AUTOMATIZATION
OF ANALYSES IN RAPID MINER

Petr Berka
University of Economics, Prague
berka@vse.cz

Seminar

Vysoká škola manažmentu v Trenčíne
International Workshop on Knowledge Management

IWKM‘2019

November, 7– 8
Bratislava 2019

Automatization of analyses in Rapid Miner

Petr Berka
(berka@vse.cz)

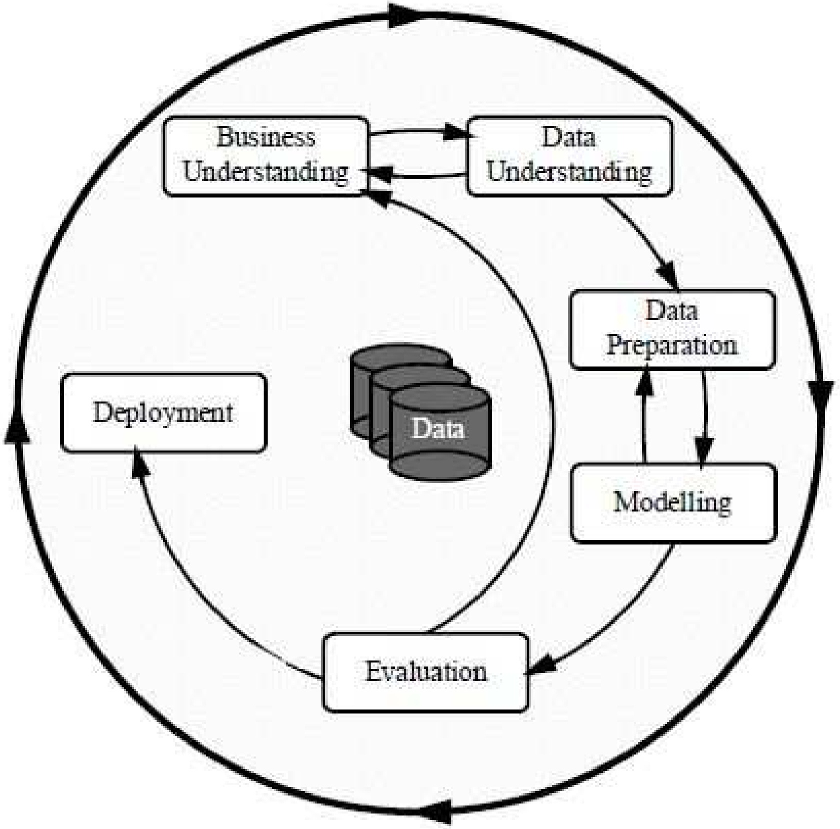
University of Economics
Prague

Automatization of KDD process…

… will allow the domain experts (knowledge workers) to perform data mining tasks without the cooperation with data mining experts, in a “do-it-yourself” way.

As the authors of the DataRobot platform believe: “Automated machine learning creates a new class of citizen data scientists with the power to create advanced machine learning models, all without having to learn to code or understand when and how to apply certain algorithms”.

KDD Process and …



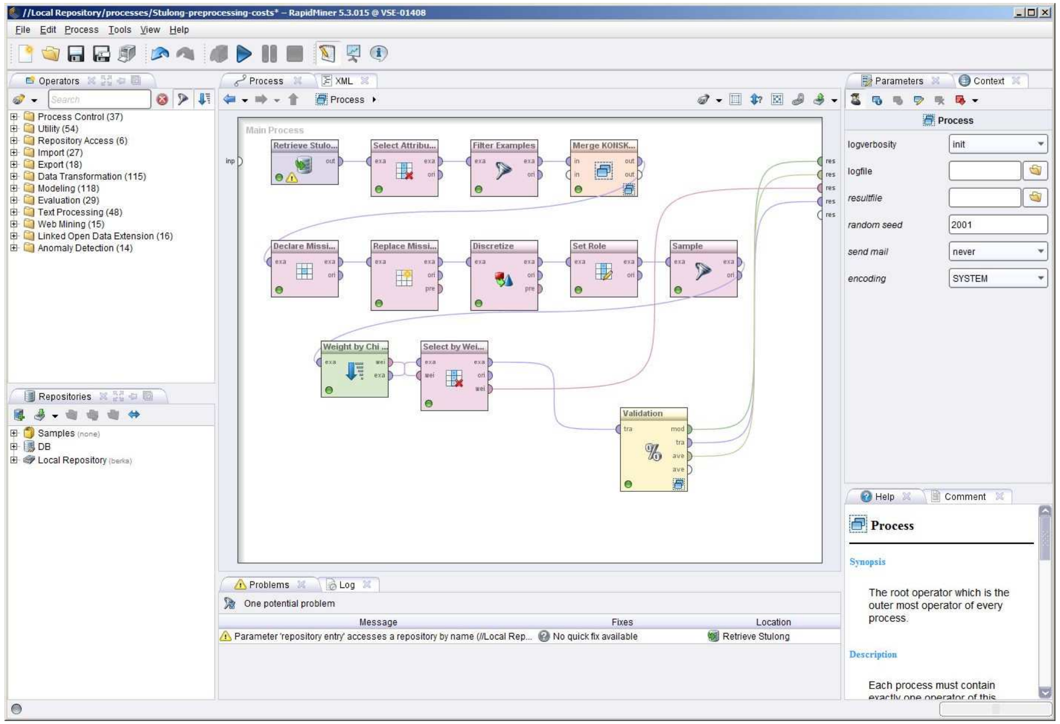
- **Business Understanding** is the initial phase that focuses on understanding the project objectives
- **Data Understanding** proceeds with activities in order to get familiar with the data
- **Data Preparation** phase covers all activities to construct the final dataset
- In the **Modeling** phase various modeling techniques are applied
- At the **Evaluation** stage the quality of the created model is assessed
- the **Deployment** phase can be as simple as generating a report or as complex as implementing a repeatable data mining process

… Possibilities for its Automatization

- **Business Understanding** and **Deployment** steps are closely related to the application domain so its automatization in a general way would be very difficult
- **Data Understanding** is already supported by computing basic characteristics of the data
- data mining automatization is oriented towards supporting the **Data Preparation** (preprocessing) and **Modeling** (learning) steps

Rapid Miner (rapidminer.com)

- a leading open-source system for knowledge discovery and data mining (www.kdnuggets.com)
- a Leader in 2016 Gartner Magic Quadrant for Advanced Analytics (www.gartner.com)
- the Top 3 Rated Predictive Analytics Software for Enterprise (www.g2crowd.com)



//Local Repository/processes/Stulong-preprocessing-costs* – RapidMiner 5.3.015 @ VSE-01408

- Process Control (37)
- Utility (54)
- Repository Access (6)
- Import (27)
- Export (18)
- Data Transformation (115)
- Modeling (118)
- Evaluation (29)
- Text Processing (48)
- Web Mining (15)
- Linked Open Data Extension (16)
- Anomaly Detection (14)

- Samples (none)
- DB
- Local Repository (berka)

Process

logverbosity	init
logfile	
resultfile	
random seed	2001
send mail	never
encoding	SYSTEM

Problems | Log

One potential problem

Message	Fixes	Location
Parameter 'repository entry' accesses a repository by name (//Local Rep...	No quick fix available	Retrieve Stulong

Process

Synopsis

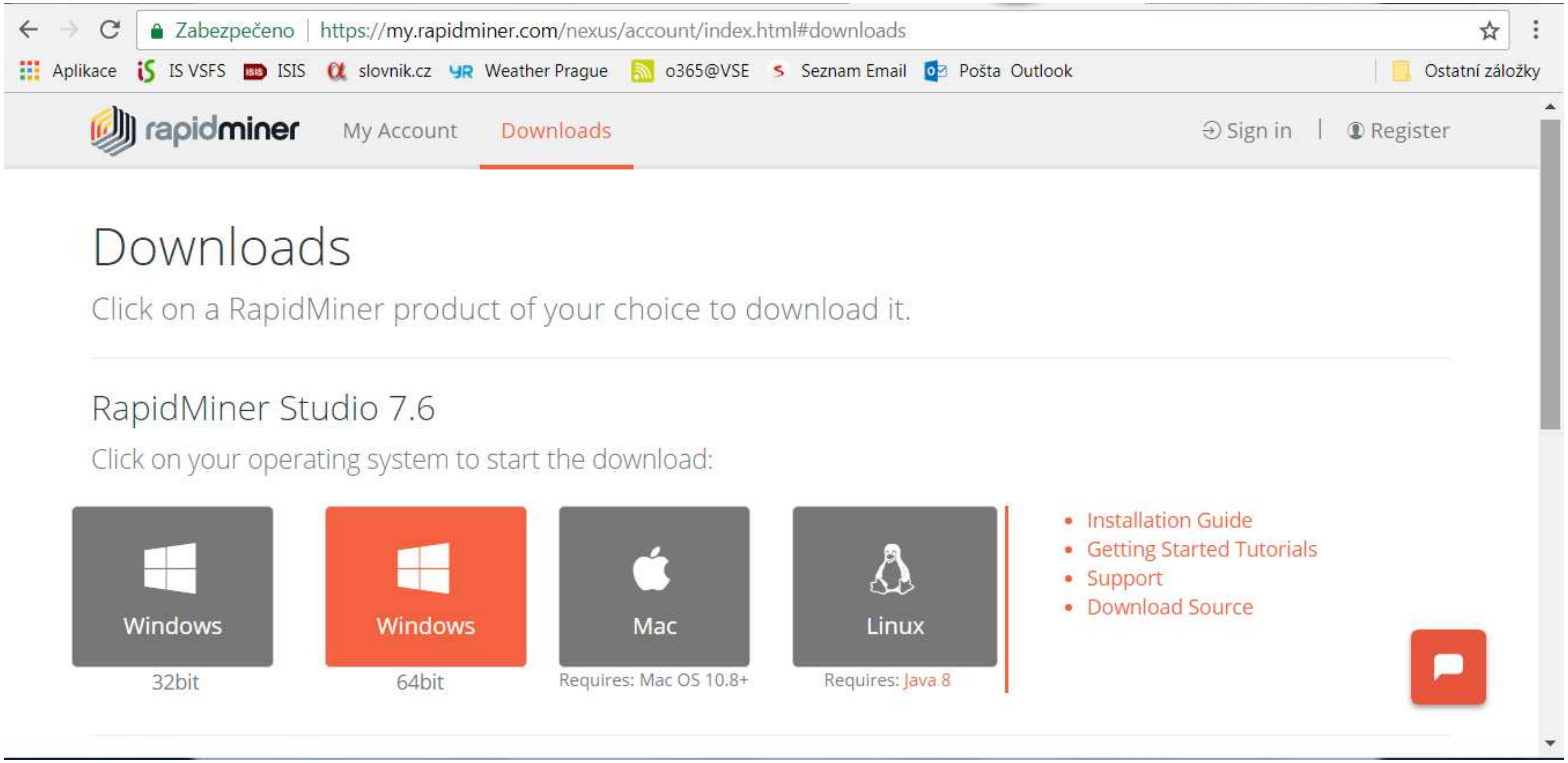
The root operator which is the outer most operator of every process.

