Orange Labs

September 2022

Tutorial

KHIOPS 10.1

KHIOPS & KHIOPS VISUALIZATION

KHIOPS COCLUSTERING & KHIOPS COVISUALIZATION

MULTI-TABLE FUNCTIONALITIES

Khiops

- Khiops
  - 3 Optimal data preparation based on discretization and value grouping
  - Scoring models for classification and regression
  - Correlation analysis between pairs of variables
- Khiops Visualization
  - Analysis of Khiops results using an interactive visualization tool
- Khiops Coclustering
  - 36 Correlation analysis of two or more variables using a hierarchical coclustering model
- Khiops Covisualization
  - Exploratory analysis of Khiops Coclustering results using an interactive visualization tool
- Multi-table functionalities
  - Multi-table database
  - Automatic feature construction
  - 60 Multi-table functionalities in Khiops and Khiops Coclustering

Khiops & Khiops Visualization

- Khiops
  - Optimal data preparation based on discretization and value grouping
  - Scoring models for classification and regression
  - Correlation analysis between pairs of variables

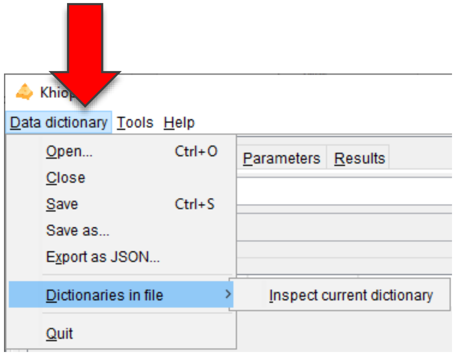
- Khiops Visualization
  - Analysis of Khiops results using an interactive visualization tool

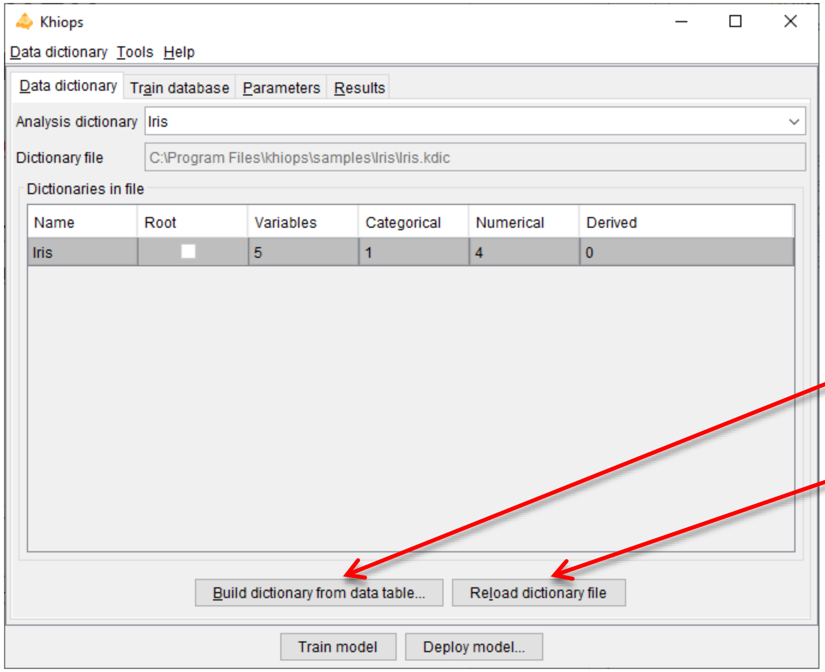
Supervised classification

- **Step 1 :** Open an existing dictionary file

  (ex: sample Iris.kdic)

  - Dictionary file: contains one or more dictionaries
  - Dictionary: description of variables of a database to use during analysis





Analysis dictionary: Iris

Dictionary file: C:\Program Files\khiops\samples\Iris\Iris.kdic

Dictionaries in file

Name	Root	Variables	Categorical	Numerical	Derived
Iris	□	5	1	4	0

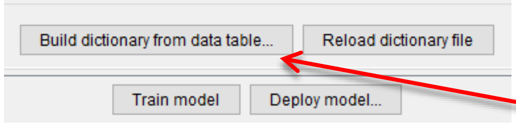
Available actions :

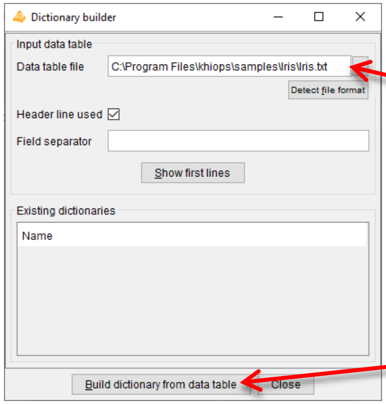
- Open, Save, Save as, Close
- Edition (menu « *Dictionary file/Inspect current dictionary* », or NotePad)
- Build dictionary from data table
- Reload dictionary file
  - useful if it has been modified from an external editor

Supervised classification

- **Step 1, bis :** Build a new dictionary from a data table

(If no available dictionary)




Dictionary builder

Input data table

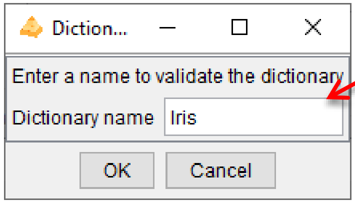
Data table file: C:\Program Files\khiops\samples\Iris\Iris.txt

Header line used ☑

Field separator:

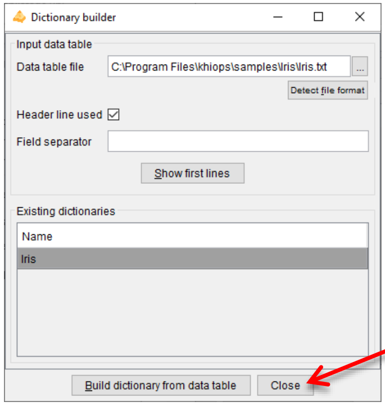
Existing dictionaries

Name


Diction...

Enter a name to validate the dictionary

Dictionary name: Iris


Dictionary builder

Input data table

Data table file: C:\Program Files\khiops\samples\Iris\Iris.txt

Header line used ☑

Field separator:

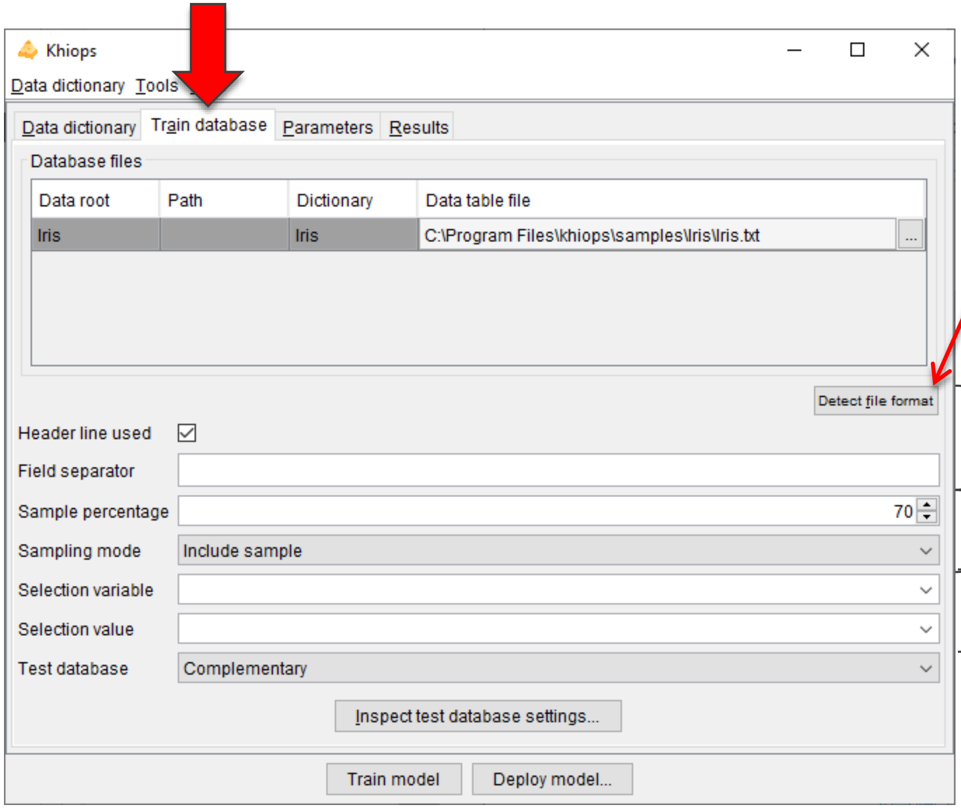
Existing dictionaries

Name
Iris

- 1. In pane Data dictionary
  - Click on button Build dictionary from data table...
- 2. In the dialog box
  - Specify the data table file
  - Build the dictionary
- 3. Specify the dictionary name
  - Repeat step 2 to build other dictionaries
- 4. Close to save dictionary file
- 5. In pane Data dictionary
  - Inspect the built dictionary
    - Check variables types
    - Select used variables

Supervised classification

- **Step 2 : Specify train database**



Khiops
Data dictionary Tools
Data dictionary | Train database | Parameters | Results
Detect file format
Inspect test database settings...
Train model
Deploy model...

Database files

Data root	Path	Dictionary	Data table file
Iris		Iris	C:\Program Files\khiops\samples\Iris\Iris.txt

Header line used ☑

Field separator

Sample percentage 70

Sampling mode Include sample

Selection variable

Selection value

Test database Complementary

Detect file format : heuristic help that scans the first few lines to guess the file format. The header line and field separator are updated on success, with a warning or an error in the log window only if necessary.

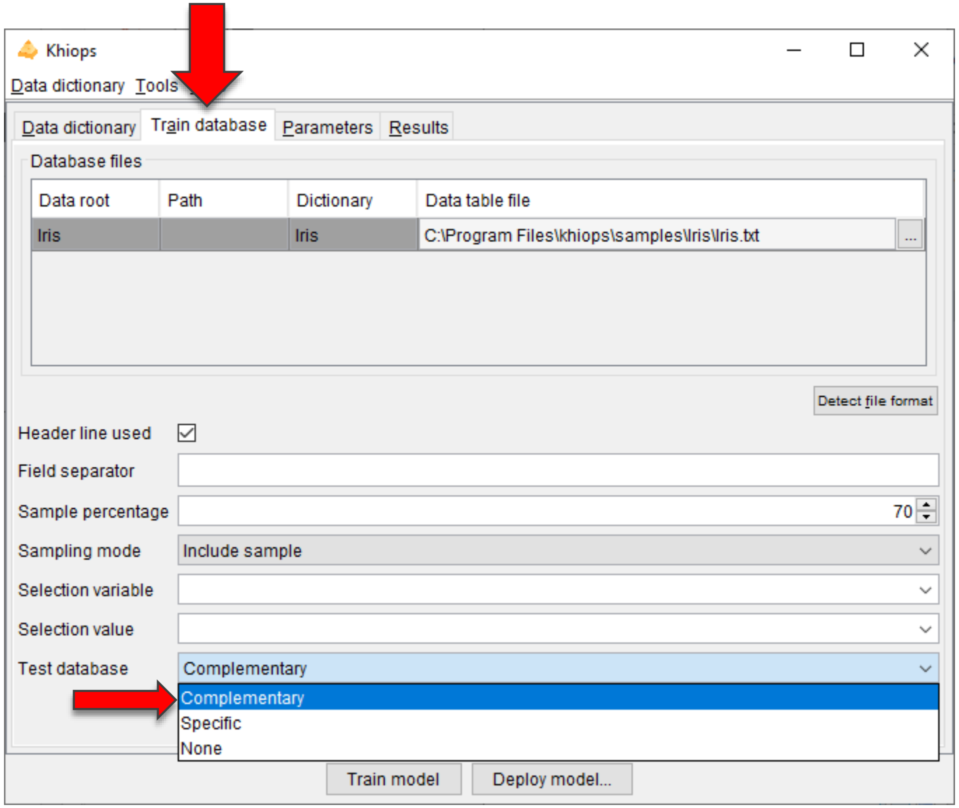
File Format

Sample percentage : default 70%

Controlled way of selecting the instances by the means of a selection variable and selection value

Supervised classification

- **Step 2, bis :** Specify test database



Data root	Path	Dictionary	Data table file
Iris		Iris	C:\Program Files\khiops\samples\Iris\Iris.txt

Header line used ☑

Field separator

Sample percentage 70

Sampling mode Include sample

Selection variable

Selection value

Test database Complementary

Complementary
Specific
None

Three possibilities :

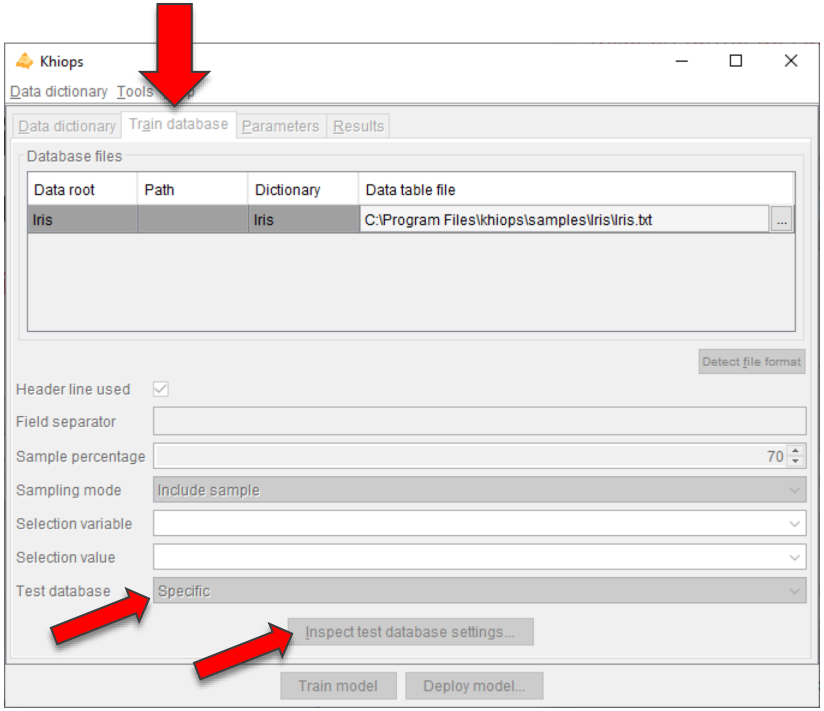
Complementary
Specific
None

Complementary (default)

The test database is the complementary of the train database according to the chosen sample percentage

Supervised classification

- **Step 2, ter :** Specify test database



Khiops

Database files

Data root	Path	Dictionary	Data table file
Iris		Iris	C:\Program Files\khiops\samples\Iris\Iris.txt

Header line used ☑

Field separator

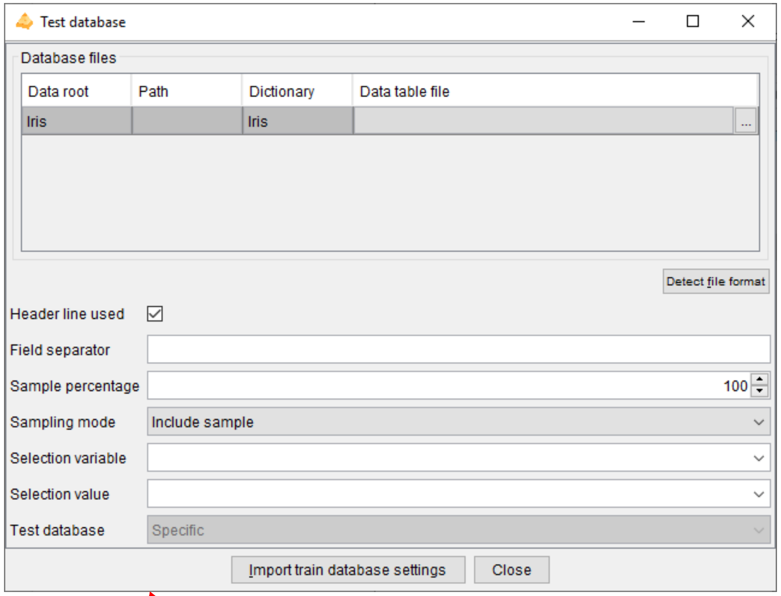
Sample percentage 70

Sampling mode Include sample

Selection variable

Selection value

Test database Specific



Test database

Database files

Data root	Path	Dictionary	Data table file
Iris		Iris	

Header line used ☑

Field separator

Sample percentage 100

Sampling mode Include sample

Selection variable

Selection value

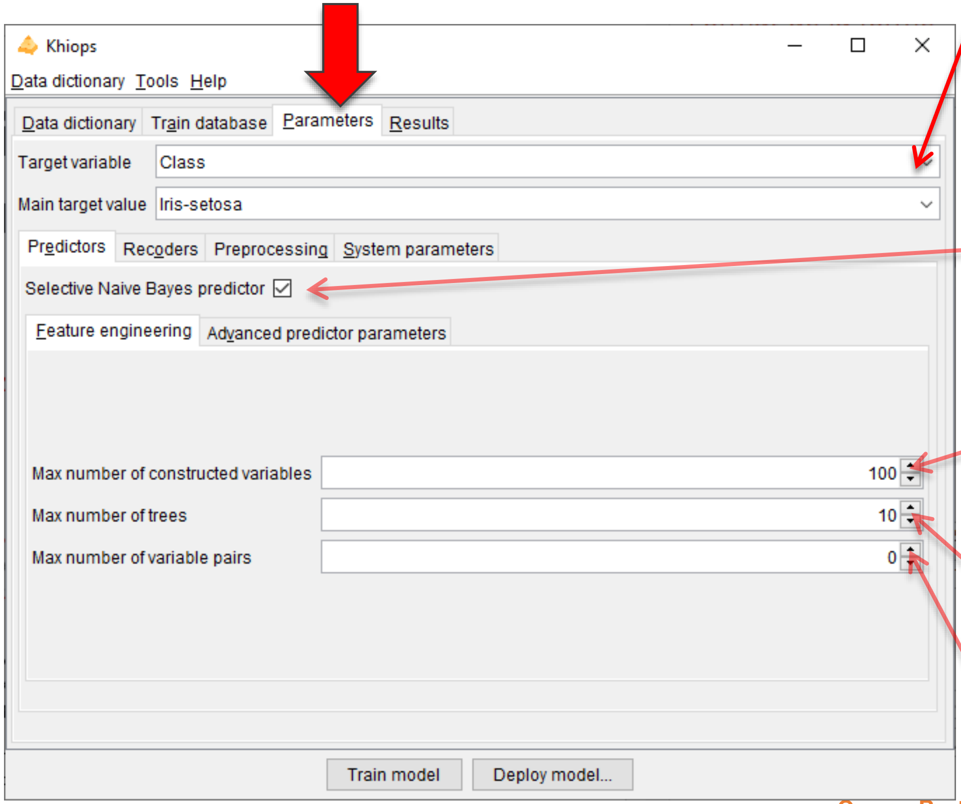
Test database Specific

Specific

The test database has it own independent specification : specific file, sampling, selection

Supervised classification

- **Step 3 :** Parameters


Khiops
Data dictionary Tools Help
Data dictionary Train database Parameters Results

Target variable: Class

Main target value: Iris-setosa

Predictors Recoders Preprocessing System parameters

Selective Naive Bayes predictor ☑

Feature engineering Advanced predictor parameters

Max number of constructed variables: 100

Max number of trees: 10

Max number of variable pairs: 0

Train model
Deploy model...

Type of selected target variable implies type of analysis

Categorical -> supervised classification
Numerical -> regression
Empty -> unsupervised analysis

Selective Naïve Bayes predictor

default true, to be set to false if only data preparation is wanted (without modeling)

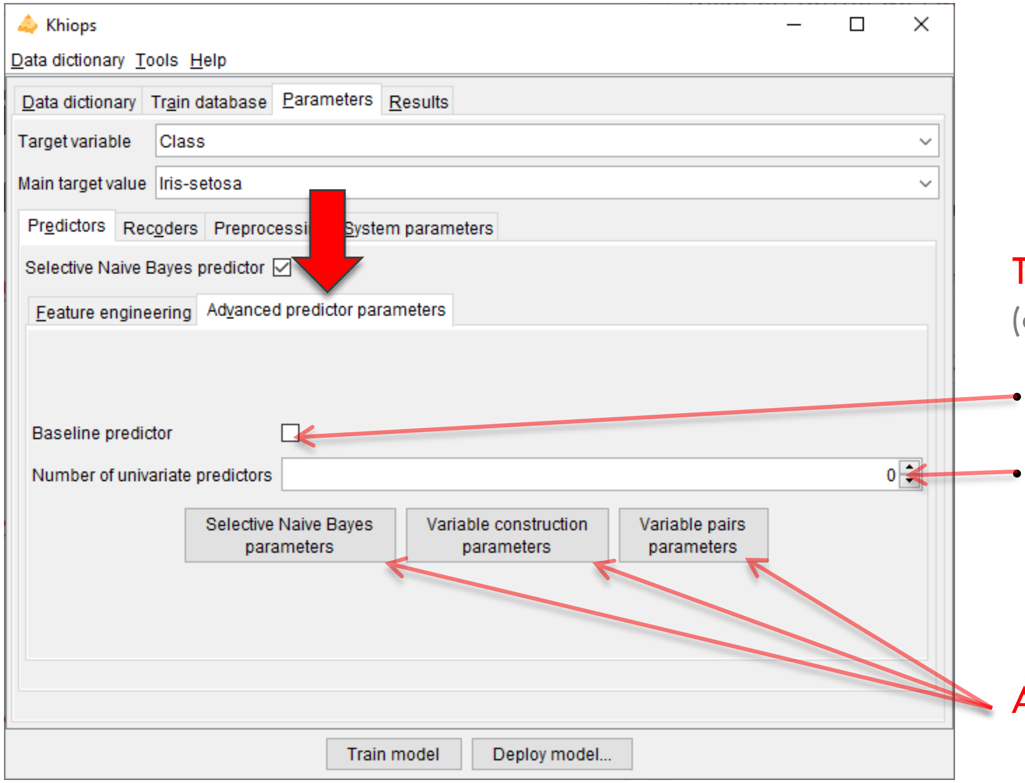
Constructed variables are computed in multi-table schema and allow to extract numerical or categorical values resulting from computing formula applied to existing variable (default 100)

The constructed trees allow to combine variables, either native or constructed (default 10)

The pairs of variables are analyzed during data preparation using a bivariate discretization method (default 0)

Supervised classification

- **Step 3 bis :** Advanced predictor parameters (optional)



Target variable: Class

Main target value: Iris-setosa

Selective Naive Bayes predictor ☑

Baseline predictor □

Number of univariate predictors: 0

Khiops
Data dictionary Tools Help
Data dictionary
Train database
Parameters
Results
Predictors
Recoders
Preprocessi
System parameters
Feature engineering
Advanced predictor parameters
Selective Naive Bayes parameters
Variable construction parameters
Variable pairs parameters
Train model
Deploy model...

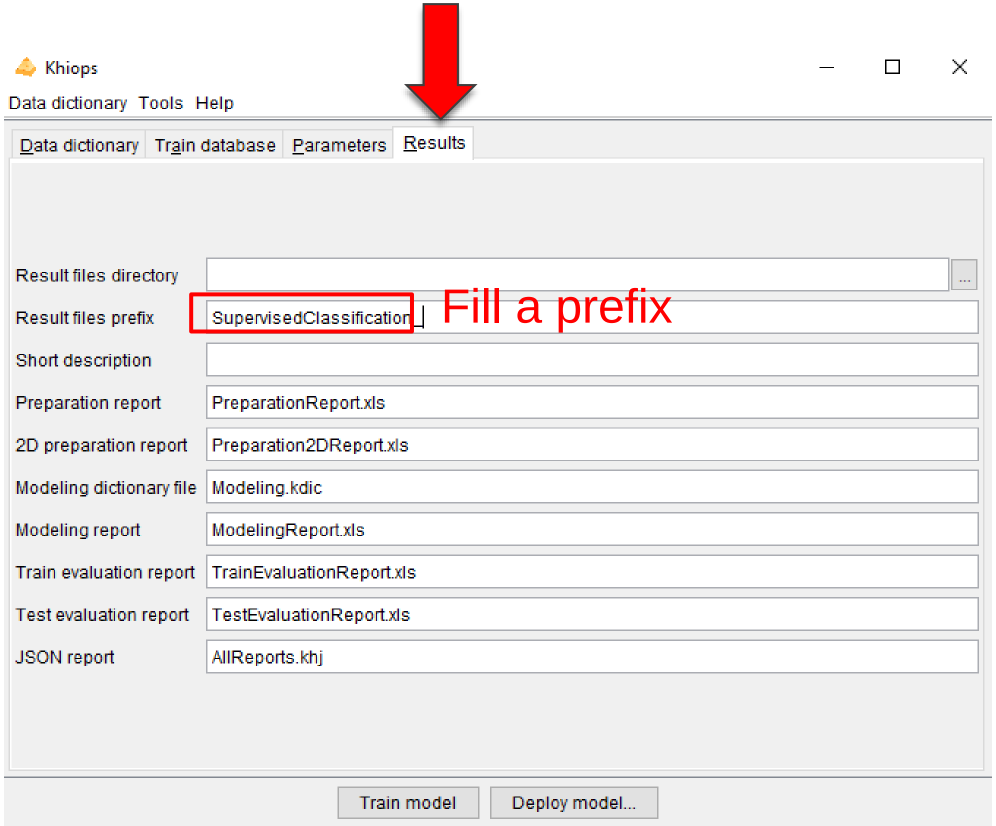
Two other optional predictors
(only in the supervised case)

- **Baseline :** prediction of the majority class (default false)
- **Univariate:** predictors exploiting one single variable (default none)

Advanced parameters to inspect

Supervised classification

- **Step 4 :** Results

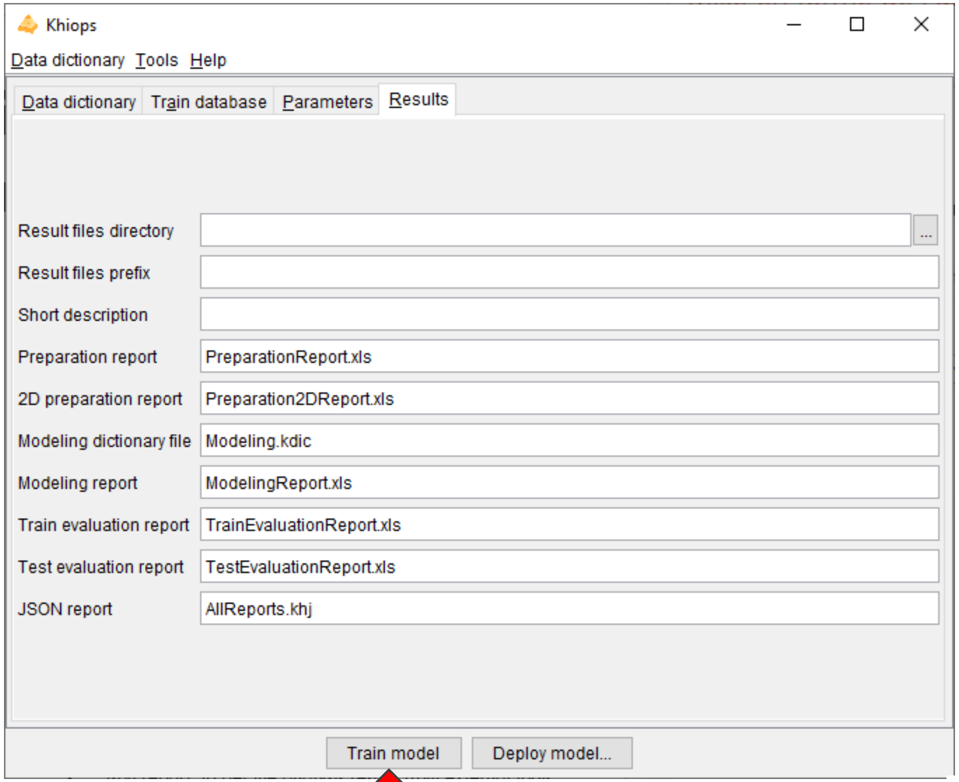


Result files directory	
Result files prefix	SupervisedClassification_
Short description	
Preparation report	PreparationReport.xls
2D preparation report	Preparation2DReport.xls
Modeling dictionary file	Modeling.kdic
Modeling report	ModelingReport.xls
Train evaluation report	TrainEvaluationReport.xls
Test evaluation report	TestEvaluationReport.xls
JSON report	AllReports.khj

- Directory where all results files are written
- Prefix (ex: in case of several experiments)
- Brief description to summarize the current analysis
- Description of trained univariate preparation models
- Description of trained bivariate preparation models
- Technical description for deployment purposes
- Description of trained models with selected variables
- Evaluation on train database
- Evaluation on test database
- Json report, to get the analysis results from external tools

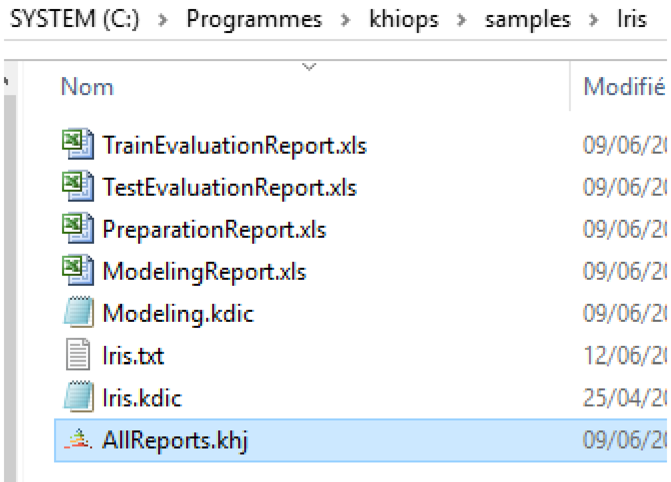
Supervised classification

- **Step 5 : Start the analysis**



Result files directory	
Result files prefix	
Short description	
Preparation report	PreparationReport.xls
2D preparation report	Preparation2DReport.xls
Modeling dictionary file	Modeling.kdic
Modeling report	ModelingReport.xls
Train evaluation report	TrainEvaluationReport.xls
Test evaluation report	TestEvaluationReport.xls
JSON report	AllReports.khj

