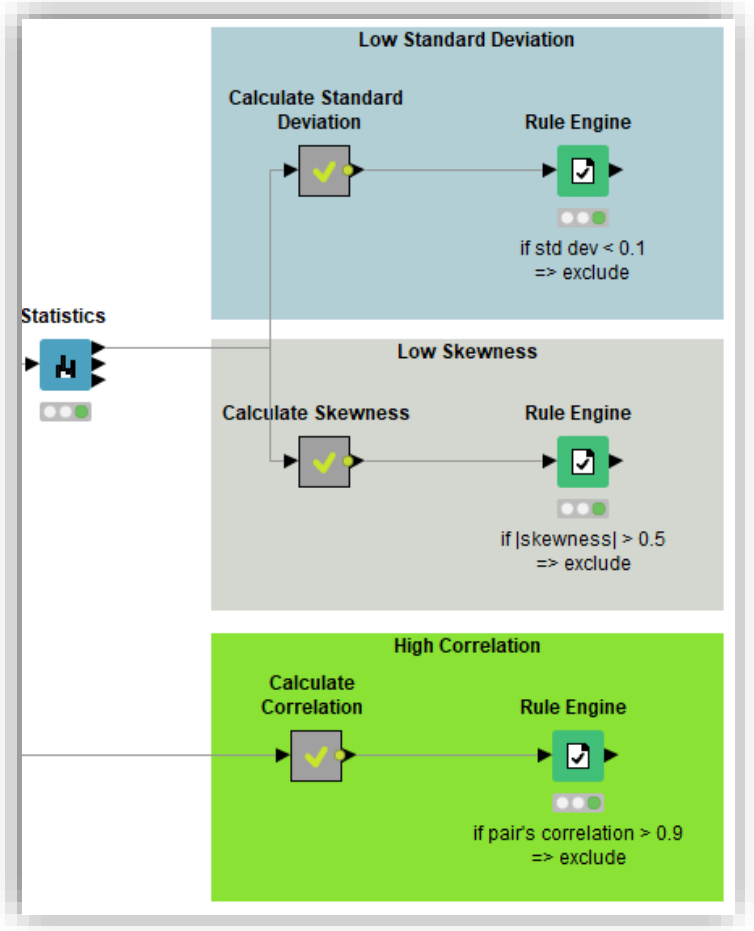
Filter Methods:

iii. Remove feature if low standard deviation

iv. Remove feature if data are highly skewed

v. Remove features that are highly correlated between each other



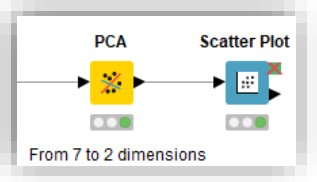


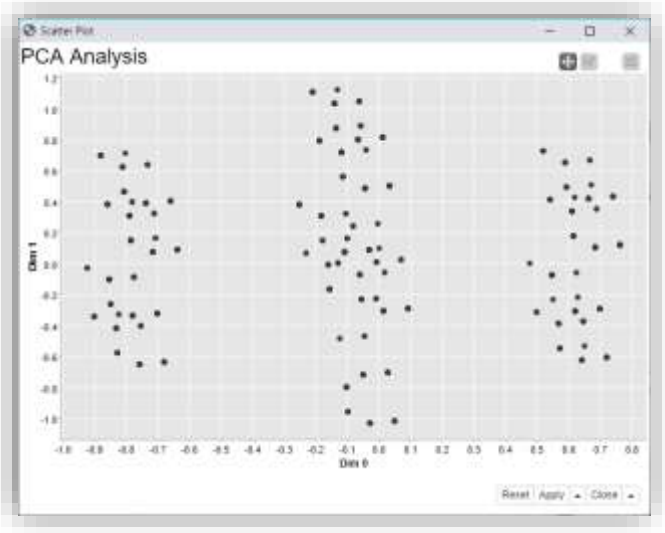
Feature Selection

Filter Methods:

vi. Principal Component Analysis (PCA)

A technique to reduce the dimension of the feature space. The goal is to reduce the number of features without losing too much information. A popular application of PCA is for visualizing higher dimensional data.



