Data Mining

Decision Trees, Clustering, Association Rule

Harry Beyel

BPI-Instruction2

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Pen-and-paper Exercises

Chair of Process
and Data Science

Decision Tree
Exercise 1

- **Consider the following training dataset. Suppose *Outcome* is our response variable.**
- **Provide a decision tree that predicts the outcome. Use the information gain criterion for splitting.**

District	Income	Previous Customer	Outcome
Suburban	High	No	Not responded
Suburban	High	Yes	Not responded
Rural	High	No	Responded
Urban	High	No	Responded
Urban	Low	No	Responded
Urban	Low	Yes	Not responded
Rural	Low	Yes	Responded
Suburban	High	Yes	Not responded
Suburban	Low	No	Responded
Urban	Low	No	Responded
Suburban	Low	Yes	Responded
Rural	High	Yes	Responded
Rural	Low	No	Responded
Urban	High	Yes	Not responded

Chair of Process and Data Science

Decision Tree
Solution 1

Compute the entropy for the whole data

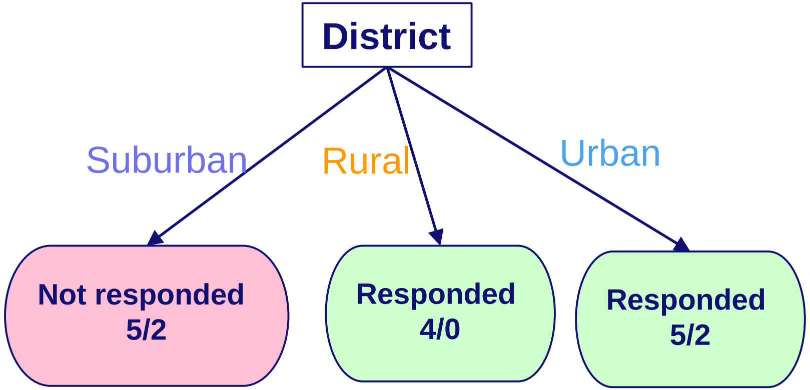
$$E = -\sum_{i=1}^{2} p_i log_2(pi) = -\frac{9}{14} log_2\left(\frac{9}{14}\right) - \frac{5}{14} log_2\left(\frac{5}{14}\right)$$

$$= 0.41 + 0.53 = 0.94$$

District	Income	Previous Customer	Outcome
Suburban	High	No	Not responded
Suburban	High	Yes	Not responded
Rural	High	No	Responded
Urban	High	No	Responded
Urban	Low	No	Responded
Urban	Low	Yes	Not responded
Rural	Low	Yes	Responded
Suburban	High	Yes	Not responded
Suburban	Low	No	Responded
Urban	Low	No	Responded
Suburban	Low	Yes	Responded
Rural	High	Yes	Responded
Rural	Low	No	Responded
Urban	High	Yes	Not responded

Chair of Process and Data Science

Decision Tree
Solution 1



$$E_1 = -\sum_{i=1}^{2} p_i log_2(pi) = -\frac{2}{5} log_2\left(\frac{2}{5}\right) - \frac{3}{5} log_2\left(\frac{3}{5}\right) = 0.97$$

$$E_2 = 0$$

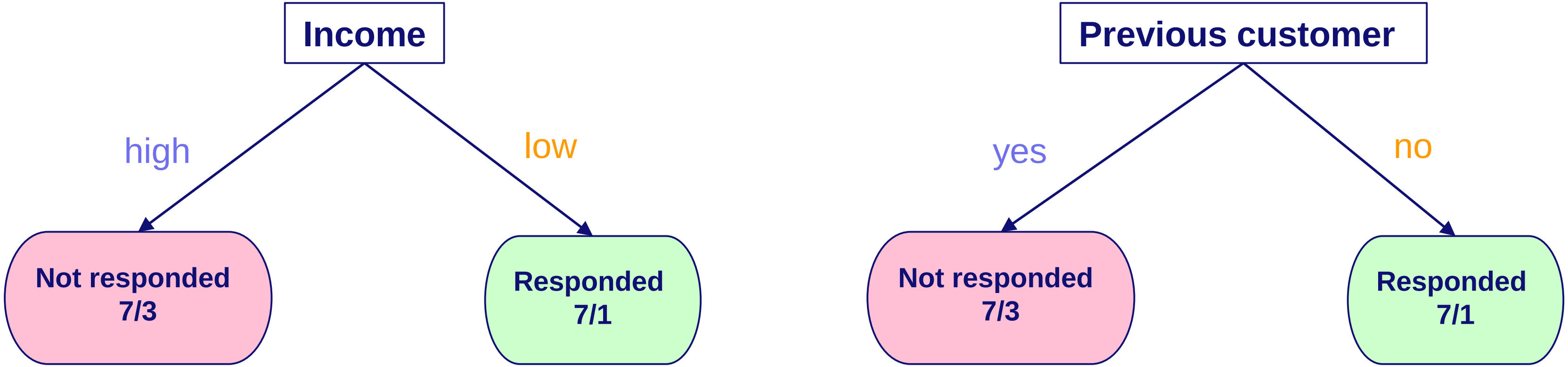
$$E_3 = -\sum_{i=1}^{2} p_i log_2(pi) = -\frac{3}{5} log_2\left(\frac{3}{5}\right) - \frac{2}{5} log_2\left(\frac{2}{5}\right) = 0.97$$

$$E_{District} = \frac{5}{14} \times 0.97 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.97 = 0.69$$

$$Information\ gain = 0.94 - 0.69 = 0.25$$

District	Income	Previous Customer	Outcome
Suburban	High	No	Not responded
Suburban	High	Yes	Not responded
Rural	High	No	Responded
Urban	High	No	Responded
Urban	Low	No	Responded
Urban	Low	Yes	Not responded
Rural	Low	Yes	Responded
Suburban	High	Yes	Not responded
Suburban	Low	No	Responded
Urban	Low	No	Responded
Suburban	Low	Yes	Responded
Rural	High	Yes	Responded
Rural	Low	No	Responded
Urban	High	Yes	Not responded

Decision Tree
Solution 1



Calculate the information gain for all the attributes:

$$E - E_{District} = 0.25$$
$$E - E_{Income} = 0.15$$
$$E - E_{Previous-Customer} = 0.15$$

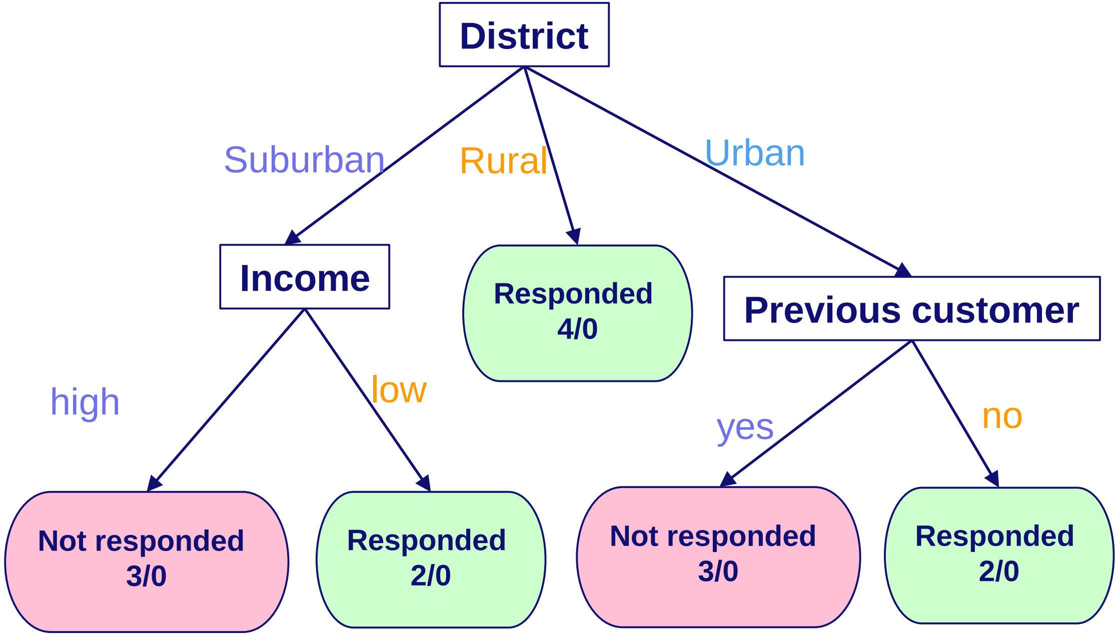
Which attribute should be selected? Why?

Decision Tree Solution 1

District	Income	Previous Customer	Outcome
Suburban	High	No	Not responded
Suburban	High	Yes	Not responded
Suburban	High	Yes	Not responded
Suburban	Low	No	Responded
Suburban	Low	Yes	Responded

District	Income	Previous Customer	Outcome
Rural	High	No	Responded
Rural	Low	Yes	Responded
Rural	High	Yes	Responded
Rural	Low	No	Responded

District	Income	Previous Customer	Outcome
Urban	High	No	Responded
Urban	Low	No	Responded
Urban	Low	Yes	Not responded
Urban	Low	No	Responded
Urban	High	Yes	Not responded



Clustering
Exercise 2

	att_1	att_2
1	2	10
2	2	5
3	8	4
4	5	8
5	7	5
6	6	4
7	1	2
8	4	9

Provide three clusters for the data instances on the left side. The distance function is Euclidean distance. Suppose initially we assign $(2, 10)$, $(5, 8)$, and $(1, 2)$ as the center of each cluster, respectively. Use the *K-means* algorithm to find the clusters.

Chair of Process and Data Science

Clustering Exercise 2

Algorithm: *k*-means. The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

- k: the number of clusters,
- D: a data set containing n objects.

Output: A set of k clusters.

Method:

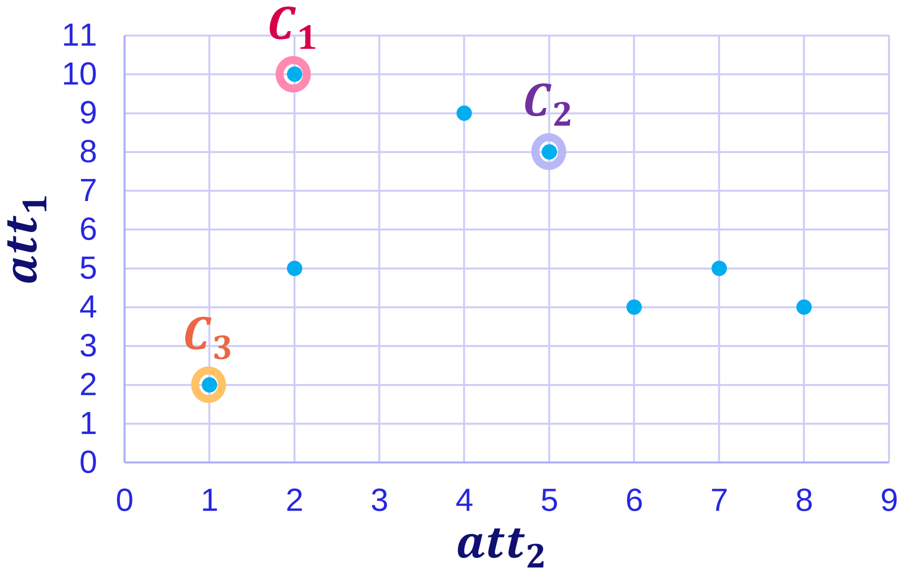
(1) arbitrarily choose k objects from D as the initial cluster centers;
(2) **repeat**
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, that is, calculate the mean value of the objects for each cluster;
(5) **until** no change;

$$d(i,j) = \sqrt{(x_{i1}-x_{j1})^2 + (x_{i2}-x_{j2})^2 + \cdots + (x_{ip}-x_{jp})^2}$$

$$d(i,j) = |x_{i1}-x_{j1}| + |x_{i2}-x_{j2}| + \cdots + |x_{ip}-x_{jp}||$$

Clustering
Solution 2

	att_1	att_2
1	2	10
2	2	5
3	8	4
4	5	8
5	7	5
6	6	4
7	1	2
8	4	9



Chair of Process and Data Science

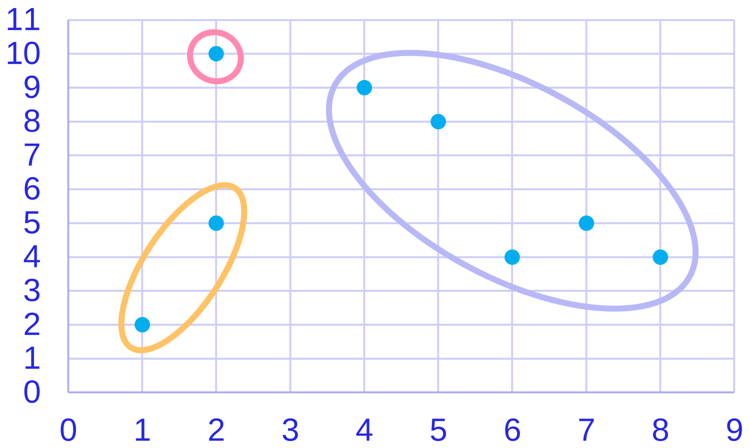
Clustering
Solution 2

	att_1	att_2	Distance from (2,10)	Distance from (5,8)	Distance from (1,2)
1	2	10	0		
2	2	5	5	4.24	3.16
3	8	4	8.49	5	7.28
4	5	8		0	
5	7	5	7.07	3.61	6.71
6	6	4	7.21	4.12	5.39
7	1	2			0
8	4	9	2.24	1.41	7.62

Assign instances to the nearest cluster.

Clustering
Solution 2

	att_1	att_2	Distance from (2,10)	Distance from (5,8)	Distance from (1,2)
1	2	10	0		
2	2	5	5	4.24	3.16
3	8	4	8.49	5	7.28
4	5	8		0	
5	7	5	7.07	3.61	6.71
6	6	4	7.21	4.12	5.39
7	1	2			0
8	4	9	2.24	1.41	7.62



Update cluster means.

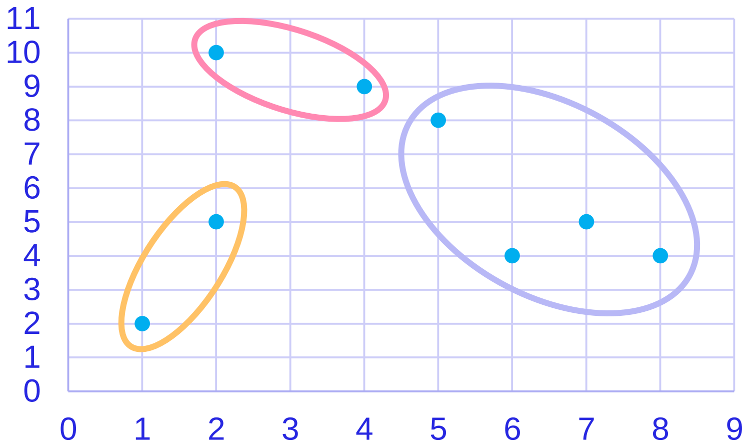
New centroids: $C_1 = (2, 10), C_2 = (6, 6), C_3 = (1.5, 3.5)$

Chair of Process and Data Science

Clustering
Solution 2

	att_1	att_2	Distance from (2,10)	Distance from (6,6)	Distance from (1.5,3.5)
1	2	10	0	5.66	6.52
2	2	5	5	4.12	1.58
3	8	4	8.49	2.83	6.52
4	5	8	3.61	2.24	5.7
5	7	5	7.07	1.41	5.7
6	6	4	7.21	2	4.53
7	1	2	8.06	6.40	1.58
8	4	9	2.24	3.61	6.04

Assign instances to the nearest cluster, update cluster means.



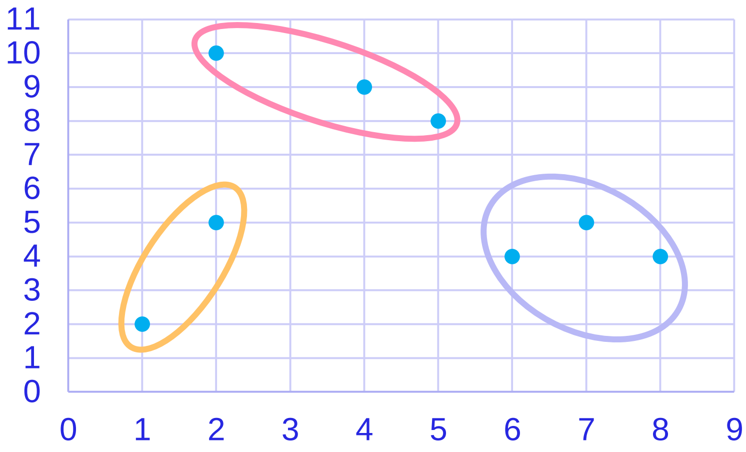
New centroids: $C_1 = (3, 9.5), C_2 = (6.5, 5.25), C_3 = (1.5, 3.5)$

Chair of Process and Data Science

Clustering
Solution 2

	att_1	att_2	Distance from (3,9.5)	Distance from (6.5,5.25)	Distance from (1.5,3.5)
1	2	10	1.11	6.54	6.52
2	2	5	4.61	4.51	1.58
3	8	4	7.43	1.95	6.52
4	5	8	2.5	3.13	5.7
5	7	5	6.02	0.56	5.7
6	6	4	6.26	1.35	4.53
7	1	2	7.76	6.39	1.58
8	4	9	1.12	4.51	6.04

Assign instances to the nearest cluster, update cluster means.