1 – Train model


SYSTEM (C:) › Programmes › khiops › samples › Iris

Nom	Modifié
TrainEvaluationReport.xls	09/06/2
TestEvaluationReport.xls	09/06/2
PreparationReport.xls	09/06/2
ModelingReport.xls	09/06/2
Modeling.kdic	09/06/2
Iris.txt	12/06/2
Iris.kdic	25/04/2
AllReports.khj	09/06/2

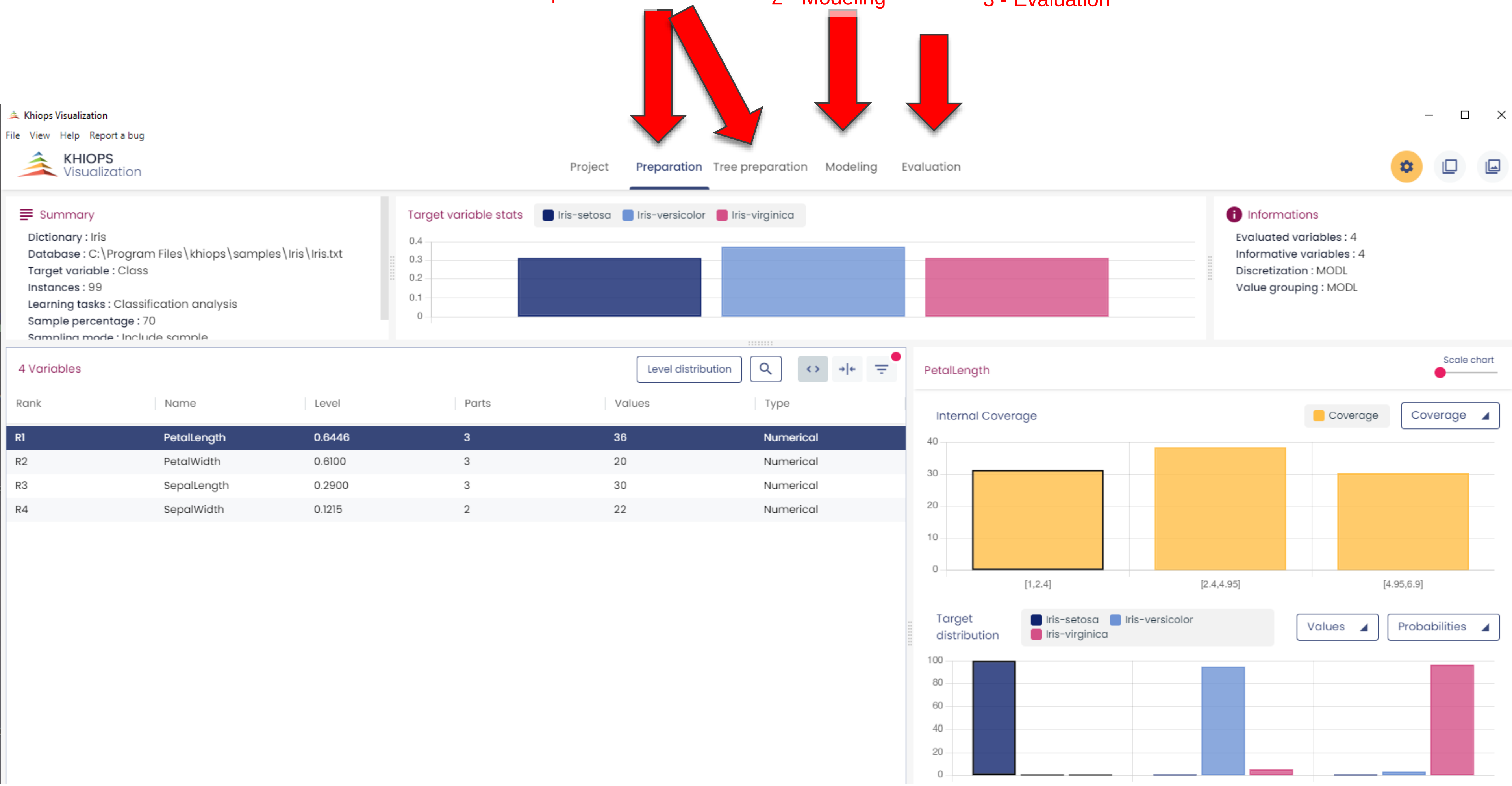
2 - Inspect the results using Khiops Visualization
(double-click on *.khj* file)

Exploratory of classification results using Khiops Visualization

1 - Preparation

2 - Modeling

3 - Evaluation



Summary

Dictionary : Iris
Database : C:\Program Files\khiops\samples\Iris\Iris.txt
Target variable : Class
Instances : 99
Learning tasks : Classification analysis
Sample percentage : 70
Sampling mode : Include sample

Informations

Evaluated variables : 4
Informative variables : 4
Discretization : MODL
Value grouping : MODL

4 Variables

Rank	Name	Level	Parts	Values	Type
R1	PetalLength	0.6446	3	36	Numerical
R2	PetalWidth	0.6100	3	20	Numerical
R3	SepalLength	0.2900	3	30	Numerical
R4	SepalWidth	0.1215	2	22	Numerical

Exploratory of classification results using Khiops Visualization

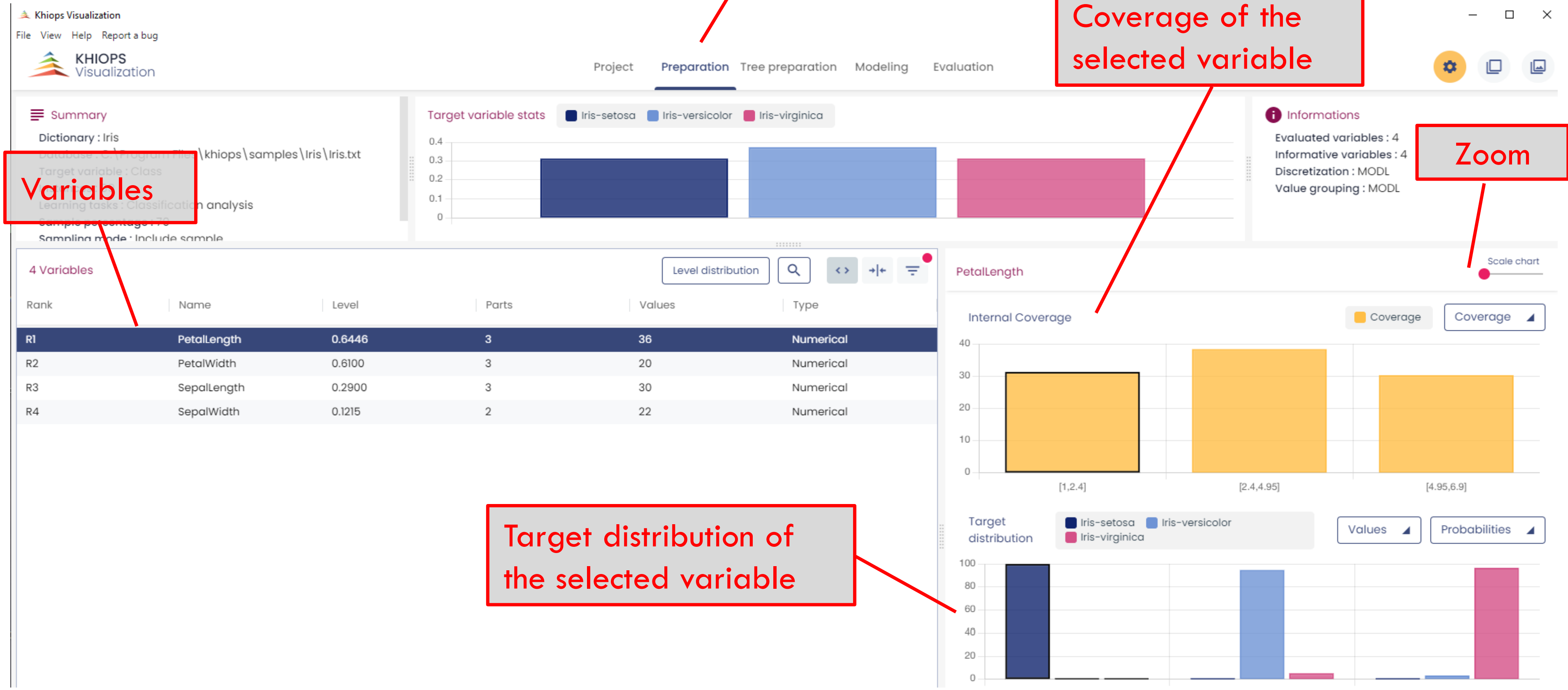
Preparation pane

Coverage of the selected variable

Zoom

Variables

Target distribution of the selected variable



Summary

Dictionary : Iris

Database : C:\Program Files\khiops\samples\Iris\Iris.txt

Target variable : Class

Instances : 150

Learning tasks : Classification analysis

Sample percentage : 70

Sampling mode : Include sample

Informations

Evaluated variables : 4

Informative variables : 4

Discretization : MODL

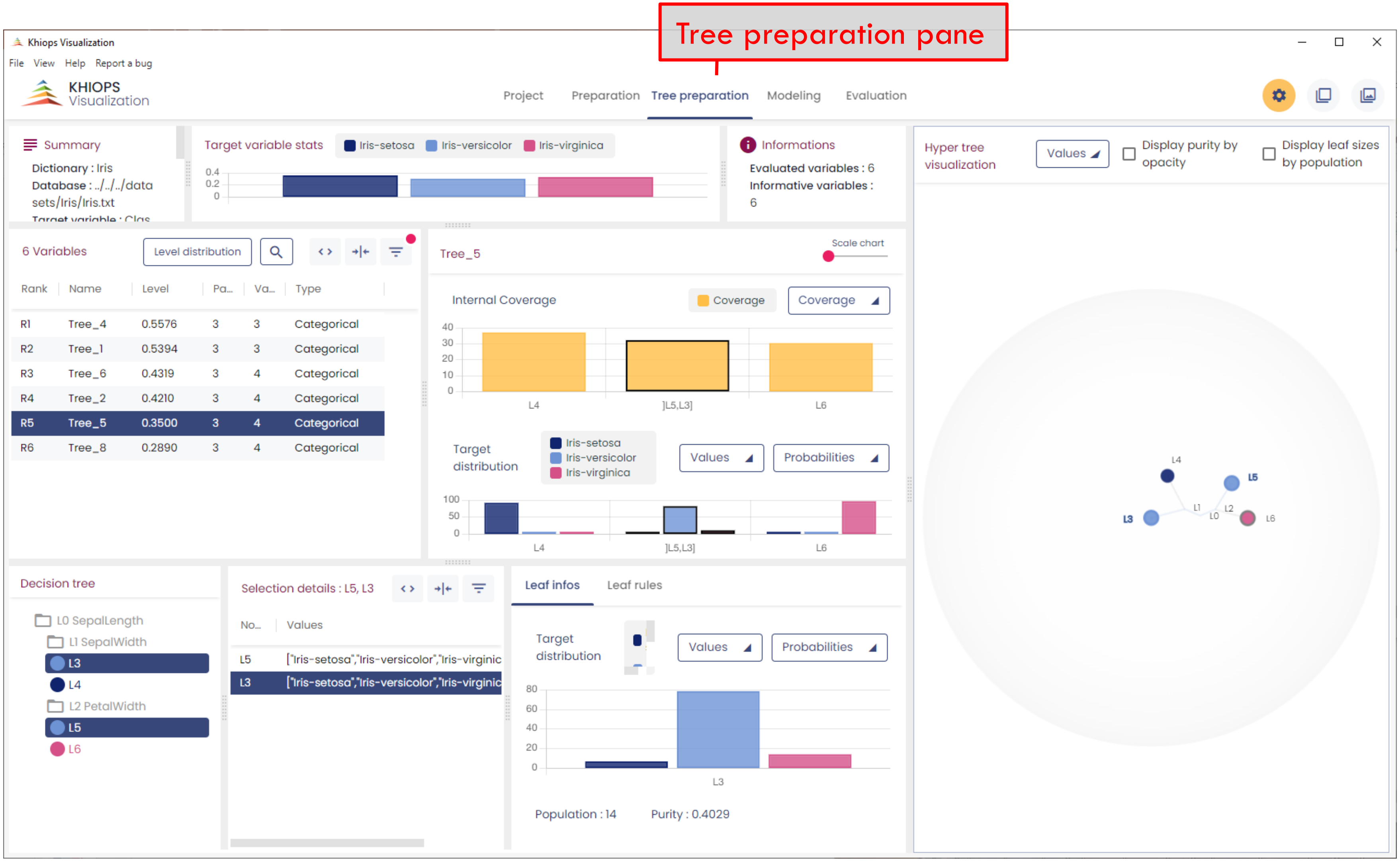
Value grouping : MODL

4 Variables

Rank	Name	Level	Parts	Values	Type
R1	PetalLength	0.6446	3	36	Numerical
R2	PetalWidth	0.6100	3	20	Numerical
R3	SepalLength	0.2900	3	30	Numerical
R4	SepalWidth	0.1215	2	22	Numerical

Exploratory of classification results using Khiops Visualization

Tree preparation pane



Summary

Dictionary : Iris
Database : ../../../data sets/Iris/Iris.txt
Target variable : Clas

Informations

Evaluated variables : 6
Informative variables : 6

6 Variables

Rank	Name	Level	Pa...	Va...	Type
R1	Tree_4	0.5576	3	3	Categorical
R2	Tree_1	0.5394	3	3	Categorical
R3	Tree_6	0.4319	3	4	Categorical
R4	Tree_2	0.4210	3	4	Categorical
R5	Tree_5	0.3500	3	4	Categorical
R6	Tree_8	0.2890	3	4	Categorical

Tree_5

Decision tree

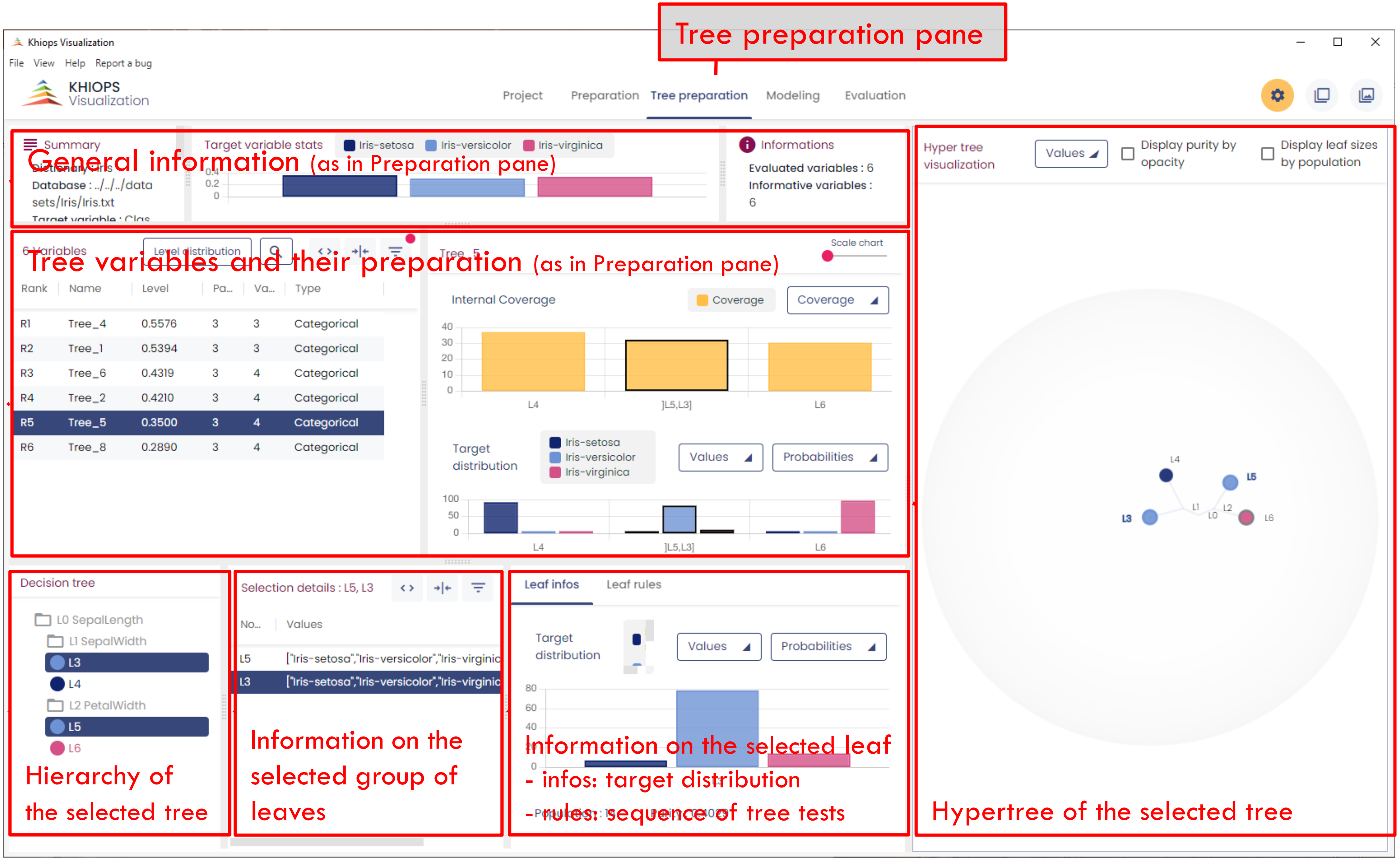
- L0 SepalLength
  - L1 SepalWidth
    - L3
    - L4
  - L2 PetalWidth
    - L5
    - L6

Selection details : L5, L3

No...	Values
L5	["Iris-setosa","Iris-versicolor","Iris-virginic
L3	["Iris-setosa","Iris-versicolor","Iris-virginic

Population : 14 Purity : 0.4029

Exploratory of classification results using Khiops Visualization



Tree preparation pane

General information (as in Preparation pane)

Tree variables and their preparation (as in Preparation pane)

Hierarchy of the selected tree

Information on the selected group of leaves

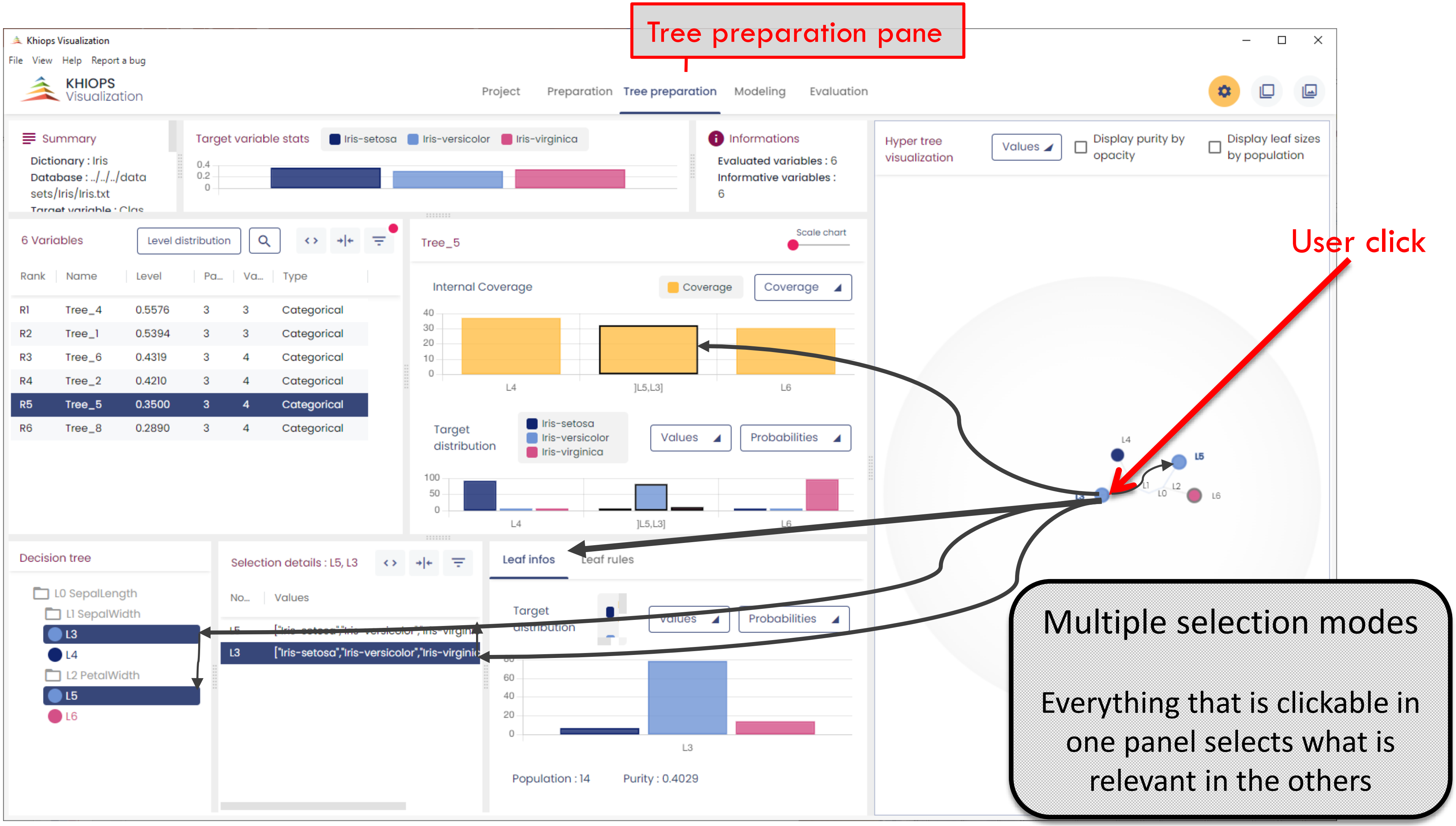
Information on the selected leaf
- infos: target distribution
- rules: sequence of tree tests

Hypertree of the selected tree

Rank	Name	Level	Pa...	Va...	Type
R1	Tree_4	0.5576	3	3	Categorical
R2	Tree_1	0.5394	3	3	Categorical
R3	Tree_6	0.4319	3	4	Categorical
R4	Tree_2	0.4210	3	4	Categorical
R5	Tree_5	0.3500	3	4	Categorical
R6	Tree_8	0.2890	3	4	Categorical

No...	Values
L5	["Iris-setosa","Iris-versicolor","Iris-virginic
L3	["Iris-setosa","Iris-versicolor","Iris-virginic

Exploratory of classification results using Khiops Visualization



Tree preparation pane

User click

Multiple selection modes

Everything that is clickable in one panel selects what is relevant in the others

Khiops Visualization

File View Help Report a bug

KHIOPS Visualization

Project Preparation **Tree preparation** Modeling Evaluation

Summary

Dictionary : Iris
Database : ../../../data sets/Iris/Iris.txt
Target variable : Clas

Informations

Evaluated variables : 6
Informative variables : 6

6 Variables

Rank	Name	Level	Pa...	Va...	Type
R1	Tree_4	0.5576	3	3	Categorical
R2	Tree_1	0.5394	3	3	Categorical
R3	Tree_6	0.4319	3	4	Categorical
R4	Tree_2	0.4210	3	4	Categorical
R5	Tree_5	0.3500	3	4	Categorical
R6	Tree_8	0.2890	3	4	Categorical

Decision tree

- L0 SepalLength
  - L1 SepalWidth
    - L3
    - L4
  - L2 PetalWidth
    - L5
    - L6

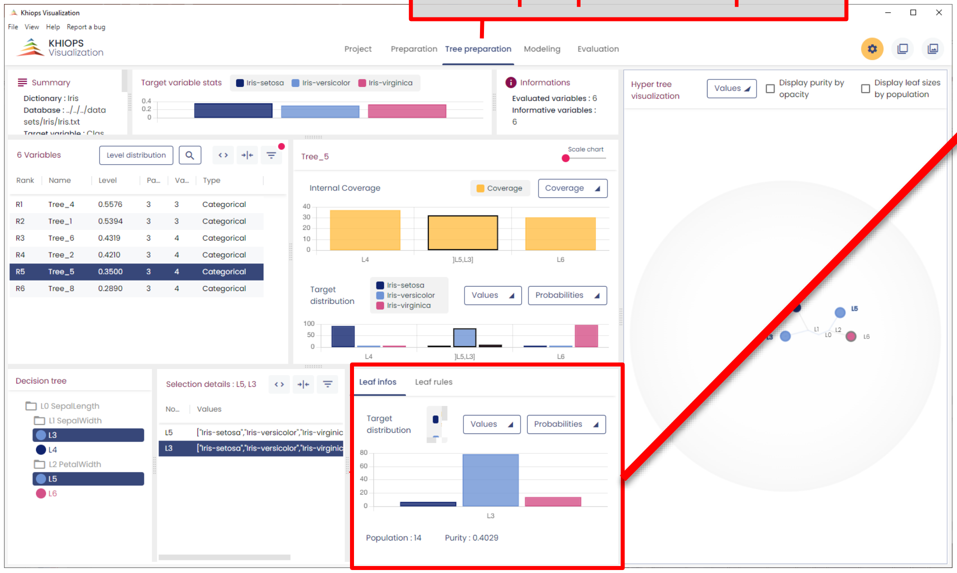
Selection details : L5, L3

No...	Values
L5	["Iris-setosa","Iris-versicolor","Iris-virgin
L3	["Iris-setosa","Iris-versicolor","Iris-virginic

Population : 14 Purity : 0.4029

Exploratory of classification results using Khiops Visualization

Tree preparation pane



Summary
Dictionary : Iris
Database : ../../../data sets/Iris/Iris.txt
Target variable : Clas

Informations
Evaluated variables : 6
Informative variables : 6

6 Variables

Rank	Name	Level	Pa...	Va...	Type
R1	Tree_4	0.5576	3	3	Categorical
R2	Tree_1	0.5394	3	3	Categorical
R3	Tree_6	0.4319	3	4	Categorical
R4	Tree_2	0.4210	3	4	Categorical
R5	Tree_5	0.3500	3	4	Categorical
R6	Tree_8	0.2890	3	4	Categorical

- L0 SepalLength
  - L1 SepalWidth
    - L3
    - L4
  - L2 PetalWidth
    - L5
    - L6

Selection details : L5, L3

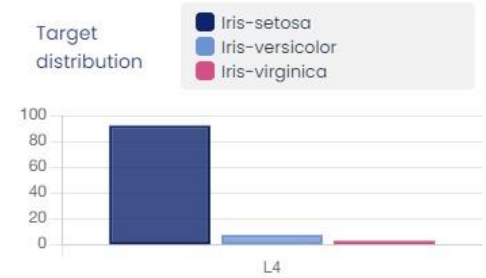
No...	Values
L5	["Iris-setosa","Iris-versicolor","Iris-virginic
L3	["Iris-setosa","Iris-versicolor","Iris-virginic

Population : 14 Purity : 0.4029

Information on selected leaf

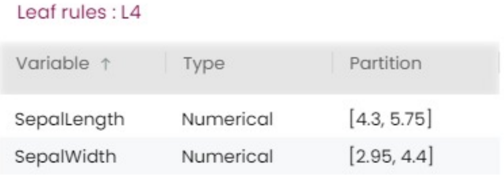
Leaf infos

Target distribution in leaf



Leaf rules

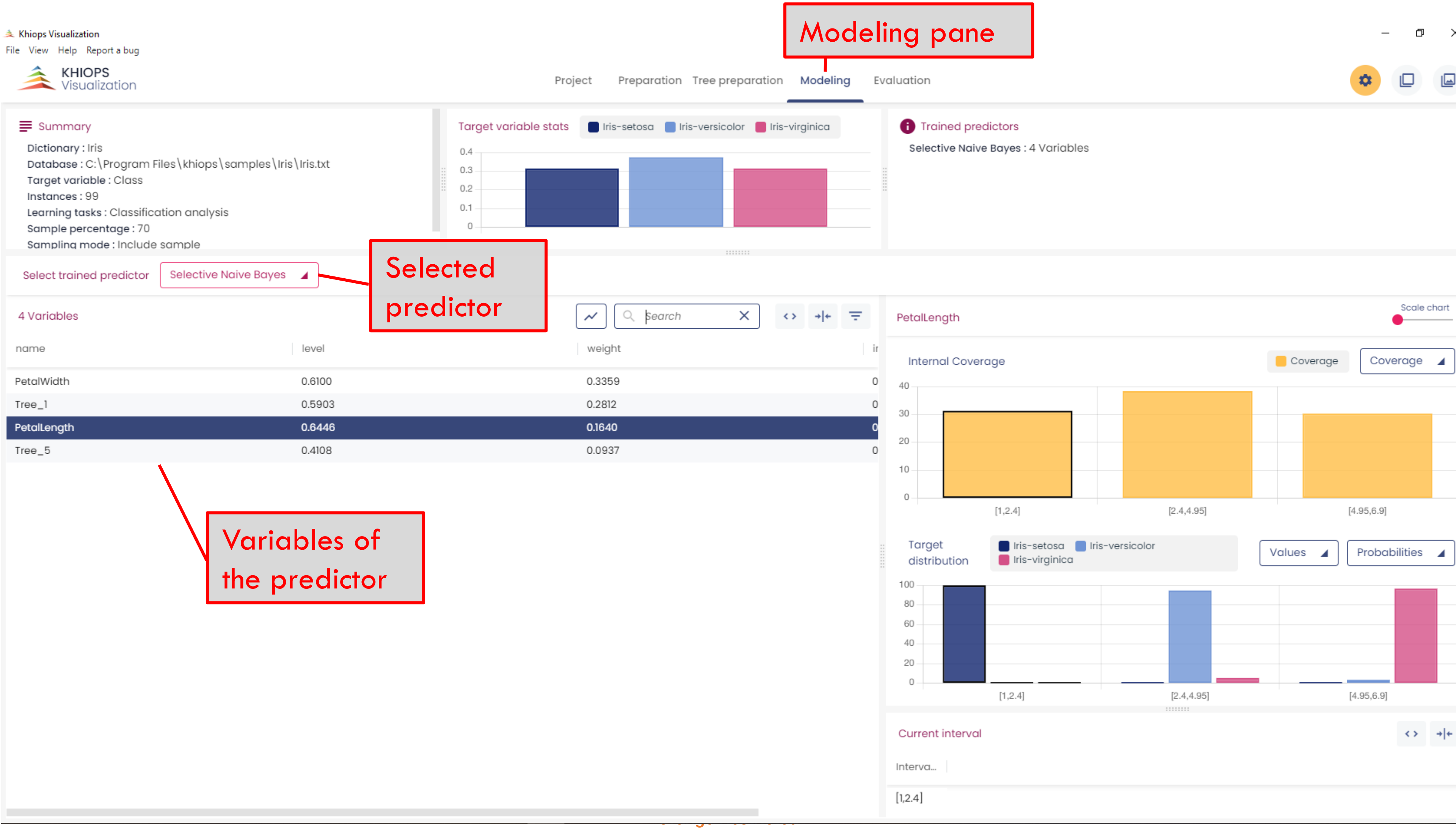
Sequence of trre rules leading to the leaf