Description

Each process must contain exactly one operator of this

Rapid Miner Downloads

https://my.rapidminer.com/nexus/account/index.html #downloads



Downloads

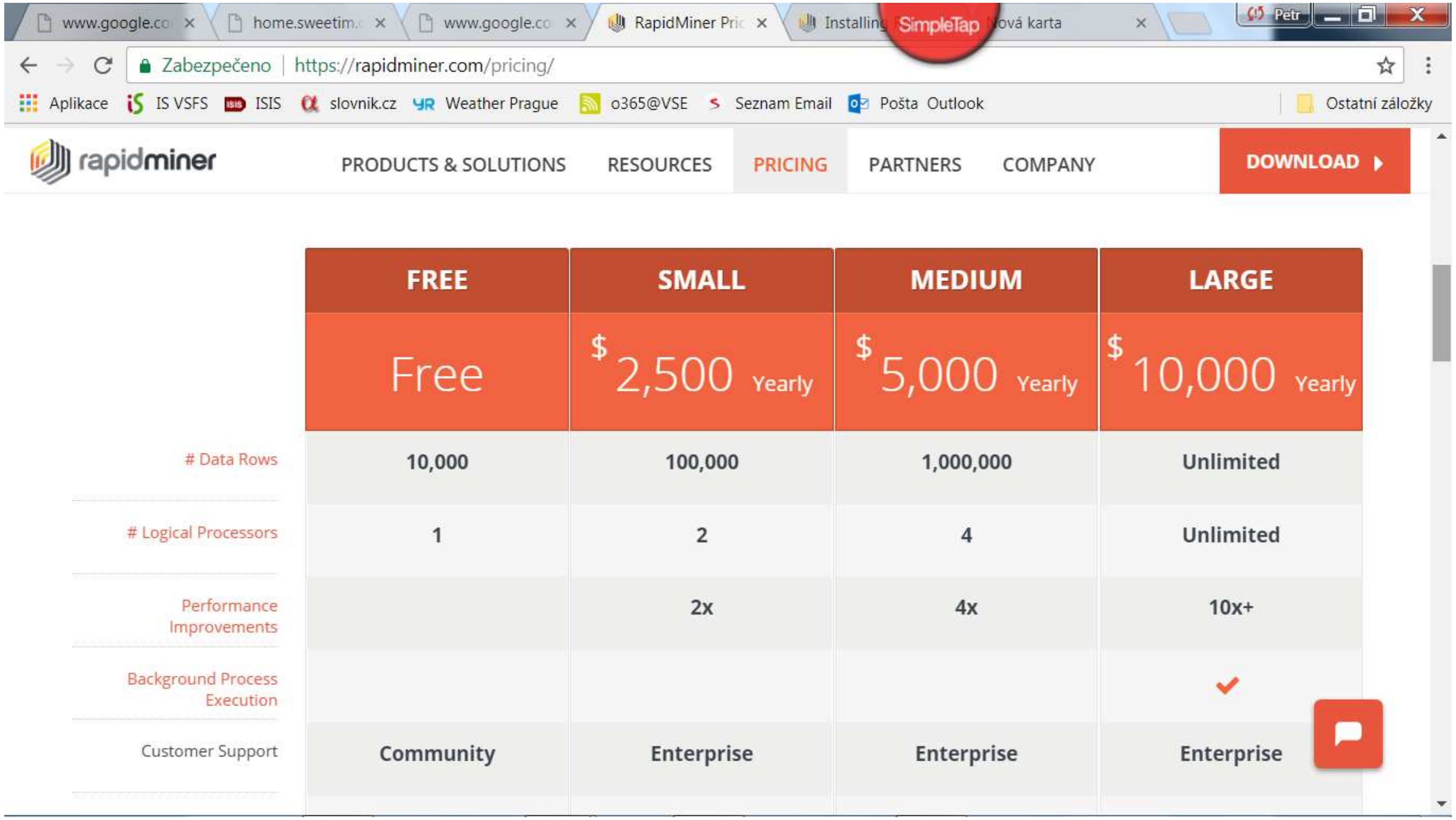
Click on a RapidMiner product of your choice to download it.

RapidMiner Studio 7.6

Click on your operating system to start the download:

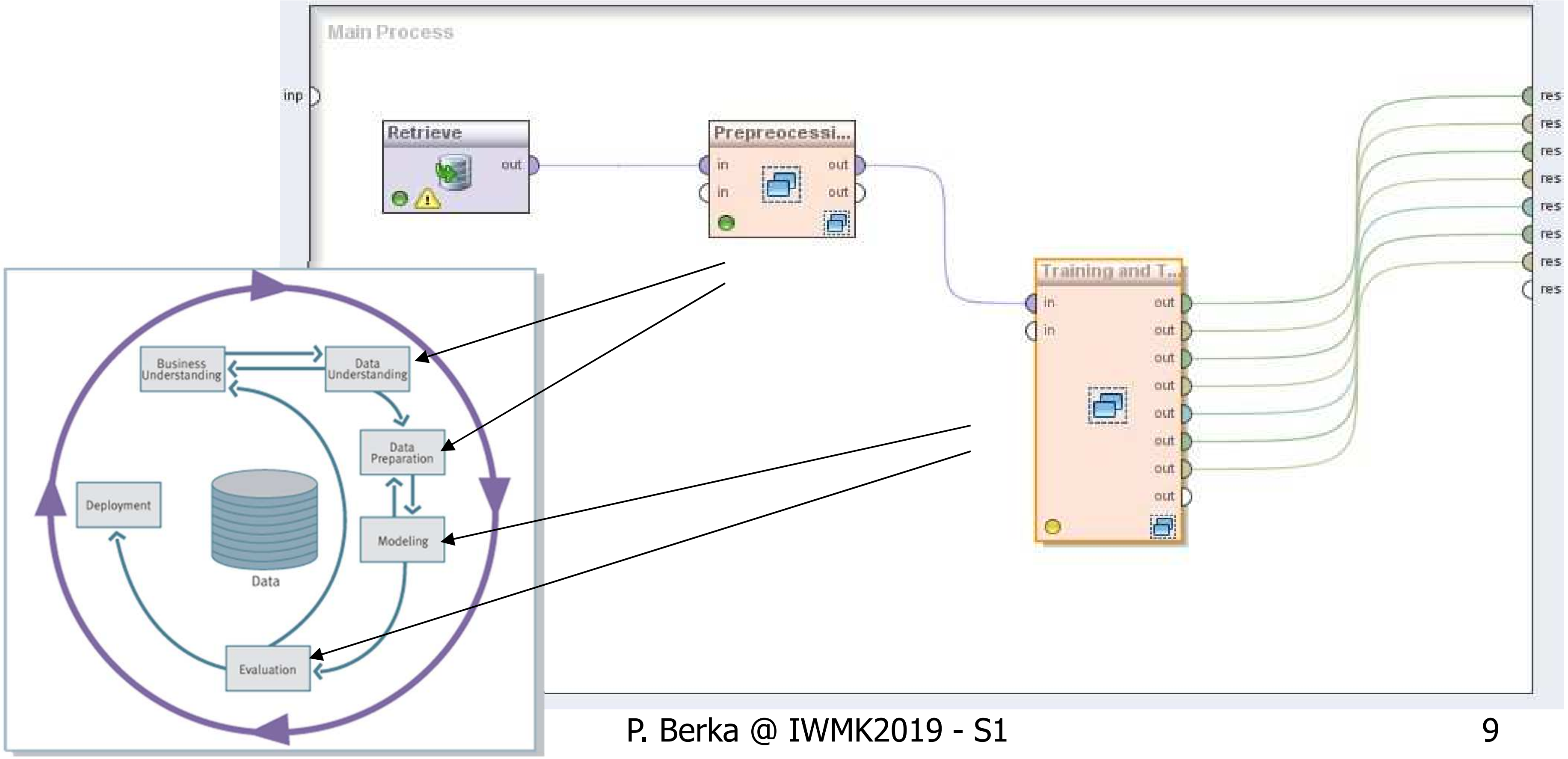
- Installation Guide
- Getting Started Tutorials
- Support
- Download Source

Rapid Miner Pricing

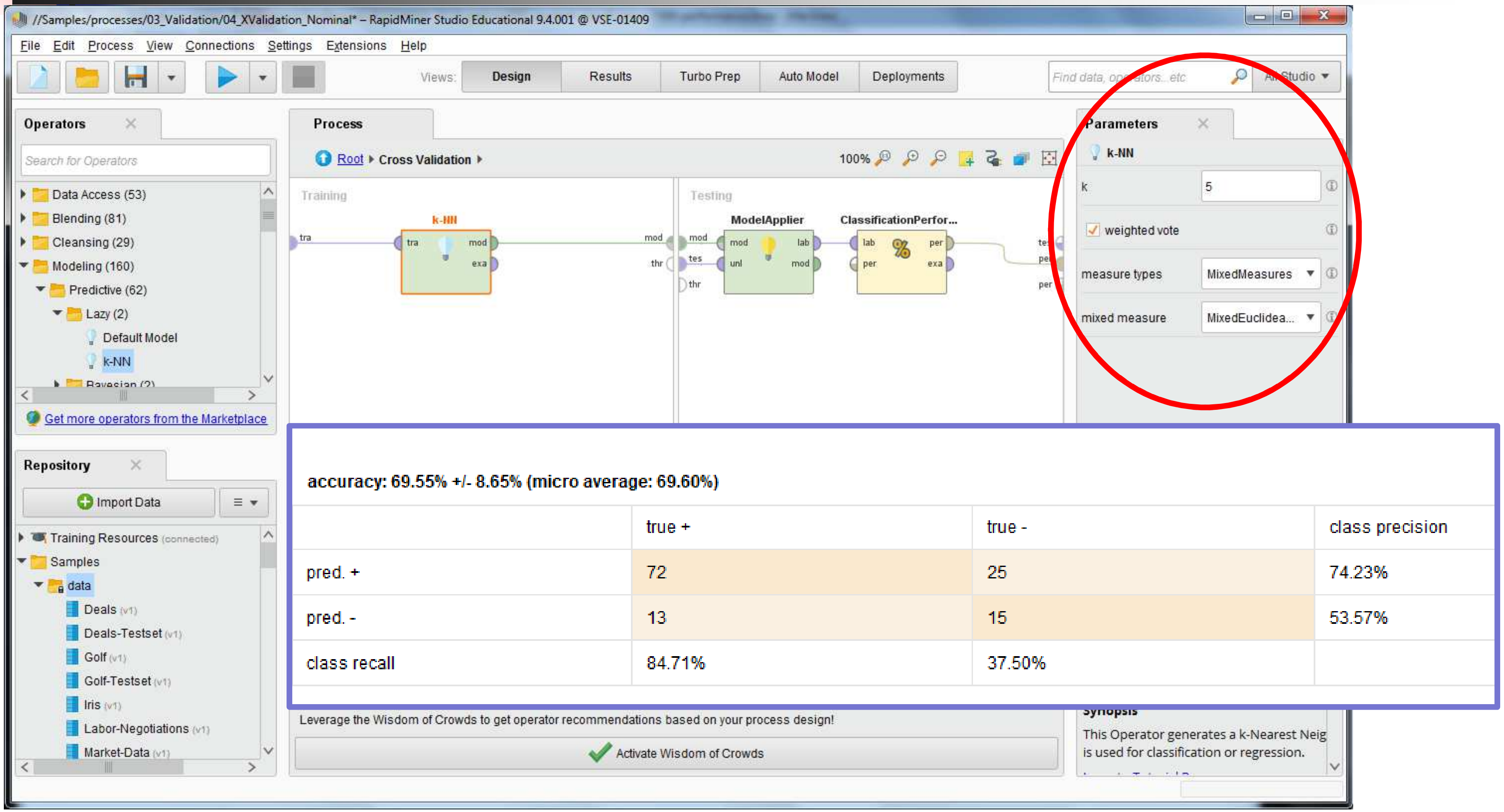


	FREE	SMALL	MEDIUM	LARGE
	Free	$ 2,500 Yearly	$ 5,000 Yearly	$ 10,000 Yearly
# Data Rows	10,000	100,000	1,000,000	Unlimited
# Logical Processors	1	2	4	Unlimited
Performance Improvements		2x	4x	10x+
Background Process Execution				✓
Customer Support	Community	Enterprise	Enterprise	Enterprise

Overview of a DM Project



„Standard“ modeling using Rapid Miner



//Samples/processes/03_Validation/04_XValidation_Nominal* – RapidMiner Studio Educational 9.4.001 @ VSE-01409

k-NN

k: 5

☑ weighted vote

measure types: MixedMeasures

mixed measure: MixedEuclidea...