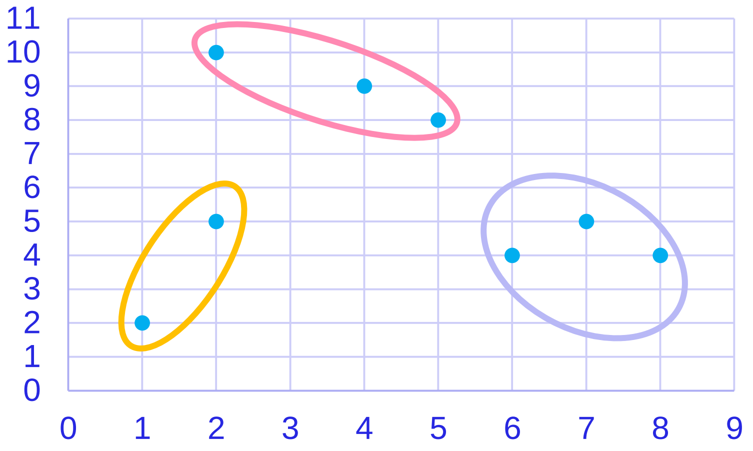
New centroids: $C_1 = (3.67, 9)$, $C_2 = (7, 4.33)$, $C_3 = (1.5, 3.5)$

Chair of Process and Data Science

Clustering
Solution 2

	att_1	att_2	Distance from (3.67,9)	Distance from (7,4.33)	Distance from (1.5,3.5)
1	2	10	1.95	7.56	6.52
2	2	5	4.33	5.04	1.58
3	8	4	6.61	1.06	6.52
4	5	8	1.66	4.17	5.7
5	7	5	5.2	0.66	5.7
6	6	4	5.52	1.06	4.53
7	1	2	7.49	6.44	1.58
8	4	9	0.33	5.54	6.04

No change!



Chair of Process and Data Science

Association rules
Exercise 3

- **Consider the following dataset:**

$$D = [\{A\}^{50}, \{A, B\}^{30}, \{A, B, C\}^{20}, \{A, D\}^{20}, \{A, F, C\}^{5}]$$

  - **Consider** $minimum\ support\ count =\ 22$**. Draw the FP-tree after removing the infrequent items.**
  - **Find the frequent itemsets (without calculating the conditional FP-trees)**

Chair of Process and Data Science

Association rules
Solution 3

$$D = [\{A\}^{50}, \{A, B\}^{30}, \{A, B, C\}^{20}, \{A, D\}^{20}, \{A, F, C\}^{5}]$$

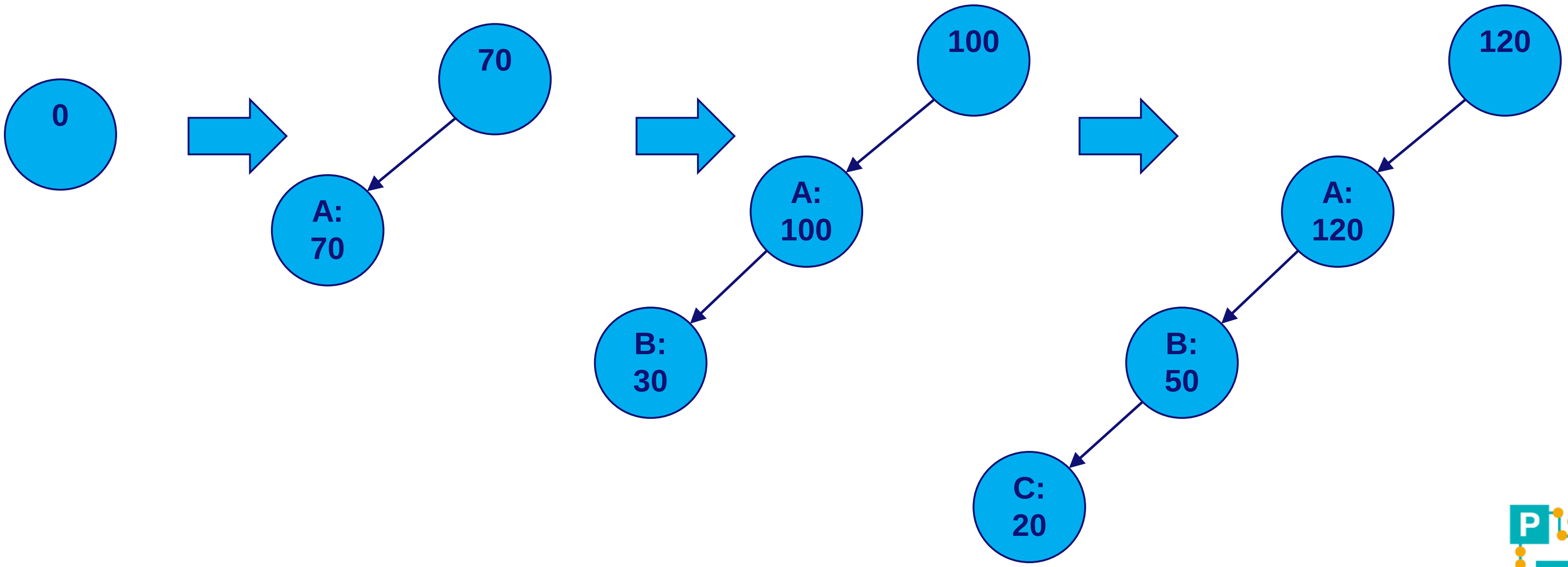
$$minimum\ support\ count = 22$$

A: 125
B: 50
C: 25
D: 20
F: 5

$$D' = [\{A\}^{50}, \{A, B\}^{30}, \{A, B, C\}^{20}, \{A\}^{20}, \{A, C\}^{5}]$$
$$= [\{A\}^{70}, \{A, B\}^{30}, \{A, B, C\}^{20}, \{A, C\}^{5}]$$

Association rules
Solution 3

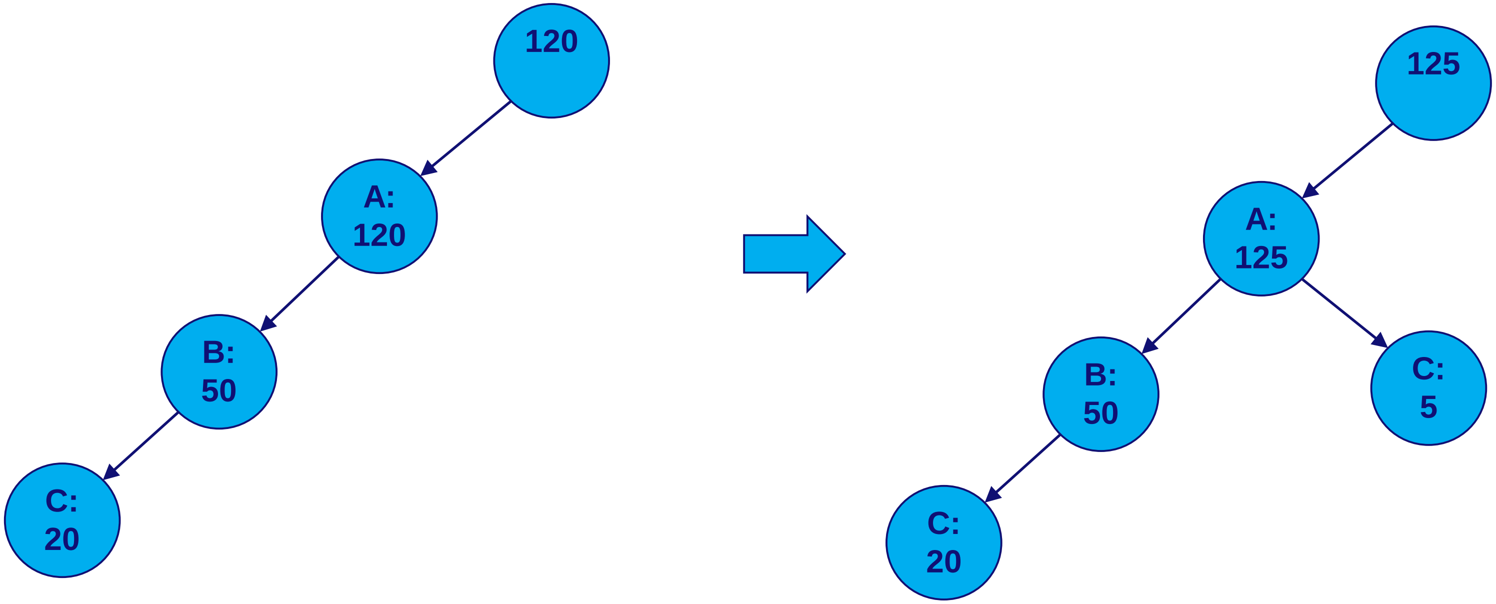
$$D' = [\{A\}^{70}, \{A, B\}^{30}, \{A, B, C\}^{20}, \{A, C\}^{5}]$$



Chair of Process and Data Science

Association rules
Solution 3

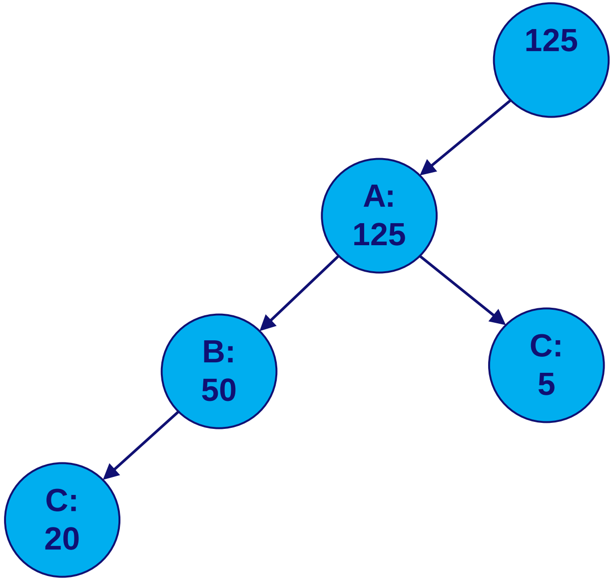
$$D' = [\{A\}^{70}, \{A, B\}^{30}, \{A, B, C\}^{20}, \{A, C\}^{5}]$$



Chair of Process and Data Science

Association rules
Solution 3

$$D' = [\{A\}^{70}, \{A,B\}^{30}, \{A,B,C\}^{20}, \{A,C\}^{5}]$$



Itemsets:
{A}: support 125
{B}: support 50
{C}: support 25
{A,B}: support 50
{A,C}: support 25
{B,C}: support 20
{A,B,C}: support 20

Frequent Itemsets:
{A}
{B}
{C}
{A,B}
{A,C}

$$minimum\ support\ count = 22$$

Association rules
Exercise 4

1. Compute $support$, $confidence$ and $lift$ for the following rules:

- $\{A, B\} \Rightarrow \{E\}$
- $\{A\} \Rightarrow \{C\}$

2. Which of the following rules satisfies the following conditions:

- $Support \geq 0.5$
- $Confidence \geq 0.8$
- $0.9 < lift < 1.1$

TID	Data items	frequency
1	A,B,E	10
2	C,A,D	25
3	C,B,D	15
4	C,A,B,E	20

Chair of Process and Data Science

Association rules
Solution 4

- $Support\ (\{A,B\} \Rightarrow\ \{E\}) = \frac{\mathbf{30}}{\mathbf{70}} = \mathbf{0.43}$
- $Confidence\ (\{A,B\} \Rightarrow\ \{E\}) = \frac{\mathbf{30}}{\mathbf{30}} = \mathbf{1}$
- $Lift\ (\{A,B\} \Rightarrow\ \{E\}) = \frac{\mathbf{30 \times 70}}{\mathbf{30 \times 30}} = \mathbf{2.3}$

- $Support\ (\{A\} \Rightarrow \{C\}) = \frac{\mathbf{45}}{\mathbf{70}} = \mathbf{0.64}$
- $Confidence\ (\{A\} \Rightarrow \{C\}) = \frac{\mathbf{45}}{\mathbf{55}} = \mathbf{0.81}$
- $Lift\ (\{A\} \Rightarrow \{C\}) = \frac{\mathbf{45 \times 70}}{\mathbf{55 \times 60}} = \mathbf{0.95}$

TID	Data items	frequency
1	A,B,E	10
2	C,A,D	25
3	C,B,D	15
4	C,A,B,E	20

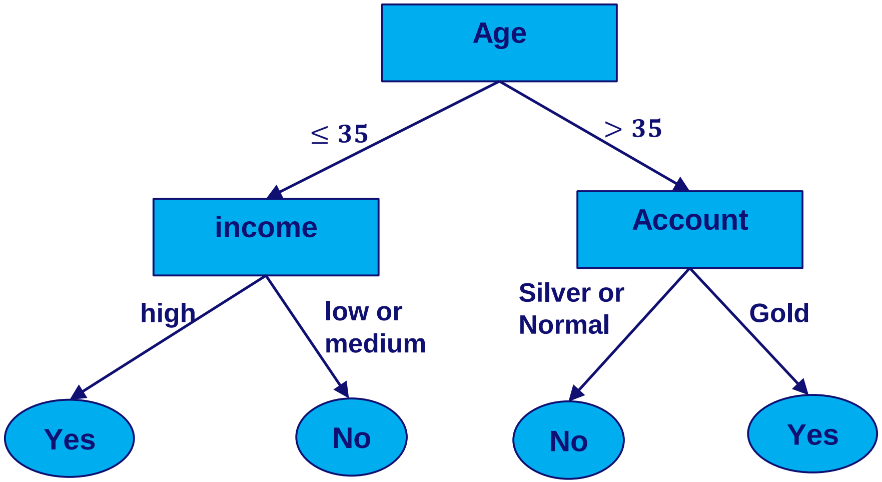
$\{A,B\} \Rightarrow\ \{E\}$ does not satisfy our conditions,
$\{A\} \Rightarrow \{C\}$, does satify the conditions

Chair of Process and Data Science

Evaluation
Exercise 5 – part 1

Create the confusion matrix for the following classifier model using the test data in the table:

ID	Age	Account type	Income	Accepted (actual class)
1	35	Silver	high	**Yes**
2	63	Gold	high	**No**
3	42	Normal	high	**No**
4	30	Normal	medium	**No**
5	35	Gold	low	**Yes**
6	56	Gold	high	**Yes**
7	23	Normal	low	**No**
8	48	Gold	medium	**No**



Chair of Process and Data Science

Evaluation
Exercise 5 – part 1

- **Confusion Matrix and performance measures**

		predicted		
		yes	no	
target	yes	TP	FN	P
	no	FP	TN	N
		P'	N'	$P+N$

Measure	Formula
$accuracy$	$\frac{TP+TN}{TP+TN+FP+FN}$
$misclassification$	$1-accuracy$
$recall$	$\frac{TP}{TP+FN}$
$precision$	$\frac{TP}{TP+FP}$
$F1-measure$	$2\times\frac{precision\times recall}{precision+recall}$

Evaluation
Solution 5 – part 1

ID	Age	Account type	Income	Accepted (actual class)	predicted class	type
1	25	Silver	high	Yes	Yes	TP
2	63	Gold	high	No	Yes	FP
3	42	Normal	high	No	No	TN
4	30	Normal	medium	No	No	TN
5	35	Gold	low	Yes	No	FN
6	56	Gold	high	Yes	Yes	TP
7	23	Normal	low	No	No	TN
8	48	Gold	medium	No	Yes	FP

		predicted		
		yes	no	
target	yes	2	1	3
	no	2	3	5
		4	4	8

Chair of Process and Data Science

Evaluation
Exercise 5 – part 2

Calculate $precision$, $recall$, $accuracy$ and $F1 - measure$ based on the confusion matrix.