Leaf rules : L4

Variable ↑	Type	Partition
SepalLength	Numerical	[4.3, 5.75]
SepalWidth	Numerical	[2.95, 4.4]

Exploratory of classification results using Khiops Visualization



Modeling pane

Selected predictor

Variables of the predictor

Khiops Visualization

Summary

Dictionary : Iris
Database : C:\Program Files\khiops\samples\Iris\Iris.txt
Target variable : Class
Instances : 99
Learning tasks : Classification analysis
Sample percentage : 70
Sampling mode : Include sample

Trained predictors

Selective Naive Bayes : 4 Variables

Select trained predictor: Selective Naive Bayes

4 Variables

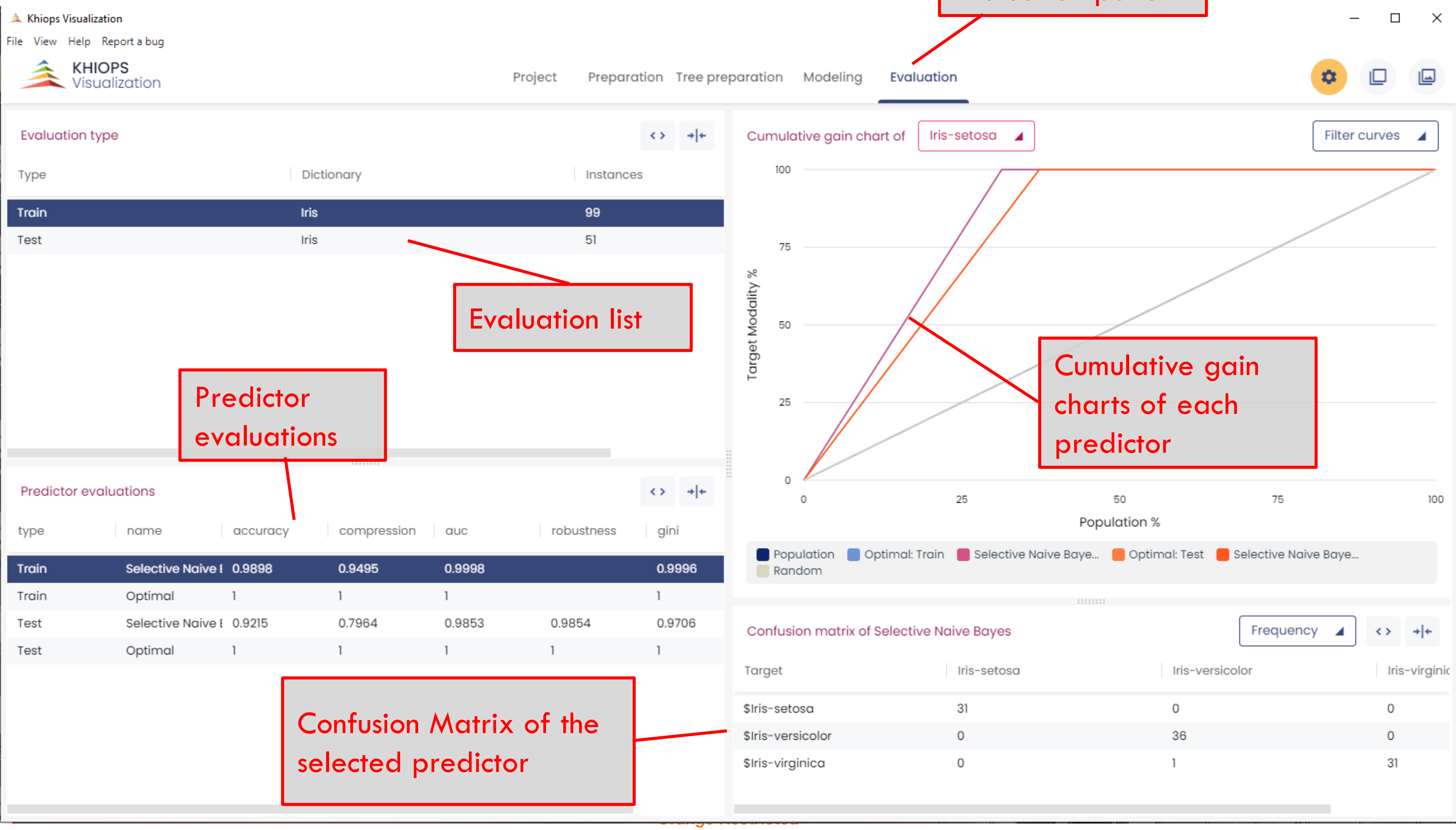
name	level	weight	ir
PetalWidth	0.6100	0.3359	0
Tree_1	0.5903	0.2812	0
PetalLength	0.6446	0.1640	0
Tree_5	0.4108	0.0937	0

PetalLength

Current interval

Interva...
[1,2.4]

Exploratory of classification results using Khiops Visualization



Evaluation type

Type	Dictionary	Instances
Train	Iris	99
Test	Iris	51

Predictor evaluations

type	name	accuracy	compression	auc	robustness	gini
Train	Selective Naive I	0.9898	0.9495	0.9998		0.9996
Train	Optimal	1	1	1		1
Test	Selective Naive I	0.9215	0.7964	0.9853	0.9854	0.9706
Test	Optimal	1	1	1	1	1

Confusion matrix of Selective Naive Bayes

Target	Iris-setosa	Iris-versicolor	Iris-virginic
$Iris-setosa	31	0	0
$Iris-versicolor	0	36	0
$Iris-virginica	0	1	31

Evaluation pane

Evaluation list

Predictor evaluations

Cumulative gain charts of each predictor

Confusion Matrix of the selected predictor

Exercises A and B …

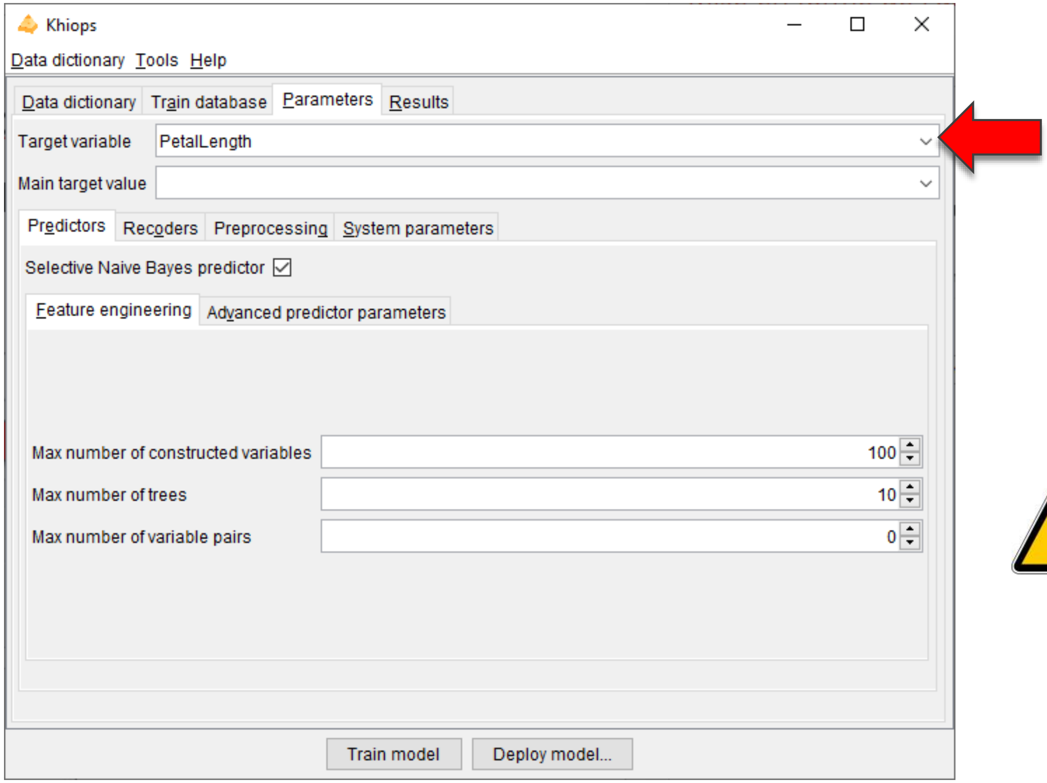
A : Perform a supervised classification on sample database Iris

B : Perform a Supervised classification on sample database Adult

➡ Interpret the analysis results

Regression (supervised)

- Same as classification with a numerical target variable



Khiops

Data dictionary Tools Help
Data dictionary Train database Parameters Results

Target variable: PetalLength

Main target value:

Predictors Recoders Preprocessing System parameters

Selective Naive Bayes predictor ☑

Feature engineering Advanced predictor parameters

Max number of constructed variables: 100

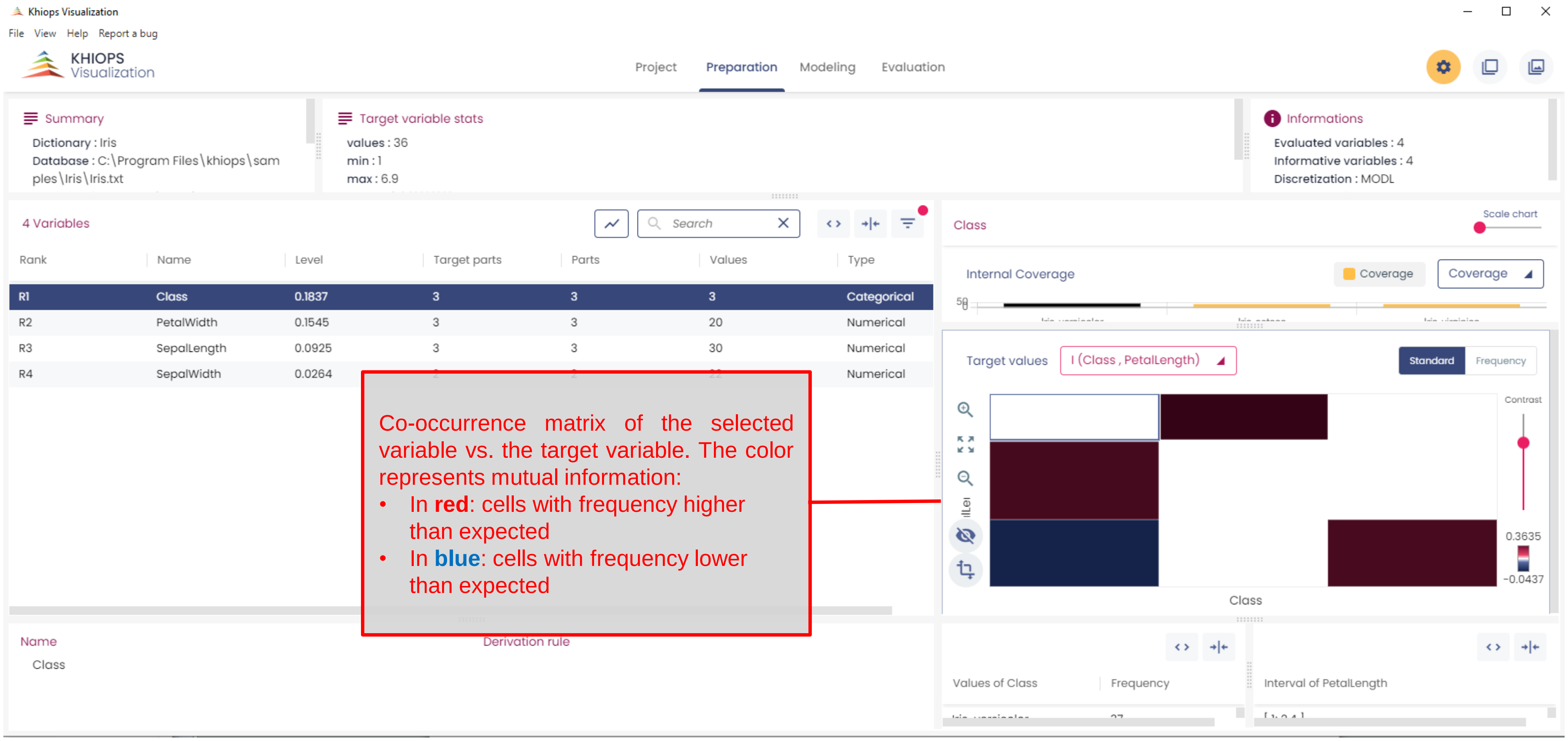
Max number of trees: 10

Max number of variable pairs: 0

Train model
Deploy model...

In this case, bivariate analysis and tree construction are not available!

Exploratory of regression results using Khiops Visualization



Summary

Dictionary : Iris
Database : C:\Program Files\khiops\samples\Iris\Iris.txt

Target variable stats

values : 36
min : 1
max : 6.9

Informations

Evaluated variables : 4
Informative variables : 4
Discretization : MODL

Rank	Name	Level	Target parts	Parts	Values	Type
R1	Class	0.1837	3	3	3	Categorical
R2	PetalWidth	0.1545	3	3	20	Numerical
R3	SepalLength	0.0925	3	3	30	Numerical
R4	SepalWidth	0.0264	[illegible]	[illegible]	[illegible]	Numerical

Name	Derivation rule
Class	

Values of Class	Frequency

Interval of PetalLength

Co-occurrence matrix of the selected variable vs. the target variable. The color represents mutual information:

- In **red**: cells with frequency higher than expected
- In **blue**: cells with frequency lower than expected

Exercise C …

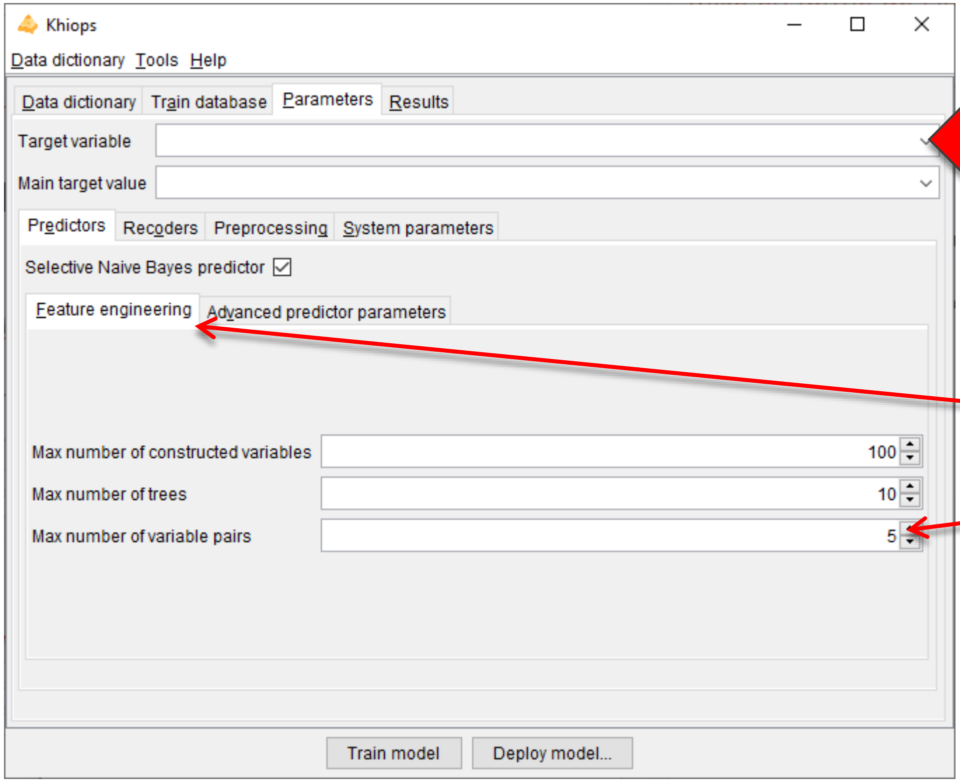
C : Perform a regression of variable PetalLength of Iris

➡ Interpret the analysis results

Correlation analysis (unsupervised, bivariate)

- **Train a correlation model between two variables**

  (categorical, numerical, both)



Target variable:

Main target value:

Selective Naive Bayes predictor ☑

Max number of constructed variables: 100

Max number of trees: 10

Max number of variable pairs: 5

1 –Target variable must be empty

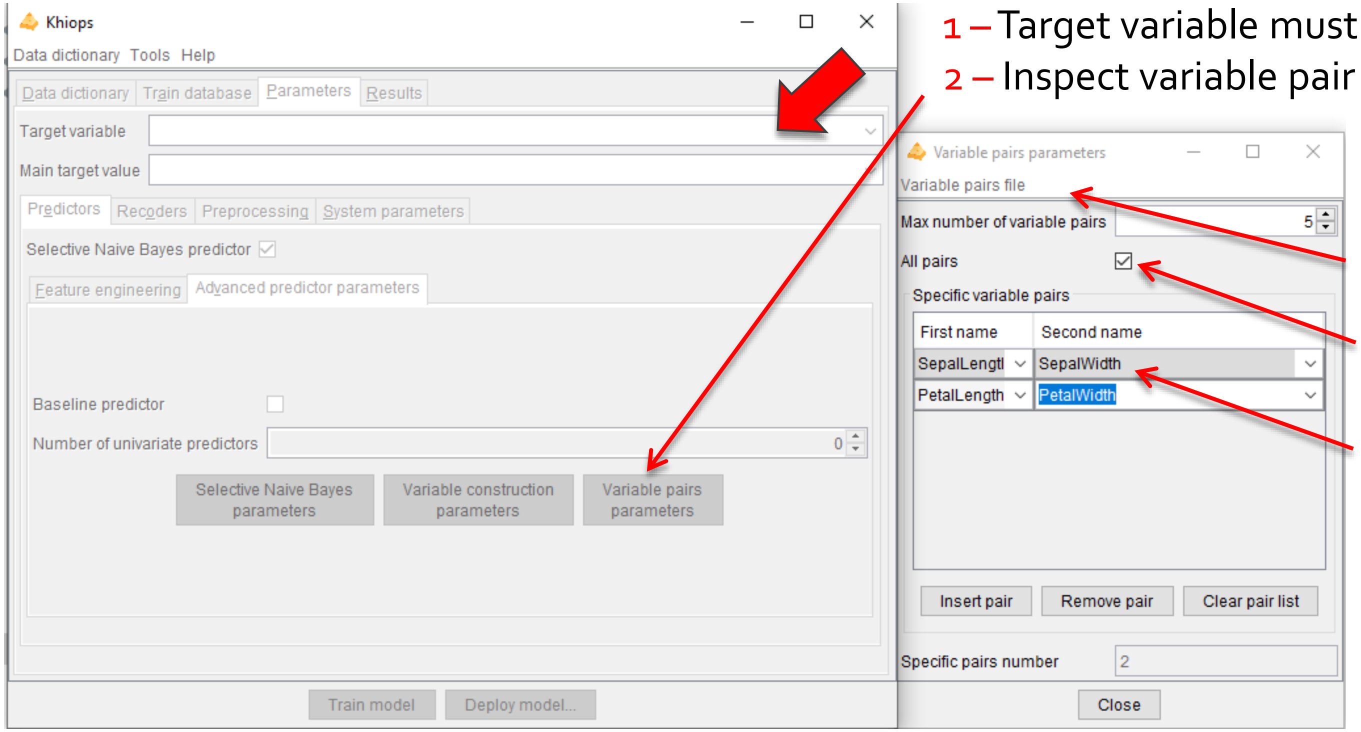
2 – Activate bivariate analysis

a – Feature engineering pane

b – Choice of a max number of pairs to analyze

Correlation analysis (unsupervised, bivariate)

- Train a correlation model : advanced parameters



Target variable:

Main target value:

Selective Naive Bayes predictor ☑

Baseline predictor ☐

Number of univariate predictors: 0

Variable pairs file

Max number of variable pairs: 5

All pairs ☑

Specific variable pairs

First name	Second name
SepalLengtl	SepalWidth
PetalLength	PetalWidth

Specific pairs number: 2

1 – Target variable must be empty

2 – Inspect variable pair parameters

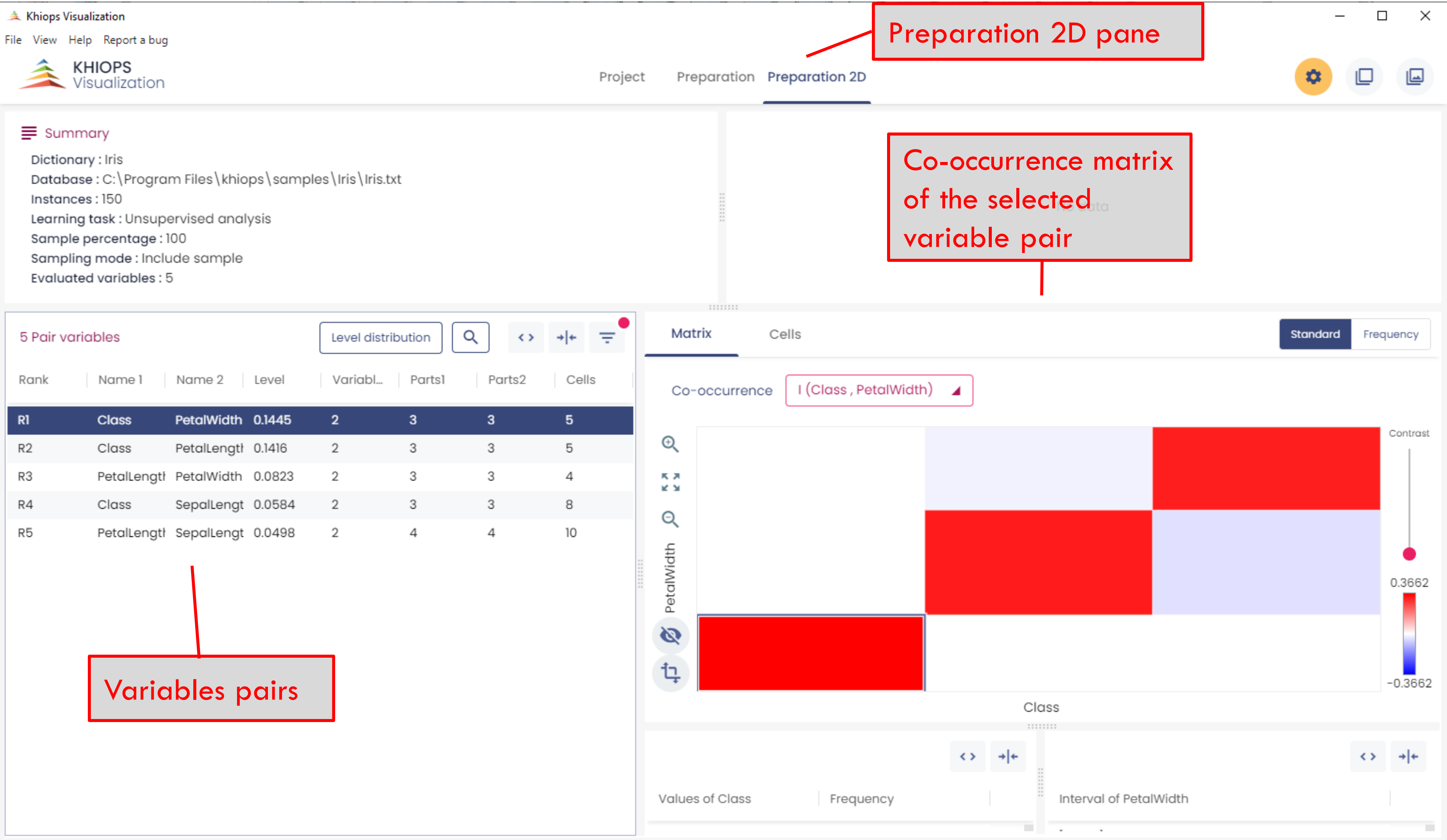
3 – Specify the pairs

a – import/export variable pairs file

b - all potential pair

c - individual pairs or families of variable pairs involving certain variables to analyze

Exploratory of correlation results using Khiops Visualization



Preparation 2D pane

Co-occurrence matrix of the selected variable pair

Variables pairs

Summary

Dictionary : Iris
Database : C:\Program Files\khiops\samples\Iris\Iris.txt
Instances : 150
Learning task : Unsupervised analysis
Sample percentage : 100
Sampling mode : Include sample
Evaluated variables : 5

5 Pair variables

Rank	Name 1	Name 2	Level	Variabl...	Parts1	Parts2	Cells
R1	Class	PetalWidth	0.1445	2	3	3	5
R2	Class	PetalLengtl	0.1416	2	3	3	5
R3	PetalLengtl	PetalWidth	0.0823	2	3	3	4
R4	Class	SepalLengt	0.0584	2	3	3	8
R5	PetalLengtl	SepalLengt	0.0498	2	4	4	10

Exercises D, E, F and G…

D : Perform the correlation analysis of the two most correlated variables of Iris
(*tip: analyze all pairs to identify the most informative*)

E : Idem with variables *PetalLength* and *PetalWidth*
(*tip: inspect the Variable pairs parameters)*

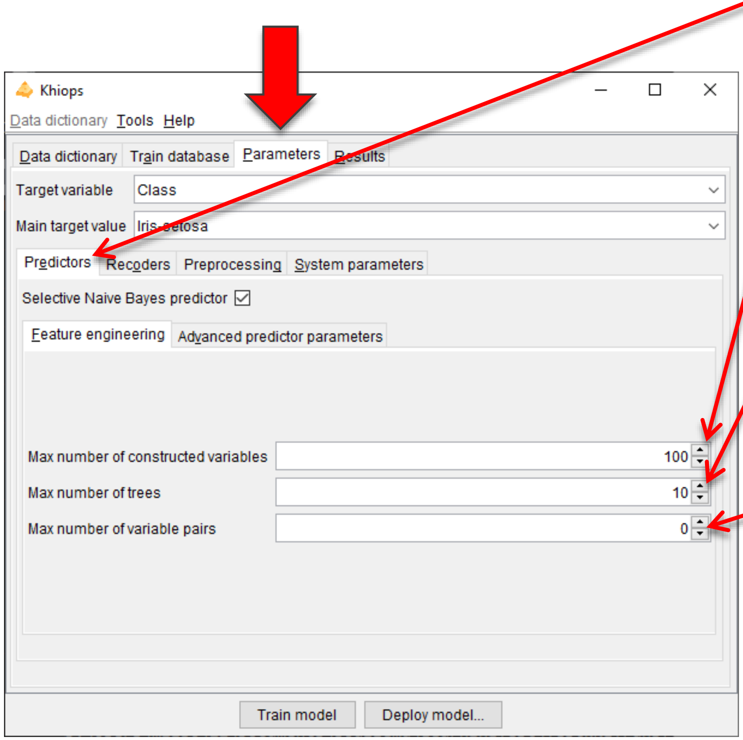
F : Idem with new constructed variables *PetalArea* and *SepalArea*
(*tip: use the derivation rule Product in dictionary,
see KhiopsGuide: sections « Derivation rules » and « Appendix »*)

G : Perform the correlation analysis of all pairs of Adult involving variable *native_country*

➡ Interpret the analysis results

Variable construction

- ## Parameters



Target variable: Class

Main target value: Iris-setosa

Selective Naive Bayes predictor ☑

Max number of constructed variables: 100

Max number of trees: 10

Max number of variable pairs: 0

Predictors

Feature engineering

- Max number of constructed variables
  - to build an analyze table from a multi-table schema (see later)
  - automatic extraction of complex information to obtain accurate classifiers
- Max number of trees
  - combines natives or constructed variables to extract complex information
  - better accuracy, at the expense of interpretability
- Max number of pairs of variable
  - to understand correlation between variables
  - use rather for exploratory analysis rather than for better accuracy

Recommendation

- start with few constructed variables, and increase incrementally
- idem for trees
  - no tree for simpler, faster and more interpretable predictors
  - more and more trees for more accurate predictors

Exercise H …

A : Perform a supervised classification on sample database Letter

Build 0, 10, 50 trees

➡ Interpret the analysis results, and the trade-off between number of trees, training time and test accuracy

Integration in information systems

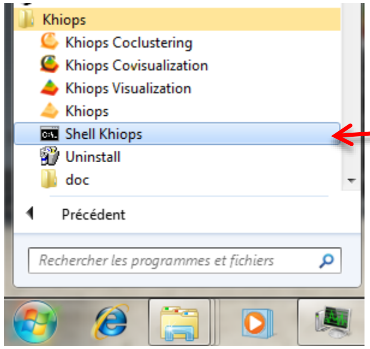
- Batch mode
  - to record and replay Khiops scripts
  - to perform any Khiops task from any programming language
  - see next slide

- Khiops Native Interface (KNI)
  - dynamic link library (DLL) for online deployment of Khiops models
  - package to download from www.khiops.com

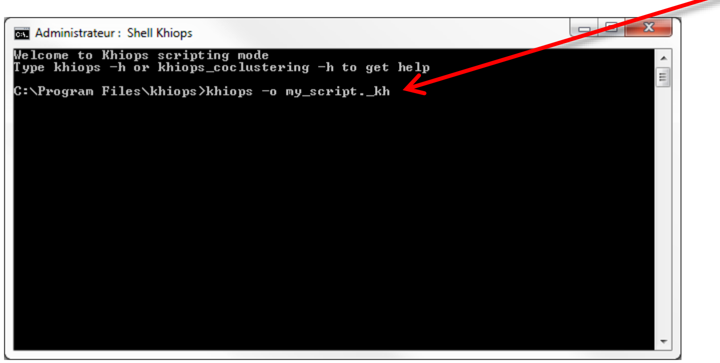
- Python Khiops Library (pykhiops)
  - to perform any Khiops task from python
  - to inspect any Khiops analysis results from python
  - python package available from www.khiops.com

