U4 MACHINE LEARNING IN PRACTICE

U4.E1 SOLVING TYPICAL MACHINE LEARNING PROBLEMS

Machine Learning Engineer

January 2021, Version 1

The Development and Research on Innovative Vocational Educational Skills project (DRIVES) is co-funded by the Erasmus+ Programme of the European Union under the agreement 591988-EPP-1-2017-1-CZ-EPPKA2-SSA-B. The European Commission support for the production of this publication does not constitute endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

LEARNING OBJECTIVES

The student is able to

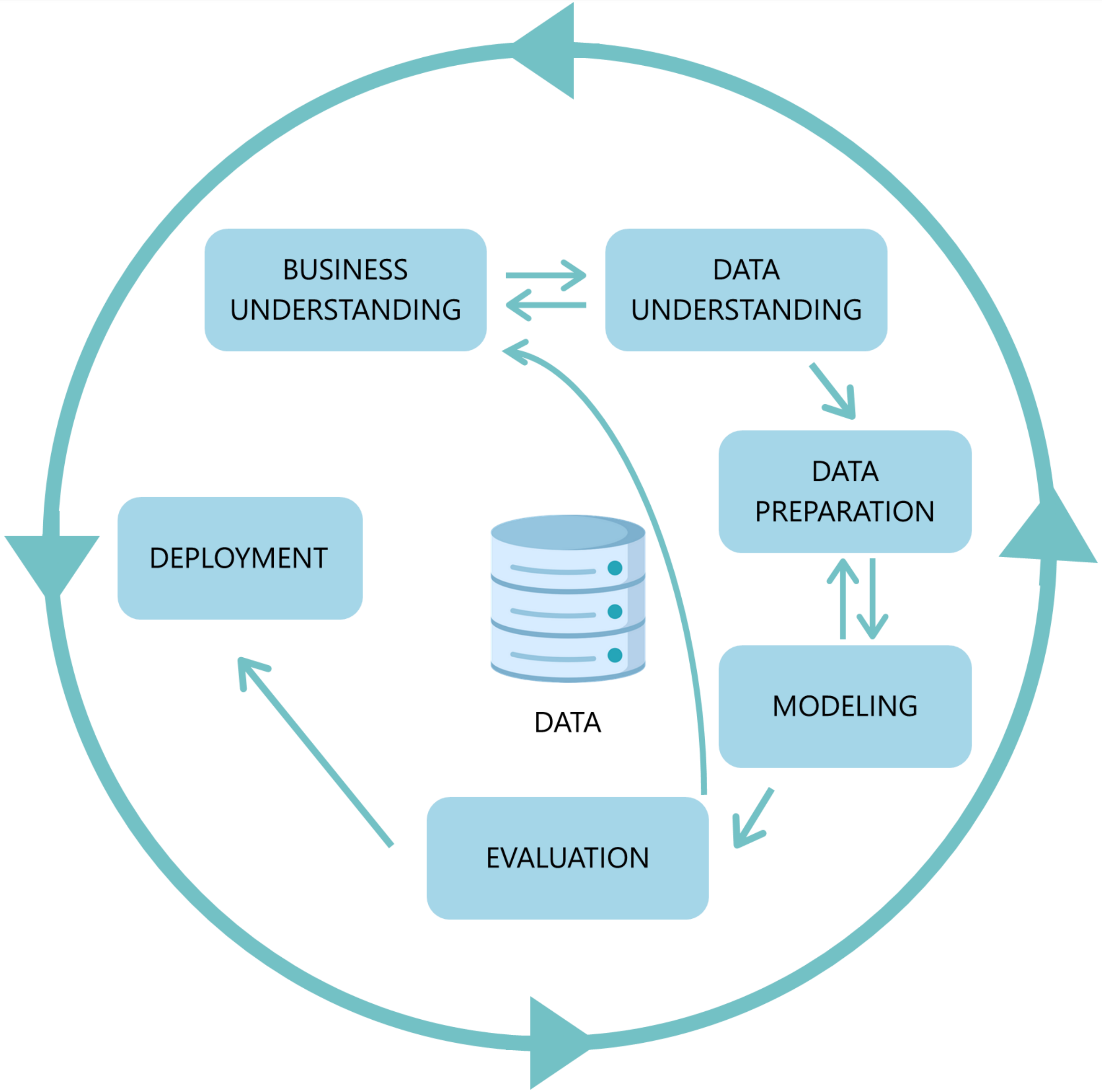
MLE.U4.E1.PC1	Recognize typical machine learning problems and its areas of intervention.
MLE.U4.E1.PC2	Know and understand the steps of the resolution approach that typically apply to machine learning problems.
MLE.U4.E1.PC3	Explore and understand the dataset as well as the main goal of the project.
MLE.U4.E1.PC4	Perform data cleaning, pre-processing and transformation.
MLE.U4.E1.PC5	Know which machine learning model to use.

LEARNING OBJECTIVES

The student is able to

MLE.U4.E1.PC6	Explore and interpret the results as well as evaluate the performance of the model.
MLE.U4.E1.PC7	Analyse and explore step by step the resolution of some machine learning problems.
MLE.U4.E1.PC8	Recognize the challenges surrounding machine learning approaches.

CRISP-DM LIFECYCLE



BUSINESS UNDERSTANDING

Business Understanding focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

1. Definition of the objectives in business terminology.

2. Definition of the objectives in technical terms.

3. Design a preliminary research plan.

1. Definition of the objectives in business terminology

- Understand client's needs and expectations;
- Uncover important factors (constraints, competing objectives);
- Identify the business units impacted by the project;
- Define business success criteria;
- Describe the problem in general terms regarding business questions and expected benefits.

DRIVES
Development and Research on Innovative Vocational Education Skills

2. Definition of the objectives in technical terms

- Identify knowledge souces and types;
- Identify software and hardware available;
- Describe relevant background;
- Translate the business questions into Data Mining goals;
- Specify the Data Mining problem type (classification, regression, clustering, etc.);
- Specify performance criteria for model assessment.

3. Design a preliminary research plan

- Define an initial process plan;
- Discuss its feasibility with involved personnel and stakeholders;
- Estimate efforts and resources;
- Identify challenges and critical steps.

Data Understanding begins with the initial data collection and proceeds with activities aimed at getting acquainted with the data, identifying problems with the quality of the data, discovering initial insights from the data or detecting interesting subsets to form hypotheses for hidden information.

1. Data acquisition.

2. Data analysis and exploration:
    - Understand the meaning of each attribute and its value in terms of business goal;
    - Analyse attribute types and ranges;
    - Compute basic statistics, such as distribution, average/mode and standard deviation, for each attribute;
    - Review the dataset's variability and assess the need to cover more cases;
    - Analyse properties of attributes and relations between them;
    - Identify data inconsistencies, duplicated instances, missing values and outliers.

MACHINE LEARNING SOFTWARES

rapidminer

Open for Innovation KNIME®

PYTORCH

WEKA
The University of Waikato

K Keras

TensorFlow

DRIVES
Development and Research on Innovative Vocational Education Skills

Waikato Environment for Knowledge Analysis (WEKA):

It is a software that allows large volumes of data to be pre-processed, different machine learning algorithms to be used and different outputs to be compared.

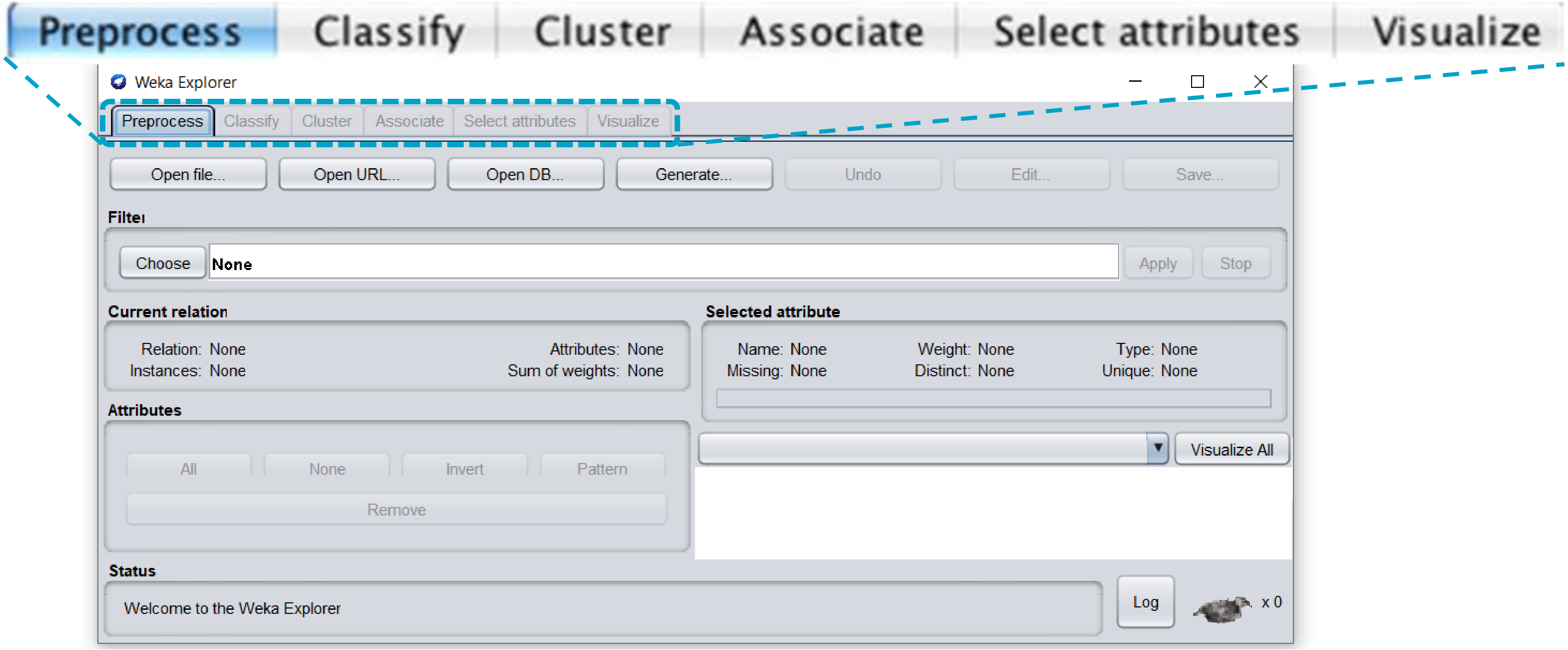
DOWNLOAD:

https://www.cs.waikato.ac.nz/ml/weka/downloading.html

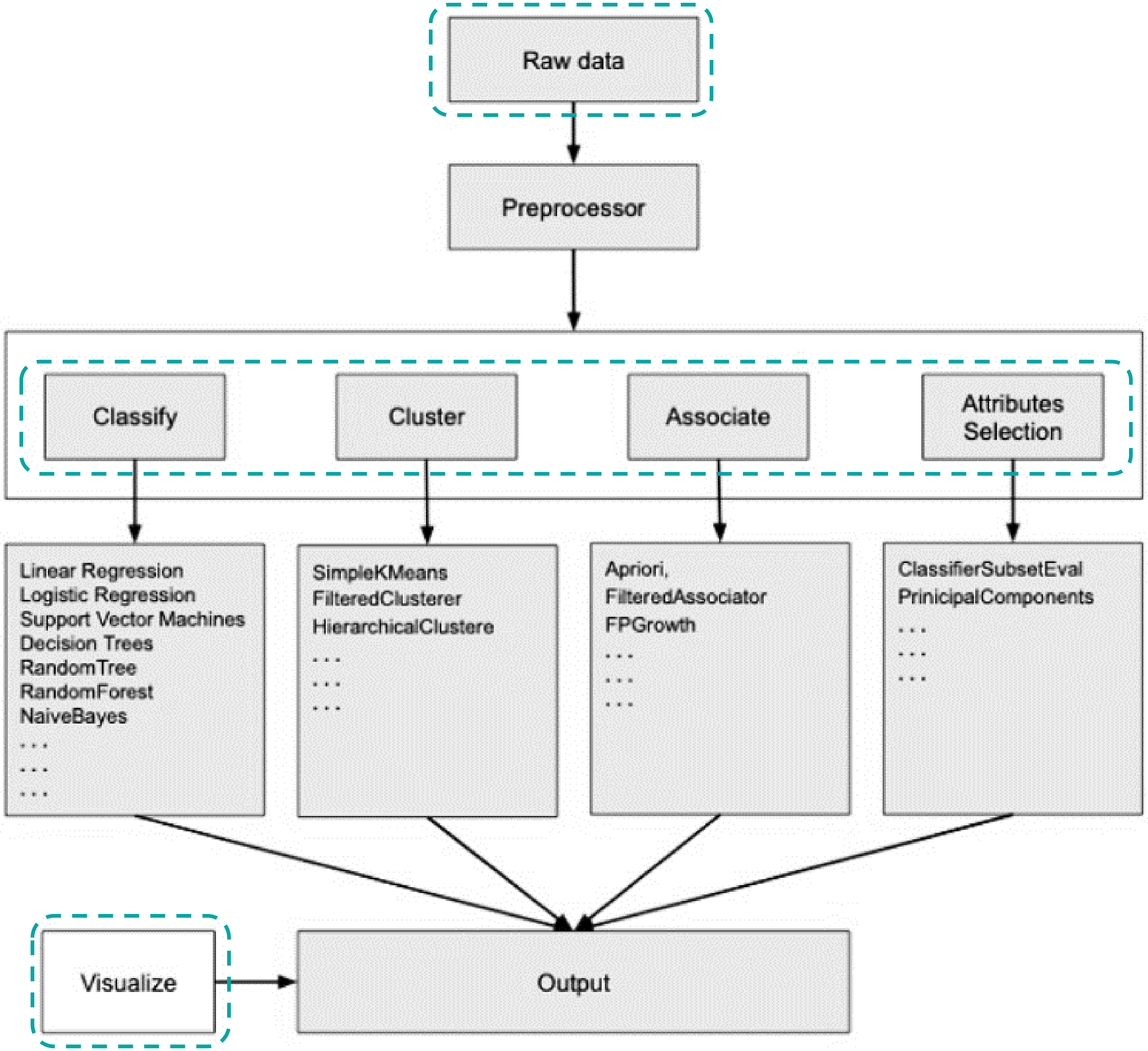
WEKA


Weka GUI Chooser
Program Visualization Tools Help
WEKA
The University of Waikato
Applications
Explorer
Experimenter
KnowledgeFlow
Workbench
Simple CLI
Waikato Environment for Knowledge Analysis
Version 3.8.4
(c) 1999 - 2019
The University of Waikato
Hamilton, New Zealand

When you open the Explorer tab, you will have access to 6 different tabs. Initially, only the preprocess tab is available since the first step in ML is data preprocessing.



WEKA

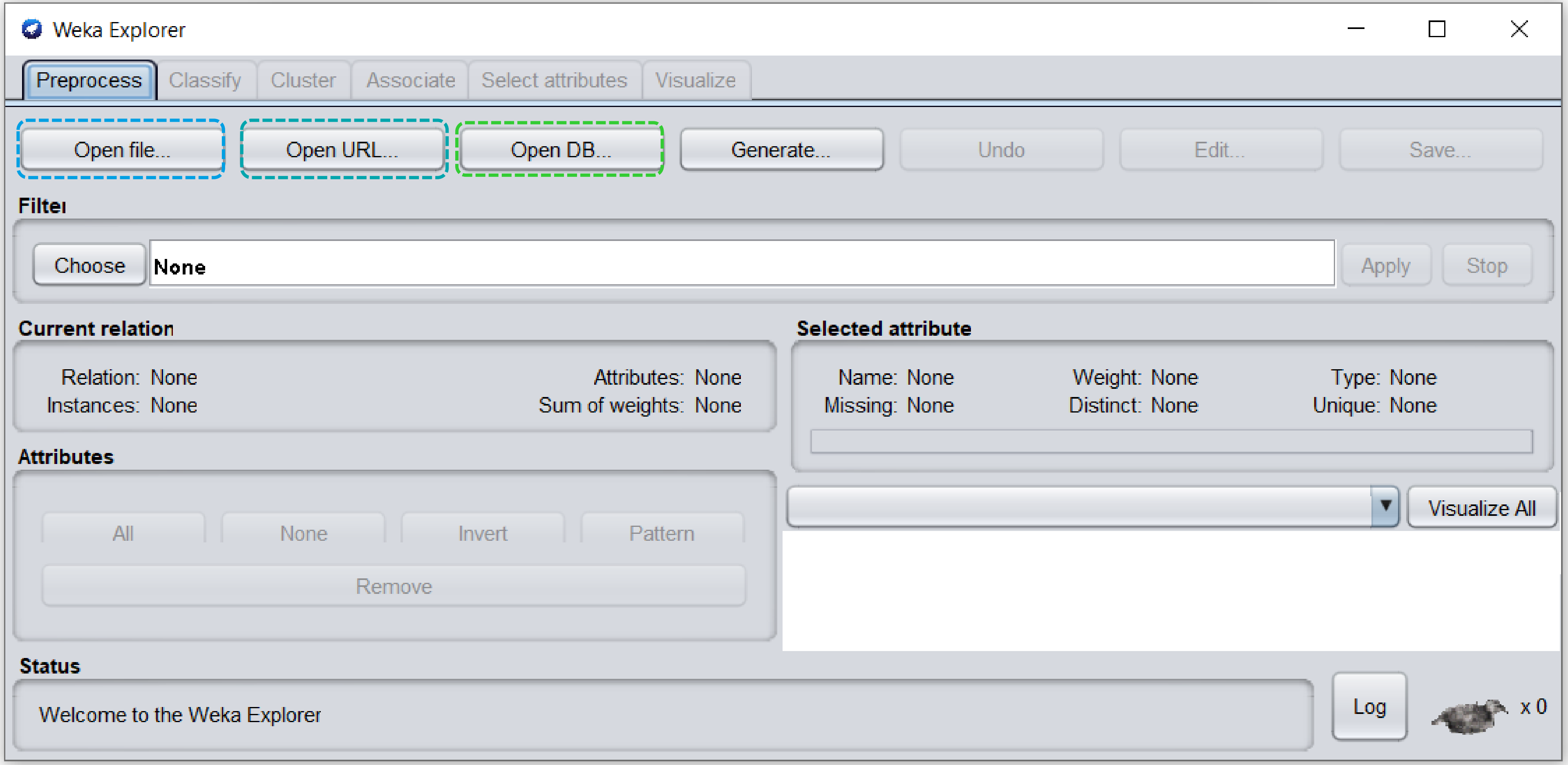


Raw data
Preprocessor
Classify
Cluster
Associate
Attributes Selection
Linear Regression
Logistic Regression
Support Vector Machines
Decision Trees
RandomTree
RandomForest
NaiveBayes
. . .
. . .
. . .
SimpleKMeans
FilteredClusterer
HierarchicalClustere
. . .
. . .
. . .
Apriori,
FilteredAssociator
FPGrowth
. . .
. . .
. . .
ClassifierSubsetEval
PrinicipalComponents
. . .
. . .
. . .
Visualize
Output

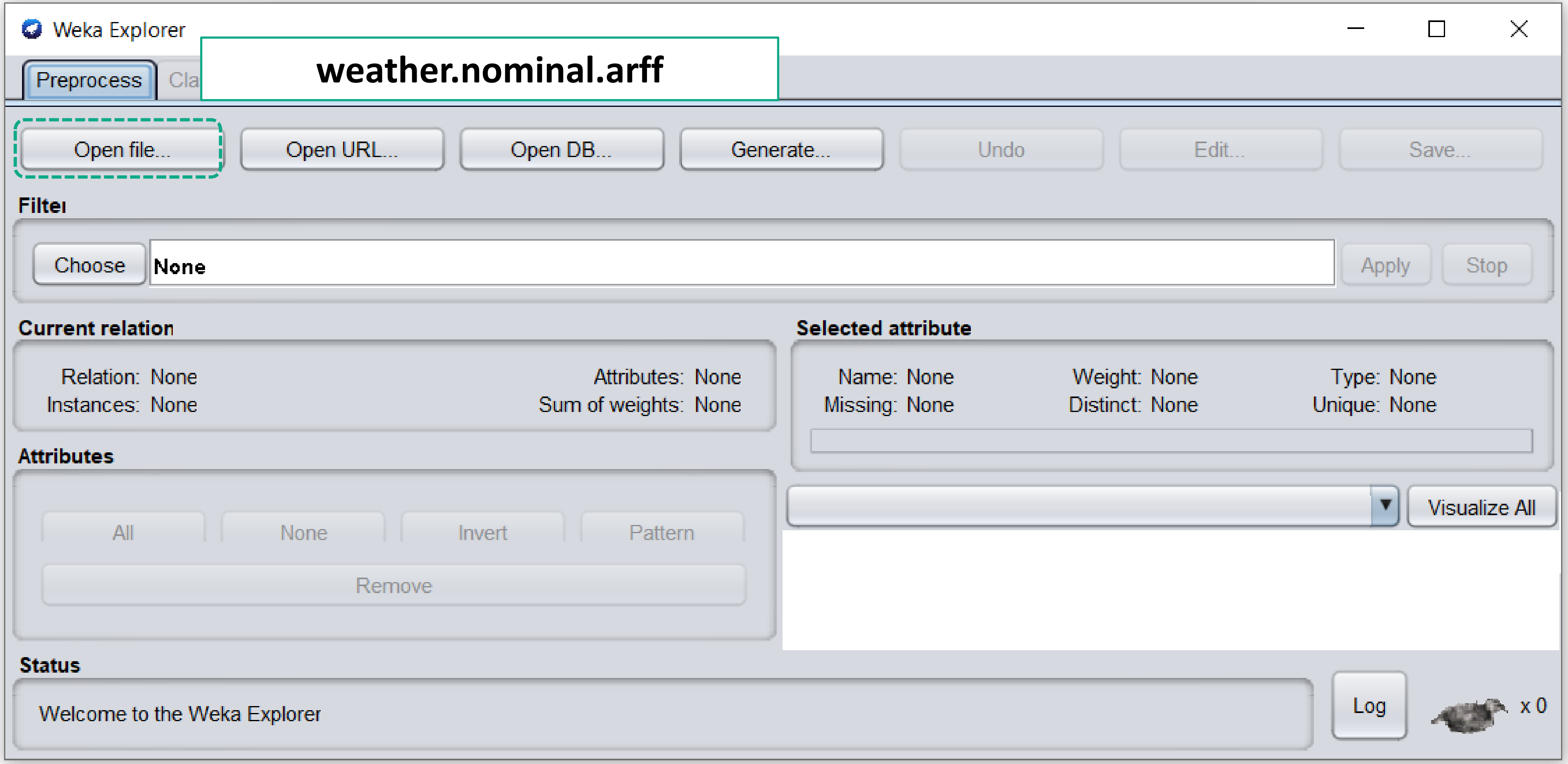
WEKA

First, the data source for the WEKA Explorer must be loaded. You can load the data from:

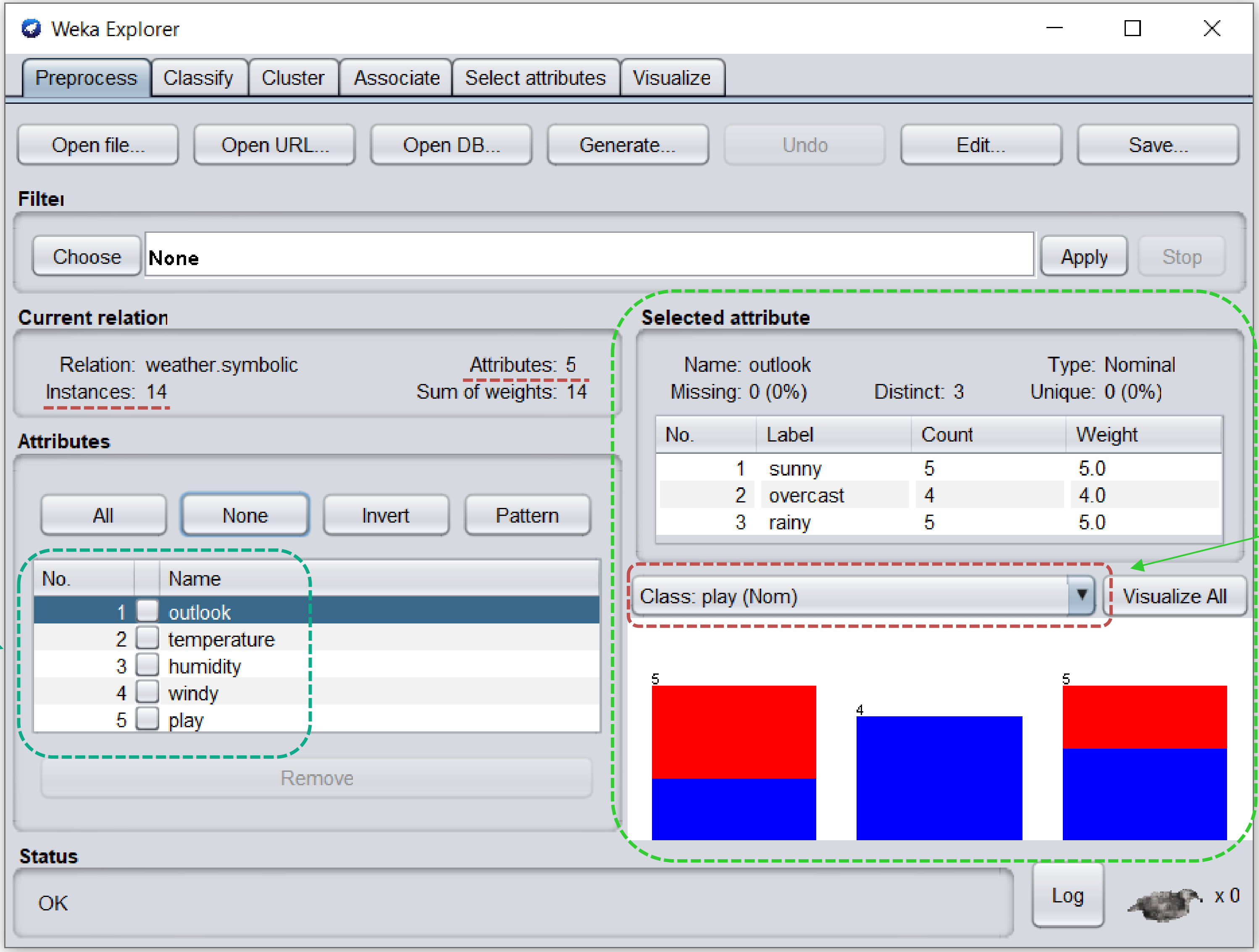
- Local File
- Web
- Database



WEKA



* The datasets are stored in the *data* folder, which is inside the software installation folder C:\Program Files\Weka-3-8-4\data



Classes

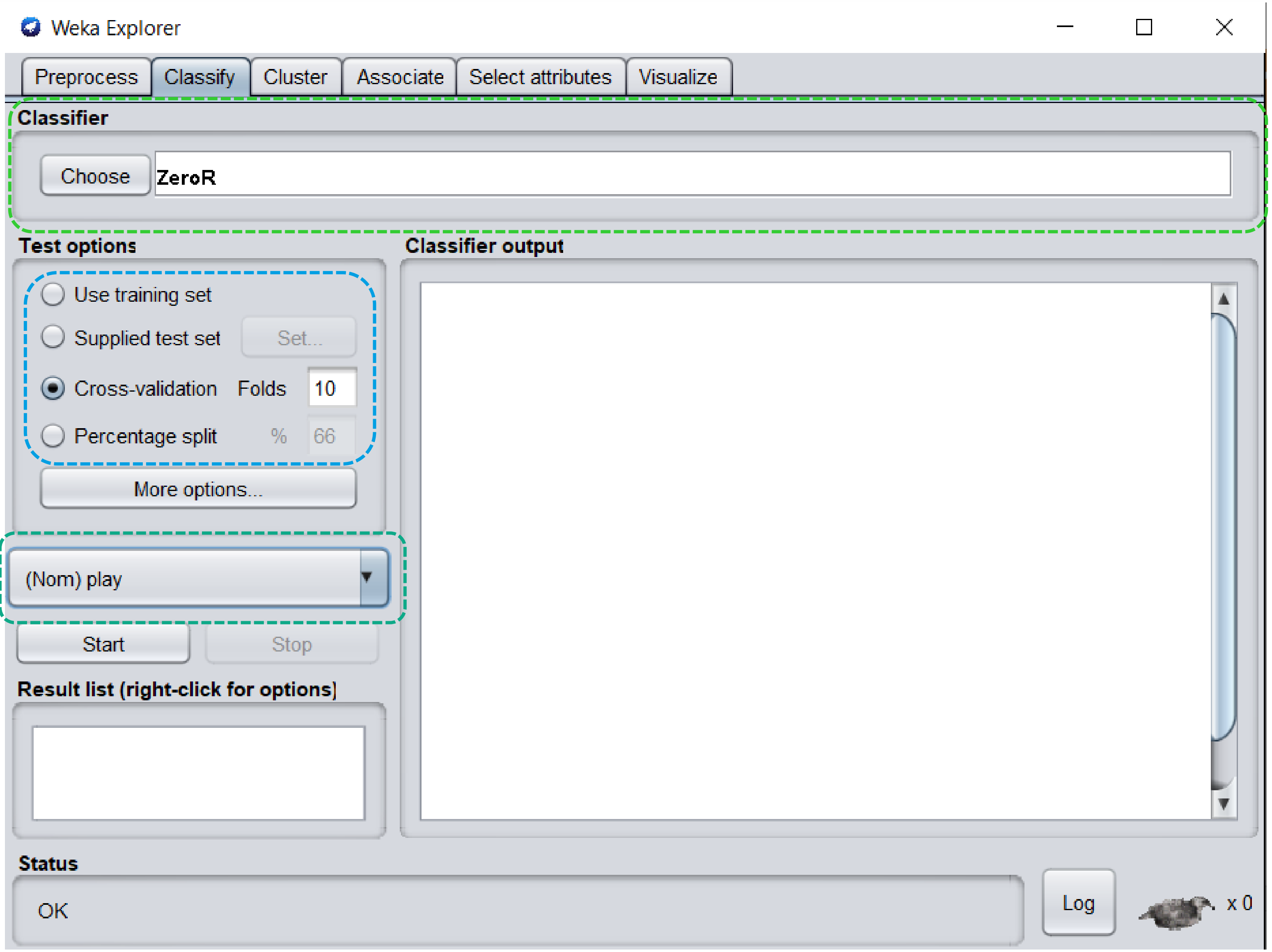
Attributes

When you press one of the attributes, you will see more details about the attribute on the right side.

CLASSIFIER

TEST OPTIONS

CLASS



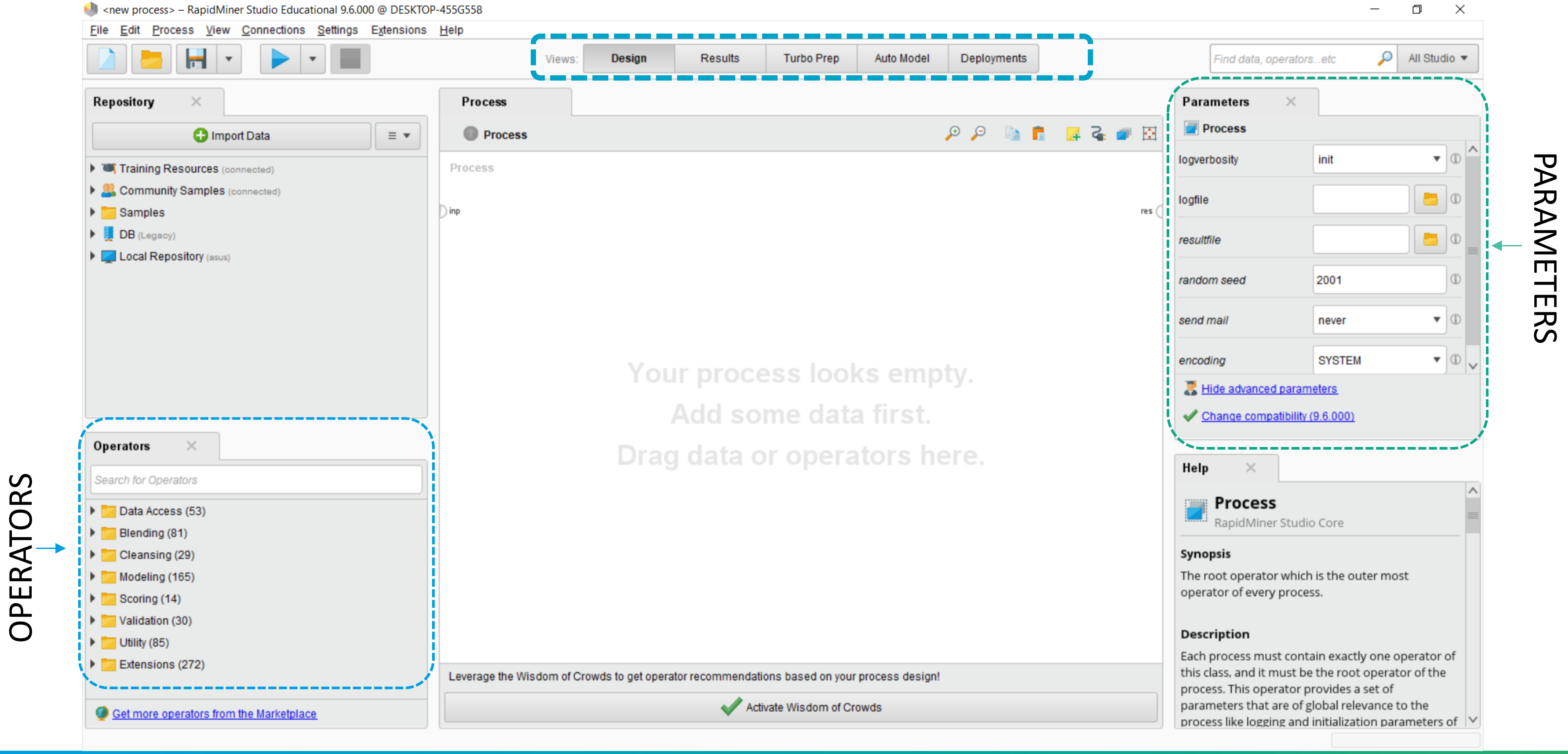
RapidMiner:

It is a commercial data analysis tool that uses machine learning and can be considered as an alternative to the Weka tool. The main objective of this tool, developed by a company with the same name, is to speed up the process of creating a predictive analysis and make it easier to apply it in practical business scenarios.

DOWNLOAD:

https://rapidminer.com/get-started/

RAPID MINER


<new process> – RapidMiner Studio Educational 9.6.000 @ DESKTOP-455G558
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Repository
Import Data
Training Resources (connected)
Community Samples (connected)
Samples
DB (Legacy)
Local Repository (asus)
Process
Your process looks empty.
Add some data first.
Drag data or operators here.
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
logverbosity init
logfile
resultfile
random seed 2001
send mail never
encoding SYSTEM
Hide advanced parameters
Change compatibility (9.6.000)
PARAMETERS
Operators
Search for Operators
Data Access (53)
Blending (81)
Cleansing (29)
Modeling (165)
Scoring (14)
Validation (30)
Utility (85)
Extensions (272)
Get more operators from the Marketplace
OPERATORS
Help
Process
RapidMiner Studio Core
Synopsis
The root operator which is the outer most operator of every process.
Description
Each process must contain exactly one operator of this class, and it must be the root operator of the process. This operator provides a set of parameters that are of global relevance to the process like logging and initialization parameters of

Data Preparation may take longer than the Data Mining process itself. The importance of data preparation is based on three aspects:

1. Real world data may be incomplete (missing values), noisy (outlier) and inconsistent (female, woman, F, W);

2. High-performance mining systems require quality data;

3. Quality data is a prerequisite for the production of effective models and quality standards.

DRIVES
Development and Research on Innovative Vocational Education Skills

DATA INTEGRATION
Integration of multiple databases or files.

DATA CLEANING
Removal of duplicates, treatment of missing values, treatment of outliers, resolution of inconsistencies, etc.

DATA TRANSFORMATION
Create attributes, rename attributes, convert data types, change data format, normalize data, etc.

DATA REDUCTION
Feature Selection, Discretization, etc.

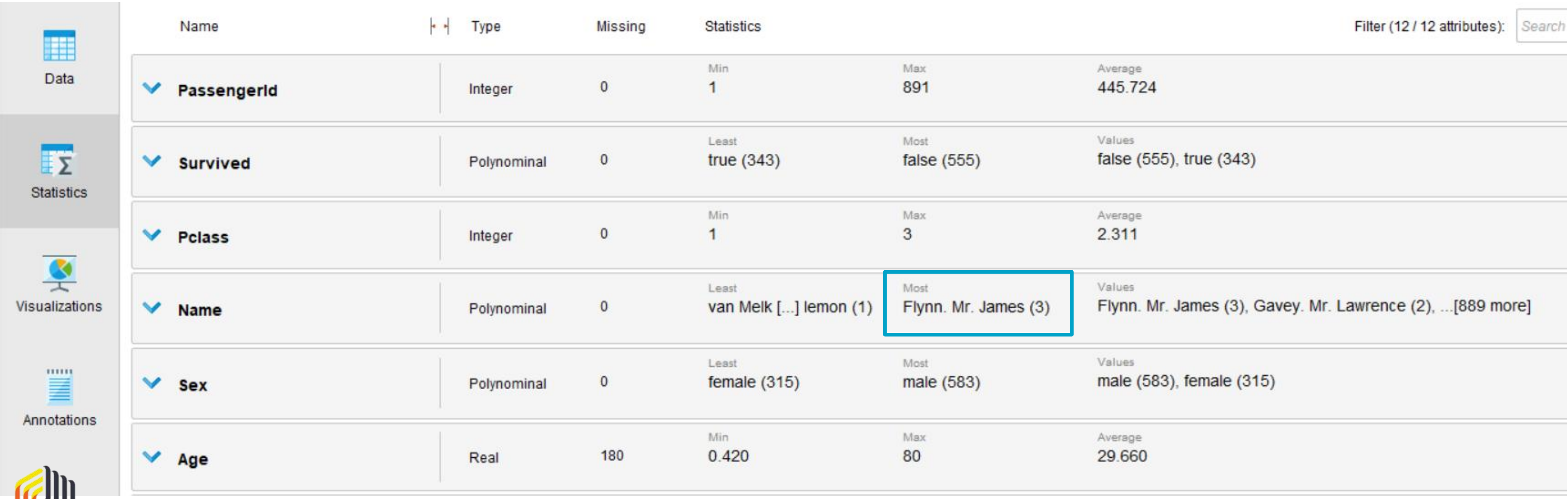
DATA SAMPLING
Oversampling, Undersampling

DATA PREPARATION

Remove Duplicates

1. Use the Titanic dataset. Drag it to the RapidMiner process window.

2. By looking at the Statistics tab we can see that there are 3 people with the same name which may indicate that there are duplicate instances in the dataset.

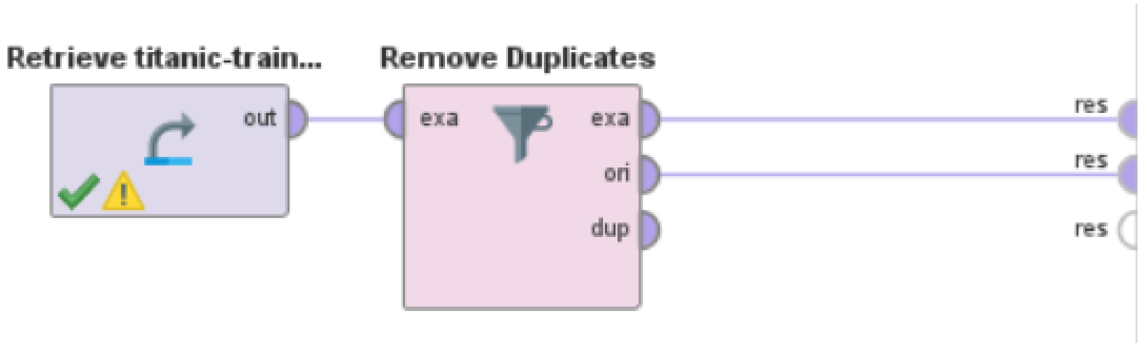


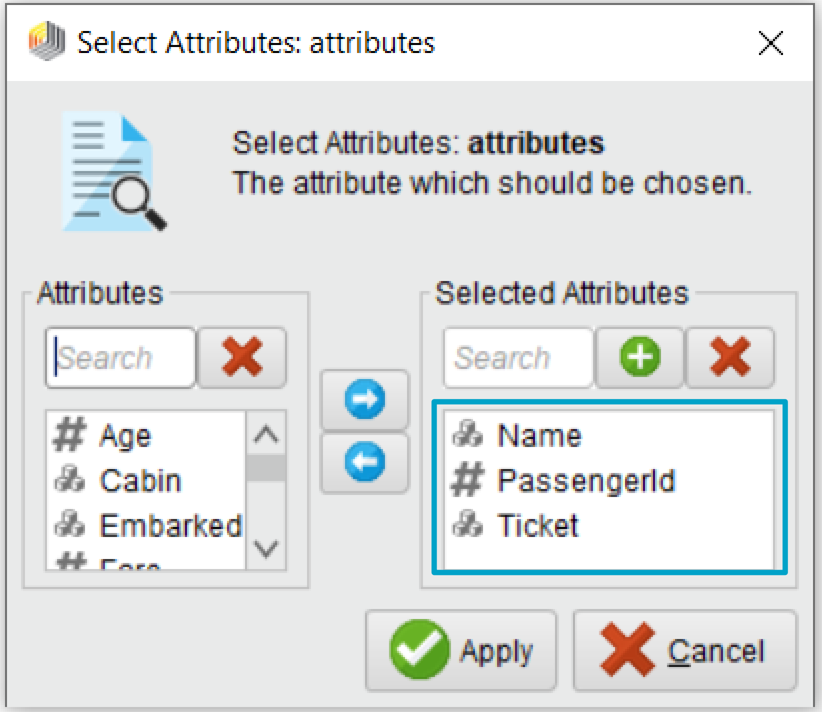
rapidminer

DRIVES
Development and Research on Innovative Vocational Education Skills

Remove Duplicates

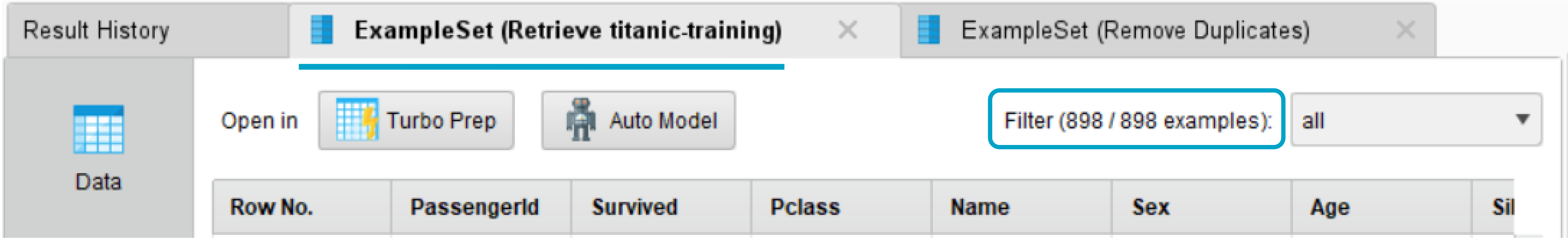
3. In the Operators window, select the "Remove Duplicates" operator and drag it to the process window. In the parameter attribute filter type choose the *subset* option. Select the attributes "PassengerID", "Name" and "Ticket". Click "Apply" and run the template.