Chair of Process and Data Science

Evaluation
Solution 5 – part 2

		predicted		
		yes	no	
target	yes	2	1	3
	no	2	3	5
		4	4	8

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{5}{8} = 0.63$$

$$precision = \frac{TP}{TP + FP} = \frac{2}{4} = 0.5$$

$$recall = \frac{TP}{TP + FN} = \frac{2}{3} = 0.67$$

$$F1 - measure = 2 \times \frac{precision \times recall}{precision + recall} = 0.57$$

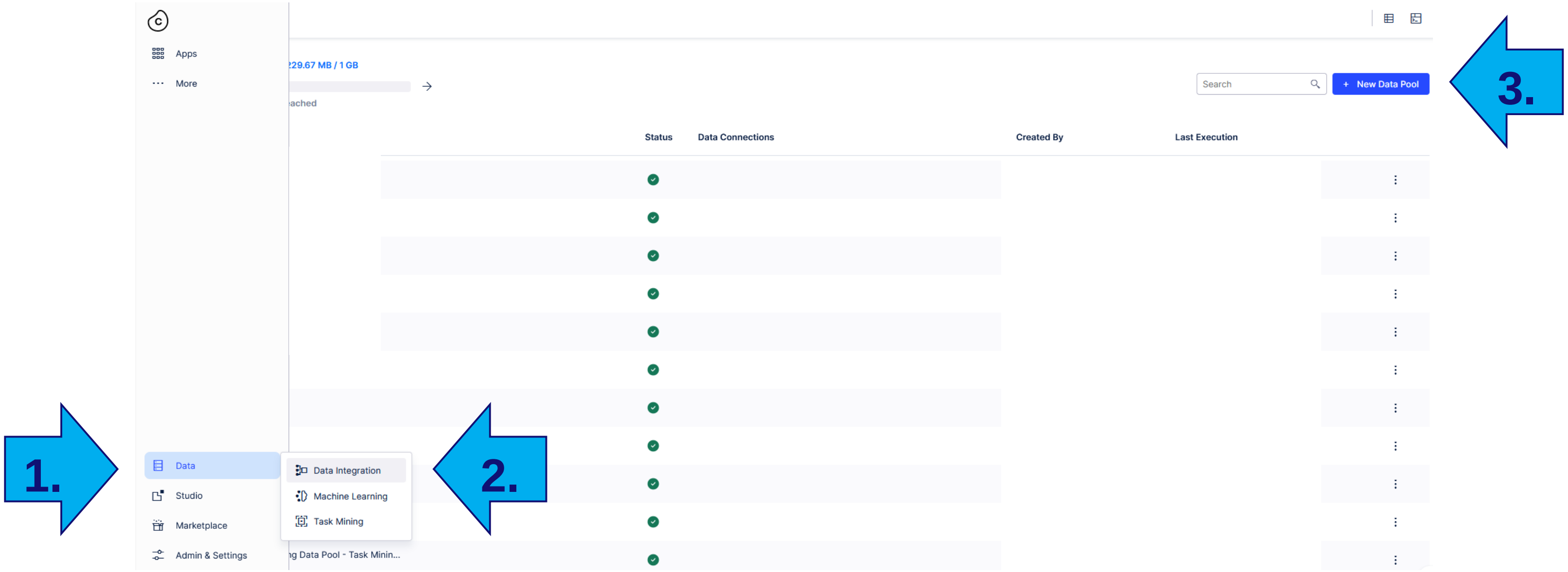
Chair of Process and Data Science

Creating Case-Situation Tables in Celonis

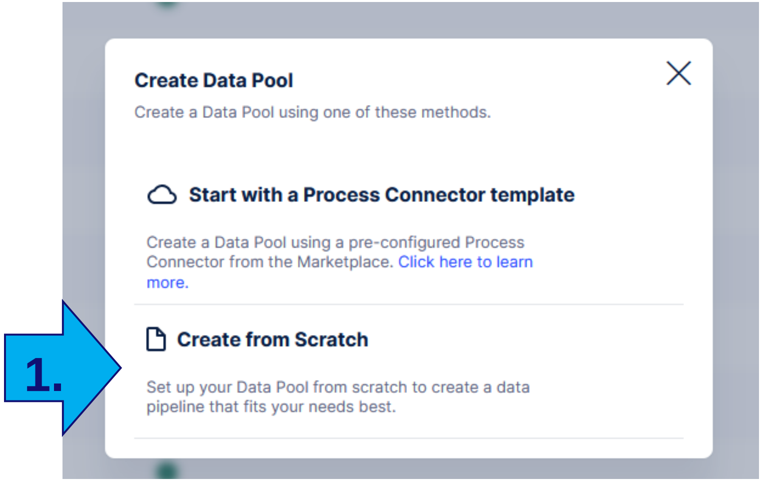
Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Create a Data Pool

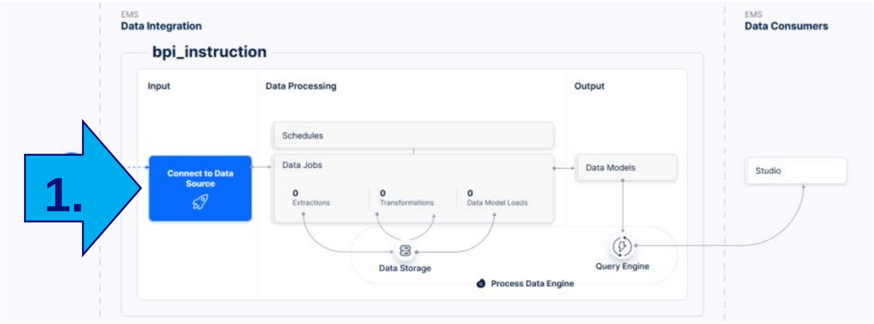


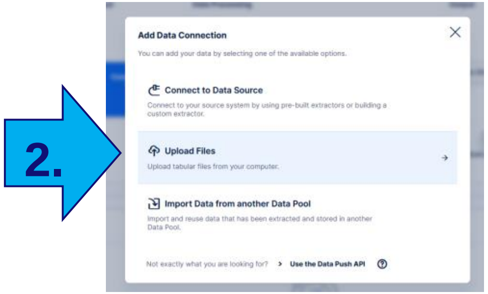
Create a Data Pool

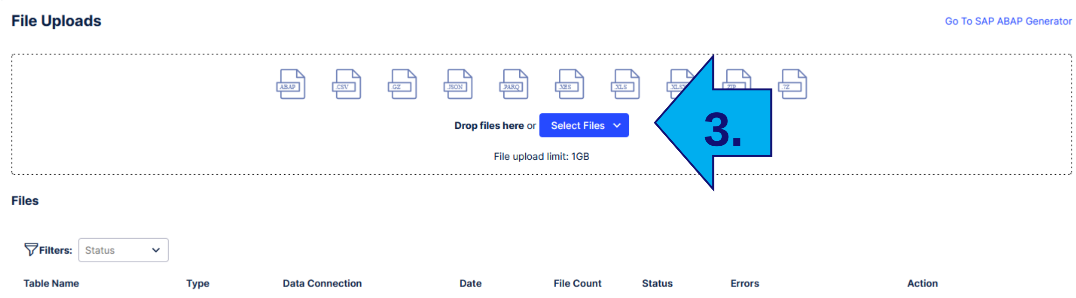


Chair of Process and Data Science

File Upload







Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

File Upload

Files

Filters: Status

Table Name	Type	Data Connection	Date	File Count	Status	Errors	Action
Case_Table	Flat File	[Global]	05-04-2023 11:33	1			
Activity_Table	Flat File	[Global]	05-04-2023 11:33	1			

1.

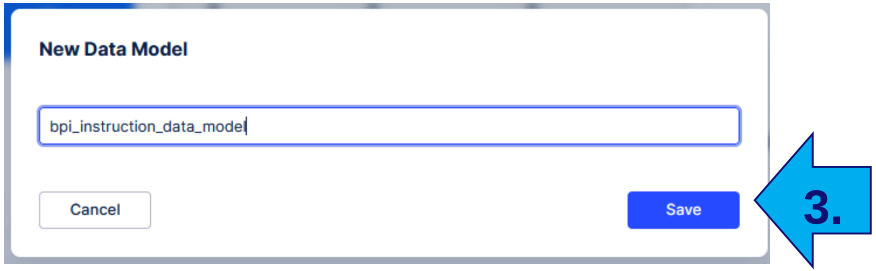
2.) Return to overview

Chair of Process and Data Science

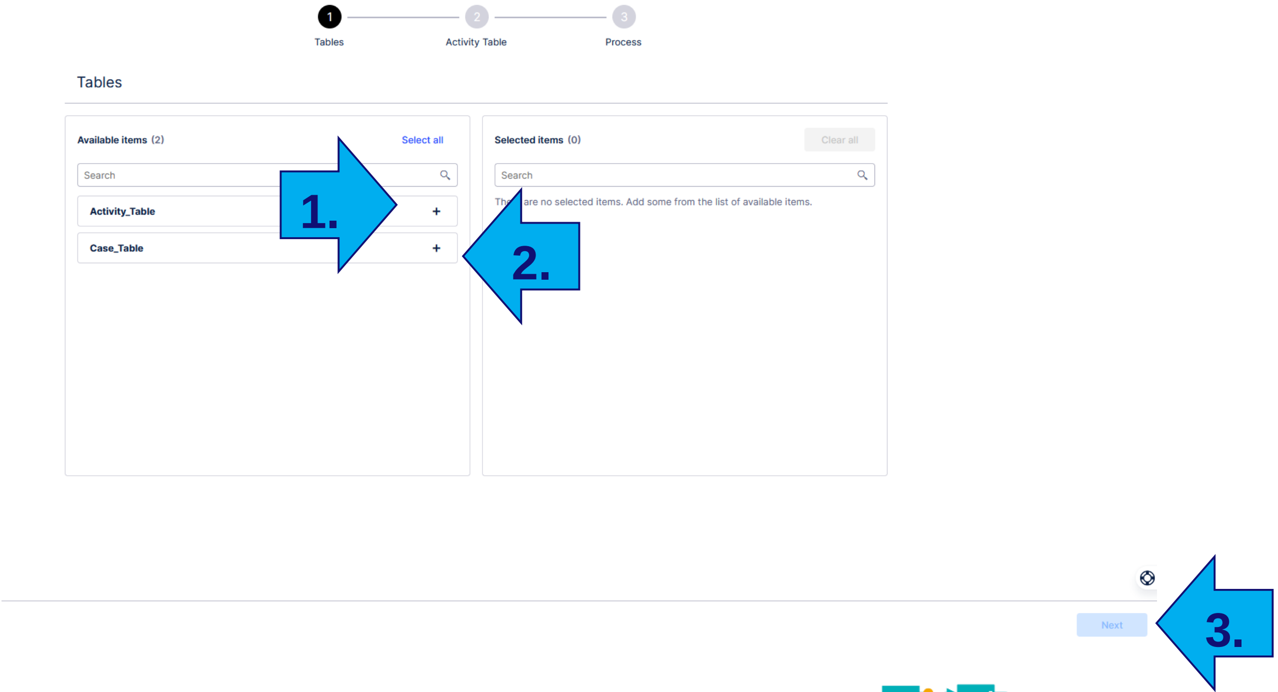
RWTH AACHEN UNIVERSITY

Data Model

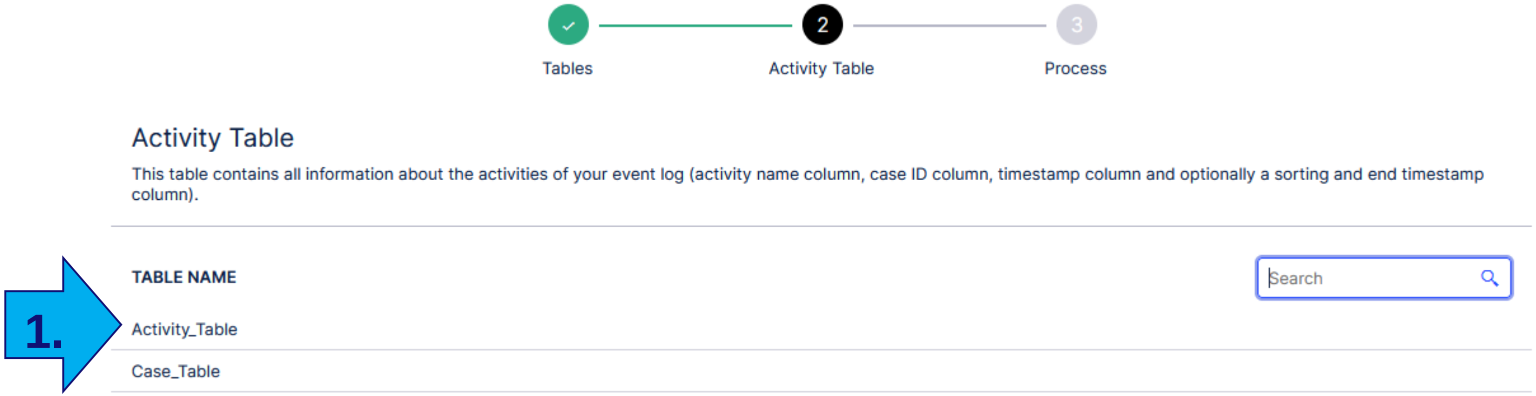

bpi_instruction
Input
Data Processing
Output
Schedules
Data Jobs
Connect to Data Source
0
Extractions
0
Transformations
Data Model Loa
Data Models
Studio
Data Storage
Query Engine
Process Data Engine
1.
2.
+ Add Data Model


New Data Model
bpi_instruction_data_model
Cancel
Save
3.

Data Model


1
2
3
Tables
Activity Table
Process
Tables
Available items (2)
Select all
Search
Activity_Table
Case_Table
Selected items (0)
Clear all
Search
are no selected items. Add some from the list of available items.
Next
1.
2.
3.

Data Model



Chair of Process and Data Science

Data Model

1. 2. 3.

ABC CASE ID	ABC ACTIVITY	ABC ORG:RESOURCE	DATE COMPLETE TIMES...	ABC EVENTORIGIN	0.1F FIRSTWITHDRAW...	0.1F CREDITSCORE	0.1F NUMBEROFTERMS	0.1F MONTHLYCOST	0.1F OFFEREDAMOUNT
Application_10003867...	A_Accepted	User_77	2016-11-25 16:33:09	Application	<Empty>	<Empty>	<Empty>	<Empty>	<Empty>
Application_10003867...	A_Complete	User_77	2016-11-25 16:35:38	Application	<Empty>	<Empty>	<Empty>	<Empty>	<Empty>
Application_10003867...	A_Concept	User_1	2016-11-25 15:32:40	Application	<Empty>	<Empty>	<Empty>	<Empty>	<Empty>
Application_10003867...	A_Create Application	User_1	2016-11-25 15:31:09	Application	<Empty>	<Empty>	<Empty>	<Empty>	<Empty>
Application_10003867...	A_Incomplete	User_29	2016-12-06 11:03:16	Application	<Empty>	<Empty>	<Empty>	<Empty>	<Empty>
Application_10003867...	A_Incomplete	User_90	2016-12-02 09:03:39	Application	<Empty>	<Empty>	<Empty>	<Empty>	<Empty>
App ✕ Case ID 57...	A ✕ Activity name	User Sorting	20 ✕ Timestamp	Appl Sorting	<Em Sorting	<Em Sorting	<Em Sorting	<Em Sorting	<Em Sorting
Application_10003867...	A_Submitted	User_1	2016-11-25 15:31:11	Application	<Empty>	<Empty>	<Empty>	<Empty>	<Empty>
Application_10003867...	A_Validating	User_133	2016-12-05 20:02:03	Application	<Empty>	<Empty>	<Empty>	<Empty>	<Empty>
Application_10003867...	A_Validating	User_53	2016-12-01 15:20:14	Application	<Empty>	<Empty>	<Empty>	<Empty>	<Empty>
Application_10003867...	O_Accepted	User_123	2016-12-06 14:58:37	Offer	<Empty>	<Empty>	<Empty>	<Empty>	<Empty>
Application_10003867...	O_Create Offer	User_77	2016-11-25 16:35:29	Offer	5000	1080	58	100.25	5000
Application_10003867...	O_Created	User_77	2016-11-25 16:35:29	Offer	<Empty>	<Empty>	<Empty>	<Empty>	<Empty>
Application_10003867...	O_Returned	User_53	2016-12-01 15:20:27	Offer	<Empty>	<Empty>	<Empty>	<Empty>	<Empty>
Application_10003867...	O_Sent (mail and online)	User_77	2016-11-25 16:35:38	Offer	<Empty>	<Empty>	<Empty>	<Empty>	<Empty>
Application_10003867...	W_Validate application	User_90	2016-12-02 09:03:39	Workflow	<Empty>	<Empty>	<Empty>	<Empty>	<Empty>
Application_1001177986	A_Accepted	User_52	2016-06-25 16:01:30	Application	<Empty>	<Empty>	<Empty>	<Empty>	<Empty>

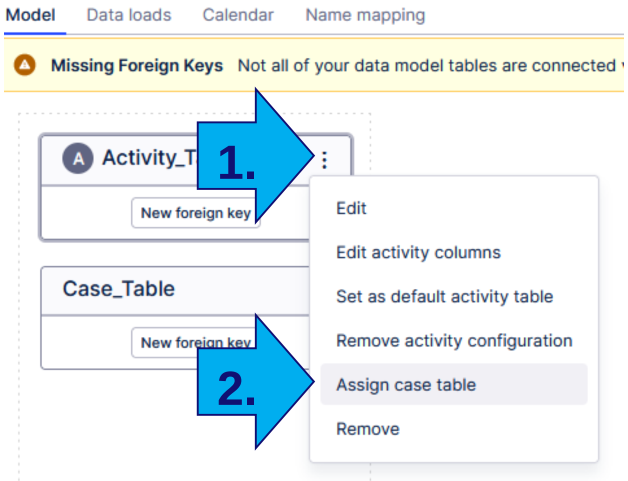
Back

4.