- JSON file exports
  - Khiops dictionaries and analysis results can be exported from the Khiops tool to exploit Khiops results from any programming language

Batch mode



- Start a Shell Khiops



```
Welcome to Khiops scripting mode
Type khiops -h or khiops_coclustering -h to get help

C:\Program Files\khiops>khiops -o my_script._kh
```

- Record a script « automatically» using Khiops user interface

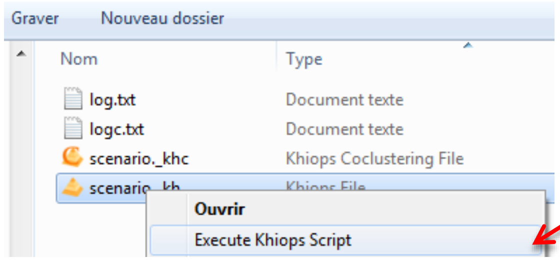
  khiops –o my_script._kh

  o = output

- Replay a script from the shell

  khiops –i my_script._kh

  i = input



Nom	Type
log.txt	Document texte
logc.txt	Document texte
scenario._khc	Khiops Coclustering File
scenario._kh	Khiops File

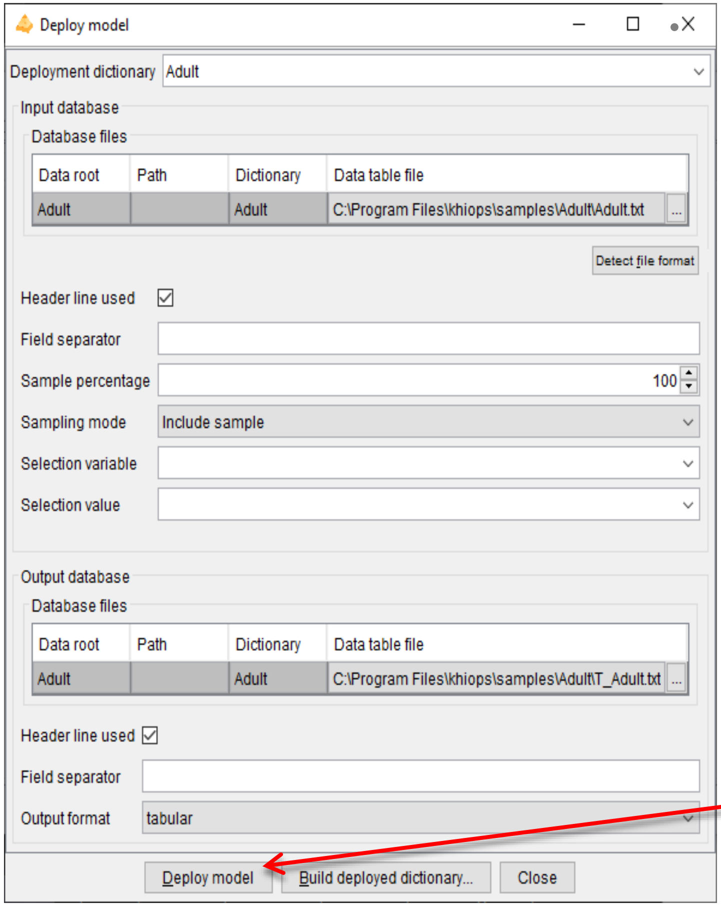
- Replay a script from Windows Explorer

  right click on script file

Exercise I …

I : Record a script file, then replay it …

Deploy a model



Deploy model

Deployment dictionary: Adult

Input database

Database files

Data root	Path	Dictionary	Data table file
Adult		Adult	C:\Program Files\khiops\samples\Adult\Adult.txt

Header line used ☑

Field separator:

Sample percentage: 100

Sampling mode: Include sample

Selection variable:

Selection value:

Output database

Database files

Data root	Path	Dictionary	Data table file
Adult		Adult	C:\Program Files\khiops\samples\Adult\T_Adult.txt

Header line used ☑

Field separator:

Output format: tabular

Steps for model deployment

- 1- Start from a modeling dictionary « *Modeling.kdic* »
  - In « Data dictionary » pane
- 2- Choose the variables to deploy
  - Inspect the modeling dictionary In « Data dictionary » pane by right-click in the "Dictionaries in file" list
  - Suppress the « *Unused* » tag from identifier variables
  - Select the prediction variables to deploy
- 3- Menu : « *Tools -> Deploy model* »
- 4- Deploy model dialog box
  - Select deployment dictionary
  - Select input database
  - Select output database
  - Click on « Deploy model » button

Exercise J …

J: Deploy a classifier on database Iris

Khiops Coclustering & Khiops Covisualization

- Khiops Coclustering
  - Correlation analysis of two or more variables using a hierarchical coclustering model
- Khiops Covisualization
  - Exploratory analysis of Khiops Coclustering results using an interactive visualization tool

Khiops Coclustering & Khiops Covisualization

- Train a coclustering model
  - Use of Khiops Coclustering back-end tool
  - Co-partition of two or more categorical or numerical variables
  - At each level of the hierarchy, the merge of clusters with the minimum information loss is performed
  - Write results in a coclustering report file « *.khcj* »

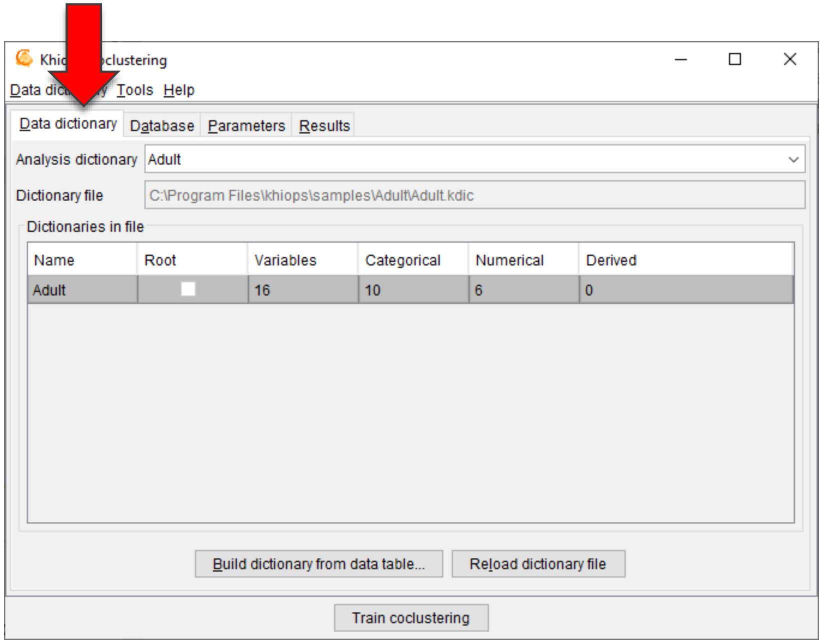
- Exploratory analysis of the results
  - Use of Khiops Covisualization tool
  - Navigation in the hierarchy of models

Train a coclustering model

- **Step 1 :** Open an existing dictionary

  (ex: sample Adult.kdic)

  - Description of variables to use during analysis



Analysis dictionary: Adult

Dictionary file: C:\Program Files\khiops\samples\Adult\Adult.kdic

Dictionaries in file

Name	Root	Variables	Categorical	Numerical	Derived
Adult	□	16	10	6	0

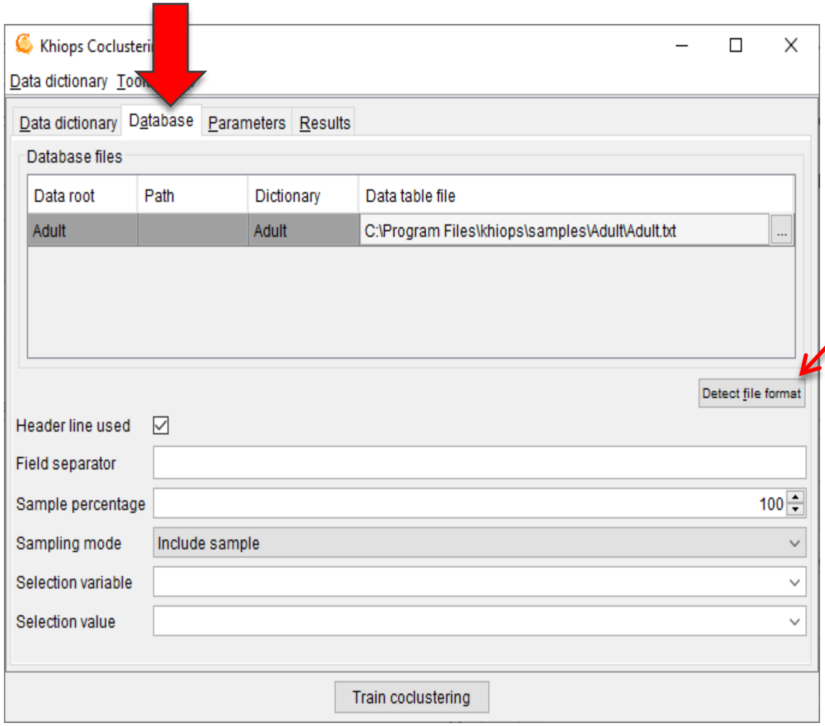
Available actions :

- Open, Save, Save as, Close
- Edition (menu « Dictionary file/Inspect current dictionary », or NotePad)
- Reload dictionary file
- Build dictionary from data table

```
Dictionary    Adult
{
              Numerical     Label;
              Numerical     age;
              Categorical   workclass;
              Numerical     fnlwgt;
              Categorical   education;
              Numerical     education_num;
              Categorical   marital_status;
              Categorical   occupation;
              Categorical   relationship;
              Categorical   race;
              Categorical   sex;
              Numerical     capital_gain;
              Numerical     capital_loss;
              Numerical     hours_per_week;
              Categorical   native_country;
              Categorical   class;
};
```

Build a coclustering model

- **Step 2 : Specification of used database**



Database files

Data root	Path	Dictionary	Data table file
Adult		Adult	C:\Program Files\khiops\samples\Adult\Adult.txt

Header line used ☑

Field separator

Sample percentage 100

Sampling mode Include sample

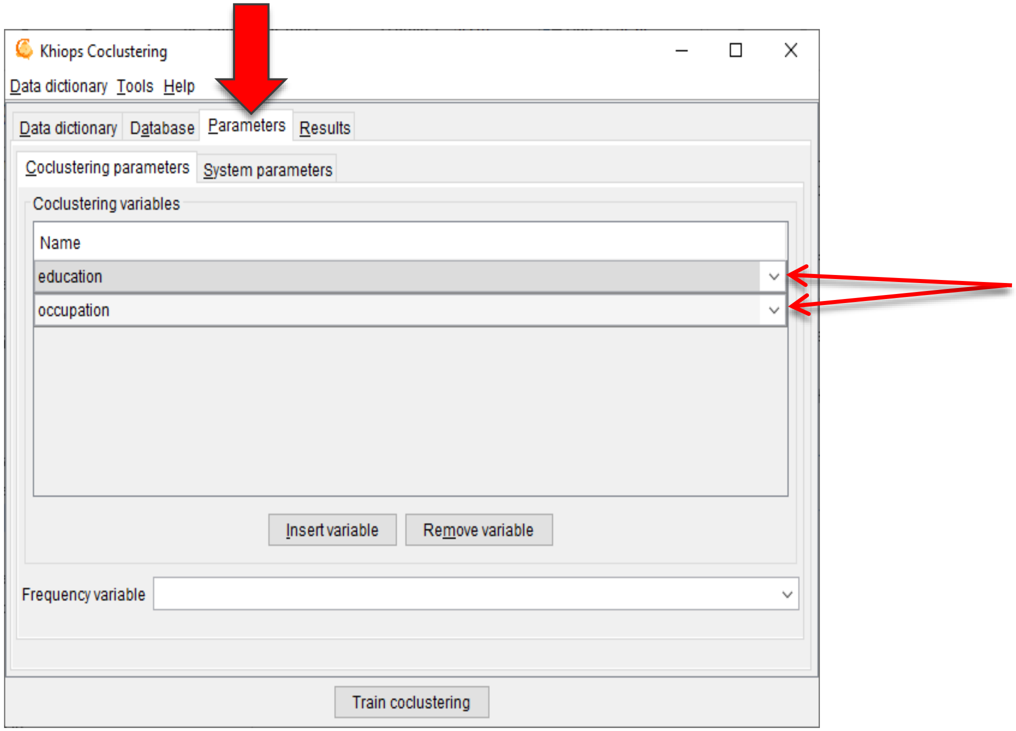
Selection variable

Selection value

Detect file format : heuristic help that scans the first few lines to guess the file format. The header line and field separator are updated on success, with a warning or an error in the log window only if necessary.

Build a coclustering model

- **Step 3 : Specification of coclustering variables**



Coclustering variables

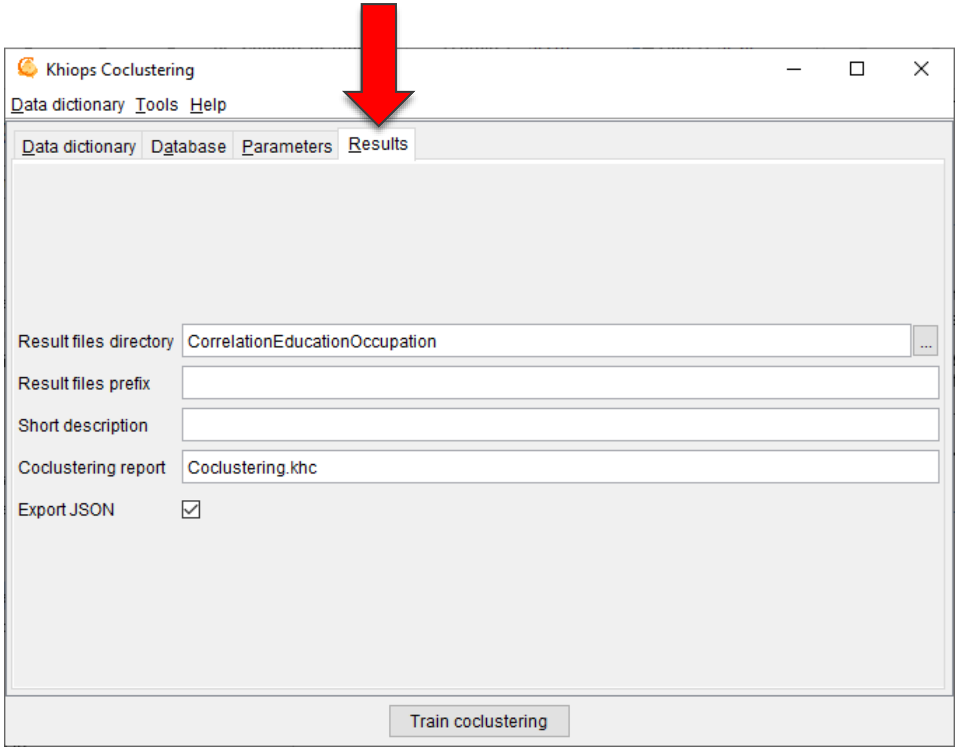
Name
education
occupation

Frequency variable

Coclustering variables

Build a coclustering model

- **Step 4 : Results**



Result files directory: CorrelationEducationOccupation

Result files prefix:

Short description:

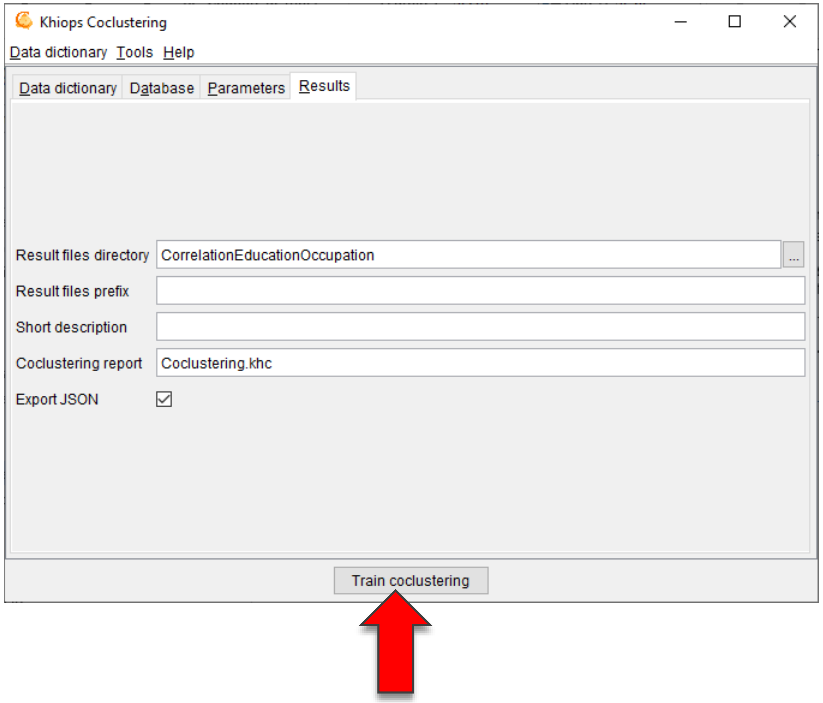
Coclustering report: Coclustering.khc

Export JSON ☑

- Directory where result file is written
- Prefix *(ex: in case of several experiments)*
- Synthetic coclustering report (cf. Khiops Covisualization)
- Json report, to get the analysis results from external tools

Build a coclustering model

- **Step 5 : Start the analysis**


Khiops Coclustering

Data dictionary | Database | Parameters | Results

Result files directory: CorrelationEducationOccupation

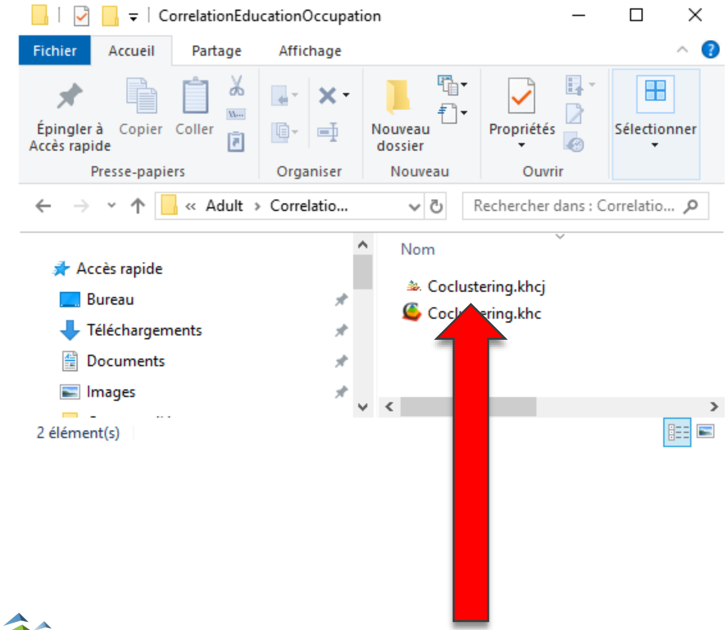
Result files prefix:

Short description:

Coclustering report: Coclustering.khc

Export JSON: ☑

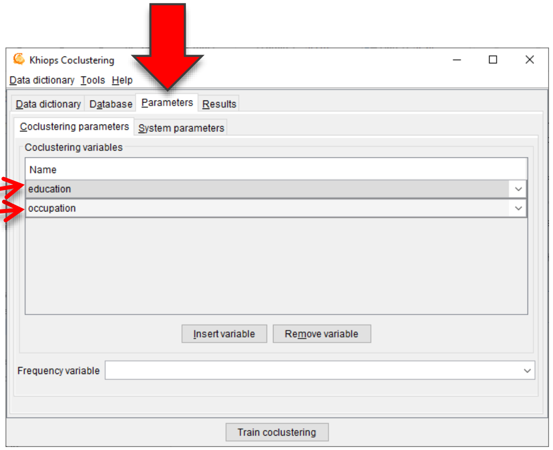
1 – Train the coclustering



2 - Inspect the results using Khiops Covisualization (double-click on *.khcj* file)

Example: base Adult education*occupation

- With Khiops Coclustering
  - Analysis of pair of variables education*occupation



Khiops Coclustering

Data dictionary Tools Help

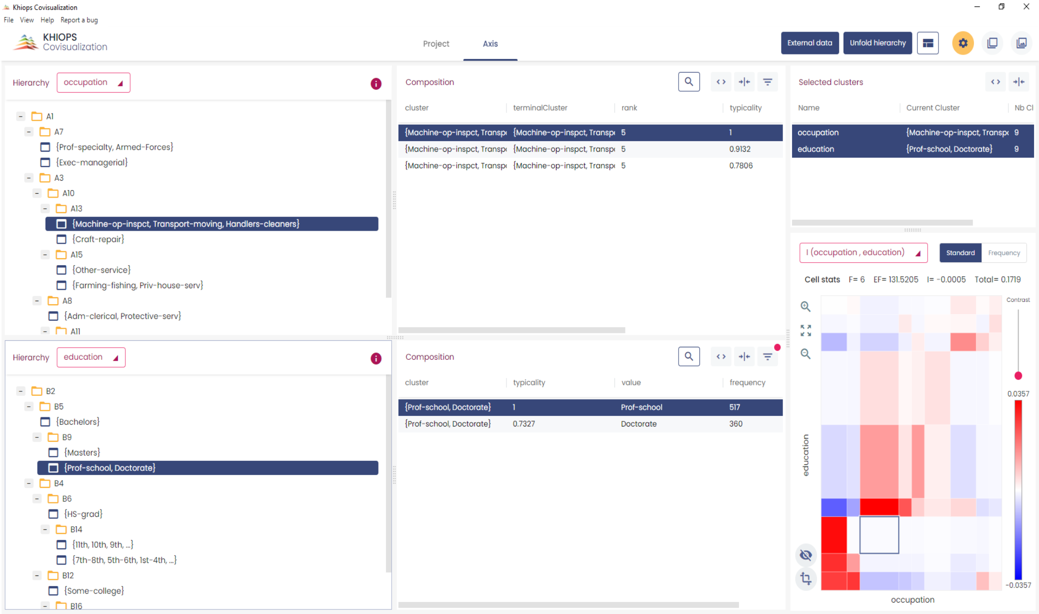
Data dictionary | Database | Parameters | Results

Coclustering parameters | System parameters

Coclustering variables

Name
education
occupation

Frequency variable:



Khiops Covisualization

Hierarchy: occupation

- A1
  - A7
    - {Prof-specialty, Armed-Forces}
    - {Exec-managerial}
  - A3
    - A10
      - A13
        - {Machine-op-inspct, Transport-moving, Handlers-cleaners}
        - {Craft-repair}
      - A15
        - {Other-service}
        - {Farming-fishing, Priv-house-serv}
    - A8
      - {Adm-clerical, Protective-serv}
      - A11

Composition

cluster	terminalCluster	rank	typicality
{Machine-op-inspct, Transp	{Machine-op-inspct, Transp	5	1
{Machine-op-inspct, Transp	{Machine-op-inspct, Transp	5	0.9132
{Machine-op-inspct, Transp	{Machine-op-inspct, Transp	5	0.7806

Selected clusters

Name	Current Cluster	Nb Cl
occupation	{Machine-op-inspct, Transp	9
education	{Prof-school, Doctorate}	9

Hierarchy: education

- B2
  - B5
    - {Bachelors}
    - B9
      - {Masters}
      - {Prof-school, Doctorate}
  - B4
    - B6
      - {HS-grad}
      - B14
        - {11th, 10th, 9th, ...}
        - {7th-8th, 5th-6th, 1st-4th, ...}
    - B12
      - {Some-college}
      - B16

Composition

cluster	typicality	value	frequency
{Prof-school, Doctorate}	1	Prof-school	517
{Prof-school, Doctorate}	0.7327	Doctorate	360

I (occupation , education)

Cell stats F= 6 EF= 131.5205 I= -0.0005 Total= 0.1719

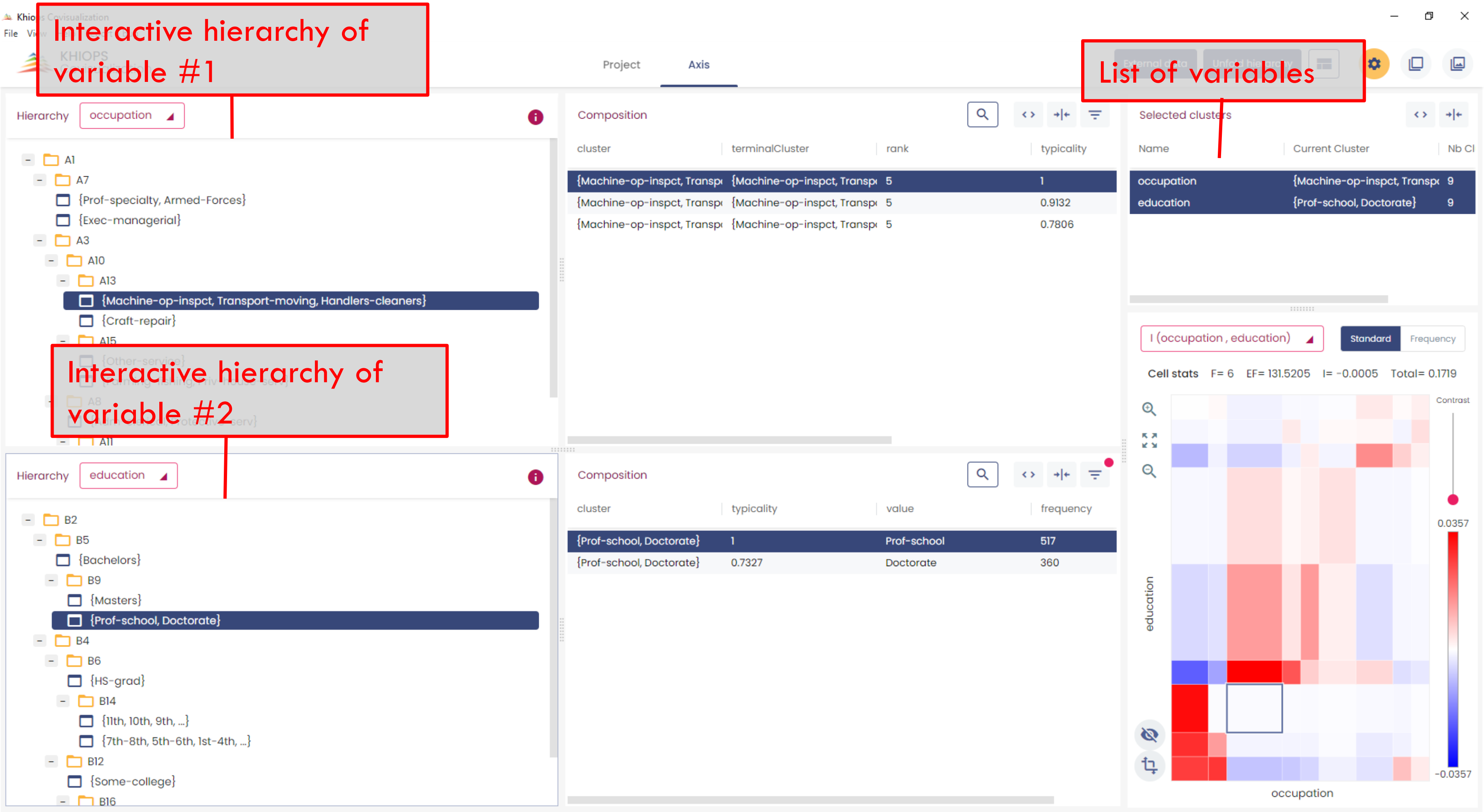
- With Khiops Covisualization
  - Exploratory analysis of the results

Exercise J …

J : Train a coclustering model
on two categorical variables of sample database Adult

➡ Explore the analysis results

Khiops Covisualization



Interactive hierarchy of variable #1

List of variables

Interactive hierarchy of variable #2

Hierarchy: occupation

- A1
  - A7
    - {Prof-specialty, Armed-Forces}
    - {Exec-managerial}
  - A3
    - A10
      - A13
        - {Machine-op-inspct, Transport-moving, Handlers-cleaners}
        - {Craft-repair}
      - A15
        - {Other-service}
    - A8
  - A11

Composition

cluster	terminalCluster	rank	typicality
{Machine-op-inspct, Transp	{Machine-op-inspct, Transp	5	1
{Machine-op-inspct, Transp	{Machine-op-inspct, Transp	5	0.9132
{Machine-op-inspct, Transp	{Machine-op-inspct, Transp	5	0.7806

Selected clusters

Name	Current Cluster	Nb Cl
occupation	{Machine-op-inspct, Transp	9
education	{Prof-school, Doctorate}	9

Hierarchy: education

- B2
  - B5
    - {Bachelors}
    - B9
      - {Masters}
      - {Prof-school, Doctorate}
  - B4
    - B6
      - {HS-grad}
      - B14
        - {11th, 10th, 9th, ...}
        - {7th-8th, 5th-6th, 1st-4th, ...}
    - B12
      - {Some-college}
      - B16