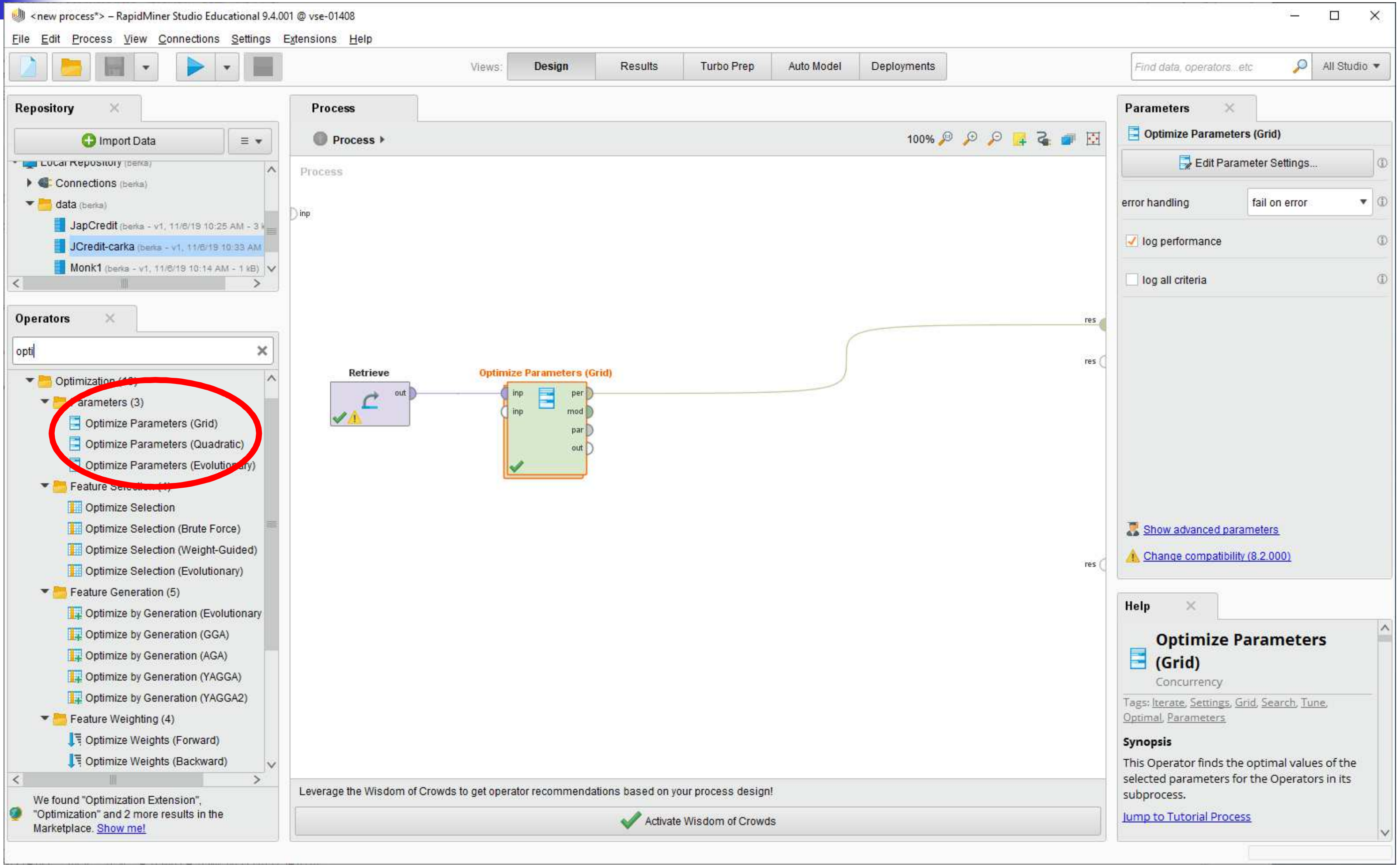
accuracy: 69.55% +/- 8.65% (micro average: 69.60%)

	true +	true -	class precision
pred. +	72	25	74.23%
pred. -	13	15	53.57%
class recall	84.71%	37.50%	

This Operator generates a k-Nearest Neig is used for classification or regression.

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Automatization no. 1: Parameter Optimization



<new process*> – RapidMiner Studio Educational 9.4.001 @ vse-01408
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
Local Repository (berka)
Connections (berka)
data (berka)
JapCredit (berka - v1, 11/6/19 10:25 AM - 3
JCredit-carka (berka - v1, 11/6/19 10:33 AM
Monk1 (berka - v1, 11/6/19 10:14 AM - 1 kB)
Operators
opti
Optimization (15)
Parameters (3)
Optimize Parameters (Grid)
Optimize Parameters (Quadratic)
Optimize Parameters (Evolutionary)
Feature Selection (4)
Optimize Selection
Optimize Selection (Brute Force)
Optimize Selection (Weight-Guided)
Optimize Selection (Evolutionary)
Feature Generation (5)
Optimize by Generation (Evolutionary
Optimize by Generation (GGA)
Optimize by Generation (AGA)
Optimize by Generation (YAGGA)
Optimize by Generation (YAGGA2)
Feature Weighting (4)
Optimize Weights (Forward)
Optimize Weights (Backward)
Process
Process ▸
100%
Process
inp
Retrieve
out
Optimize Parameters (Grid)
inp
inp
per
mod
par
out
res
res
res
Parameters
Optimize Parameters (Grid)
Edit Parameter Settings...
Show advanced parameters
Change compatibility (8.2.000)
Help
Activate Wisdom of Crowds

error handling: fail on error

☑ log performance

□ log all criteria

Optimize Parameters (Grid)

Concurrency

Tags: Iterate, Settings, Grid, Search, Tune, Optimal, Parameters

Synopsis

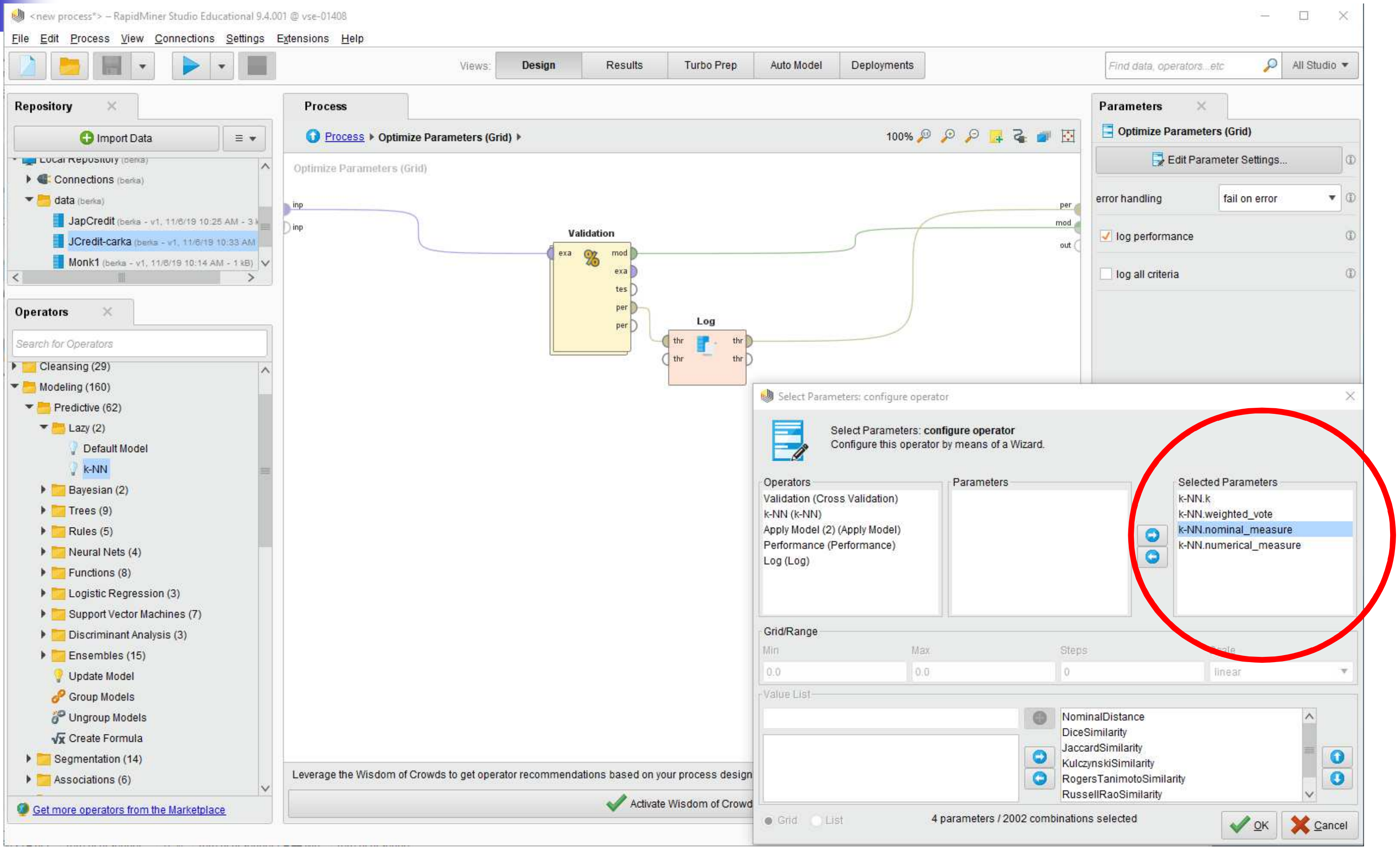
This Operator finds the optimal values of the selected parameters for the Operators in its subprocess.

Jump to Tutorial Process

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

We found "Optimization Extension", "Optimization" and 2 more results in the Marketplace. Show me!

Automatization no. 1: Parameter Optimization



<new process*> – RapidMiner Studio Educational 9.4.001 @ vse-01408

Process ▸ Optimize Parameters (Grid) ▸

Optimize Parameters (Grid)

error handling: fail on error

☑ log performance

☐ log all criteria

Leverage the Wisdom of Crowds to get operator recommendations based on your process design

Select Parameters: configure operator

Select Parameters: **configure operator**
Configure this operator by means of a Wizard.

Operators:
- Validation (Cross Validation)
- k-NN (k-NN)
- Apply Model (2) (Apply Model)
- Performance (Performance)
- Log (Log)

Selected Parameters:
- k-NN.k
- k-NN.weighted_vote
- k-NN.nominal_measure
- k-NN.numerical_measure

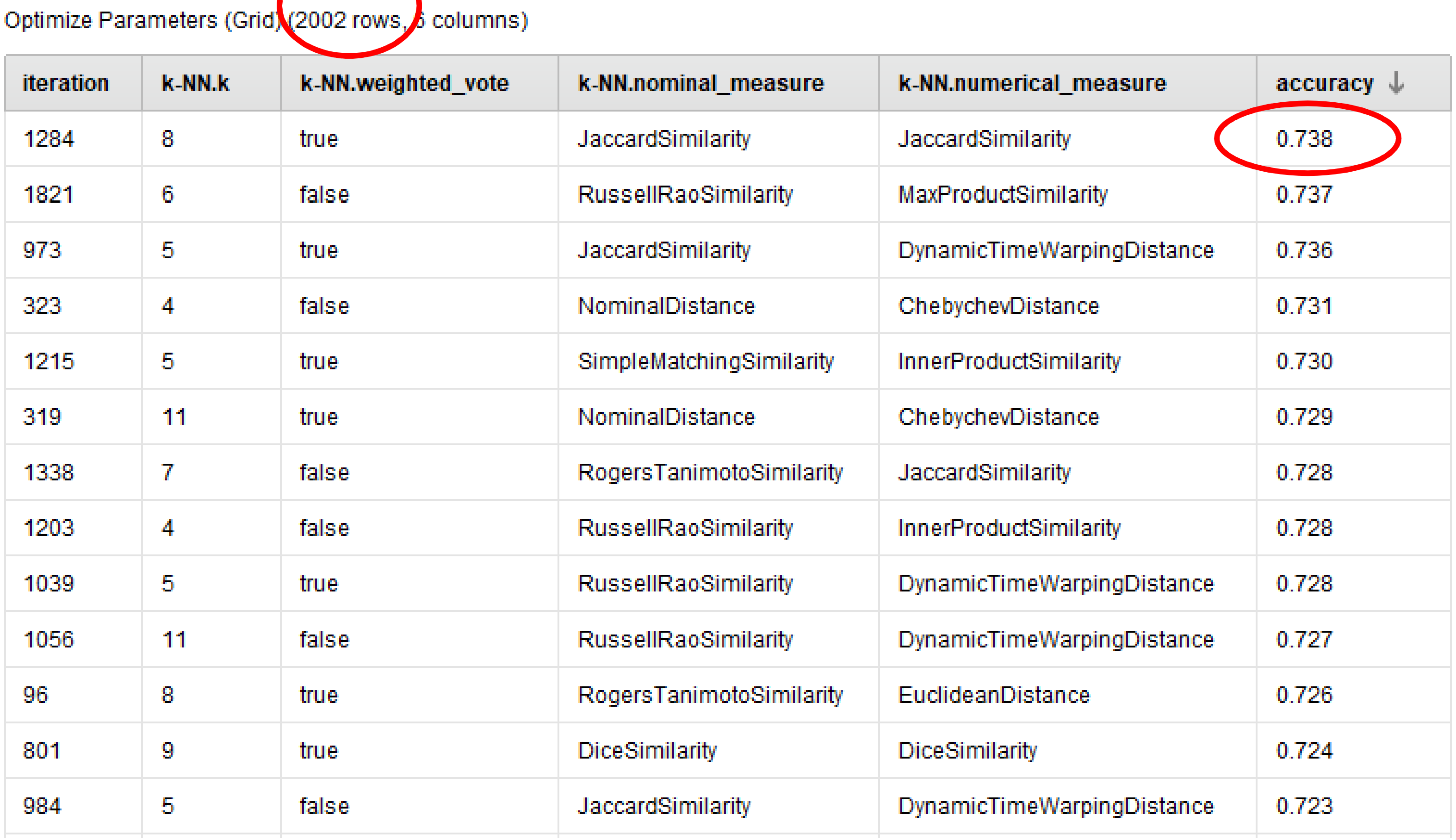
Grid/Range: Min 0.0, Max 0.0, Steps 0, Scale linear