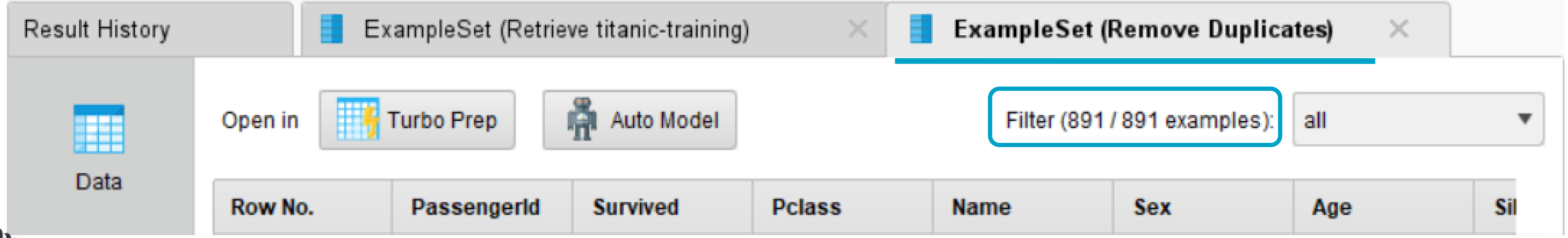




rapidminer

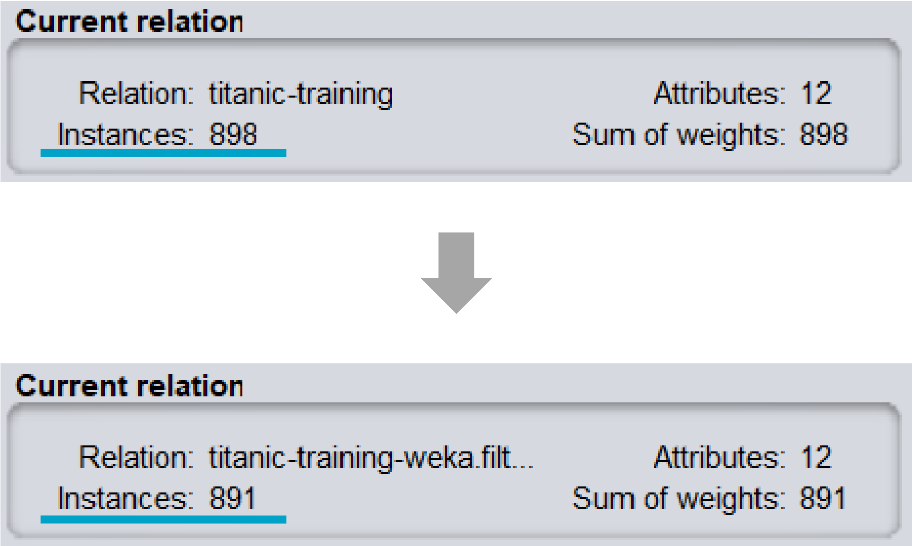
Remove Duplicates





Remove Duplicates

In the “Preprocess” window, in *filters – unsupervised – instance,* choose the “RemoveDuplicates” operator and click “Apply”.



Missing Attributes

Recover Missing Values

Please contact the participants and ask them to fill in the missing values

Remove Missing Values

Delete instances that contain missing values
*** If the sample is large enough, it is likely to be able to remove the data without significant loss of statistical power.*

RapidMiner: *Filter Examples*

Weka: *RemoveWithValues*

Impute Missing Values

Replace missing values with alternative values

RapidMiner: *Replace Missing Values* or *Impute Missing Values*

Weka: *ReplaceMissingValues*

Outliers

Maintain Outliers

In some cases, outliers do not originate from data errors and correspond to natural aberrant values

Remove Outliers

Remove instances that contain outliers

*** If the sample is large enough, it is likely to be able to remove the data without significant loss of statistical power.*

RapidMiner: *Detect Outlier + Filter Examples*

Weka: *InterquartileRange + RemoveWithValues*

Replace Outliers

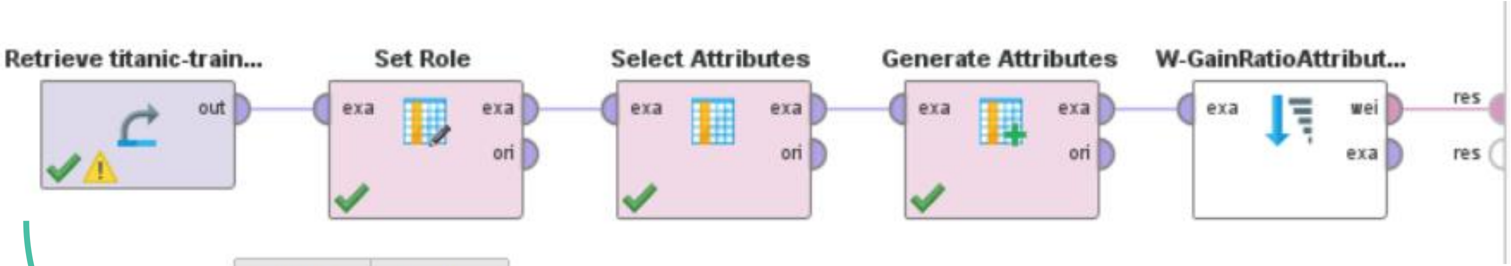
Replace the outliers with the highest or second lowest value in the observations, except for the outliers.

RapidMiner: *Replace Missing Values* or *Impute Missing Values*

Weka: *ReplaceMissingValues*

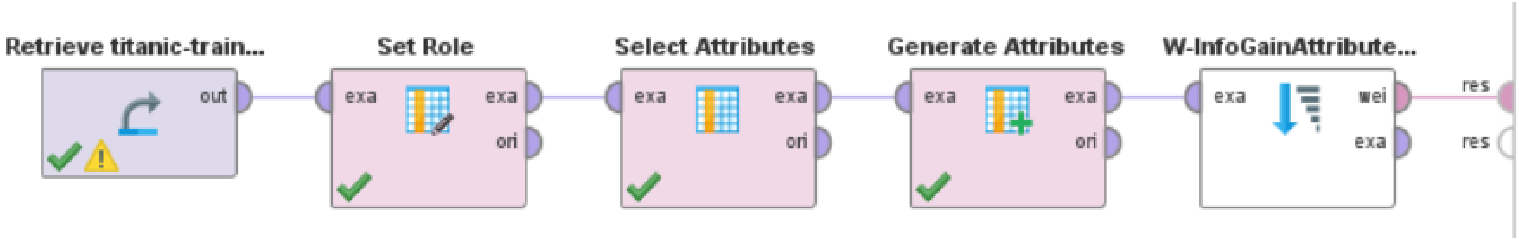
Feature Selection

Find the "W-GainRatioAttributeEval" operator, drag it into the process and run the model. Then try the "W-InfoGainAttributeEval" operator.



attribute	weight
Sex	0.231
Fare	0.071
Pclass	0.059
Age	0.033
Birth	0.033
SibSp	0.025
Parch	0.021
Embarked	0.020
Cabin	0.004
Passeng...	0

attribute	weight
Sex	0.216
Fare	0.097
Pclass	0.085
Cabin	0.029
SibSp	0.026
Embarked	0.022
Parch	0.016
Age	0.012
Birth	0.012
PassengerId	0



Feature Selection

In the window "Select attributes", in Attribute Evaluator, choose the option "GainRatioAttributeEval". Click on "Start". Then try the option "InfoGainAttributeEval".

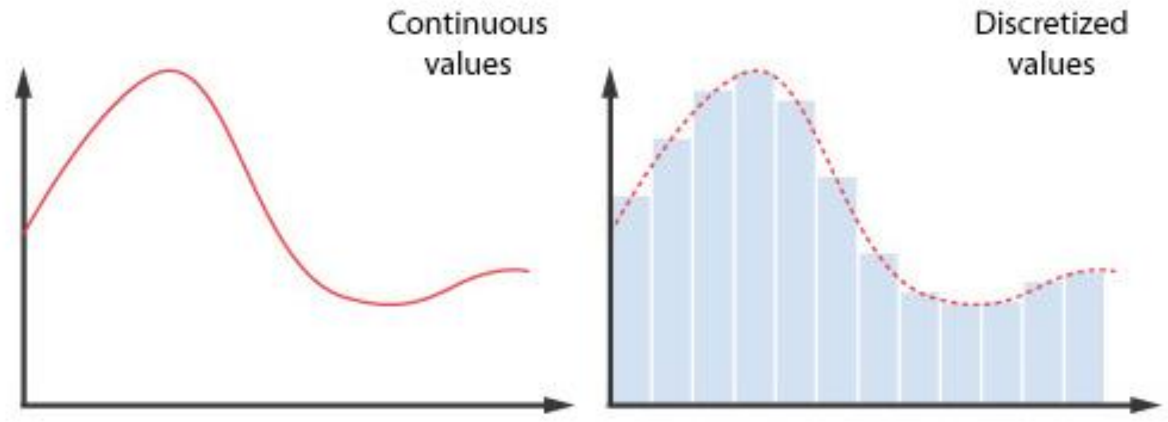
```
=== Attribute Selection o

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 2 Survived):
        Gain Ratio feature evaluator

Ranked attributes:
 0.23249    4 Sex
 0.06999    8 Fare
 0.05824    3 Pclass
 0.03337    5 Age
 0.03337   11 birth
 0.02509    6 SibSp
 0.01938    7 Parch
 0.01903   10 Embarked
 0.00413    9 Cabin
 0          1 PassengerId

Selected attributes: 4,8,3,5,11,6,7,10,9,1 : 10
```

```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 2 Survived):
        Information Gain Ranking Filter

Ranked attributes:
 0.2177    4 Sex
 0.0962    8 Fare
 0.0838    3 Pclass
 0.0292    9 Cabin
 0.0265    6 SibSp
 0.0209   10 Embarked
 0.0154    7 Parch
 0.0117   11 birth
 0.0117    5 Age
 0          1 PassengerId

Selected attributes: 4,8,3,9,6,10,7,11,5,1 : 10
```

Discretization/Binning

The objective of discretization is to transform a continuous attribute into a discrete attribute. In several Data Mining algorithms, it is necessary to use discretized data since these algorithms can only handle discrete attributes.

Discretization reduces the impact that small fluctuations in the data have on the model, often small fluctuations are just noise. Each "bin" soothes the fluctuations/noise.



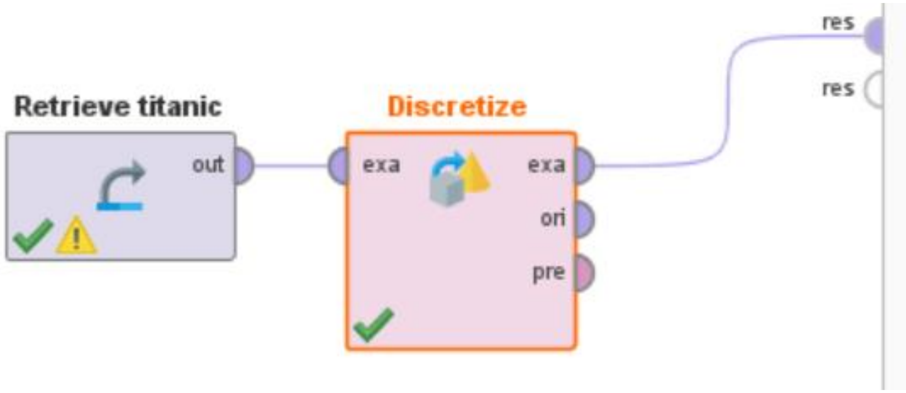
DRIVES

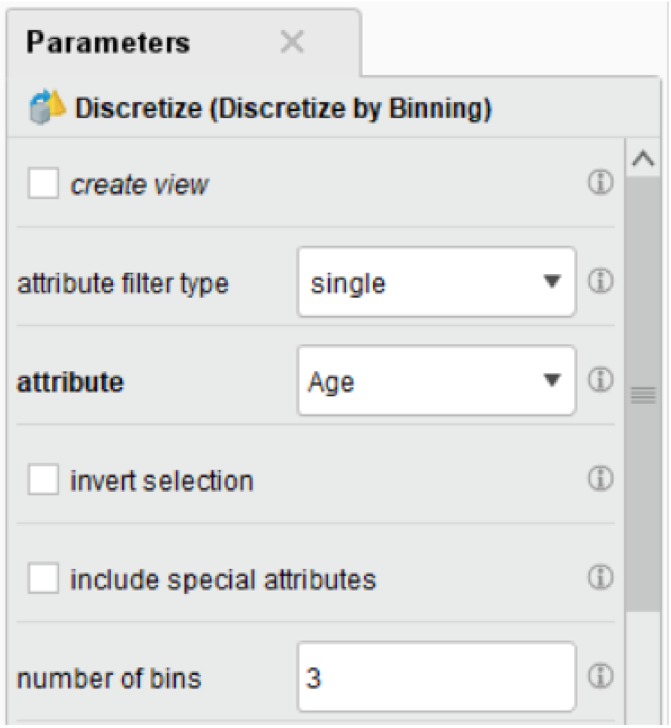
Development and Research on Innovative Vocational Education Skills

Discretization/Binning

1. Find the "Discretize by Binning" operator and drag it to the process window.

2. In the parameters window, select the value "single" for the parameter attribute filter type, choose the attribute that will suffer the discretization and the number of bins.

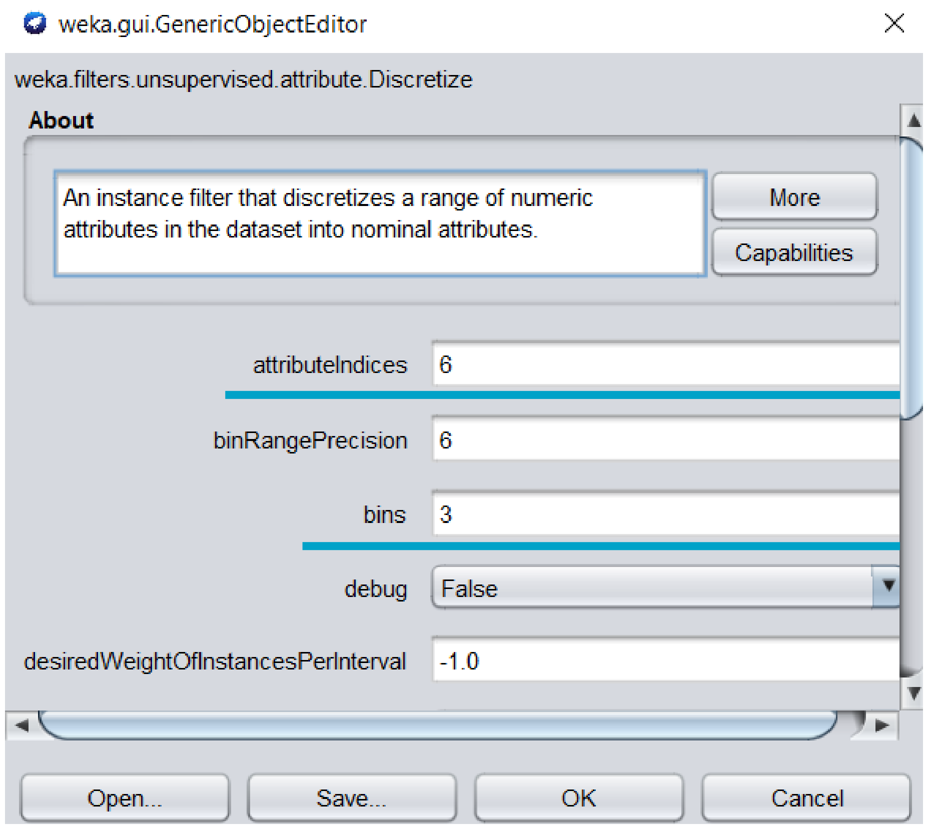


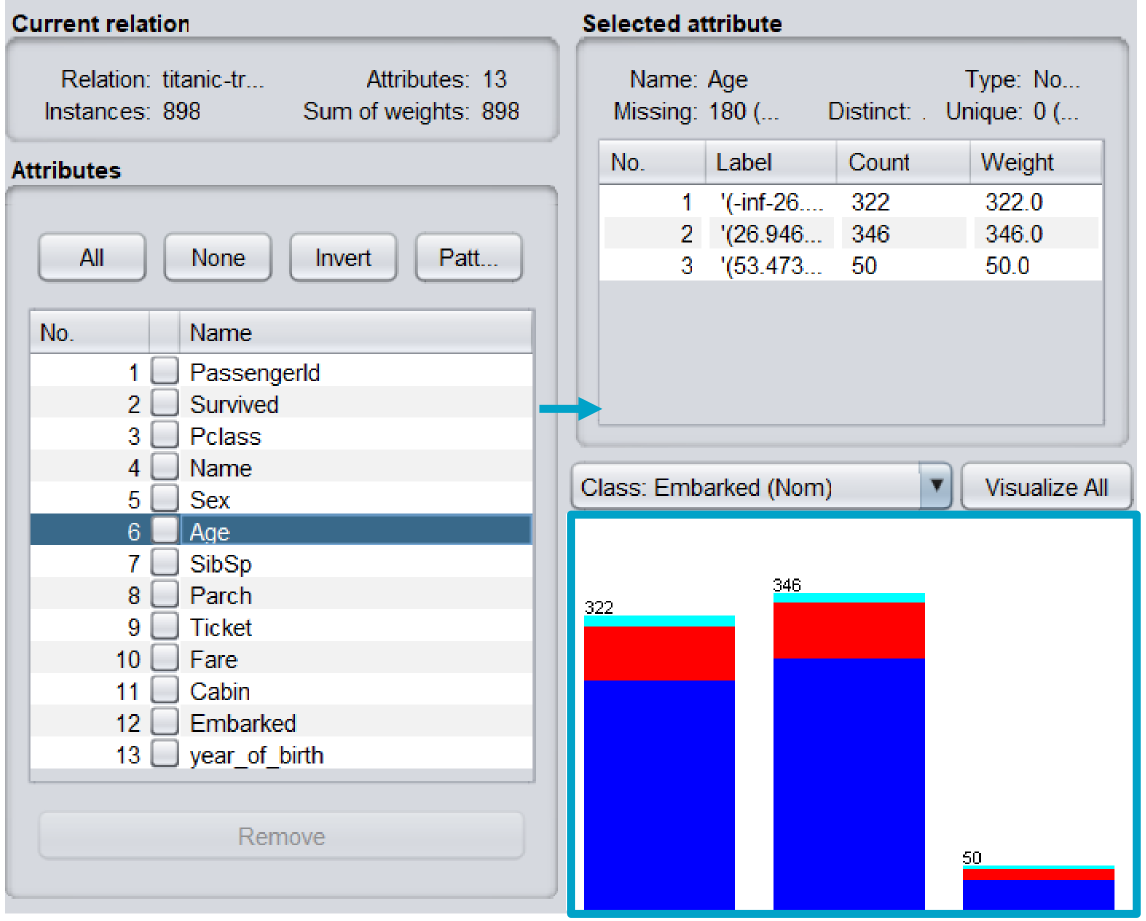


rapidminer

Discretization/Binning

In the Preprocess window, in filters - unsupervised - attribute, choose the filter "Discrete". Double-click the filter.





DRIVES
Development and Research on Innovative Vocational Education Skills

Normalization

When there are attributes with disparate value ranges or at different scales, attributes with values at a higher scale may unrealistically overshadow a significant or equally important attribute (but at a lower scale). Thus, attributes are normalized to transform all attributes on the same scale.

Data normalization allows a new scale to be assigned to an attribute so that the values of that attribute can fall in a new scale in a specific range from 0 to 1 for example.

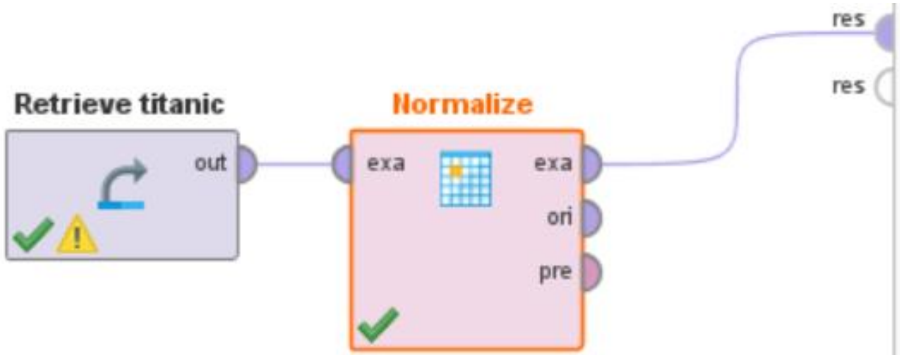
$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

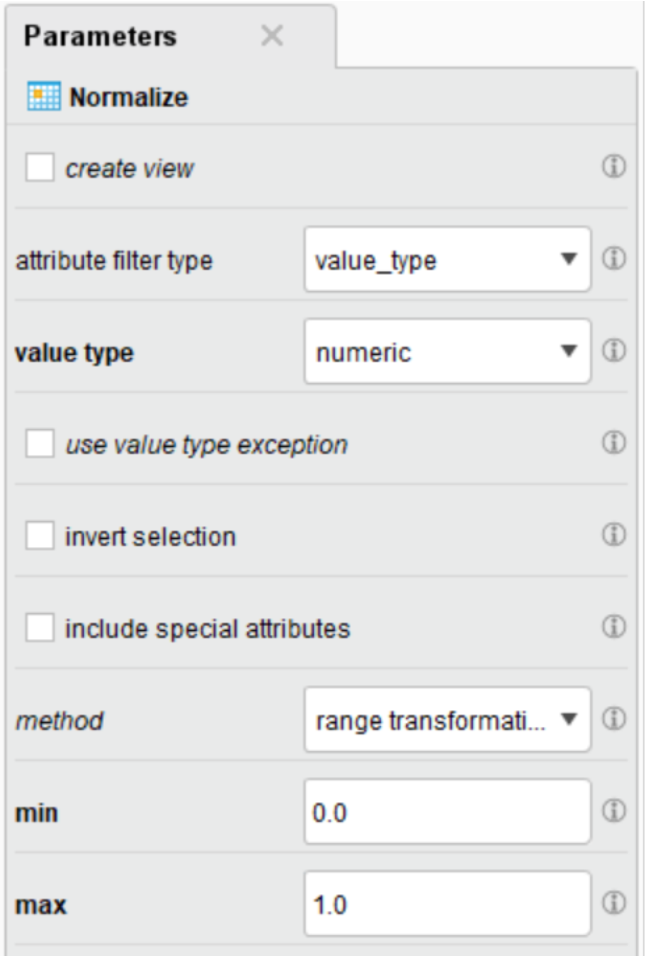
→ Min-Max Normalization

Normalization

1. Find the "Normalize" operator and drag it to the process window.

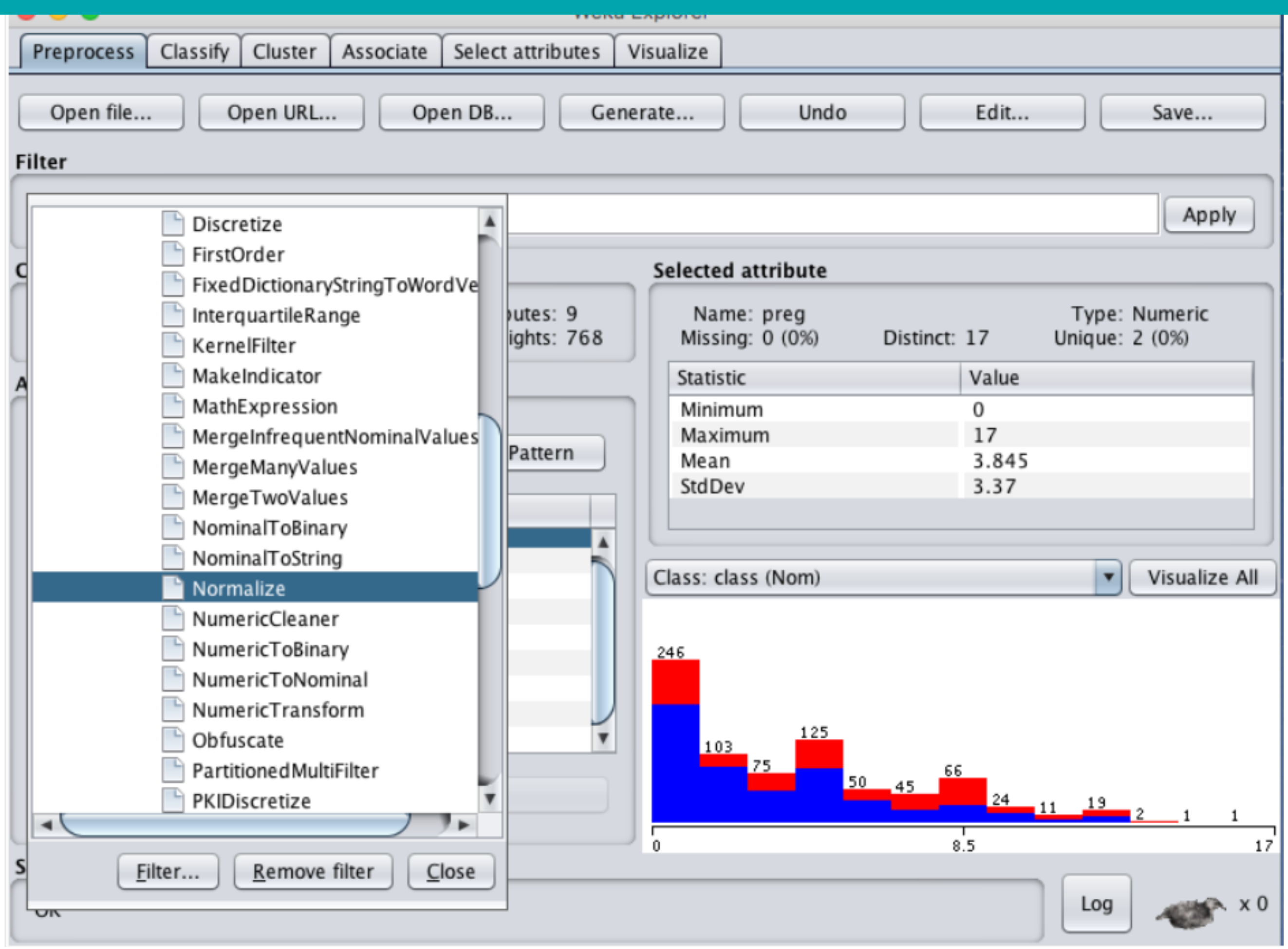
2. In the parameters window, set the parameters attribute filter type, value type and method.





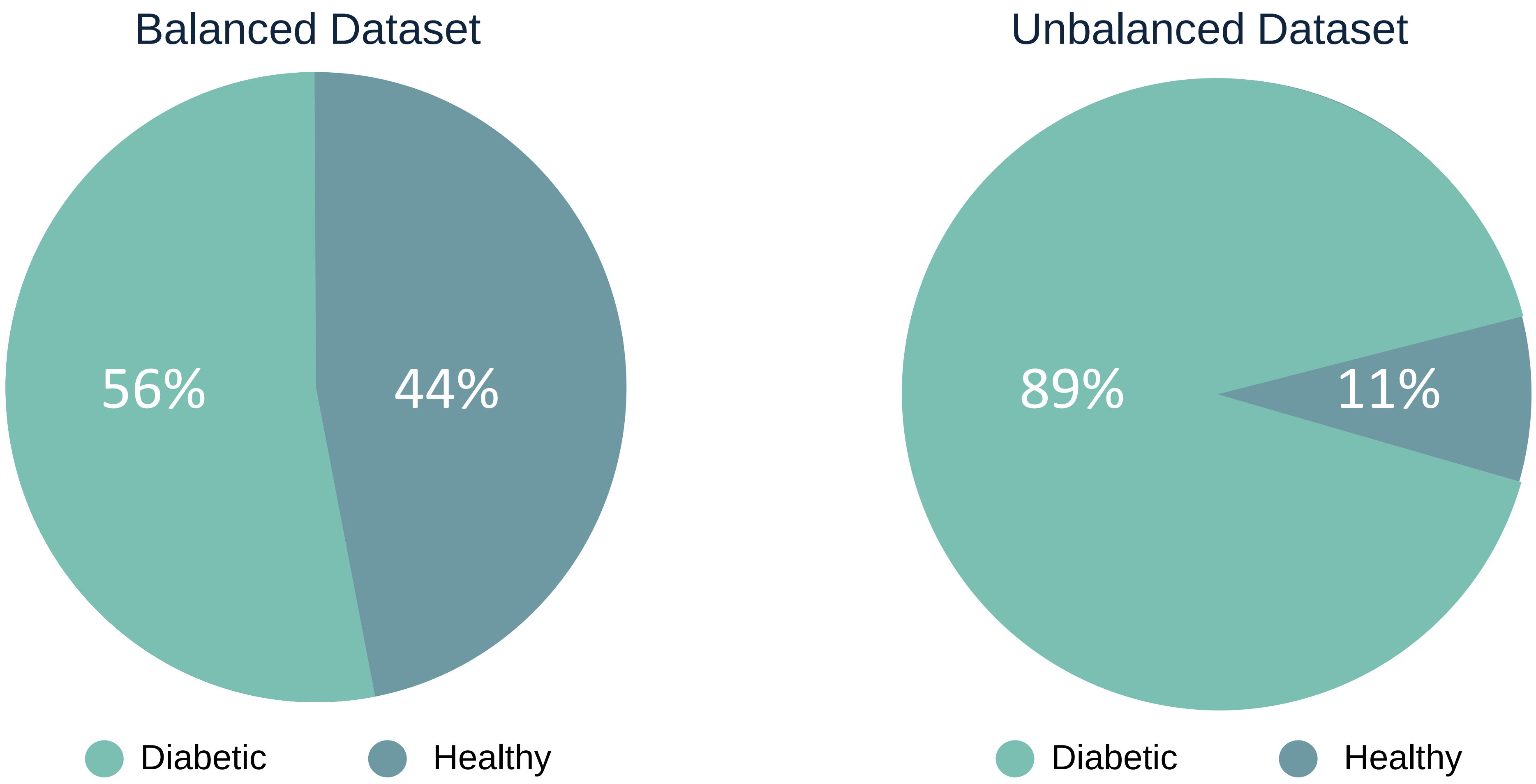
Normalization

In the Preprocess window, in filters - unsupervised - attribute, choose the "Normalize" filter.



DRIVES
Development and Research on Innovative Vocational Education Skills

Data Sampling



DRIVES
Development and Research on Innovative Vocational Education Skills

Data Sampling

In these cases, the algorithm receives significantly more examples from a class, which leads it to be skewed to that specific class. Due to the disparity of classes, the algorithm is then prone to categorize instances into the majority class and does not learn what makes the other class "different", nor does it understand the underlying patterns that allow classes to be distinguished.

Classifiers generated from unbalanced datasets have high false negative rates for the less common classes.

As there are few instances of the minority class, the associated error is reduced, giving at the same time the false sense that we are building a highly accurate model. Both the inability to predict rare events, i.e., the minority class, and the misleading accuracy decrease the performance of the prediction models built.

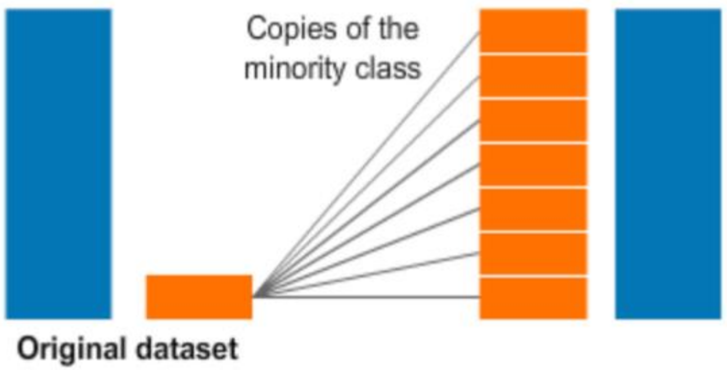
DRIVES
Development and Research on Innovative Vocational Education Skills

Data Sampling

Alteration of the class distributions in the data set, with the aim of reducing the imbalance and obtaining better classifiers than those obtained from the original distribution.

Undersampling
Removal of cases from the majority class

Oversampling
Replication of cases from the minority class

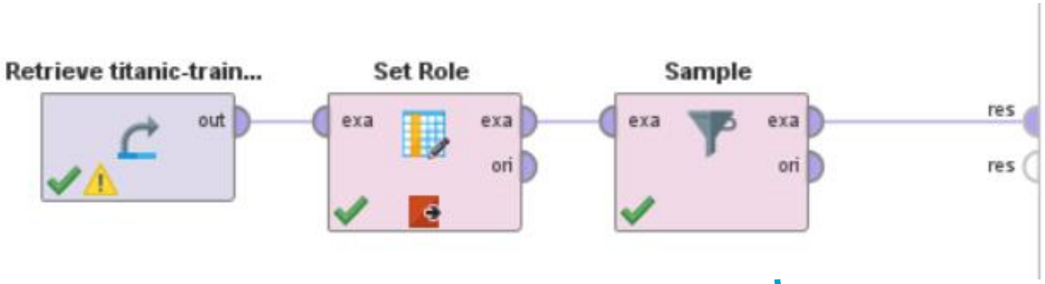


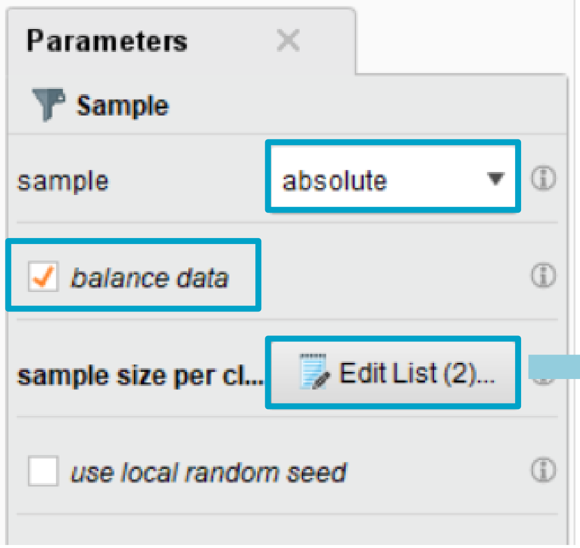
*** substantial loss of statistical power may occur*

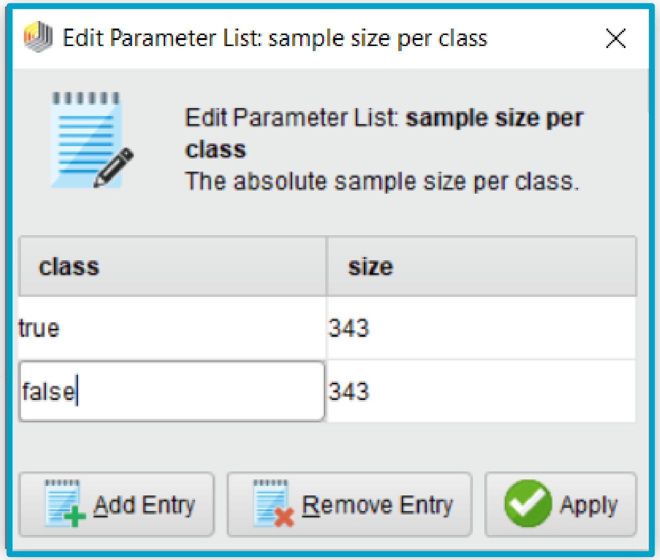
Data Sampling - Undersampling

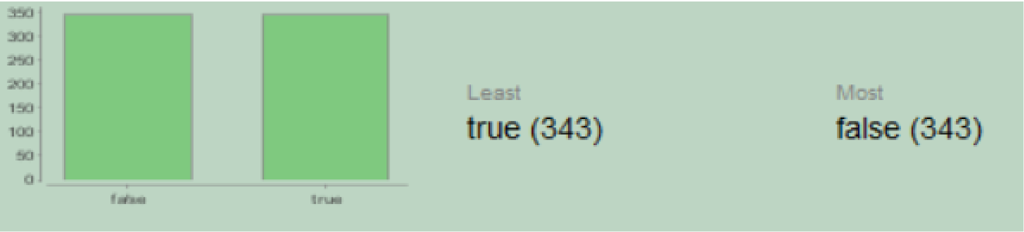
1. Find the "Sample" operator and drag it to the process window.

2. In the parameters window, set the parameters *sample*, *balance data* and *sample size per class* according to the instructions below.







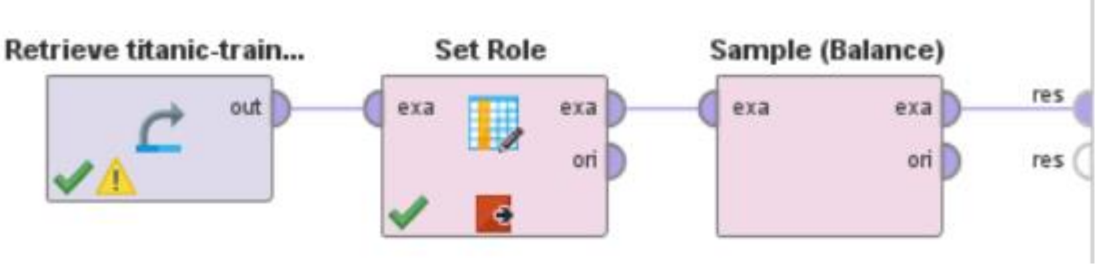


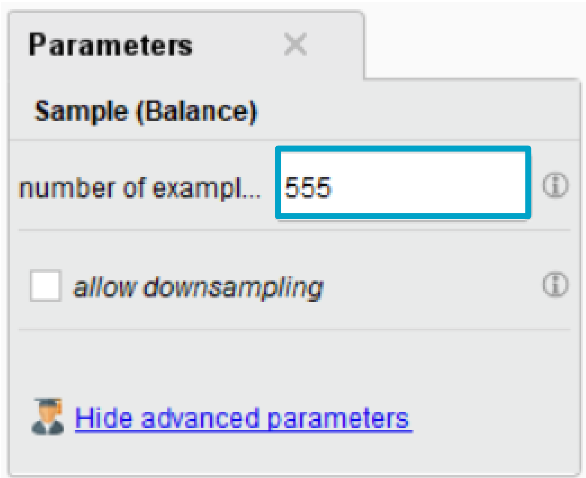
Data Sampling - Oversampling

1. Install the Mannheim RapidMiner Toolbox extension in Extensions-> Marketplace.

2. Find the operator "Sample (Balance)" and drag it to the process window.

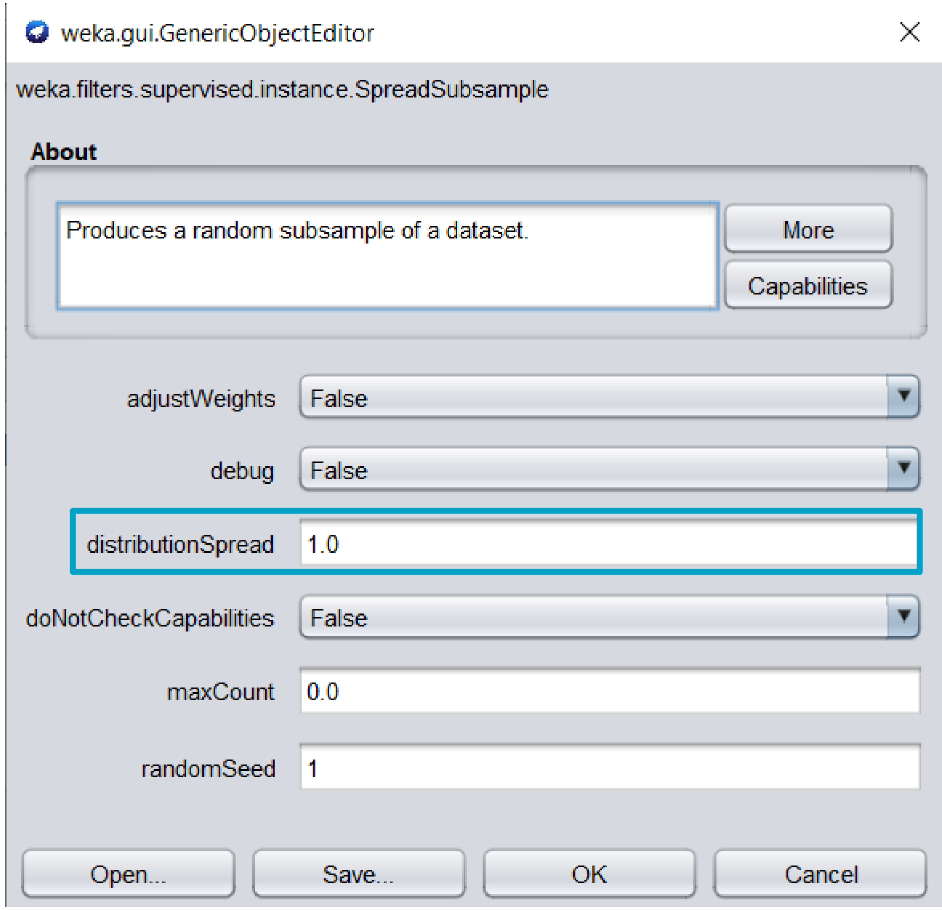
3. In the parameters window, set the number of examples for the number of instances of the majority class.