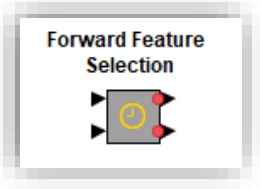


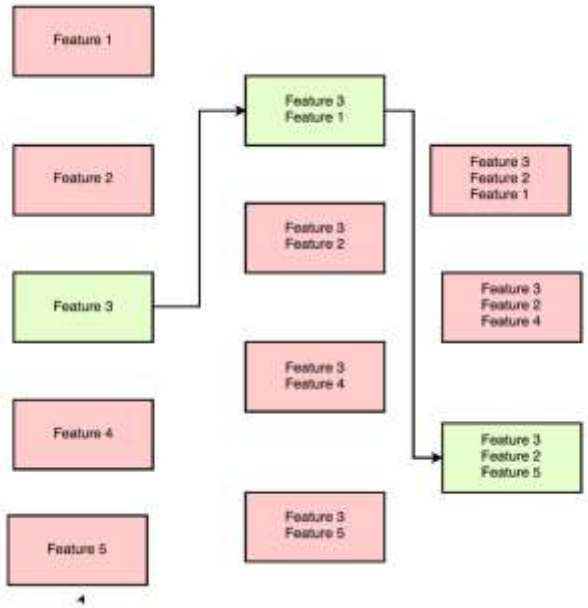
Feature Selection

Wrapper Methods:

Use a ML algorithm to select the most important features! Select a set of features as a search problem, prepare different combinations, evaluate and compare them!
Measure the “usefulness” of features based on the classifier performance.

Sequential Forward Selection



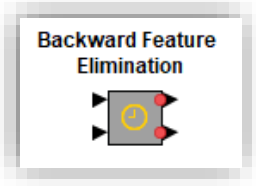


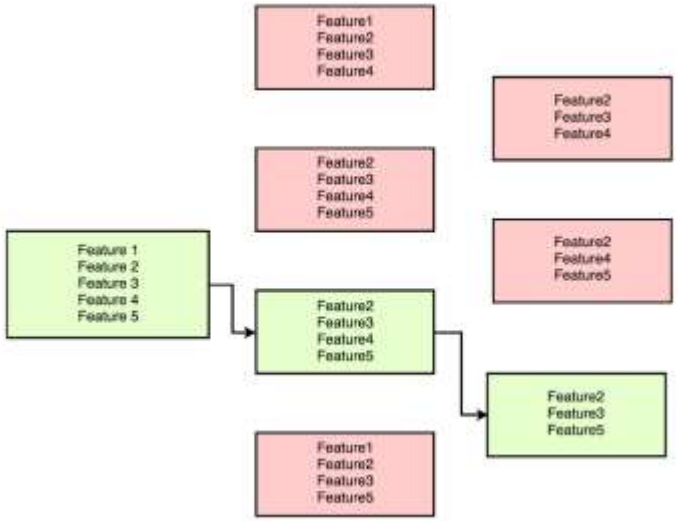
Feature Selection

Wrapper Methods:

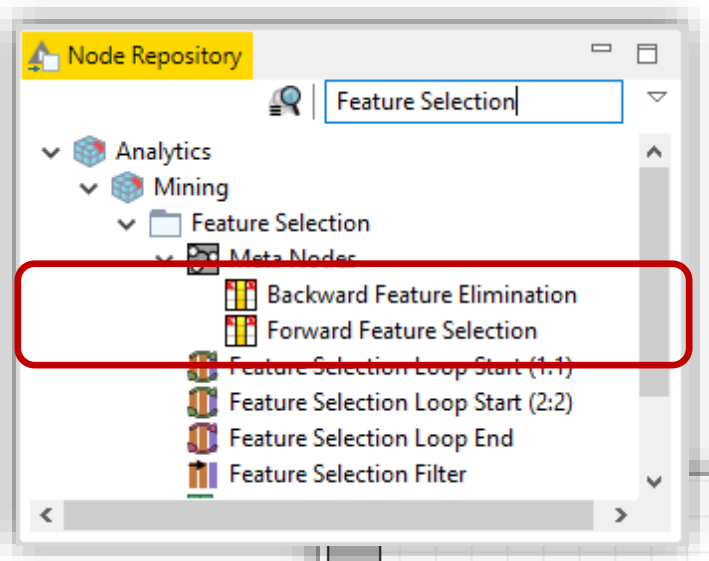
Use a ML algorithm to select the most important features! Select a set of features as a search problem, prepare different combinations, evaluate and compare them!
Measure the “usefulness” of features based on the classifier performance.