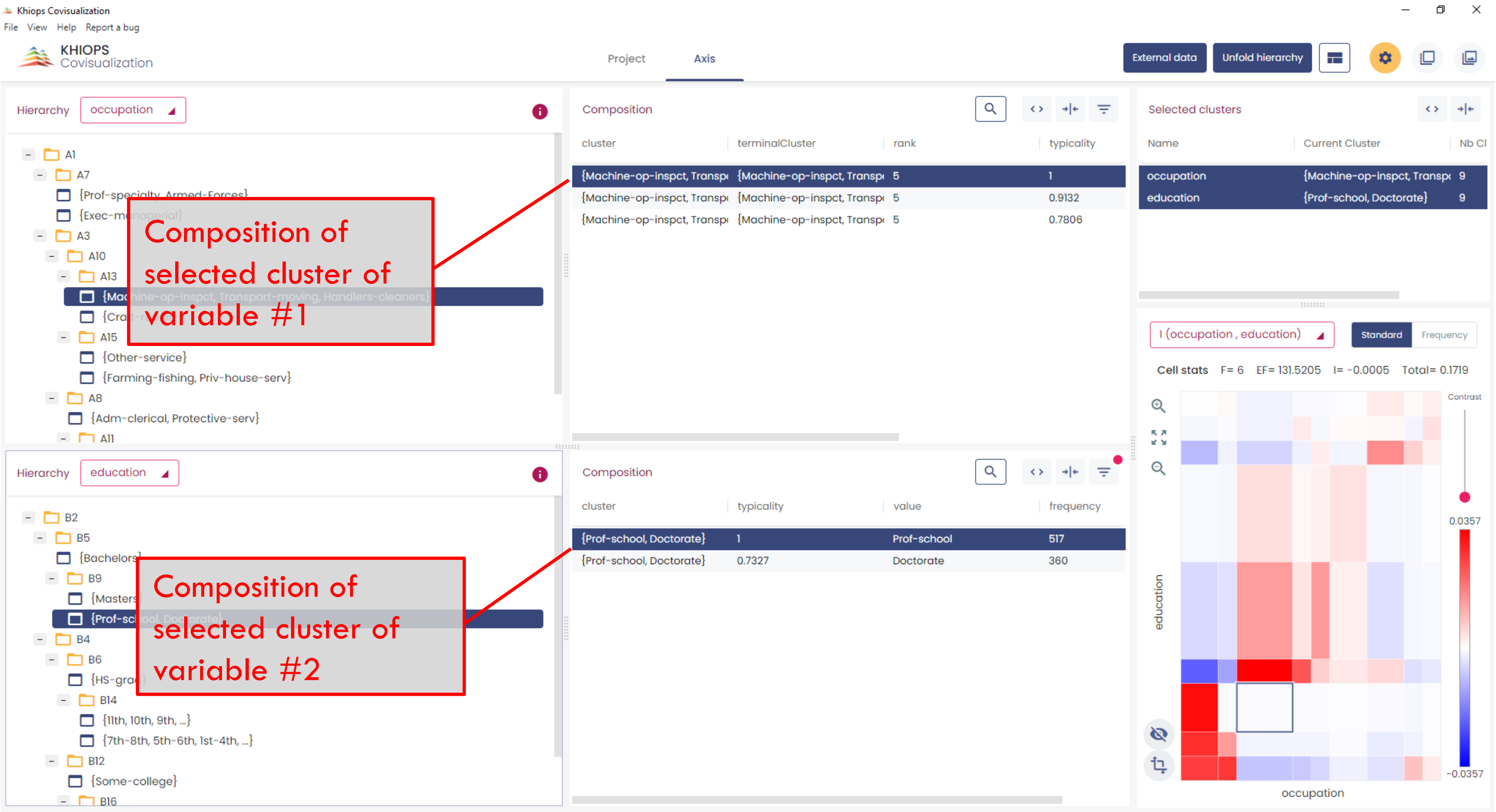
Composition

cluster	typicality	value	frequency
{Prof-school, Doctorate}	1	Prof-school	517
{Prof-school, Doctorate}	0.7327	Doctorate	360

I (occupation , education) — Standard | Frequency

Cell stats F= 6 EF= 131.5205 I= -0.0005 Total= 0.1719

Khiops Covisualization



Composition of selected cluster of variable #1

Composition of selected cluster of variable #2

Composition

cluster	terminalCluster	rank	typicality
{Machine-op-inspct, Transp	{Machine-op-inspct, Transp	5	1
{Machine-op-inspct, Transp	{Machine-op-inspct, Transp	5	0.9132
{Machine-op-inspct, Transp	{Machine-op-inspct, Transp	5	0.7806

Selected clusters

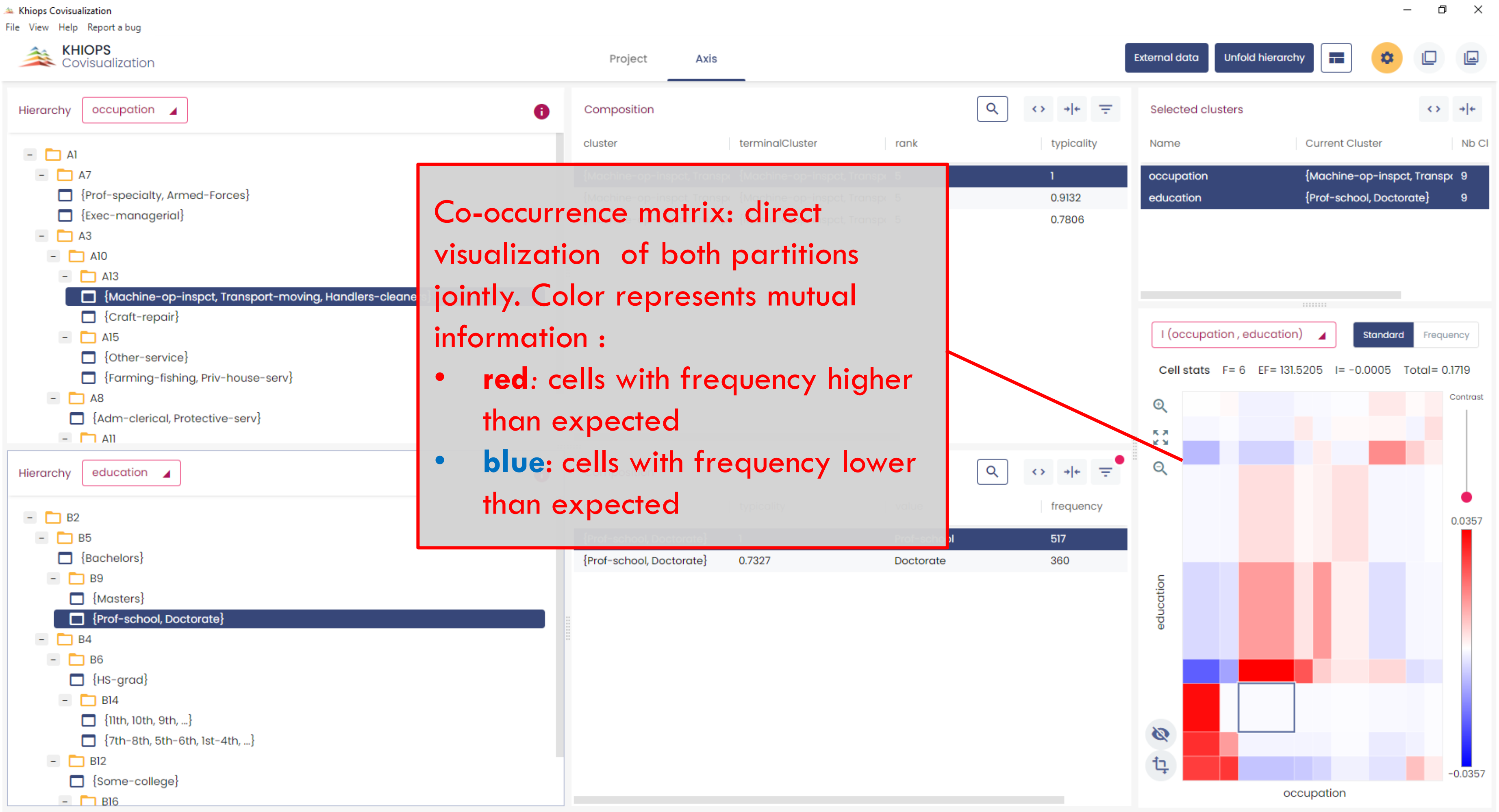
Name	Current Cluster	Nb Cl
occupation	{Machine-op-inspct, Transp	9
education	{Prof-school, Doctorate}	9

Cell stats F= 6 EF= 131.5205 I= -0.0005 Total= 0.1719

Composition

cluster	typicality	value	frequency
{Prof-school, Doctorate}	1	Prof-school	517
{Prof-school, Doctorate}	0.7327	Doctorate	360

Khiops Covisualization



Hierarchy occupation:

- A1
  - A7
    - {Prof-specialty, Armed-Forces}
    - {Exec-managerial}
  - A3
    - A10
      - A13
        - {Machine-op-inspct, Transport-moving, Handlers-cleane
        - {Craft-repair}
      - A15
        - {Other-service}
        - {Farming-fishing, Priv-house-serv}
    - A8
      - {Adm-clerical, Protective-serv}
      - A11

Composition:

cluster	terminalCluster	rank	typicality
{Machine-op-inspct, Transpo	{Machine-op-inspct, Transpo	5	1
{Machine-op-inspct, Transpo	{Machine-op-inspct, Transpo	5	0.9132
	{Machine-op-inspct, Transpo	5	0.7806

Selected clusters:

Name	Current Cluster	Nb Cl
occupation	{Machine-op-inspct, Transpo	9
education	{Prof-school, Doctorate}	9

Cell stats F= 6 EF= 131.5205 I= -0.0005 Total= 0.1719

Hierarchy education:

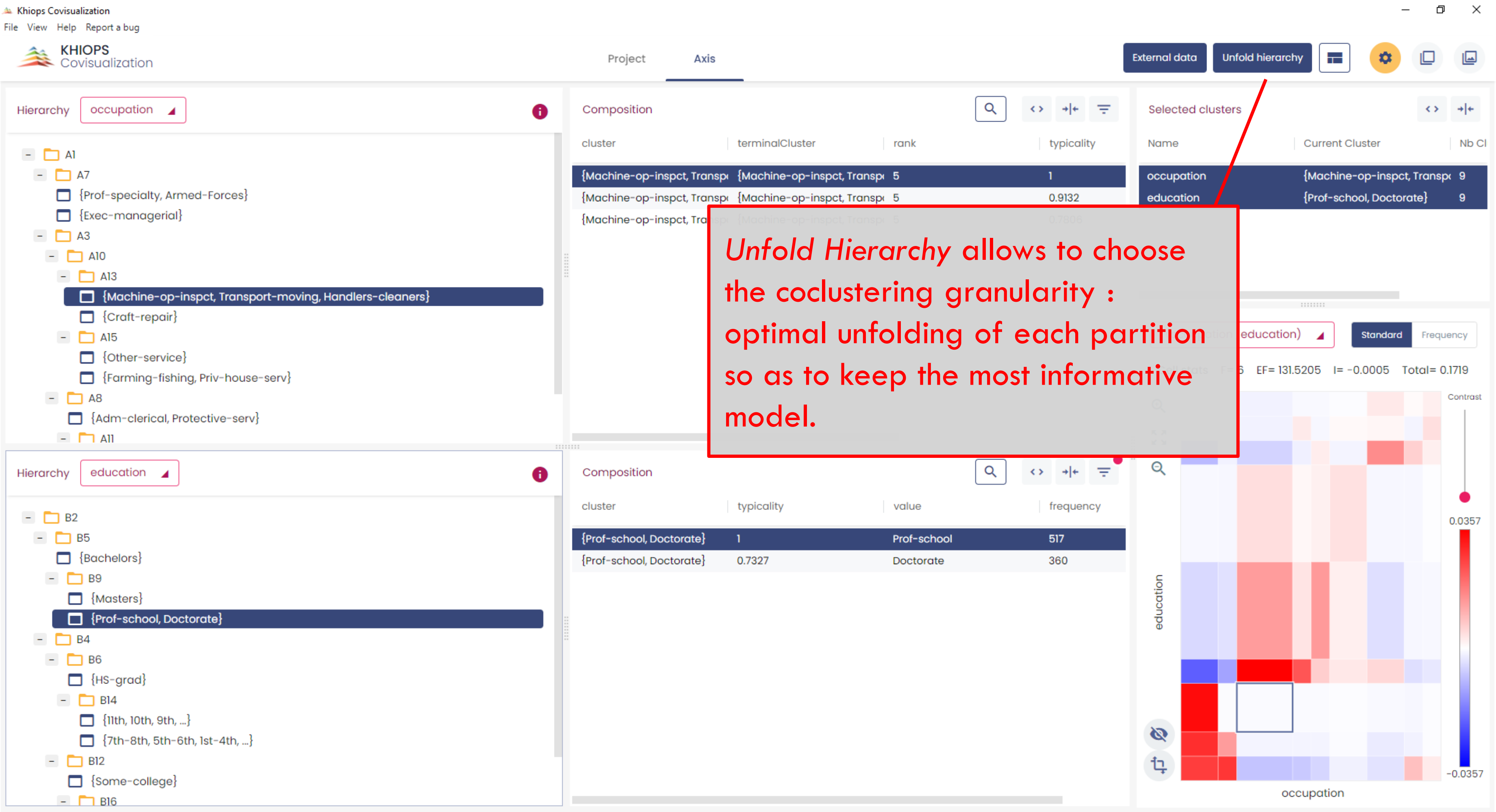
- B2
  - B5
    - {Bachelors}
    - B9
      - {Masters}
      - {Prof-school, Doctorate}
  - B4
    - B6
      - {HS-grad}
      - B14
        - {11th, 10th, 9th, ...}
        - {7th-8th, 5th-6th, 1st-4th, ...}
    - B12
      - {Some-college}
      - B16

	typicality	value	frequency
{Prof-school, Doctorate}	1	Prof-school	517
{Prof-school, Doctorate}	0.7327	Doctorate	360

Co-occurrence matrix: direct visualization of both partitions jointly. Color represents mutual information :

- **red**: cells with frequency higher than expected
- **blue**: cells with frequency lower than expected

Khiops Covisualization



Composition

cluster	terminalCluster	rank	typicality
{Machine-op-inspct, Transp	{Machine-op-inspct, Transp	5	1
{Machine-op-inspct, Transp	{Machine-op-inspct, Transp	5	0.9132
{Machine-op-inspct, Transp	{Machine-op-inspct, Transp	5	0.7806

Selected clusters

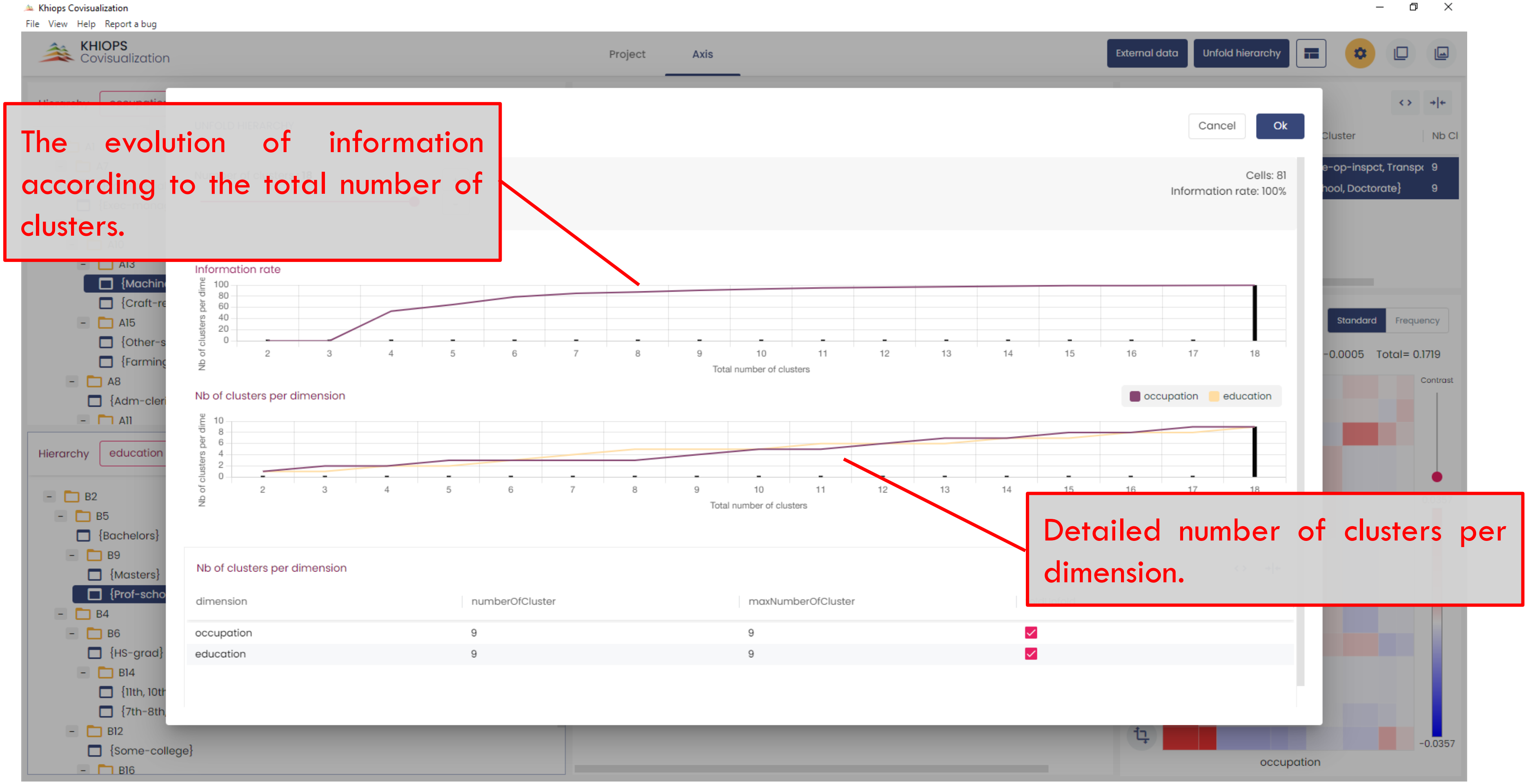
Name	Current Cluster	Nb Cl
occupation	{Machine-op-inspct, Transp	9
education	{Prof-school, Doctorate}	9

Composition

cluster	typicality	value	frequency
{Prof-school, Doctorate}	1	Prof-school	517
{Prof-school, Doctorate}	0.7327	Doctorate	360

Unfold Hierarchy allows to choose the coclustering granularity : optimal unfolding of each partition so as to keep the most informative model.

Khiops Covisualization



Cells: 81
Information rate: 100%

Nb of clusters per dimension

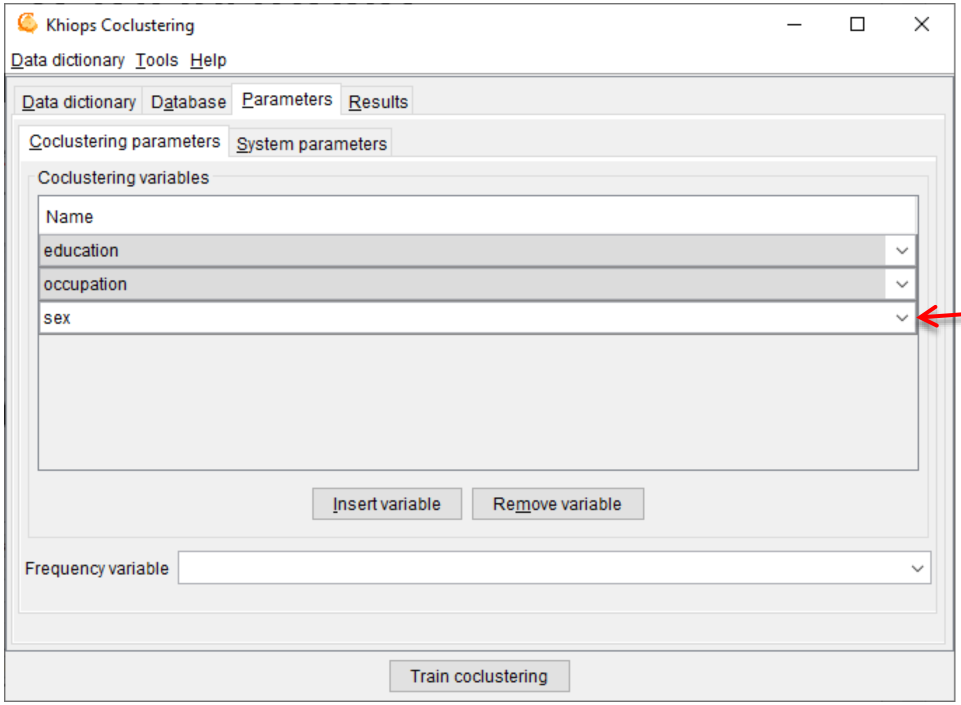
dimension	numberOfCluster	maxNumberOfCluster	
occupation	9	9	☑
education	9	9	☑

The evolution of information according to the total number of clusters.

Detailed number of clusters per dimension.

Training a triclustering

- Same as coclustering (Step 3) by inserting a third variable



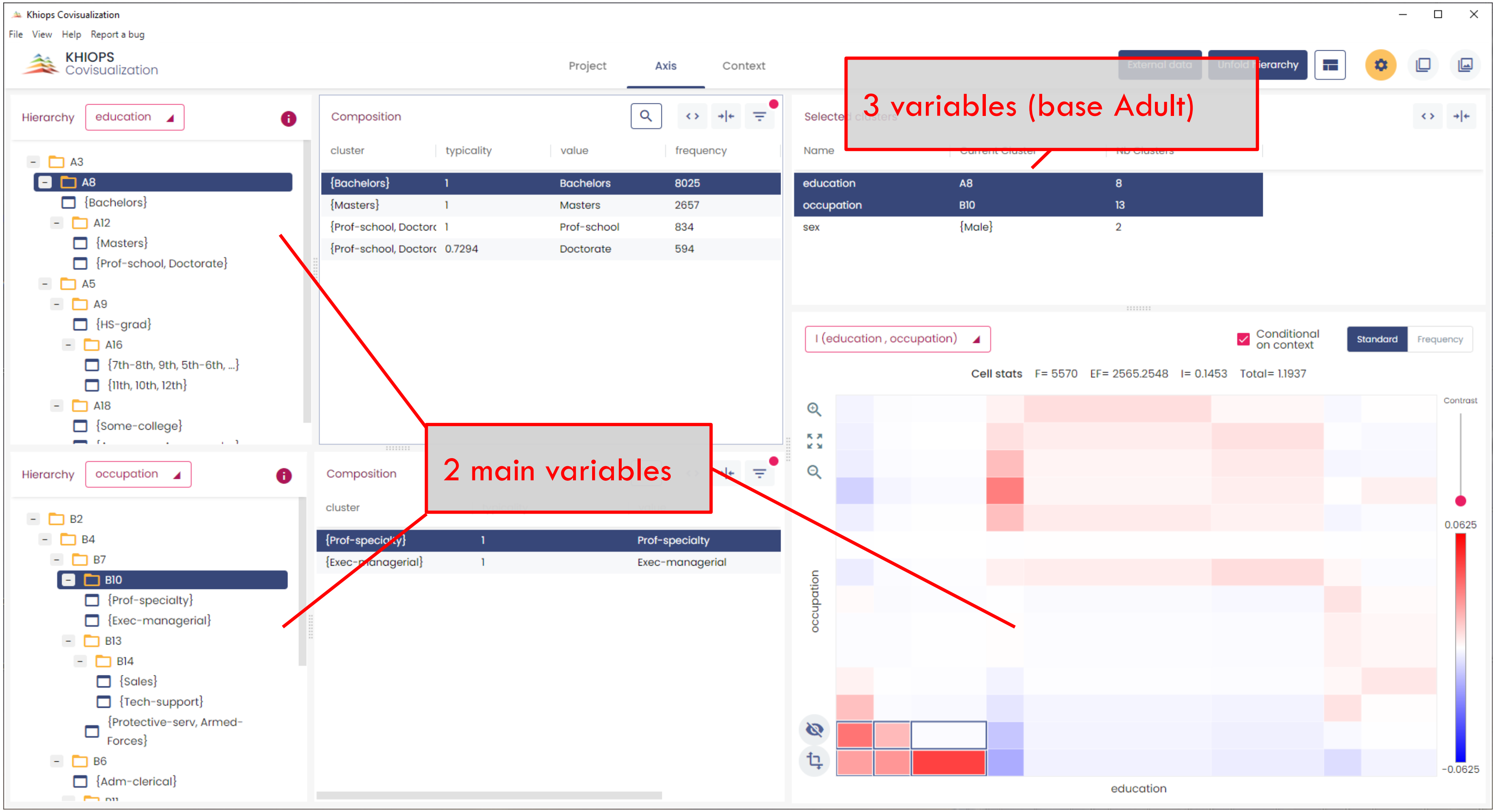
Coclustering variables

Name
education
occupation
sex

Frequency variable

The third variable

Exploring a triclustering



3 variables (base Adult)

2 main variables

Hierarchy: education

- A3
  - A8
    - {Bachelors}
    - A12
      - {Masters}
      - {Prof-school, Doctorate}
  - A5
    - A9
      - {HS-grad}
      - A16
        - {7th-8th, 9th, 5th-6th, ...}
        - {11th, 10th, 12th}
    - A18
      - {Some-college}

cluster	typicality	value	frequency
{Bachelors}	1	Bachelors	8025
{Masters}	1	Masters	2657
{Prof-school, Doctorc	1	Prof-school	834
{Prof-school, Doctorc	0.7294	Doctorate	594

Name	Current Cluster	Nb Clusters
education	A8	8
occupation	B10	13
sex	{Male}	2

I (education , occupation)

Cell stats F= 5570 EF= 2565.2548 I= 0.1453 Total= 1.1937

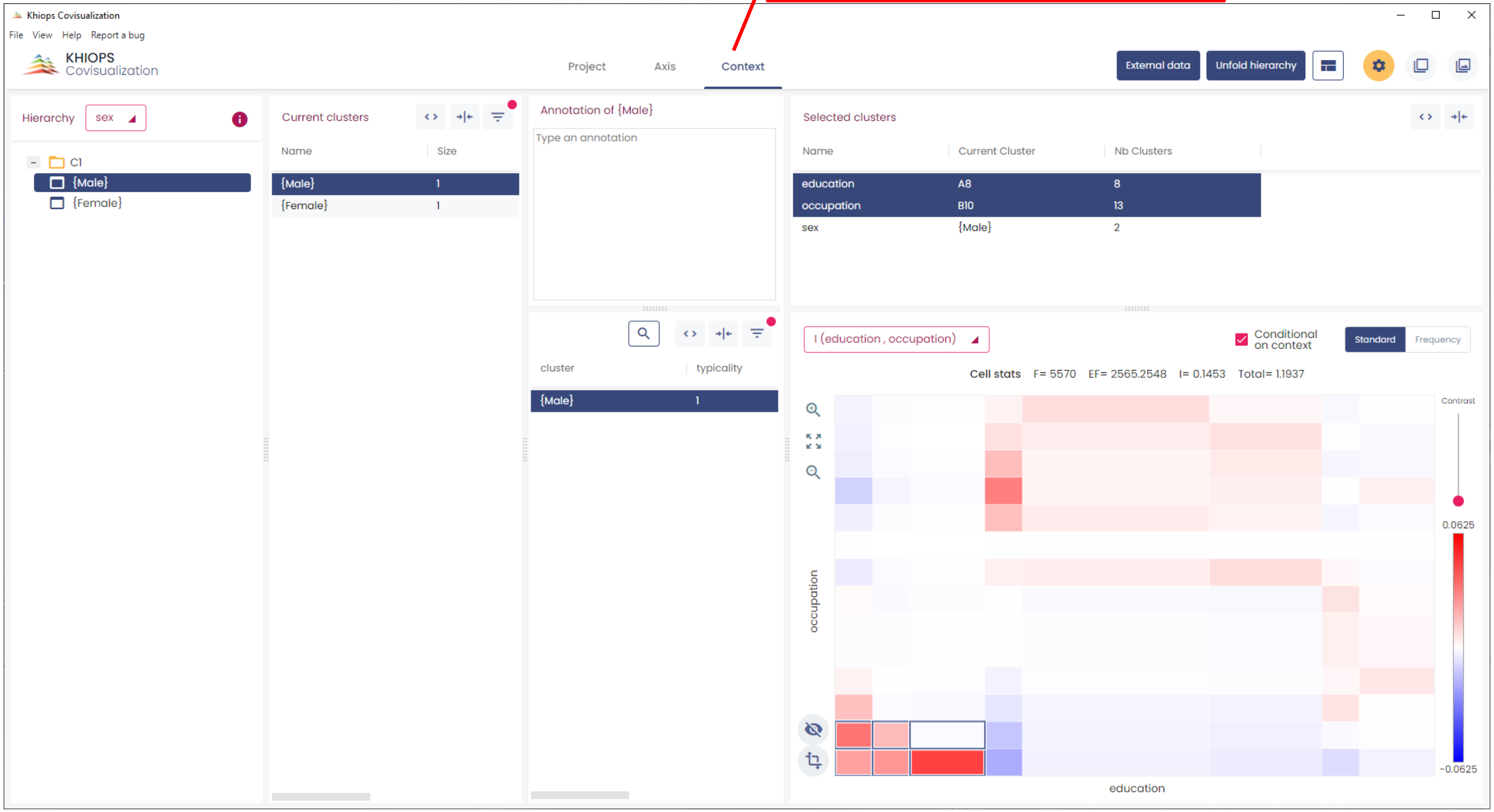
Hierarchy: occupation

- B2
  - B4
    - B7
      - B10
        - {Prof-specialty}
        - {Exec-managerial}
      - B13
        - B14
          - {Sales}
          - {Tech-support}
        - {Protective-serv, Armed-Forces}
    - B6
      - {Adm-clerical}
      - B11

cluster	typicality	value
{Prof-specialty}	1	Prof-specialty
{Exec-managerial}	1	Exec-managerial