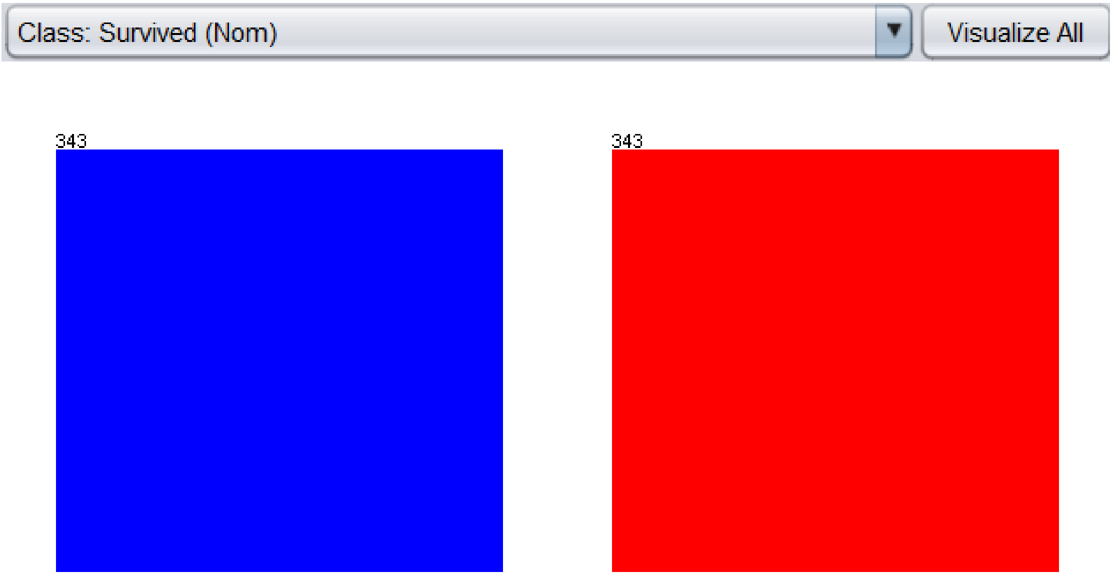




Data Sampling - Undersampling

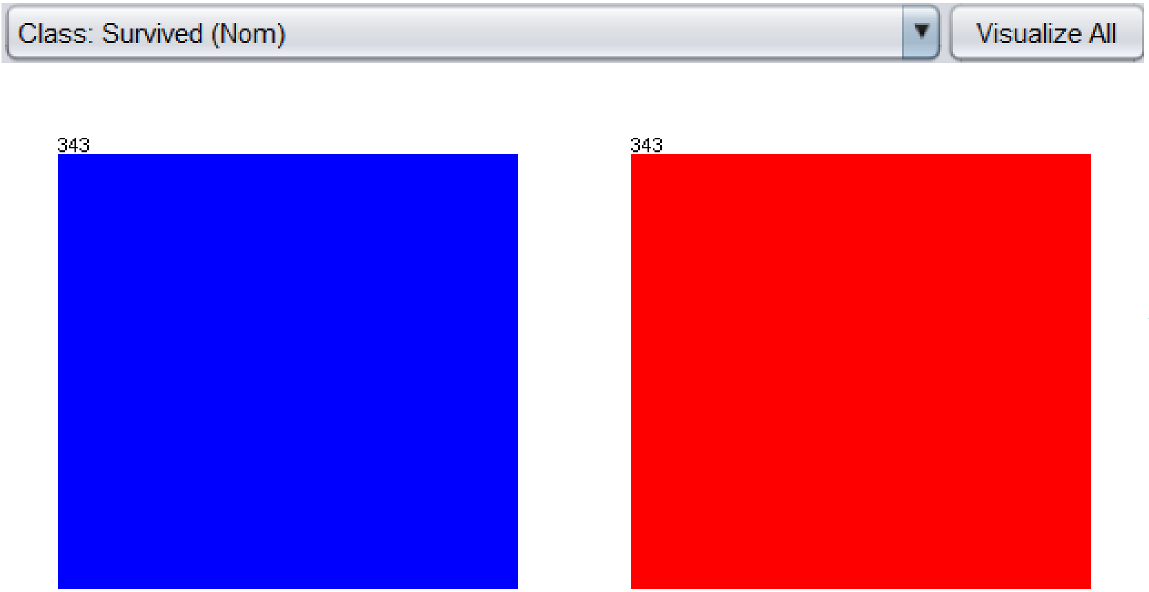
1. In *Filters-> supervised -> instance -> SpreadSubsample* ***distributionSpread*** = 1.0

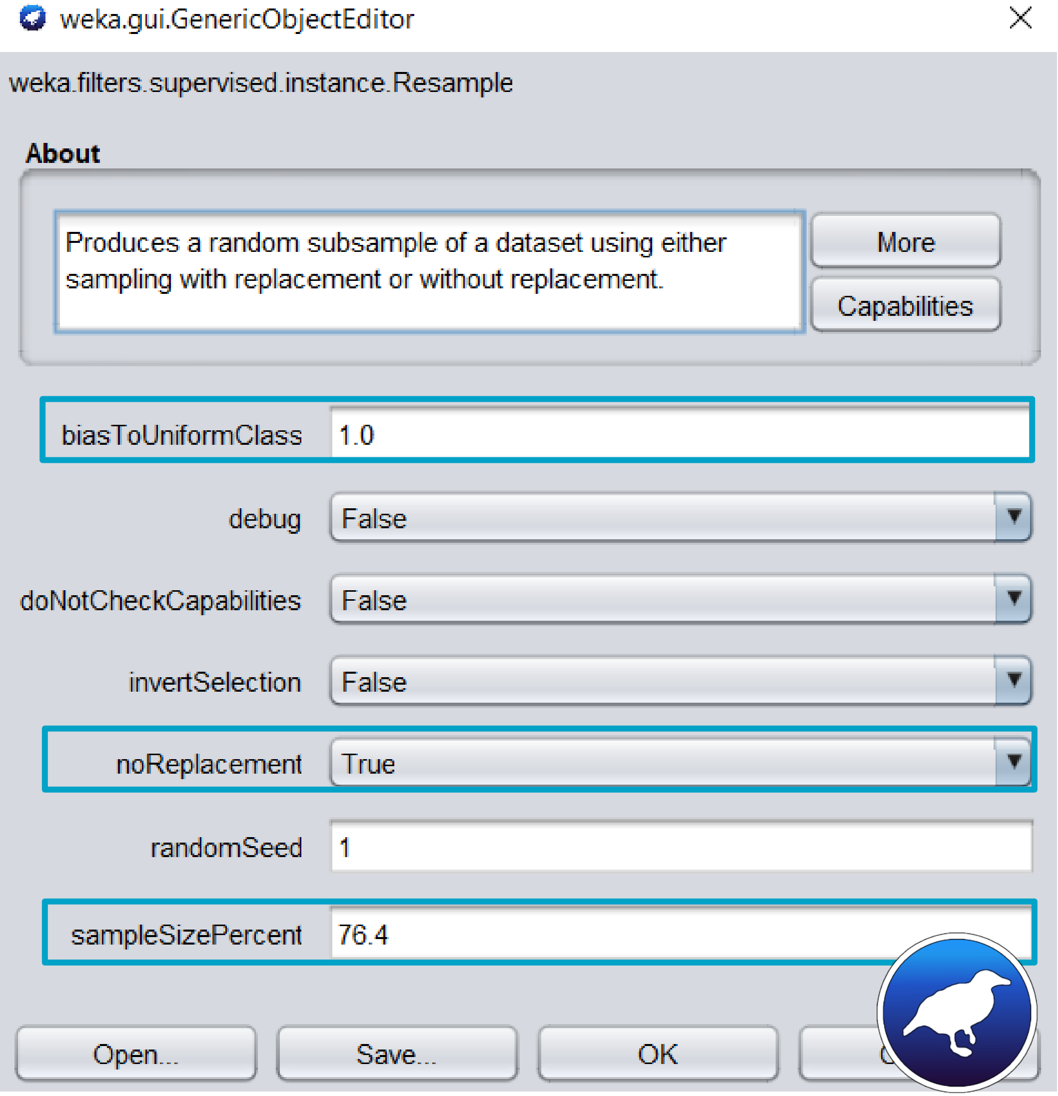




Data Sampling - Undersampling

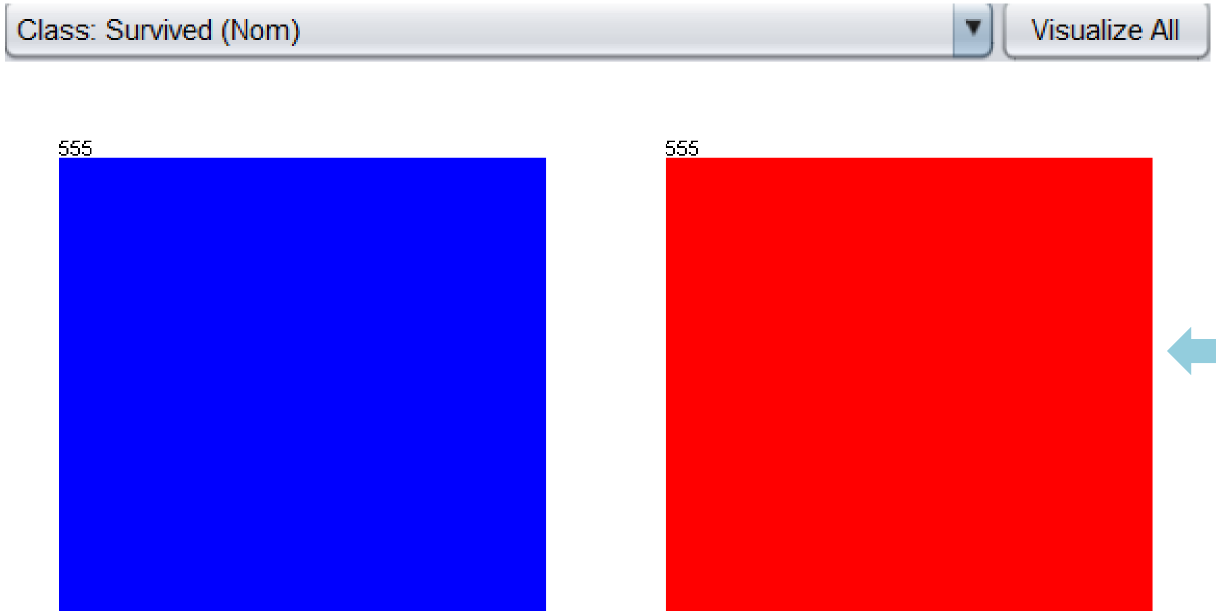
2. In *Filters-> supervised -> instance -> Resample*
biasToUniformClass = 1.0
sampleSizePercent = X, where X/2 is the percentage of data that belongs to the minority class
noReplacement = true

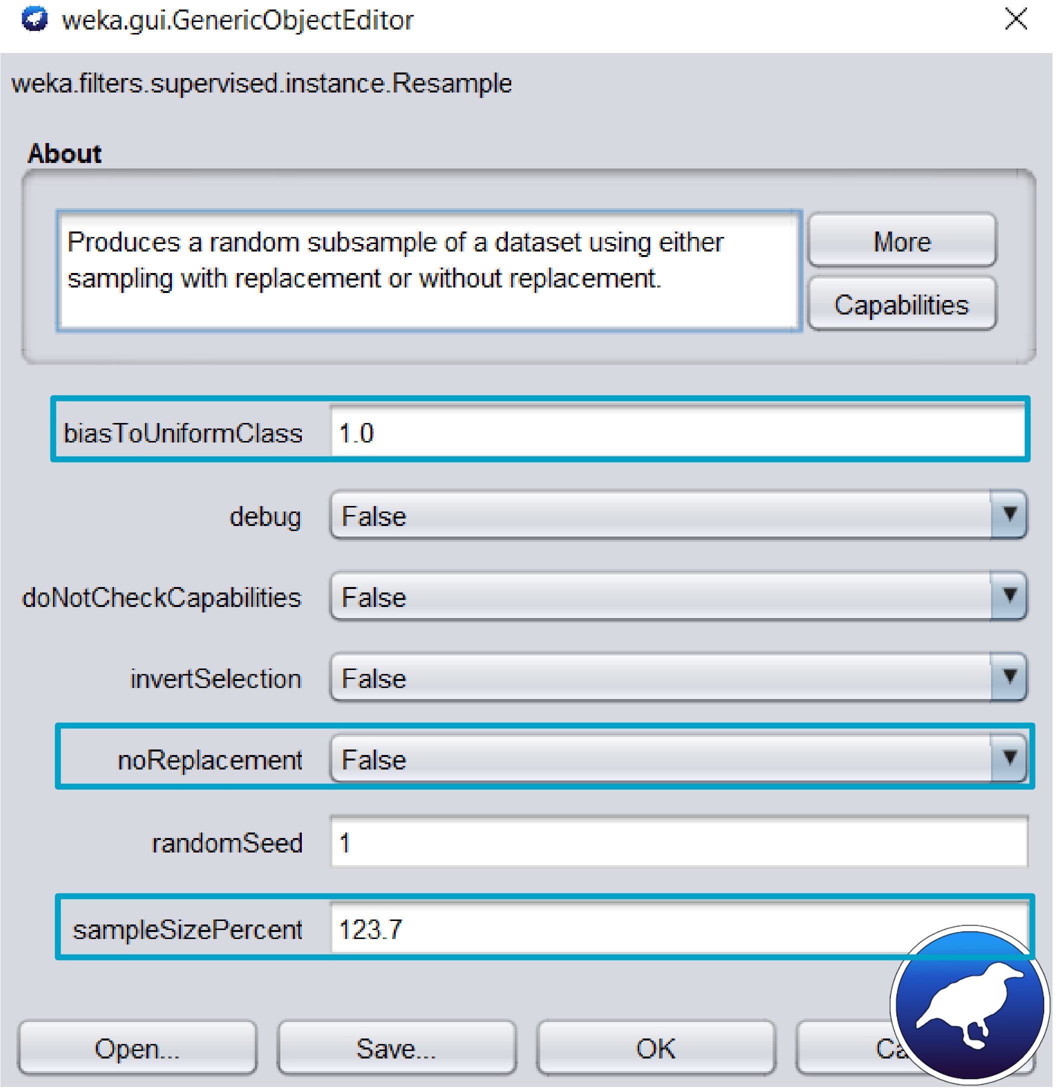




Data Sampling - Oversampling

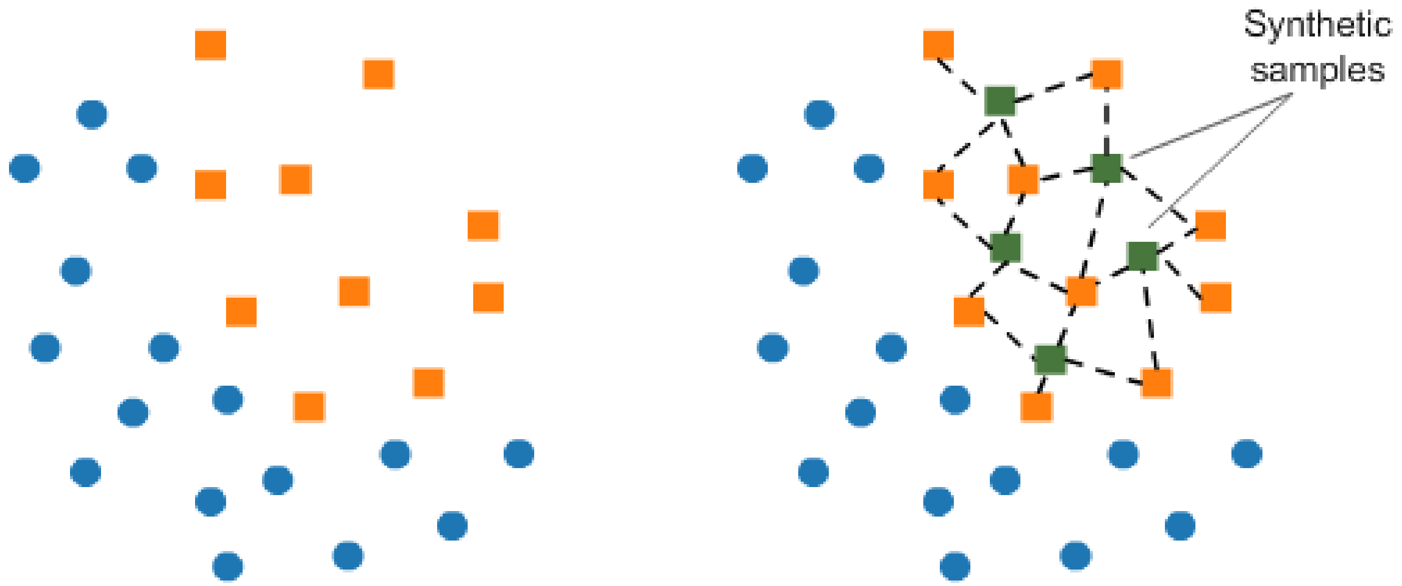
1. In *Filters-> supervised -> instance -> Resample*
biasToUniformClass = 1.0
sampleSizePercent = Y, where Y/2 is the percentage of data that belongs to the majority class
noReplacement = false





Data Sampling – SMOTE (Synthetic Minority Over-sampling Technique)

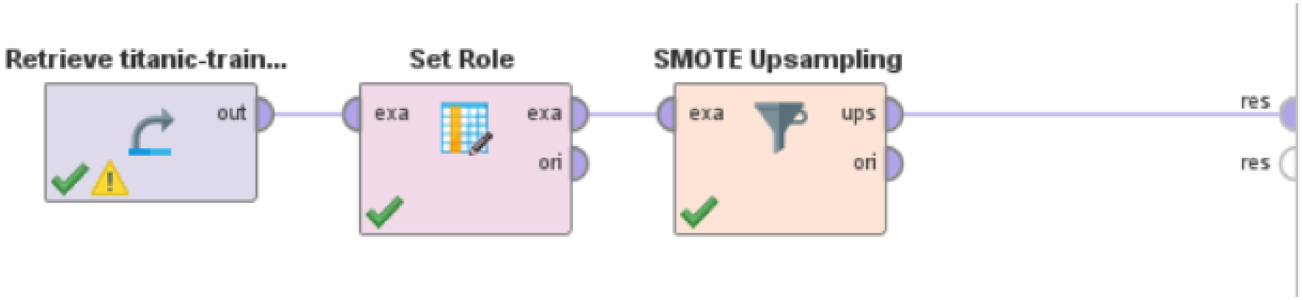
It is a technique based on the k nearest neighbor, judged by the Euclidean distance between data points in the feature space.

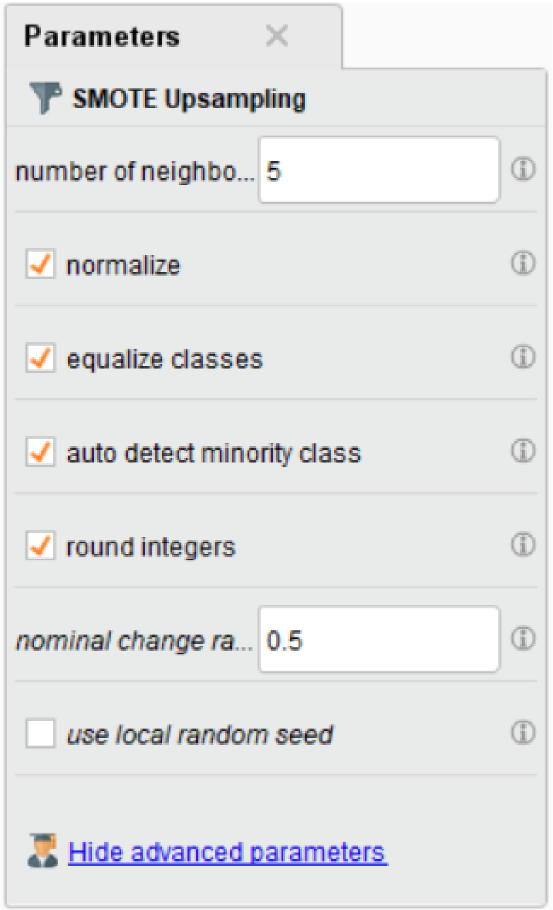


Data Sampling - SMOTE

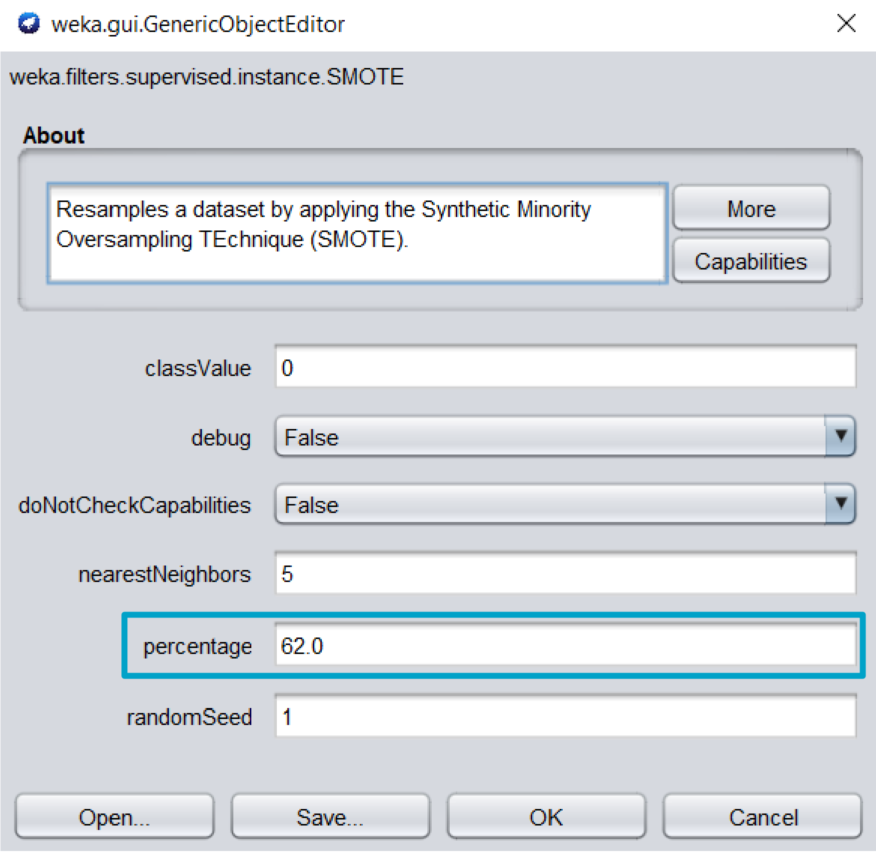
1. Install the Operator Toolbox extension through Extensions -> MarketPlace.

2. Find the operator "SMOTE Upsampling" and drag it to the process window.





Data Sampling - SMOTE

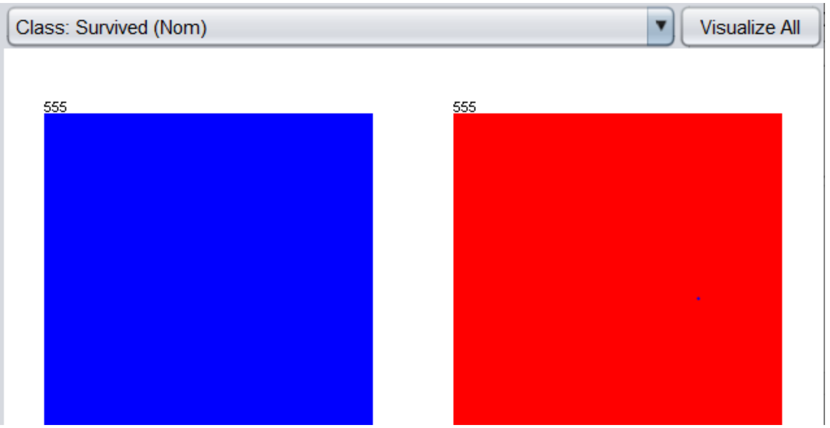


1. Install the SMOTE extension through Tools -> Package manager.

2. In Filters-> supervised -> instance -> SMOTE

percentage = x, where x = ((max-min)*100)/min where max is the number of instances that belong to the majority class and min is the number of instances that belong to the minority class.

nearestNeighbors = k, where k indicates how many nearby instances (around the new instance) are used to build a synthetic instance. The default value is 5.



At the **Modeling** stage, algorithms are used to determine patterns in the data previously processed. As a result, several modeling techniques are selected and applied, and their parameters are calibrated to the optimum values. In this way:

1. Based on the defined objectives, modeling techniques should be selected for the previously prepared data set.

2. Some scenarios should be defined to test and verify the quality and validity of the model.

3. Finally, the models should be executed in the prepared data set.

SUPERVISED LEARNING

Classification or Regression

UNSUPERVISED LEARNING

Dimensionality Reduction or Clustering

SEMI-SUPERVISED LEARNING

Predictions in the medical field (tests and diagnostics are expensive and time consuming and only part of the population has them)

REINFORCEMENT LEARNING

Gaming, Finance Sector, Manufacturing, Inventory Management, Robot Navigation

1. Evaluation of the results achieved:

- Understand the results and verify their impact on the data mining objective initially defined;
- Verify the result against existing literature in order to see whether innovative and useful discoveries have been made;
- Draw relevant conclusions from the results achieved;
- Analyze whether there are new objectives that can be addressed in the future.

2. Review the data mining process to identify possible failures, neglected factors, changes in steps or unexpected options.

3. Refine the process and analyze the implementation potential.

DRIVES
Development and Research on Innovative Vocational Education Skills

Training Dataset

Weka
Use Training Set

RapidMiner
Multiply operator

Supplied Test Set

Weka
Supplied Test set

RapidMiner
Drag the dataset

Percentage Split

Weka
Percentage Split

RapidMiner
Split operator

Cross Validation

Weka
Cross-validation

RapidMiner
Cross Validation operator

The **Confusion Matrix** is a table with four different combinations of predicted and actual values.

Confusion Matrix

		ACTUAL	
		Positives (1)	Negatives (0)
PREDICTED	Positives (1)	TP	FP
	Negatives (0)	FN	TN

Accuracy measures the ability of the model to capture true positive as positive and true negative as negative. It can be a useful measure if there is the same number of samples per class, but if, on the contrary, the set of samples is unbalanced, the accuracy is not an adequate measure.

Accuracy

		ACTUAL	
		Positives (1)	Negatives (0)
PREDICTED	Positives (1)	TP	FP
	Negatives (0)	FN	TN

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Classification Error measures the number of instances incorrectly classified by the model, that is, the number of False Positives, also known as Type I error, and the number of False Negatives, also known as Type II error.

Classification Error

ACTUAL

PREDICTED	Positives (1)	Negatives (0)
Positives (1)	TP	FP
Negatives (0)	FN	TN

$$Classification\ Error = \frac{FP + FN}{TP + FP + FN + TN}$$

Precision measures the accuracy of the model against the predicted positives and determines how many of them are actually positive. Precision is a good measure if the cost of False Positives is high (e.g.: SPAM detection).

Precision

		ACTUAL	
		Positives (1)	Negatives (0)
PREDICTED	Positives (1)	TP	FP
	Negatives (0)	FN	TN

$$Precision = \frac{TP}{TP + FP}$$

Recall also called **Sensitivity** or **True Positive Rate** calculates how many of the true positives the model captures as being positive. Recall should be the metric to be use when there is a high cost associated with false negatives (e.g. medical diagnosis).

Recall

		ACTUAL	
		Positives (1)	Negatives (0)
PREDICTED	Positives (1)	TP	FP
	Negatives (0)	FN	TN

$$Recall = \frac{TP}{TP + FN}$$

The **F1 score** is adequate when it is necessary to find a balance between Precision and Recall and when there is an uneven distribution of the class.

F1 score

PREDICTED \ ACTUAL	Positives (1)	Negatives (0)
Positives (1)	TP	FP
Negatives (0)	FN	TN

$$F1\,score = 2 * \frac{precision * recall}{precision + recall}$$

Specificity or **True Negative Rate** calculates how many of the true negatives the model captures as being negative. Consider the example of a medical examination to diagnose a disease, the Specificity relates to the ability of the test to correctly reject healthy patients. A test with a higher Specificity has a lower error rate of Type I.

Specificity

$$Specificity = \frac{TN}{FP + TN}$$

		ACTUAL	
		Positives (1)	Negatives (0)
PREDICTED	Positives (1)	TP	FP
	Negatives (0)	FN	TN

Fall-out or **False Positive Rate** calculates how many false positives the model was unable to capture as being negative.

Fall-out

$$FPR = \frac{FP}{FP + TN} = 1 - specificity$$

		ACTUAL	
		Positives (1)	Negatives (0)
PREDICTED	Positives (1)	TP	FP
	Negatives (0)	FN	TN

K statistic is a measure of the reliability among evaluators and the discrepancy between them, taking into account the possibility that the agreement may occur by chance.

K statistic

		ACTUAL	
		Positives (1)	Negatives (0)
PREDICTED	Positives (1)	TP	FP
	Negatives (0)	FN	TN

$$K\ statistic = \frac{accuracy\ - p_e}{1 - \ p_e}$$

$$p_e = p_{yes} + \ p_{no}$$

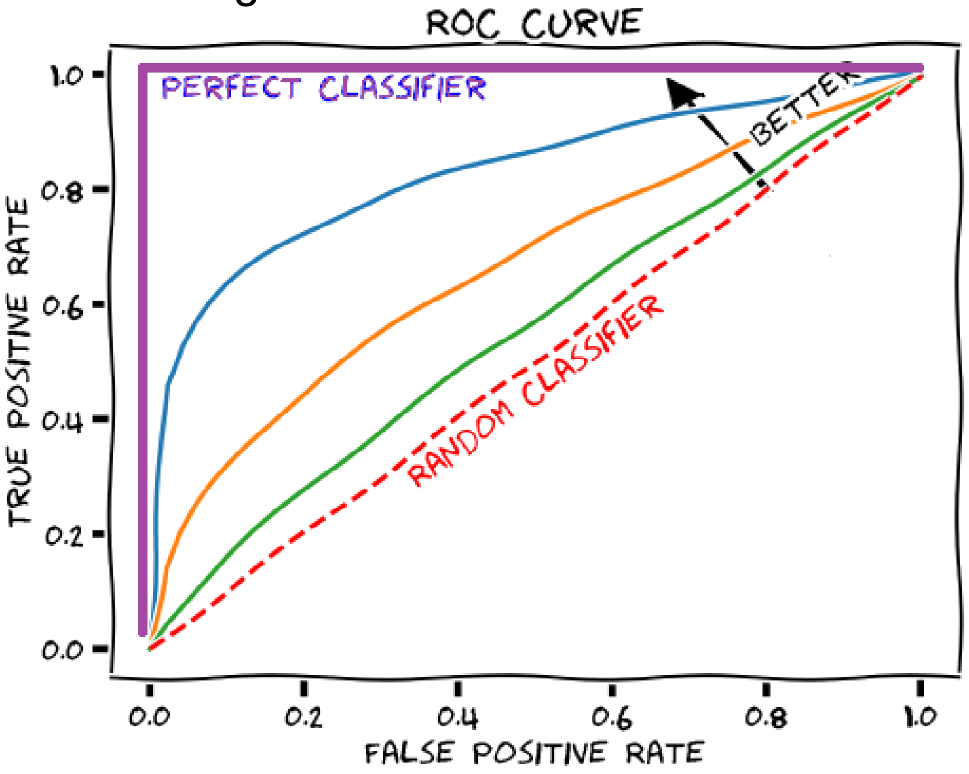
$$p_{yes} = \frac{TP + FP}{TP + FP + FN + TN} * \frac{TP + FN}{TP + FP + FN + TN}$$

$$p_{no} = \frac{TN + FN}{TP + FP + FN + TN} * \frac{TN + FP}{TP + FP + FN + TN}$$

Receiver Operating Characteristic (ROC) is a probability curve, and Area Under the Curve (AUC) is a separability measure that informs the ability of the model to distinguish classes. The higher the AUC, the better the model predicts 0s as being 0s and 1s as being 1s.

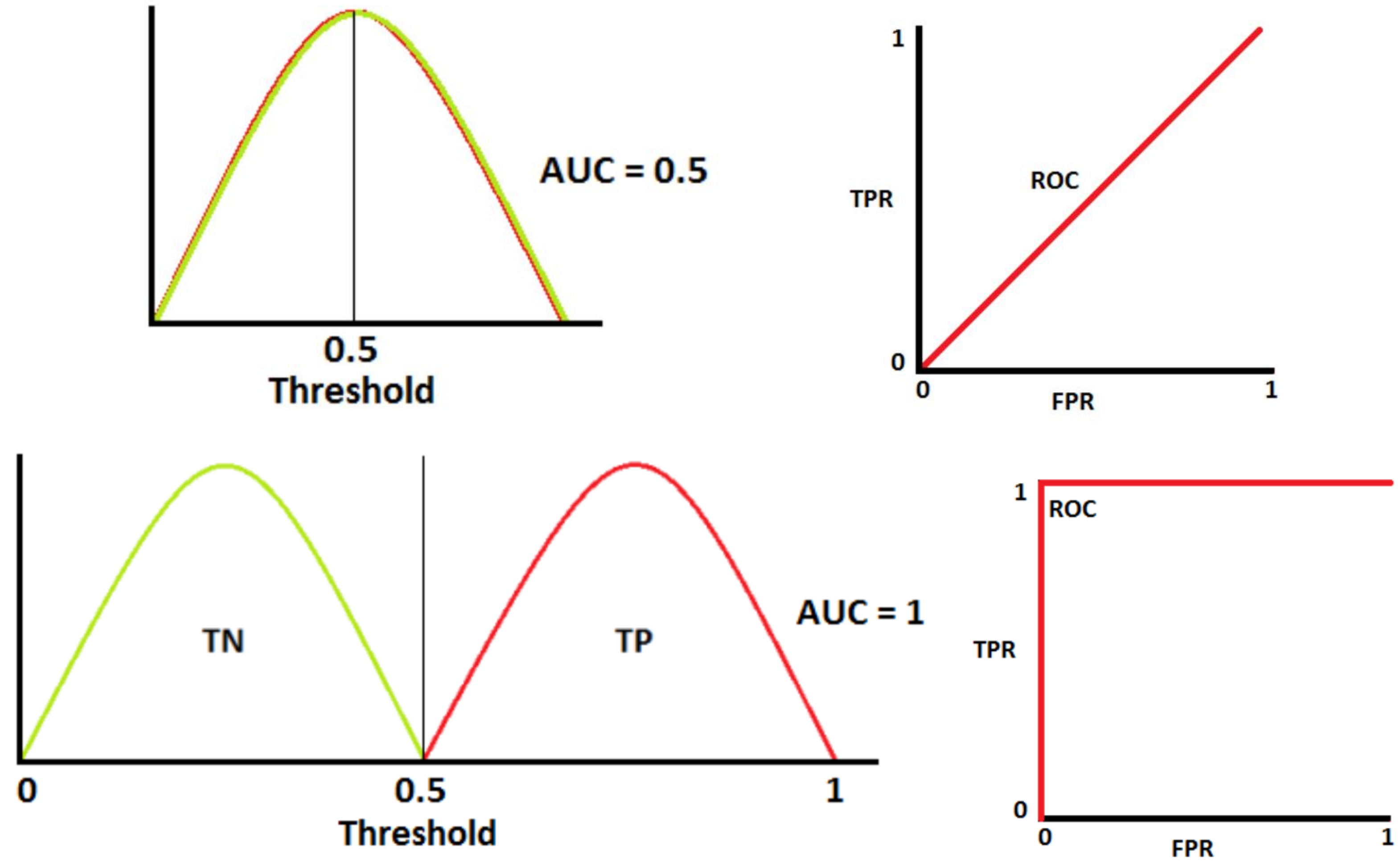
ROC AUC

Recall/Sensitivity

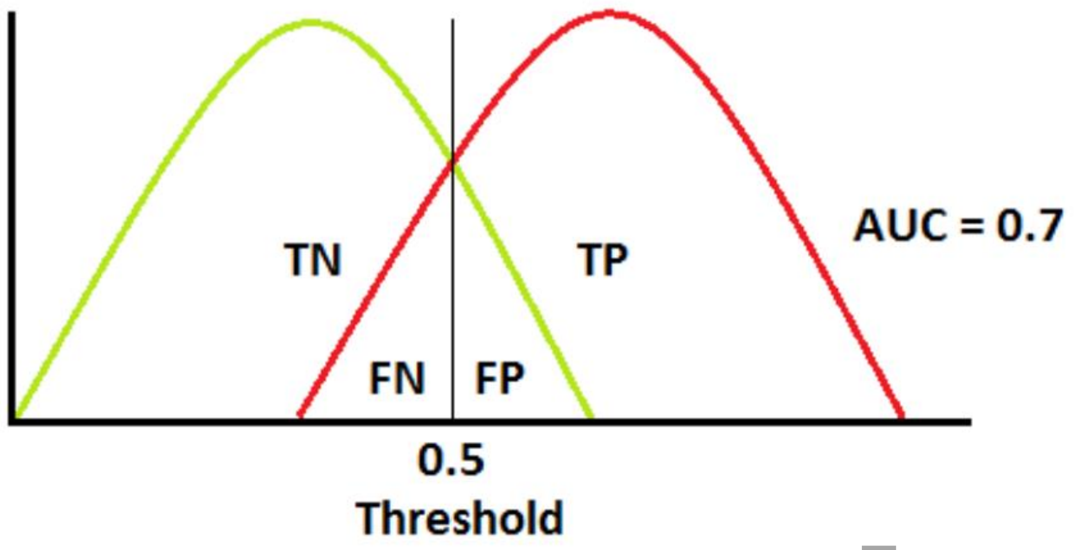


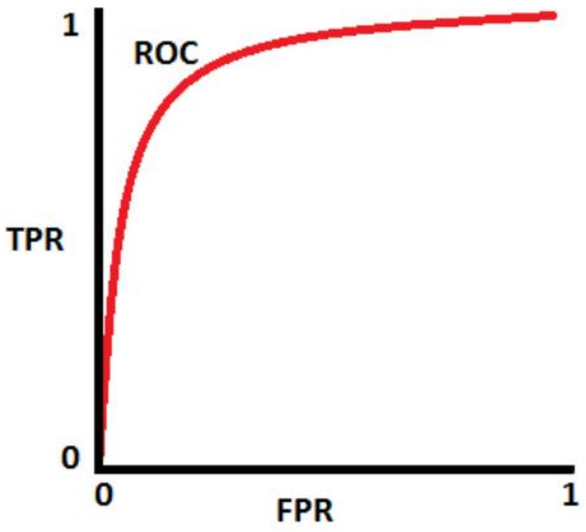
1 - Recall/Sensitivity

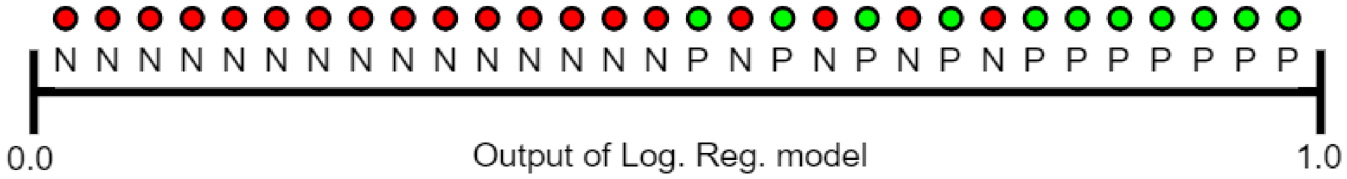
Receiver Operating Characteristic (ROC)



Receiver Operating Characteristic (ROC)







Deployment concerns the tatics to organize, present, and deploy the results of evaluation. Deployment can be as simple as generating a report or as complex as implementing a repeatable data mining process.

1. Implementation of the final models in a real environment.

2. Monitoring and maintenance of the Data Mining models.

WEKA

Waikato Environment for Knowledge Analysis (WEKA):

It is a software that allows large volumes of data to be pre-processed, different machine learning algorithms to be used and different outputs to be compared.

DOWNLOAD:

https://www.cs.waikato.ac.nz/ml/weka/downloading.html

[1] Open Weka / Explorer and load the *“contact-lens.arff”* data set.

[a] How many instances/records does the data set have?
[b] How many attributes/columns does the data set have?
[c] How many and what are the possible values for the "*act*" attribute?
[d] What are the possible values for the attribute "*contact-lens*"?
[e] Which attribute has "*reduced*" as one of the values?

[2] Open Weka/Explorer and load the “*iris.arff*” data set.

[a] How many instances/records does the data set have?
[b] How many attributes/columns does the data set have?
[c] Does the "*iris-setosa*" class tend to have higher or lower "*sepal.length*" values?
[d] Does the “*iris-viginica*” class tend to have higher or lower “*petal.width*” values?
[e] Which of these attributes alone appears to give a better indication of the "*class*"?

DRIVES

Development and Research on Innovative Vocational Education Skills

[3] Open the Weka/Explorer and load the "*weather.nominal.arff*" data set.

[a] What are the attributes of this data set?

[b] The use of classification algorithms may bring specific knowledge through the data presented. Indicate an objective that can be achieved by applying classification algorithms when executed on similar but previously unknown data.

[4] Open Weka/Explorer and load the "*glass.arff*" dataset.

[a] Open the "*Classify*" tab and select the *J48 ("trees")* algorithm.

[b] Observe the "*Confusion Matrix*" and indicate which are the biggest failures in the classification process.

[c] How many "*headlamps*" were classified as "*build wind float*"?

[d] What is the number of instances correctly classified as "*vehic wind non-float*"?

[e] What is the number of instances correctly classified as "*vehic wind float*"?

[f] In the list of obtained results, click with the right button and select "*Visualize tree*". Copy the results to the solution sheet and briefly describe the classification process of the algorithm.

[5] Open Weka/Explorer and load the "labor.arff" data set.

[a] Run the *J48* classification algorithm with the default parameters. Indicate the percentage of correctly classified instances.

[b] Open the *J48* algorithm configuration and set the "*unpruned*" option to "*True*". Run the classification again and indicate the percentage of correctly classified instances.

[6] Open Weka/Explorer and reload the "glass.arff" data set.

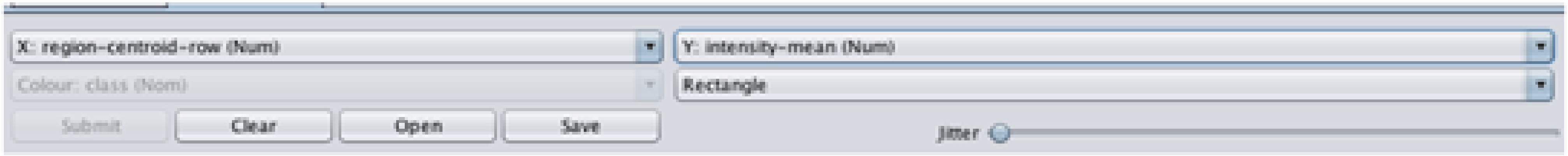
[a] Remove the "*Fe*" attribute. What is the result of the classification?

[b] Remove all attributes except "*Ri*" and "*Mg*". What is the result of the classification?

EXERCISES

[7] On the Weka home screen open the “*package manager*” (Tools -> Package Manager). Install the “*UserClassifier 1.0.3*” package. Open Weka/Explorer and load the "*segment-challenge.arff*" data set. In the Classify tab, set "*segment-test.arff*" as test set.

[a] Use the trees -> UserClassifier and click *Start*; Then change to the Data Visualizer tab and select the following options (another value can be used instead of the rectangle):



Select the possible groups to define and then, determine the result of the classification.

[b] Compare the results obtained with this method of creating a decision tree with the results of the *J48* algorithm.

[8] Open the Weka/Explorer and load the "*segment-challenge.arff*" data set.

[a] Use the *J48* algorithm as classifier and the "*segment-test.arff*" data set as test set. What is the value of the classification?

[b] Use the "*Use training set*" option to determine the classification result. Why should this option not be used to determine the quality and applicability of algorithms to data?

[c] Choose *J48* as the classifier and change the division percentages ("*Percentage Split*") of training and test groups in: 10%, 20%, 40%, 60% and 80%. What do you observe?

[d] Repeat the previous question using 90%, 95%, 98% and 99%. What happens to the number of correctly classified instances? And what happens to the percentage of instances correctly classified? Explain this variation.

[e] Although a percentage of 98% for the training and 2% for the test give a 100% rating, does this mean that the model built is the most suitable for the problem presented?

[f] Based on the experiences above, what is the best estimate of the true accuracy of *J48* for this data set?

EXERCISES

[9] Open the Weka/Explorer and load the "*iris.arff*" data set.

[a] Selecting "*Percentage Split*" at 80% how many instances will be used for training and how many will be used for testing? (The Weka rounds to the nearest integer).

[b] Changing the "*Random seed*" between 1,2,3,4 and 5, and keeping the "*Percentage Split*" at 80%, indicate the minimum and maximum value of incorrectly classified instances (Click on the 'more options' button).

[c] What is the average percentage of correctly classified instances?

[d] If you repeated the exercise [13/b] with 10 "*random seed*" instead of 5 what would be the effect on the average?

[10] Open the Weka/Explorer and load the "*diabetes.arff*" data set.

[a] This data set features 3 classes with 50 instances each. What will be the hit percentage of the ZeroR algorithm?

[b] What is the result of the base line classification when the "*Percentage Split*" method is used at 66%?

[11] Open the Weka/Explorer and load the "*heart-c.arff*" data set. This dataset describes risk factors for heart disease. The attribute *num* represents the class attribute (binary): class < 50 means no disease; class > 50_1 indicates increased level of heart disease. This problem fits into a classification task whose main purpose is to predict heart disease from other attributes in the dataset.

[a] For each attribute, find the following information:

i. The type of attribute, e.g. nominal, ordinal, numeric.
ii. Percentage of missing values in the data.
iii. Max, min, mean and standard deviation.
iv. Are there records that have a value for an attribute that no other record has?

[b] Investigate the possibility of using the Weka AttributeSelection filter to select a subset of attributes with good predictive capability. Save the dataset with the selected attributes in the file heart-c1.arff.