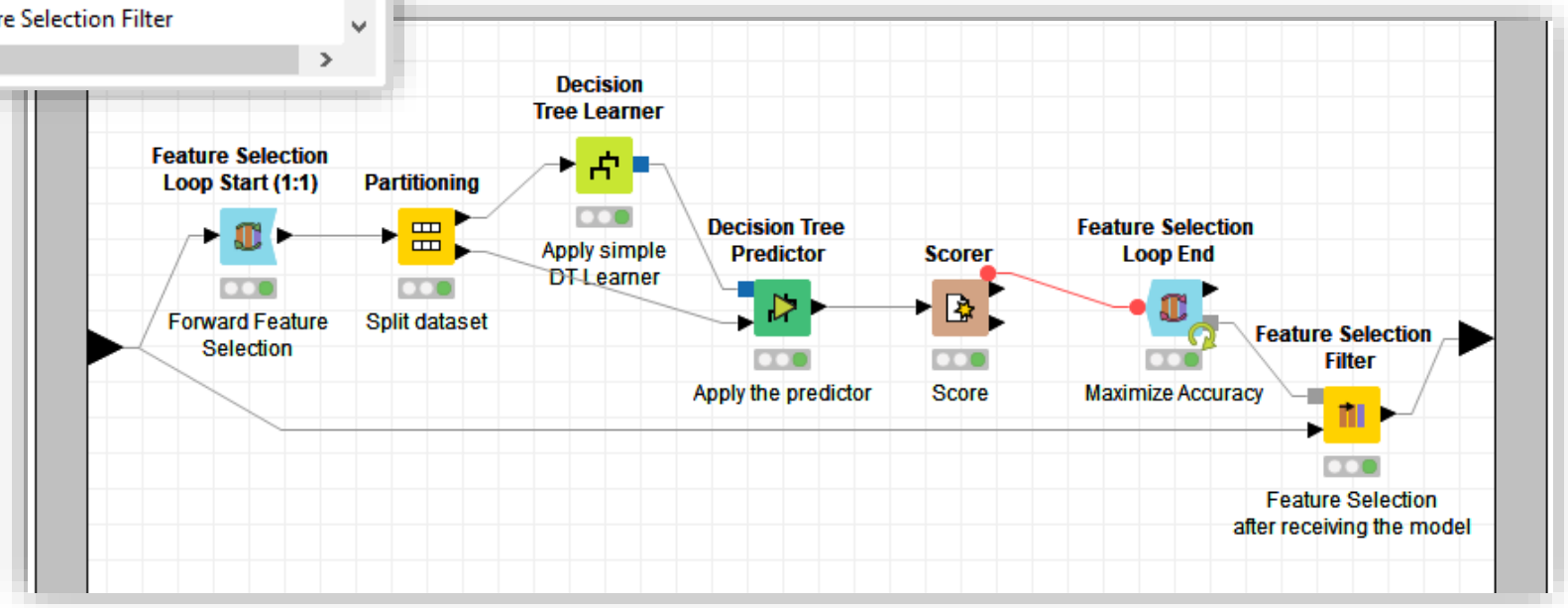
Backward Feature Elimination



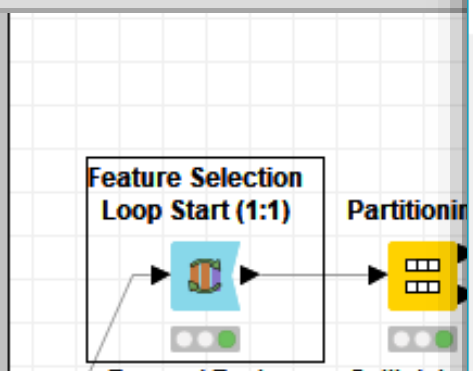


Feature Selection - Wrapper Methods





Feature Selection Loop Start



Dialog - 0:96:95 - Feature Selection Loop Start (1:1) (Forward Feature)

File

Options | Flow Variables | Memory Policy

The list on the left contains 'static' columns such as the target column.
The columns to choose from need to be in the list on the right.