Value List:
- NominalDistance
- DiceSimilarity
- JaccardSimilarity
- KulczynskiSimilarity
- RogersTanimotoSimilarity
- RussellRaoSimilarity

○ Grid ○ List

4 parameters / 2002 combinations selected

Automatization no. 1: Parameter Optimization



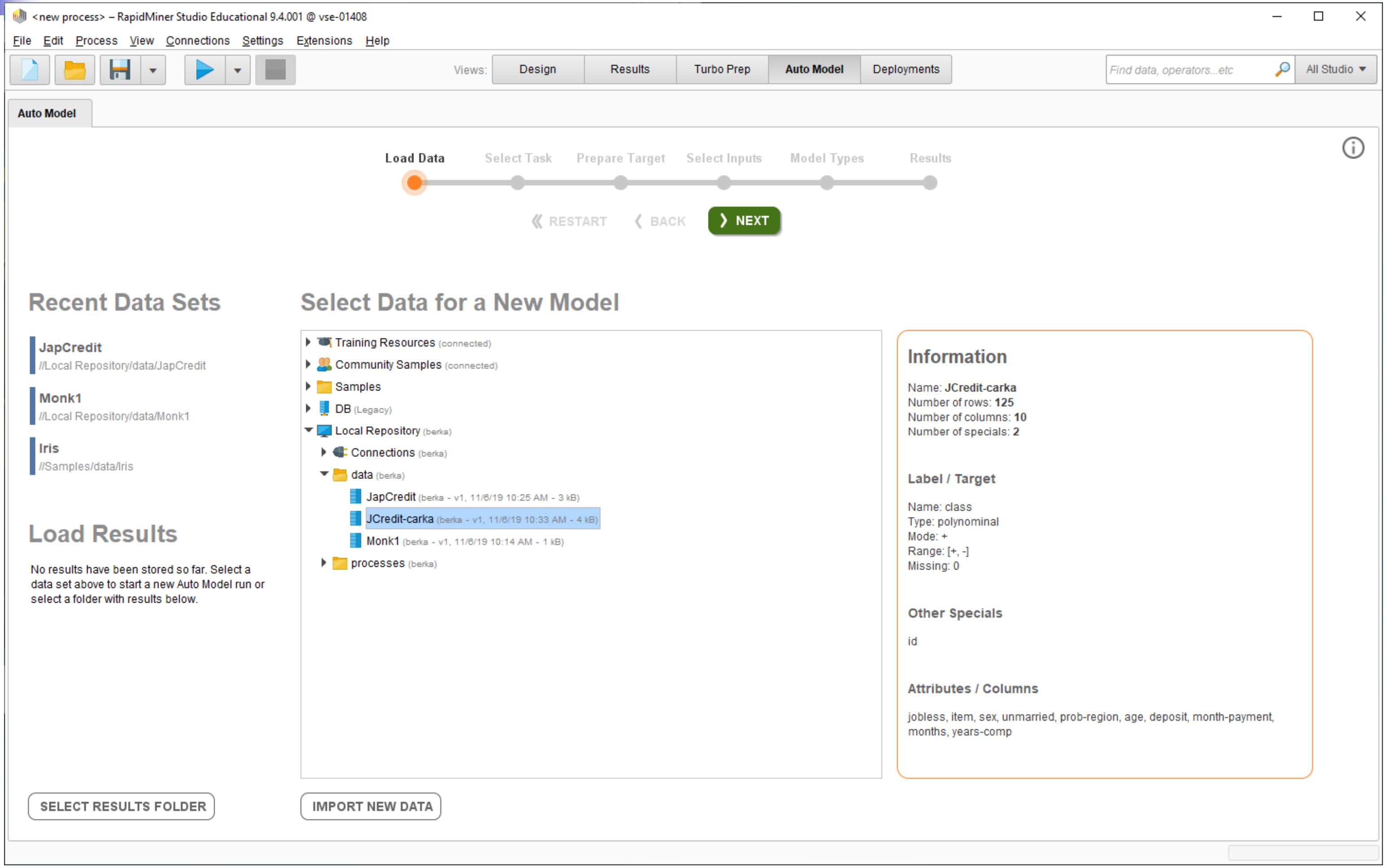
Optimize Parameters (Grid) (2002 rows, 6 columns)

iteration	k-NN.k	k-NN.weighted_vote	k-NN.nominal_measure	k-NN.numerical_measure	accuracy ↓
1284	8	true	JaccardSimilarity	JaccardSimilarity	0.738
1821	6	false	RussellRaoSimilarity	MaxProductSimilarity	0.737
973	5	true	JaccardSimilarity	DynamicTimeWarpingDistance	0.736
323	4	false	NominalDistance	ChebychevDistance	0.731
1215	5	true	SimpleMatchingSimilarity	InnerProductSimilarity	0.730
319	11	true	NominalDistance	ChebychevDistance	0.729
1338	7	false	RogersTanimotoSimilarity	JaccardSimilarity	0.728
1203	4	false	RussellRaoSimilarity	InnerProductSimilarity	0.728
1039	5	true	RussellRaoSimilarity	DynamicTimeWarpingDistance	0.728
1056	11	false	RussellRaoSimilarity	DynamicTimeWarpingDistance	0.727
96	8	true	RogersTanimotoSimilarity	EuclideanDistance	0.726
801	9	true	DiceSimilarity	DiceSimilarity	0.724
984	5	false	JaccardSimilarity	DynamicTimeWarpingDistance	0.723

Automatization no. 2: Auto Model

- **Auto Model** finds the best model using multiple machine learning algorithms and hyperparameter optimization

Auto Model step 1



<new process> – RapidMiner Studio Educational 9.4.001 @ vse-01408
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Auto Model
Load Data Select Task Prepare Target Select Inputs Model Types Results
RESTART BACK NEXT

Recent Data Sets

JapCredit
//Local Repository/data/JapCredit

Monk1
//Local Repository/data/Monk1

Iris
//Samples/data/Iris

Load Results

No results have been stored so far. Select a data set above to start a new Auto Model run or select a folder with results below.

SELECT RESULTS FOLDER

Select Data for a New Model

- Training Resources (connected)
- Community Samples (connected)
- Samples
- DB (Legacy)
- Local Repository (berka)
  - Connections (berka)
  - data (berka)
    - JapCredit (berka - v1, 11/6/19 10:25 AM - 3 kB)
    - JCredit-carka (berka - v1, 11/6/19 10:33 AM - 4 kB)
    - Monk1 (berka - v1, 11/6/19 10:14 AM - 1 kB)
  - processes (berka)

IMPORT NEW DATA

Information

Name: **JCredit-carka**
Number of rows: **125**
Number of columns: **10**
Number of specials: **2**

Label / Target

Name: class
Type: polynominal
Mode: +
Range: [+, -]
Missing: 0

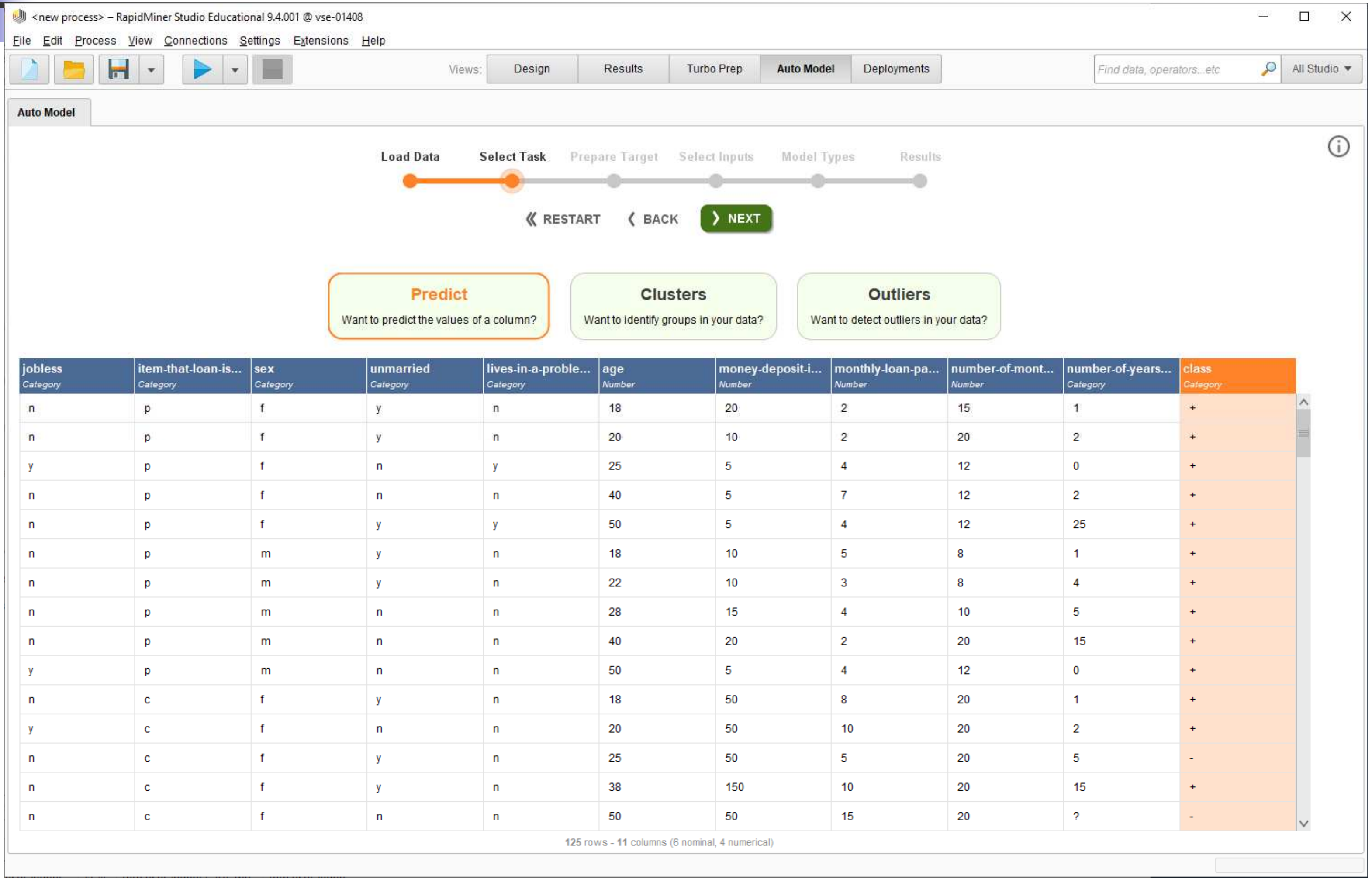
Other Specials

id

Attributes / Columns

jobless, item, sex, unmarried, prob-region, age, deposit, month-payment, months, years-comp

Auto Model step 2


<new process> – RapidMiner Studio Educational 9.4.001 @ vse-01408
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Auto Model
Load Data
Select Task
Prepare Target
Select Inputs
Model Types
Results
RESTART
BACK
NEXT

Predict
Want to predict the values of a column?