[c] Consider the following methods to deal with missing values and investigate each possibility in Weka.

i. Replace the missing values with the mean of the attribute if the attribute is numeric. Otherwise, replace the missing values with the mode of the attribute (if the attribute is nominal). Use the ReplaceMissingValues filter to do this transformation. Save the data set you have obtained without missing values to the heart-c2.arff file.
ii. Investigate the possibility of using 10-fold cross-validation linear regression to estimate the missing values for each attribute. Note that linear regression can only be applied to numeric attributes. Present the resulting equation and estimate the missing values through this equation. Save the data set you have obtained without missing values in the heart-c3.arff file.

[d] Delete the discrepant records and save the obtained data set without outliers in the heart-c34.arff file. Investigate the possibility of using the Weka - Unsupervised - Attribute - InterquartileRange filter to detect outliers and the Weka - Unsupervised - Instance - RemoveWithValues filter to delete outliers (don't forget to configure the parameters *attributeIndex*, which refers to the outlier index, and *nominalIndices*, which corresponds to the location (first or last) of the nominal value of the attribute you want to remove).

[e] The last step is to use classification algorithms available in Weka to discover hidden patterns in the data. You must repeat the steps described below for each of the data sets created during preprocessing, in addition to using the original dataset.

i. Start with the OneR classifier. Compare the accuracy of the classifier obtained in the training set with the estimated accuracy obtained through the 10 fold-cross validation method. How do you explain this difference?
ii. Use the JRip classifier to create a classifier with and without rule pruning. Which one is the best? Justify your answer.

iii. Use the J48 classifier, i.e. the Weka version of the C4.5 classifier of the decision tree.

- Explore the use of different J48 parameters, such as pruning ("unpruned") and minimum number of records on leaves ("minNumObj").
- Describe the patterns you have obtained and compare them with the conclusions reached in the previous questions.

RapidMiner:

It is a commercial data analysis tool that uses machine learning and can be considered as an alternative to the Weka tool. The main objective of this tool, developed by a company with the same name, is to speed up the process of creating a predictive analysis and make it easier to apply it in practical business scenarios.

DOWNLOAD:

https://rapidminer.com/get-started/

CONTEXT

Sara is a regional sales manager for a national supplier of fossil fuels for home heating.

Recent volatility in market prices for specific heating oil, along with a large variability in the size of each home heating oil order, has been of concern to Sara.

She feels the need to know the types of behavior and other factors that may influence the demand for heating oil in the domestic market.

What factors are related to heating oil use and how can knowledge of these factors be used to better manage inventory and anticipate demand?

Data Mining can help Sara understand these factos and interactions.

BUSINESS UNDERSTANDING

Sara's goal is to better understand how her company can succeed in the home heating oil market.

She recognizes that there are many factors that influence heating oil consumption and believes that by investigating the relationship between these various factors, she will be able to better monitor and respond to the demand for heating oil. Sara decided to select correlation as a way to model the relationship between the factors she wants to investigate.

Correlation is a statistical measure that measures how strong the relationships are between attributes in a data set.

DATA UNDERSTANDING

Using Sara's employer data, extracted primarily from the company's billing database, a data set was created consisting of the following attributes:

- **Insulation**: density rating that ranges from 1 to 10 and indicates how thick each house's insulation is. A house with a density rating of one is poorly insulated, while a house with a density of ten has excellent insulation.
- **Temperature**: average outside ambient temperature for each house in the most recent year, measured in degrees Fahrenheit.
- **Heating_Oil**: total number of heating oil units purchased by the owner of each house in the most recent year.
- **Num_Occupants**: total number of occupants living in each house.
- **Avg_Age**: average age of the occupants living in each house.
- **Home_Size**: rating, on a scale of 1 to 8, of the overall size of the home. The higher the number, the larger the house.

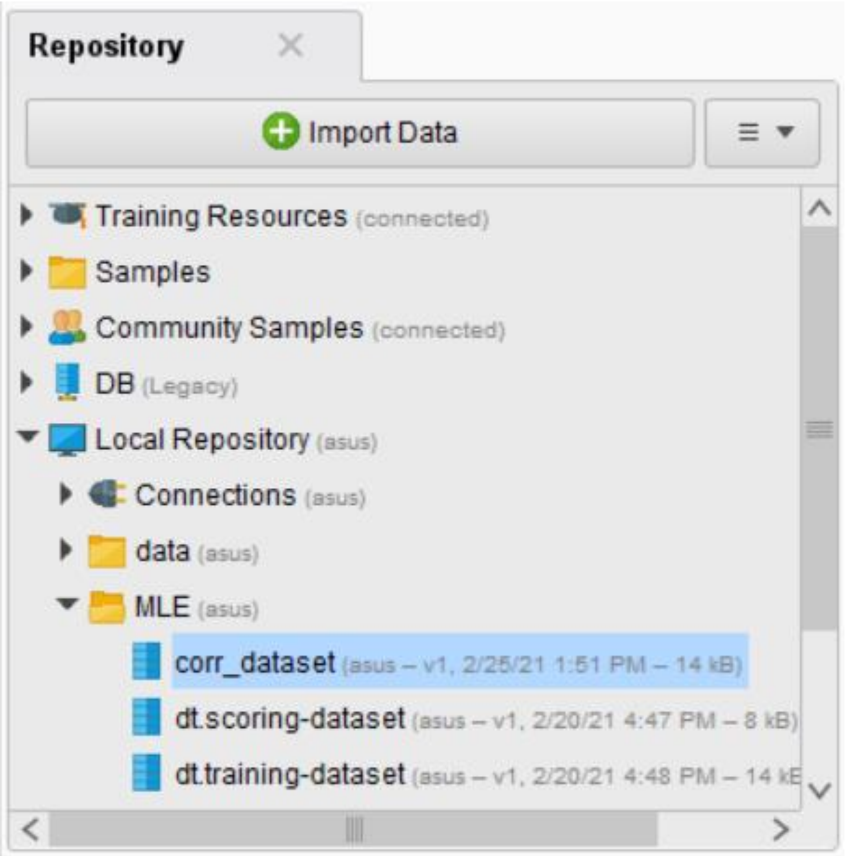
CORRELATION WITH RAPIDMINER

DATA PREPARATION

Download the dataset: corr_dataset.csv

1. Import the CSV to the RapidMiner repository (Import Data -> My Computer)

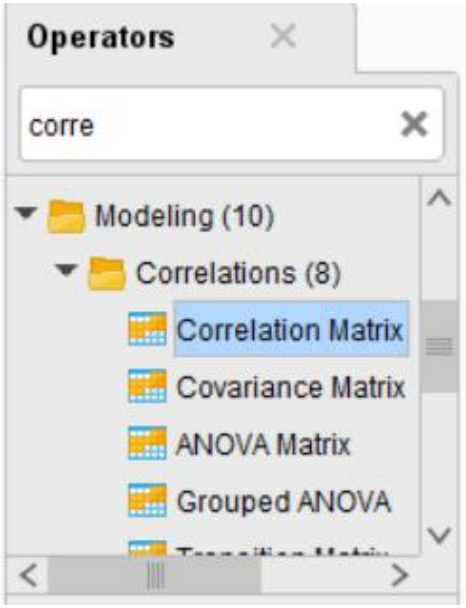
2. Check the results view and inspect the imported CSV data (Data, Statistics)



MODELING

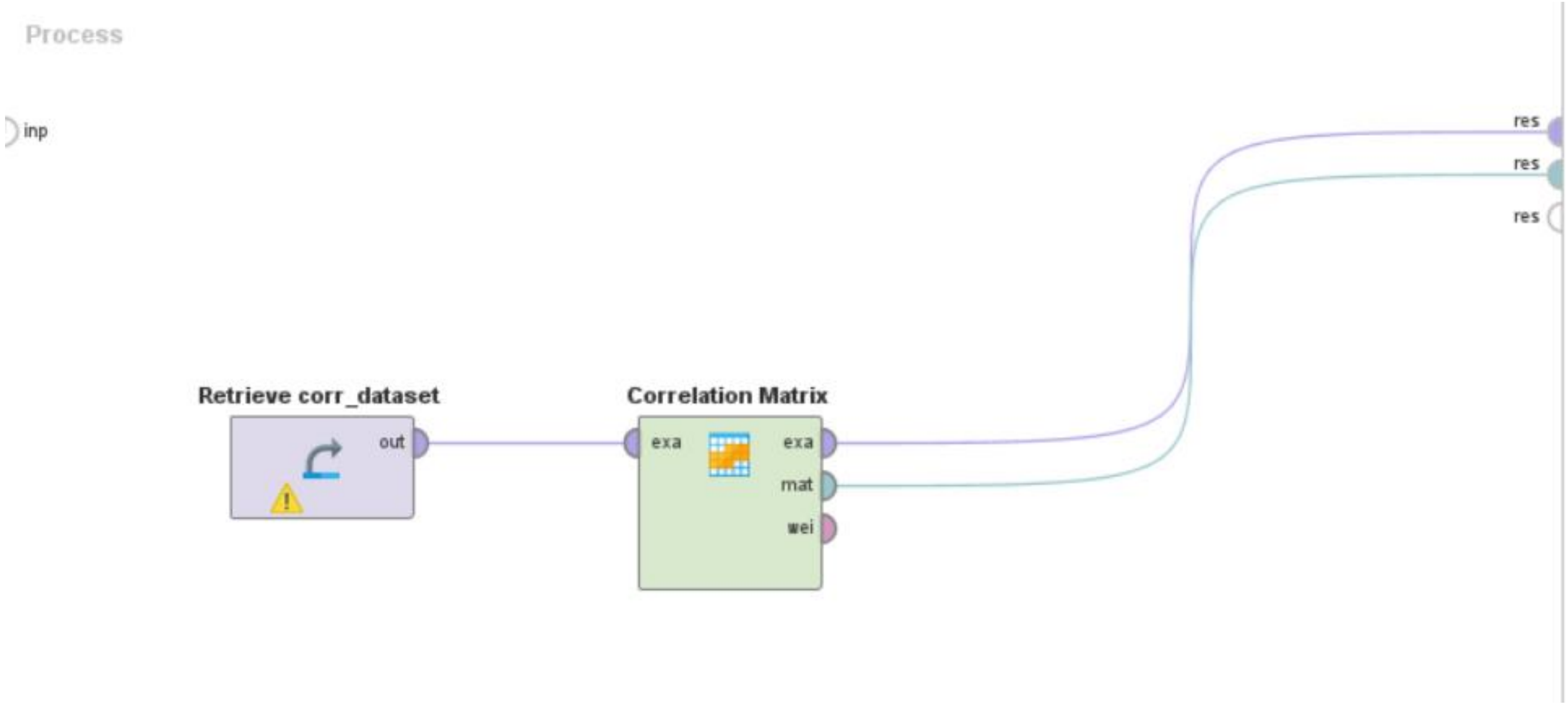
1. Switch to the design perspective and drag the dataset into the process window.

2. In the Operators tab (Data Mining tools section), in the lower left corner, use the search box and type the word 'correlation'. The tool you need is called 'Correlation Matrix'. Drag it to the process window and drop it.



MODELING

3. Establish the connections as shown in the figure. Click Run.



MODELING

Correlation Matrix

Attribut...	Insulation	Temper...	Heating...	Num_O...	Avg_Age	Home_...
Insulation	1	-0.794	0.736	-0.013	0.643	0.201
Tempera...	-0.794	1	-0.774	0.013	-0.673	-0.214
Heating_...	0.736	-0.774	1	-0.042	0.848	0.381
Num_Oc...	-0.013	0.013	-0.042	1	-0.048	-0.023
Avg_Age	0.643	-0.673	0.848	-0.048	1	0.307
Home_S...	0.201	-0.214	0.381	-0.023	0.307	1

EVALUATION

CORRELATION COEFFICIENTS

between 0 and 1

Positive Correlations

between 0 and -1

Negative Correlations

DRIVES
Development and Research on Innovative Vocational Education Skills

EVALUATION

Attribut...	Insulation	Temper...	Heating...	Num_O...	Avg_Age	Home_...
Insulation	1	-0.794	0.736	-0.013	0.643	0.201
Tempera...	-0.794	1	-0.774	0.013	-0.673	-0.214
Heating_...	0.736	-0.774	1	-0.042	0.848	0.38[illegible]
Num_Oc...	-0.013	0.013	-0.042	1	-0.048	-0.023
Avg_Age	0.643	-0.673	0.848	-0.048	1	0.307
Home_S...	0.201	-0.214	0.381	-0.023	0.307	1

The *Heating_Oil consumption* and *Insulation rating level* attributes have a positive correlation of 0.736

What does this mean?

Correlations that are positive mean that as the value of one attribute increases, the value of the other attribute also increases. But a positive correlation also means that as the value of one attribute decreases, the value of the other attribute also decreases.

EVALUATION

When the attribute values move in the same direction, the correlation is positive.

↑ Heating Oil use rises	↑ Insulation rating also rises		↓ Heating Oil use falls	↓ Insulation rating also falls

When the attribute values move in opposite directions, the correlation is negative.

↑ Temperature rises	↓ Insulation rating falls		↓ Temperature falls	↑ Insulation rating rises

EVALUATION

Correlation coefficients not only allow us to determine the relationship between attributes, but also tell us something about the strength of the correlation.

-1 0 1

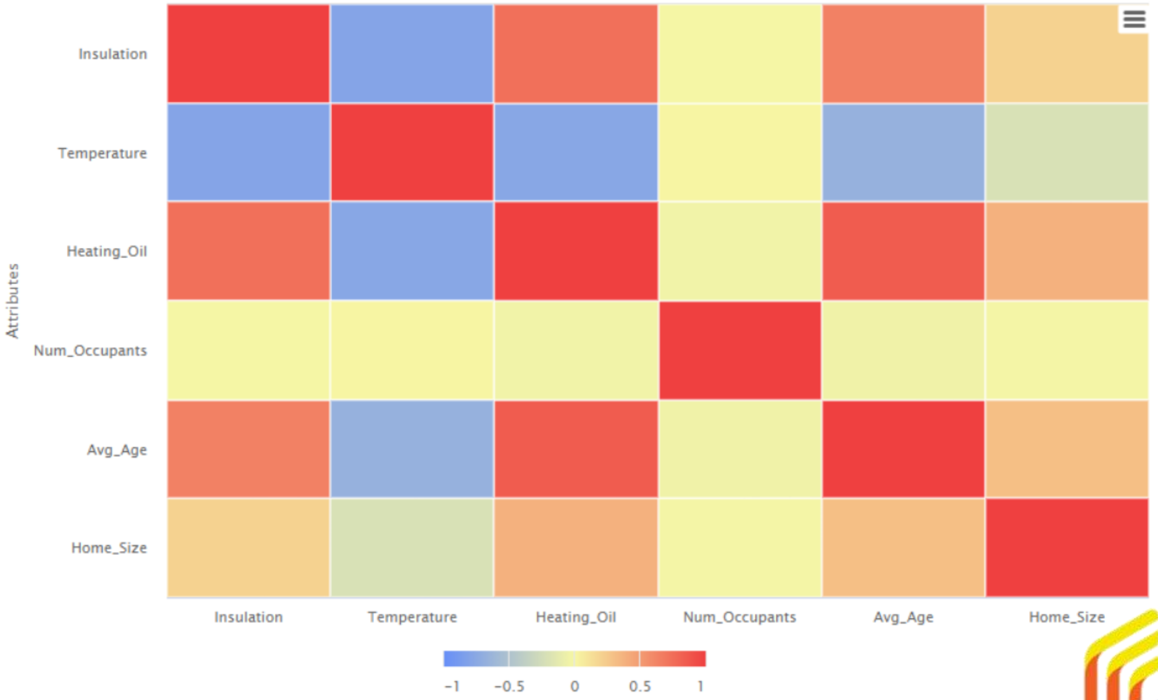
-1 ← -0.8	-0.8 ← -0.6	-0.6 ← -0.4	-0.4 ← 0	0 → 0.4	0.4 → 0.6	0.6 → 0.8	0.8 → 1.0
Very Strong Correlation	Strong Correlation	Some Correlation	No correlation	No correlation	Some correlation	Strong correlation	Very strong correlation

The closer a correlation coefficient is to 1 or -1, the stronger the correlation of the attributes.

EVALUATION

RapidMiner helps recognize strong correlations by color-coding both the *Data* tab and *the Matrix Visualization* tab.

Attribut...	Insulation	Temper...	Heating...	Num_O...	Avg_Age	Home_...
Insulation	1	-0.794	0.736	-0.013	0.643	0.201
Tempera...	-0.794	1	-0.774	0.013	-0.673	-0.214
Heating_...	0.736	-0.774	1	-0.042	0.848	0.381
Num_Oc...	-0.013	0.013	-0.042	1	-0.048	-0.023
Avg_Age	0.643	-0.673	0.848	-0.048	1	0.307
Home_S...	0.201	-0.214	0.381	-0.023	0.307	1



DRIVES
Development and Research on Innovative Vocational Education Skills

DEPLOYMENT

With this study, it was possible to see that the two most strongly correlated attributes are *Heating_Oil* and *Avg_Age*, with a coefficient of 0.848.

As the average age of the occupants of a house increases, so does the use of heating oil in that house. **Why?**

The assumption that a correlation proves causality is dangerous and often false!

DRIVES
Development and Research on Innovative Vocational Education Skills

DEPLOYMENT

The correlation coefficient between *Avg_Age* and *Temperature* is -0.673, i.e, a strong negative correlation.

"As the age of the residents of a house increases, the outside temperature decreases; and as the temperature increases, the age of the residents decreases."

Although statistically there is a correlation between these two attributes, there is no logical reason why the average age of the occupants of a house should have any effect on the external temperature of the house and vice versa.

The assumption that a correlation proves causality is dangerous and often false!

DEPLOYMENT

Another false interpretation is that correlation coefficients are percentages (%).

A correlation coefficient of 0.776 ≠ 77.6% variability between these attributes.

The mathematical formula underlying the calculation of the correlation coefficients measures only the strength of the interaction between the attributes, as indicated by the proximity of 1 or -1.

DEPLOYMENT

The concept of deployment in Data Mining means to do something with the results of the model, that is, to take some action based on what the model has learned.

There are several things Sara can do to act on the model obtained:

- Remove the **Num_Occupants** attribute
- Investigate the role of home **insulation**
- Increase the **granularity** of the data set
- Add **attributes** to the data set

DEPLOYMENT

Remove the **Num_Occupants** attribute

The number of people living in a house might logically seem like a variable that influences energy use, but this attribute did not correlate significantly with any other attributes.

Investigate the role of home **insulation**

The Insulation attribute was highly correlated with a number of other attributes. This means that there may be an opportunity to partner with a company that specializes in adding insulation to existing homes or even create your own.

DEPLOYMENT

Increase the **granularity** of the data set

This dataset has low granularity attributes such as the average annual temperature. Temperatures fluctuate throughout the year and therefore monthly, or even weekly, measurements would show more detailed results that are closer to reality.

Add **attributes** to the data set

For example, perhaps the number of instruments that consume heating oil in each house, such as furnaces and/or boilers, would add something to Sara's study.

DRIVES
Development and Research on Innovative Vocational Education Skills

CORRELATION

[1] What are the main limitations of correlation models?

[2] What is a correlation coefficient and how is it interpreted?

[3] What is the difference between a negative correlation and a positive correlation?

[4] How is the strength of a correlation measured? What are the limits to that strength?

[5] Access the file *mpg_dataset.csv*. Run the Data Preparation step. Don't forget to check for outliers and missing values. Once properly processed, save the data to a .csv file that allows you to run the correlation process in the rapidminer.

[6] Document which attributes may influence or explain fuel consumption/efficiency in a given vehicle (mpg).

CONTEXT

Peter is the city manager of a medium-sized but constantly growing city. Like most municipalities, the city has limited resources in the face of the needs it encounters.

Peter knows that the citizens of the community are active in various community organizations such as churches, social clubs, and hobby enthusiasts, and he believes that these groups can work together to meet some of the community's needs.

Before he starts asking community organizations to start working together, Peter needs to find out if there are natural associations between the different types of groups.

Data Mining can help Peter understand these associations.

BUSINESS UNDERSTANDING

Peter's goal is to identify and take advantage of existing connections in his local community to do some work that benefits the whole community.

Peter and his family are involved in a broad group of community organizations, so he is aware, in a more general sense, of the diversity of the groups as well as their interests, goals, and potential contributions.

Identifying individuals to work with in each church, social club, or political organization will be overwhelming without first categorizing the organizations into groups and looking for associations among them.

Association rules are a Data Mining methodology that seeks to find frequent links between attributes in a data set.

BUSINESS UNDERSTANDING

Association rules are common when doing shopping basket analysis. Merchants and suppliers in various industries use this data mining approach to try to find which products are frequently bought together. For example, when buying a smartphone, accessories such as screen protectors, chargers or earbuds are often recommended.

Recommended items are identified through association rule techniques between items that previous customers have bought together with the item that the client is buying. This happens when the association is so frequent in the dataset that the association can be considered a rule. Thus, the name of this Data Mining approach is "association rules".

DATA UNDERSTANDING

Using Peter's knowledge of the local community, a short questionnaire was created and administered online via a website. The leaders of each organization were invited to participate in the study and share it with their group members. After the questionnaire was completed, a data set was created consisting of the following attributes:

- **Elapsed_Time:** time the person spent to complete the questionnaire. It is expressed in decimal minutes (4.5 in this attribute would be four minutes and thirty seconds).
- **Time_in_Community:** time the person has lived in the area for 0-2 years, 3-9 years or 10+ years. It is recorded in the data set as "Short", "Medium", or "Long", respectively.
- **Gender:** gender of the person.
- **Working:** yes/no answer indicating whether or not the person is currently in paid employment.
- **Age:** age of the person in years.
- **Family:** yes/no response indicating whether or not the person is a member of a family-oriented community organization, such as recreational or sports leagues for children, genealogy

DATA UNDERSTANDING

- **Hobbies:** yes/no response indicating whether or not the person is currently a member of a hobby-oriented community organization, such as amateur radio, outdoor recreation, motorcycle/bicycle riding.

- **Social_Club:** yes/no response indicating whether or not the person is a member of a community social organization.

- **Political:** yes/no answer indicating whether or not the person is a member of a political organization with regular meetings in the community, such as a political party.

- **Professional:** yes/no answer indicating whether or not the person is a member of a professional organization with local committee meetings, such as a committee of a law or medical society, a small business group.

- **Religious:** yes/no answer indicating whether or not the person is currently a member of a church in the community

- **Support_Group:** yes/no response indicating whether or not the person is a member of a

DRIVES
Development and Research on Innovative Vocational Education Skills

DATA PREPARATION

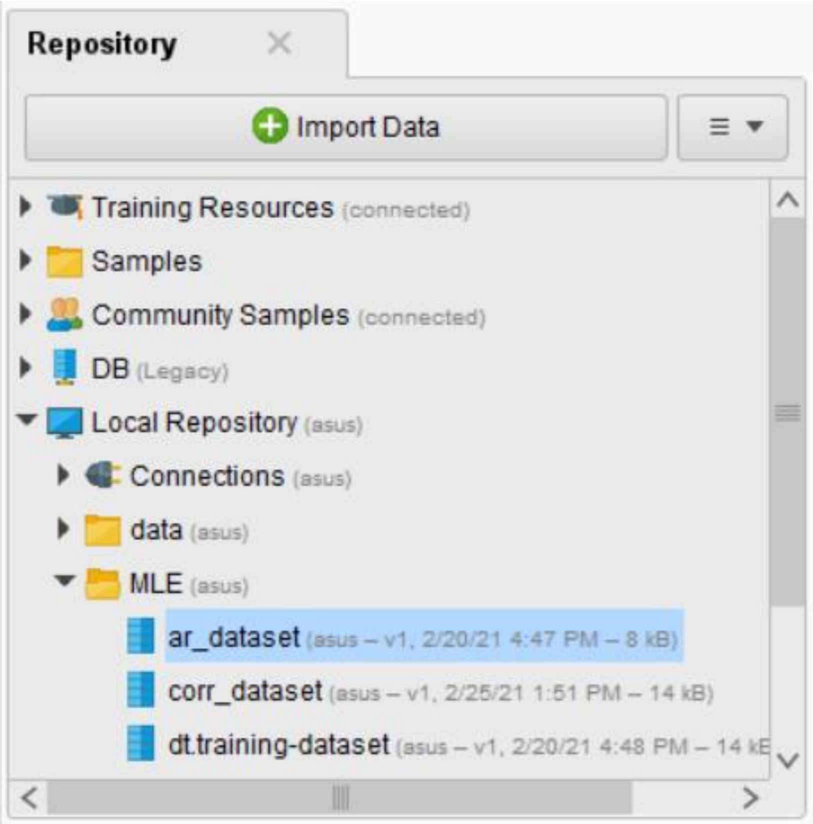
Download the dataset: ar_dataset.csv

1. Import the CSV to the RapidMiner repositor (Import Data -> My Computer).

2. Verify the results view and inspect the CSV data imported (Data, Statistics).

3. Drag the **ar_dataset** to a new process window in the RapidMiner.

4. Execute the model to inspect the data.

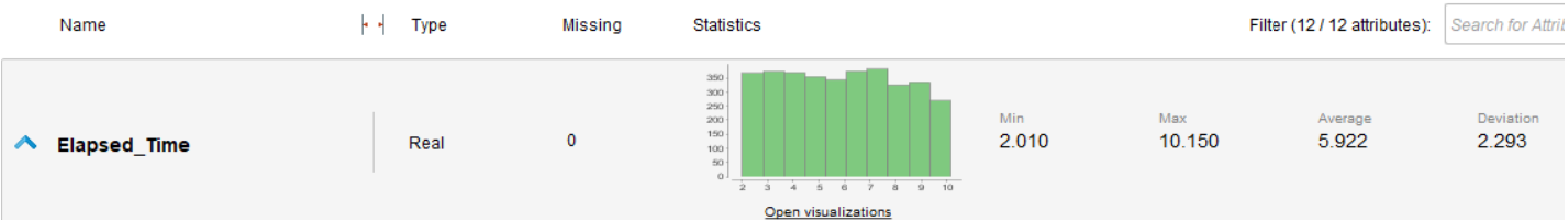


DATA PREPARATION

5. Select the “Results” view and choose the “Statistics” option. Note that:

- There is no missing value for any of the 12 attributes.

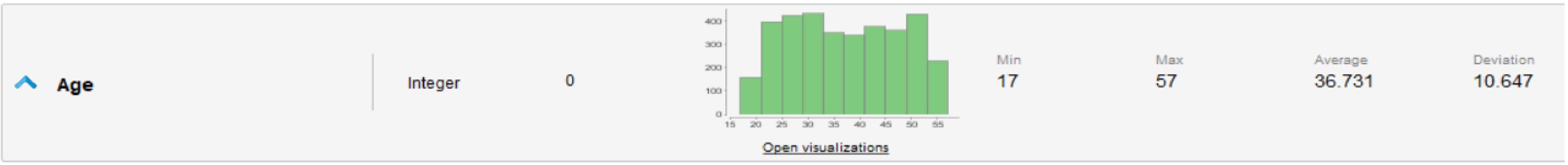
- For the numerical data, RapidMiner presents the minimum, maximum, mean and standard deviation value for each attribute.



DATA PREPARATION

- Any value less than two standard deviations below the mean or two standard deviations above the mean is statistically considered an outlier. For example, in the “Age” attribute, the mean is 36.731, while the standard deviation is 10.647. Two standard deviations above the mean would be 58.025 (36.731+(2*10.647)) and two standard deviations below the mean would be 15.437 (36.731-(2*10.647)).

- By looking at the Min and Max value, you can see that the attribute has a range of 17 to 57, so all instances are within two standard deviations above and below the mean -> there are no outliers.



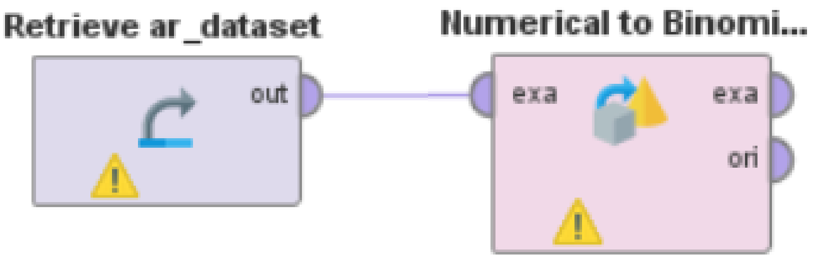
It is important to know that while two standard deviations is a guideline, it is not a universal rule.

DATA PREPARATION

- Attributes of type yes/no were recorded as 0 or 1 and imported as 'integer’.

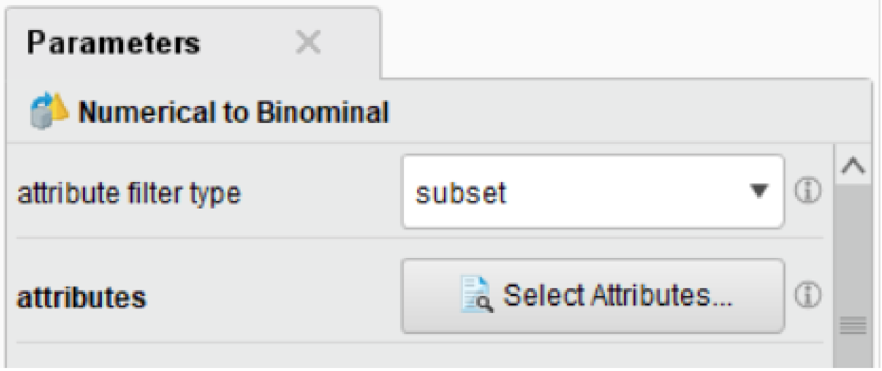
RapidMiner's association rule operators require attributes to be of data type 'binominal’.

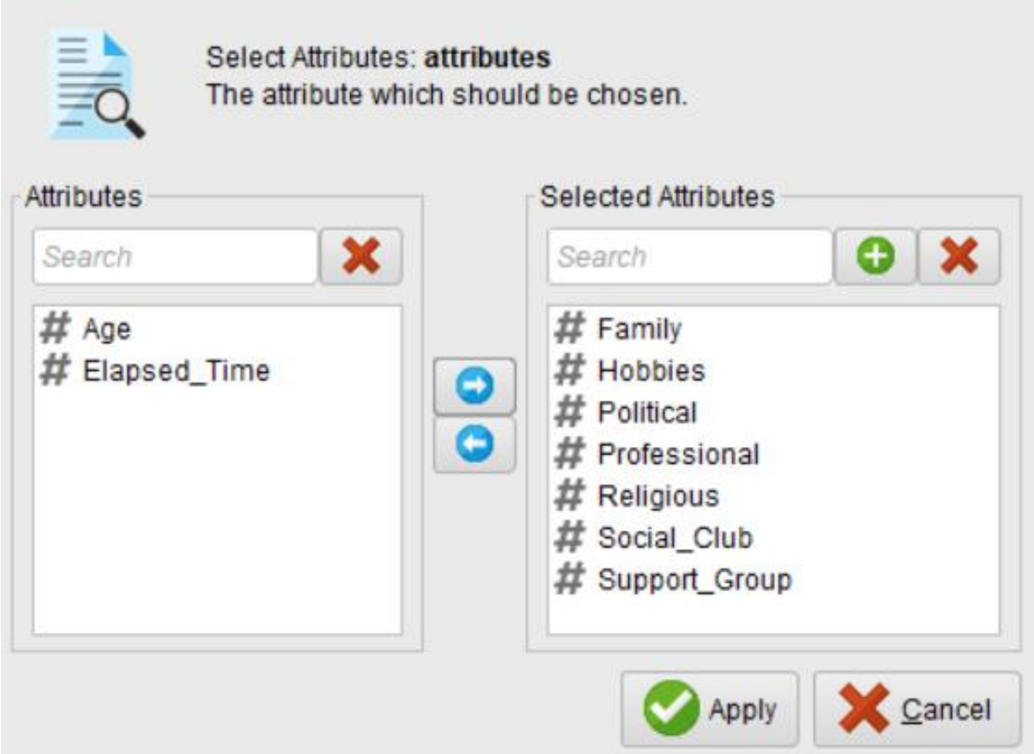
6. Go back to the "Design" view. In the Operators box, search for "Numerical to Binomial" and add this operator to the process window.



DATA PREPARATION

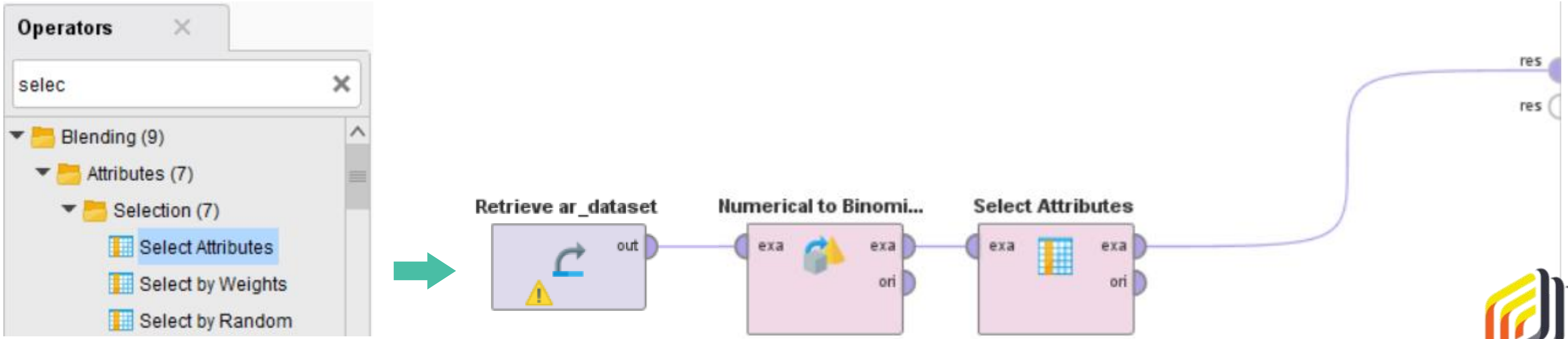
7. In the process window, click on the "Numerical to Binomial" operator. On the right side panel, entitled *Parameters*, change the attribute filter type to "subset" and then select the "Select Attributes" option. Select the following attributes for inclusion: Family, Hobbies, Social_Club, Political, Professional, Religious, Support_Group.





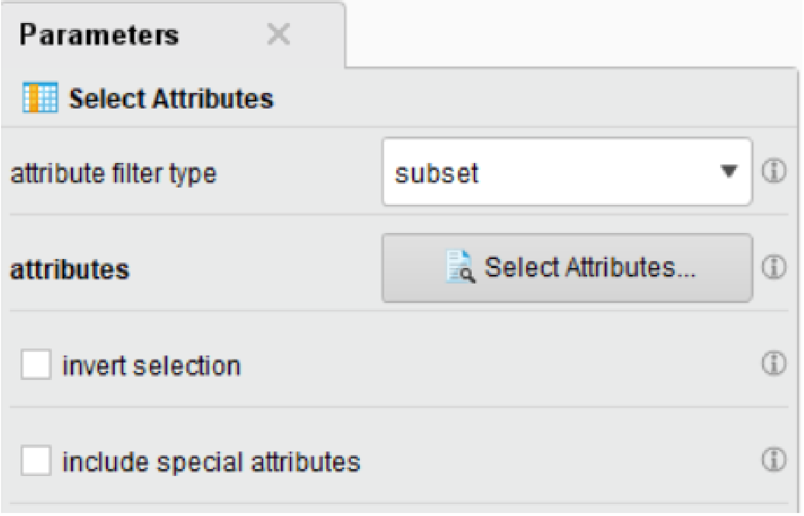
DATA PREPARATION

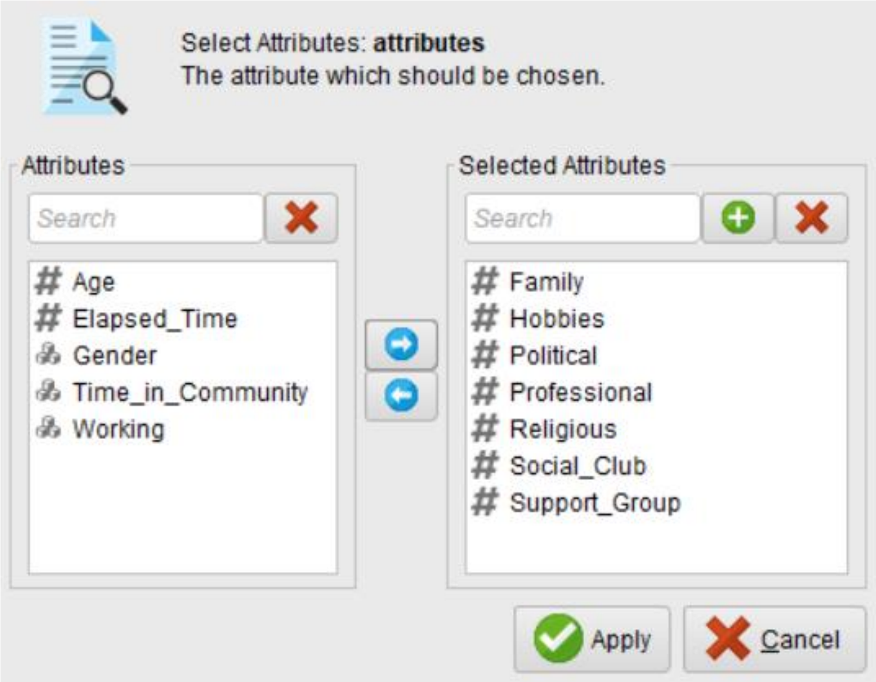
8. The number of attributes in the data set needs to be reduced. The time that each person took to complete the survey is not relevant in the context of the problem, as well as other attributes such as gender and age. Add a Select Attributes type operator and drag it to the process window.



DATA PREPARATION

9. In the process window, click on the Select Attributes operator. On the right side panel, entitled *Parameters*, change the attribute filter type to "subset" and then select the "Select Attributes" option. Select the following attributes for inclusion: Family, Hobbies, Social_Club, Political, Professional, Religious, Support_Group.





DATA PREPARATION

10. Click in the ‘play’ button to run the model.

Row No.	Family	Hobbies	Social_Club	Political	Professional	Religious	Support_Gr...
1	true	false	false	false	false	false	false
2	false	false	false	false	false	true	true
3	true	true	false	false	true	false	false
4	false	false	false	false	false	false	false
5	false	false	false	true	true	false	true
6	false	false	false	false	true	false	false
7	false	false	false	false	false	false	true
8	true	true	true	false	false	true	false

Values of 1 or 0 are now reflected as 'true' or 'false’, respectively.

In RapidMiner, the 'binominal' data type is used instead of 'binomial'. Binomial means one of two numbers (usually 0 and 1). Binominal, on the other hand, means one of two values that can be either numeric or character-based.

MODELING

RapidMiner features several association rule operators. In this example the FP-Growth operator will be used.

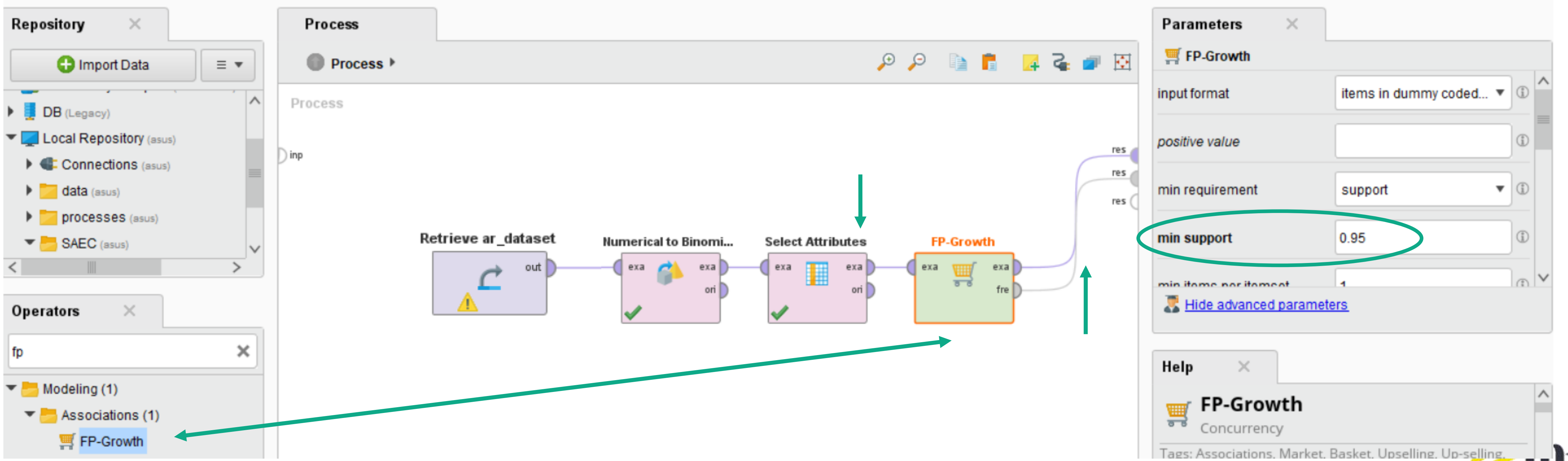
FP (*Frequency Pattern*)

↓

Without having the frequency of attribute combinations, we could not determine whether any of the patterns in the data occur often enough to be considered a rule.

MODELING

1. Drag operator *FP-Growth* into the process. Note the *min support* parameter on the right side. Make sure that the *exa* and *fre* ports are connected to the *res* ports.



Exa port → will generate an example tab (observations and statistics from the dataset)

Fre port → will generate an array of any frequent patterns that the operator can find in the data

MODELING

2. Run the model and select the results tab.

No. of Sets: 6

Total Max. Size: 2

Min. Size: 1

Max. Size: 2

Contains Item:

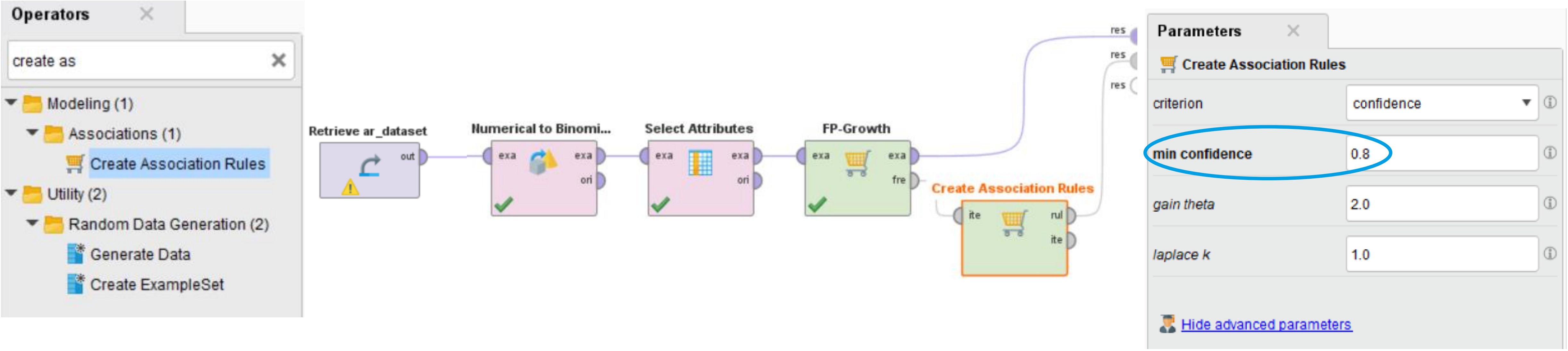
Update View

Size	Support	Item 1	Item 2
1	0.419	Religious	
1	0.390	Family	
1	0.324	Professional	
1	0.300	Hobbies	
2	0.225	Religious	Family
2	0.239	Religious	Hobbies

Religious organizations may have some natural connections to family and hobbies organizations.