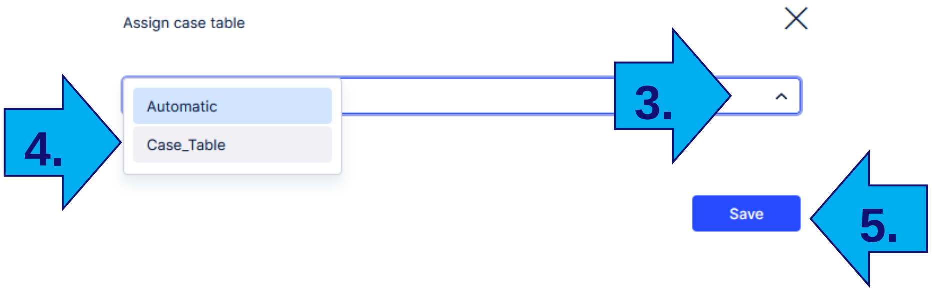
Finish

Chair of Process and Data Science

Data Model

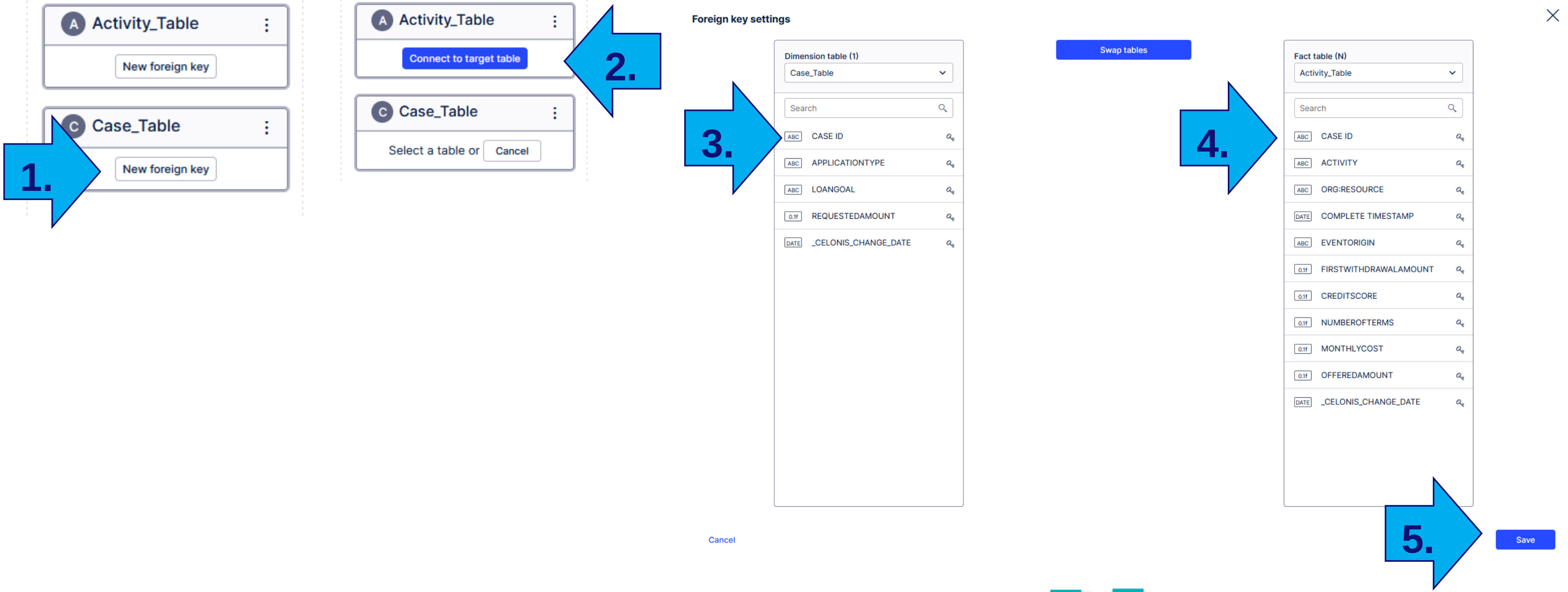


Model
Data loads
Calendar
Name mapping
Missing Foreign Keys
Not all of your data model tables are connected
Activity_T
New foreign key
Case_Table
New foreign key
1.
2.
Edit
Edit activity columns
Set as default activity table
Remove activity configuration
Assign case table
Remove



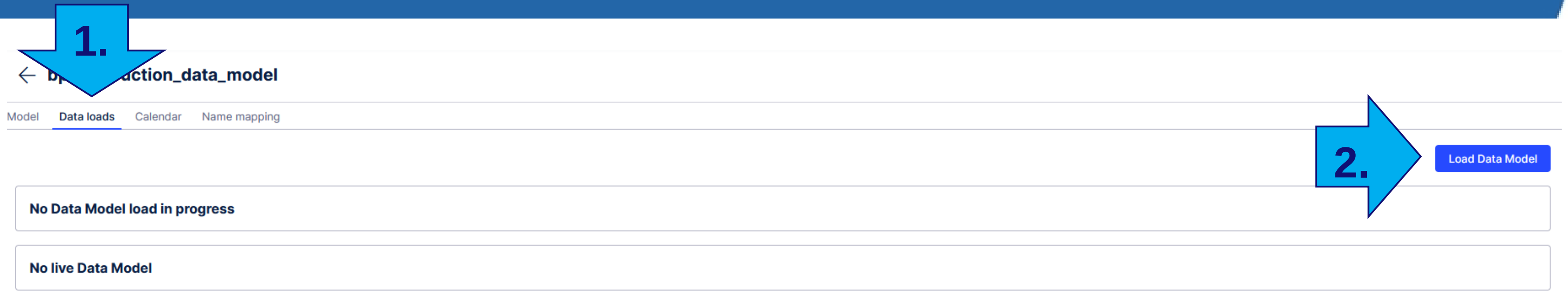
Assign case table
Automatic
Case_Table
3.
4.
Save
5.

Data Model



Activity_Table
New foreign key
Case_Table
New foreign key
1.
Activity_Table
Connect to target table
Case_Table
Select a table or
Cancel
2.
Foreign key settings
Dimension table (1)
Case_Table
Search
CASE ID
APPLICATIONTYPE
LOANGOAL
REQUESTEDAMOUNT
_CELONIS_CHANGE_DATE
3.
Swap tables
Fact table (N)
Activity_Table
Search
CASE ID
ACTIVITY
ORG:RESOURCE
COMPLETE TIMESTAMP
EVENTORIGIN
FIRSTWITHDRAWALAMOUNT
CREDITSCORE
NUMBEROFTERMS
MONTHLYCOST
OFFEREDAMOUNT
_CELONIS_CHANGE_DATE
4.
Cancel
5.
Save

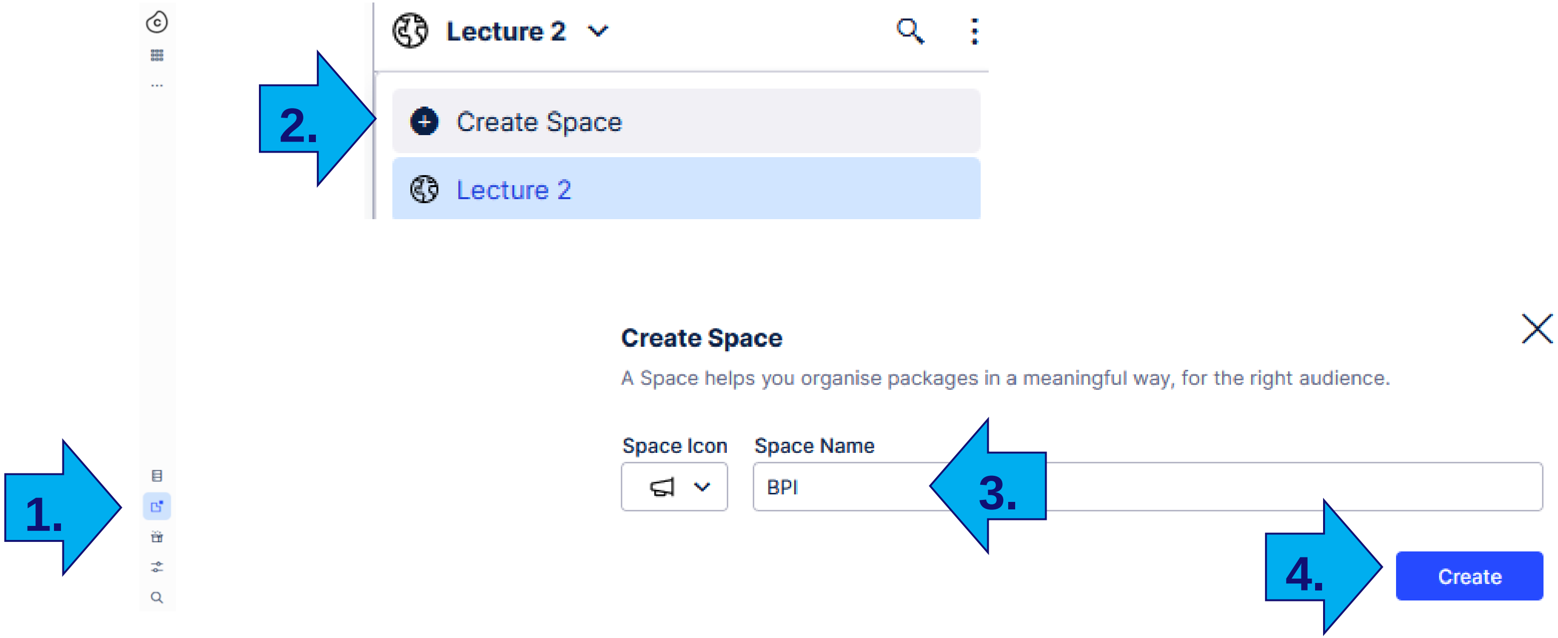
Data Model



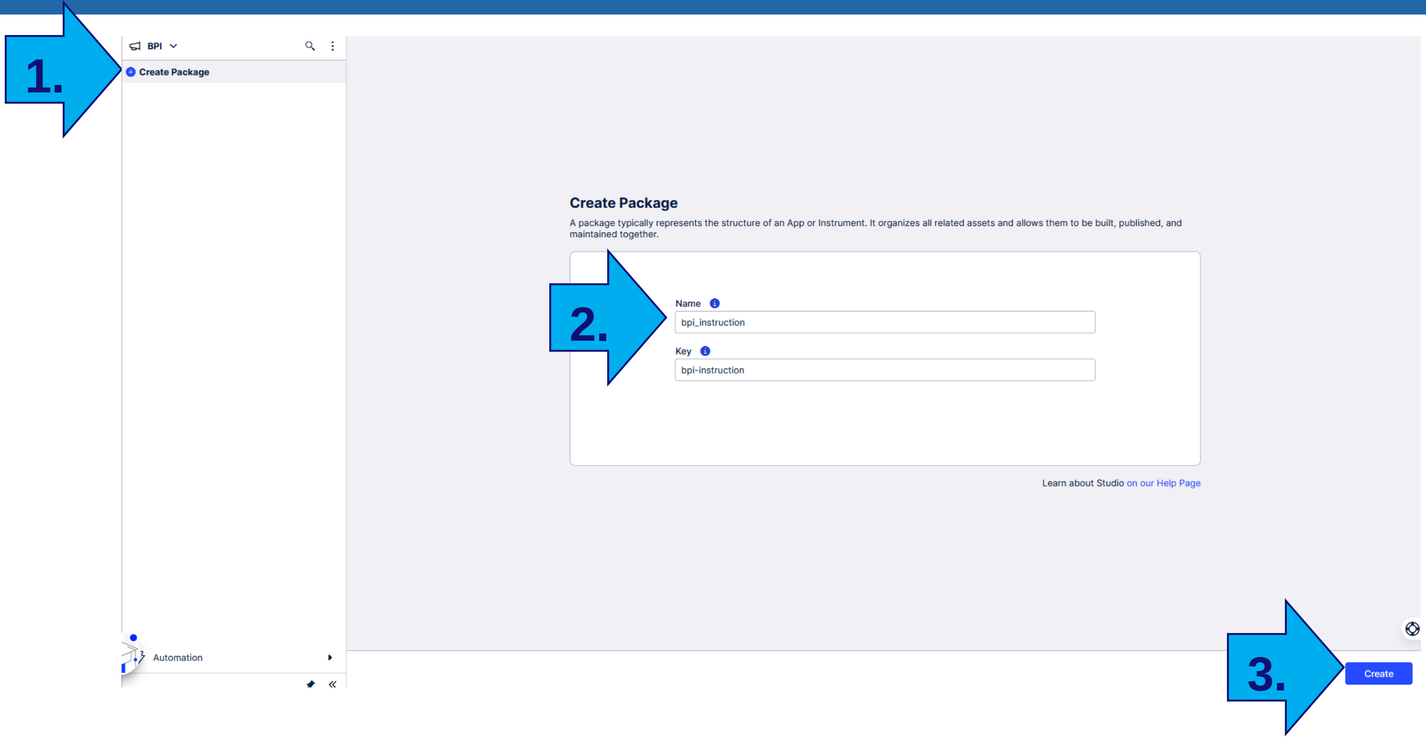
Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

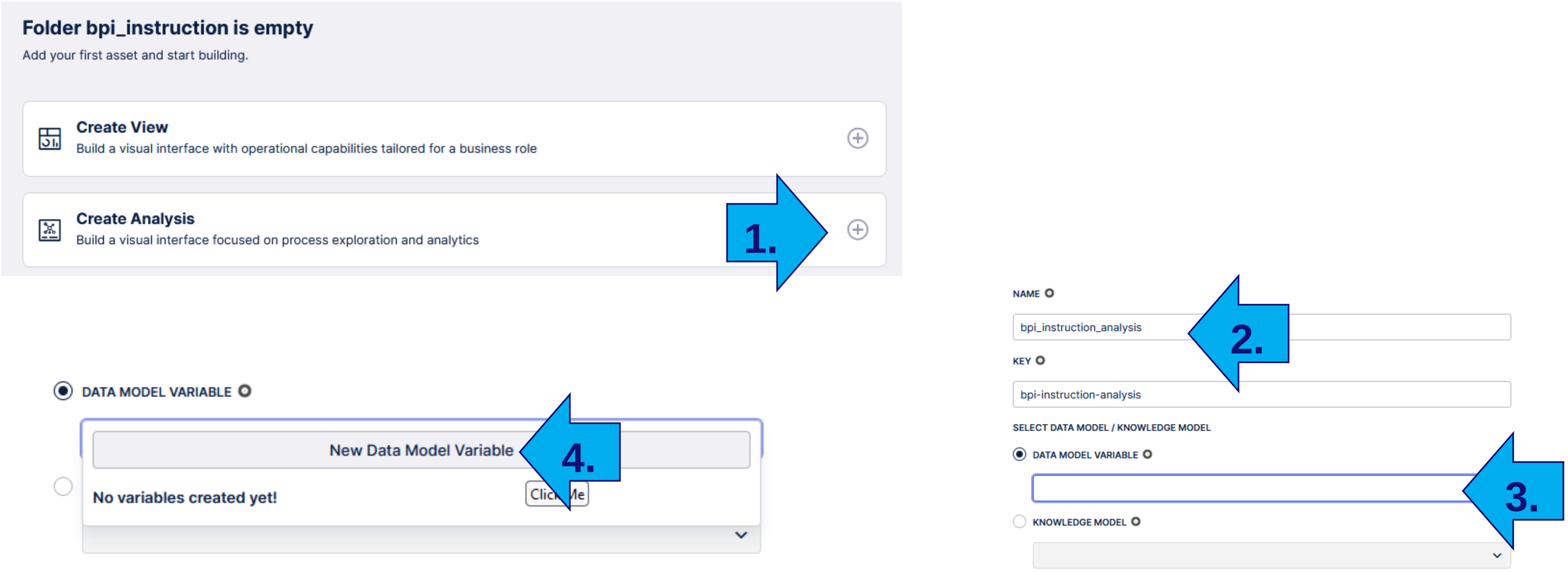
Studio


1.
2.
Lecture 2
Create Space
Lecture 2
Create Space
A Space helps you organise packages in a meaningful way, for the right audience.
Space Icon
Space Name
BPI
3.
4.
Create