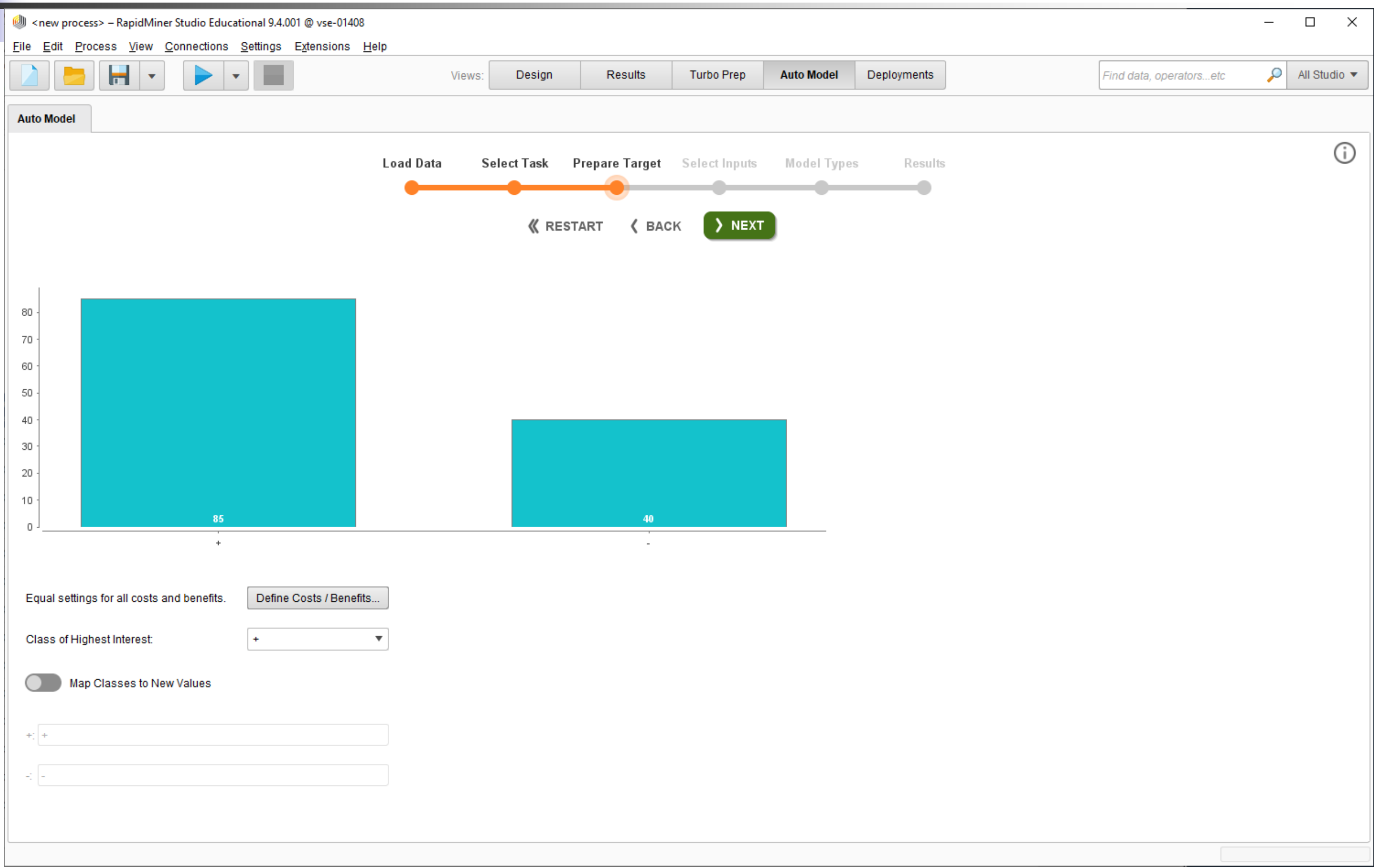
Clusters
Want to identify groups in your data?

Outliers
Want to detect outliers in your data?

jobless	item-that-loan-is...	sex	unmarried	lives-in-a-proble...	age	money-deposit-i...	monthly-loan-pa...	number-of-mont...	number-of-years...	class
Category	*Category*	*Category*	*Category*	*Category*	*Number*	*Number*	*Number*	*Number*	*Category*	*Category*
n	p	f	y	n	18	20	2	15	1	+
n	p	f	y	n	20	10	2	20	2	+
y	p	f	n	y	25	5	4	12	0	+
n	p	f	n	n	40	5	7	12	2	+
n	p	f	y	y	50	5	4	12	25	+
n	p	m	y	n	18	10	5	8	1	+
n	p	m	y	n	22	10	3	8	4	+
n	p	m	n	n	28	15	4	10	5	+
n	p	m	n	n	40	20	2	20	15	+
y	p	m	n	n	50	5	4	12	0	+
n	c	f	y	n	18	50	8	20	1	+
y	c	f	n	n	20	50	10	20	2	+
n	c	f	y	n	25	50	5	20	5	-
n	c	f	y	n	38	150	10	20	15	+
n	c	f	n	n	50	50	15	20	?	-

125 rows - 11 columns (6 nominal, 4 numerical)

Auto Model step 3



<new process> – RapidMiner Studio Educational 9.4.001 @ vse-01408
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Auto Model
Load Data
Select Task
Prepare Target
Select Inputs
Model Types
Results
RESTART
BACK
NEXT
80
70
60
50
40
30
20
10
0
85
40
+
-

Equal settings for all costs and benefits. Define Costs / Benefits...

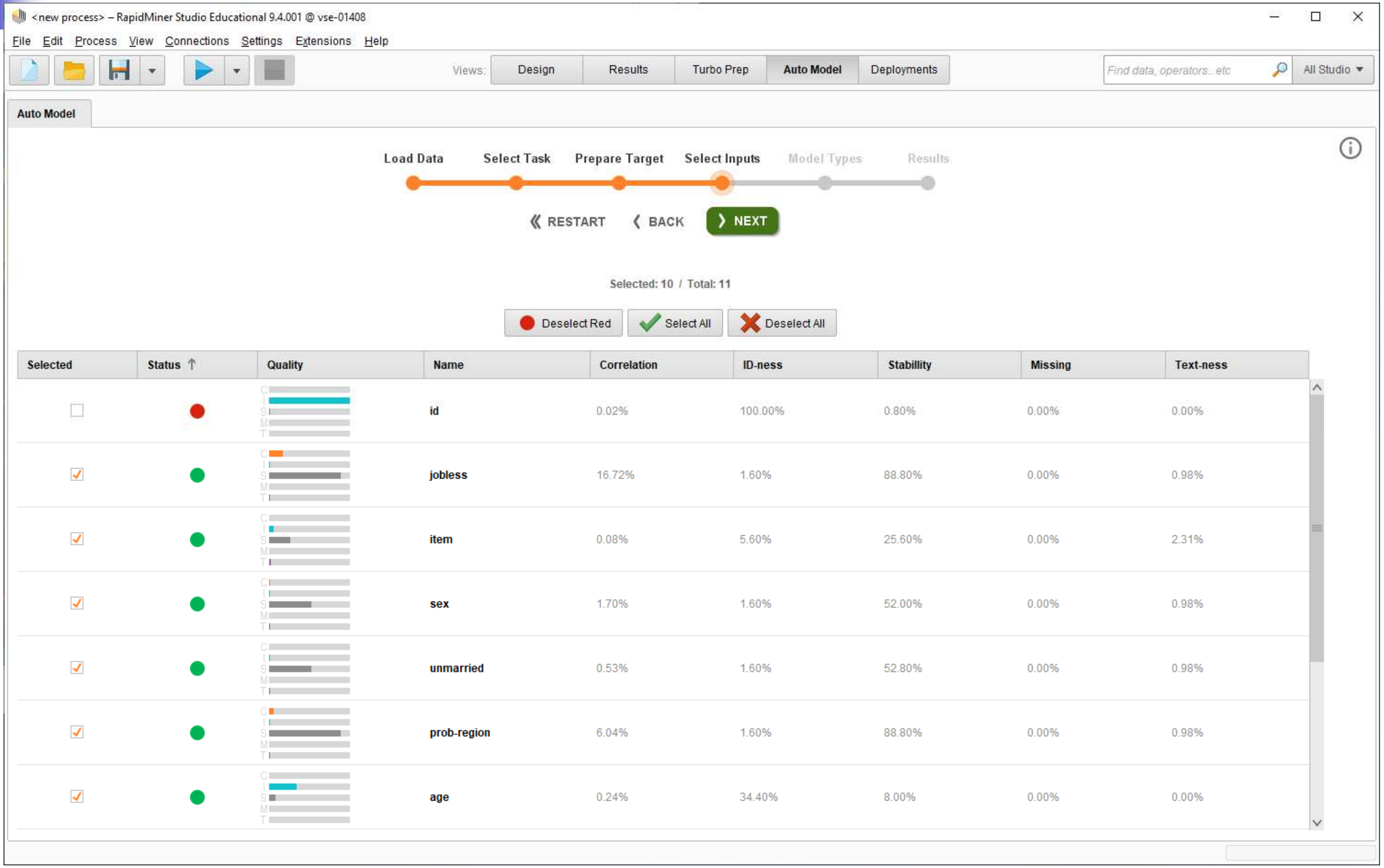
Class of Highest Interest: +

Map Classes to New Values

+: +

-: -

Auto Model step 4

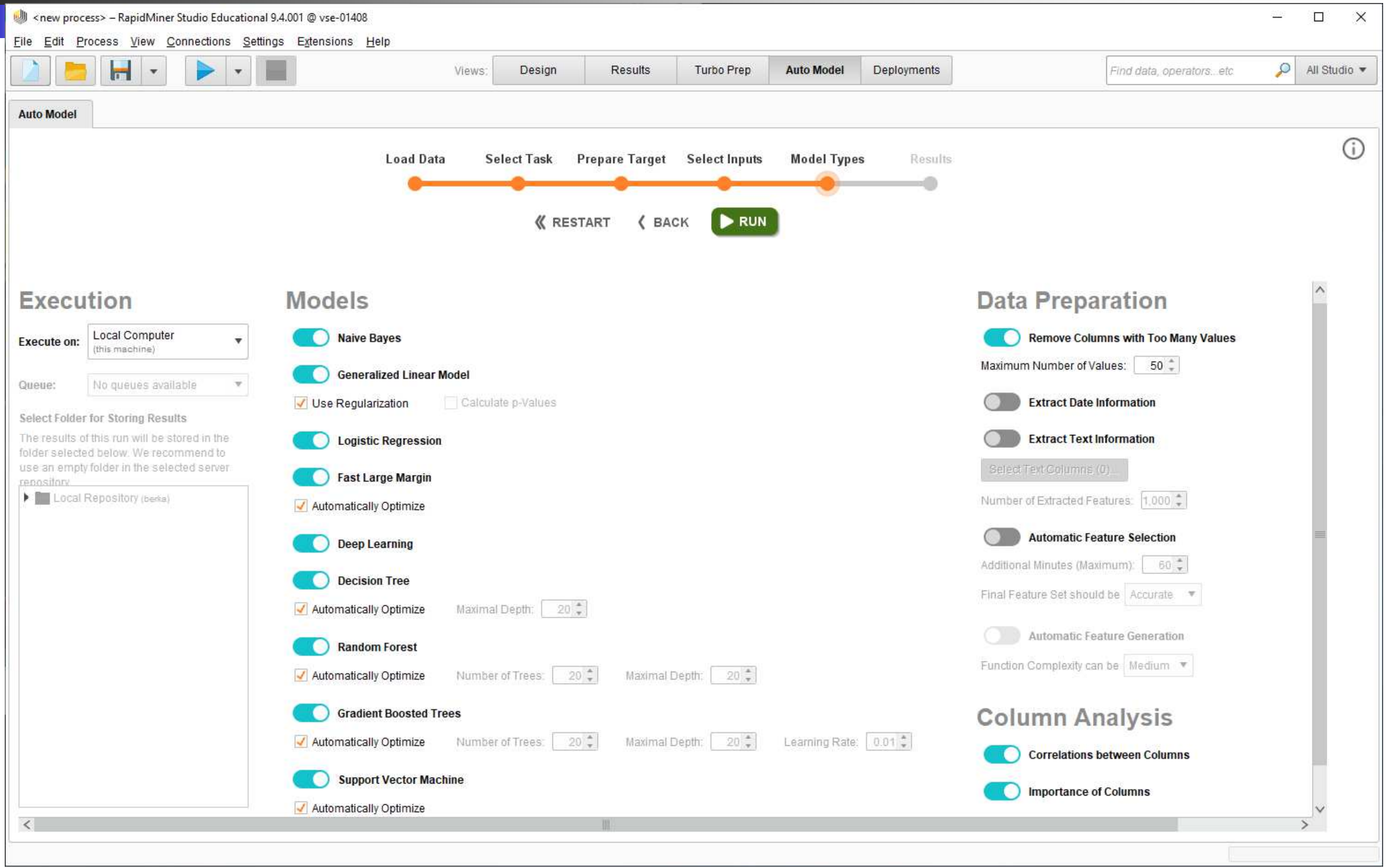


<new process> – RapidMiner Studio Educational 9.4.001 @ vse-01408
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Auto Model
Load Data
Select Task
Prepare Target
Select Inputs
Model Types
Results
RESTART
BACK
NEXT
Deselect Red
Select All
Deselect All

Selected: 10 / Total: 11

Selected	Status	Quality	Name	Correlation	ID-ness	Stabillity	Missing	Text-ness
□	●		id	0.02%	100.00%	0.80%	0.00%	0.00%
☑	●		jobless	16.72%	1.60%	88.80%	0.00%	0.98%
☑	●		item	0.08%	5.60%	25.60%	0.00%	2.31%
☑	●		sex	1.70%	1.60%	52.00%	0.00%	0.98%
☑	●		unmarried	0.53%	1.60%	52.80%	0.00%	0.98%
☑	●		prob-region	6.04%	1.60%	88.80%	0.00%	0.98%
☑	●		age	0.24%	34.40%	8.00%	0.00%	0.00%

Auto Model step 5


<new process> – RapidMiner Studio Educational 9.4.001 @ vse-01408
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Auto Model
Load Data Select Task Prepare Target Select Inputs Model Types Results
RESTART BACK RUN

Execution

Execute on: Local Computer (this machine)

Queue: No queues available

Select Folder for Storing Results

The results of this run will be stored in the folder selected below. We recommend to use an empty folder in the selected server repository.