Exploring a triclustering

Context pane



Current clusters

Name	Size
{Male}	1
{Female}	1

Selected clusters

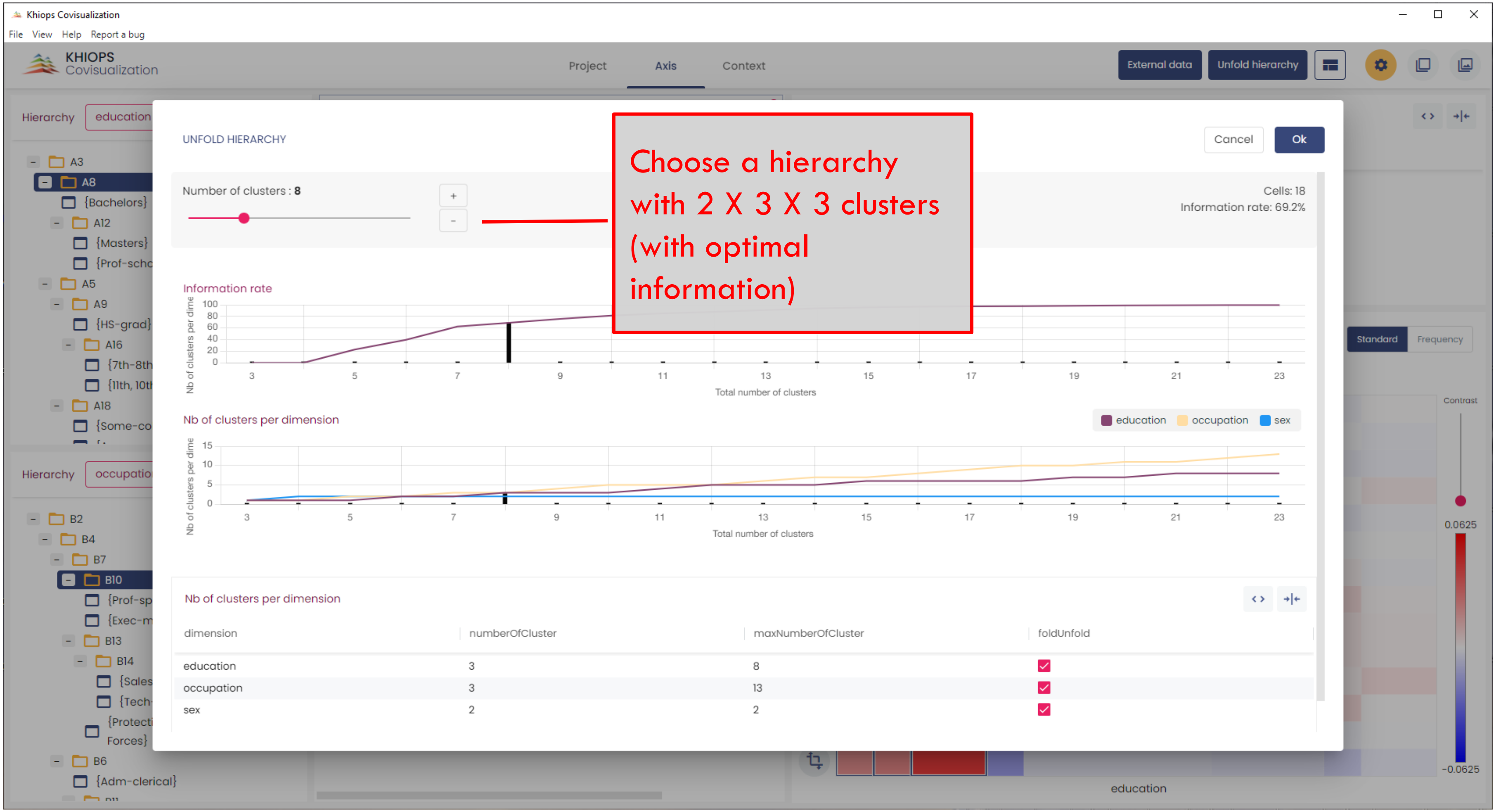
Name	Current Cluster	Nb Clusters
education	A8	8
occupation	B10	13
sex	{Male}	2

cluster	typicality
{Male}	1

☑ Conditional on context

Cell stats F= 5570 EF= 2565.2548 I= 0.1453 Total= 1.1937

Exploring a triclustering



Choose a hierarchy with 2 X 3 X 3 clusters (with optimal information)

UNFOLD HIERARCHY

Number of clusters : **8**

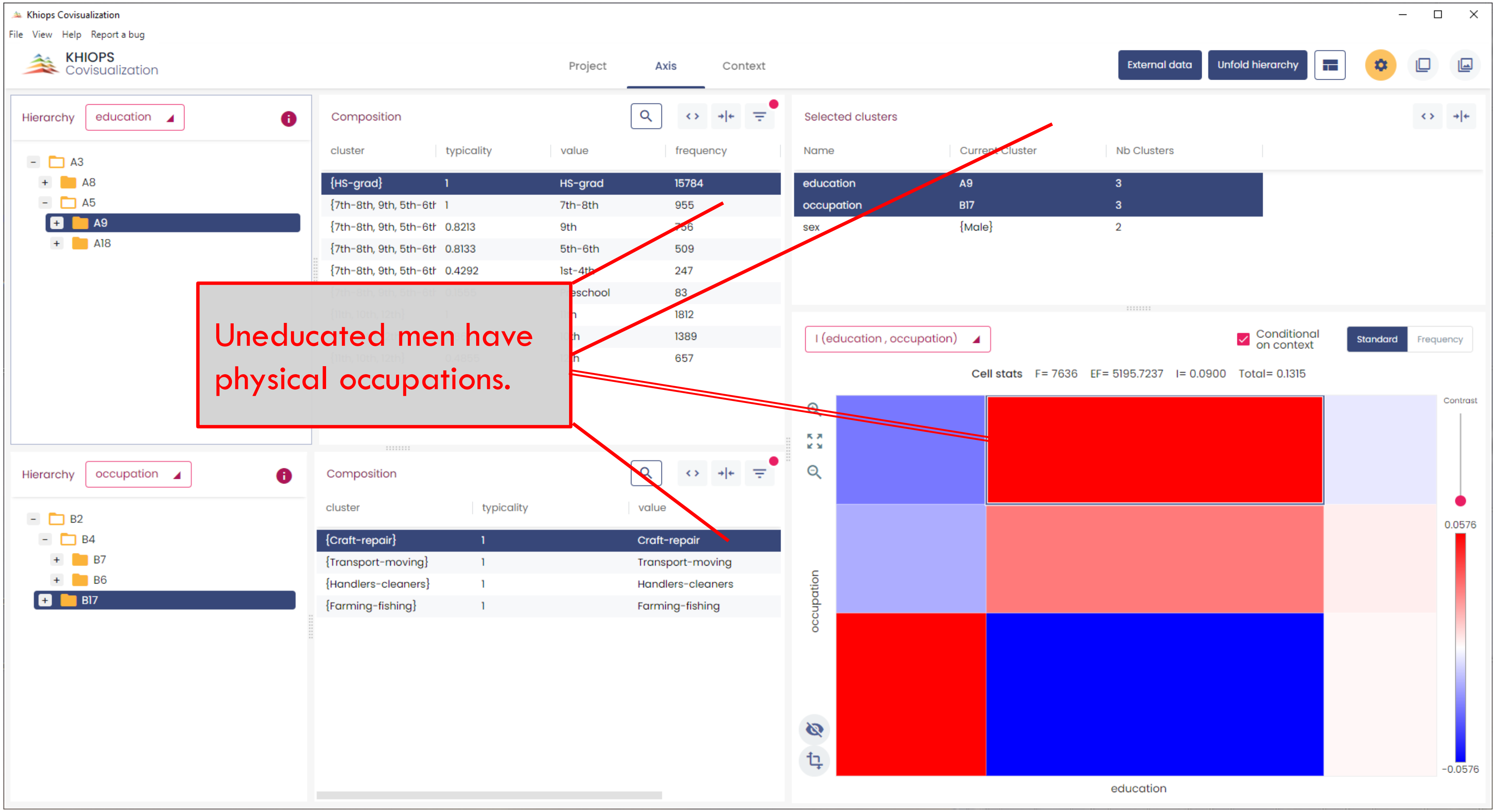
Cells: 18

Information rate: 69.2%

Nb of clusters per dimension

dimension	numberOfCluster	maxNumberOfCluster	foldUnfold
education	3	8	☑
occupation	3	13	☑
sex	2	2	☑

Exploring triclustering



cluster	typicality	value	frequency
{HS-grad}	1	HS-grad	15784
{7th-8th, 9th, 5th-6tł	1	7th-8th	955
{7th-8th, 9th, 5th-6tł	0.8213	9th	756
{7th-8th, 9th, 5th-6tł	0.8133	5th-6th	509
{7th-8th, 9th, 5th-6tł	0.4292	1st-4th	247
{7th-8th, 9th, 5th-6tł	0.1555	Preschool	83
{11th, 10th, 12th}	1	11th	1812
[illegible]	[illegible]	10th	1389
{11th, 10th, 12th}	0.4855	12th	657

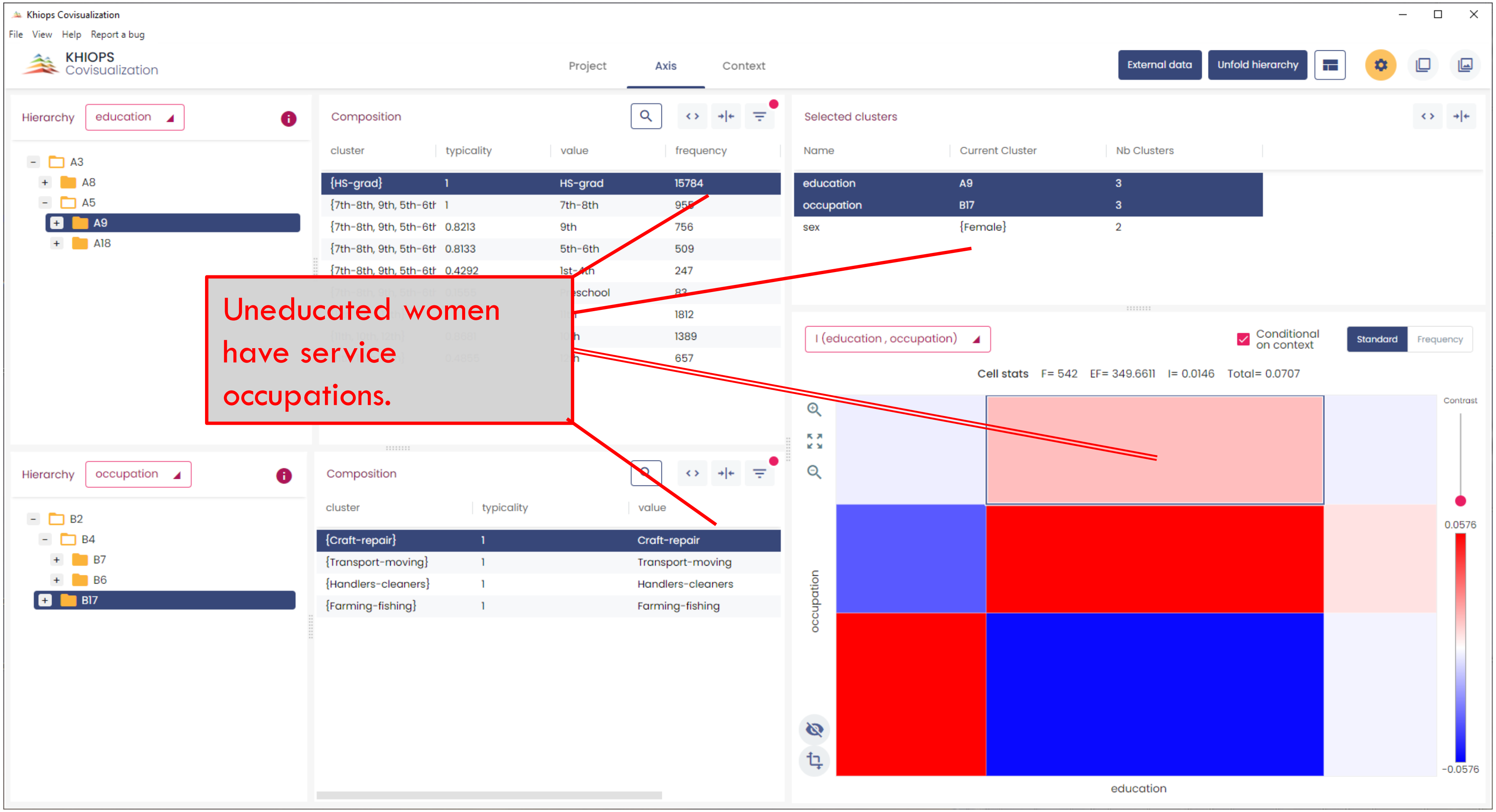
Name	Current Cluster	Nb Clusters
education	A9	3
occupation	B17	3
sex	{Male}	2

Cell stats F= 7636 EF= 5195.7237 I= 0.0900 Total= 0.1315

cluster	typicality	value
{Craft-repair}	1	Craft-repair
{Transport-moving}	1	Transport-moving
{Handlers-cleaners}	1	Handlers-cleaners
{Farming-fishing}	1	Farming-fishing

Uneducated men have physical occupations.

Exploring a triclustering



Composition

cluster	typicality	value	frequency
{HS-grad}	1	HS-grad	15784
{7th-8th, 9th, 5th-6th	1	7th-8th	955
{7th-8th, 9th, 5th-6th	0.8213	9th	756
{7th-8th, 9th, 5th-6th	0.8133	5th-6th	509
{7th-8th, 9th, 5th-6th	0.4292	1st-4th	247
{7th-8th, 9th, 5th-6th	0.1555	Preschool	83
{11th, 10th, 12th}	[illegible]	11th	1812
{11th, 10th, 12th}	0.8681	10th	1389
{11th, 10th, 12th}	0.4855	12th	657

Name	Current Cluster	Nb Clusters
education	A9	3
occupation	B17	3
sex	{Female}	2

I (education , occupation)

☑ Conditional on context

Standard Frequency

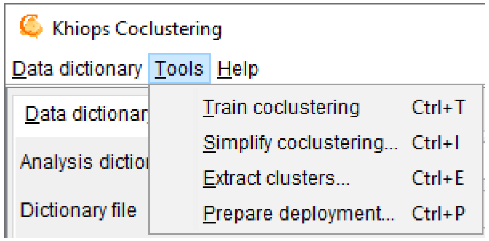
Cell stats F= 542 EF= 349.6611 I= 0.0146 Total= 0.0707

Composition

cluster	typicality	value
{Craft-repair}	1	Craft-repair
{Transport-moving}	1	Transport-moving
{Handlers-cleaners}	1	Handlers-cleaners
{Farming-fishing}	1	Farming-fishing

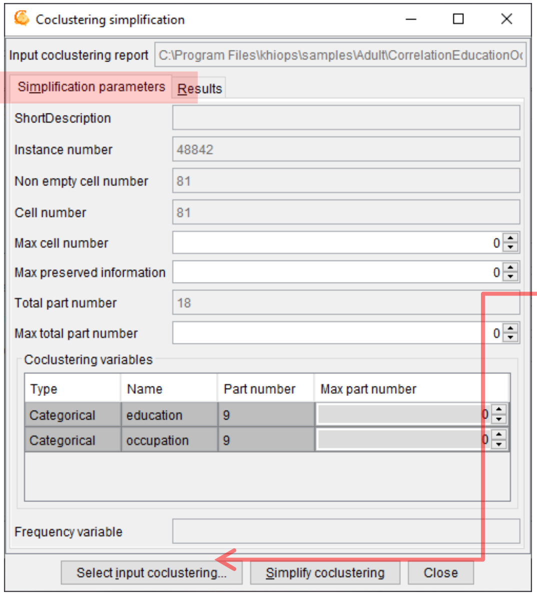
Uneducated women have service occupations.

Exploiting a coclustering model


Khiops Coclustering
Data dictionary
Tools
Help
Data dictionar
Analysis dictio
Dictionary file
Train coclustering Ctrl+T
Simplify coclustering... Ctrl+I
Extract clusters... Ctrl+E
Prepare deployment... Ctrl+P

- Tools menu
  - Train coclustering
    - Input: dictionary and database file
    - Train a coclustering model
  - Simplify coclustering
    - Input: coclustering model
    - Build a simplify coclustering model given user constraints
  - Extract clusters
    - Input: coclustering model
    - Extract clusters in a text file for a given coclustering variable
  - Prepare deployment
    - Input: dictionary and coclustering model
    - Enables the deployment of a coclustering model on new data by the means of a Khiops deployment dictionary
    - See multi-table section of the tutorial

Simplifying a coclustering model


Coclustering simplification
Simplification parameters
Results
Select input coclustering...
Simplify coclustering
Close

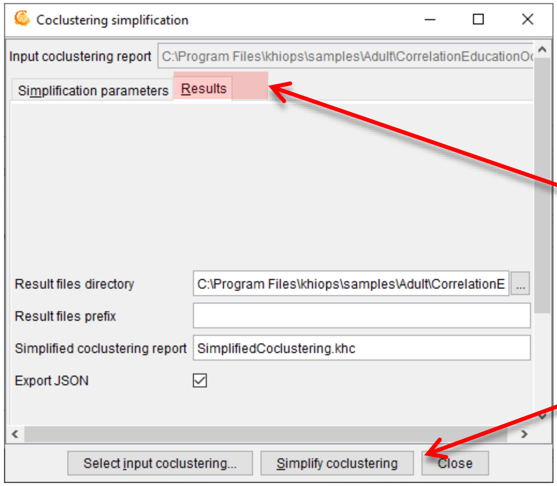
Input coclustering report: C:\Program Files\khiops\samples\Adult\CorrelationEducationOc

ShortDescription	
Instance number	48842
Non empty cell number	81
Cell number	81
Max cell number	0
Max preserved information	0
Total part number	18
Max total part number	0

Coclustering variables

Type	Name	Part number	Max part number
Categorical	education	9	0
Categorical	occupation	9	0

Frequency variable


Coclustering simplification
Simplification parameters
Results
Select input coclustering...
Simplify coclustering
Close

Input coclustering report: C:\Program Files\khiops\samples\Adult\CorrelationEducationOc

Result files directory	C:\Program Files\khiops\samples\Adult\CorrelationE
Result files prefix	
Simplified coclustering report	SimplifiedCoclustering.khc
Export JSON	☑

- Steps for coclustering model simplification

- 1- Select input coclustering (.khc)

- 2- Specify user simplification constraints
  - Max cell number :
    - max number of cells to keep in the simplified coclustering
  - Max preserved information
    - max percentage of information to keep in the simplified coclustering
  - Max total part number
    - max for the sum of the part number per coclustering variable
  - Per coclustering variables (in the array)

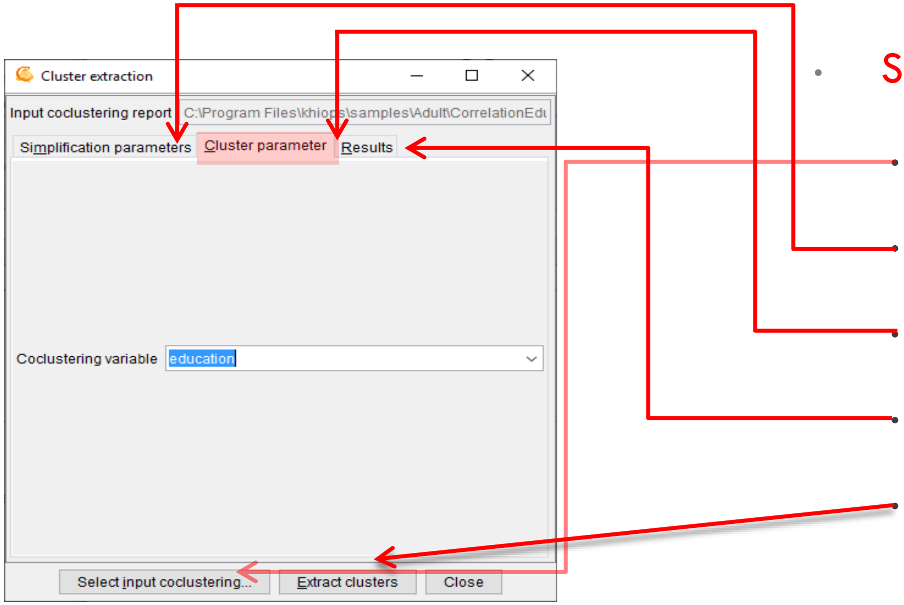
    Max part number
    - max number of part to keep for this variable in the simplified coclustering

  (0 : no constraint)

- 3- Select result files directory

- 4- Click on « *Simplify coclustering* »

Extracting clusters in a text file



Cluster extraction

Input coclustering report: C:\Program Files\khiops\samples\Adult\CorrelationEdu

Coclustering variable: education

- Steps for cluster extraction
  - 1- Select input coclustering (.khc)
  - 2- Specify user simplification constraints
  - 3- Select coclustering variable containing the clusters
  - 4- Select result files directory
  - 5- Click on « *Extract clusters* »

Output cluster file

- Text file with header line and separator tabulation
- Columns:
  - Cluster: name of the cluster (group of values)
  - Value: name of the value contained in the cluster
  - Frequency: frequency of the value
  - Typicality: interest measure of the value within its cluster

Exercise K, L …

K : Simplify previously built adult coclustering model

Keep 50% of the information in the model

➡ Explore the simplified analysis results with Khiops covisualization

L : Extract clusters from variable education of adult coclustering model

➡ Inspect the cluster file with a text editor

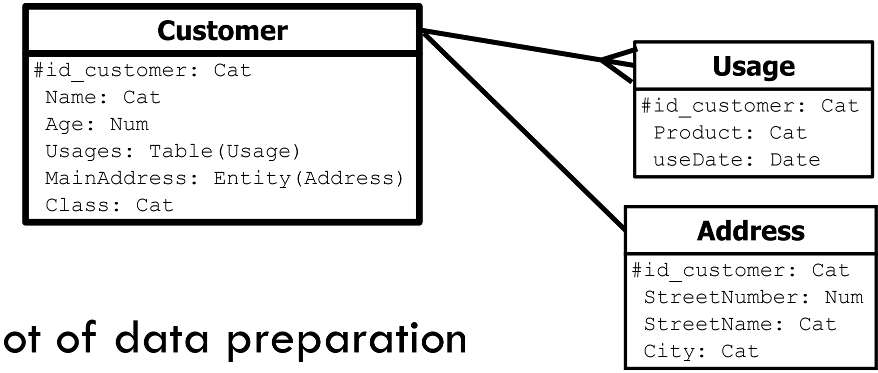
Multi-table functionalities

- Multi-table functionalities
  - Multi-table database
  - Automatic feature construction
  - Multi-table functionalities in Khiops and Khiops Coclustering

Why extending to multi-table?

- Why extending to multi-table?
  - Most data mining tools work on instances*variables flat tables
  - Real data often have a structure coming from databases
  - The input representation is richer using multi-table specification
  - Data mining methods may benefit from explicit richer domain description

- Real data is usually structured
  - Example
    - Marketing: Customer with shopping list
    - Web analytics: cookie with web log
    - Telecommunications: Customer with call detail records
    - Bioinformatics: DNA segments with ordered list of nucleotides
    - …



Customer

```
#id_customer: Cat
 Name: Cat
 Age: Num
 Usages: Table(Usage)
 MainAddress: Entity(Address)
 Class: Cat
```

Usage

```
#id_customer: Cat
 Product: Cat
 useDate: Date
```

Address

```
#id_customer: Cat
 StreetNumber: Num
 StreetName: Cat
 City: Cat
```

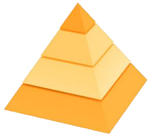
- Data mining with structured data requires a lot of data preparation
  - Constructing a representation in a flat table
    - Expert knowledge necessary to constructed new variables
    - Time expensive process to get a flat table usable for data analysis
  - This process is unreliable
    - Risk of missing informative variables
    - Risk of constructing and selecting irrelevant variables

Khiops multi-table

- Khiops can deal with multi-table databases
  - star schema: one root entity and several 0-1 or 0-n secondary entities
  - snowflake schemas and beyond
- Impact on Khiops
  - Multi-table dictionary
    - to describe star-schema input representation
  - Multi-table database
    - to store input data on multiple files
  - Feature construction language
    - to drive automatic feature construction
  - Sort functionality on large files
  - Powerful analytic functionalities
    - Automatic feature construction
    - Recoding of multi-table databases to get a flattened representation
    - Modeling and deployment at the multi-table level
- Impact on Khiops Coclustering
  - Deployment of coclustering models
    - For example, given a text*word coclustering model, assign new texts to their closest cluster

Khiops multi-table

- Khiops can deal with multi-table databases
  - star schema: one root entity and several 0-1 or 0-n secondary entities
  - snowflake schemas and beyond

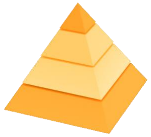


- Impact on Khiops
  - Multi-table dictionary
    - to describe star-schema input representation
  - Multi-table database
    - to store input data on multiple files
  - Feature construction language
    - to drive automatic feature construction
  - Sort functionality on large files
  - Powerful analytic functionalities
    - Automatic feature construction
    - Recoding of multi-table databases to get a flattened representation
    - Modeling and deployment at the multi-table level

- All other Khiops functionalities are available similarly
  - Classification, regression, correlation analysis
  - Deployment, recoding, evaluation
  - …

Khiops multi-table

- Khiops can deal with multi-table databases
  - star schema: one root entity and several 0-1 or 0-n secondary entities
  - snowflake schemas and beyond

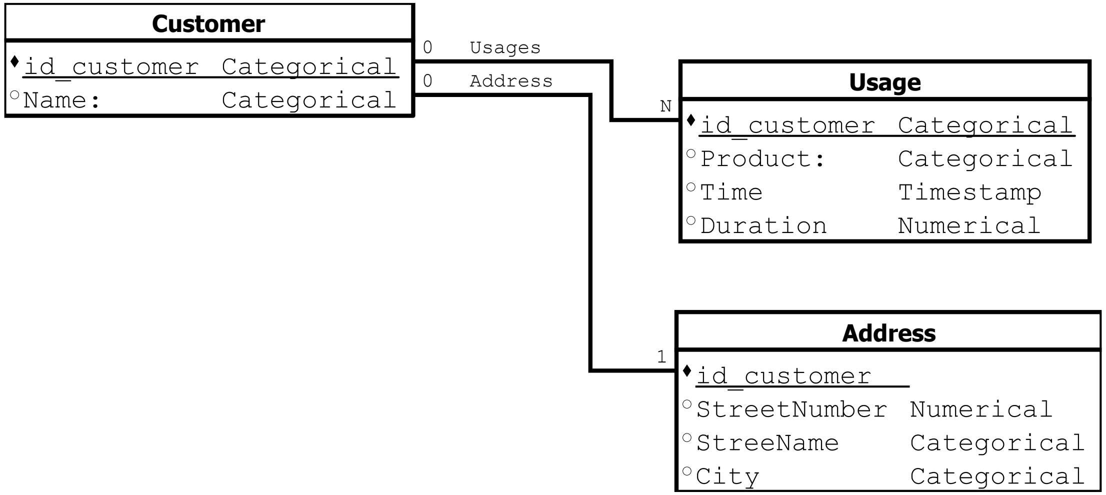


- Impact on Khiops
  - Multi-table dictionary
    - to describe star-schema input representation
  - Multi-table database
    - to store input data on multiple files
  - Feature construction language
    - to drive automatic feature construction
  - Sort functionality on large files
  - Powerful analytic functionalities
    - Automatic feature construction
    - Recoding of multi-table databases to get a flattened representation
    - Modeling and deployment at the multi-table level

- All other Khiops functionalities are available similarly
  - Classification, regression, correlation analysis
  - Deployment, recoding, evaluation
  - …

Star schema

- One root entity
  - secondary tables in 0-1 relationship: Entity
  - secondary tables in 0-n relationship: Table



Customer	
♦id_customer	Categorical
○Name:	Categorical

0 Usages

0 Address

N

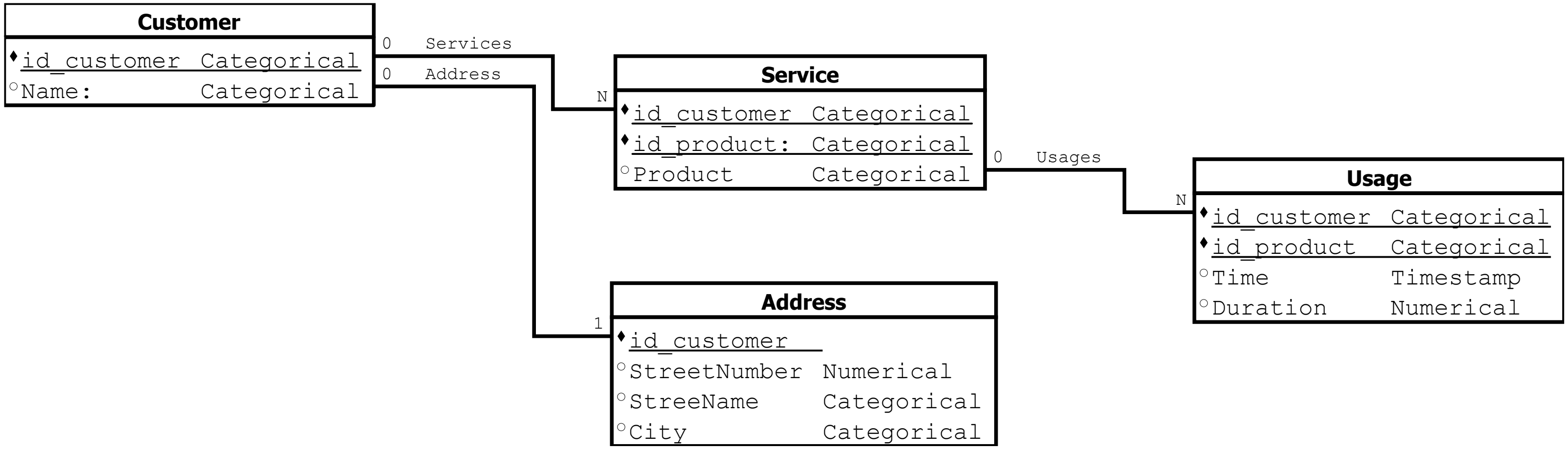
Usage	
♦id_customer	Categorical
○Product:	Categorical
○Time	Timestamp
○Duration	Numerical

1

Address	
♦id_customer	
○StreetNumber	Numerical
○StreeName	Categorical
○City	Categorical

Snowflake schema

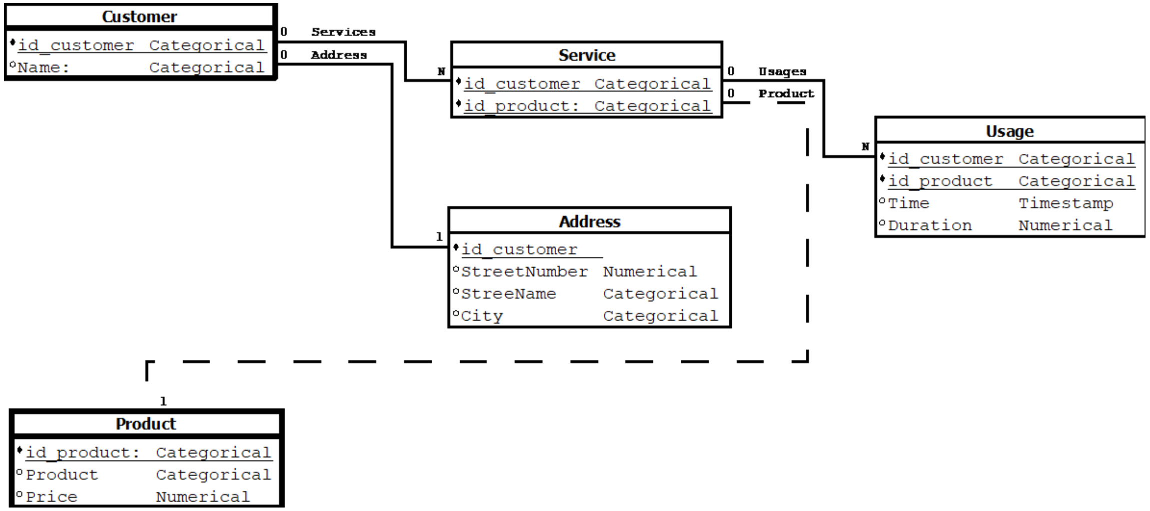
- One root entity
  - secondary tables in 0-1 relationship: Entity
  - secondary tables in 0-n relationship: Table
- Each table may have secondary tables



- Example in samples/Customer
  - detailed explanations in sample

External tables

- One root entity
  - secondary tables in 0-1 relationship: Entity
  - secondary tables in 0-n relationship: Table
- Each table can have secondary tables
- External tables
  - to reuse common table shared by all analysis entities
  - can be referenced from any table, with specific keys



Customer
- ♦id_customer Categorical
- ○Name: Categorical

Service
- ♦id_customer Categorical
- ♦id_product: Categorical

Usage
- ♦id_customer Categorical
- ♦id_product Categorical
- ○Time Timestamp
- ○Duration Numerical

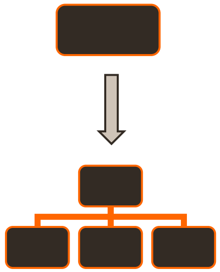
Address
- ♦id_customer
- ○StreetNumber Numerical
- ○StreeName Categorical
- ○City Categorical

Product
- ♦id_product: Categorical
- ○Product Categorical
- ○Price Numerical

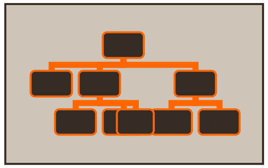
- Example in samples/CustomerExtended
  - detailed explanations in sample

Multi-table schemas: synthesis

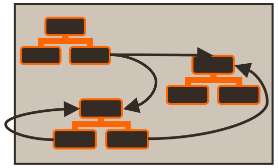
- Khiops 8.0:
  - from mono-table to star schema
    - Automatic variable construction
    - a technological disruption



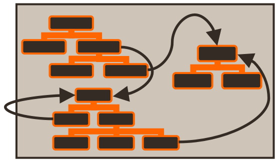
- Khiops 9.0:
  - extended data schema



    - Snowflake schema



    - External data



    - Multiple snowflake schema

Example of a multi-table database

French road accidents database

The **`AccidentsSummary`** is described using the following star schema:

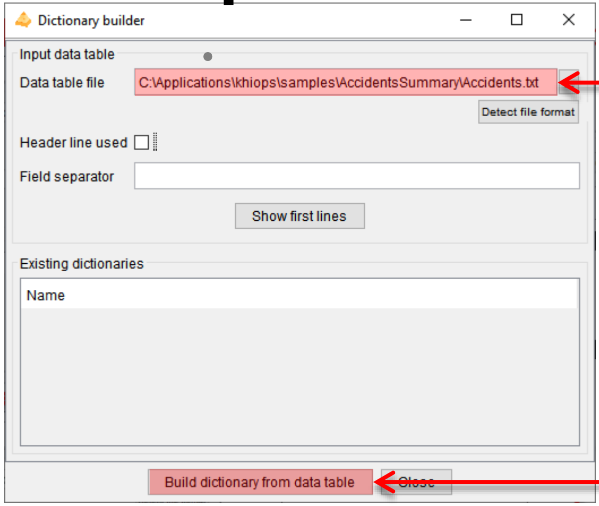
```
Accident
|
| -- 1:n -- Vehicle
```

Each accident has associated one or more vehicles. In the Khiops dictionary Accident-Vehicle 1:n relationship is described with the **`Table`** keyword. The key linking both tables is **`AccidentId`**.