Studio


1.
BPI
Create Package
Automation
Create Package
A package typically represents the structure of an App or Instrument. It organizes all related assets and allows them to be built, published, and maintained together.
2.
Name
bpi_instruction
Key
bpi-instruction
Learn about Studio on our Help Page
3.
Create

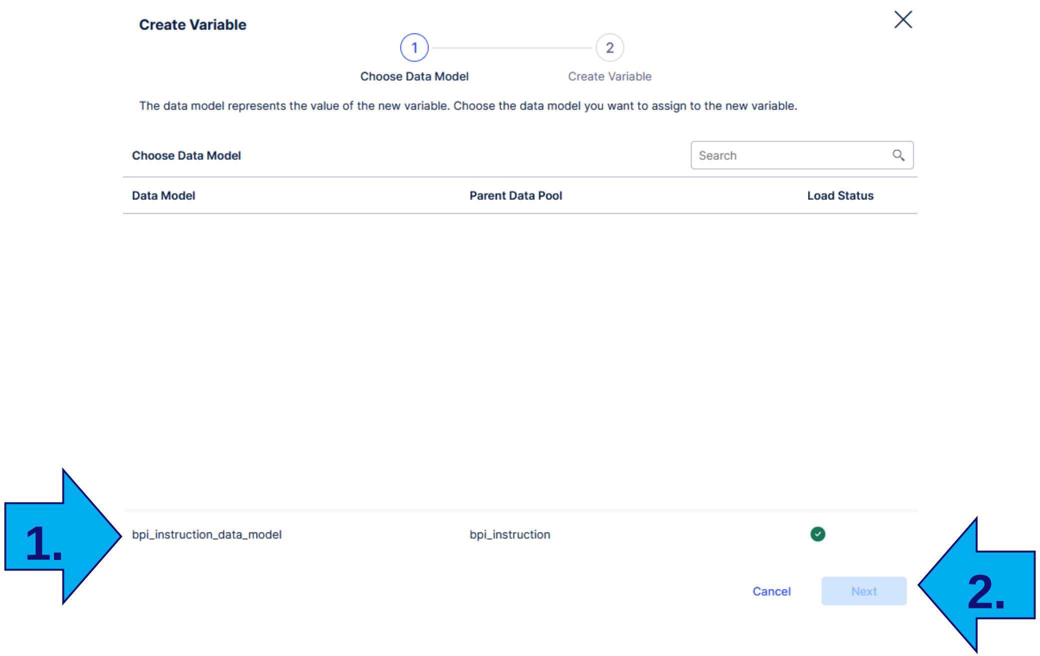
Studio

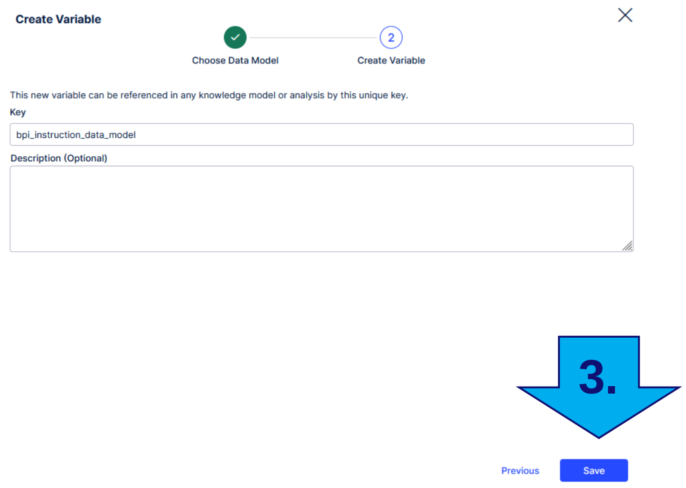


Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Studio

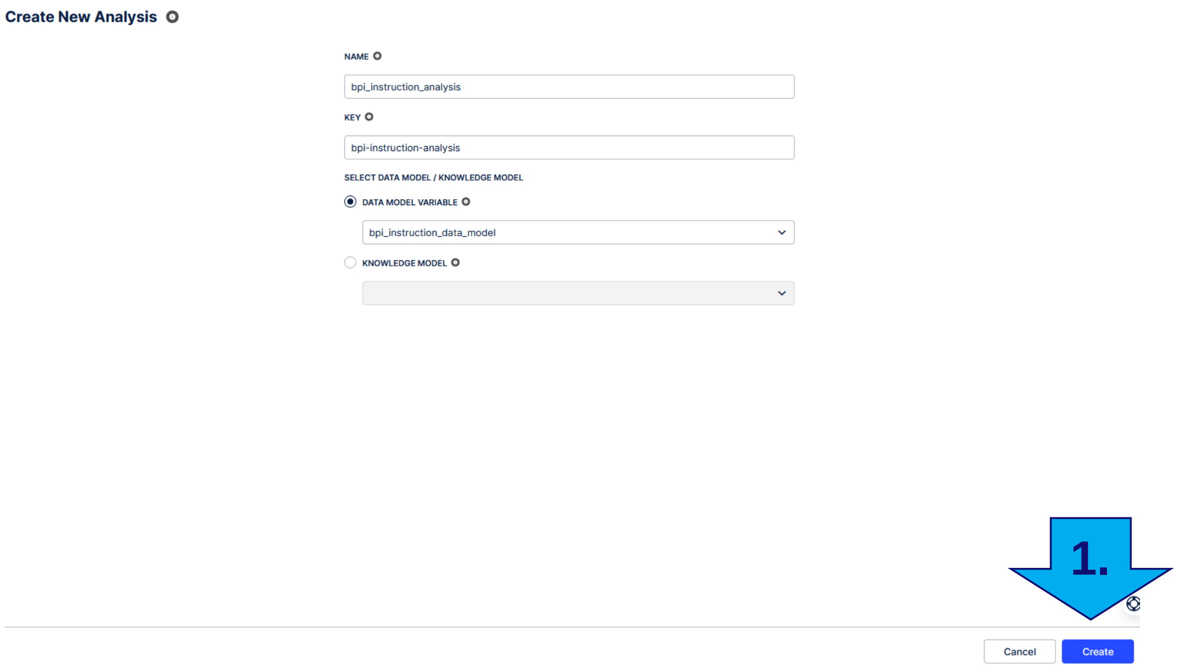




Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

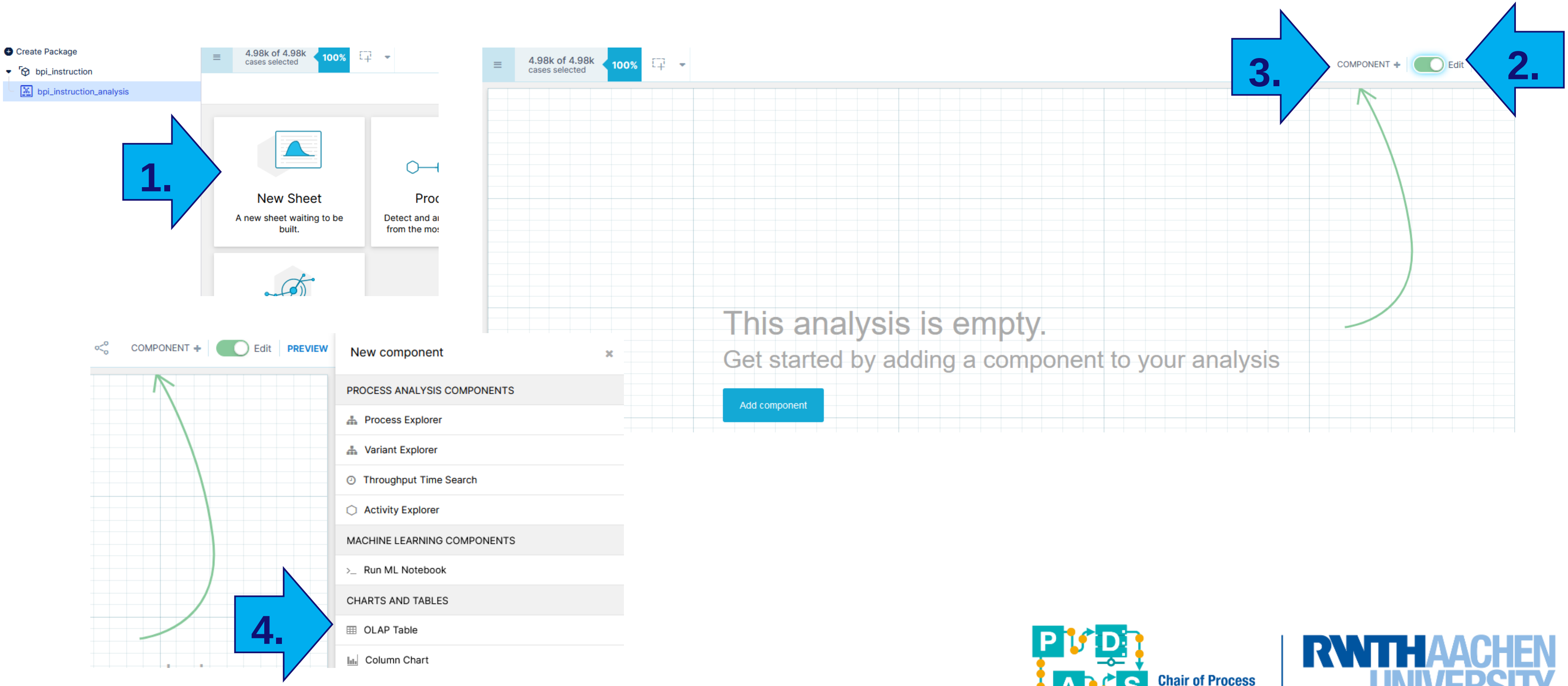
Studio



Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

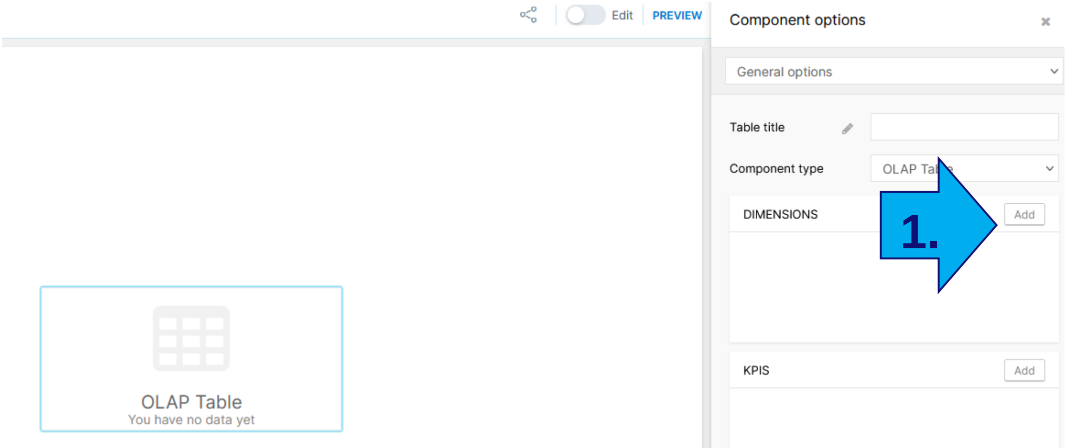
Analysis



Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

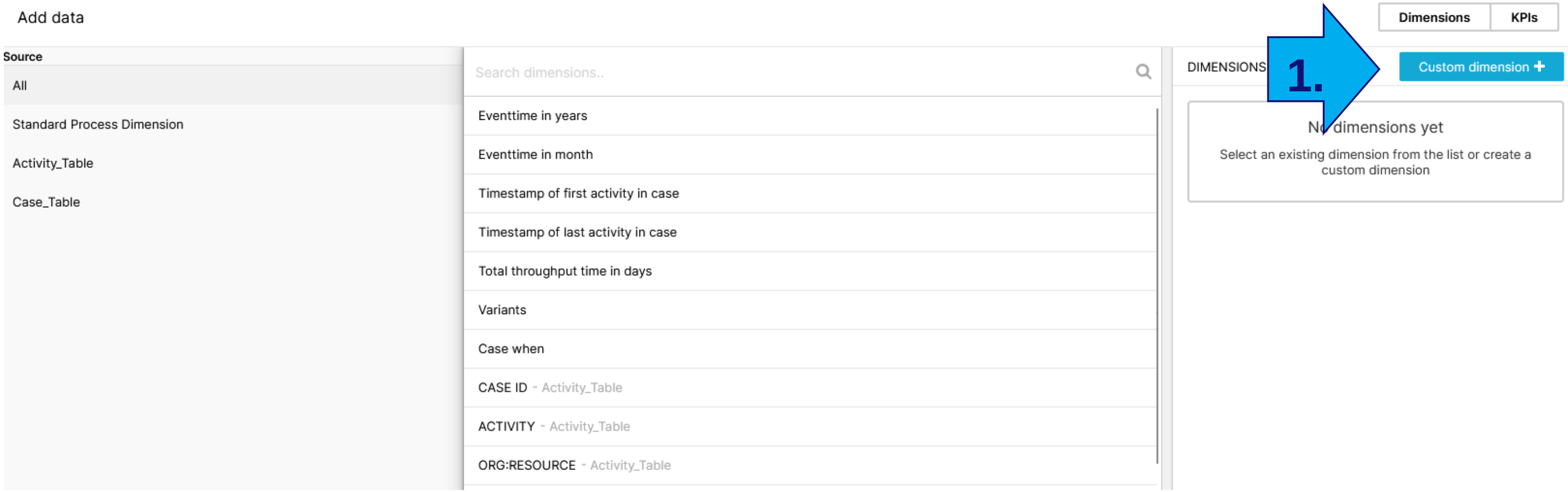
Analysis



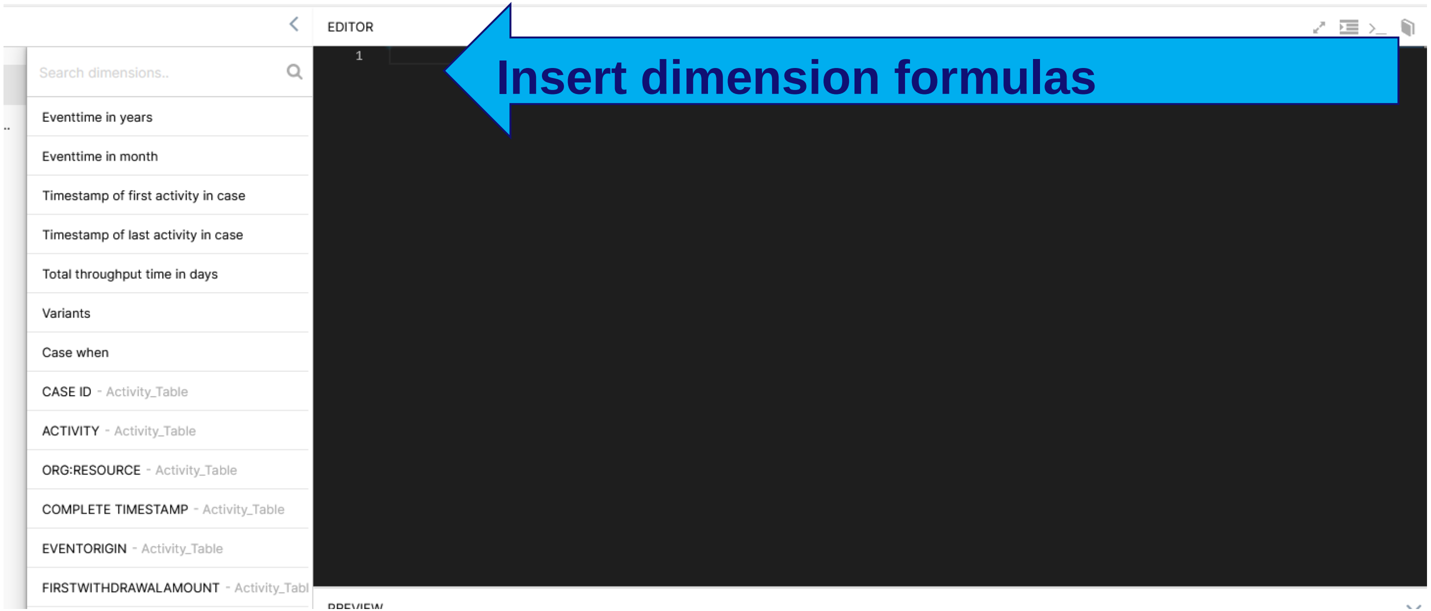
Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Analysis


Add data
Dimensions
KPIs
Source
All
Standard Process Dimension
Activity_Table
Case_Table
Search dimensions..
Eventtime in years
Eventtime in month
Timestamp of first activity in case
Timestamp of last activity in case
Total throughput time in days
Variants
Case when
CASE ID - Activity_Table
ACTIVITY - Activity_Table
ORG:RESOURCE - Activity_Table
DIMENSIONS
1.
Custom dimension
No dimensions yet
Select an existing dimension from the list or create a custom dimension