Local Repository (berka)

Models

- Naive Bayes: on
- Generalized Linear Model: on
  - ☑ Use Regularization ☐ Calculate p-Values
- Logistic Regression: on
- Fast Large Margin: on
  - ☑ Automatically Optimize
- Deep Learning: on
- Decision Tree: on
  - ☑ Automatically Optimize Maximal Depth: 20
- Random Forest: on
  - ☑ Automatically Optimize Number of Trees: 20 Maximal Depth: 20
- Gradient Boosted Trees: on
  - ☑ Automatically Optimize Number of Trees: 20 Maximal Depth: 20 Learning Rate: 0.01
- Support Vector Machine: on
  - ☑ Automatically Optimize

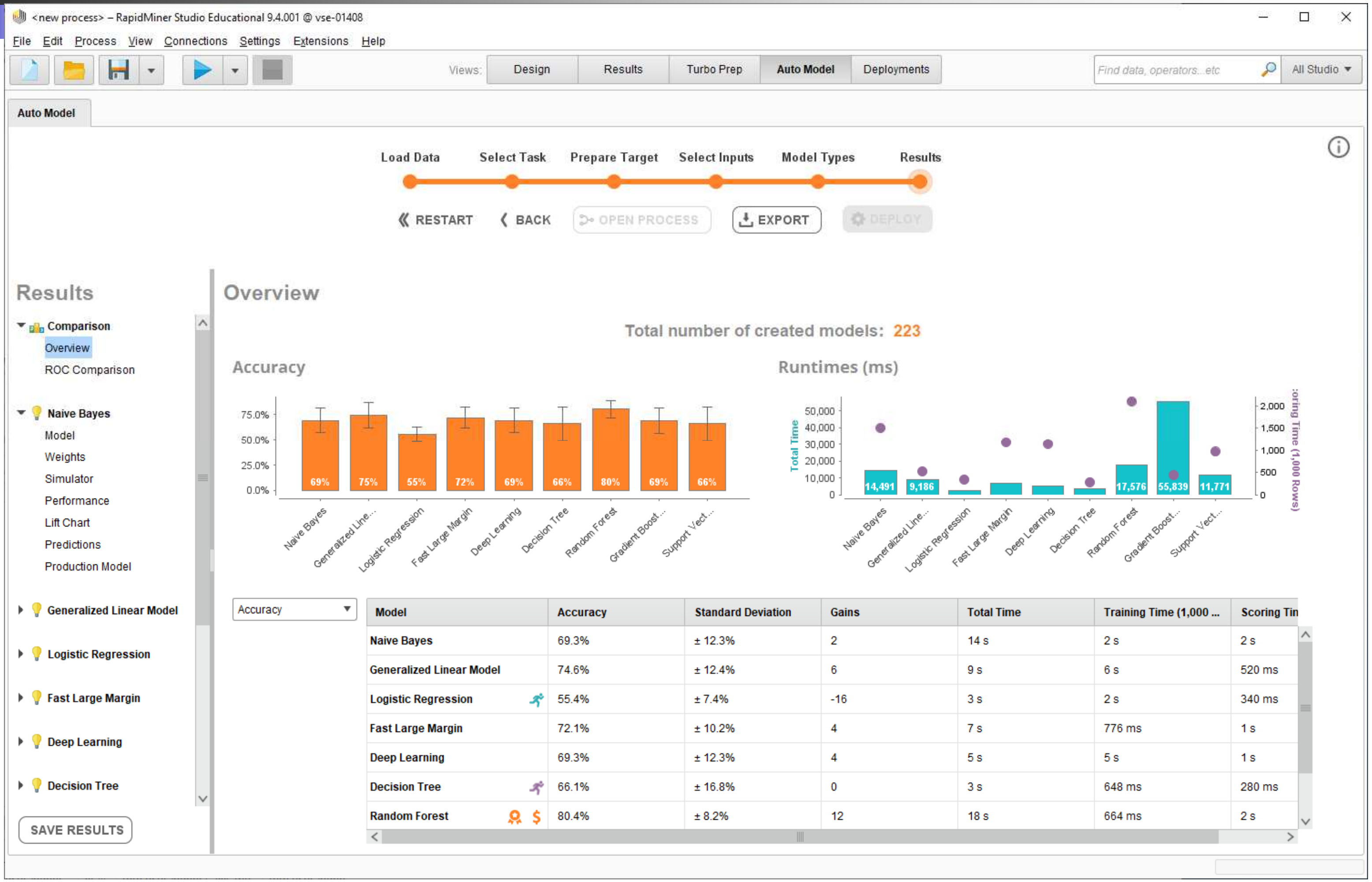
Data Preparation

- Remove Columns with Too Many Values: on
  - Maximum Number of Values: 50
- Extract Date Information: off
- Extract Text Information: off
  - Select Text Columns (0)...
  - Number of Extracted Features: 1,000
- Automatic Feature Selection: off
  - Additional Minutes (Maximum): 60
  - Final Feature Set should be Accurate
- Automatic Feature Generation: off
  - Function Complexity can be Medium

Column Analysis

- Correlations between Columns: on
- Importance of Columns: on

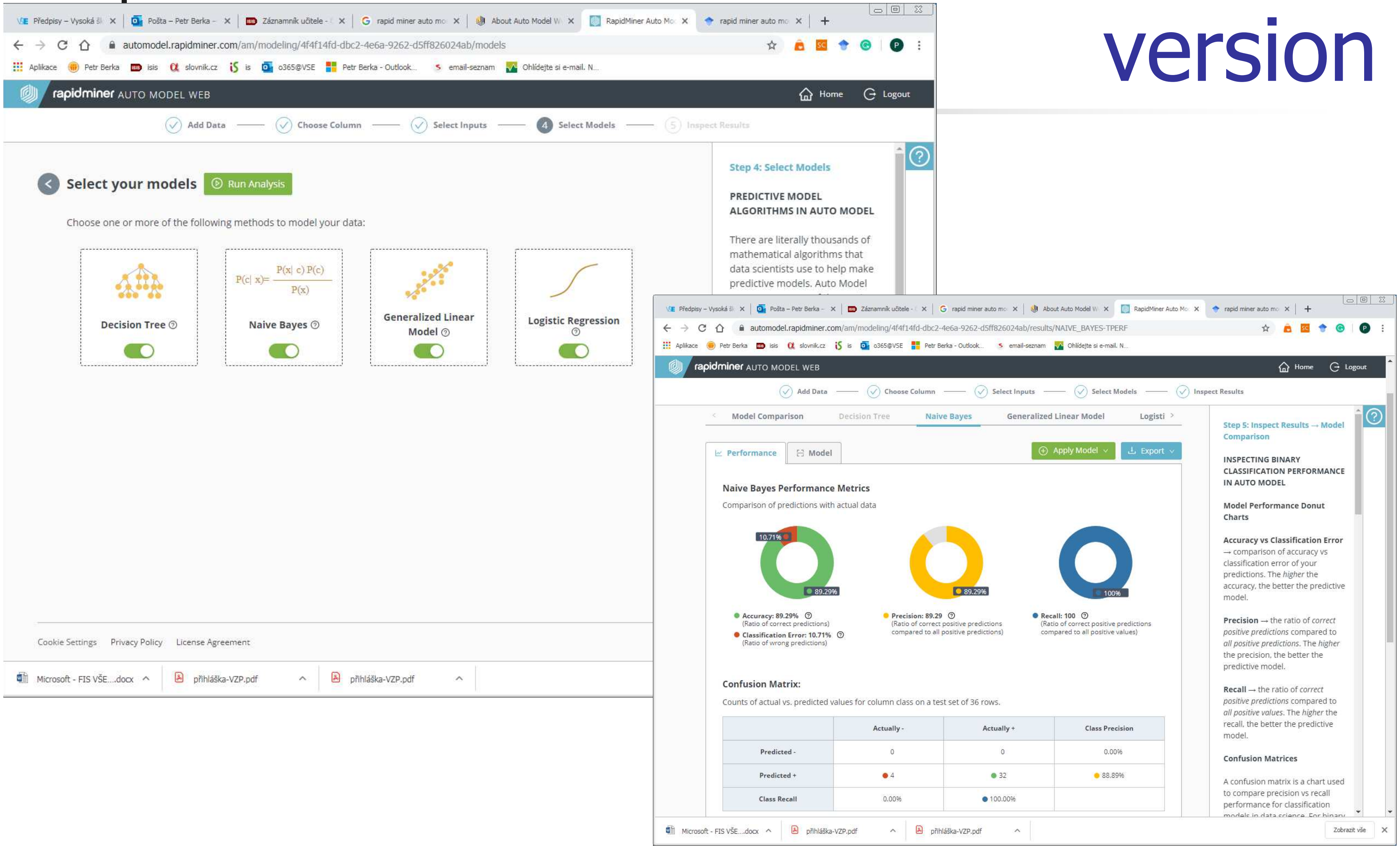
Auto Model step 6



Total number of created models: 223

Model	Accuracy	Standard Deviation	Gains	Total Time	Training Time (1,000 ...	Scoring Tin
Naive Bayes	69.3%	± 12.3%	2	14 s	2 s	2 s
Generalized Linear Model	74.6%	± 12.4%	6	9 s	6 s	520 ms
Logistic Regression	55.4%	± 7.4%	-16	3 s	2 s	340 ms
Fast Large Margin	72.1%	± 10.2%	4	7 s	776 ms	1 s
Deep Learning	69.3%	± 12.3%	4	5 s	5 s	1 s
Decision Tree	66.1%	± 16.8%	0	3 s	648 ms	280 ms
Random Forest $	80.4%	± 8.2%	12	18 s	664 ms	2 s

Auto Model (web/cloud) version



automodel.rapidminer.com/am/modeling/4f4f14fd-dbc2-4e6a-9262-d5ff826024ab/models
rapidminer AUTO MODEL WEB
Home
Logout
Add Data
Choose Column
Select Inputs
4 Select Models
5 Inspect Results

Select your models

Run Analysis

Choose one or more of the following methods to model your data:

Step 4: Select Models

PREDICTIVE MODEL ALGORITHMS IN AUTO MODEL

There are literally thousands of mathematical algorithms that data scientists use to help make predictive models. Auto Model

Naive Bayes Performance Metrics

Comparison of predictions with actual data

● Accuracy: 89.29% (Ratio of correct predictions)

● Classification Error: 10.71% (Ratio of wrong predictions)

● Precision: 89.29 (Ratio of correct positive predictions compared to all positive predictions)

● Recall: 100 (Ratio of correct positive predictions compared to all positive values)

Confusion Matrix:

Counts of actual vs. predicted values for column class on a test set of 36 rows.