Manual Selection | Wildcard/Regex Selection

Static Columns

Filter

S quality

Enforce exclusion

Variable Columns ('Features')

Filter

D fixed acidity
D volatile acidity
D citric acid
D residual sugar
D chlorides
D free sulfur dioxide
D total sulfur dioxide
D pH

Enforce inclusion

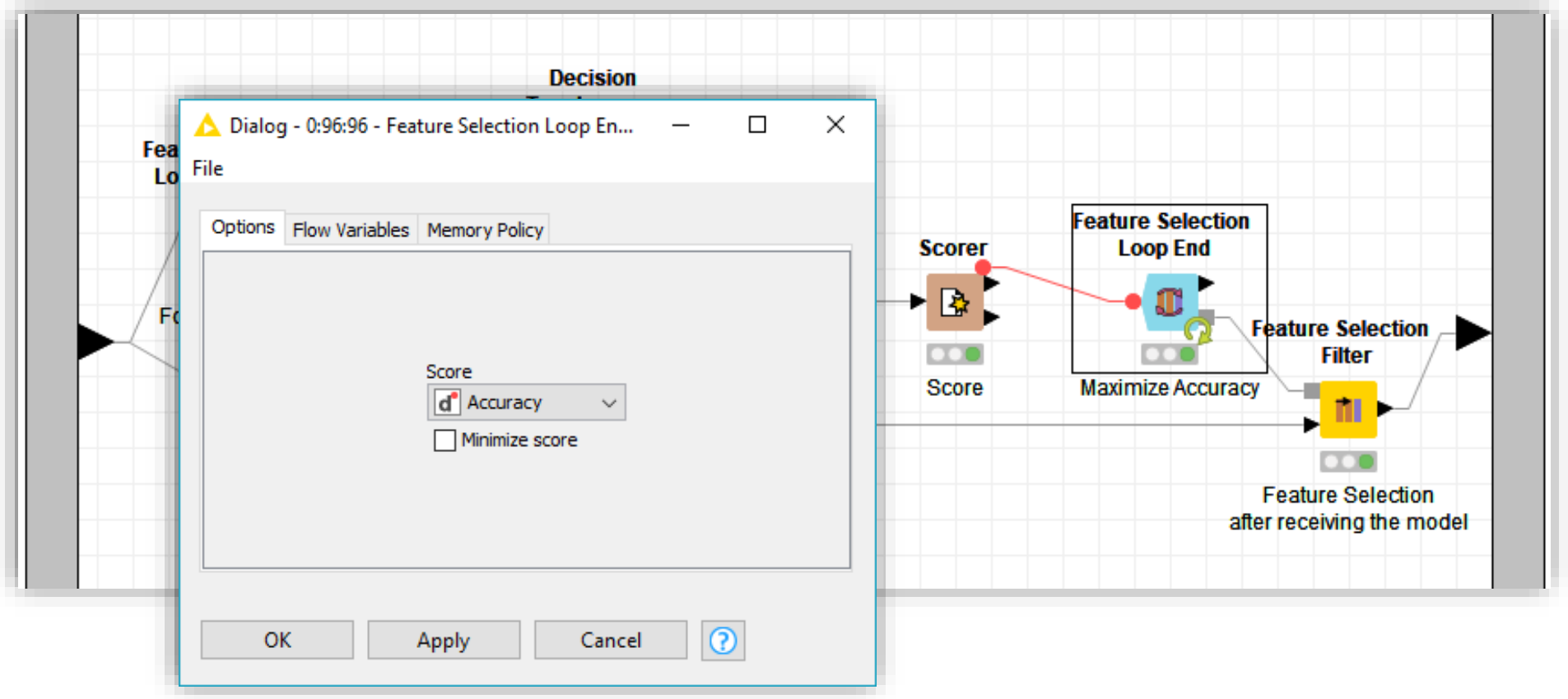
Feature selection strategy: Forward Feature Selection