Analysis



Chair of Process and Data Science
RWTH AACHEN UNIVERSITY

Accessing Columns in Table

- **For later formulas, you need to refer to the columns of your table**

- **You refer to columns with YourTable.YourColumn**

Pull Up Aggregation

- **Pull-Up-functions are used to achieve nested aggregations and filters on aggregations**
- **The formula always consists of "PU_X (target_table, source_table.column …)"**
- **A *1:N* relationship between the target table and the table of the specified source column is required.**
  - **Therefore, your case table is, most of the time, the target table, your activity table the source table**

Analysis

Function Name	General Formula
Case When	Case When … Then … Else … End
COUNT	Count([Distinct] Column)
CALC_REWORK	CALC_REWORK(Condition, Column)
PU_AGGREGATION	PU_FIRST(/LAST) (Group, Column, Condition, Order)
CALC_THROUGHPUTTIME	CALC_THROUGHPUTTIME(Trace Crop, Transformed Timestamp to Integer)
DAYS_BETWEEN	DAYS_BETWEEN (Column, Column)
MATCH_ACTIVITIES	MATCH_ACTIVITIES(Column, STARTING[], NODE[], EXCLUDING[], ENDING [some Activities])

More formulas and explanations: https://docs.celonis.com/en/pql-function-library.html

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Analysis

Function Name	Role in Creating the Situation Table’s Variables
Case When	Creating conditional structures in other functions
COUNT/COUNT Distinct	Extracting trace features based on the times certain activities occur in the trace/ or checks just the existence of these activities in the trace
CALC_REWORK	Extracting trace features based on the times certain activities occur in the trace
PU_FIRST, PU_LAST	Extracting trace attributes based on the first/last trace’s event attribute
CALC_THROUGHPUTTIME	Extracting the throughput time feature of a trace(between start and end events or any other pairs of the trace)
DAYS_BETWEEN	Extracting a time feature of a trace between a pair of its events (in days)
MATCH_ACTIVITIES	Extracting a binary trace attribute whether it starts/ ends with certain activities or includes /excludes these activities

More formulas and explanations: https://docs.celonis.com/en/pql-function-library.html

Analysis

Dimension Name	Dimension Formula
Case ID	"Case_Table"."CASE ID"
Application Type	"Case_Table"."APPLICATIONTYPE"
Loan Goal	"Case_Table"."LOANGOAL"
Requested Amount	"Case_Table"."REQUESTEDAMOUNT"
First Credit Score	PU_FIRST ("Case_Table", "Activity_Table"."CREDITSCORE")
Last Credit Score	PU_LAST ("Case_Table", "Activity_Table"."CREDITSCORE")

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Analysis

Dimension Name	Dimension Formula
Having Call to Complete Application	COUNT(DISTINCT CASE WHEN "Activity_Table"."ACTIVITY"='W_Call incomplete files' THEN 1 END) Or alternatively: CASE WHEN MATCH_ACTIVITIES("Activity_Table"."ACTIVITY", NODE ['W_Call incomplete files'])=1 THEN 1 ELSE 0 END

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Analysis

Dimension Name	Dimension Formula
Number of Offers	COUNT(CASE WHEN "Activity_Table"."ACTIVITY"='O_Create Offer' THEN "Activity_Table"."ACTIVITY" END) Or alternatively: CALC_REWORK ("Activity_Table"."ACTIVITY"='O_Create Offer',"Activity_Table"."ACTIVITY")

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Analysis

Dimension Name	Dimension Formula
First Offer	PU_FIRST ("Case_Table", "Activity_Table"."OFFEREDAMOUNT")
Last Offer	PU_LAST ("Case_Table", "Activity_Table"."OFFEREDAMOUNT")
First Offers' Monthly Cost	PU_FIRST ("Case_Table","Activity_Table"."MONTHLYCOST")
Last Offers' Monthly Cost	PU_LAST ("Case_Table","Activity_Table"."MONTHLYCOST")

Analysis

Dimension Name	Dimension Formula
Throughput Times	DAYS_BETWEEN (MIN("Activity_Table"."COMPLETE TIMESTAMP"), MAX("Activity_Table"."COMPLETE TIMESTAMP")) Or alternatively: CALC_THROUGHPUT (CASE_START TO CASE_END, REMAP_TIMESTAMPS("Activity_Table"."COMPLETE TIMESTAMP", DAYS))

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Analysis

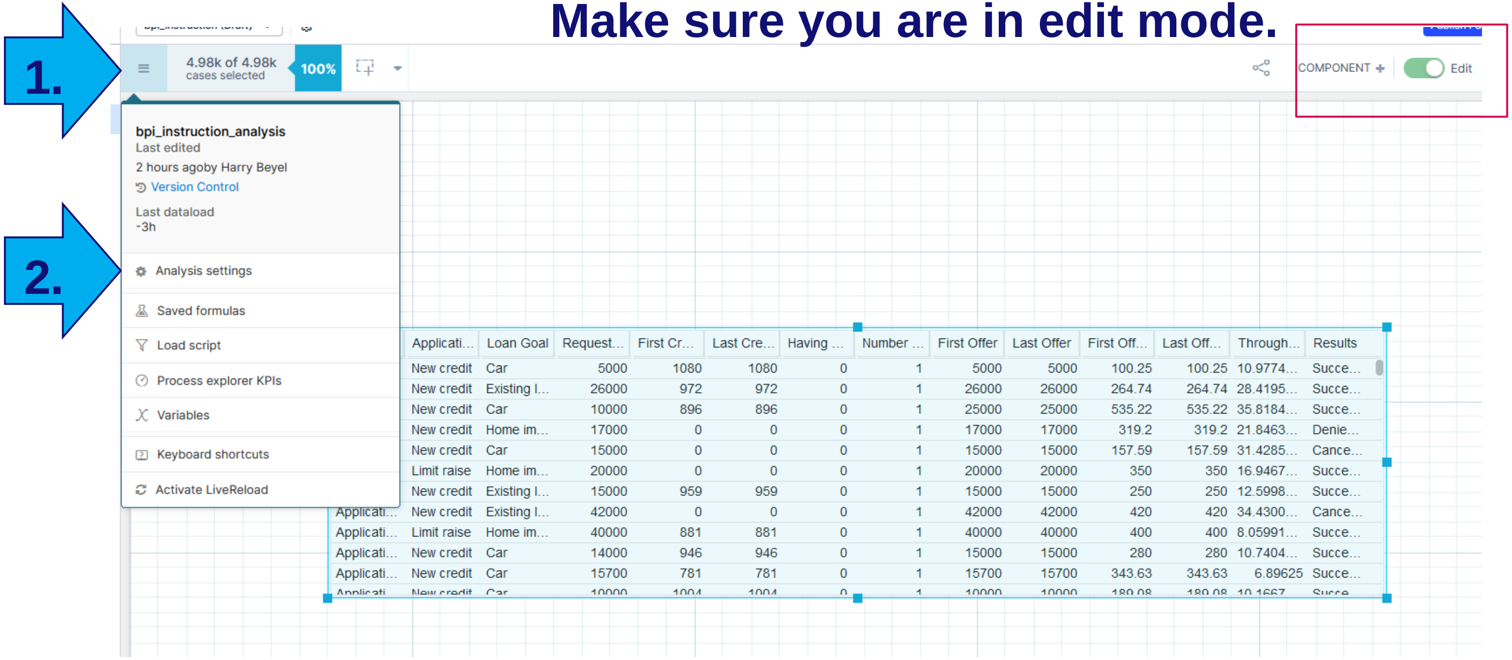
Dimension Name	Dimension Formula
Results	CASE WHEN MATCH_ACTIVITIES ("Activity_Table"."ACTIVITY", NODE ['A_Cancelled']) = 1 THEN 'Cancelled by client' WHEN MATCH_ACTIVITIES ("Activity_Table"."ACTIVITY", NODE ['A_Denied']) = 1 THEN 'Denied by the bank' WHEN MATCH_ACTIVITIES ("Activity_Table"."ACTIVITY", NODE ['A_Pending']) = 1 THEN 'Successful completion' END

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

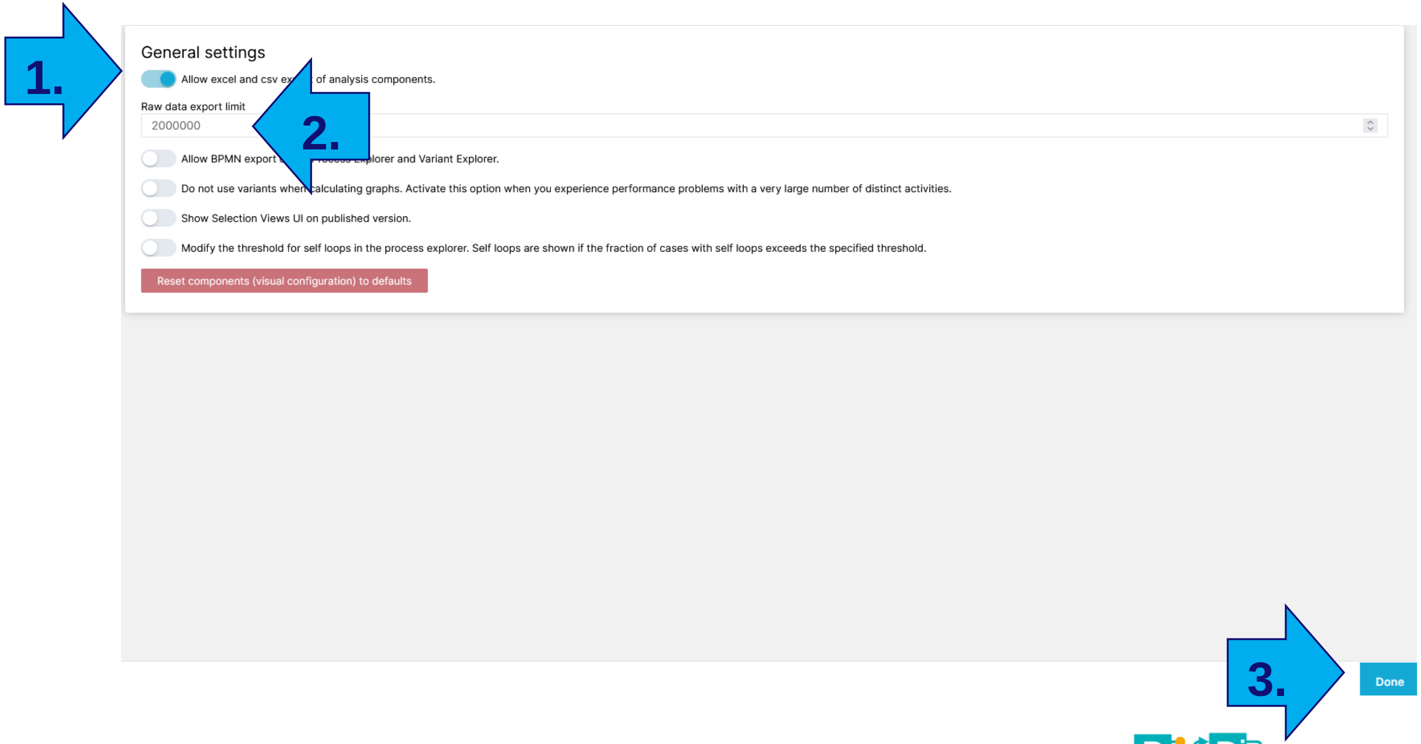
Export Situation Table

Make sure you are in edit mode.

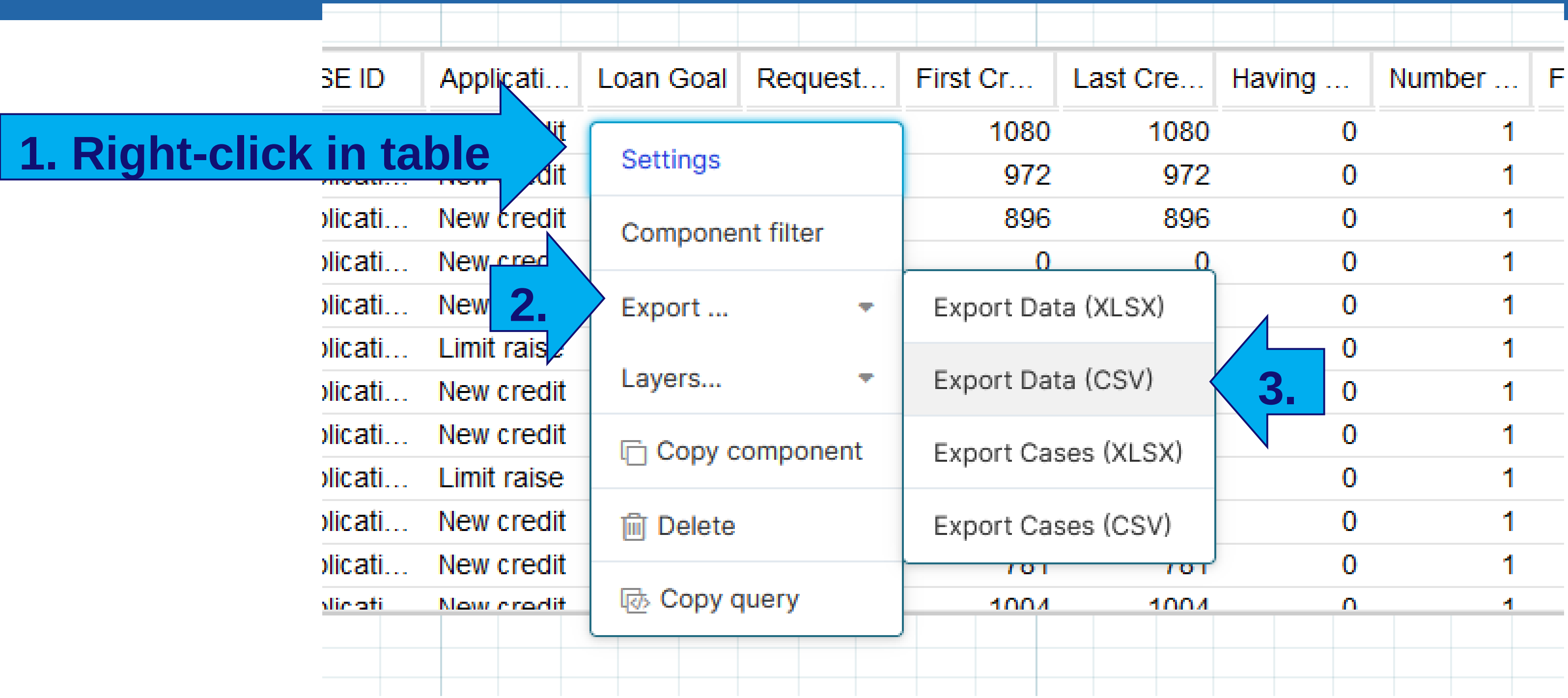


Chair of Process and Data Science
RWTH AACHEN UNIVERSITY

Export Situation Table



Export Situation Table



Export Situation Table

You can find your CSV-file in your download folder.