Objective: predict fatal traffic accidents (target variable: `Gravity` field of `Accident` table)

Build a multi-table dictionary

- **Step 1:** Build one dictionary per data table



Dictionary builder

Input data table

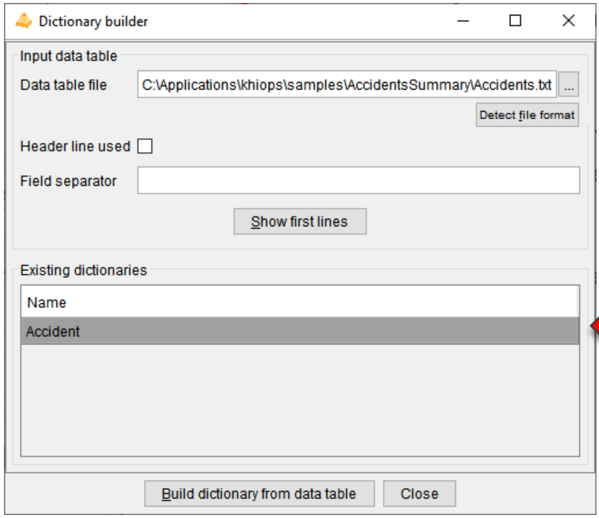
Data table file: C:\Applications\khiops\samples\AccidentsSummary\Accidents.txt

Header line used ☐

Field separator:

Existing dictionaries

Name



Dictionary builder

Input data table

Data table file: C:\Applications\khiops\samples\AccidentsSummary\Accidents.txt

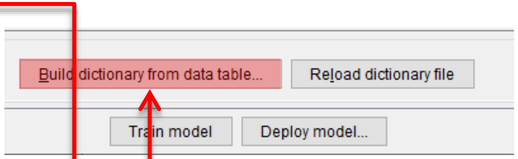
Header line used ☐

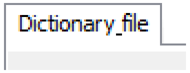
Field separator:

Existing dictionaries

Name
Accident

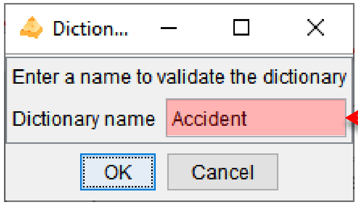
a : Build the first dictionary for the `Accidents.txt` table





1. In pane
   - Click on button *Build dictionary…*

2. Build the first dictionary
   - Specify the data table file: `Accidents.txt`
   - Build the dictionary
   - Specify the dictionary name : `Accident`



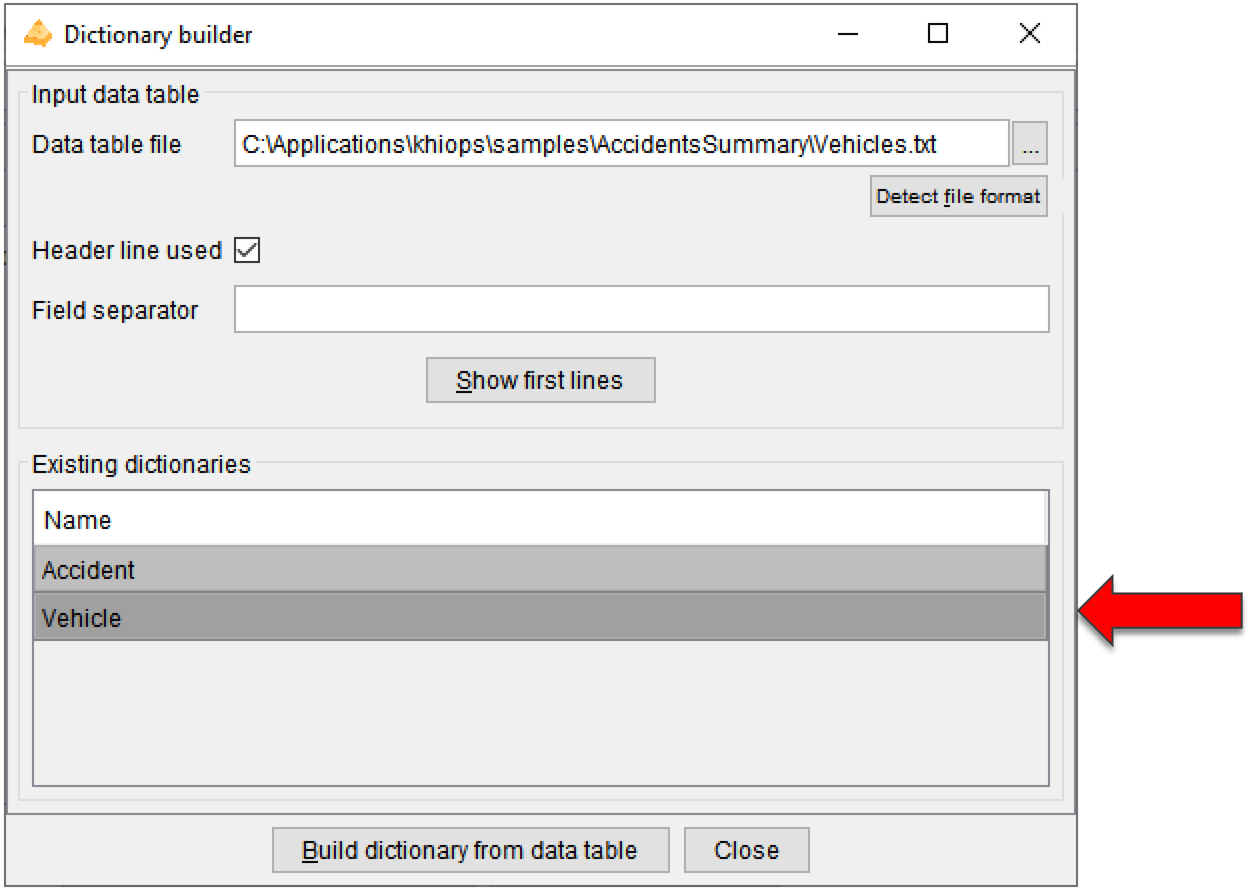
Diction...

Enter a name to validate the dictionary

Dictionary name: Accident

Build a multi-table dictionary

- **Step 1:** Build one dictionary per data table
  - b : Repeat for the `Vehicles.txt` table



Dictionary builder

Input data table

Data table file: C:\Applications\khiops\samples\AccidentsSummary\Vehicles.txt

Header line used ☑

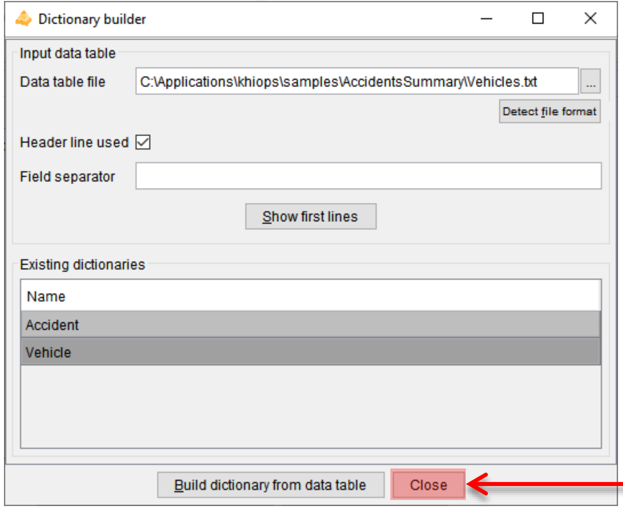
Field separator:

Existing dictionaries

Name
Accident
Vehicle

Build a multi-table dictionary

- **Step 1:** save the constructed dictionary into a `.kdic` file



Dictionary builder

Input data table

Data table file: C:\Applications\khiops\samples\AccidentsSummary\Vehicles.txt

Header line used ☑

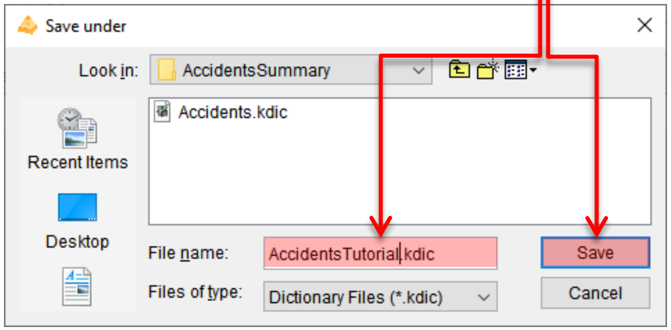
Field separator:

Existing dictionaries

Name
Accident
Vehicle

4. Save in dictionary file
   - Close
   - Specify the dictionary file name:

     `AccidentsTutorial.kdic`
   - Save



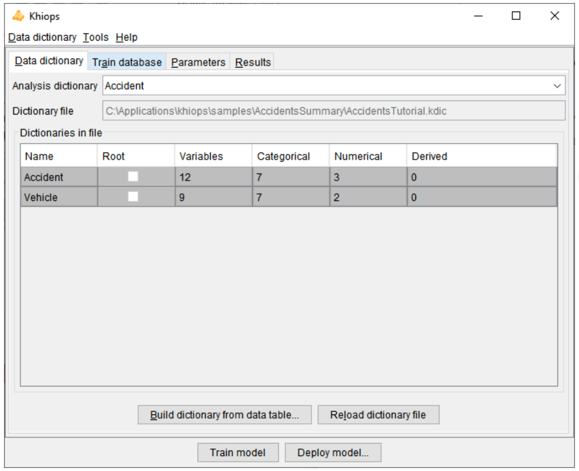
Save under

Look in: AccidentsSummary

Accidents.kdic

File name: AccidentsTutorial.kdic

Files of type: Dictionary Files (*.kdic)



Khiops

Analysis dictionary: Accident

Dictionary file: C:\Applications\khiops\samples\AccidentsSummary\AccidentsTutorial.kdic

Dictionaries in file

Name	Root	Variables	Categorical	Numerical	Derived
Accident	□	12	7	3	0
Vehicle	□	9	7	2	0

Build a multi-table dictionary

- **Step 2:** Describe the table relationships in the `.kdic` file

5. Open the dictionary file with a text editor

5.1 Specify the root entity

5.2 Fix the types of the fields in **green**

```
Root Dictionary Accident(AccidentId)
{
        Categorical AccidentId;
        Categorical Gravity;
        Date   Date;
        Time   Hour;
        Categorical Light;
        Categorical Department;
        Categorical Commune;
        Categorical InAgglomeration;
        Categorical IntersectionType;
        Categorical Weather;
        Categorical CollisionType;
        Categorical PostalAddress;
        Table(Vehicle) Vehicles;
};

Dictionary  Vehicle(AccidentId, VehicleId)
{
        Categorical AccidentId;
        Categorical VehicleId;
        Categorical Direction;
        Categorical Category;
        Numerical   PassengerNumber;
        Categorical FixedObstacle;
        Categorical MobileObstacle;
        Categorical ImpactPoint;
        Categorical Maneuver;
};
```

5.3 Specify the key fields for each entity

Key fields must be Categorical and not derived

5.4 Specify the relation between the root entity and the secondary entity

Add a variable per relation to root dictionary

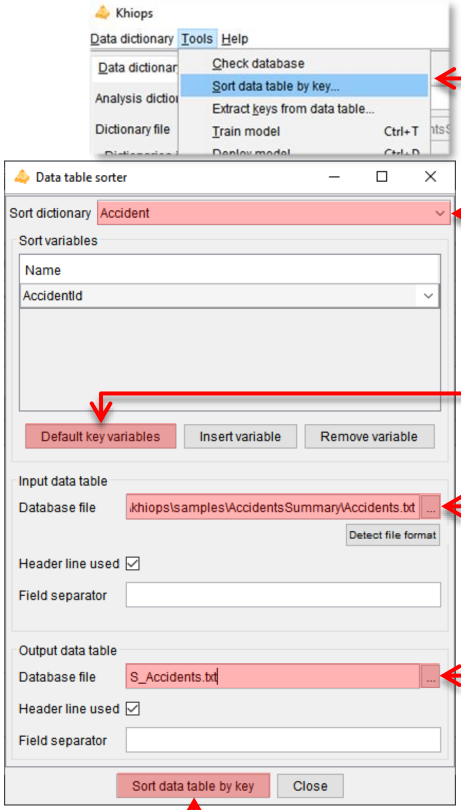
- *Table* for 0-n relationship
- *Entity* for 0-1 relationship

6. Save the dictionary file

Sort data table files (if necessary)

For multi-table analyses data table files must be sorted by their keys

- Sorting is done only once before any Khiops analysis
  - Note: Records of the root table **must** be unique by key
- It is necessary for efficiency, specially when treating large databases
  - Records of the root and secondary tables are read synchronously from their data table files



Sort dictionary: Accident

Sort variables

Name
AccidentId

Input data table

Database file: khiops\samples\AccidentsSummary\Accidents.txt

Header line used ☑

Field separator:

Output data table

Database file: S_Accidents.txt

Header line used ☑

Field separator:

1. Open the multi-table dictionary file

   This allows to obtain the definition of the tables to sort

2. Menu: *Tools* → *Sort data table by key*

   In the *Data table sorter* window, for each data table to sort

2.1 Specify the sort dictionary

2.2 Specify the sort variables

Default key variables to use the keys defined in dictionary

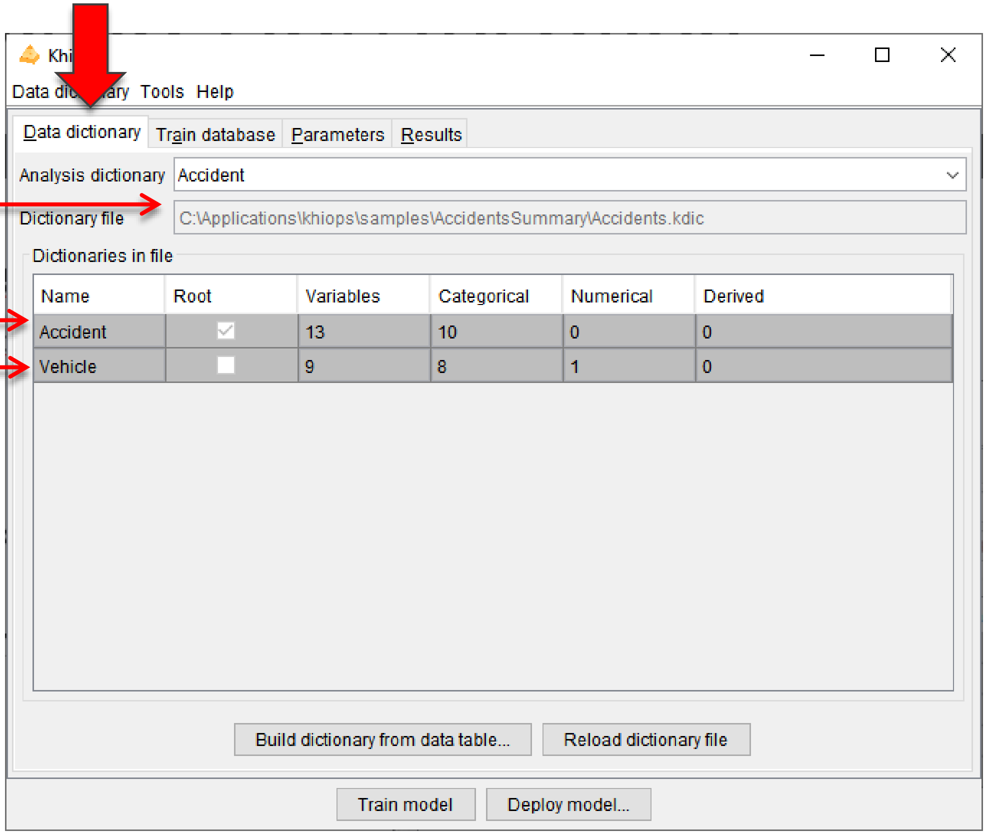
2.3 Specify the input and output data table files

2.4 Sort

- The output file is sorted by key
- All native variables are kept (used or not in the dictionary)
- Derived variables are ignored

Supervised classification

- **Step 1, bis :** *Open the Accidents.kdic dictionary file*



Analysis dictionary: Accident

Dictionary file: C:\Applications\khiops\samples\AccidentsSummary\Accidents.kdic

Name	Root	Variables	Categorical	Numerical	Derived
Accident	☑	13	10	0	0
Vehicle	□	9	8	1	0

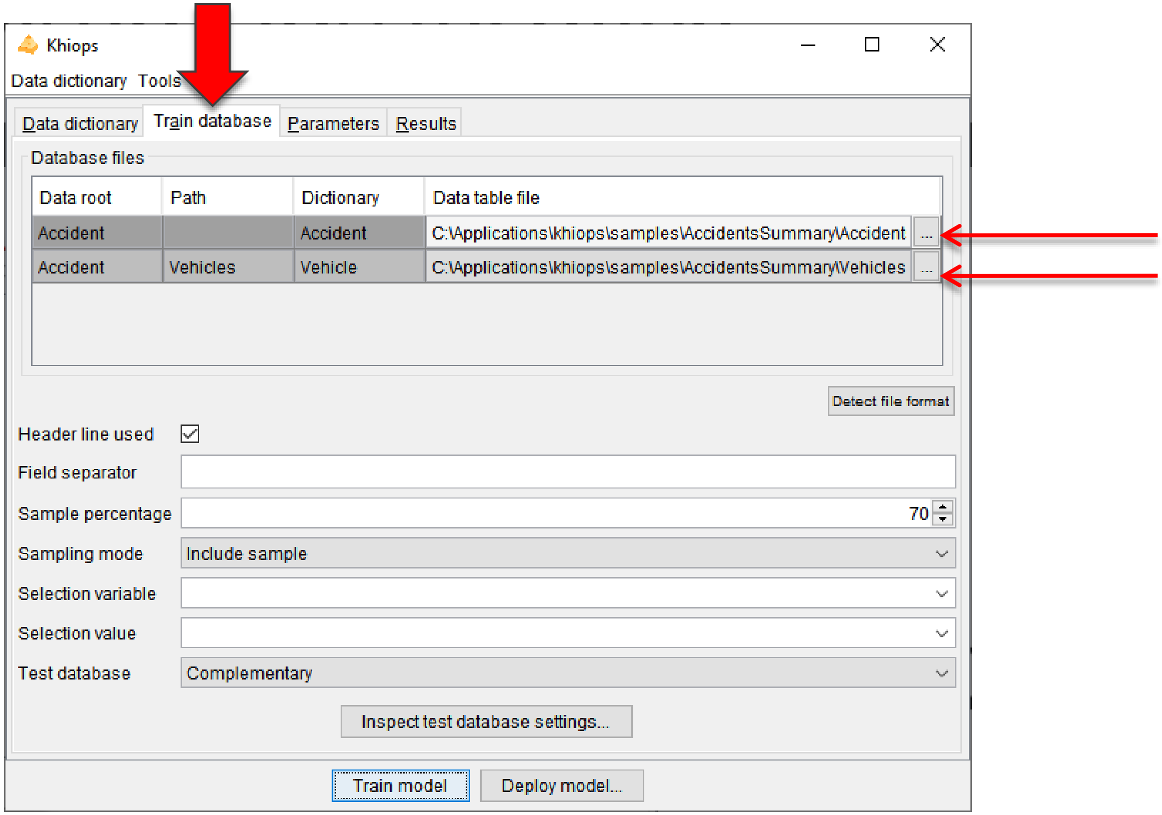
Analysis dictionary

Root entity

Secondary entity

Supervised classification

- **Step 2 :** Specify train and test databases
  - Specify the root and secondary data table files



Database files

Data root	Path	Dictionary	Data table file
Accident		Accident	C:\Applications\khiops\samples\AccidentsSummary\Accident
Accident	Vehicles	Vehicle	C:\Applications\khiops\samples\AccidentsSummary\Vehicles

Header line used	☑
Field separator	
Sample percentage	70
Sampling mode	Include sample
Selection variable	
Selection value	
Test database	Complementary

Root data table files

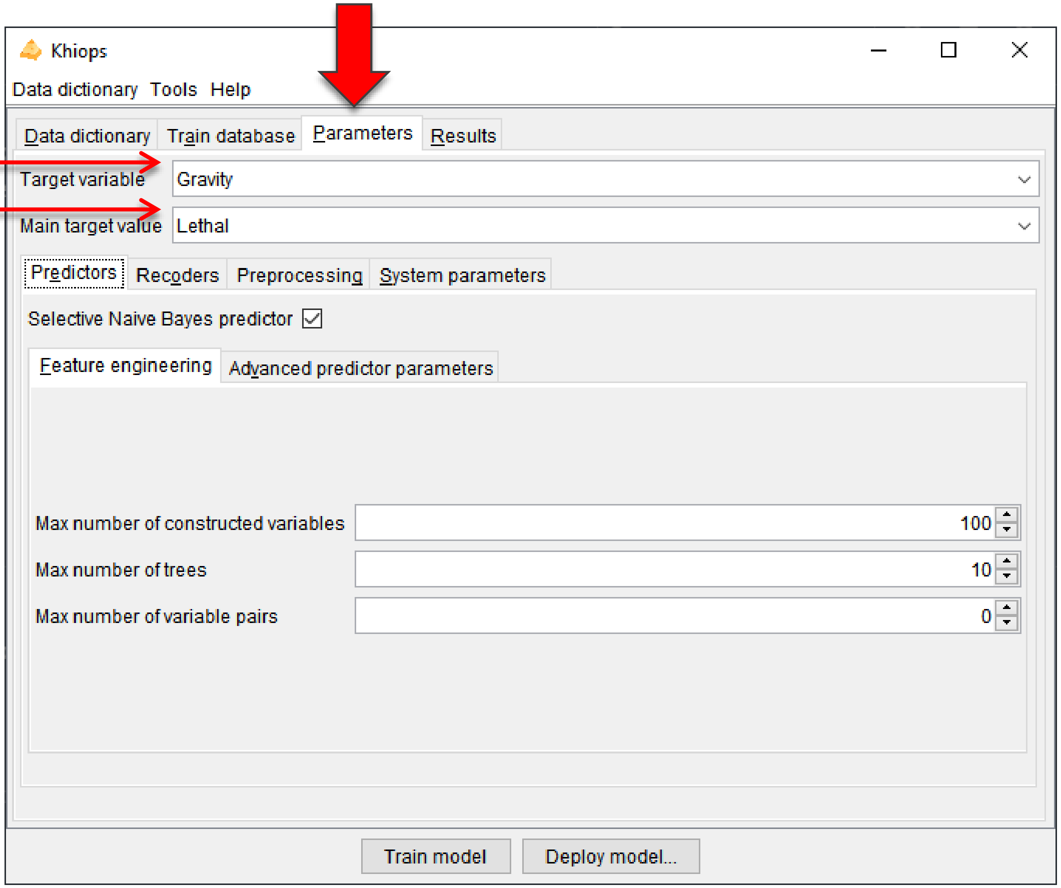
Secondary data table file

Supervised classification

- **Step 3 :** Parameters

Target variable

Main target value



Target variable: Gravity

Main target value: Lethal

Selective Naive Bayes predictor ☑

Max number of constructed variables: 100

Max number of trees: 10

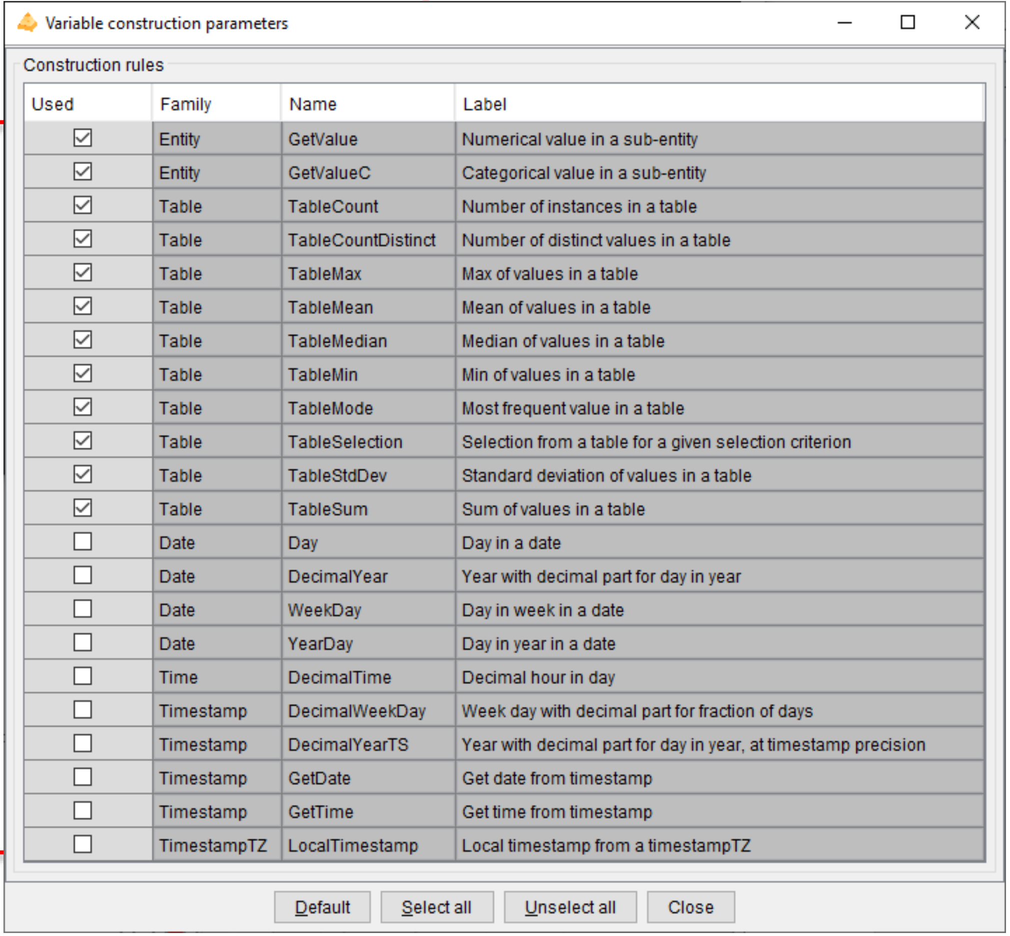
Max number of variable pairs: 0

Supervised classification

- **Step 4 :** Variable construction parameters

Optional

Choice of construction rules

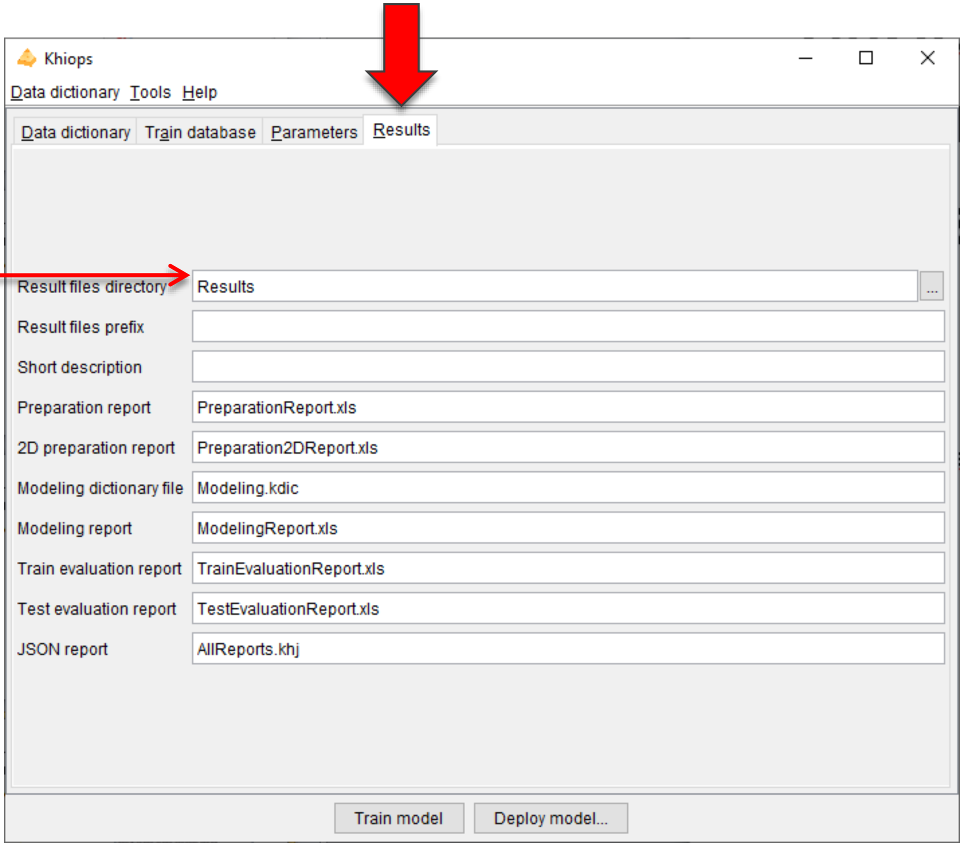


Used	Family	Name	Label
☑	Entity	GetValue	Numerical value in a sub-entity
☑	Entity	GetValueC	Categorical value in a sub-entity
☑	Table	TableCount	Number of instances in a table
☑	Table	TableCountDistinct	Number of distinct values in a table
☑	Table	TableMax	Max of values in a table
☑	Table	TableMean	Mean of values in a table
☑	Table	TableMedian	Median of values in a table
☑	Table	TableMin	Min of values in a table
☑	Table	TableMode	Most frequent value in a table
☑	Table	TableSelection	Selection from a table for a given selection criterion
☑	Table	TableStdDev	Standard deviation of values in a table
☑	Table	TableSum	Sum of values in a table
☐	Date	Day	Day in a date
☐	Date	DecimalYear	Year with decimal part for day in year
☐	Date	WeekDay	Day in week in a date
☐	Date	YearDay	Day in year in a date
☐	Time	DecimalTime	Decimal hour in day
☐	Timestamp	DecimalWeekDay	Week day with decimal part for fraction of days
☐	Timestamp	DecimalYearTS	Year with decimal part for day in year, at timestamp precision
☐	Timestamp	GetDate	Get date from timestamp
☐	Timestamp	GetTime	Get time from timestamp
☐	TimestampTZ	LocalTimestamp	Local timestamp from a timestampTZ