Use threshold for number of features

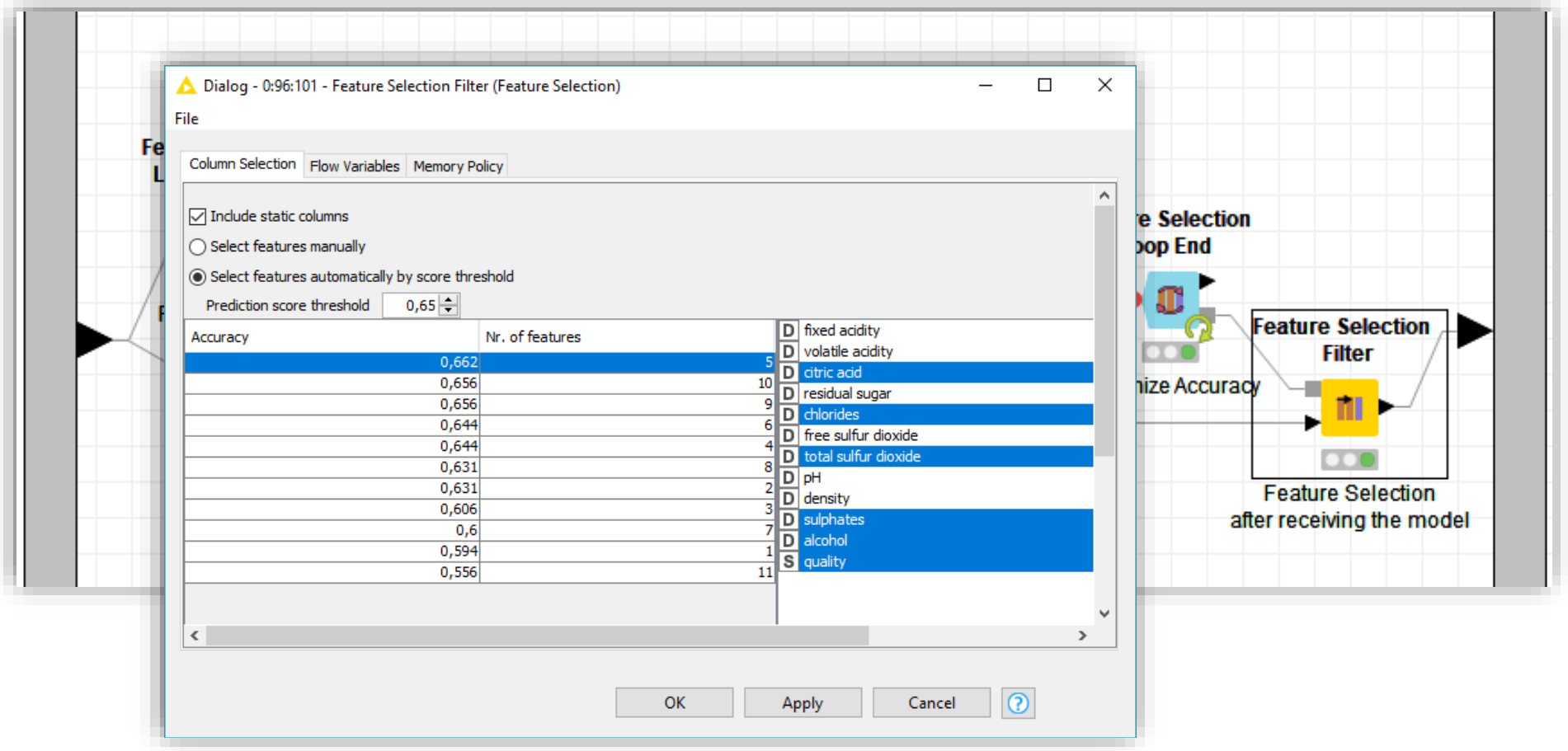
Select threshold for number of features: 20

OK | Apply | Cancel

Feature Selection Loop End



Feature Selection Filter



Hands On


KNIME Analytics Platform
File Edit View Node Help
KNIME Explorer
My-KNIME-Hub (hub.knime.com)
EXAMPLES (knime@hub.knime.com)
LOCAL (Local Workspace)
Aulas Exemplos
Aulas Exercícios
Example Workflows
ML Projects
KNIME_Aula_Exercicio
Workflow Coach
Recommended Nodes
File Reader
CSV Reader
Excel Reader (XLS)
Table Creator
Database Reader (legacy)
Table Reader
List Files
Node Repository
IO
Manipulation
Views
Analytics
DB
Other Data Types
Structured Data
Scripting
Tools & Services
Workflow Control
Workflow Abstraction
Reporting
KNIME_Aula_Exercicio
Description
KNIME Hub Search
Console
Outline
HANDS ON