	Actually -	Actually +	Class Precision
Predicted -	0	0	0.00%
Predicted +	● 4	● 32	● 88.89%
Class Recall	0.00%	● 100.00%	

Step 5: Inspect Results → Model Comparison

INSPECTING BINARY CLASSIFICATION PERFORMANCE IN AUTO MODEL

Model Performance Donut Charts

Accuracy vs Classification Error → comparison of accuracy vs classification error of your predictions. The *higher* the accuracy, the better the predictive model.

Precision → the ratio of *correct positive predictions* compared to *all positive predictions*. The *higher* the precision, the better the predictive model.

Recall → the ratio of *correct positive predictions* compared to *all positive values*. The *higher* the recall, the better the predictive model.

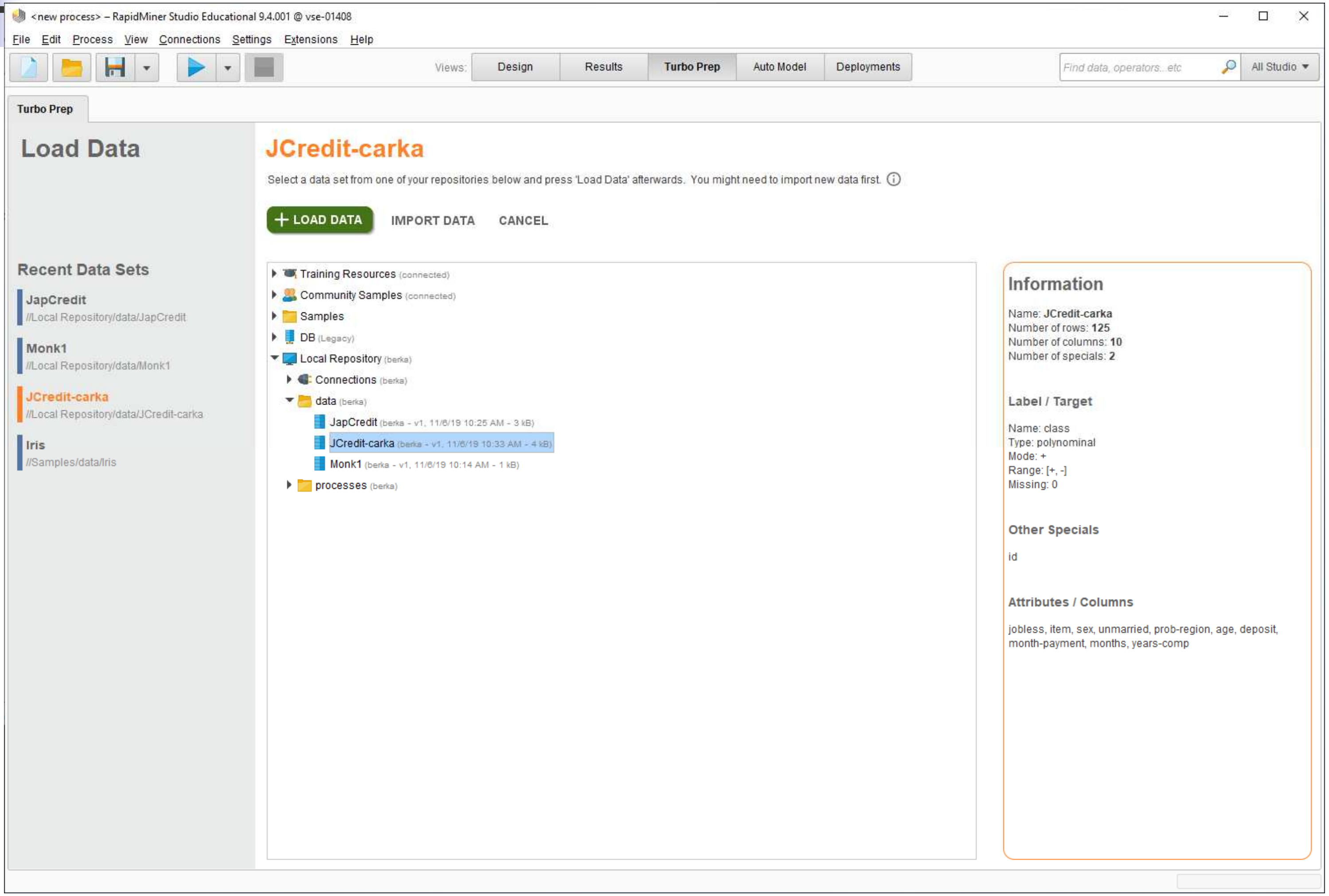
Confusion Matrices

A confusion matrix is a chart used to compare precision vs recall performance for classification models in data science. For binary

Automatization no. 3: Turbo Prep

- **Turbo Prep** aims at intuitive data preparation. This extension allows to interactively explore and visualize the data, simplifies data cleansing (automatically removes low quality and correlated data columns) and merges multiple datasets together by automatically identifying matching columns to merge.

Turbo Prep step 1



<new process> – RapidMiner Studio Educational 9.4.001 @ vse-01408

File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Turbo Prep

Load Data

Recent Data Sets

JapCredit
//Local Repository/data/JapCredit

Monk1
//Local Repository/data/Monk1

JCredit-carka
//Local Repository/data/JCredit-carka

Iris
//Samples/data/Iris

JCredit-carka

Select a data set from one of your repositories below and press 'Load Data' afterwards. You might need to import new data first.

- Training Resources (connected)
- Community Samples (connected)
- Samples
- DB (Legacy)
- Local Repository (berka)
  - Connections (berka)
  - data (berka)
    - JapCredit (berka - v1, 11/6/19 10:25 AM - 3 kB)
    - JCredit-carka (berka - v1, 11/6/19 10:33 AM - 4 kB)
    - Monk1 (berka - v1, 11/6/19 10:14 AM - 1 kB)
  - processes (berka)

Information

Name: **JCredit-carka**
Number of rows: **125**
Number of columns: **10**
Number of specials: **2**

Label / Target

Name: class
Type: polynominal
Mode: +
Range: [+, -]
Missing: 0

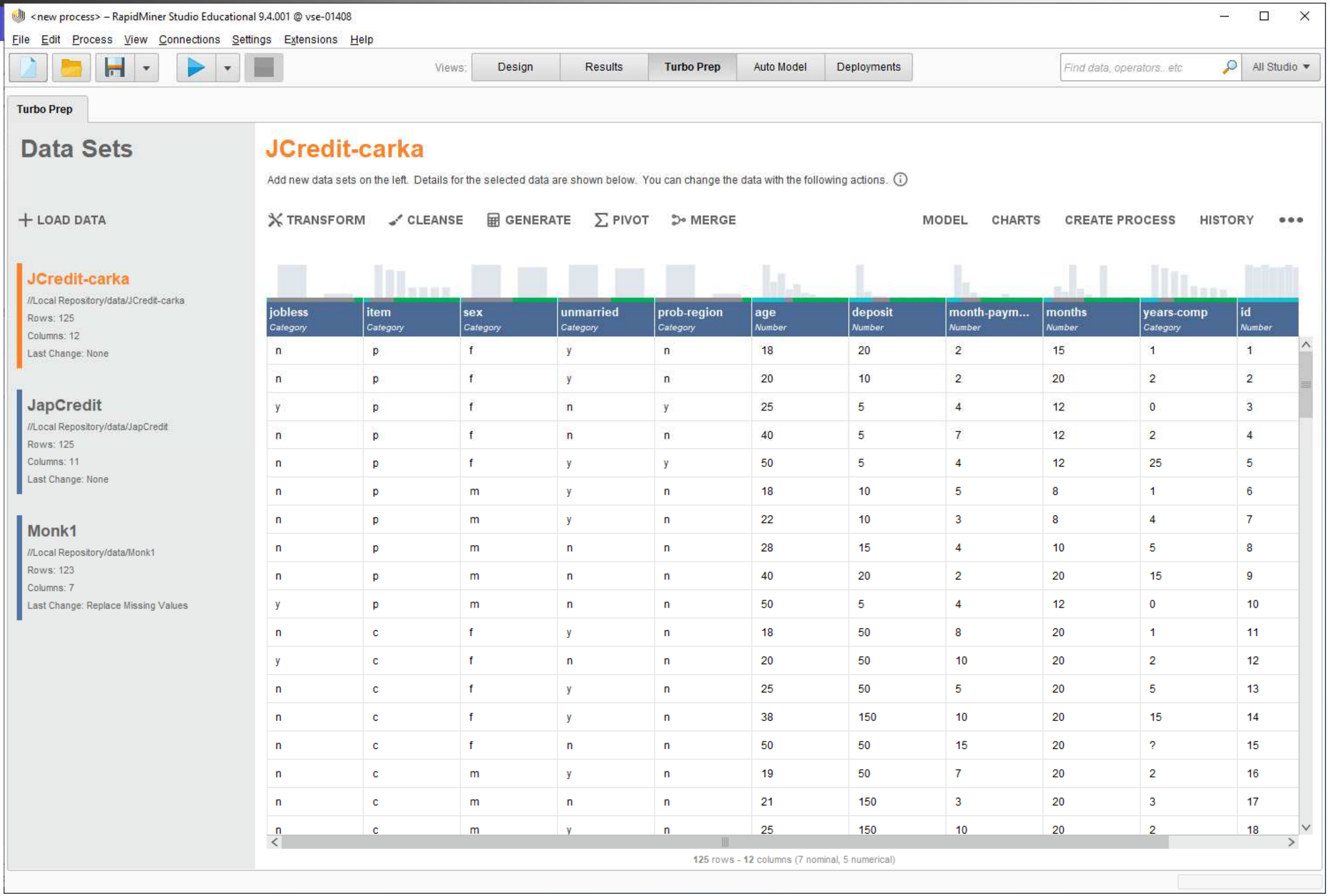
Other Specials

id

Attributes / Columns

jobless, item, sex, unmarried, prob-region, age, deposit, month-payment, months, years-comp

Turbo Prep step 2



<new process> – RapidMiner Studio Educational 9.4.001 @ vse-01408
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Turbo Prep

Data Sets

+ LOAD DATA

JCredit-carka
//Local Repository/data/JCredit-carka
Rows: 125
Columns: 12
Last Change: None

JapCredit
//Local Repository/data/JapCredit
Rows: 125
Columns: 11
Last Change: None

Monk1
//Local Repository/data/Monk1
Rows: 123
Columns: 7
Last Change: Replace Missing Values