Chair of Process and Data Science

Working with Situation Tables from Celonis

Chair of Process and Data Science

About the data

1. CASE ID: The unique identifier assigned to each loan application
2. Application Type: The type of each loan application
3. Loan Goal: The goal of each loan application
4. Requested Amount The amount each loan application asks to be paid
5. First Credit Score: The first recorded credit of the application
6. Last Credit Score: The last recorded credit of the application
7. Having Call to Complete Application: Indicates if the loan application process includes the activity “W_Call incomplete files” which means the bank calls the applicant to provide further data/documents.
8. Number Of Offers: The number of times the bank offers a loan to an application. These offers are logged when the event’s activity is “O_Create Offer”.
9. First Offer: The amount of the first offer the bank provides to an application.
10. Last Offer: The amount of the last offer the bank provides to an application.
11. First Offer’s Monthly Cost: The monthly cost of the first offer the bank provides to an application.
12. Last Offer’s Monthly Cost: The monthly cost of the last offer the bank provides to an application.
13. Throughput Time: The time spent on each application from its first event to the last
14. Results: The outcome of each application. This could be one of the following:
    i. Successful completion: the application includes the “A_Pending” activity. This activity describes a situation in which all documents have been received and the assessment is positive. The loan is confirmed and signed to be paid to the customer.
    ii. Denied by the bank: the application includes “A_Denied” activity. This activity describes a situation in which the application doesn't match the acceptance criteria.
    iii. Cancelled by client: the application includes the “A_Cancelled” activity. This activity describes a situation in which the applicant does not get back to the bank after an offer was sent out

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

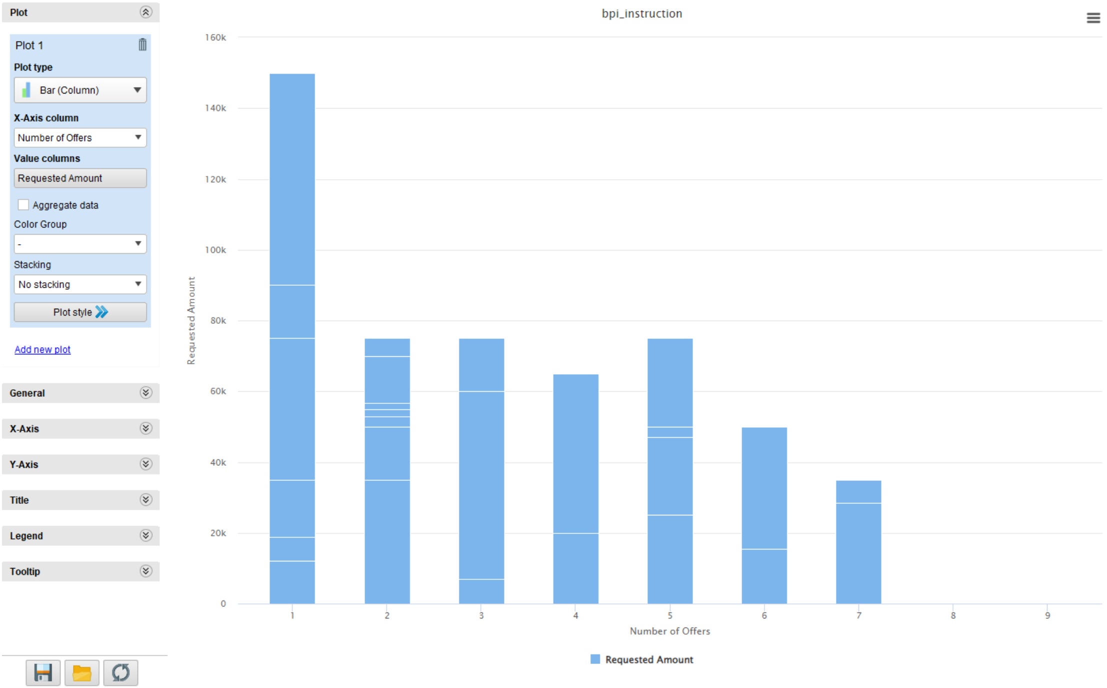
Import the data

- Import the data in RapidMiner
- No date format needed

Visualization

- **Click on Visualization and create a bar graph (column), with an x-axis showing *Numbers of Offers*, and *Requested Amount* as value column**
- **What are your observations?**

Solution



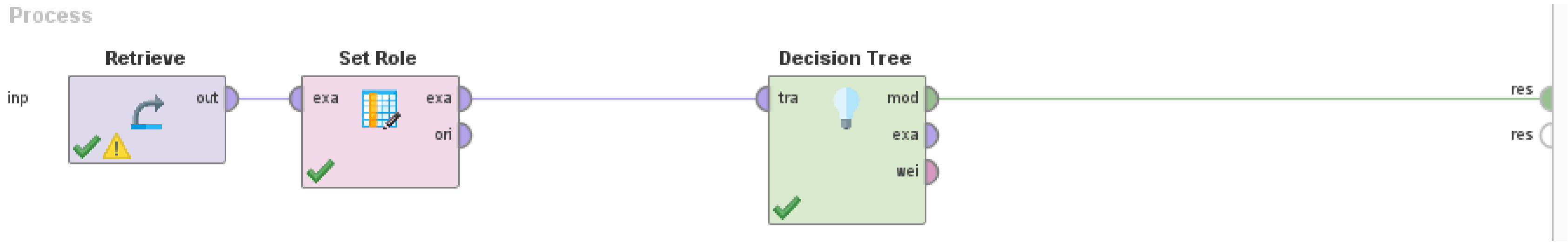
Plot
Plot 1
Plot type
Bar (Column)
X-Axis column
Number of Offers
Value columns
Requested Amount
Aggregate data
Color Group
-
Stacking
No stacking
Plot style
Add new plot
General
X-Axis
Y-Axis
Title
Legend
Tooltip
bpi_instruction
160k
140k
120k
100k
80k
60k
40k
20k
0
Requested Amount
1
2
3
4
5
6
7
8
9
Number of Offers
Requested Amount

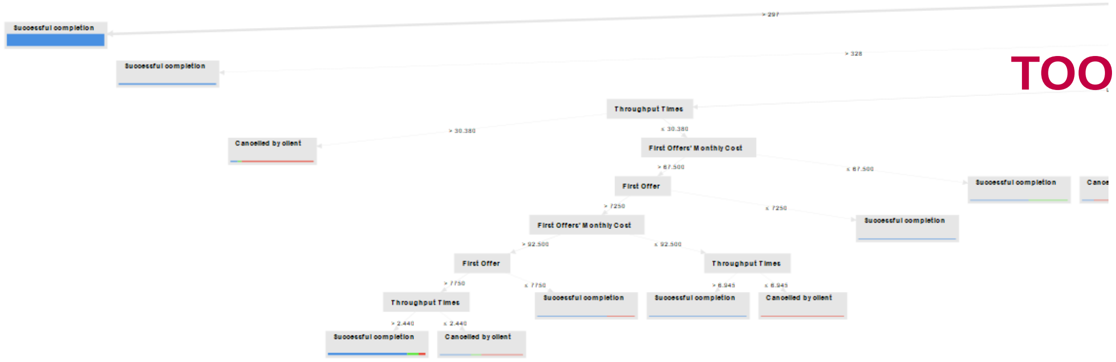
Decision Tree

- We want to predict *Result*. Therefore, we have to set the role.
- As the target role, select *label*.
- Connect the *Decision Tree* with the *Set Role* module and run the process.

Chair of Process and Data Science

Solution





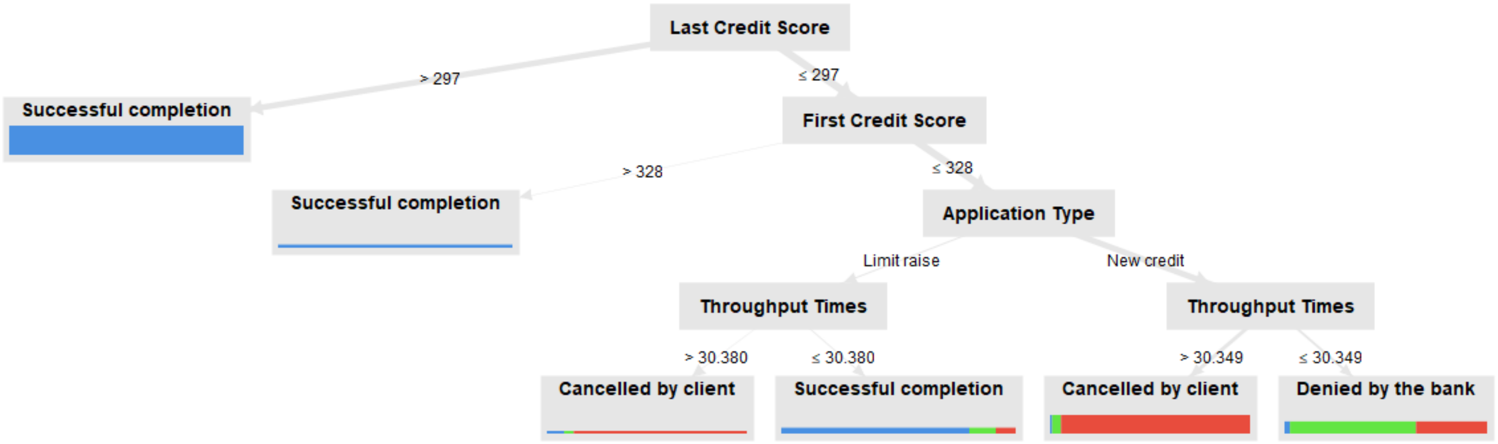
TOO BIG

Decision Tree – Parameters

- **For most modules, you can set different parameters.**
- **For now, set the maximal depth to 5**
- **Rerun the process**

Chair of Process and Data Science | RWTH AACHEN UNIVERSITY

Solution



Last Credit Score
> 297
≤ 297
Successful completion
First Credit Score
> 328
≤ 328
Successful completion
Application Type
Limit raise
New credit
Throughput Times
Throughput Times
> 30.380
≤ 30.380
> 30.349
≤ 30.349
Cancelled by client
Successful completion
Cancelled by client
Denied by the bank

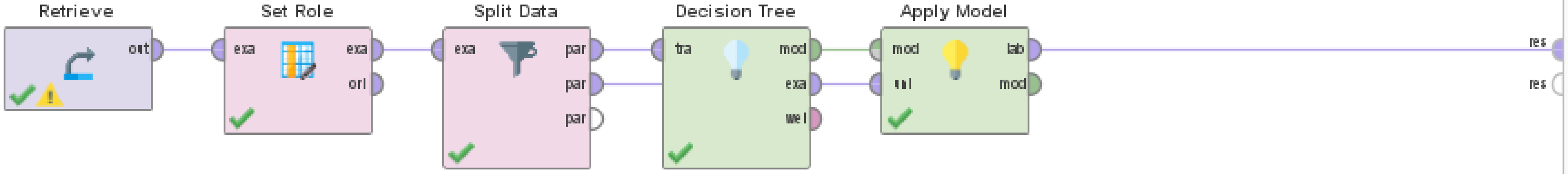
Decision Tree –Check results

- **To evaluate your results, you need to split the data. Otherwise, you will evaluate on data your model knows.**
- **To split data, use the *Split Data* module. Specify as partitions 0.8 and 0.2**
- **Next, apply the decision tree to the unseen data, using *Apply Model***
- **Run the process**

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Solution



Row No.	Results	prediction(Results)	confidence(Successful completion)	confidence(Denied by the bank)	confidence(Cancelled by client)
1	Successful completion	Successful completion	0.806	0.118	0.075
2	Successful completion	Successful completion	0.806	0.118	0.075
3	Successful completion	Successful completion	1	0	0
4	Successful completion	Successful completion	1	0	0
5	Successful completion	Successful completion	1	0	0
6	Successful completion	Successful completion	1	0	0
7	Denied by the bank	Cancelled by client	0.014	0.266	0.720
8	Cancelled by client	Cancelled by client	0.014	0.266	0.720
9	Successful completion	Cancelled by client	0.014	0.266	0.720

Chair of Process and Data Science

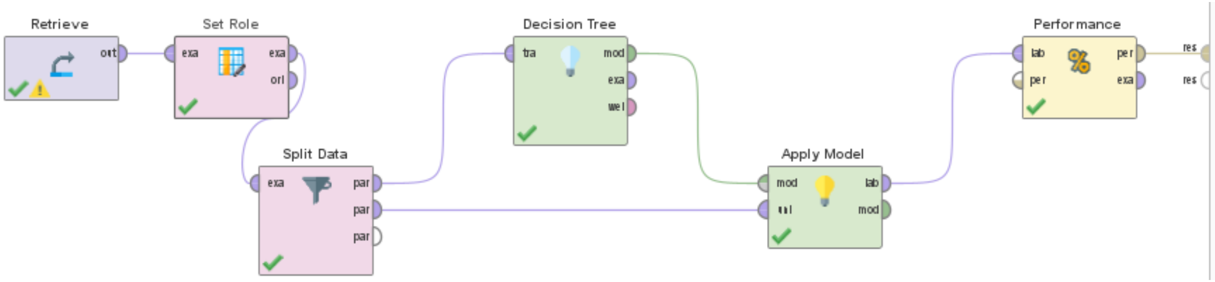
Decision Tree –Evaluating

- **Checking every data instance by hand is impossible.**
- **Use the *Performance* module and link it with the labeled data from the *Apply Model* module**
- **Rerun the process.**

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Solution



accuracy: 85.46%

	true Successful completion	true Denied by the bank	true Cancelled by client	class precision
pred. Successful completion	524	7	11	96.68%
pred. Denied by the bank	0	0	0	0.00%
pred. Cancelled by client	8	119	328	72.09%
class recall	98.50%	0.00%	96.76%	

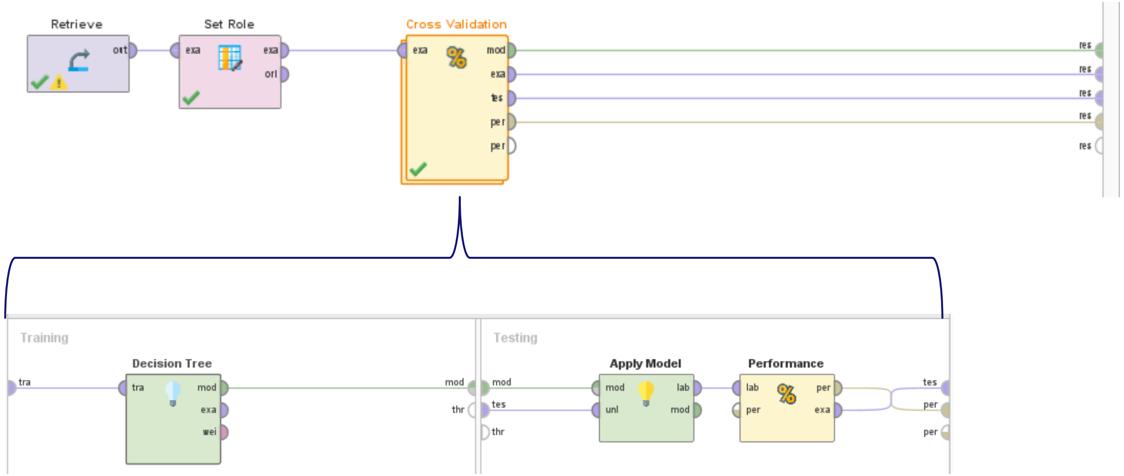
Chair of Process and Data Science

Decision Tree – Cross Validation

- Instead of doing the previous steps, you can also use the *Cross Validation* operator
- You only have three modules on the “main” process page: *Retrieve, Set Role,* and *Cross Validation.* Link the first four outputs of the operator to the process output
- By double-clicking on *Cross Validation,* you can see the operator’s subprocess
- For training, use the prev. decision tree, for testing, the operators *Apply Model* and *Performance*
- Rerun the process

Chair of Process and Data Science

Solution



accuracy: 90.04% +/- 2.08% (micro average: 90.04%)

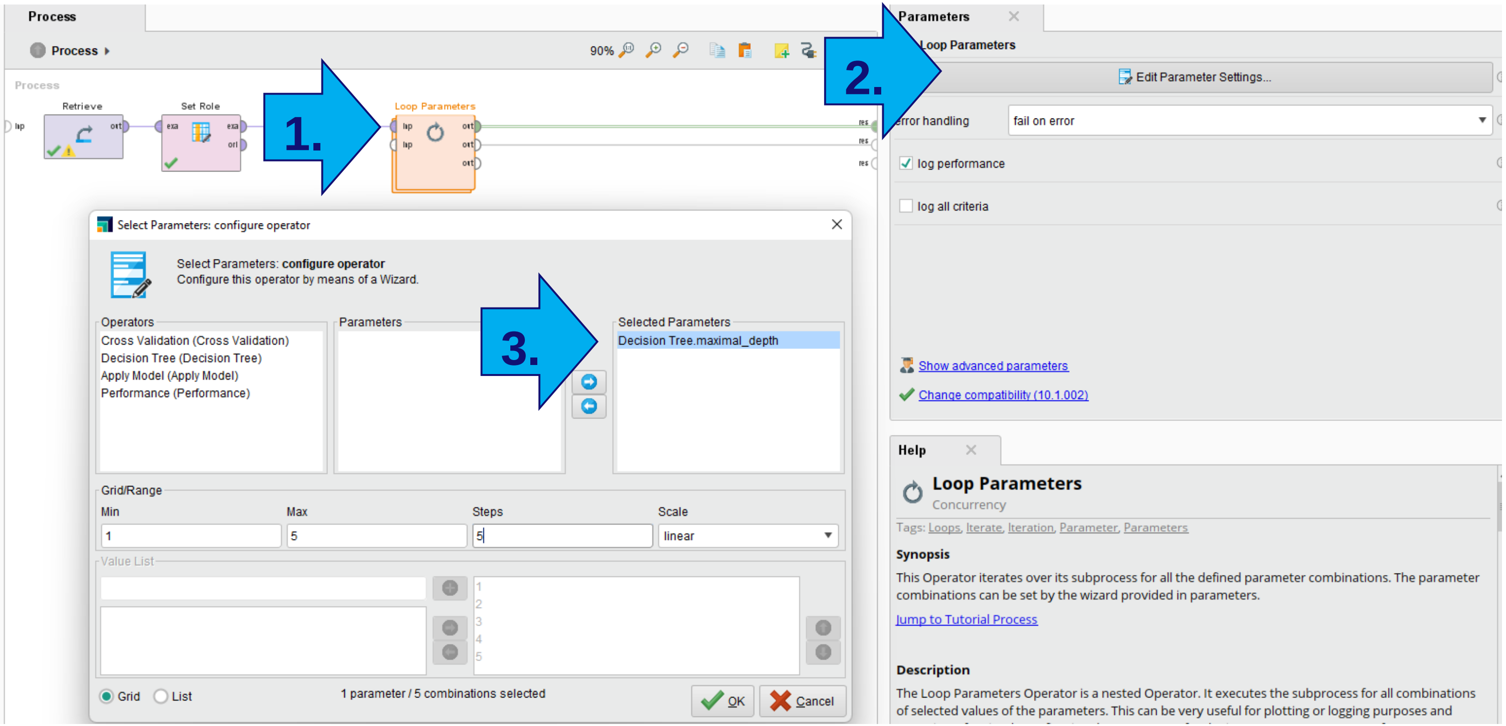
	true Successful completion	true Denied by the bank	true Cancelled by client	class precision
pred. Successful completion	2612	40	31	97.35%
pred. Denied by the bank	21	474	263	62.53%
pred. Cancelled by client	25	116	1400	90.85%
class recall	98.27%	75.24%	82.64%	