MODELING

3. We can use the Create Association Rules operator to investigate these relationships. This operator uses data from the pattern frequency matrix and looks for patterns that occur often enough to be considered rules. Look for this operator, drag it to the process (as in the image) and run the model.



MODELING

Result: No association rules were found.

The CRISP-DM process is cyclical in nature and sometimes it is necessary to go back and forth between steps before creating a model that produces results.

EVALUATION

Confidence Percentage

How confident are we that when an attribute is flagged as true, the associated attribute will also be flagged as true?

Premise → Conclusion

Support Percentage

It is the number of times the rule occurred, divided by the number of observations in the dataset (in percent).

EVALUATION

Example:

10x	Milk in 7 Cookies in 4	Milk and cookies together in 3

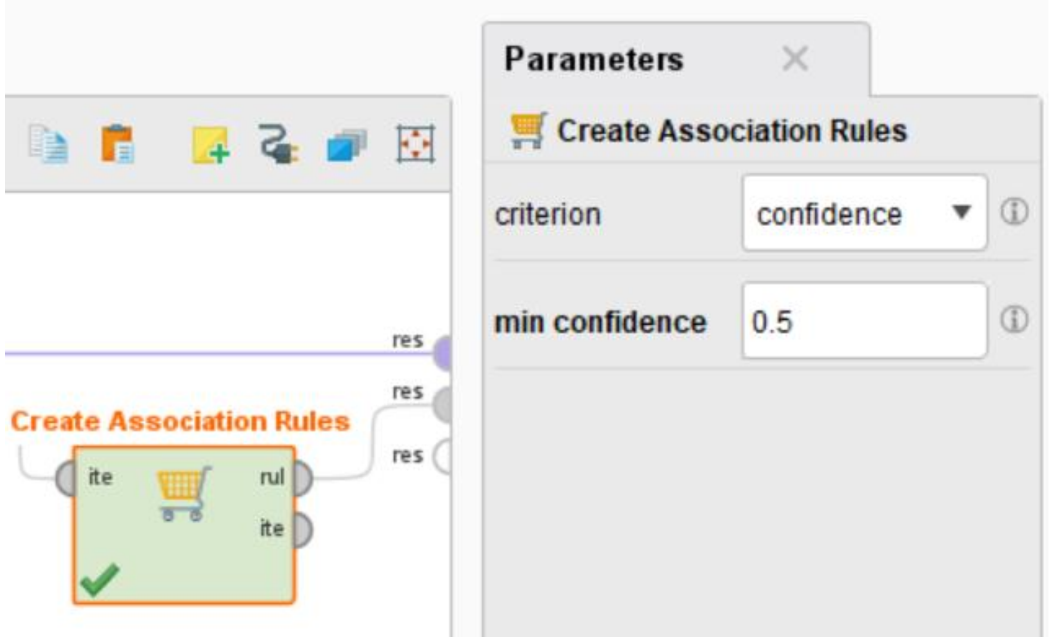
Cookies → Milk
They could have matched on 4 carts,
but they only matched on 3
3/4 → 0.75 → 75% confidence

Milk → Cookies
They could have matched on 7 carts,
but they only matched on 3
3/7 → 0.429 → 43% confidence

3/10 → 0.3 → 30% of support

EVALUATION

In the Design tab, click in the Create Association Rules Operator and change the minimum confidence parameter to 0.5 → any association with at least 50 percent confidence should be displayed as a rule.



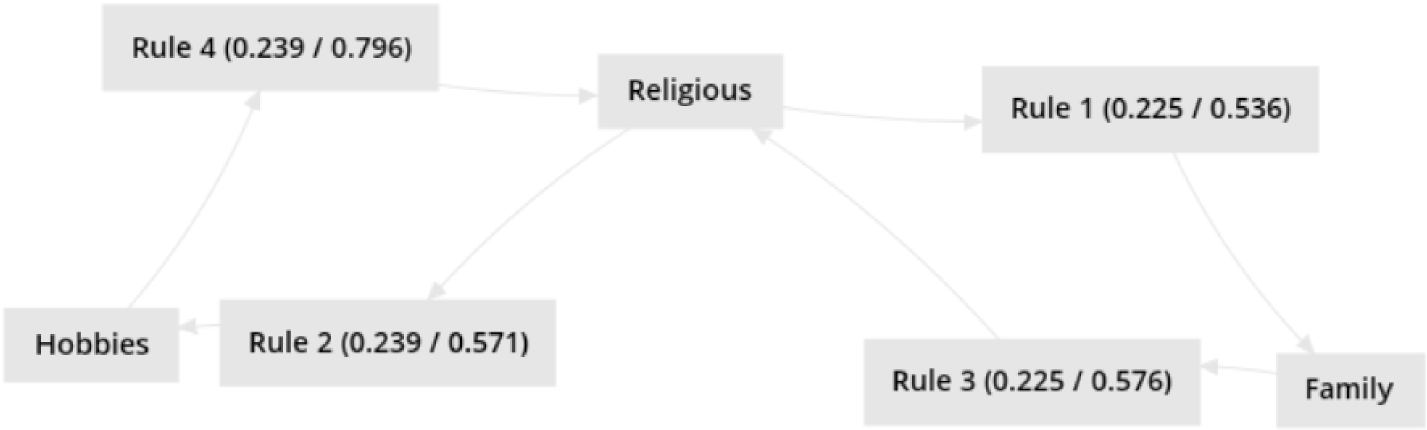
EVALUATION

No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Convicti...
1	Religious	Family	0.225	0.536	0.863	-0.613	0.061	1.376	1.316
2	Religious	Hobbies	0.239	0.571	0.873	-0.598	0.113	1.902	1.630
3	Family	Religious	0.225	0.576	0.881	-0.555	0.061	1.376	1.371
4	Hobbies	Religious	0.239	0.796	0.953	-0.361	0.113	1.902	2.852

Min. Criterion:

confidence

Min. Criterion Value:



EVALUATION

- The hunch that religious, family and hobby organizations are related was correct;
- Rule number 4 has a confidence percentage of almost 80%;
- The other associations have lower confidence percentages, but are still very good;
- We can see that each of the four rules are supported by more than 20% of the observations in the dataset;
- % support: rule 1 = rule 3 and rule 2 = rule 4
- % confidence: rule 1 = rule 2 = rule 3 = rule 4

DRIVES
Development and Research on Innovative Vocational Education Skills

DEPLOYMENT

Are there connections between the types of community groups?

Yes, the church, family, and hobby organizations in the community have some members in common.

It seems that Peter will have better luck finding groups that will collaborate on projects around town involving church, hobby, and family related organizations.

ASSOCIATION RULES

[1] What are association rules? What are they good for?

[2] What are the two main metrics calculated in association rules and how are they calculated?

[3] What type of data must the attributes of a dataset be to use the Frequent Pattern operator in RapidMiner?

[4] Download and import the order.csv dataset to RapidMiner. Perform the steps in the Data Preparation step on your data set as needed. Make sure all your variables have consistent data and that your data types are appropriate for the FP-Growth operator.

[5] Generate association rules for the dataset. Modify the values of *min confidence* and *min support* in order to identify the optimal levels to get interesting rules with reasonable confidence and support values. Analyze other measures of rule strength, such as LaPlace or Conviction. What rules have you found? Which attributes are most strongly associated? Are any of your association rules good enough to the point that you can rely on them to make decisions? Why?

ASSOCIATION RULES

[6] Create a new association rules model using the same dataset, but this time use the RapidMiner WFPGrowth operator. In order to be able to use this operator, first install "Weka Extension" from Extensions-> Marketplace. Present the results and discuss them. (Tips: (1) This operator shall establish its own rules without the assistance of other operators; (2) The support and confidence parameters of this operator shall be identified as U and C, respectively).

[7] The Apriori algorithm is often used for associations in the data mining process. Search for Apriori (W-Apriori) in the operators of RapidMiner and add it to your dataset in a new process. Use the Help tab in the lower right corner of RapidMiner to learn about this operator's parameters and functions. Present your results and discuss them.

DRIVES
Development and Research on Innovative Vocational Education Skills

CONTEXT

Olivia is a program director for a large health insurance provider.

Recently, Olivia has been reading medical journals and other articles and found a strong emphasis on the influence of gender, weight and cholesterol on the development of coronary heart disease.

Olivia has decided to propose that her company offer weight and cholesterol control programs to individuals who receive health insurance through the company.

As she considers where her efforts might be most effective, she begins to wonder if there are groups of individuals who are at higher risk of high weight and high cholesterol, and if these groups exist, where the natural dividing lines between the different groups occur.

Data Mining can help Olivia understand these groups.

BUSINESS UNDERSTANDING

Olivia's goal is to identify and contact insured people at high risk of coronary heart disease due to their high weight and/or high cholesterol levels. Olivia knows that people at low risk, i.e. those with low weight and low cholesterol, are unlikely to participate in the programs that her company will offer.

She also understands that there are likely to be policyholders with high weight and low cholesterol, others with high weight and high cholesterol, and others with low weight and high cholesterol. She further recognizes that there are likely to be many people somewhere in between these types.

In order to achieve her goal, Olivia needs to search among thousands of policyholders to find groups of people with similar characteristics and to set up programs that are relevant and attractive to people in these different groups.

DATA UNDERSTANDING

Using the insurance company's database, Olivia extracted three attributes for 547 randomly selected individuals. The three attributes are the person's weight in pounds as recorded at the person's most recent medical examination, their last cholesterol level as determined by blood test, and their gender. As is typical in many datasets, the gender attribute uses 0 to indicate Female and 1 to indicate Male.

We will use this data to create a clustering model to help Olivia understand how her company's customers seem to group together based on their weights, genders, and cholesterol levels.

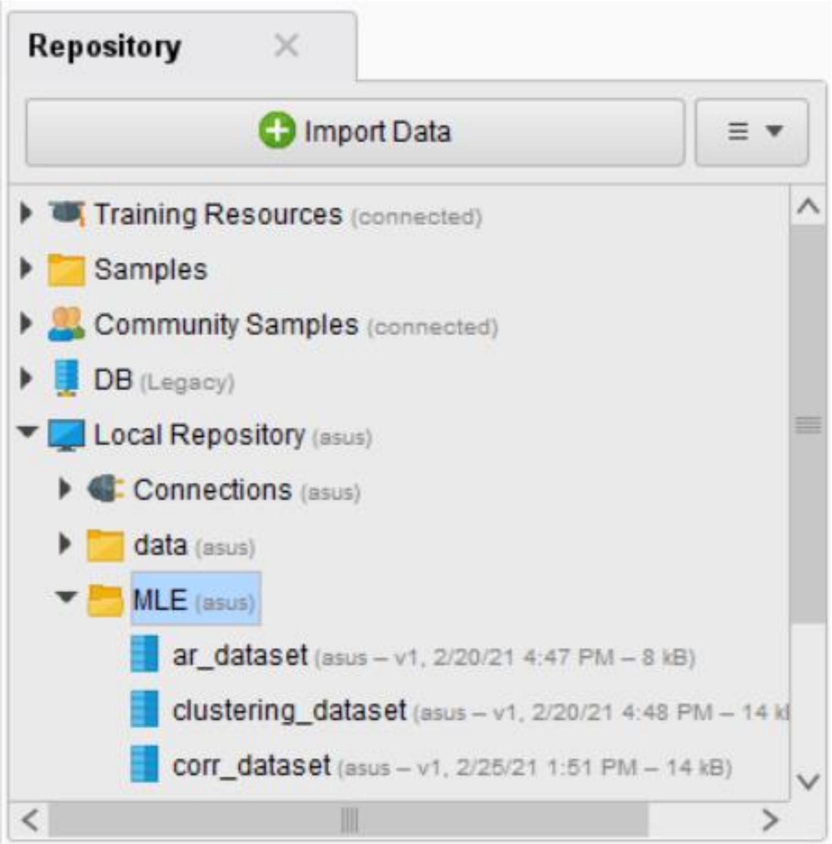
It should be remembered that, in doing so, averages are particularly susceptible to undue influence from extreme values, so it is very important to identify inconsistent data when using the k-Means clustering data mining methodology.

DATA PREPARATION

Download the dataset: clustering_dataset.csv

1. Import the CSV into the RapidMiner repository (Import Data -> My Computer);

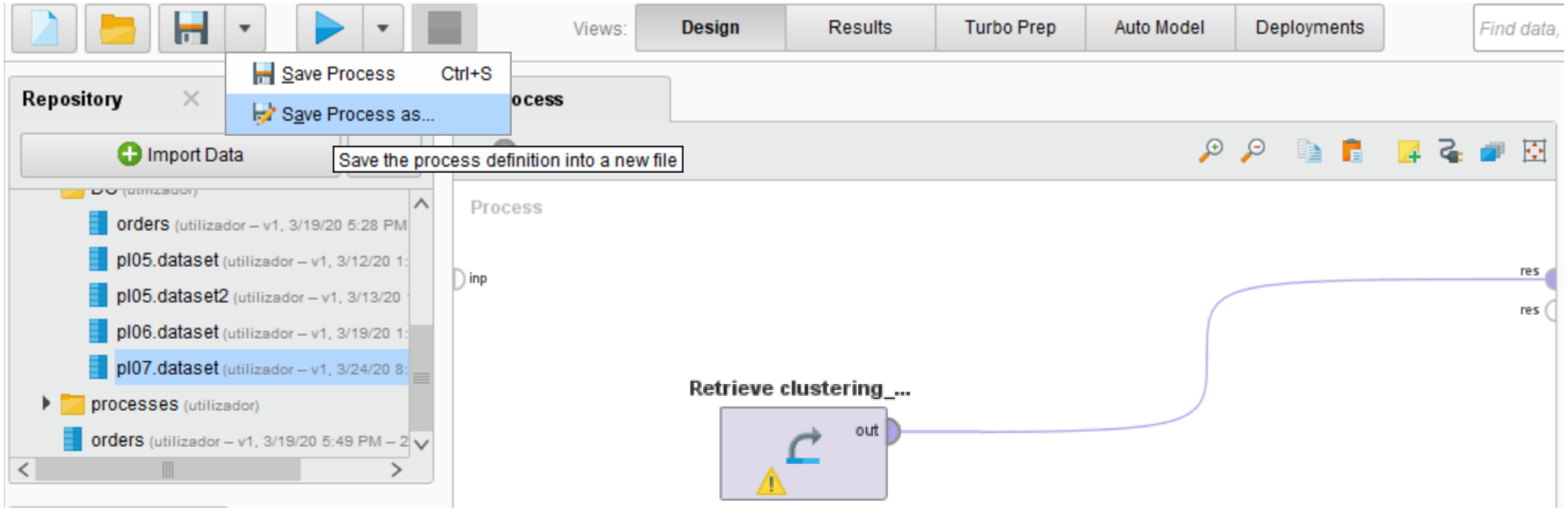
2. Verify the results *view* and inspect the CSV data imported (Data, Statistics);



DATA PREPARATION

3. Drag the clustering_dataset into a new process window in RapidMiner;

4. Run the model to inspect the data and save the process;



DRIVES
Development and Research on Innovative Vocational Education Skills

DATA PREPARATION

5. Select the "Results" view and choose the "Statistics" option. Note that:

- There are no missing values for any of the 12 attributes.
- None of the values appear to be inconsistent.

DRIVES
Development and Research on Innovative Vocational Education Skills

MODELING

The “k” in k-means refers to the number of sets, groups or clusters. The aim of this data mining methodology is to analyze each observation for individual attribute values and compare them to the averages of potential clusters of other observations in order to identify natural groups that are similar to each other.

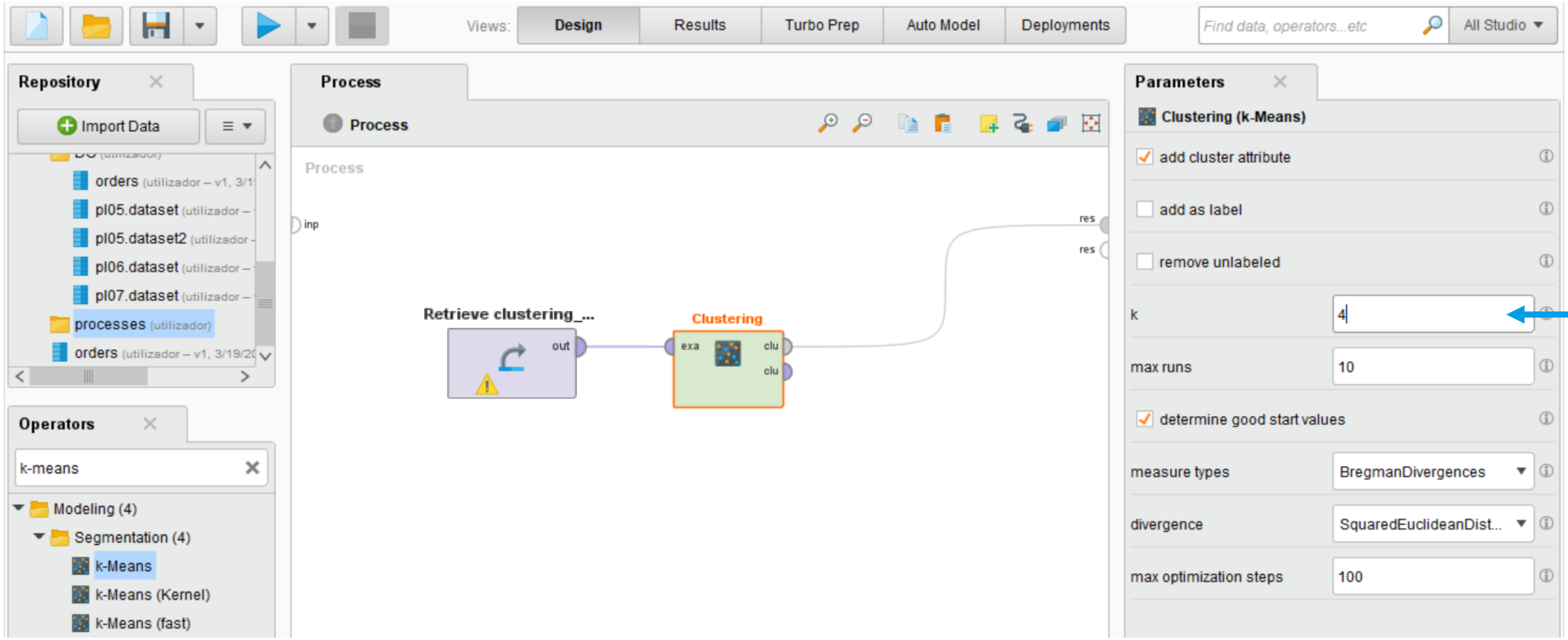
The k-means algorithm does this by sampling a set of dataset observations, averaging each attribute for the observations in that sample, and then comparing the other attributes in the dataset with the averages in that sample.

The algorithm repeatedly does this to find the best matches and then formulates groups of observations that become clusters. As the calculated averages become increasingly similar, clusters are formed, and each observation whose attribute values are similar to the cluster averages becomes a member of that cluster.

DRIVES
Development and Research on Innovative Vocational Education Skills

MODELING

1. Find and drag k-means operator into the process window. Regarding the k value, since there are likely to be at least four potentially different groups, let's change the value of k to 4.

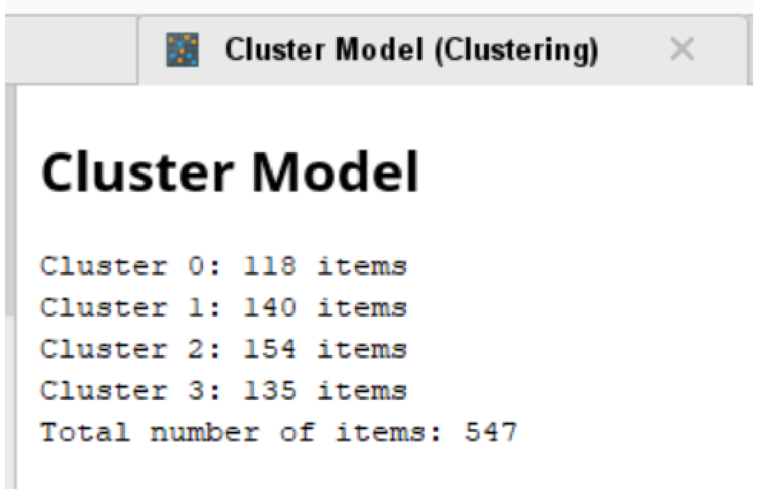


K-MEANS CLUSTERING WITH RAPIDMINER

DRIVES
Development and Research on Innovative Vocational Education Skills

MODELING

2. Run the model. Next, an initial report of the number of items that remained in each of the four clusters is presented. In this particular model, the clusters are fairly well balanced.

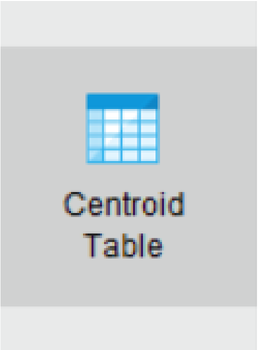


At this point we could go back and adjust the number of clusters, the value of 'max-runs' or even try out other parameters presented by the k-Means operator.

EVALUATION

Remember that Olivia's main goal is to try to find natural breaks between different types of heart disease risk groups. Using the k-Means operator in RapidMiner, we have identified four groups, and can now evaluate their usefulness.

1. Select the "Centroid Table" option. This window contains the averages for each attribute in each of the four clusters created.



Attribute	cluster_0	cluster_1	cluster_2	cluster_3
Weight	152.093	106.850	184.318	127.726
Cholesterol	185.907	119.536	218.916	154.385
Gender	0.441	0.543	0.591	0.459

EVALUATION

Attribute	cluster_0	cluster_1	cluster_2	cluster_3
Weight	152.093	106.850	184.318	127.726
Cholesterol	185.907	119.536	218.916	154.385
Gender	0.441	0.543	0.591	0.459

- *cluster* 2 has the highest average "Weight" and "Cholesterol";
- with 0 representing Female and 1 representing Male, a mean of 0.591 indicates that we have more males than females in this cluster.

EVALUATION

High cholesterol and high weight are two key indicators of the risk of heart disease that policyholders can do something about.

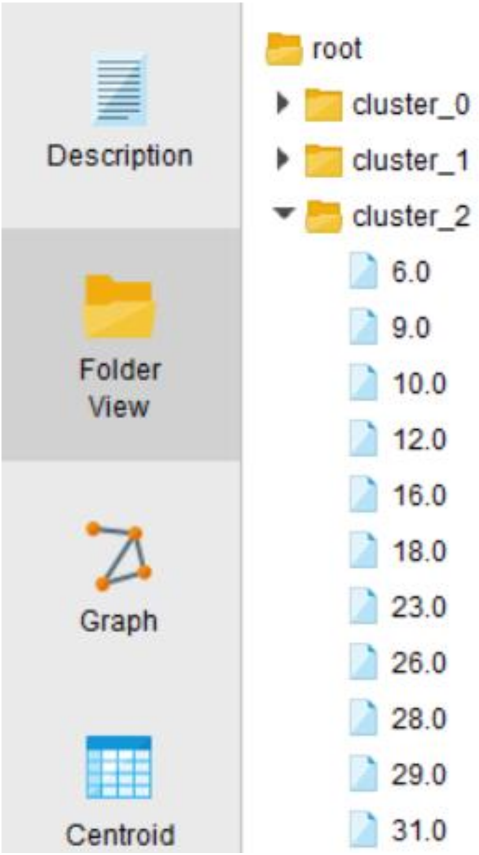
What does this mean?

Olivia should start with cluster 2 members when promoting her new programs and then extend to cluster 0 and 3 members, who are respectively the members with the highest averages for these two key risk factor attributes.

EVALUATION

Olivia knows that cluster 2 is where she's going to focus her first efforts, but how does she know who she's going to be in contact with? Who are the members of this high-risk cluster?

2. Select the "Folder View" option to access this type of information.

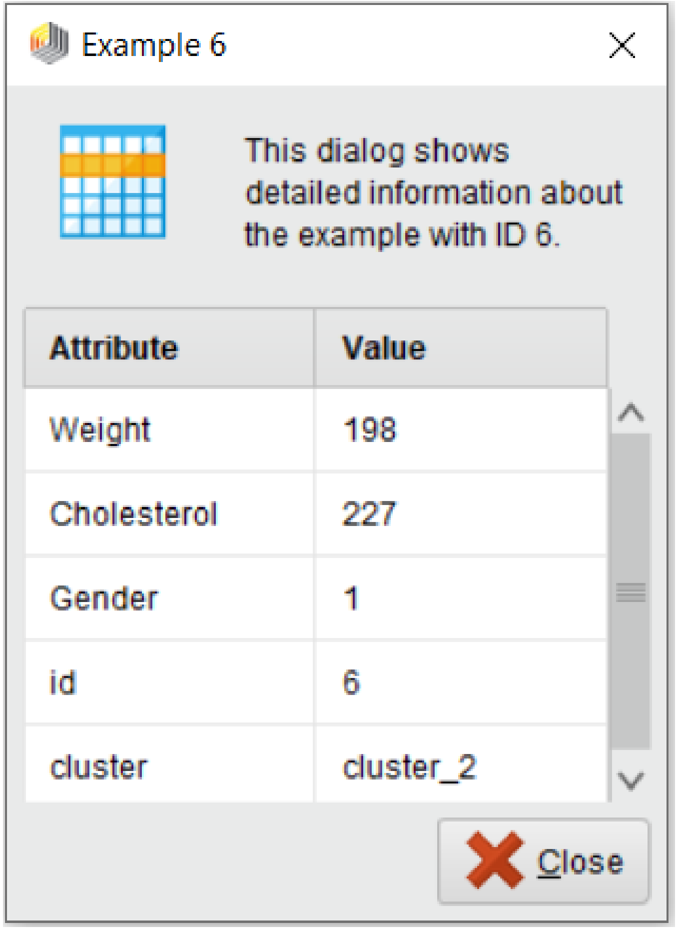


EVALUATION

3. Click on top of an observation to see its details.

The average for cluster 2 was just over 184 for weight and just under 219 for cholesterol. The person shown in Observation 6 is heavier and has higher cholesterol than the average for this highest risk group.

This is a person that Olivia can help!


Example 6

This dialog shows detailed information about the example with ID 6.

Attribute	Value
Weight	198
Cholesterol	227
Gender	1
id	6
cluster	cluster_2

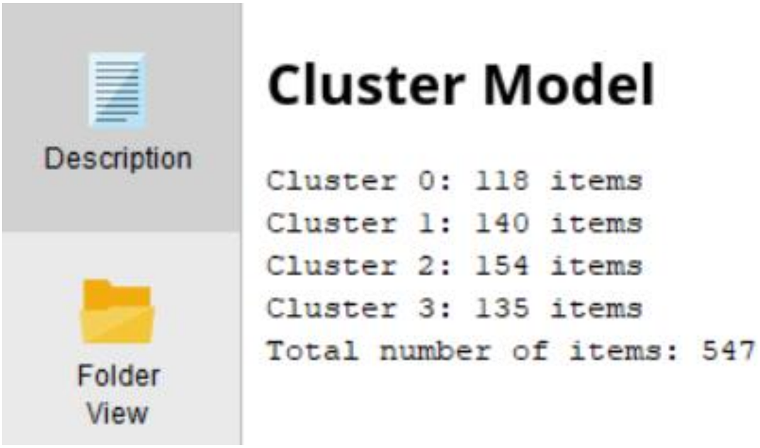
Close

EVALUATION

From the Cluster Model description, we know that there are 154 members in the dataset that fall into this group.

Clicking on each of them is a time consuming and inefficient process.

We can help Olivia extract the observations from cluster 2 fairly quickly and easily.



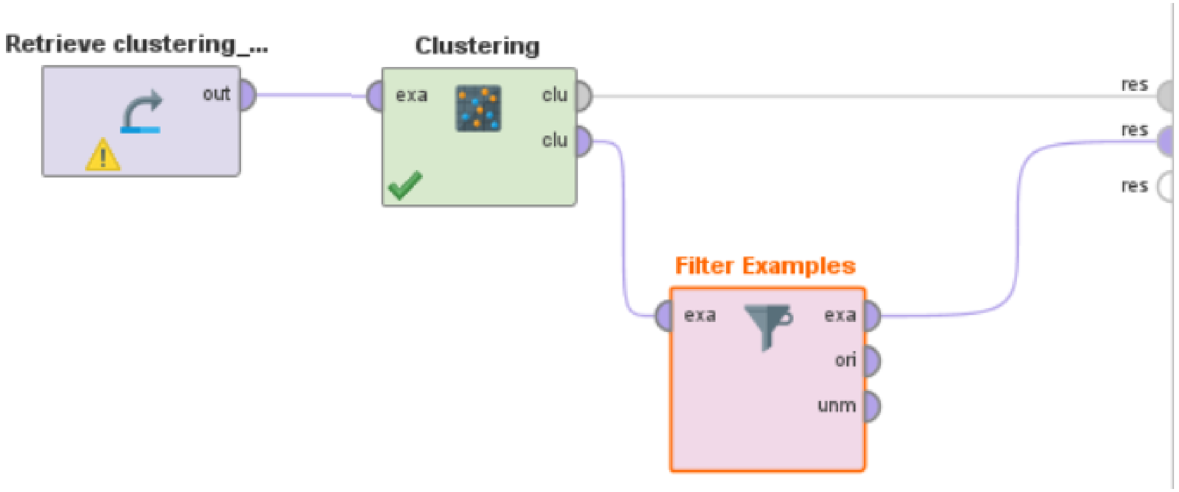
DRIVES
Development and Research on Innovative Vocational Education Skills

DEPLOYMENT

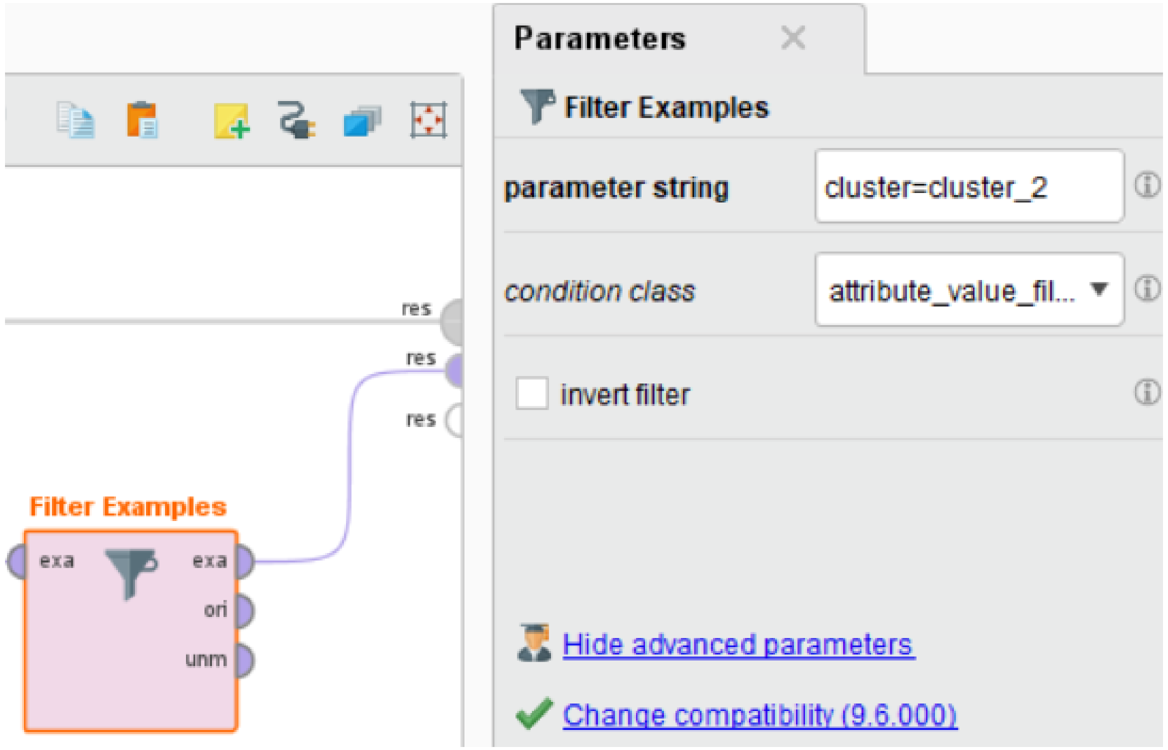
We can help Olivia extract the observations from cluster 2 fairly quickly and easily.

1. Go back to the Design perspective in RapidMiner.

2. Find and drag the "Filter Examples" operator and connect it to the k-Means Clustering operator. Connect the second *'clu'* (cluster) port to the *'exa'* port of the "Filter Examples" operator and connect the *'exa'* port of the "Filter Examples" to the final *'res'* port.



DEPLOYMENT

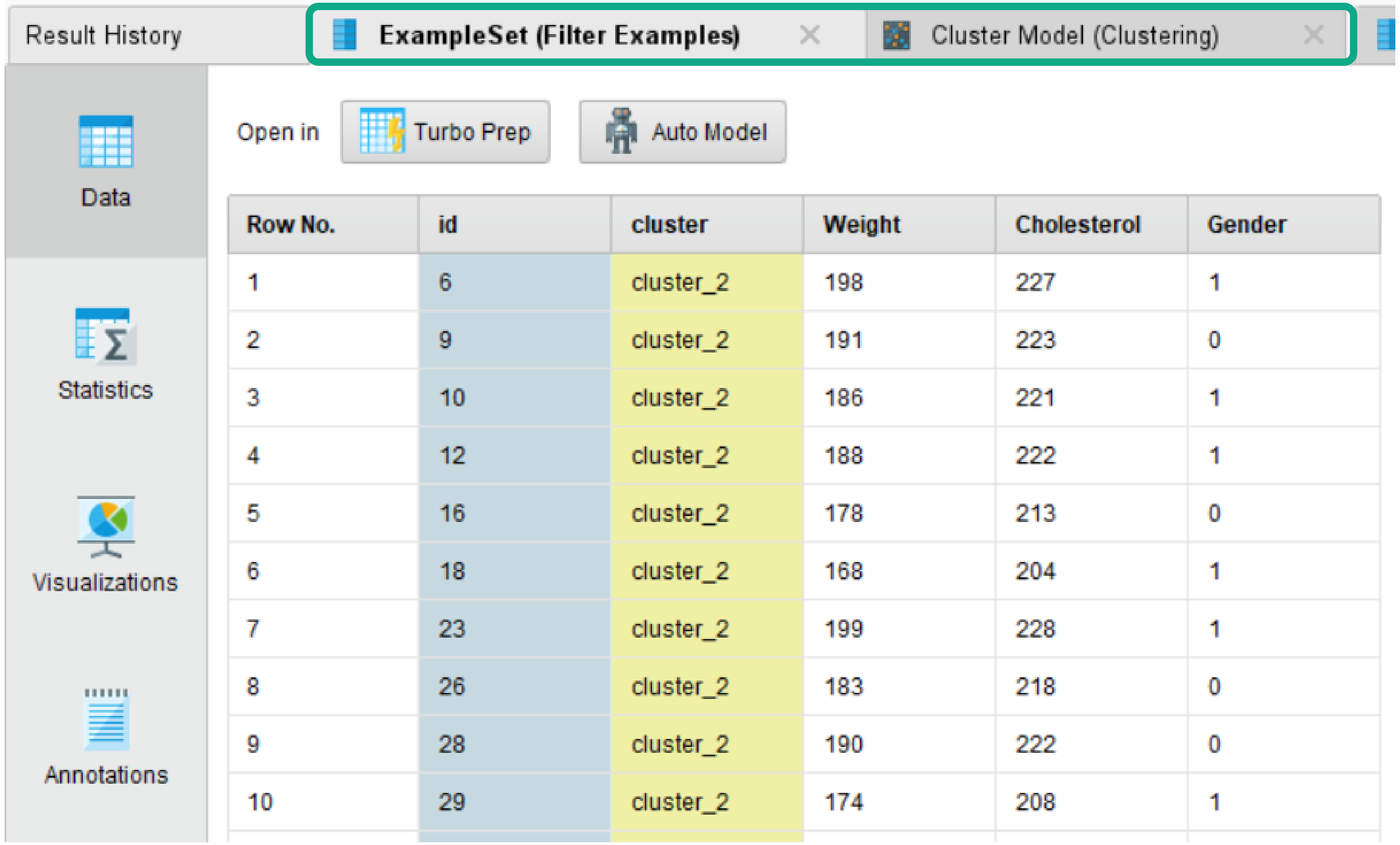


3. In the "condition class" field, select the 'attribute_value_filter' option, and for the "parameter string" field, enter the following:

cluster=cluster_2

This parameter refers to the "cluster" attribute and tells RapidMiner to filter out all observations where the value of that attribute is cluster_2. This means that only observations in the dataset that are classified as cluster_2 will be kept.

DEPLOYMENT



Row No.	id	cluster	Weight	Cholesterol	Gender
1	6	cluster_2	198	227	1
2	9	cluster_2	191	223	0
3	10	cluster_2	186	221	1
4	12	cluster_2	188	222	1
5	16	cluster_2	178	213	0
6	18	cluster_2	168	204	1
7	23	cluster_2	199	228	1
8	26	cluster_2	183	218	0
9	28	cluster_2	190	222	0
10	29	cluster_2	174	208	1

4. Execute the model.

In addition to the "Cluster Model" tab, there is the "ExampleSet" tab, which contains only the 154 observations that belong to cluster 2.

DRIVES
Development and Research on Innovative Vocational Education Skills

DEPLOYMENT

The high-risk group has weights between 167 and 203, and cholesterol levels between 204 and 235.

Data

Statistics

Visualizations

Annotations

Filter (5 / 5 attributes): Search for Attributes

Name	Type	Missing	Statistics		
id	Integer	0	Min 6	Max 543	Average 271.727
cluster	Nominal	0	Least cluster_3 (0)	Most cluster_2 (154)	Values cluster_2 (154), cluster_0 (0), ...[2 more]
Weight	Integer	0	Min 167	Max 203	Average 184.318
Cholesterol	Integer	0	Min 204	Max 235	Average 218.916
Gender	Integer	0	Min 0	Max 1	Average 0.591

DEPLOYMENT

Olivia can use these numbers to start contacting potential participants. To do so, she should access her company's database and perform an SQL query like this one:

```
SELECT First_Name, Last_Name, Policy_Num, Address, Phone_Num
FROM PolicyHolders_view
WHERE Weight >= 167
AND Cholesterol >= 204;
```

Through this query, Olivia is able to obtain the contact list of every person who falls into the highest risk group (cluster 2) in the hope of raising awareness, educating policyholders and modifying behaviors that will lead to a lower incidence of heart disease.

DRIVES
Development and Research on Innovative Vocational Education Skills

SUMMARY

- k-Means clustering is a Data Mining model that mainly fits into Classification. In this example, it does not necessarily predict which policyholders will or will not develop heart disease. Instead, it deals with known indicators of the attributes in a dataset and clusters them based on how similar these attributes are to the group averages.

- k-Means is an effective way to group observations on the basis of what is typical or normal for a group. In addition, it helps to understand where one group starts and the other ends, or, in other words, where there are natural breaks between groups in a given dataset.

- Although quite simple in its configuration and definition, k-Means clustering is a powerful and flexible method for finding natural groups of observations in a data set.

k-MEANS CLUSTERING

[1] What does the 'k' in k-Means clustering mean?

[2] How do you identify clusters? What is the process that RapidMiner uses to define and place observations in a given cluster?

[3] What does the Centroid Table reveal to the user? How do you interpret the values in this table?

[4] Think of a problem that can be solved by grouping observations into clusters. Search the internet for a dataset that can be used and applied to a k-Means model. Suggestion: go to the UCI - Machine Learning Repository website and choose a dataset whose Default Task is Clustering.

(a) Import the data into RapidMiner. Be sure to ensure that it is in the CSV format. Perform the Data Preparation step. This may include data inconsistency components, missing values, or changing data types;

k-MEANS CLUSTERING

(b) Add a k-means clustering operator to the dataset in RapidMiner and change the parameters as needed (especially the k value, to fit the problem at hand);
(c) Study the Centroid Table, Folder View, and other evaluation tools;
(d) Report the steps taken and the evidence found. Discuss the iterations in the model and how it is possible to answer the initial problem.

[5] Try the same dataset with different k-means operators like Kernel or Fast. To what extent do they differ from the original model. Do these operators change the original clusters? If so, to what extent?

DRIVES
Development and Research on Innovative Vocational Education Skills

CONTEXT

Remember Sara, the regional sales manager from the example in the correlations lesson? Her business is expanding, with more and more new customers, and she wants to make sure that the company will be able to meet this level of demand.

Sara knows that there is some correlation between the attributes in her data set and now wonders if she can use the same data set to predict heating oil usage for new customers.

The new customers have not yet started using heating oil. Sara wants to know how much oil needs to be kept in stock to meet the demand of these new customers.

Data mining can help Sara examine the various attributes and quantities of oil consumption from previous cases to anticipate and respond to the needs of new customers.

BUSINESS UNDERSTANDING

Sara's new goal is quite clear: she wants to anticipate the demand for heating oil.

Sara has a dataset with 1,218 observations, the same used in the correlation lesson, which provides a profile of attributes for each household, along with the annual heating oil consumption of those households. She wants to use the data from this dataset as training data to build a model that can predict the consumption of new customers.

To meet Sara's goal, we will use a **linear regression** model, a statistical modeling approach that computes a relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables) and then uses that relationship to make the prediction.

DATA UNDERSTANDING

Therefore, the dataset used in correlations will be used to train the model. Recall that this dataset is composed of the following attributes:

- **Insulation**: density rating that ranges from 1 to 10 and indicates how thick each house's insulation is. A house with a density rating of one is poorly insulated, while a house with a density of ten has excellent insulation.

- **Temperature**: average outside ambient temperature for each house in the most recent year, measured in degrees Fahrenheit.

- **Heating_Oil**: total number of heating oil units purchased by the owner of each home in the most recent year.

DRIVES
Development and Research on Innovative Vocational Education Skills

DATA UNDERSTANDING

- **Num_Occupants:** total number of occupants living in each house.
- **Avg_Age:** average age of the occupants living in each house.
- **Home_Size:** rating, on a scale of 1 to 8, of the overall size of the home. The higher the number, the larger the house.

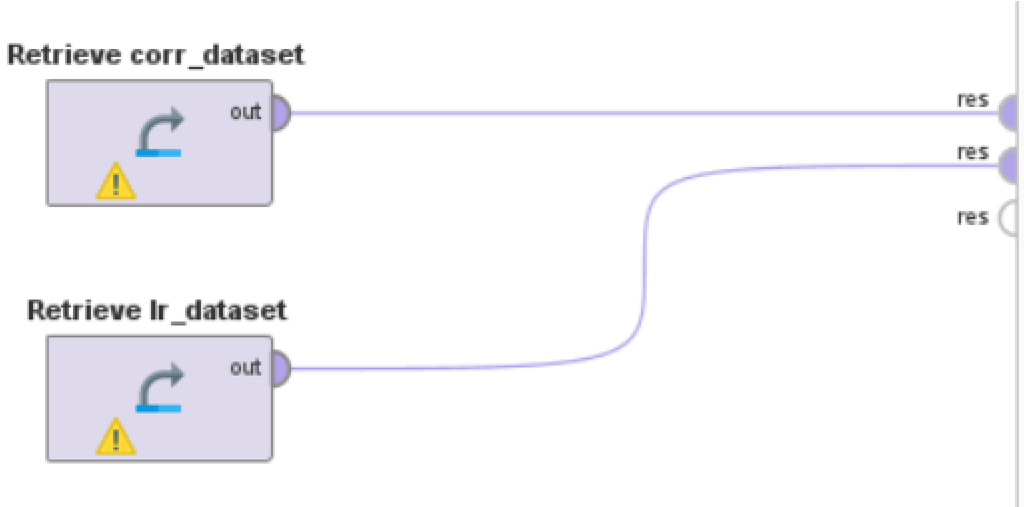
Sara has assembled in a CSV file the new customers' data containing all these attributes, except of course the *Heating_Oil* attribute. This dataset will be the dataset used to test the linear regression model.

