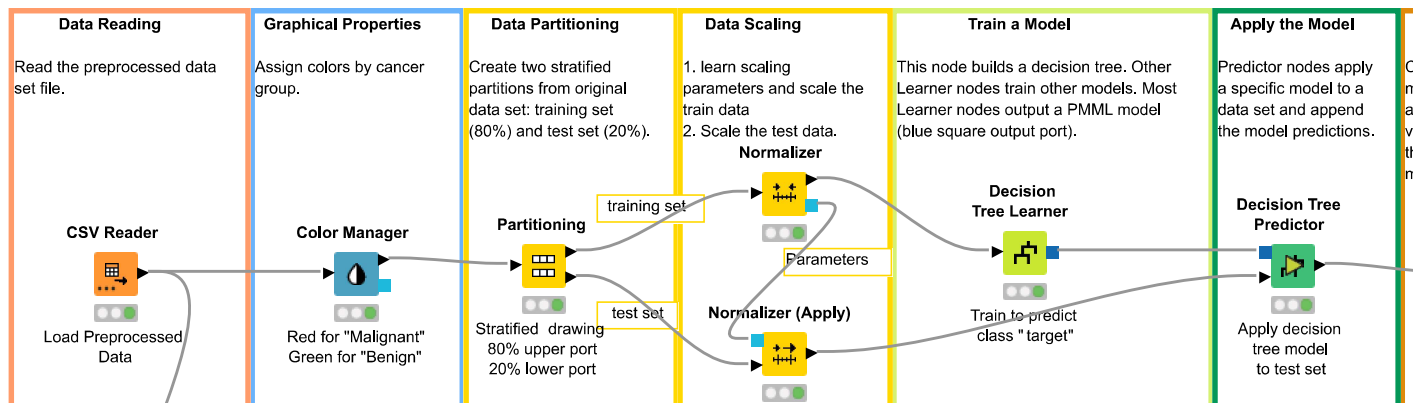# Model Training for Classification
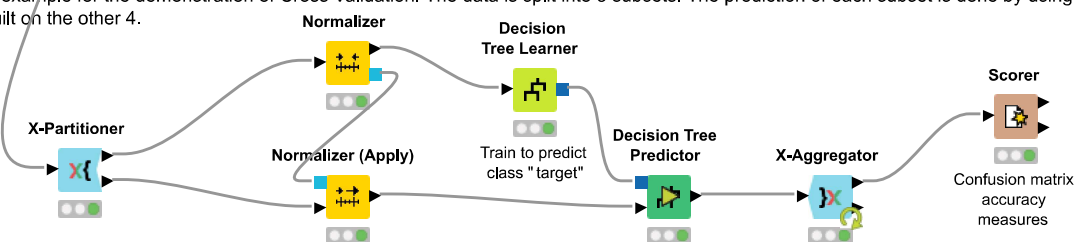
This workflow demonstrates how a simple classifier is built and applied to new data.
Task Classification of Benign and Malignant Breast Cancer

Data Reading

Read the preprocessed data set file.

CSV Reader

Load Preprocessed Data

Graphical Properties

Assign colors by cancer group.

Color Manager

Red for "Malignant"
Green for "Benign"

Data Partitioning

Create two stratified partitions from original data set: training set (80%) and test set (20%).

Partitioning

Stratified drawing
80% upper port
20% lower port

training set

test set

Data Scaling

1. learn scaling parameters and scale the train data
2. Scale the test data.

Normalizer

Parameters

Normalizer (Apply)

Train a Model

This node builds a decision tree. Other Learner nodes train other models. Most Learner nodes output a PMML model (blue square output port).

Decision Tree Learner

Train to predict class "target"

Apply the Model

Predictor nodes apply a specific model to a data set and append the model predictions.

Decision Tree Predictor

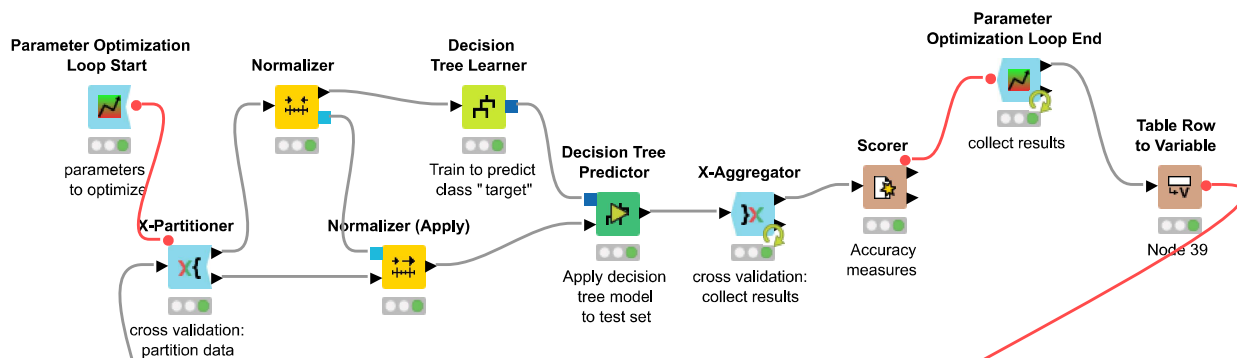Apply decision tree model to test set

Cross Validation with Decision Tree

A simple example for the demonstration of Cross Validation. The data is split into 5 subsets. The prediction of each subset is done by using an Dtree model built on the other 4.
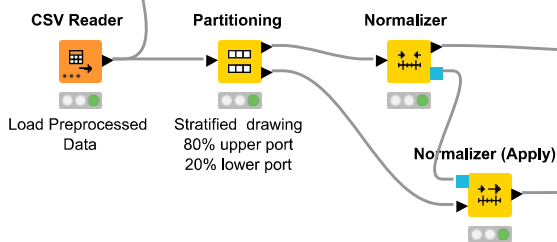
X-Partitioner

Normalizer

Normalizer (Apply)

Decision Tree Learner

Train to predict class "target"

Decision Tree Predictor

X-Aggregator

Scorer

Confusion matrix accuracy measures

Cross ValidaCross Validation with Decistion Tree and Parameter Optimisation

A simple example for the demonstration of Cross Validation combined with Parameter Optimization. The data is split into 5 subsets. The prediction of each subset is done by using an Decision Tree model built on the other 4. The min_samples_node parameter for the Decision Tree Learner is varied and controlled by a flow variable. The Parameter Optimisation Loop Start node varies the value of the flow variable for each iteration. The Parameter Optimization Loop End node collects the results and output the value for the optimized parameter yelding maximum accuracy as an objective function.

Parameter Optimization Loop Start

parameters to optimize

Normalizer

Decision Tree Learner

Train to predict class "target"

X-Partitioner

cross validation: partition data

Normalizer (Apply)

Decision Tree Predictor

Apply decision tree model to test set

X-Aggregator

cross validation: collect results

Scorer

Accuracy measures

Parameter Optimization Loop End

collect results

Table Row to Variable

Node 39

Load Data and Pre-Processing

CSV Reader

Load Preprocessed Data

Partitioning

Stratified drawing
80% upper port
20% lower port

Normalizer

Normalizer (Apply)

Evaluation Fine-Tuned Decision Tree Model

Decision Tree Learner

Train to predict class "target"

Decision Tree Predictor

Apply decision tree model to test set

Scorer

Confusion matrix accuracy measures