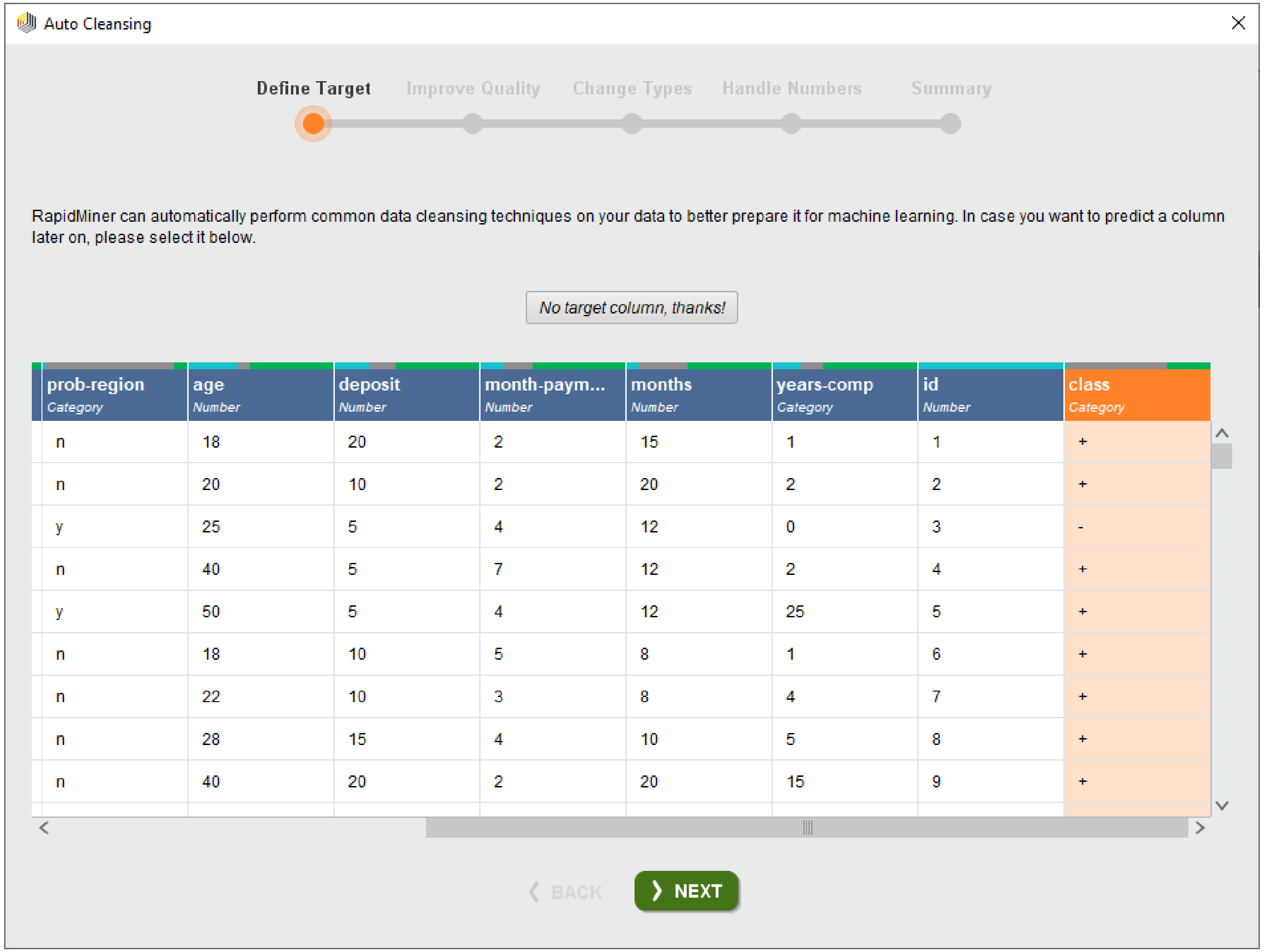
JCredit-carka

Add new data sets on the left. Details for the selected data are shown below. You can change the data with the following actions.

jobless	item	sex	unmarried	prob-region	age	deposit	month-paym...	months	years-comp	id
Category	Category	Category	Category	Category	Number	Number	Number	Number	Category	Number
n	p	f	y	n	18	20	2	15	1	1
n	p	f	y	n	20	10	2	20	2	2
y	p	f	n	y	25	5	4	12	0	3
n	p	f	n	n	40	5	7	12	2	4
n	p	f	y	y	50	5	4	12	25	5
n	p	m	y	n	18	10	5	8	1	6
n	p	m	y	n	22	10	3	8	4	7
n	p	m	n	n	28	15	4	10	5	8
n	p	m	n	n	40	20	2	20	15	9
y	p	m	n	n	50	5	4	12	0	10
n	c	f	y	n	18	50	8	20	1	11
y	c	f	n	n	20	50	10	20	2	12
n	c	f	y	n	25	50	5	20	5	13
n	c	f	y	n	38	150	10	20	15	14
n	c	f	n	n	50	50	15	20	?	15
n	c	m	y	n	19	50	7	20	2	16
n	c	m	n	n	21	150	3	20	3	17
n	c	m	y	n	25	150	10	20	2	18

125 rows - 12 columns (7 nominal, 5 numerical)

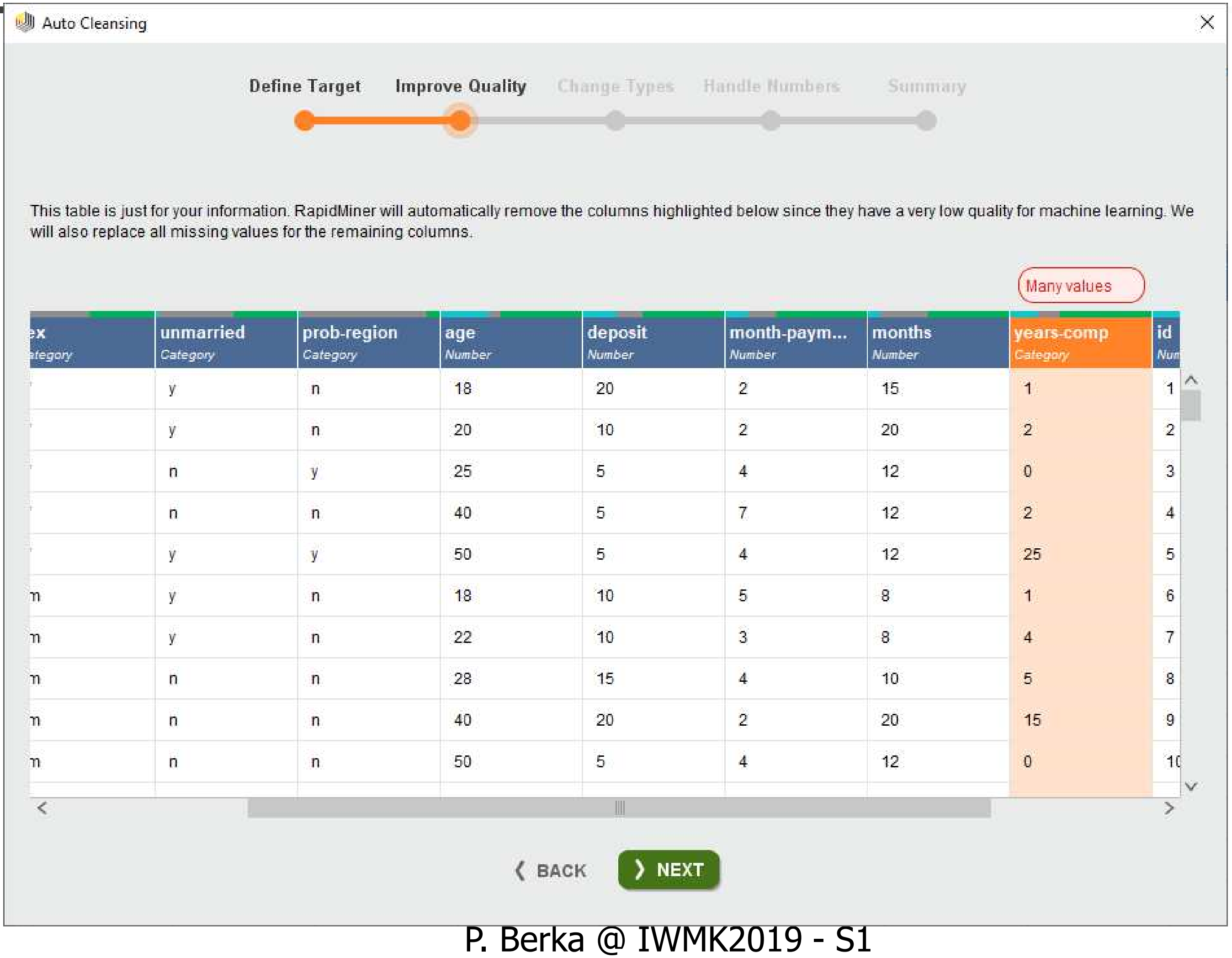
Turbo Prep step 3



RapidMiner can automatically perform common data cleansing techniques on your data to better prepare it for machine learning. In case you want to predict a column later on, please select it below.

prob-region Category	age Number	deposit Number	month-paym... Number	months Number	years-comp Category	id Number	class Category
n	18	20	2	15	1	1	+
n	20	10	2	20	2	2	+
y	25	5	4	12	0	3	-
n	40	5	7	12	2	4	+
y	50	5	4	12	25	5	+
n	18	10	5	8	1	6	+
n	22	10	3	8	4	7	+
n	28	15	4	10	5	8	+
n	40	20	2	20	15	9	+

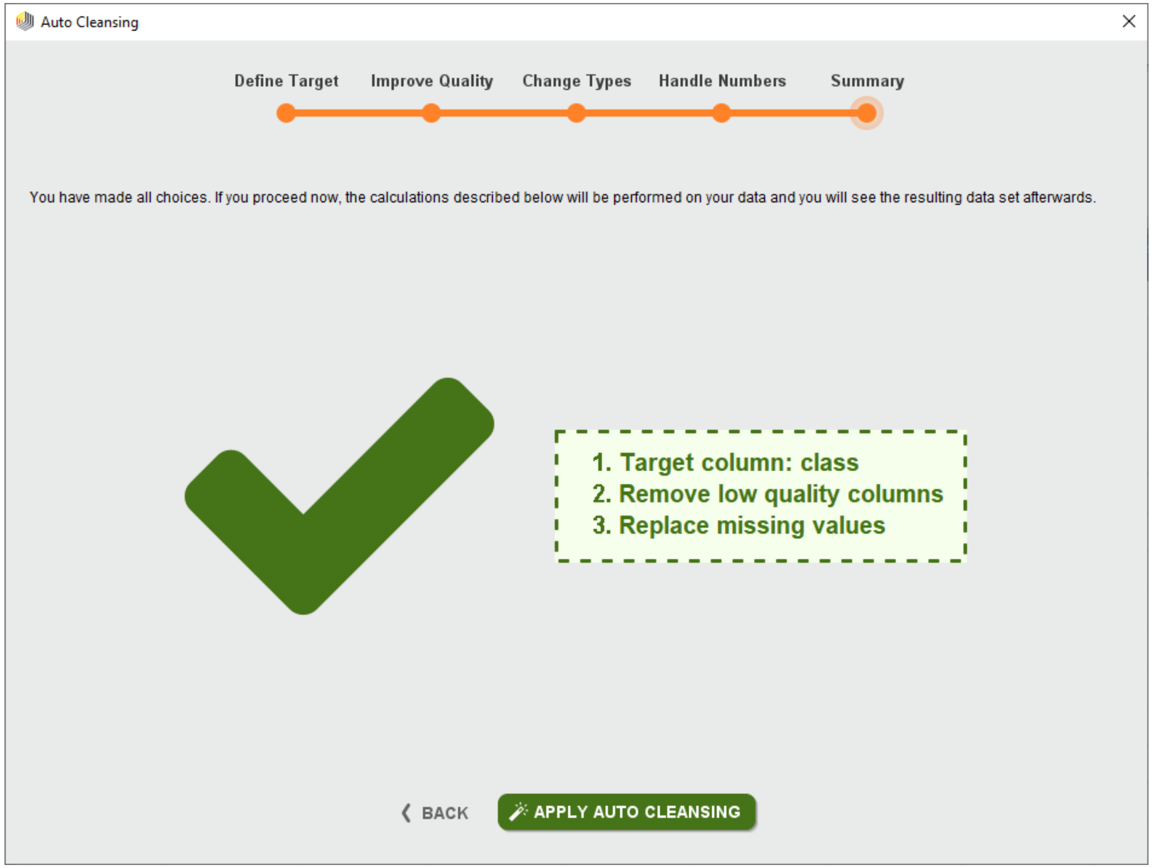
Turbo Prep step 4



This table is just for your information. RapidMiner will automatically remove the columns highlighted below since they have a very low quality for machine learning. We will also replace all missing values for the remaining columns.

ex ategory	unmarried Category	prob-region Category	age Number	deposit Number	month-paym... Number	months Number	years-comp Category	id Num
	y	n	18	20	2	15	1	1
	y	n	20	10	2	20	2	2
	n	y	25	5	4	12	0	3
	n	n	40	5	7	12	2	4
	y	y	50	5	4	12	25	5
m	y	n	18	10	5	8	1	6
m	y	n	22	10	3	8	4	7
m	n	n	28	15	4	10	5	8
m	n	n	40	20	2	20	15	9
m	n	n	50	5	4	12	0	10

Turbo Prep step 5



You have made all choices. If you proceed now, the calculations described below will be performed on your data and you will see the resulting data set afterwards.

1. Target column: class
2. Remove low quality columns
3. Replace missing values

Automatization of analyses in Rapid Miner

- Optimized parameters node always available
- Auto Model and Turbo Prep available in paid version and within free academic license

Resources

1. Get Started with Fully Automated Data Science, 2019. *Rapid Miner.* [online] Available at: <httphttps://rapidminer.com/products/#automated> [Accessed 1 October 2019].
2. HOFMANN, M. and KLINKENBERG, R., 2013. RapidMiner: Data Mining Use Cases and Business Analytics Applications. Chapman and Hall/CRC.