DATA PREPARATION

Download the dataset: corr_dataset.csv + lr_dataset.csv

1. Import the *datasets* to the RapidMiner repository (Import Data -> My Computer).

2. Switch to the design perspective and drag the two datasets into the process window. Connect both *out* ports to the *res* ports, as shown in the figure below, and then run the model.



DRIVES
Development and Research on Innovative Vocational Education Skills

DATA PREPARATION

The ranges for all attributes in the test data must be within the ranges for the corresponding attributes in the training data.

Insulation	Integer	0	Min 2	Max 10	
Temperature	Integer	0	Min 38	Max 90	
Heating_Oil	Integer	0	Min 114	Max 301	
Num_Occupants	Integer	0	Min 1	Max 10	Average 3.113
Avg_Age	Real	0	Min 15.100	Max 72.200	Average 42.706
Home_Size	Integer	0	Min 1	Max 8	Average 4.649

training data – corr_dataset

WHY?

A training data set cannot be used to predict an attribute in the test data with observations whose values are outside the values in the training data set.

DRIVES
Development and Research on Innovative Vocational Education Skills

DATA PREPARATION

The ranges are the same for all attributes except the *Avg_Age* attribute. The test data have observations where *Avg_Age* is slightly below the lower bound of the training data set of 15.1, and some observations where *Avg_Age* is slightly above the upper bound of the training set of 72.2.

			Min	Max	Average
Insulation	Integer	0	2	10	5.989
Temperature	Integer	0	38	90	63.962
Num_Occupants	Integer	0	1	10	5.489
Avg_Age	Real	0	15	73	44.040
Home_Size	Integer	0	1	8	4.495

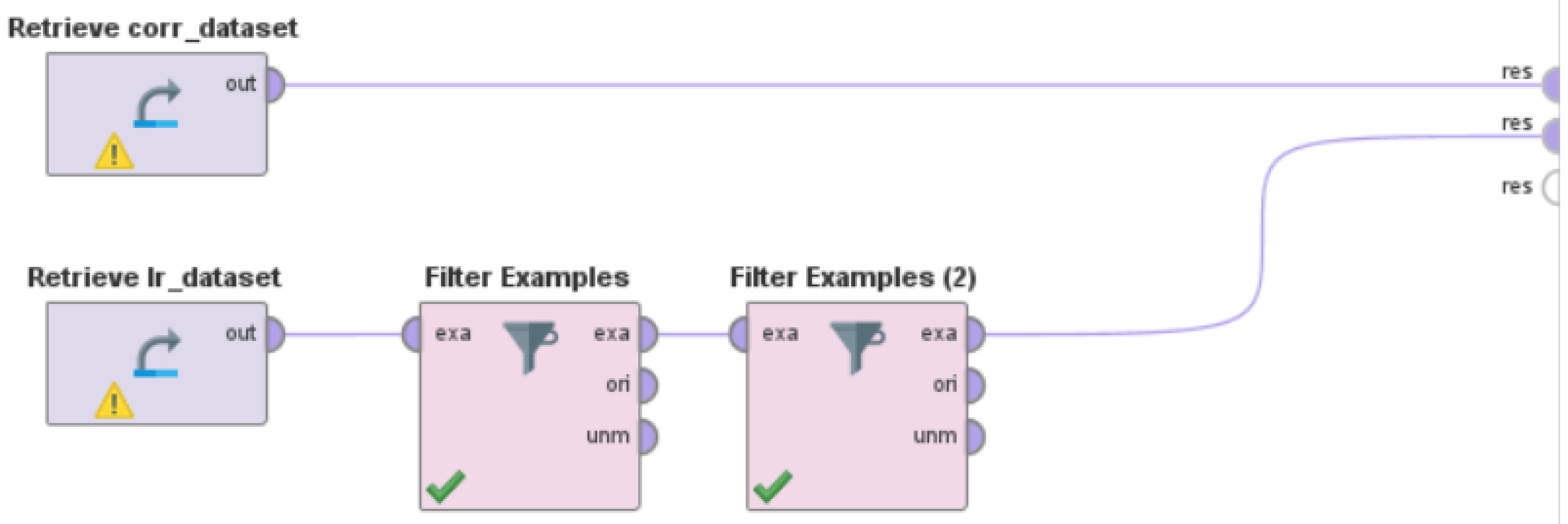
test data – lr_dataset

DATA PREPARATION

It is necessary to remove these observations from the test data set.

3. Go back to the design perspective. In the Operators tab in the lower left corner, use the search box to find the 'Filter Examples' operator. Drag two operators of that type into the process window. Set the 'condition class' parameter to 'attribute_value_filter' and the 'parameter string' parameter to:

Avg_Age >= 15.1
Avg_Age <= 72.2



DATA PREPARATION

4. Run the model. The test dataset now has 42,042 observations. Recheck the attribute ranges to make sure that none of the test attributes have ranges outside the values of the training attributes.

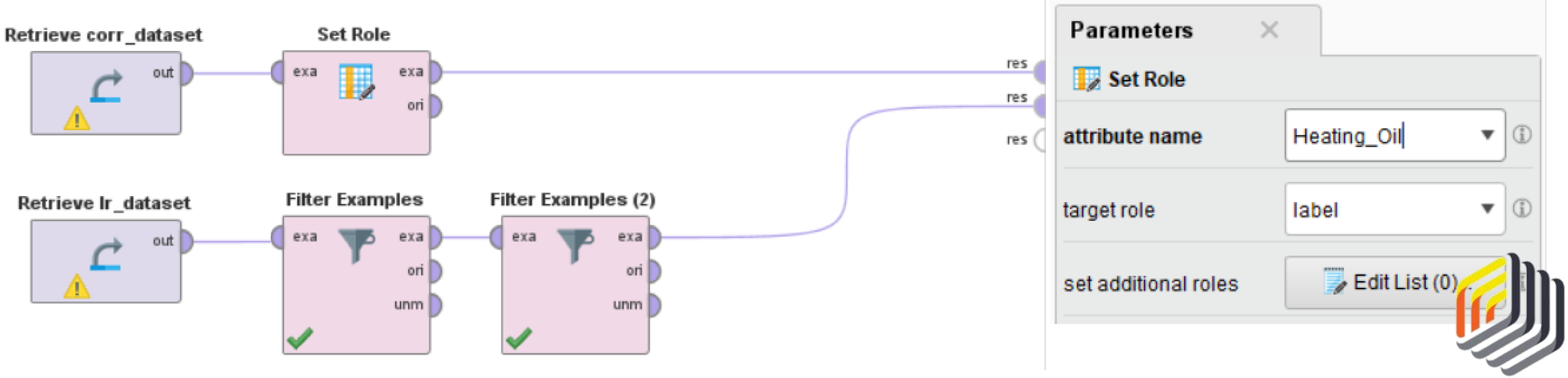
Insulation	Integer	0	Min 2	Max 10	Average 5.988
Temperature	Integer	0	Min 38	Max 90	Average 63.949
Num_Occupants	Integer	0	Min 1	Max 10	Average 5.489
Avg_Age	Real	0	Min 15.100	Max 72.200	Average 43.674
Home_Size	Integer	0	Min 1	Max 8	Average 4.497

DRIVES
Development and Research on Innovative Vocational Education Skills

DATA PREPARATION

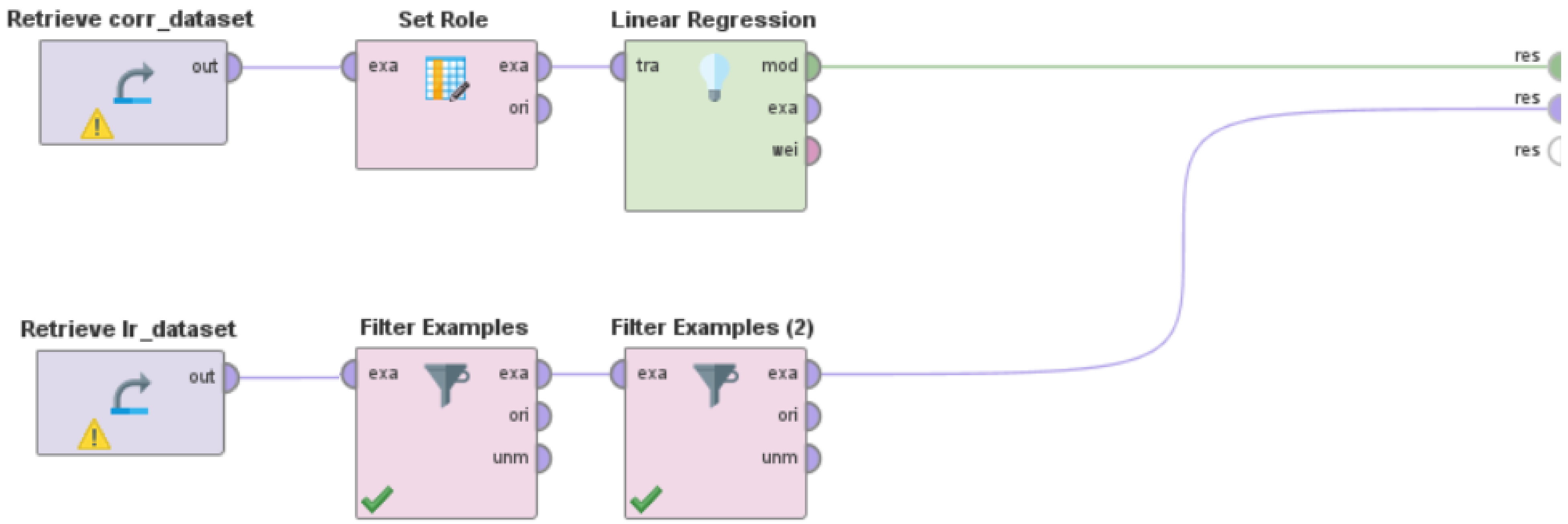
Linear regression is a predictive model and therefore needs an attribute to be designated as label - this is the target attribute, the one that it is intended to be predicted.

5. Go back to the Design perspective. Look for the "Set Role" operator and drag it into the process window. Associate this operator with the training flow. Change the parameters to indicate *Heating_Oil* as the target attribute for this model.



MODELING

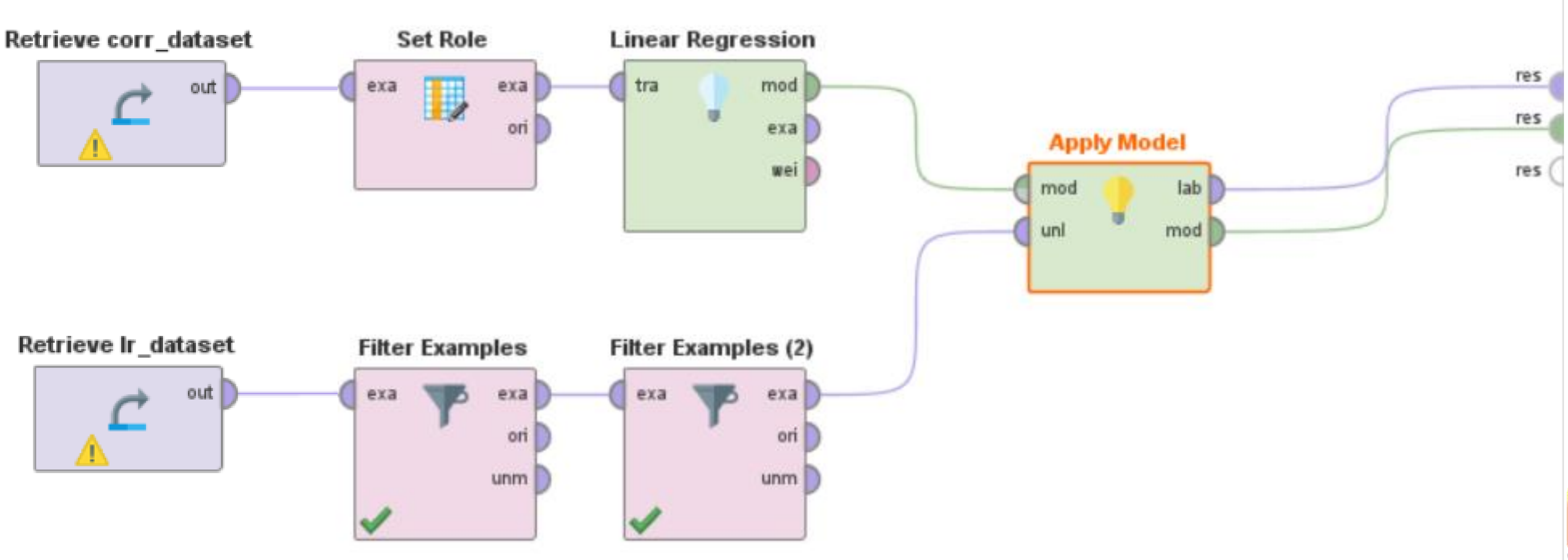
1. Find the 'Linear Regression' operator and drag it to the process window. Associate this operator with the training flow, as shown in the figure below.



DRIVES
Development and Research on Innovative Vocational Education Skills

MODELING

2. The final step in completing the model is to use an 'Apply Model' operator to connect the training flow to the test flow. Find this operator and drag it into the process window. Make sure to connect the *lab* and *mod* ports to the *res* ports as illustrated in the figure.



DRIVES
Development and Research on Innovative Vocational Education Skills

EVALUATION

1. Run the model. The fact that there are two outputs of the 'Apply Model' operator connected to the *res* ports will result in two tabs in the results perspective. Let's examine the *LinearRegression* tab first.

Result History | LinearRegression (Linear Regression)

Data

Description

Annotations

Attribute	Coeffici...	Std. Error	Std. Coe...	Tolerance	t-Stat	p-Value	Code
Insulation	3.323	0.420	0.164	0.431	7.906	0.000	****
Tempera...	-0.869	0.071	-0.262	0.405	-12.222	0	****
Avg_Age	1.968	0.065	0.527	0.491	30.217	0	****
Home_Si...	3.173	0.311	0.131	0.914	10.210	0	****
(Intercept)	134.511	7.589	?	?	17.725	0	****

DRIVES
Development and Research on Innovative Vocational Education Skills

EVALUATION

Linear regression modeling aims to determine how close a given observation is to an imaginary line that represents the mean or center of all points in the data set.

Independent Variable
Insulation, p.e.

What we want to predict (label)
Heating_Oil

$$y = mx + b$$

Constant
Intercept value in the table
134.511

Coeficiente do atributo
Second column of the table for
an Insulation = 3.323
(biggest coefficient of all)

If we had a house with an insulation density of 5, our formula using these insulation values would be y=(5×3.323)+134.511

EVALUATION

How can we set up this linear formula when we have several independent variables?

The result of the LinearRegression operator has only four attributes. What happened to the *Num_Occupants* attribute?

EVALUATION

The result of the LinearRegression operator has only four attributes. What happened to the *Num_Occupants* attribute?

Num_Occupants was not a statistically significant variable for predicting heating oil use in this dataset and was therefore removed by the RapidMiner.

When RapidMiner evaluated the influence that each attribute in the dataset had on heating oil use for each residence represented in the training dataset, the number of occupants had such a small influence that its weight in the formula was set to zero.

EVALUATION

How can we set up this linear formula when we have several independent variables?

$$y = mx + mx + mx \ldots + b$$

For instance:
- Insulation: 6
- Temperature: 67
- Avg_Age: 35.4
- Home_Size: 5

$$y = (6 * 3.323) + (67 * -0.869) + (35.4 * 1.968) + (5 * 3.173) + 134.511 = 181.758$$

The forecast for the annual number of heating oil units ordered (y) for this house is 181,758, i.e., basically 182 units.

DRIVES
Development and Research on Innovative Vocational Education Skills

DEPLOYMENT

Still in the results view, switch to the ExampleSet tab. We can see that the model developed in RapidMiner made a quick and effective prediction of the number of heating oil units that each of Sara's company's new customers are likely to use in their first year.

$$(5 * 3.323) + (69 * -0.869) + (70.1 * 1.968) + (7 * 3.173) + 134.511 = 251.321$$

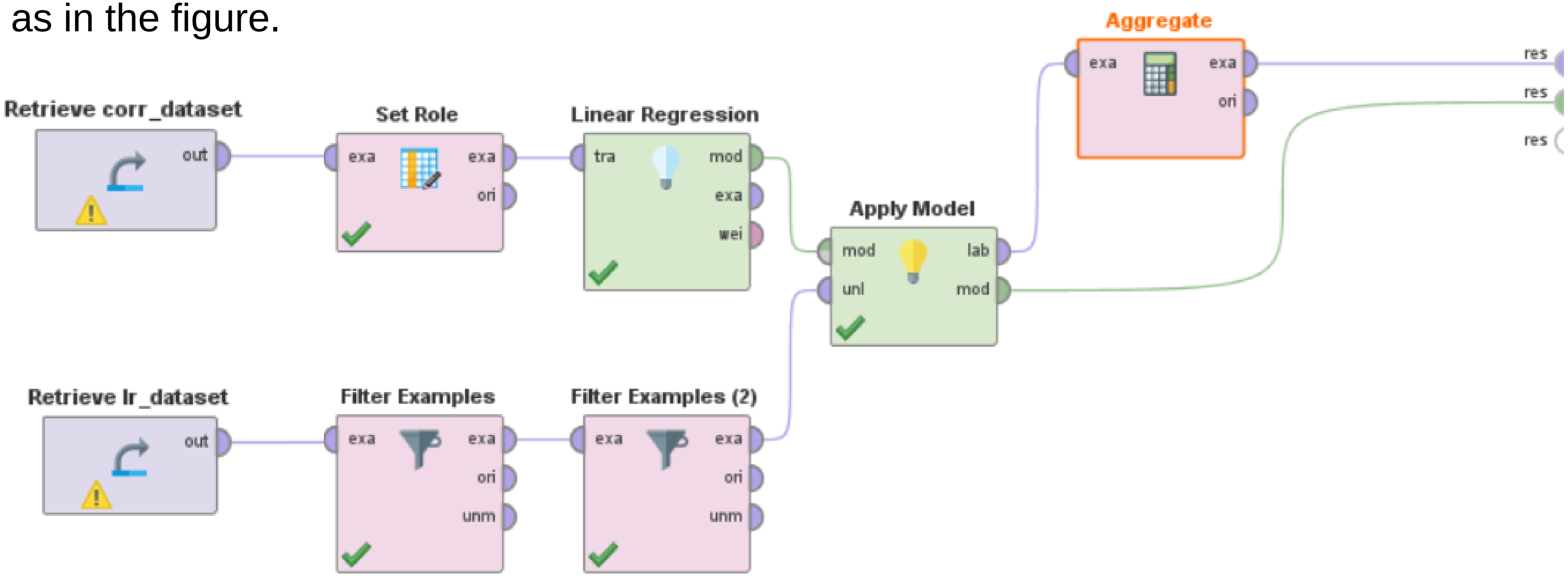
Row No.	prediction(H...	Insulation	Temperature	Num_Occup...	Avg_Age	Home_Size
1	251.321	5	69	10	70.100	7
2	216.028	5	80	1	66.700	1
3	226.087	4	89	9	67.800	7
4	209.529	7	81	9	52.400	6
5	164.669	4	58	8	22.900	7
6	180.512	4	58	6	37.400	3
7	221.188	6	51	2	51.600	3
8	164.001	2	73	5	37.400	4

DRIVES
Development and Research on Innovative Vocational Education Skills

DEPLOYMENT

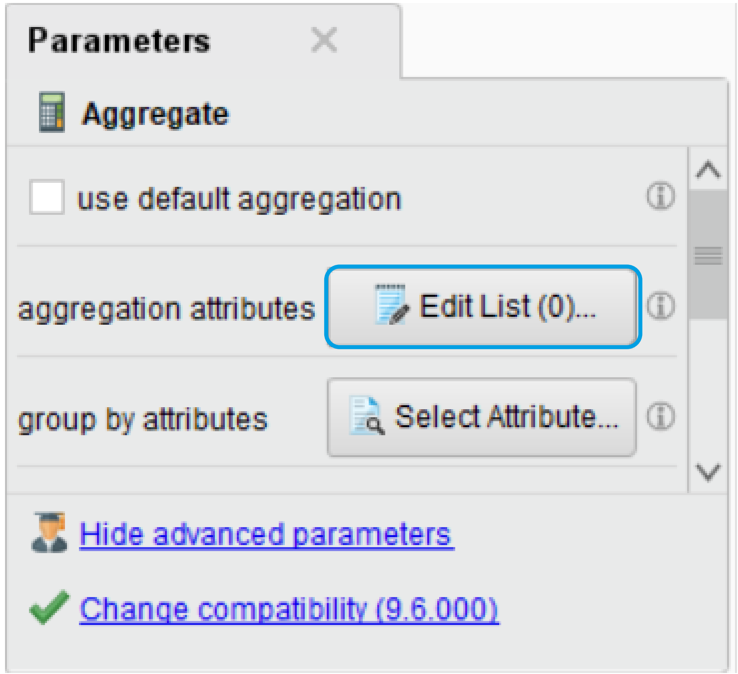
Sara now has a forecast of the oil consumption for the homes of each of the new customers, except for those with *Avg_Age* values outside the range. This will tell Sara the total number of new heating oil units that the company will need to supply in the coming year.

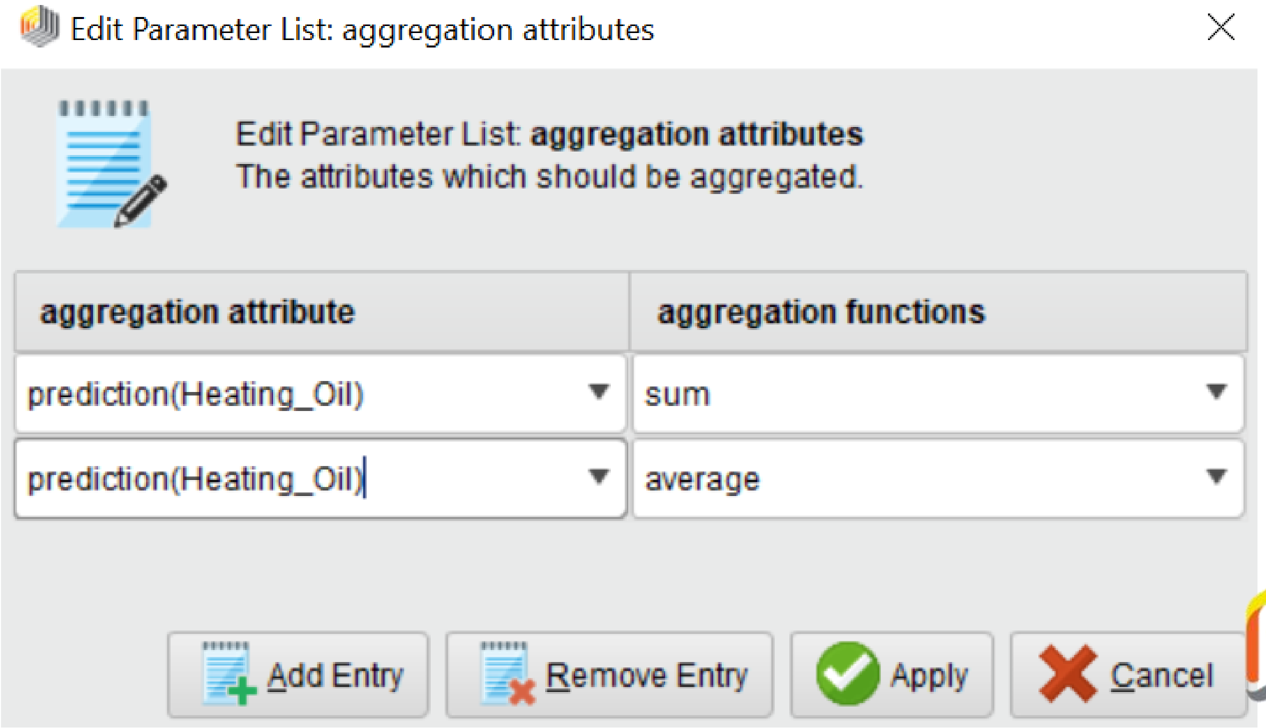
1. Go back to the Design tab, look for the 'Aggregate' operator and add it between the *lab* and *res* ports, as in the figure.



DEPLOYMENT

2. In the 'Parameters' tab, click on the *Edit List* button. Set the prediction attribute (Heating_Oil) as the aggregation attribute and the aggregation function as "sum". If you want, you can add other aggregations, such as the average.





DEPLOYMENT

3. Click "*OK*" to return to the main process window and then run the model. In the results section, select the *ExampleSet (Aggregate)* tab and select the *Data* option.

Row No.	sum(predict...	average(pre...
1	8368087.536	199.041

From these results, we can see that Sara's company will probably sell approximately 8,368,088 units of heating oil to the new customers. The company can expect that, on average, the new customers will order about 200 units each.

SUMMARY

- Linear regression is a predictive model that uses training and test datasets to generate numerical predictions. It is important to remember that linear regression uses numerical data for all its attributes.

- Each attribute in the dataset is statistically evaluated for its ability to predict the label type attribute. Attributes with poor predictive ability are removed from the model.

- As more data is collected, it can be added to the training dataset to make it more robust or expand the ranges for some attributes to include even more values. It is very important to remember that the ranges for the scoring attributes must be within the ranges of the training attributes to ensure valid predictions.

Linear Regression

[1] Linear regression requires all attributes to be of a certain data type. What is this data type? What is the data type of the predicted attribute when it is calculated?

[2] Why are attribute ranges so important when performing data mining through linear regression?

[3] What are linear regression coefficients? What does 'weight' mean in this context?

[4] What is the mathematical formula for linear regression and how is it organized?

[5] Download the "NBA_dataset" and select some attributes (at least four) to store data about each athlete. Some of the possible attributes you might consider may be annual salary, points_per_game, height, weight, age, etc. The goal of this exercise will be to predict the athletes' salary, so this should be a required attribute. [Note: Remember that linear regression only works with numerical data].

DRIVES
Development and Research on Innovative Vocational Education Skills

Linear Regression

[6] Divide the dataset observations into two parts: the training part and the test part. Make sure that you have at least 20 observations in the training dataset and at least 20 in the testing dataset. Since we will try to predict the salary of the athletes in the test dataset, you don't need to fill the salary column for these athletes. Save two CSV files (training and test), load them into RapidMiner and drag them into a new process window.

[7] Repeat the steps in RapidMiner as instructed and after running the model, in the Results section, examine the attribute coefficients and the athlete's salary predictions in the test set.

[8] Report the results by answering the following questions:

- [a] Which attributes have more weight?
- [b] Have attributes been removed from the dataset because they were not effective predictors? If so, which one(s) and why do you think it was not effective for the prediction?
- [c] Look up the salaries of some of the athletes in the test data and compare the actual salary to the predicted one. Is it close?

CONTEXT

Arthur works for a large online retail store. His company will soon be launching a next generation eReader and they want to maximize the effectiveness of their marketing.

Arthur noticed that some people were more eager to buy the previous generation device, while others seemed happy to wait to buy the electronic device later. So, he wonders what motivates some people to buy the product as soon as it comes out, while others are less motivated to buy the product.

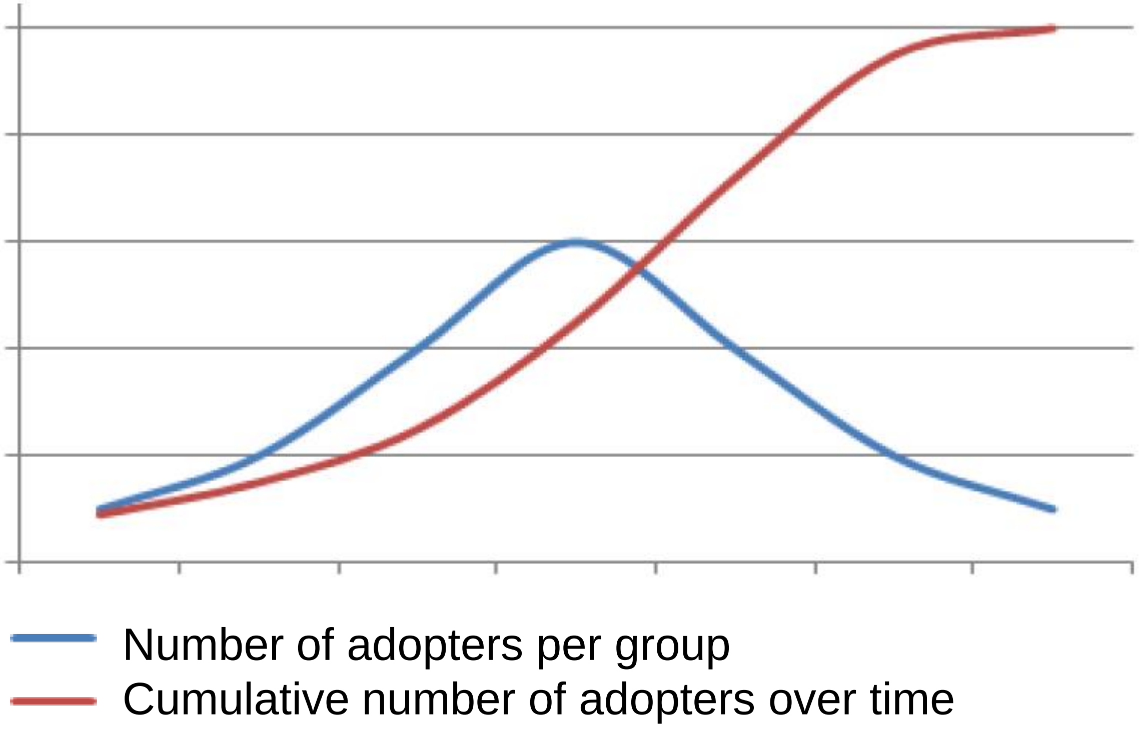
The company where Arthur works also sells other products, such as books (paper and digital), music, and electronic products of various kinds. Arthur believes that, by extracting customer data on general consumer behavior from the site, he will be able to find out which customers will buy the new eReader early, which one will buy next and which one will buy later.

Data mining can help Arthur predict when the customer will be ready to purchase the next generation eReader, enabling him to target his marketing to the people most willing to respond to ads and promotions.

BUSINESS UNDERSTANDING

Arthur also wants to understand how customer behavior on his company's website might indicate the timing of the purchase of the new eReader.

Diffusion of Innovation Theory (Rogers, 1960s)



The adoption of new technology or innovation tends to follow the 'S' curve, starting with a smaller group of more entrepreneurial and innovative technology-acquiring customers, followed by larger groups of medium-sized adopters (majority of adopters), followed by smaller groups of late adopters.

BUSINESS UNDERSTANDING

Following Rogers' theory, it was decided to categorize the company's customers who will eventually buy the new eReader into one of four groups:

Innovators → Early Adopters → Early Majority → Late Majority

Arthur hopes that by looking at the customer activity on the company's website, it will be possible to predict approximately when each person is most likely to purchase an eReader. Data mining can help Arthur figure out which activities are the best predictors of which category each customer falls into.

DATA UNDERSTANDING

Training Dataset

It contains the site activities of customers who bought the company's previous generation eReader and the time when they bought it.

Test Dataset

Composed of attributes of current customers that are expected to purchase the new eReader.

Arthur hopes to find out which category each customer will fall into in the test data set, based on the profiles and the purchase time of the customers in the training data set.

DATA UNDERSTANDING

The datasets have the following attributes:

- **User_ID**: a unique, numeric identifier assigned to each customer who has an account on the company's site.
- **Gender**: the gender of the customer. In the dataset, an 'M' for male and an 'F' for female is recorded. The Decision Tree operator can handle non-numeric data types.
- **Age**: the age of the client at the time the data was extracted from the site's database.
- **Marital_Status**: the marital status of the client. In the dataset: married -> M, single -> S
- **Site_Activity**: indication of how active each customer is on the company's website (rarely, regularly, or frequently).
- **Browsed_Electronics_12Mo**: Yes/No attribute, indicating whether or not the customer has searched for electronic products on the company's website in the past year.
- **Bought_Electronics_12Mo**: Yes/No attribute, indicating whether or not the customer has purchased an electronic item from the company's website in the past year.
- **Bought_Digital_Media_18Mo**: Yes/No attribute indicating whether or not the customer has purchased any form of digital media in the last year and a half. This attribute does not include purchases of digital books.

DATA UNDERSTANDING

- **Bought_Digital_Books:** Yes/No attribute, indicating whether or not the customer has bought a digital book since the beginning, not just last year.
- **Payment_Method:** Identifies the method by which the customer pays for his/her purchases. In cases where the customer has paid in more than one way, the most frequent mode or method of payment is used. There are four options available:
  - Bank transfer - payment via electronic check or other form of bank transfer directly from the bank to the company;
  - Website account - the customer has set up a credit card or permanent electronic funds transfer to his/her account, so that purchases are charged directly to the account at the time of purchase.
  - Credit card - the customer enter a credit card number and authorization every time he/she buys something on the website.
  - Monthly Billing – the customer makes periodically purchases and receives a paper or an electronic invoice that he/she pays later by sending a check or through the company's website payment system.

DRIVES
Development and Research on Innovative Vocational Education Skills

DATA UNDERSTANDING

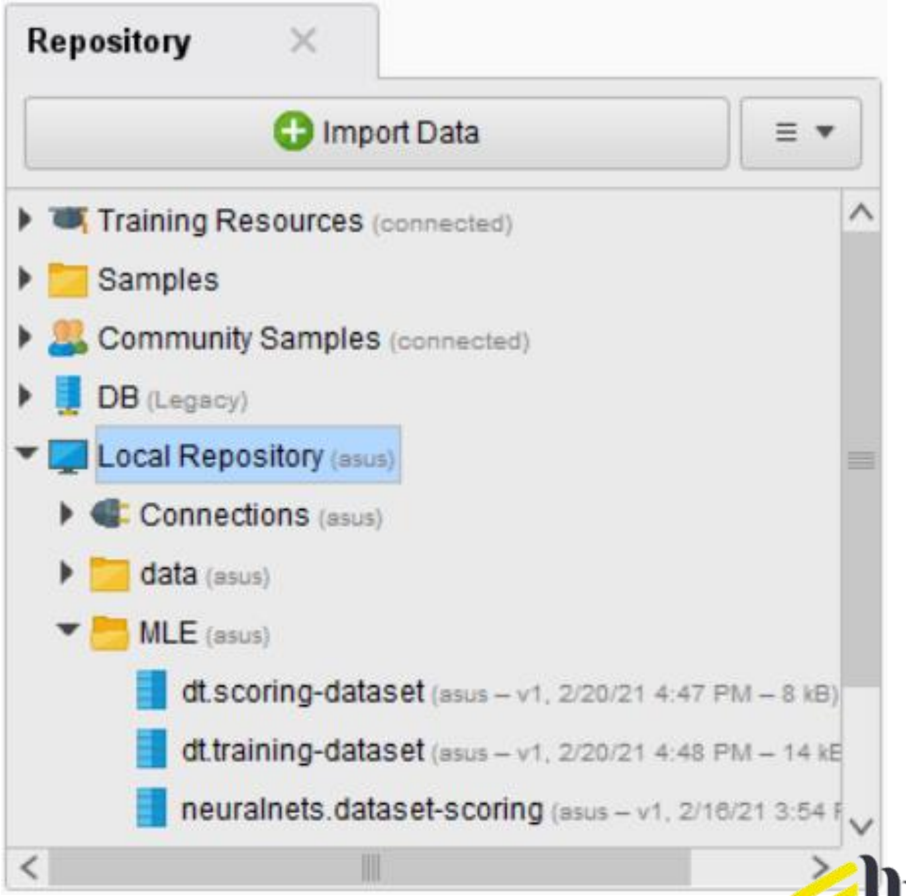
- **eReader_Adoption:** This attribute exists only in the training dataset and concerns information about customers who have purchased the previous generation of eReader:

  - Those who purchased within a week of product launch will be labeled as "*Innovator*".
  - Those who bought after the first week, but within the second or third week, are recorded as "*Early Adopter*“.
  - Those who bought after three weeks, but in the first two months, are "*Early Majority*“.
  - Those who bought it after the first two months are "*Late Majority*" This attribute will be used as label when we apply the training data to the test data.

DRIVES
Development and Research on Innovative Vocational Education Skills

DATA PREPRATION

Download do dataset: dt.training-dataset.csv
dt.scoring-dataset.csv

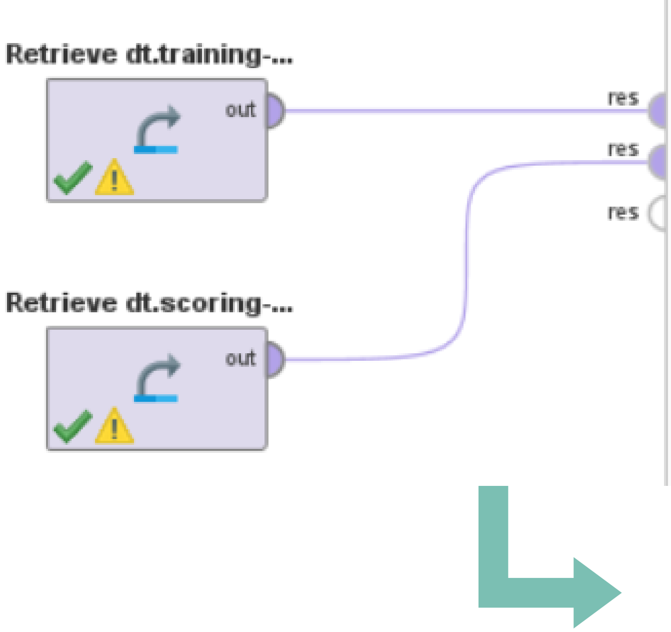


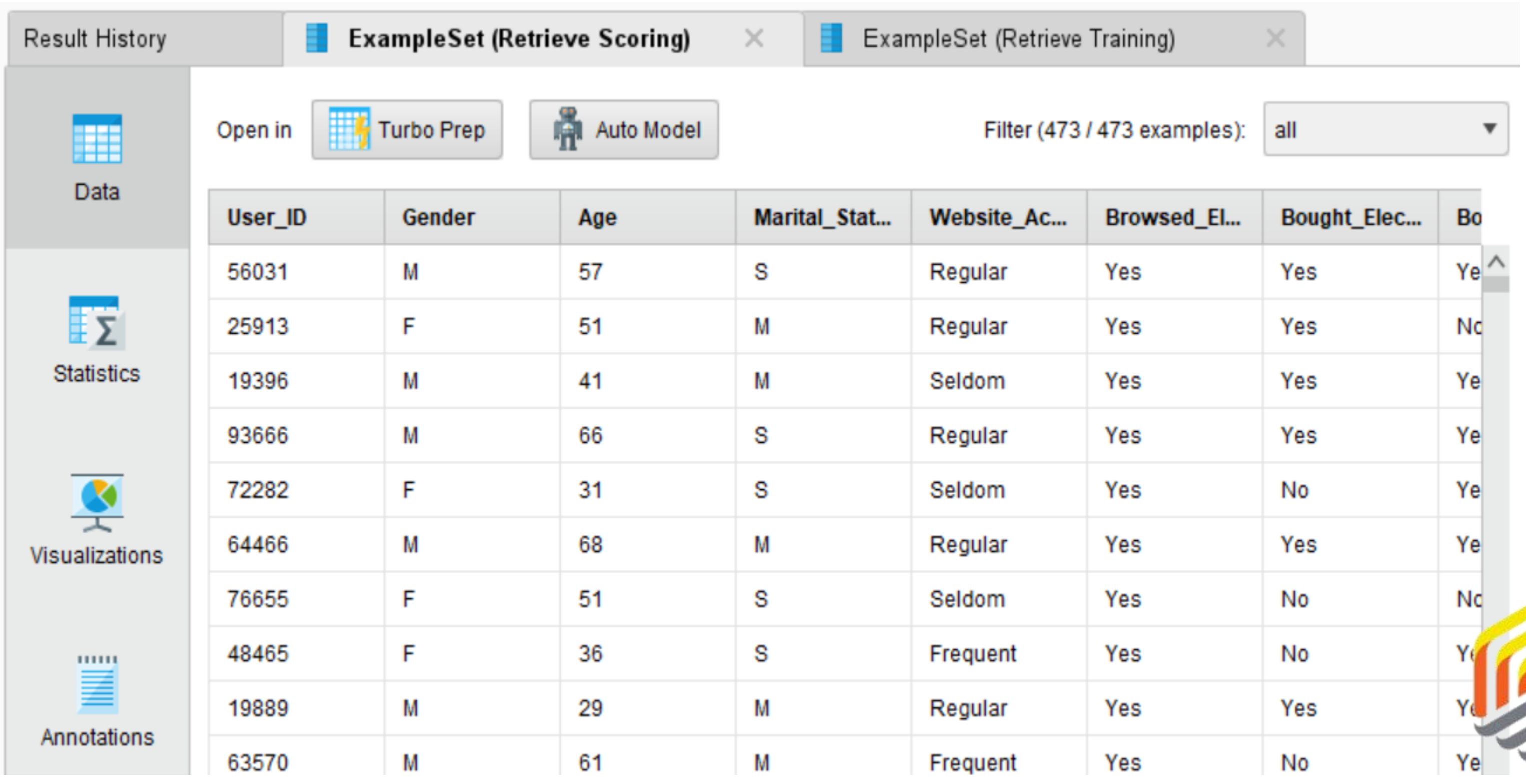
1. Import the dataset into the RapidMiner repository (Import Data -> My Computer).

2. Check the results view and inspect the CSV data you have imported. You don't need to worry about attribute data types because the decision tree operator can handle all types of data.

DATA PREPRATION

3. Connect both *out* ports to the *res* ports, as shown in the figure below, and then run the model. Examine the data and familiarize yourself with the attributes shown in the table.





User_ID	Gender	Age	Marital_Stat...	Website_Ac...	Browsed_El...	Bought_Elec...	Bo
56031	M	57	S	Regular	Yes	Yes	Ye
25913	F	51	M	Regular	Yes	Yes	No
19396	M	41	M	Seldom	Yes	Yes	Ye
93666	M	66	S	Regular	Yes	Yes	Ye
72282	F	31	S	Seldom	Yes	No	Ye
64466	M	68	M	Regular	Yes	Yes	Ye
76655	F	51	S	Seldom	Yes	No	No
48465	F	36	S	Frequent	Yes	No	Y
19889	M	29	M	Regular	Yes	Yes	Y
63570	M	61	M	Frequent	Yes	No	Ye

DRIVES
Development and Research on Innovative Vocational Education Skills

DATA PREPRATION

Apparently, there are no inconsistent data or missing values, however there is still some data preparation to be done.