Chair of Process and Data Science

Decision Tree – Play with Parameters

- **As seen earlier, we can change multiple parameters.**
- **Instead of changing parameters one by one and rerunning the process, we can do that with an operator: *Loop parameters***
- **Drag *Cross Validation* into *Loop Parameters***
- **Setting for *Loop Parameters:* Selected parameter is maximal depth, min=1, max=5, in 5 steps**
- **Input for *Loop Parameters*: *Set Role***
- **Output for *Cross Validation: Model and performance***

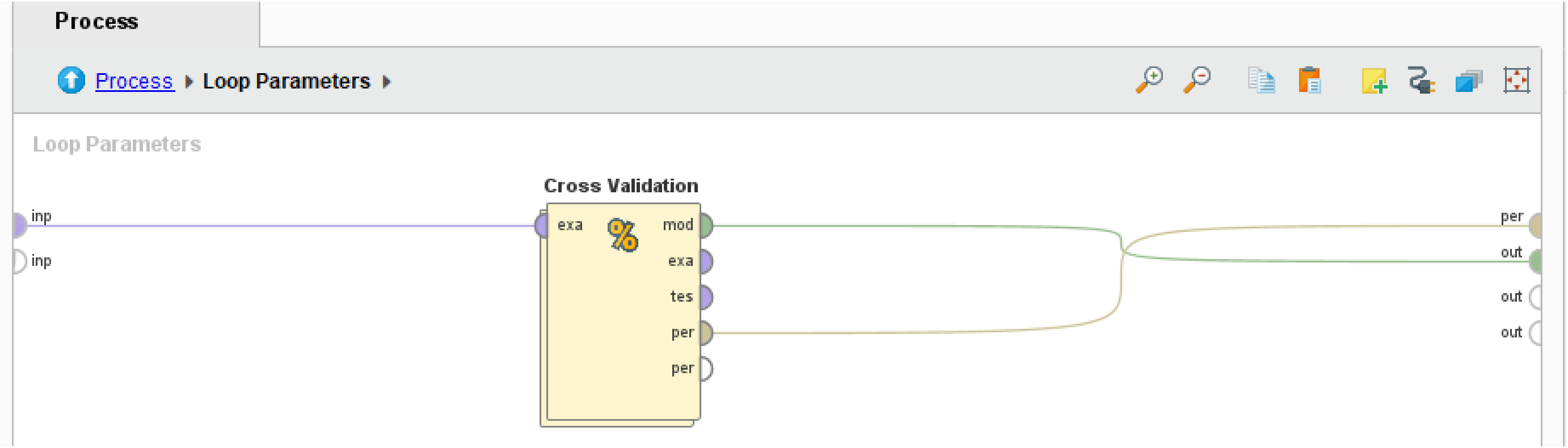
Solution



Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

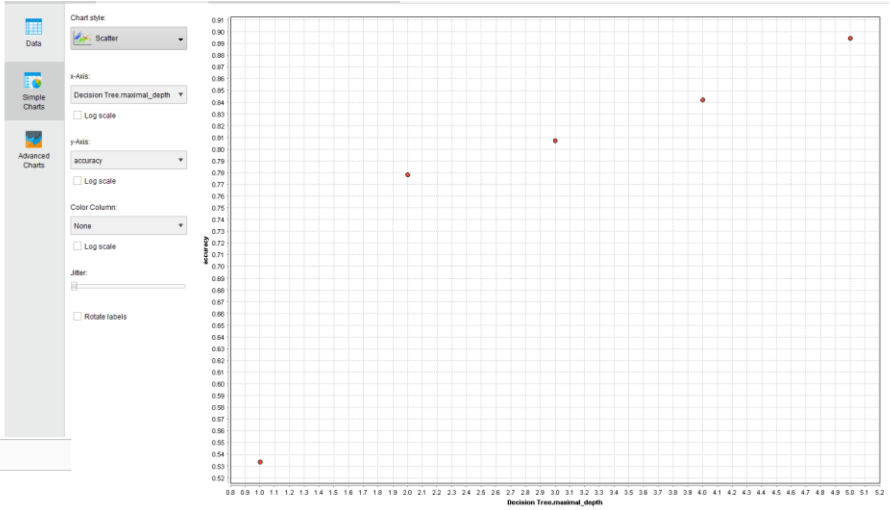
Solution



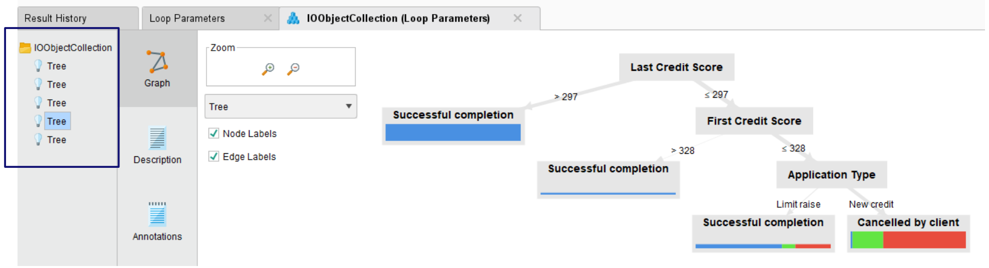
Process
Process
Loop Parameters
Loop Parameters
Cross Validation
inp
inp
exa
mod
exa
tes
per
per
per
out
out
out

Solution

iteration	Decision Tree.maximal_depth	accuracy ↑
1	1	0.534
2	2	0.778
3	3	0.807
4	4	0.842
5	5	0.895



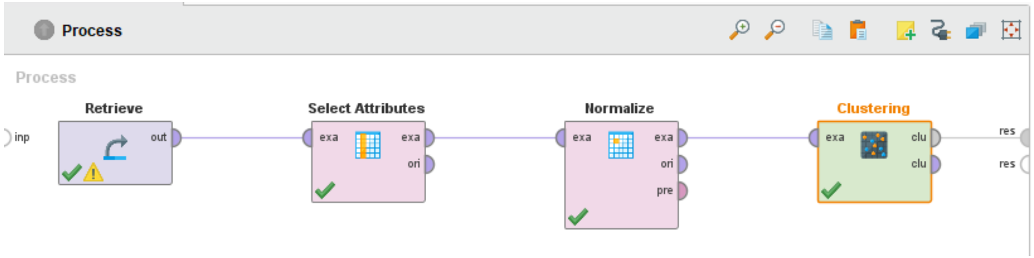
Inspect different trees:



Clustering – Getting Started

- Using *Select Attributes*, select the attributes *Numbers of Offers, Requested Amount, Throughput Times*
- Normalize the values using Z-transformation (advanced parameters)
- Let’s cluster similar instances. To group them, we use the operator *Clustering (k-Means)*, with 4 centroids
- Run the process

Solution

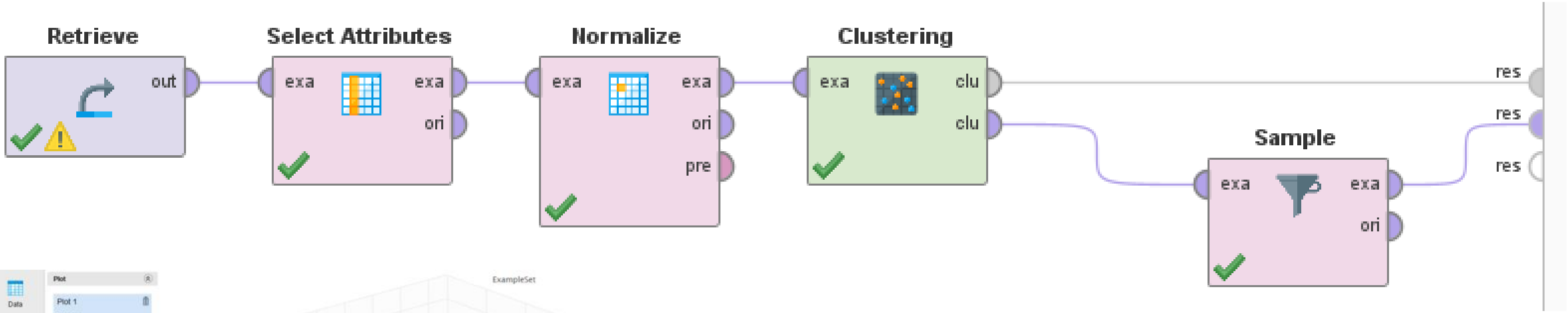


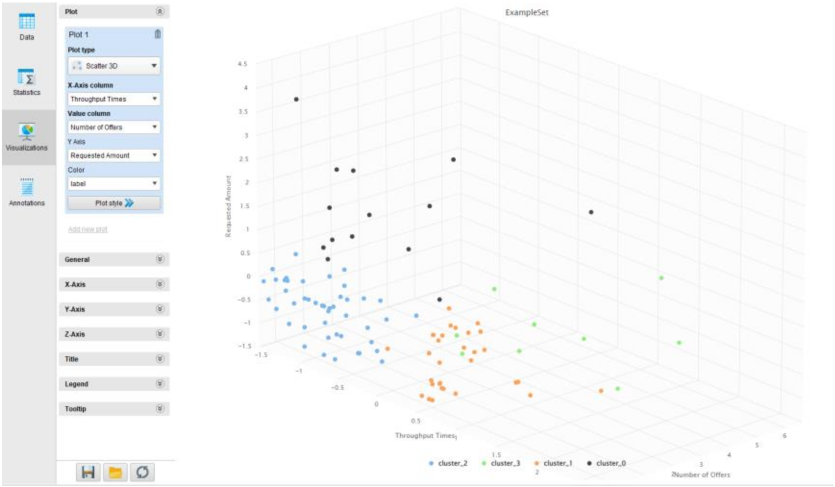
Attribute	cluster_0	cluster_1	cluster_2	cluster_3
Requested Amount	1.975	-0.281	-0.351	0.077
Number of Offers	-0.151	-0.128	-0.249	3.025
Throughput Times	0.016	0.923	-0.772	0.675

Clustering – Visualize Clusters

- **If we want to visualize the clusters, we cannot display every instance → Sampling is needed**
- **Connect the output clustered set with *Sample* operator. Set the sample size to 100 and the local random seed to 2023 (advanced parameter).**
- **Create a 3D-visualization of the clustered samples.**

Solution





Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

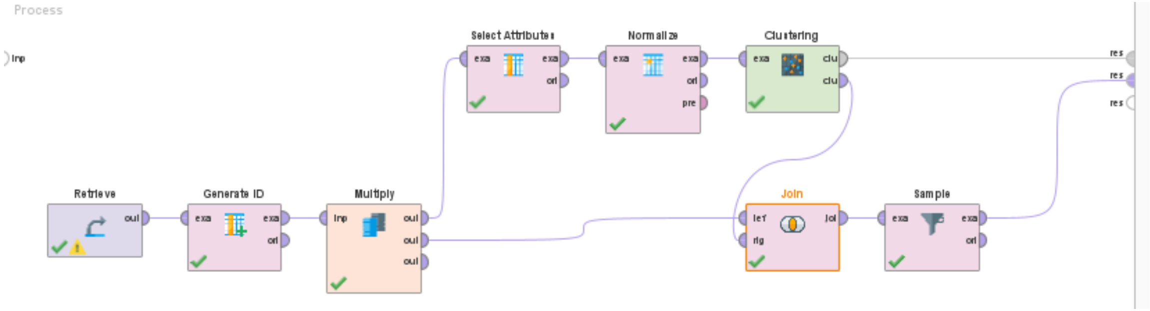
Clustering –Drawing Conclusions

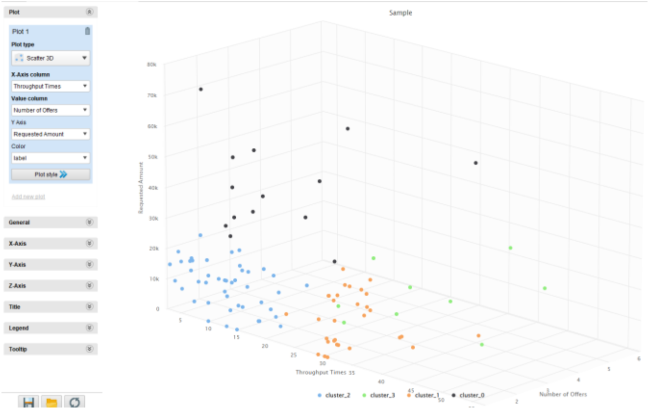
- **To draw conclusions related to unnormalized, or even original data, we have to join the clustered data with the original one**
- **To join data, we generate ids. Then we use the *Multiply* operator to multiply the original data with ids.**
- **Then, we run the process.**
- **Create a 3D-Visualization as before.**

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Solution





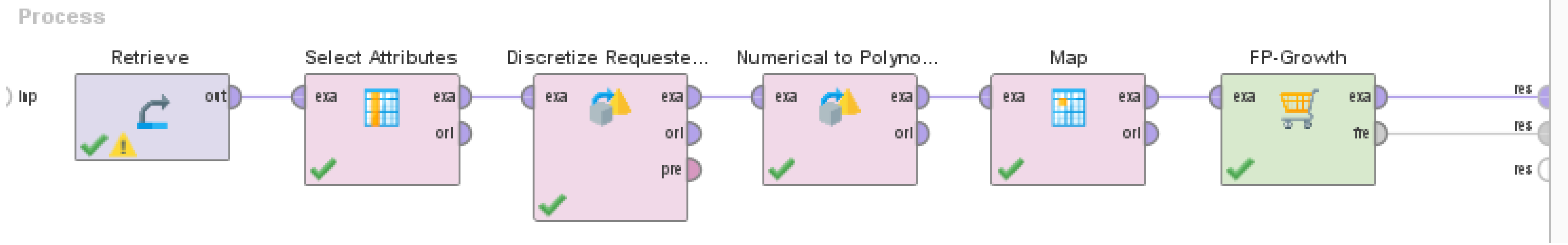
Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Frequent Itemsets

- We want to know which of the following variable values appear together: *Requested Amount, Having a Call to Complete* Application, Results
- In this process, we discretize the variables:
  - Requested amount:
    - Upper limit 9000: low request
    - Upper limit 20,000: medium request
    - Upper limit Infinity: High request
  - Having a Call to Complete Application:
    - Two operators are needed: *Numerical to Polynomial, Map*
    - If there was a call, denote it with "called," otherwise with "not called"
- Use the module *FP-Growth*, by setting the frequency to 100

Solution



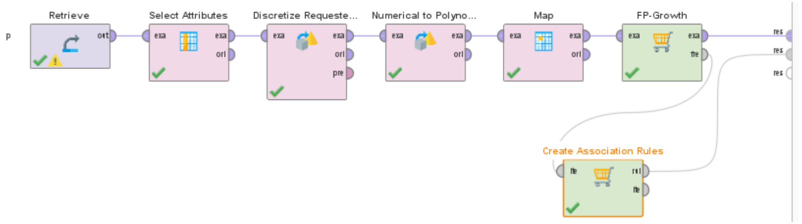
Size	Support ↓	Item 1	Item 2	Item 3
1	0.915	Not called		
1	0.534	Successful completion		
2	0.474	Not called	Successful completion	
1	0.377	medium request		
1	0.362	low request		
2	0.348	Not called	medium request	
1	0.340	Cancelled by client		

Association Rules

- Add to your former process the module *Create Association Rules*
- Set the minimum confidence to 0.8
- Connect the frequent set output with this module and run the process

Chair of Process and Data Science
RWTH AACHEN UNIVERSITY

Solution



No.	Premises	Conclusion	Support	Confiden...	LaPlace	Gain	p-s	Lift	Convicti...
15	low request, Denied by the bank	Not called	0.047	0.967	0.998	-0.050	0.003	1.056	2.562
14	medium request, Cancelled by client	Not called	0.118	0.956	0.995	-0.129	0.005	1.045	1.929
13	Cancelled by client	Not called	0.322	0.948	0.987	-0.358	0.011	1.036	1.631
12	low request, Cancelled by client	Not called	0.124	0.945	0.994	-0.138	0.004	1.032	1.539

Chair of Process and Data Science