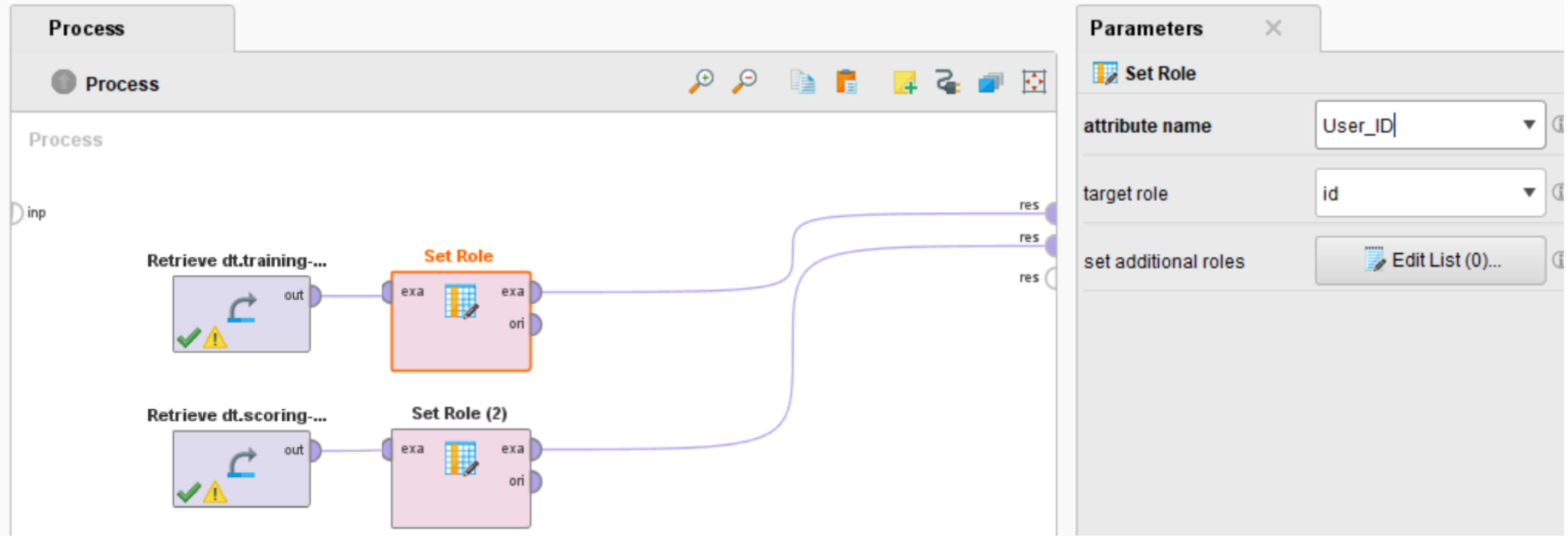
1st User_ID attribute
User_ID attribute
It serves only as an identifier of the client in the dataset and therefore should not be included in the model as an independent variable.

Select Attributes Remove the attribute	**OR**	**Set Role** New way to handle a non-predictive attribute

DRIVES
Development and Research on Innovative Vocational Education Skills

DATA PREPRATION

4. Find and add two *Set Role* operators to each of the flows (training and test). In the parameters tab (right side), set the role of the *User_ID* attribute to 'id' (for both Set Role operators). This causes the attribute to remain in the dataset, but not be considered as a predictor for the label attribute.



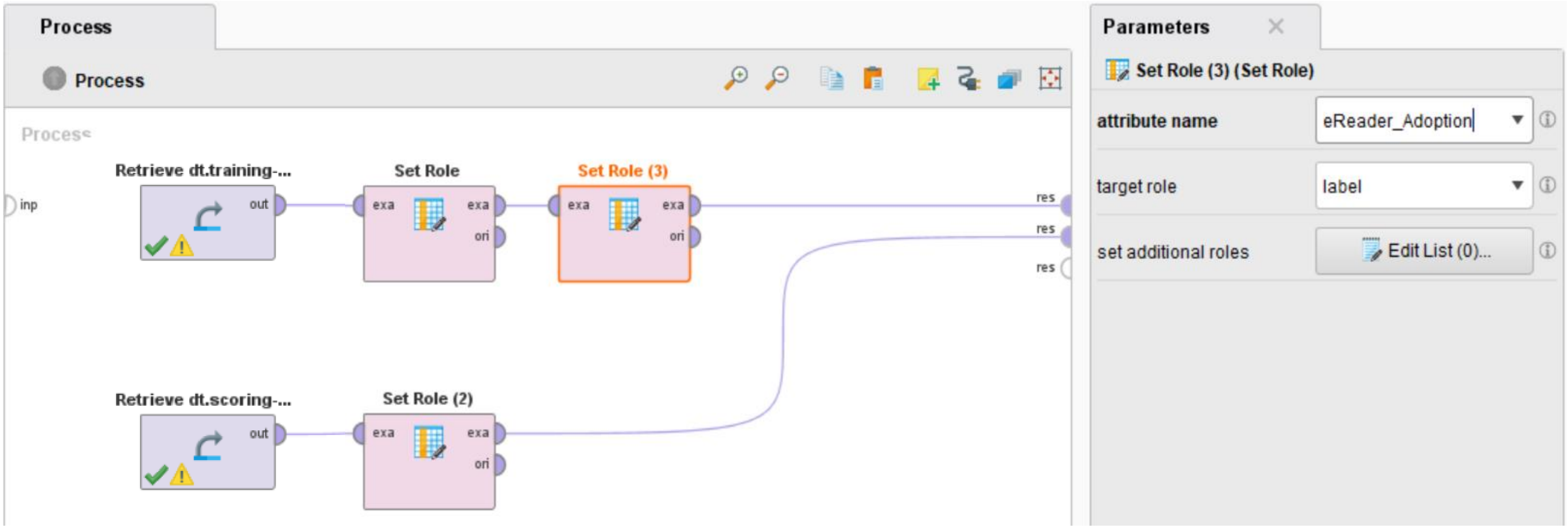
DRIVES
Development and Research on Innovative Vocational Education Skills

DATA PREPRATION

2^{nd} eReader_Adoption attribute
eReader_Adoption attribute

As with the other predictive models, the label attribute must be defined.

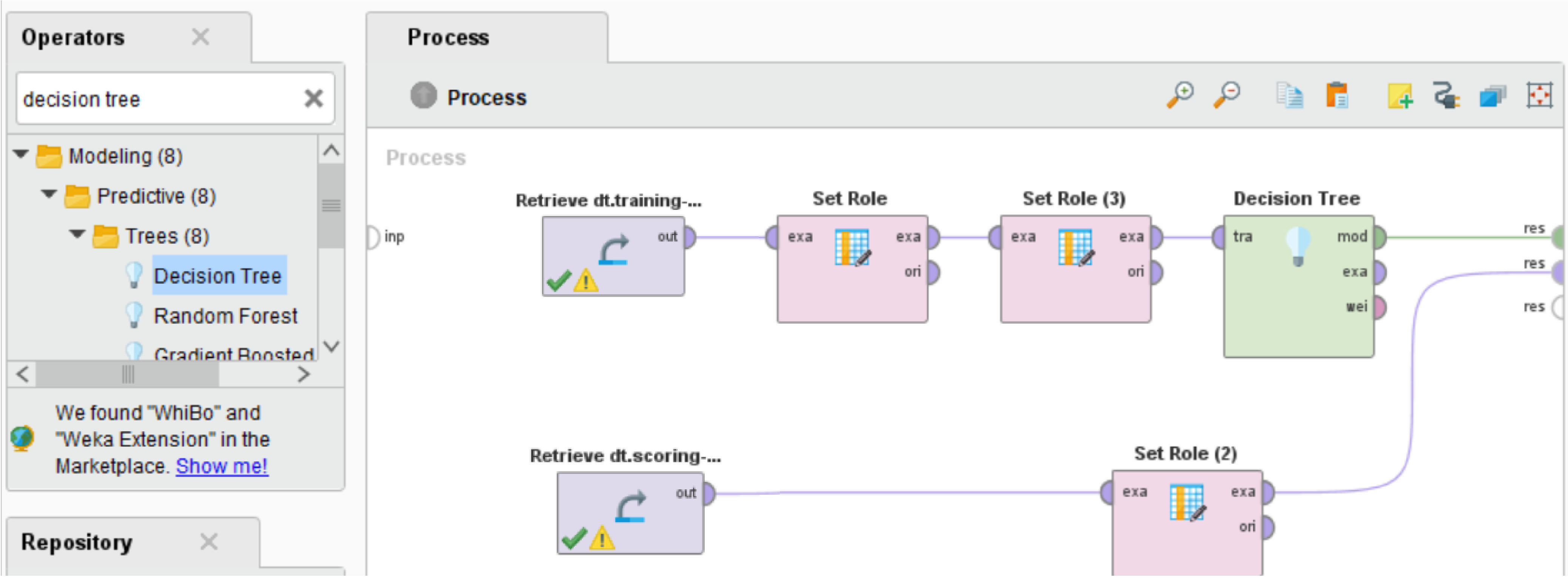
5. Add a Set Role operator to the training flow and set the "eReader_Adoption" attribute to 'label'.



DRIVES
Development and Research on Innovative Vocational Education Skills

DATA PREPRATION

6. Then, search for "Decision Tree“ in the Operators tab. Select the basic *Decision Tree* operator and add it to your training flow.



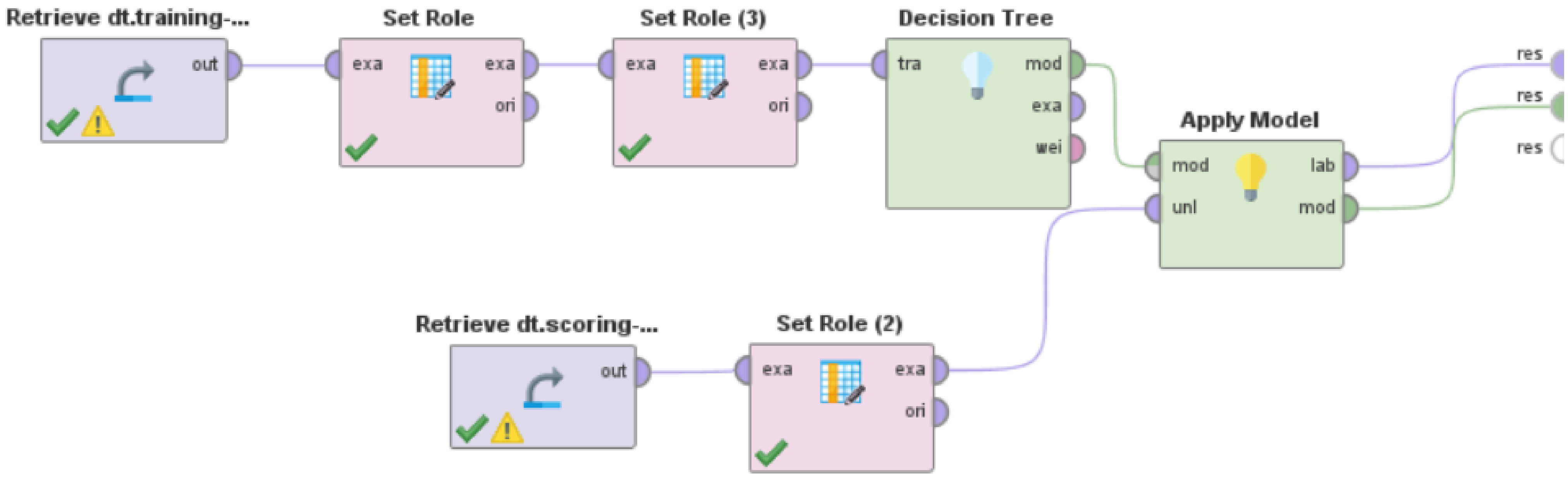
DATA PREPRATION

7. Run the model and switch to the *Tree (Decision Tree)* tab in the results perspective. You can see the preliminary tree, consisting of nodes (completely gray rectangles) and leaves (gray rectangles with a colored line at the bottom).

The nodes are attributes that serve as good predictors for the label attribute. The leaves are the endpoints that show us the distribution of the categories of our label attribute that follow the branch of the tree to the point of that leaf.

MODELING

1. Switch to the Design perspective. In the Operators tab look for the 'Apply Model' operator and drag it into the process window, joining the training and scoring flows. Make sure that both *lab* and *mod* ports are connected to the *res* ports in order to generate the desired results.



MODELING

2. Execute the model. Click on the 'ExampleSet' tab next to the 'Tree' tab. The tree was applied to the test data and, as a result, the confidence attributes were created by RapidMiner, along with a prediction attribute.

Id **User_ID**	Integer	0	Min 10153	Max 99694	Average 54647.074
Prediction **prediction(eReader_Adoption)**	Polynominal	0	Least Innovator (37)	Most Early Adopter (153)	Values Early Adopter (153), Late Majority (14
Confidence_Early Majority **confidence(Early Majority)**	Real	0	Min 0	Max 1	Average 0.287
Confidence_Late Majority **confidence(Late Majority)**	Real	0	Min 0	Max 1	Average 0.294
Confidence_Early Adopter **confidence(Early Adopter)**	Real	0	Min 0	Max 1	Average 0.288
Confidence_Innovator **confidence(Innovator)**	Real	0	Min 0	Max 1	Average 0.131
Gender	Polynominal	0	Least F (221)	Most M (252)	Values M (252), F (221)

MODELING

3. Switch to the 'Data View' option where the forecast for each customer's adoption group is displayed, along with the confidence percentages for each forecast. There are four confidence attributes, corresponding to the four possible values in the label attribute (*eReader_Adoption*).

Row No.	User_ID	prediction(e...	confidence(...	confidence(...	confidence(...	confidence(I...	Gender	Age	Marital_Stat...	Website_A
1	56031	Early Adopter	0.071	0	0.500	0.429	M	57	S	Regular
2	25913	Early Adopter	0.273	0.045	0.545	0.136	F	51	M	Regular
3	19396	Late Majority	0.061	0.879	0.030	0.030	M	41	M	Seldom
4	93666	Early Majority	1	0	0	0	M	66	S	Regular
5	72282	Late Majority	0.061	0.879	0.030	0.030	F	31	S	Seldom
6	64466	Early Majority	0.750	0.250	0	0	M	68	M	Regular
7	76655	Late Majority	0.065	0.879	0.056	0	F	51	S	Seldom
8	48465	Innovator	0	0.111	0	0.889	F	36	S	Frequent
9	19889	Late Majority	0	0.500	0.500	0	M	29	M	Regular
10	63570	Early Majority	1	0	0	0	M	61	M	Frequent
11	63239	Early Adopter	0.273	0.045	0.545	0.136	M	47	S	Regular
12	67603	Early Majority	0.950	0	0	0.050	F	62	S	Regular

MODELING

How to interpret these values?

Confidence percentages add up to a total of 100 per cent and measure how confident we are that the forecast will come true. The forecast corresponds to the category that produced the highest confidence percentage.

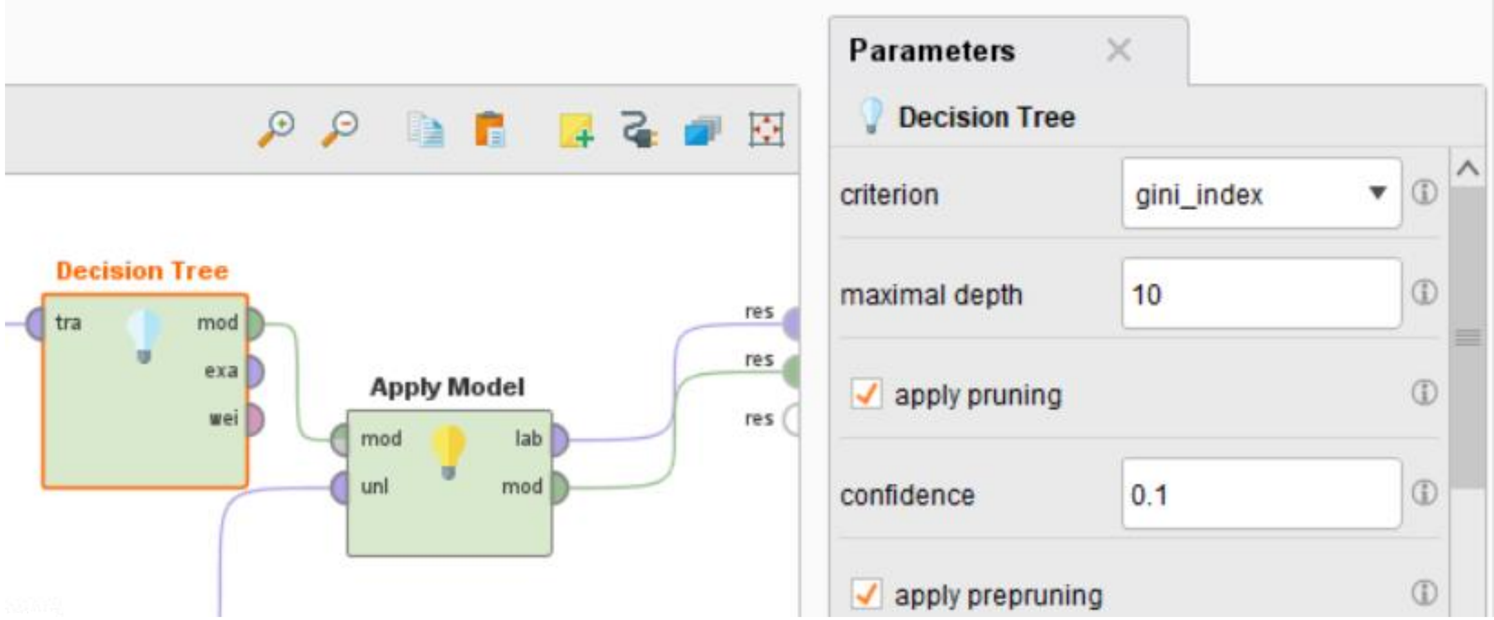
5	72282	Late Majority	0.061	0.879	0.030	0.030
6	64466	Early Majority	0.750	0.250	0	0
7	76655	Late Majority	0.065	0.879	0.056	0

RapidMiner is fairly (but not 100%) convinced that the person 64466 (line 6) is going to be a member of the 'early majority' (75%). Despite some uncertainty, RapidMiner is completely convinced that this person will not be an 'early adopter' (0%) nor an 'innovator' (0%).

MODELING

Remember that CRISP-DM is cyclical in nature, and that in some modeling techniques, especially those with less structured data, some trial-and-error may reveal more interesting patterns in the data.

4. Go back to the Design perspective, click on the 'Decision Tree' operator and change the 'criterion' parameter to 'gini_index'. Run the model.



EVALUATION

By analyzing the results, we can see that the tree has even more detail when using the *gini_index* criterion.

We could further modify the tree by going back to the Design splitter and changing the *minimal size for split* or the *minimal size for a leaf*.

Even with the default values for these parameters, we can see that the Gini algorithm alone is more sensitive than the Gain Ratio algorithm in identifying nodes and leaves.

DRIVES
Development and Research on Innovative Vocational Education Skills

EVALUATION

1. Switch to the 'ExampleSet' tab and choose the 'Data View' option. Switching the algorithm underlying the tree has, in some cases, changed our confidence in the prediction.

Row No.	User_ID	prediction(e...	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)	Gender	Age
1	56031	Early Adopter	0.200	0	0.600	0.200	M	57
2	25913	Early Adopter	0	0	0.875	0.125	F	51
3	19396	Late Majority	0.061	0.879	0.030	0.030	M	41
4	93666	Innovator	0.333	0	0	0.667	M	66
5	72282	Late Majority	0.061	0.879	0.030	0.030	F	31
6	64466	Early Majority	0.750	0.250	0	0	M	68
7	76655	Late Majority	0.333	0.667	0	0	F	51
8	48465	Innovator	0	0.250	0	0.750	F	36
9	19889	Early Majority	0.500	0	0.500	0	M	29
10	63570	Early Majority	1	0	0	0	M	61
11	63239	Early Majority	0.667	0	0.167	0.167	M	47
12	67603	Early Majority	0.917	0	0.042	0.042	F	62

EVALUATION

Let's analyze the customer in line 2 (ID 25913) as an example. According to the Gain Ratio criteria, this customer was calculated as having at least some percentage probability of landing in any of the four adopter categories. There was 54.5% certainty that he/she would be an *early adopter*, but almost 27.3% certainty that he/she could also become a member of the *early majority*.

Gain Ratio

Row No.	User_ID	prediction(...	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)	Gender	Age
1	56031	Early Adopter	0.071	0	0.500	0.429	M	57
2	25913	Early Adopter	0.273	0.045	0.545	0.136	F	51

Gini Index

Row No.	User_ID	prediction(e...	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)	Gender	Age
1	56031	Early Adopter	0.200	0	0.600	0.200	M	57
2	25913	Early Adopter	0	0	0.875	0.125	F	51

During the implementation phase, Arthur will have to decide which of the categories the customer belongs to. But perhaps, using the Gini Index criteria, it might be possible to help him decide.

EVALUATION

According to the Gini Index criteria, this customer has an 87.5% chance of being an *Early Adopter* and only a 12.5% chance of being an *Innovator*. Note that the odds of him/her becoming part of the *Early Majority* and the *Late Majority* have dropped to zero.

While customer 25913 may not be at the top of Arthur's list when the implementation is launched, it will likely be positioned higher than it would be if it were under the Gain Ratio criteria.

Gain Ratio

Row No.	User_ID	prediction(...	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)	Gender	Age
1	56031	Early Adopter	0.071	0	0.500	0.429	M	57
2	25913	Early Adopter	0.273	0.045	0.545	0.136	F	51

Gini Index

Row No.	User_ID	prediction(e...	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)	Gender	Age
1	56031	Early Adopter	0.200	0	0.600	0.200	M	57
2	25913	Early Adopter	0	0	0.875	0.125	F	51

DRIVES
Development and Research on Innovative Vocational Education Skills

EVALUATION

Note that while the Gini Index criteria changed some of the predictions, it did not affect all of them. Check again the person with the ID 64466. There is no difference in this person's predictions under either algorithm. Sometimes the confidence level in a prediction using a decision tree is so high that a more sensitive underlying algorithm does not change the values of that prediction at all.

Gain Ratio

Row No.	User_ID	prediction(eReader_Adoption)	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)
6	64466	Early Majority	0.750	0.250	0	0
7	76655	Late Majority	0.065	0.879	0.056	0

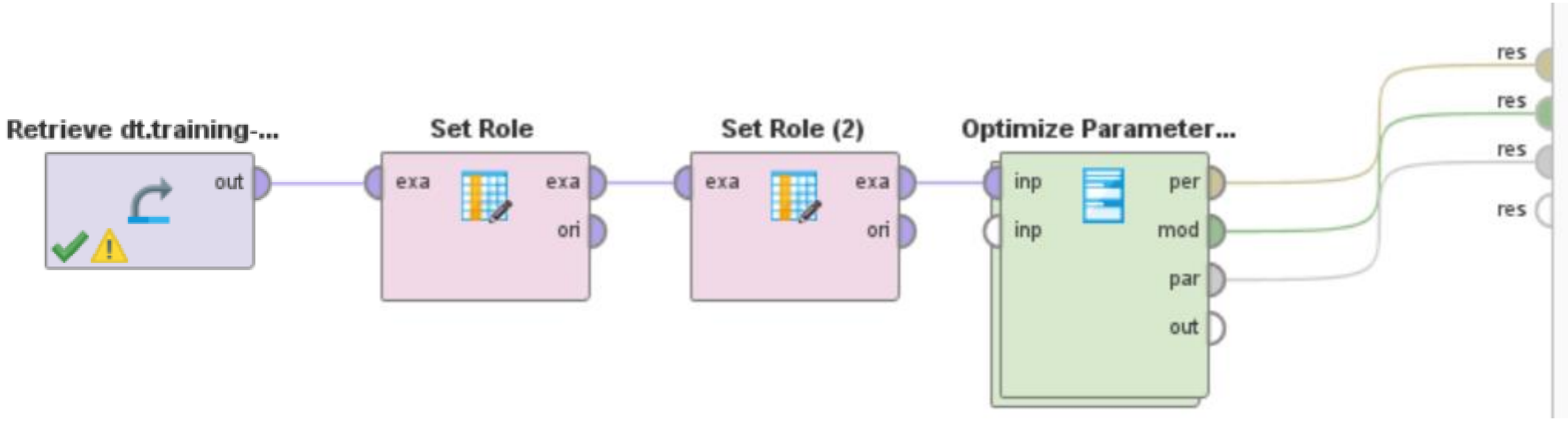
Gini Index

Row No.	User_ID	prediction(e...	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)
6	64466	Early Majority	0.750	0.250	0	0
7	76655	Late Majority	0.333	0.667	0	0

EVALUATION

After this initial approach, it is important to realize that the parameters that have been set in the *Decision Tree* operator are probably not the most appropriate to achieve the best possible result. Thus, it is important to try to find the best values that the parameters can have in order to maximize the performance of the model.

2. Create a new process and drag the training dataset into this process. As before, use the two *Set Role* operators, one for the ID and one for the label attribute. Find the *Optimize Parameters (Grid)* operator and drag it into the process. Connect the first 3 ports of this operator to the *res* ports.

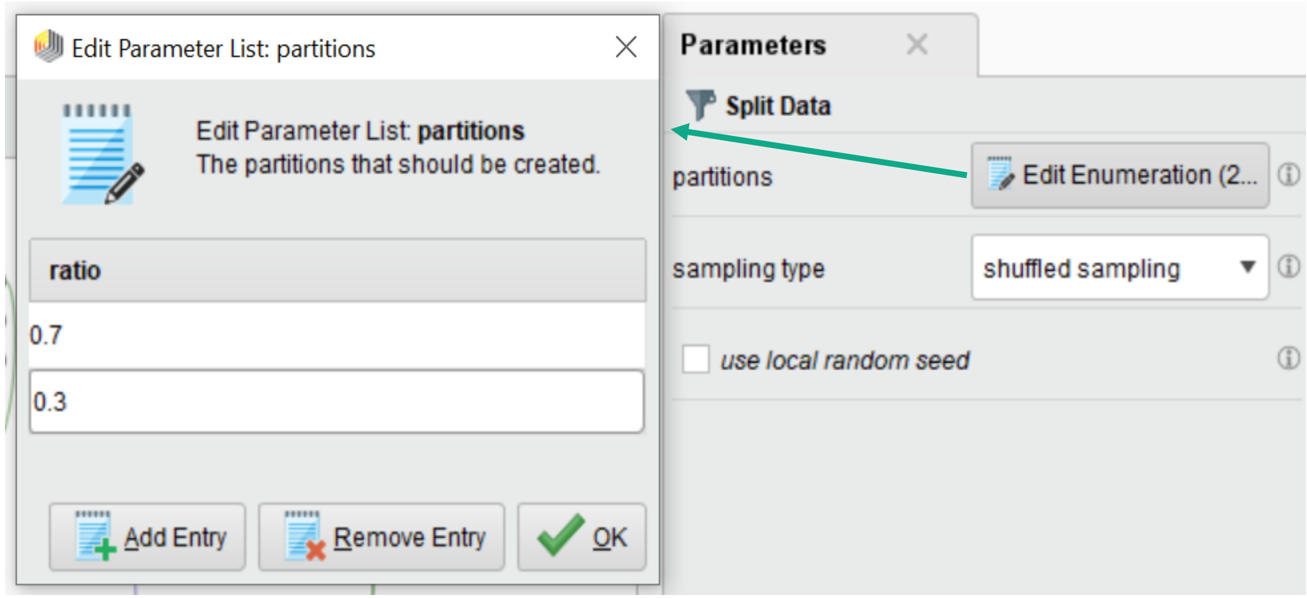


DRIVES
Development and Research on Innovative Vocational Education Skills

EVALUATION

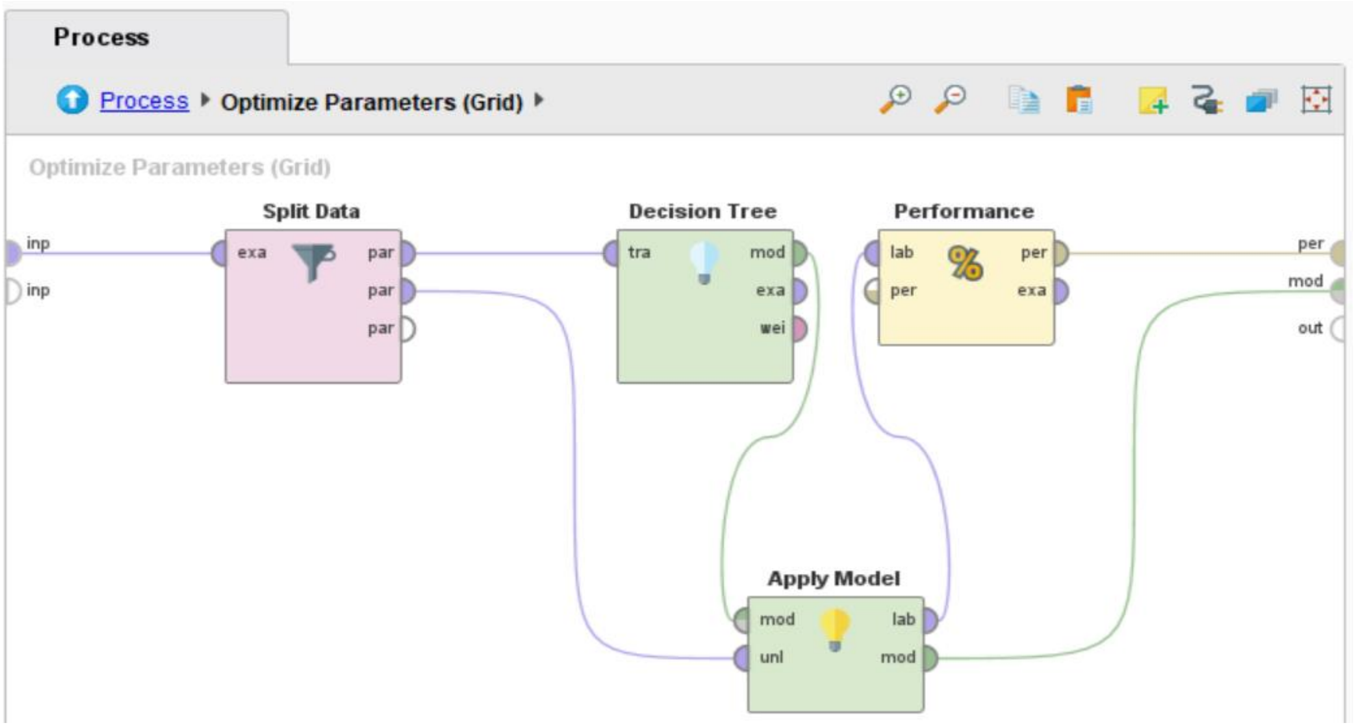
The *Optimize Parameters (Grid)* operator is a nested operator. It executes the subprocess that it incorporates for all the possible combinations of values of the selected parameters and then returns the optimal values for these parameters. What is still missing is to incorporate the subprocess that we want to repeat, within the optimization operator, that is, the classification with the Decision Tree algorithm.

3. Double click on the *Optimize Parameters (Grid)* operator. A new subprocess window will open. Start the subprocess with a *Split Data* operator, since for this case you will need to split the dataset, so that you can later evaluate the accuracy of the model. Set the parameters as shown in the picture.



EVALUATION

4. Next, add a *Decision Tree* and *Apply Model* operator as shown in the figure. This time, the *Performance* operator will be added to allow you to statistically evaluate the performance of the classification model. This evaluation will be carried out by default through accuracy.



EVALUATION

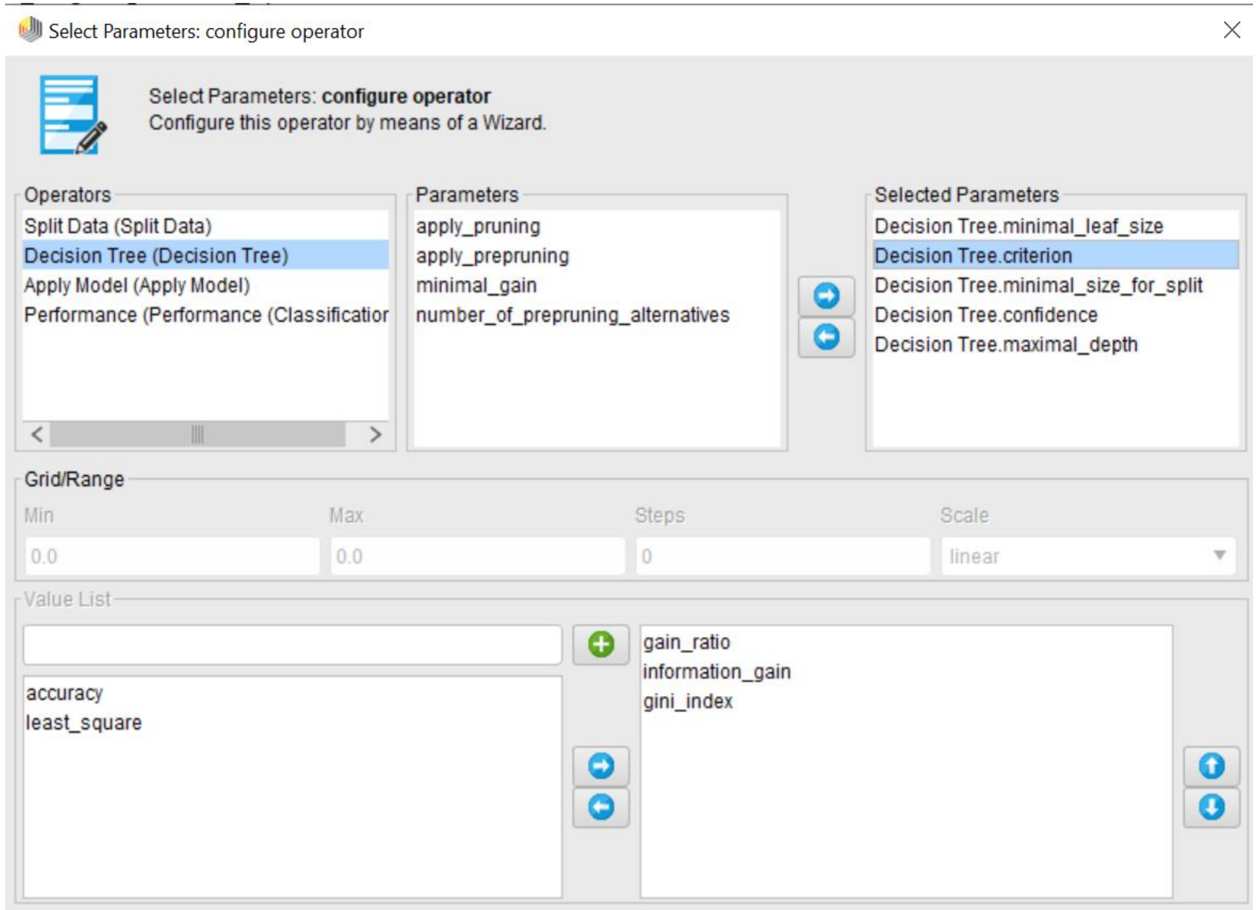
Now, in the optimization operator, it is necessary to indicate which are the parameters that we want to optimize, in this case, the parameters associated to the *Decision Tree* operator.

5. Go back by clicking in "Process" in the process bar.

6. Click in the *Optimize Parameters (Grid)* operator and, in the parameters panel placed on the righ side, click in *Edit Parameters Settings.*

EVALUATION

A new window appears, where it is possible to choose the parameters to optimize. First, let’s choose the *Decision Tree* in the *Operators* list and then select the desired *Parameters*, sending them to the right side list. It is important to bear in mind that in what regards the *criterion,* it is necessary to remove the *accuracy* and the *least_square* values from the list, because they do not apply to our model.



DRIVES
Development and Research on Innovative Vocational Education Skills

EVALUATION

7. Run the model. Note that, the more optimization parameters are selected, the slower the execution of the model will be. In the *Results* window, there are several tabs. The *ParameterSet separator* presents the best result (*accuracy*) obtained during all interations and which parameter values were used to obtain that result. The *Optimize Parameters* separator shows the iterations performed for each parameter.

ParameterSet

```
Parameter set:

Performance:
PerformanceVector [
-----accuracy: 74.40%
ConfusionMatrix:
True:   Early Majority  Late Majority   Early Adopter   Innovator
Early Majority: 59      5       21      5
Late Majority:  3       62      4       2
Early Adopter:  9       1       46      9
Innovator:      1       1       3       19
]
Decision Tree.criterion = gain_ratio
Decision Tree.minimal_size_for_split    = 21
Decision Tree.maximal_depth     = 39
Decision Tree.confidence        = 0.25000005
```

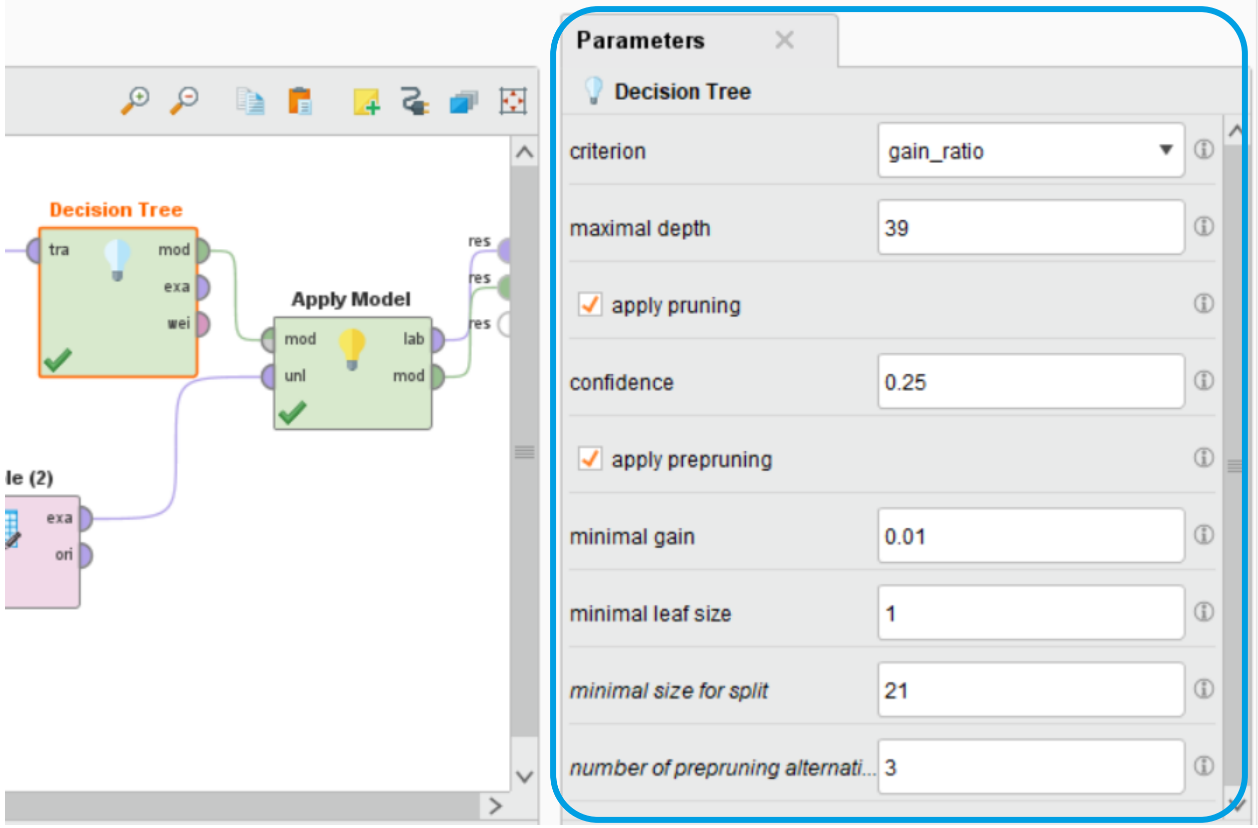
Optimize Parameters (Grid) (3993 rows, 6 columns)

iteration	Decision Tree.criterion	Decision Tree.minimal_size_...	Decision Tree.maximal_...	Decision Tree.confidence	accuracy
1001	information_gain	31	80	0.100	0.700
501	gini_index	11	39	0.050	0.632
1002	gini_index	31	80	0.100	0.632
1	gain_ratio	1	-1	0.000	0.636
502	gain_ratio	21	39	0.050	0.688
1003	gain_ratio	41	80	0.100	0.652
503	information_gain	21	39	0.050	0.664
2	information_gain	1	-1	0.000	0.588
1004	information_gain	41	80	0.100	0.672
504	gini_index	21	39	0.050	0.684
1005	gini_index	41	80	0.100	0.656
505	gain_ratio	31	39	0.050	0.696
1006	gain_ratio	51	80	0.100	0.688
3	gini_index	1	-1	0.000	0.588

DRIVES
Development and Research on Innovative Vocational Education Skills

EVALUATION

Once we figure out the optimized values of the parameters of the *Decision Tree* operator, we can go back to the previous process to try to obtain better results in the classification of the test dataset.



8. Go back to the process and replace the parameters' values of *criterion*, *minimal_size_for_split*, *maximal_depth* and *confidence* for the values found with the optimization process.

DRIVES
Development and Research on Innovative Vocational Education Skills

EVALUATION

Row No.	User_ID	prediction(e...	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)	Gender	Age
1	56031	Early Adopter	0.200	0	0.600	0.200	M	57
2	25913	Early Adopter	0	0	0.875	0.125	F	51
3	19396	Late Majority	0.061	0.879	0.030	0.030	M	41
4	93666	Innovator	0.333	0	0	0.667	M	66
5	72282	Late Majority	0.061	0.879	0.030	0.030	F	31
6	64466	Early Majority	0.750	0.250	0	0	M	68
7	76655	Late Majority	0.333	0.667	0	0	F	51
8	48465	Innovator	0	0.250	0	0.750	F	36

Row No.	User_ID	prediction(e...	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)	Gender	Age
1	56031	Innovator	0	0	0.357	0.643	M	57
2	25913	Early Adopter	0	0	0.800	0.200	F	51
3	19396	Late Majority	0.061	0.879	0.030	0.030	M	41
4	93666	Early Majority	1	0	0	0	M	66
5	72282	Late Majority	0.061	0.879	0.030	0.030	F	31
6	64466	Early Majority	0.815	0.111	0	0.074	M	68
7	76655	Late Majority	0.063	0.874	0.049	0.014	F	51
8	48465	Innovator	0	0.111	0	0.889	F	36

EVALUATION

By updating the parameter values of the *Decision Tree* operator according to the insights found, there was, as expected, an increase in the confidence of the predictions made by the classifier.

This is in accordance with the logic behind the CRISP-DM methodology, which states that the Data Mining process is cyclical, and that it is possible to go back as many times as necessary to redo and readjust the final model in order to obtain the best results.

With these results, Arthur now has the information and knowledge needed to achieve the goals initially proposed.

DRIVES
Development and Research on Innovative Vocational Education Skills

DEPLOYMENT

Arthur's goal is to find out which customers are expected to purchase the new eReader and in what timeframe, based on the latest release of the company's digital reader.

The Decision Tree operator has allowed him to make this prediction and to determine how reliable the predictions are. Arthur has also been able to determine which attributes have the most predictive power in the eReader adoption.

But how can Arthur use this newly discovered knowledge?

The simplest and most direct answer is that he now has a list of customers and their likely adoption timelines for the new eReader. These customers can be identified by the *User_ID*, which enables Arthur to initiate a target marketing process that is timely and relevant to each individual.

DEPLOYMENT

Those who are most likely to purchase the product early on (*Early Adopter*) may be contacted and encouraged to purchase as soon as the new product comes out and may even want the option to pre-order the new device.

Those who are less likely (*Early Majority*) may need some persuasion, perhaps an offer or a discount on another product with the purchase of the new eReader.

Those who are less likely (*Late Majority*), may be passively targeted for marketing, or perhaps not at all, if marketing budgets are tight and money has to be spent on encouraging the customers who are most likely to buy the product.

On the other hand, perhaps very little marketing is needed for *Innovators*, as they are expected to be the most likely to purchase the eReader in the first place.

DEPLOYMENT

Arthur now has a tree that shows him which attributes are most important in determining the likelihood of purchase for each group.

New marketing campaigns can then use this information to focus on increasing the level of activity on the site, or to associate electronics products with eReader discounts on the company's site.

These types of cross-category promotions can be further refined to attract buyers of a specific gender or age group.

With this Data Mining analysis, Arthur now has a wealth of new insights that will help him promote the next generation of eReader.

SUMMARY

- Decision trees are excellent predictive models when the target attribute is **categorical** in nature and when the data set is of **mixed types**.

- Decision trees have the advantage of effectively addressing **attributes with missing or inconsistent values** by not handling them - decision trees work around these data and generate useful results.

- Decision trees consist of **nodes** and **leaves**, which represent the **best predictive attributes** in the data set. These nodes and leaves lead to confidence percentages based on the attributes of the training data set, which can then be applied to similarly structured test data to generate predictions for the test observations (*scoring*).

- Decision trees tell us **what the prediction is**, **how confident** we can be in the prediction, and **how we got to the prediction**. The "how we got to the prediction" part is shown in a graphic representation of the tree.

Decision Trees

[1] What characteristics of dataset attributes might lead you to choose a decision tree data mining methodology rather than a linear regression approach?

[2] What are the confidence percentages for, and why is it important to consider them, other than just considering the prediction attribute?

[3] What are the main advantages of using decision trees compared to other Data Mining techniques?

[4] Download the dataset "titanic_dataset". Import the "titanic-training" data into the RapidMiner repository. Perform the Data Understanding phase.

- [a] What was the percentage of surviving passengers?
- [b] What was the main age group of the passengers who were on the Titanic?
- [c] Have more children or more adults survived?

Decision Trees

[5] Perform the Data Preparation step. Don't forget to place the Set Role operator on the attributes that justify its application.

[6] Using RapidMiner, create a first process using the parameter optimization operator to discover the optimum values for the Decision Tree operator parameters.

[7] In an Excel sheet include some people in the test dataset (titanic-scoring.csv) (you can use information from people you know). Save this Excel sheet as a CSV file and import it into the RapidMiner repository.

Decision Trees

[8] In a new process, repeat the steps in RapidMiner as described earlier to apply the Decision Tree model to the test dataset ("titanic-scoring").

(a) Run the model using the default parameters. After running the model, in the results section, examine the predictions and confidence percentages on the test set. Report the nodes in the tree, and discuss whether the people you entered would be survivors, deceased, or unknown.

(b) Run the model again, but now using the parameter values found in Exercise 6. Report the differences in the tree structure. Discuss whether the chances of survival of you and the people you know increase.

(c) Repeat Exercises 6 and 8(b) until you are satisfied with the results. Present all the attempts, as well as the results obtained and their comparisons.

DRIVES
Development and Research on Innovative Vocational Education Skills

CONTEXT

Peter is an analyst of the statistical performance of a professional team. The management believes that by adding two to four excellent athletes, the team will have an excellent chance to reach the league championship.

Peter needs to identify the best options from a list of 59 athletes. None of the athletes should be excluded without assessing their potential ability to add productivity to the team.

Peter needs to quickly evaluate the past performance of these athletes and make recommendations on the basis of his analysis. We will help Peter by building a neural network, a data mining methodology that predicts categories or classifications in essentially the same way as decision trees but has a greater ability to find the strength of the connections between attributes.

Data mining can help Peter evaluate the different athletes and respond to the needs of the team.

BUSINESS UNDERSTANDING

Peter wants to extract a dataset from all the athletes in the current league to use Data Mining techniques to find the most likely athletes to bring the most excitement, score and defense to the team in order to get to the league championship.

The management informed Peter that their goal was to push for next season's championship, and that they were willing to do everything they could financially to bring in the best athletes.

With the objectives of his superiors in mind, Peter is willing to evaluate each of the 59 prospective athletes in order to make his best recommendations. He knows that the best athletes usually have strong connections between two or more performance areas, while the most typical athletes may have strength in one area, but weaknesses in others.

DATA UNDERSTANDING

Using the data of the league and his knowledge of the league's athletes, Peter prepared a training dataset of 263 observations and 19 attributes. The test data set represents the list of 59 athletes as potential signings. The attributes that make up the dataset are:

- **Player_Name:** corresponds to the name of the athlete. In the data preparation phase, we will assign the role of 'id' to this attribute, since it is not predictive, but it is important to keep it in place so that Peter can easily identify the athletes.
- **Position_ID:** there are 12 possible positions for the sport that Peter's team is playing. Each of them is represented as an integer ranging between 0 and 11 in the datasets.
- **Shots:** total number of shots or scoring opportunities that each athlete had in the most recent season.
- **Makes:** number of times the athlete scored points when shooting during the most recent season.
- **Personal_Points:** number of points that the athlete scored individually during the most recent season.

DATA UNDERSTANDING

- **Total_Points:** total number of points that the athlete has contributed to scoring during the most recent season. Each time an athlete scores a personal point, his/her total points increase by one, and each time an athlete contributes to a teammate's score, his/her total points also increase by one.
- **Assists:** number of times the athlete helped his/her team to get the ball away from the opposing team during the most recent season.
- **Concessions:** number of times that the athlete's play directly led the opposing team to concede an offensive advantage in the most recent season.
- **Blocks:** number of times the athlete directly and independently blocked the shot of the opposing team during the most recent season.
- **Block_Assists:** number of times an athlete has collaborated with a teammate to block the shot of the opposing team during the most recent season.
- **Fouls**: number of times, in the most recent season, that the athlete has committed a foul.

DATA UNDERSTANDING

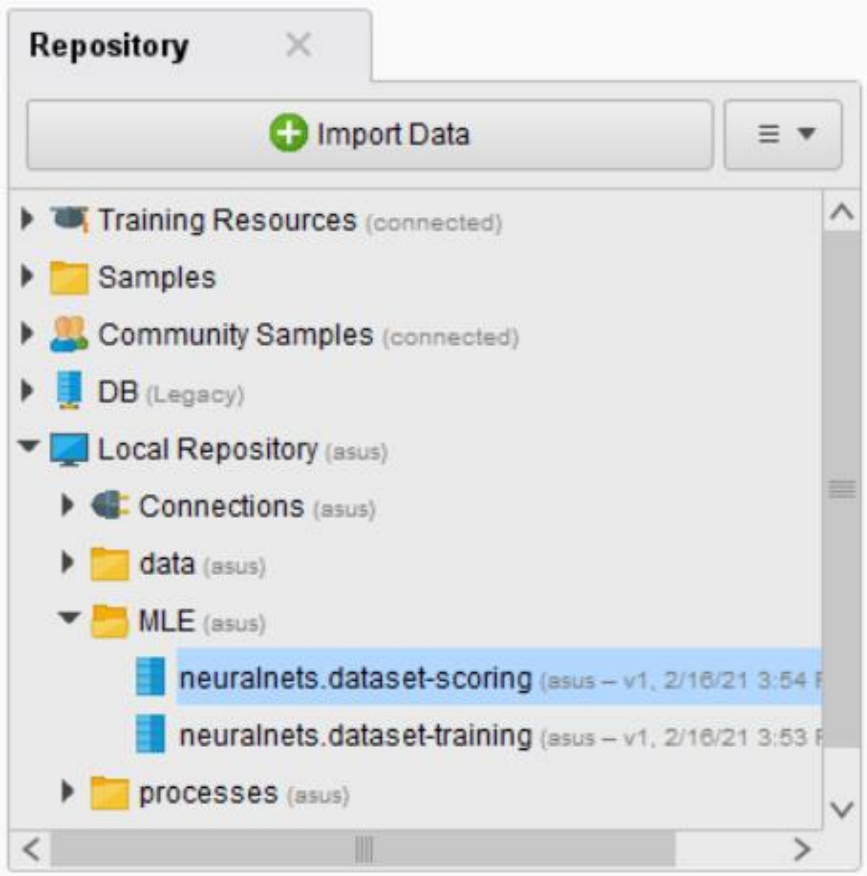
- **Years_Pro:** in the training *dataset*, this is the number of years the athlete has played at the professional level. In the test *dataset*, this is the number of years the athlete has had experience, including years as a professional, if any, and years in amateur leagues.
- **Career_Shots:** represents the same as the *Shots* attribute, except that it is cumulative for the athlete's entire career. All the career attributes (Career_) are an attempt to assess the athlete's ability to perform consistently over time.
- **Career_Makes:** represents the same as the *Makes* attribute, except that it is cumulative for the athlete's entire career.
- **Career_PP:** represents the same as the *Personal_Points* attribute, except that it is cumulative for the athlete's entire career.
- **Career_TP:** represents the same as the *Total_Points* attribute, except that it is cumulative for the athlete's entire career.
- **Career_Assists:** represents the same as the *Assists* attribute, except that it is cumulative for the athlete's entire career.
- **Career_Con:** represents the same as the *Concessions* attribute, except that it is cumulative for the athlete's entire career.

DATA UNDERSTANDING

- **Team_Value:** categorical attribute that summarizes the value of an athlete to his/her team. It is only present in the training data as it will serve as a "label" in the test data set. There are four categories available:
  - Role Player: athlete who is good enough to play at the professional level, and who may be really good in one category, but not excellent overall.
  - Contributor: athlete who contributes to defense and offense in different categories and who can be counted on to help the team win on a regular basis.
  - Franchise Player: athlete whose skills are so broad, strong and consistent that the team will want to hold on to them for a long time.
  - Superstar: rare athlete whose talents are so superior that he/she makes the difference in every game. Most teams in the league have such an athlete, but the teams that are always fighting for the league title have two or three.

DATA PREPARATION

Download the datasets: neuralnets.dataset-training.csv
neuralnets.dataset-scoring.csv



1. Import the *datasets* into the RapidMiner repository (Import Data -> My Computer).

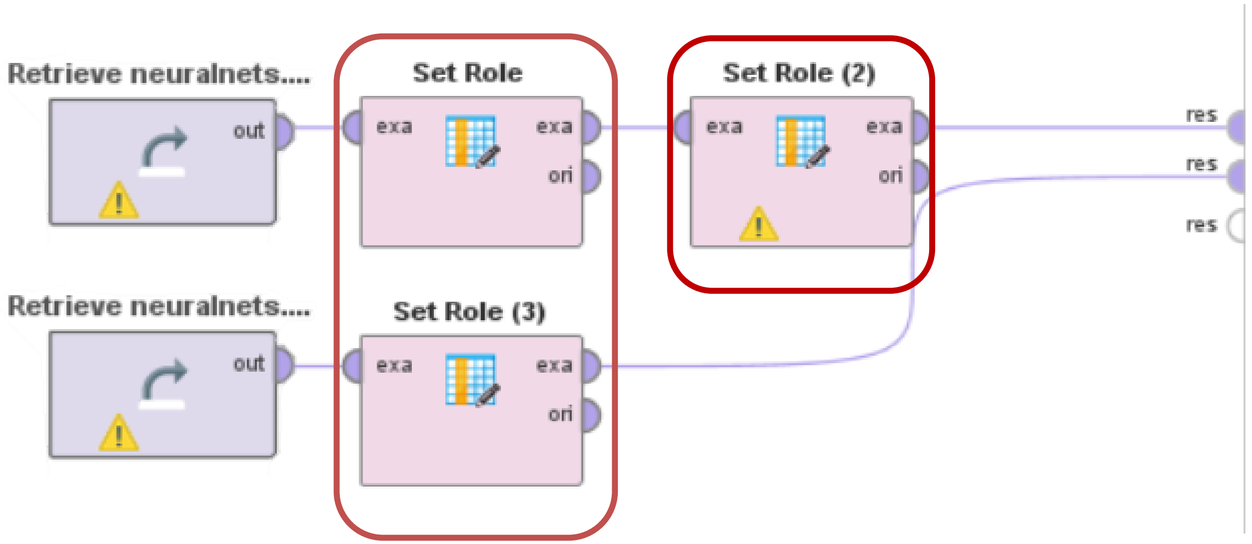
2. In the design perspective, drag the two datasets into the process window. Connect both *out* ports to the *res* ports and then run the model. Check for missing values and categorical attributes. <u>Neural network models do not work well with missing values or categorical data.</u>

DATA PREPARATION

3. Note that although the range of some attributes in the test dataset is not included in the range of attributes in the training dataset, neural networks do not require this transformation. They use the concept of fuzzy logic that allows you to deal with imprecise and uncertain situations. This is a probability-based inferential approach to data comparison, which allows to infer, based on probabilities, the strength of the relationship between the attributes of the datasets.

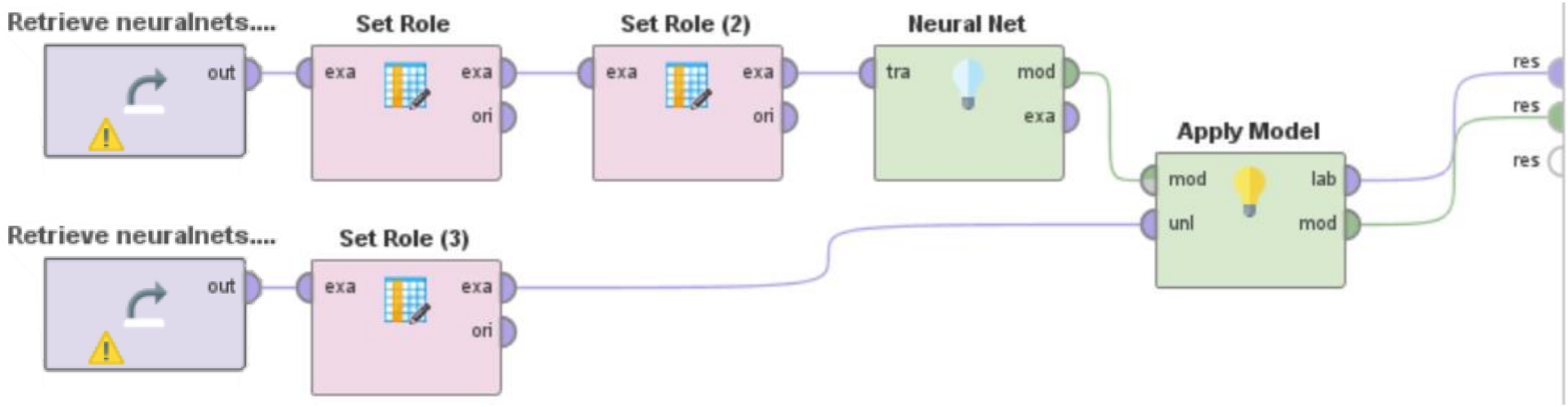
DATA PREPARATION

4. Add two operators of the *Set Role* type, one for each flow. Use these operators to set the *Player_Name* attribute role to 'id’, so that it is not included in the prediction calculations of the neural network. Drag a third operator of the *Set Role* type and place it in the training flow, set the *Team_Value* attribute to the 'label' of the model.



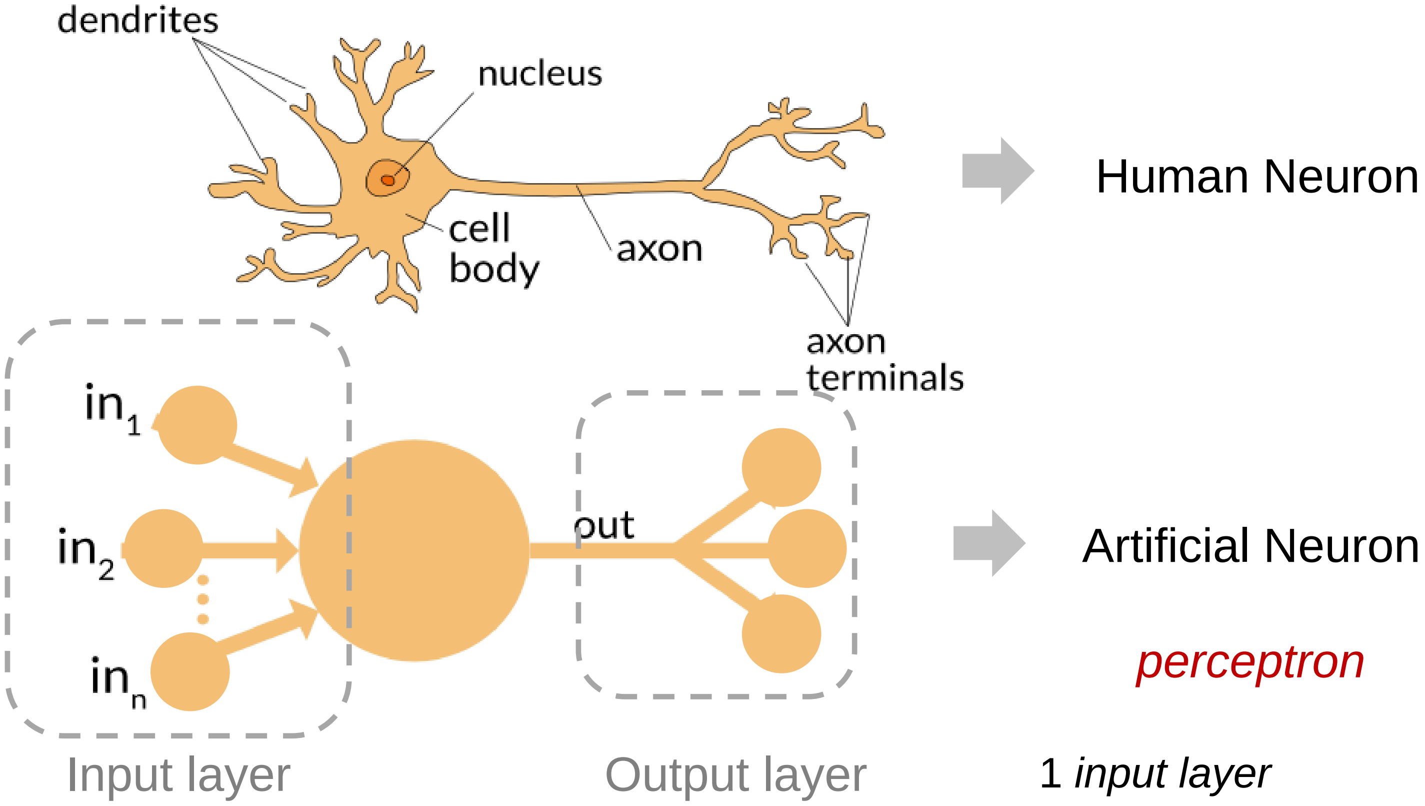
MODELING

1. In the Operators tab look for the 'Neural Net' operator and add it to the training flow. Also look for the 'Apply Model' operator and drag it into the process window, joining the training and scoring flows. Make sure that both the *lab* and *mod* ports are connected to the *res* ports in order to generate the desired results.



MODELING

2. Run the model. In the results you will see the graphical model (*ImprovedNeuroNet (Net)* tab) and the predictions (*ExampleSet* tab).



perceptron

1 *input layer*
1 *output layer*

Neural networks are computational models inspired by the human brain, consisting of a network architecture composed of artificial neurons to detect hidden patterns in data.

DRIVES
Development and Research on Innovative Vocational Education Skills

MODELING

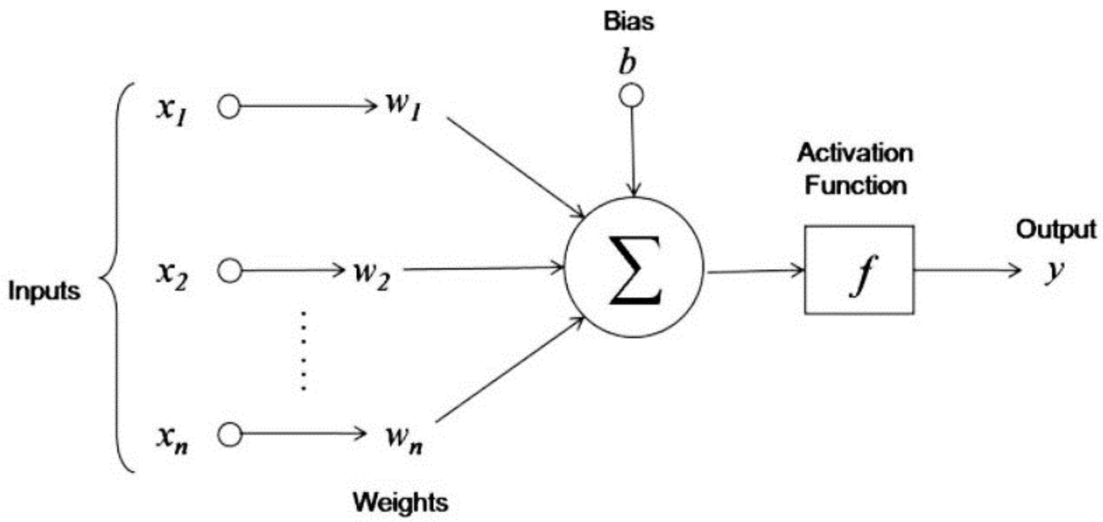
The circles in the graph of the neural network are **nodes** or **neurons**. A node combines the data input with a set of **coefficients** or **weights**, which **amplify** or **dampen** that input, thereby assigning meaning to the inputs in relation to the task the algorithm is trying to learn. The thicker and darker the connection between the nodes, the stronger the affinity between those nodes.

An Artificial Neural Network (ANN) is typically used to model **complex** and **non-linear** relationships between input and output variables. This is possible due to the existence of more than one layer, in addition to the input and output layers, called the **hidden** layer (*multi-layer perceptron*). A hidden layer contains a layer of nodes that connects inputs from previous layers and applies an activation function. Neural nets use the hidden layer to compare all attributes in a dataset with all other attributes.

MODELING

How Neural Networks work?

A node combines the data input from the previous column with a **coefficient** or **weight** that determines the connection between two neurons. These input-weight products are summed and then a biased value - **bias** - is added to the total calculated value. Finally, an **activation function** is applied to the obtained value to transform the previously calculated total value to a number between 0 and 1 in order to determine if and to what extent this signal should progress further through the network to affect the final result. The RapidMiner operator uses a sigmoid function to do this transformation.



The type of output node(s) is sigmoid if the learning data describes a classification task (the # of output nodes is equal to the possible classes) and linear if the learning data describes a regression task (it presents only one node as output).

MODELING

A "***feed-forward***" neural network is an artificial neural network in which the connections between nodes do not form a directed loop. In this network, information moves in only one direction, forward, from the input nodes, through the hidden nodes (if any) to the output nodes.

A multi-layer perceptron network uses "***back propagation***" to train the network. The back propagation algorithm is a supervised learning method that can be divided into two phases: **propagation** and **weight updating**. The two phases are repeated until the network performance is good enough. The output values are compared with the correct answer to calculate the error value. Using this information, the algorithm adjusts the weights of each link in order to reduce the error value. After repeating this process for a sufficiently large number of training cycles (*training_cycles*) or reaching an error value below the set error value (*error_epsilon*), the network training phase ends.

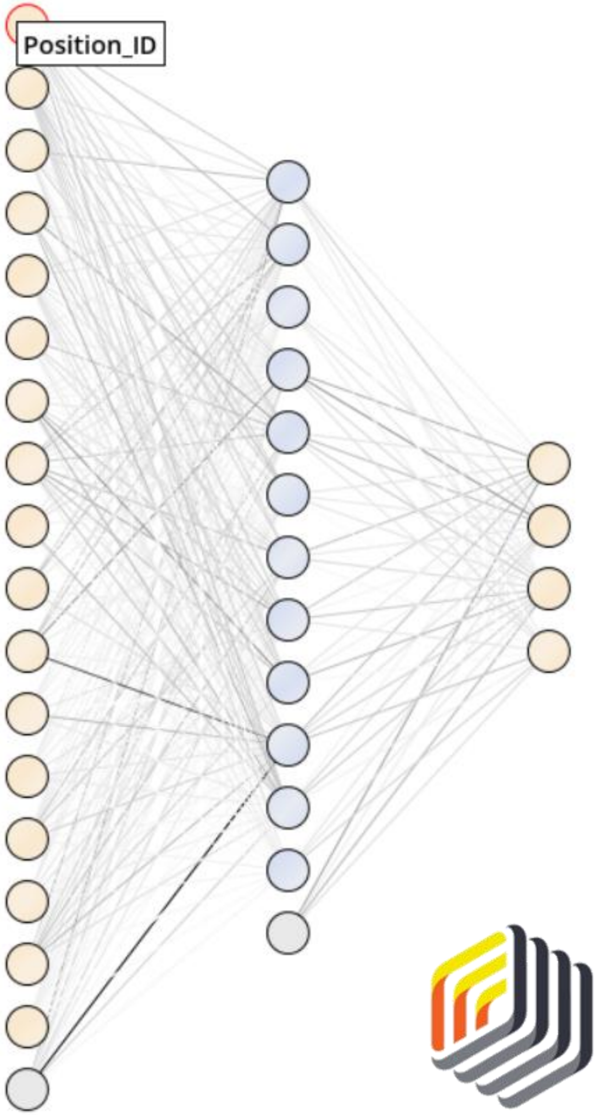
EVALUATION

The graph starts at the left, with a node for each of the predictor attributes. If you click on these nodes on the left side, the name of the attribute it represents will be revealed. The first layer of nodes nearest to the input is called the ***input layer***.

The ***hidden layer*** performs the comparison between all these attributes.

The layer of nodes on the right is called the ***output layer*** and represents the four possible values for the attribute to predict (label): *Role_Player*, *Contributor*, *Franchise Player* or *Superstar*.

Graphical view of the neuronal network showing different neurons with different strengths and the four nodes for each of the possible *Team_Value* categories.



EVALUATION

1. Switch to the 'ExampleSet' tab. Again, as with the previous predictive models, we can see that four new special attributes have been generated by RapidMiner. Each of the 59 athletes has a prediction of their *Team_Value* category, with their respective confidence percentages.

Name	Type	Missing		
Id: Player_Name	Polynominal	0	Least: Zachary Lawson (1)	Most: Alan Hunter (1)
Prediction: prediction(Team_Value)	Polynominal	0	Least: Superstar (7)	Most: Role Player (22)
Confidence_Superstar: confidence(Superstar)	Real	0	Min: 0.000	Max: 0.941
Confidence_Contributor: confidence(Contributor)	Real	0	Min: 0.003	Max: 0.813
Confidence_Franchise Player: confidence(Franchise Player)	Real	0	Min: 0.007	Max: 0.586
Confidence_Role Player: confidence(Role Player)	Real	0	Min: 0.000	Max: 0.933
Position_ID	Integer	0	Min: 0	Max: 11

DRIVES
Development and Research on Innovative Vocational Education Skills

EVALUATION

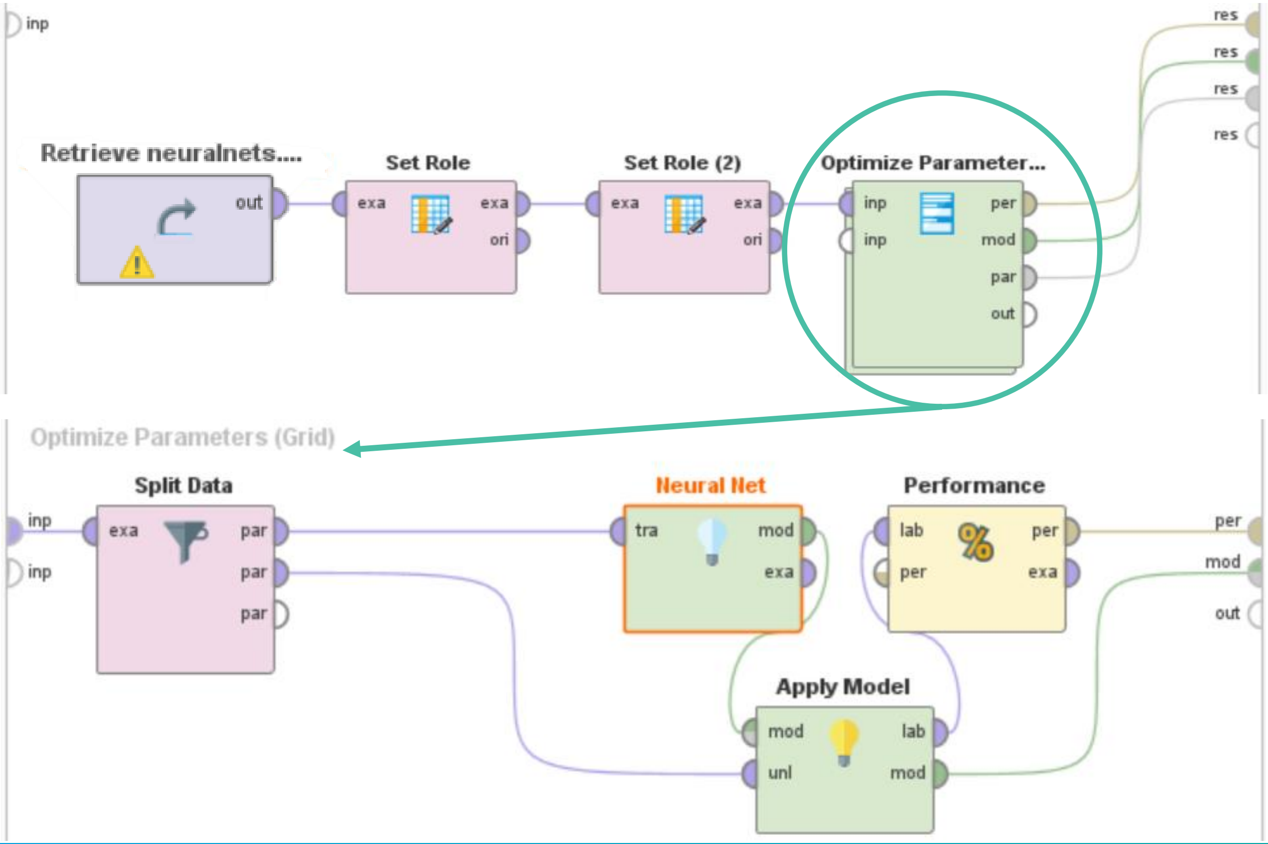
So far, the results of this type of predictive model are quite familiar as they are similar to some models already studied. At this point, all 59 athletes are categorized in a predictive manner and it is also known how confident RapidMiner is in these predictions.

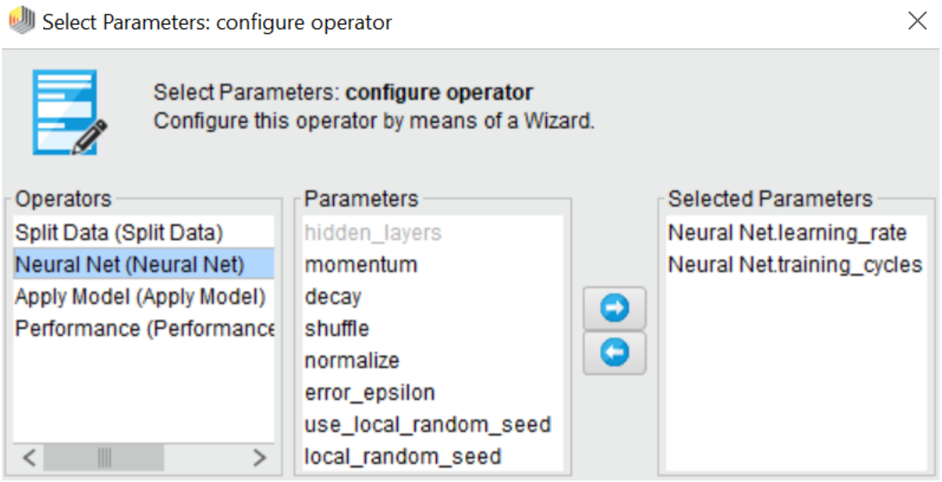
Row No.	Player_Name	prediction(T...	confidence(Superstar)	confidence(Contributor)	confidence(Franc...	confidence(Role
1	Gary Price	Franchise Pl...	0.338	0.299	0.363	0.000
2	Raul Little	Contributor	0.011	0.719	0.269	0.000
3	Roman Rich...	Contributor	0.193	0.519	0.289	0.000
4	Geoffrey Lloyd	Contributor	0.137	0.547	0.316	0.000
5	Jesus Huff	Contributor	0.160	0.539	0.301	0.000
6	Jan Becker	Franchise Pl...	0.236	0.331	0.433	0.000
7	John Mcguire	Superstar	0.826	0.003	0.170	0.000
8	Robert Hollo...	Superstar	0.906	0.007	0.087	0.000
9	Herbert Watk...	Franchise Pl...	0.122	0.373	0.505	0.000
10	Stewart Chav...	Contributor	0.093	0.468	0.439	0.000
11	Ralph Sharp	Franchise Pl...	0.155	0.315	0.530	0.001
12	Drew Kelley	Superstar	0.765	0.033	0.202	0.000
13	Jessie Strick...	Franchise Pl...	0.229	0.344	0.427	0.000

EVALUATION

2. Create a new process to optimize the parameters of the Neural Net operator. Drag the "Optimize Parameters (Grid)" operator and follow the steps until you get the process and subprocess below.

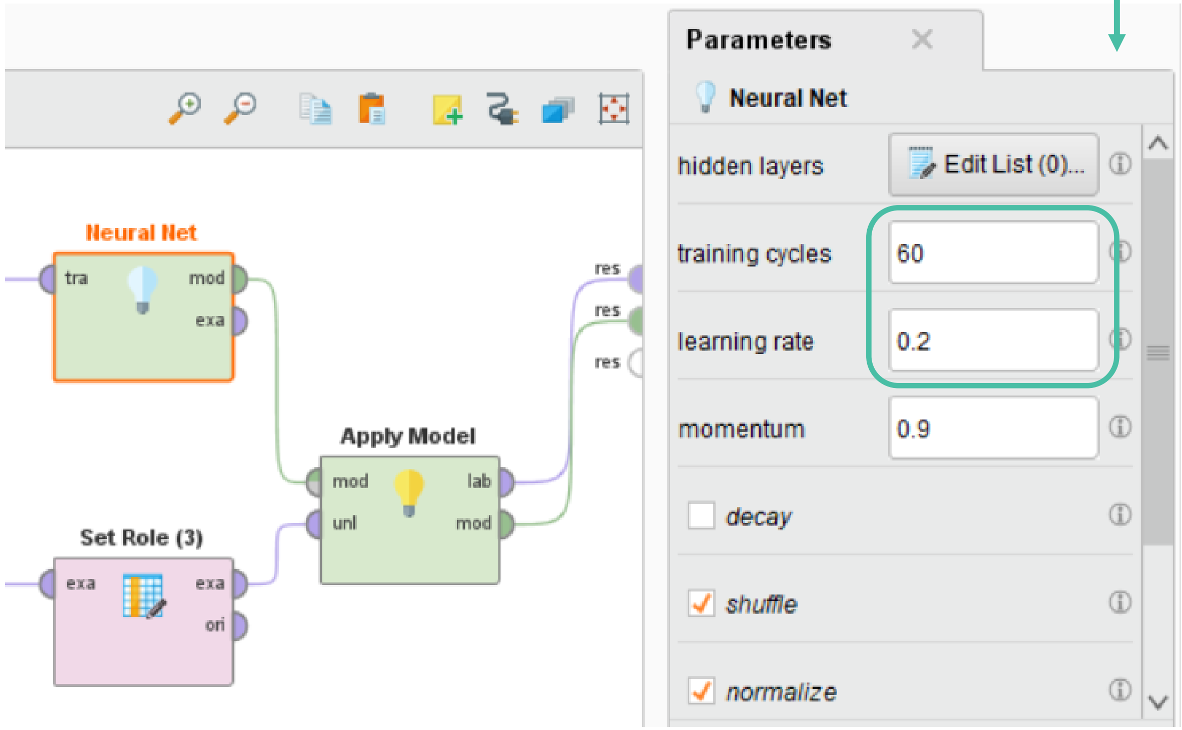
We will use the 0.8/0.2 partition because it is the distribution that most closely matches our data division.





EVALUATION

3. Replace the parameters of the "Neural Net" operator with the best values obtained in the optimization process, run the model and observe the results obtained.



Optimize Parameters (Grid) (121 rows, 4 columns)

iteration	Neural ...	Neural ...	acc... ↓
69	0.200	60	0.633
113	0.200	100	0.595
46	0.100	41	0.582
59	0.300	51	0.582
102	0.200	90	0.582
37	0.300	31	0.570
24	0.100	21	0.570
106	0.600	90	0.570
104	0.400	90	0.570
14	0.200	11	0.557
38	0.400	31	0.557
68	0.100	60	0.557
71	0.400	60	0.557

DEPLOYMENT

Peter wanted to evaluate these 59 players' perspectives quickly and easily on the basis of their past performance. He can implement his model by responding to the club with a number of different outputs from our neural network. First, he can double-click the *prediction(Team_Value)* column header to bring all the Superstars to the top. (Superstar is the last of our values in alphabetical order, so it is first in reverse alphabetical order).

Row No.	Player_Name	prediction(Team_Val... ↓	confidence(Superstar)	confidence(Contributor)	confidence(Franchise ...	confidence(Role Play
1	Gary Price	Superstar	0.993	0.007	0.000	0.000
7	John Mcguire	Superstar	1.000	0.000	0.000	0.000
8	Robert Hollo...	Superstar	0.998	0.001	0.001	0.000
12	Drew Kelley	Superstar	0.999	0.000	0.001	0.000
14	Gerald Luna	Superstar	1.000	0.000	0.000	0.000
15	Fred Clarke	Superstar	0.908	0.056	0.033	0.003
22	Ian Tucker	Superstar	0.998	0.002	0.000	0.000
28	Rodolfo Jaco...	Superstar	1.000	0.000	0.000	0.000
36	Johnny Denn...	Superstar	0.978	0.022	0.000	0.000

DEPLOYMENT

At the top, 9 athletes with the potential to be superstars are now shown. Furthermore, 3 of them - John Mcguire, Gerald Luna and Rodolfo Jacobs - reach 100% confidence.

Peter may want to go ahead and quickly recommend to the club management that they contact these three athletes. Drew Kelly is also extremely close, with only a small chance of being a Franchise Player rather than a Superstar. Even Franchise players are athletes with a massive upside, so the risk of pursuing this player is minimal.

Peter knows these players are probably already on the radar of a number of other teams. Perhaps he should look for alternatives that are not so obvious to every club. Peter can win by thinking creatively, and his experience has taught him that sometimes the best player acquisitions are not always the most obvious.

DEPLOYMENT

1. Click on *confidence(Franchise_Player)* twice. There are 5 players out of the 59 considered Franchise Players. Perhaps Peter could suggest to management that a solid player might be Jan Becker or Samuel French. These players may be easier to sign, because probably fewer teams have contacted them, and they may be cheaper in terms of salary than most Superstar players.

Row No.	Player_Name	prediction(Team_Value)	confidence(Superstar)	confidence(Contributor)	confidence(Franch...	confidence(Rol
6	Jan Becker	Franchise Player	0.085	0.009	0.905	0.001
41	Samuel French	Franchise Player	0.012	0.049	0.904	0.036
13	Jessie Strickland	Franchise Player	0.131	0.105	0.757	0.007
17	Jerry Reed	Franchise Player	0.030	0.199	0.755	0.017
24	Harvey Dean	Franchise Player	0.238	0.008	0.741	0.013

DEPLOYMENT

Consider Harvey Dean on line 24. Dean is expected to be a Franchise Player, so Peter knows that he can play consistently at a high level. He would be a solid, long-term acquisition for any team. Although the Franchise Player percentage is 74%, our neural network predicts that there is almost a 24% chance that Dean will rise to the Superstar level. With 9 years of experience, Dean may be ready to reach the peak of his career next season. While he was not the first or most obvious choice in the dataset, Dean certainly seems like an athlete worth considering.

While the model and its predictions provided a great deal of information to consider, it is clear that Peter must continue to use his knowledge, experience, and assessment of other factors not included in the datasets to make his final recommendations. Amateur players, for example, have performance statistics that may not be representative of their ability to perform at a professional level.

SUMMARY

- Neural networks attempt to mimic the human brain by using artificial "neurons" to compare attributes with each other and look for strong connections. By receiving attribute values, processing them, and generating neurons, this data mining model can provide predictions and confidence percentages. Neural networks are not as limited in terms of value ranges compared to other methodologies.

- In their graphical representation, neural networks are drawn using nodes connected together. The thicker or darker the line between the nodes, the stronger the connection represented by that neuron. Stronger neurons are equal to stronger predictive capability. The computer is able to read the network and apply the model to test data to make predictions.

- Between the prediction and confidence percentages, we can use neural networks to find interesting observations that may not be obvious, but that represent good opportunities to solve problems.

Neural Networks

[1] Where did the name neural networks come from? What are the features of the model that make it "neuronal"?

[2] What advantage(s) do neural networks have over other prediction models?

[3] How should the confidence percentages be used in conjunction with the neural network predictions?

[4] What are the layers that make up the neural networks and what do they represent?

[5] Download the dataset "credit-training_dateset" and import it into the RapidMiner repository. Perform the Data Understanding phase.

(a) What levels of credit risk exist?

(b) What is the average loan amount?

Neural Networks

[6] Create your own test dataset using the attributes of the training dataset as a guide. Enter at least 20 observations. You can enter data for people you know (you may need to estimate some attribute values, e.g. credit score) or you can simply test different values for each attribute. For example, you can choose to enter four consecutive observations with the same values in all attributes except the credit score, where you can increase the credit score for each observation by 100, from 400 to 800.

[7] Perform the Data Preparation step. Do not forget to place the Set Role operator on the attributes that justify its application, taking into account that the objective is to predict credit risk.

[8] In a new process, repeat the steps in RapidMiner as described above to apply the neural network model to the test dataset. You may choose to run a process instead to discover the optimal values of the neural network operator parameters.

[9] Run the model and analyze the predictions for each test observation. Report the results, including interesting or unexpected results.

DRIVES
Development and Research on Innovative Vocational Education Skills

- CRISP-DM is a popular methodology used for increasing the success of a DM project and is composed by six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

- There are different ML softwares available to conduct DM projects. In this lecture, we explore the application of ML algorithms on both Weka and RapidMiner.

- Business Understanding focuses on the definition of the project objective from a business perspective, then converting it into a DM problem definition.

- Data Understanding involves acquisition, analysis and exploration of data.

- Data Preparation involves data integration, cleaning, transformation, reduction, and sampling.

- Modeling consists in the application of different ML algorithms.

- Evaluation regards the assessment of the quality of the results obtained by the models and the verification of their impact on the DM objective initially defined. There are several evaluation metrics available to assess the models such as Recall, Accuracy, and AUC.

- Deployment concerns the implementation, monitoring and maintenance of the final models.

REFERENCE TO AUTHORS

Diana Ferreira

- PhD student in Biomedical Engineering
- Research Collaborator of the Algoritmi Research Center

Regina Sousa

- PhD student in Biomedical Engineering
- Research Collaborator of the Algoritmi Research Center

José Machado

- Associate Professor with Habilitation at the University of Minho
- Integrated Researcher of the Algoritmi Research Center

REFERENCE TO AUTHORS

DRIVES
Development and Research on Innovative Vocational Education Skills

António Abelha

- Assistant Professor at the University of Minho
- Integrated Researcher of the Algoritmi Research Center

iD

Victor Alves

- Assistant Professor at the University of Minho
- Integrated Researcher of the Algoritmi Research Center

iD [0000-0003-1819-7051](0000-0003-1819-7051)

REFERENCE TO AUTHORS

This Training Material has been certified according to the rules of **ECQA – European Certification and Qualification Association.**

The Training Material was developed within the international job role committee **“Machine Learning Engineer”:**

UMINHO – University of Minho (https://www.uminho.pt/PT)

The development of the training material was partly funded by the EU under Blueprint Project DRIVES.

DRIVES

Development and Research on Innovative Vocational Education Skills

Thank you for your attention

DRIVES project is project under **The Blueprint for Sectoral Cooperation on Skills in Automotive Sector**, as part of New Skills Agenda.

The aim of the Blueprint is **to support an overall sectoral strategy and to develop concrete actions to address short and medium term skills needs.**

Follow DRIVES project at:

More information at:

www.project-drives.eu

The Development and Research on Innovative Vocational Educational Skills project (DRIVES) is co-funded by the Erasmus+ Programme of the European Union under the agreement 591988-EPP-1-2017-1-CZ-EPPKA2-SSA-B. The European Commission support for the production of this publication does not constitute endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.