Supervised classification

- **Step 5 :** Analysis results

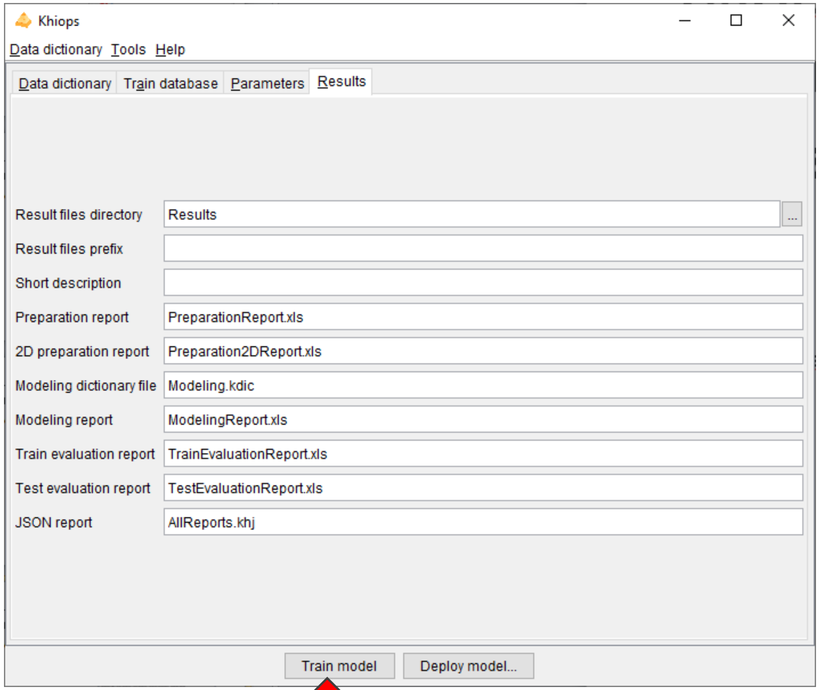
Results files directory



Result files directory	Results
Result files prefix	
Short description	
Preparation report	PreparationReport.xls
2D preparation report	Preparation2DReport.xls
Modeling dictionary file	Modeling.kdic
Modeling report	ModelingReport.xls
Train evaluation report	TrainEvaluationReport.xls
Test evaluation report	TestEvaluationReport.xls
JSON report	AllReports.khj

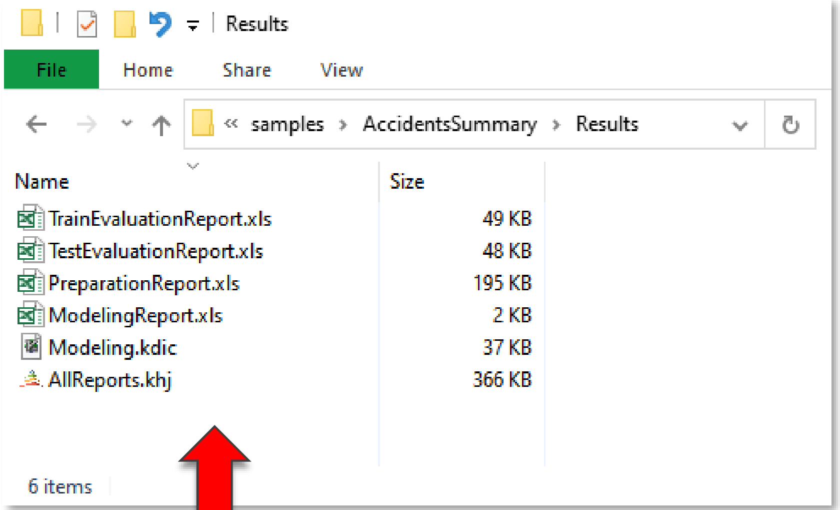
Supervised classification

- **Step 6 : Start the analysis**


Khiops
Data dictionary Tools Help
Data dictionary Train database Parameters Results
Train model
Deploy model...

Result files directory	Results
Result files prefix	
Short description	
Preparation report	PreparationReport.xls
2D preparation report	Preparation2DReport.xls
Modeling dictionary file	Modeling.kdic
Modeling report	ModelingReport.xls
Train evaluation report	TrainEvaluationReport.xls
Test evaluation report	TestEvaluationReport.xls
JSON report	AllReports.khj

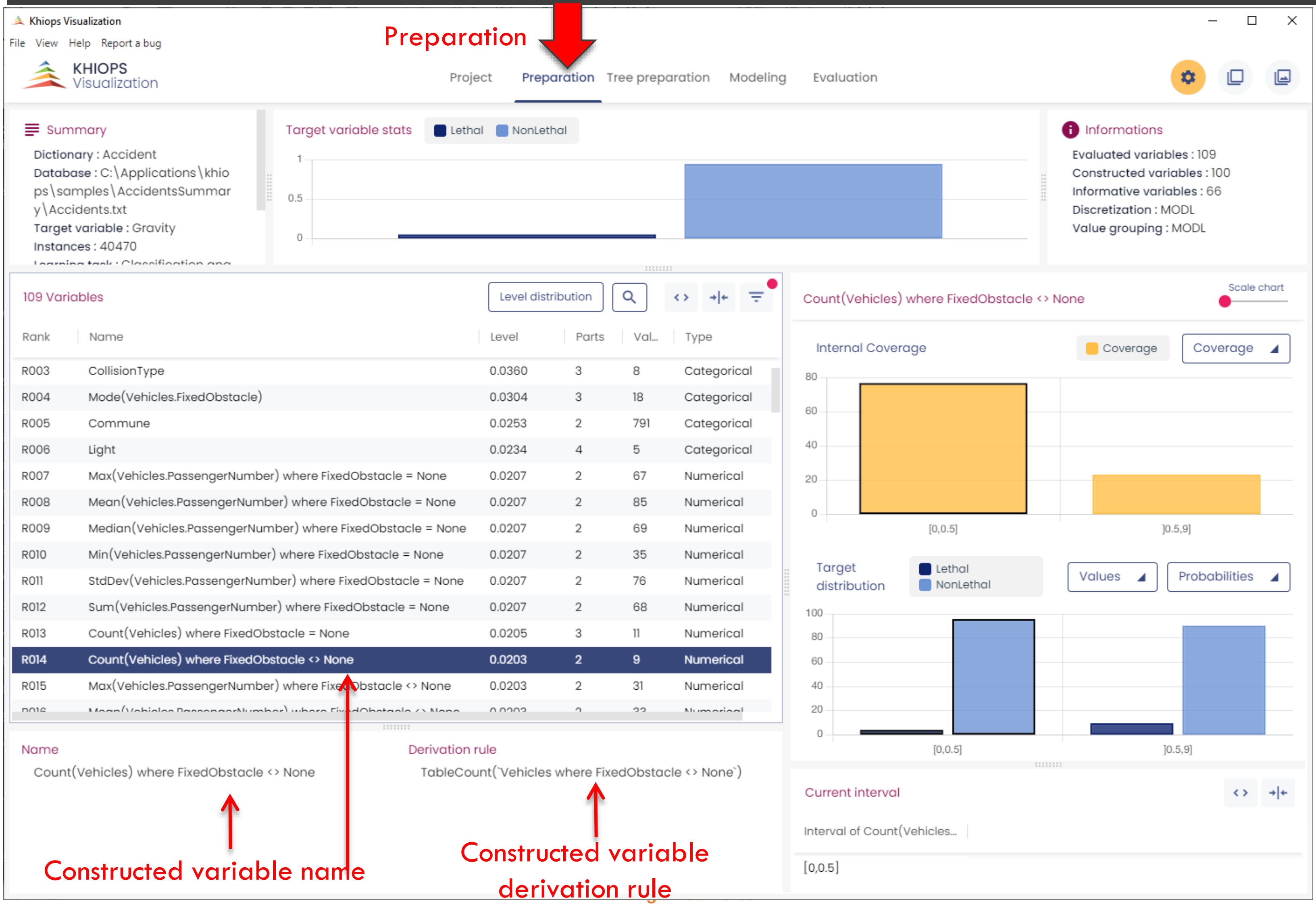
1 - Train model



Name	Size
TrainEvaluationReport.xls	49 KB
TestEvaluationReport.xls	48 KB
PreparationReport.xls	195 KB
ModelingReport.xls	2 KB
Modeling.kdic	37 KB
AllReports.khj	366 KB

2 - Inspect the results using Khiops Visualization
(double-click on *.khj* file)

Exploratory of classification results using Khiops Visualization



Preparation

Summary

Dictionary : Accident
Database : C:\Applications\khiops\samples\AccidentsSummary\Accidents.txt
Target variable : Gravity
Instances : 40470
Learning task : Classification ana

Informations

Evaluated variables : 109
Constructed variables : 100
Informative variables : 66
Discretization : MODL
Value grouping : MODL

109 Variables

Rank	Name	Level	Parts	Val...	Type
R003	CollisionType	0.0360	3	8	Categorical
R004	Mode(Vehicles.FixedObstacle)	0.0304	3	18	Categorical
R005	Commune	0.0253	2	791	Categorical
R006	Light	0.0234	4	5	Categorical
R007	Max(Vehicles.PassengerNumber) where FixedObstacle = None	0.0207	2	67	Numerical
R008	Mean(Vehicles.PassengerNumber) where FixedObstacle = None	0.0207	2	85	Numerical
R009	Median(Vehicles.PassengerNumber) where FixedObstacle = None	0.0207	2	69	Numerical
R010	Min(Vehicles.PassengerNumber) where FixedObstacle = None	0.0207	2	35	Numerical
R011	StdDev(Vehicles.PassengerNumber) where FixedObstacle = None	0.0207	2	76	Numerical
R012	Sum(Vehicles.PassengerNumber) where FixedObstacle = None	0.0207	2	68	Numerical
R013	Count(Vehicles) where FixedObstacle = None	0.0205	3	11	Numerical
R014	Count(Vehicles) where FixedObstacle <> None	0.0203	2	9	Numerical
R015	Max(Vehicles.PassengerNumber) where FixedObstacle <> None	0.0203	2	31	Numerical
R016	Mean(Vehicles.PassengerNumber) where FixedObstacle <> None	0.0203	2	33	Numerical

Count(Vehicles) where FixedObstacle <> None

Name: Count(Vehicles) where FixedObstacle <> None

Derivation rule: TableCount(`Vehicles where FixedObstacle <> None`)

Current interval

Interval of Count(Vehicles...

[0,0.5]

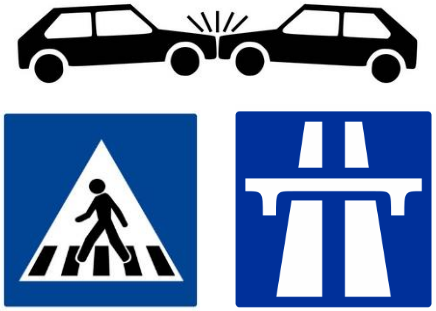
Constructed variable name

Constructed variable derivation rule

French road accidents database (full version)

This the full version of the `AccidentsSummary` dataset.
It is described using the following snowflake schema:

```
Accident
 |
 | -- 1:n -- Vehicle
 |              |
 |              |-- 1:n -- User
 |
 | -- 1:1 -- Place
```



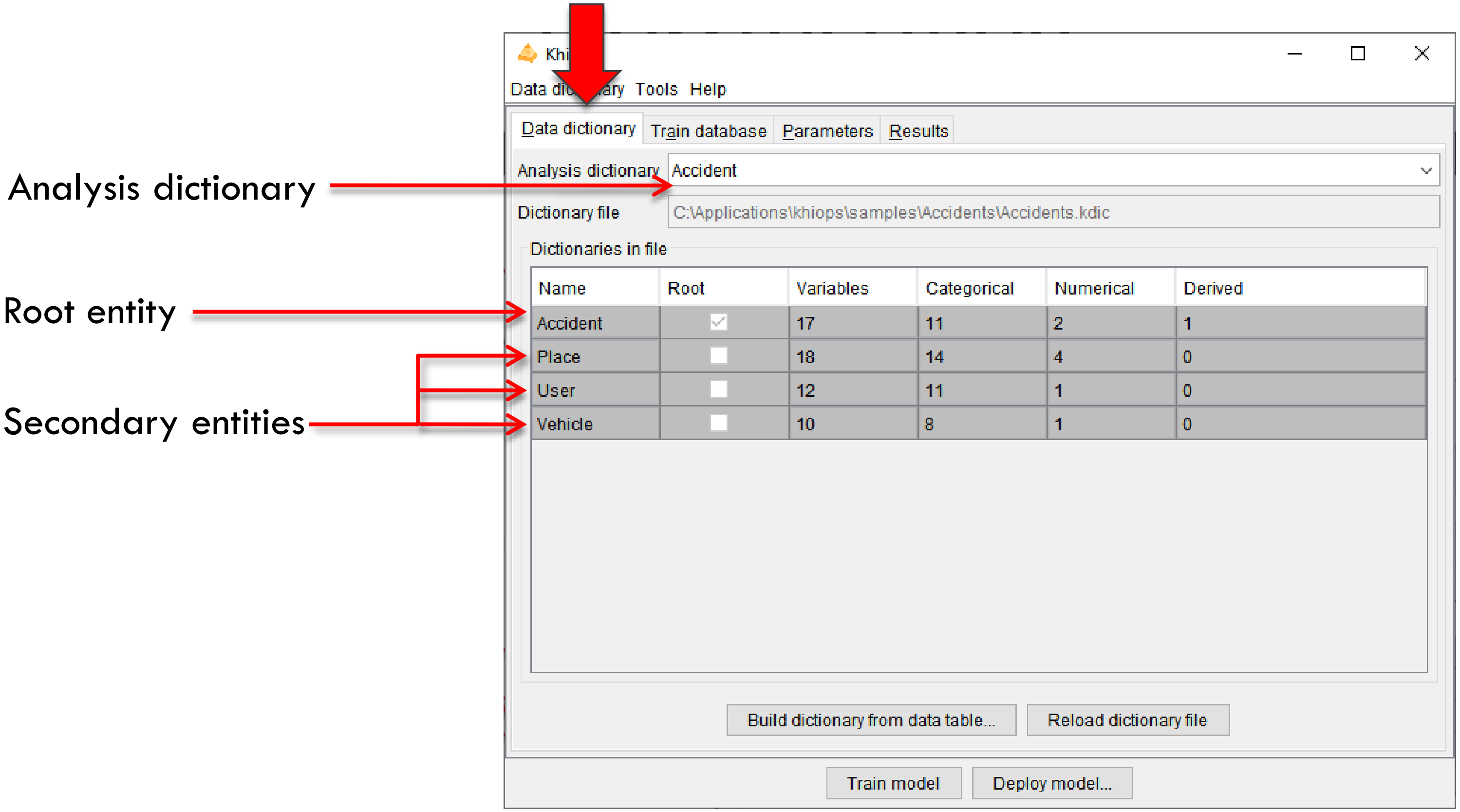
Each accident has associated one or more vehicles and one unique place. The vehicles involved in an accident have in turn associated one or more road users (passengers and pedestrians).

In the Khiops dictionary the Accident-Place relationship (1:1) is described with the `Entity` keyword, whereas the Accident-Vehicle and Vehicle-User relationships (1:n) with the `Table` keyword.

Objective: predict fatal traffic accidents (target variable: `Gravity` field of `Accident` table)

Supervised classification

- **Step 1 :** Open the *Accidents.kdic* dictionary file



Analysis dictionary

Root entity

Secondary entities

Analysis dictionary: Accident

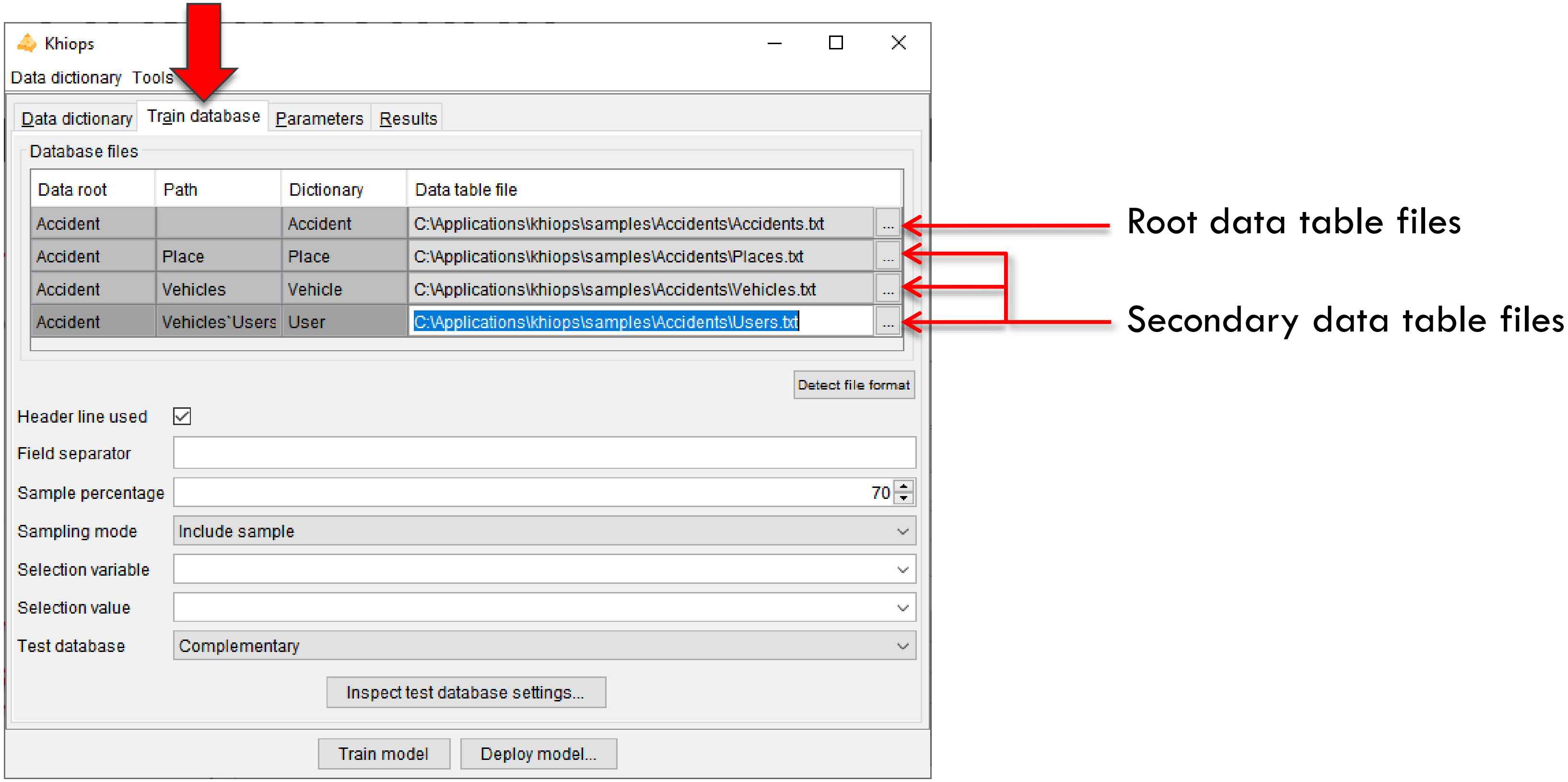
Dictionary file: C:\Applications\khiops\samples\Accidents\Accidents.kdic

Dictionaries in file

Name	Root	Variables	Categorical	Numerical	Derived
Accident	☑	17	11	2	1
Place	□	18	14	4	0
User	□	12	11	1	0
Vehicle	□	10	8	1	0

Supervised classification

- **Step 2 : Specify train and test databases**
  - Root and other data table files have to be specified



Database files

Data root	Path	Dictionary	Data table file
Accident		Accident	C:\Applications\khiops\samples\Accidents\Accidents.txt
Accident	Place	Place	C:\Applications\khiops\samples\Accidents\Places.txt
Accident	Vehicles	Vehicle	C:\Applications\khiops\samples\Accidents\Vehicles.txt
Accident	Vehicles`Users	User	C:\Applications\khiops\samples\Accidents\Users.txt

Header line used ☑

Field separator

Sample percentage 70

Sampling mode: Include sample

Selection variable

Selection value

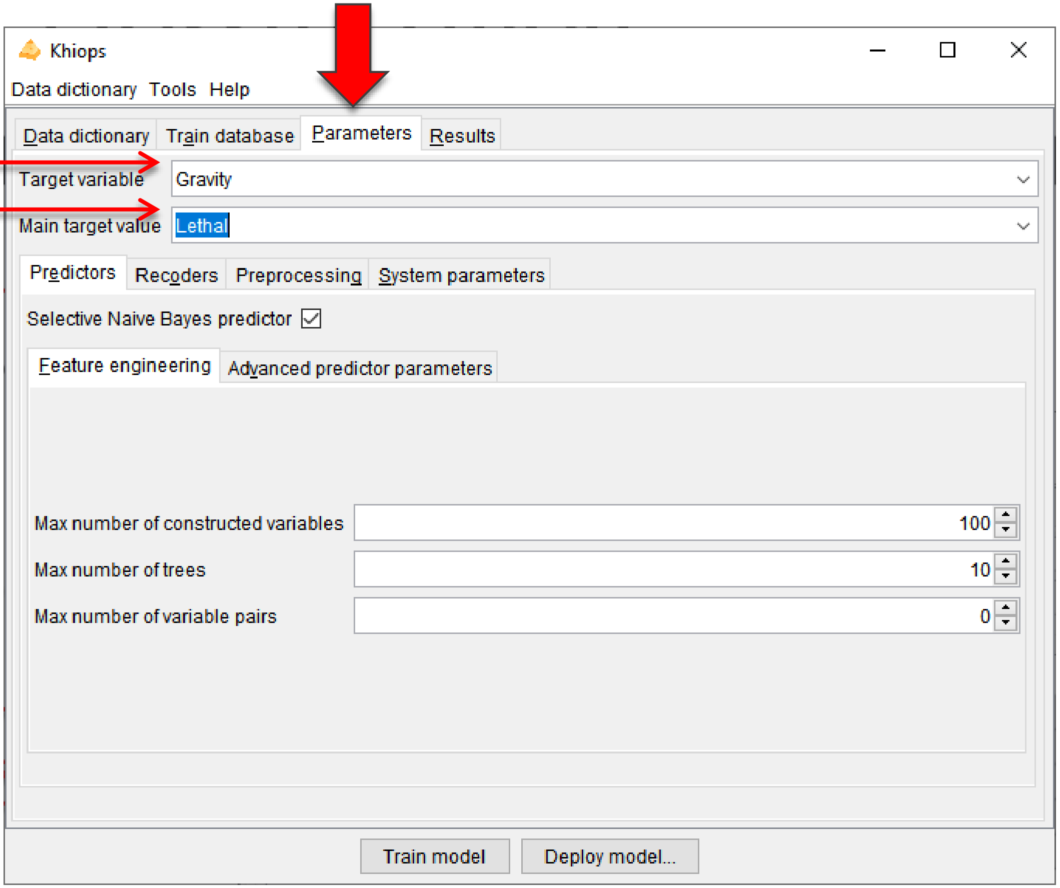
Test database: Complementary

Supervised classification

- **Step 3 :** Parameters

Target variable

Main target value



Target variable: Gravity

Main target value: Lethal

Selective Naive Bayes predictor ☑

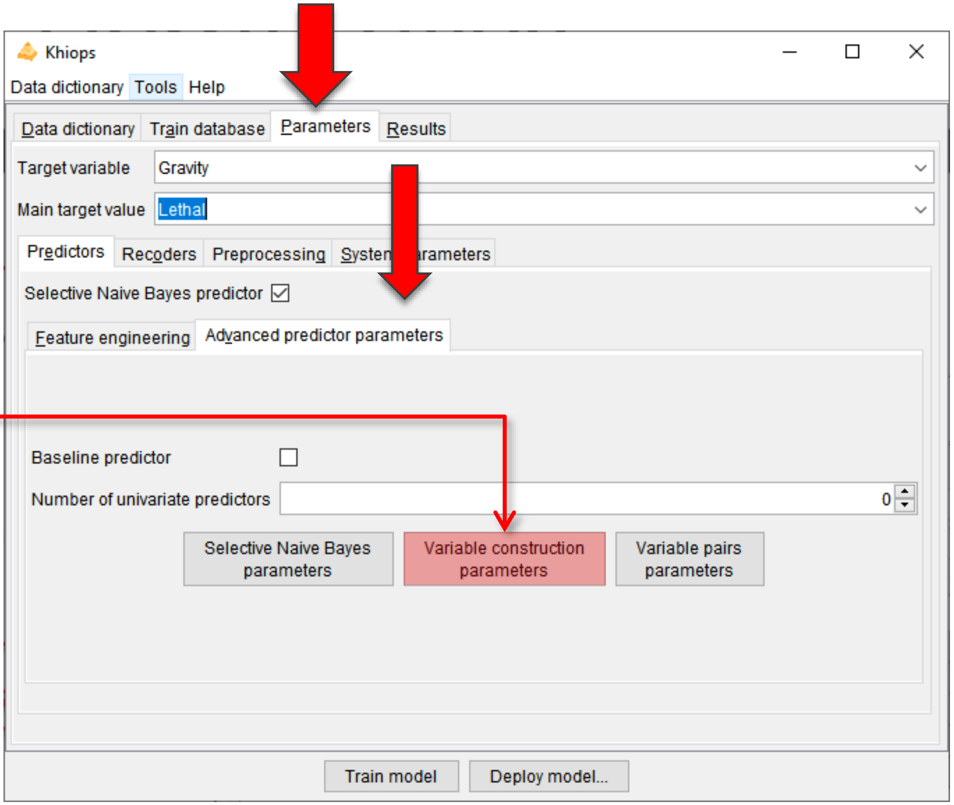
Max number of constructed variables: 100

Max number of trees: 10

Max number of variable pairs: 0

Supervised classification

- **Step 4 :** Variable construction parameters



Khiops

Data dictionary Tools Help

Data dictionary | Train database | Parameters | Results

Target variable: Gravity

Main target value: Lethal

Predictors | Recoders | Preprocessing | System parameters

Selective Naive Bayes predictor ☑

Feature engineering | Advanced predictor parameters

Baseline predictor □

Number of univariate predictors: 0

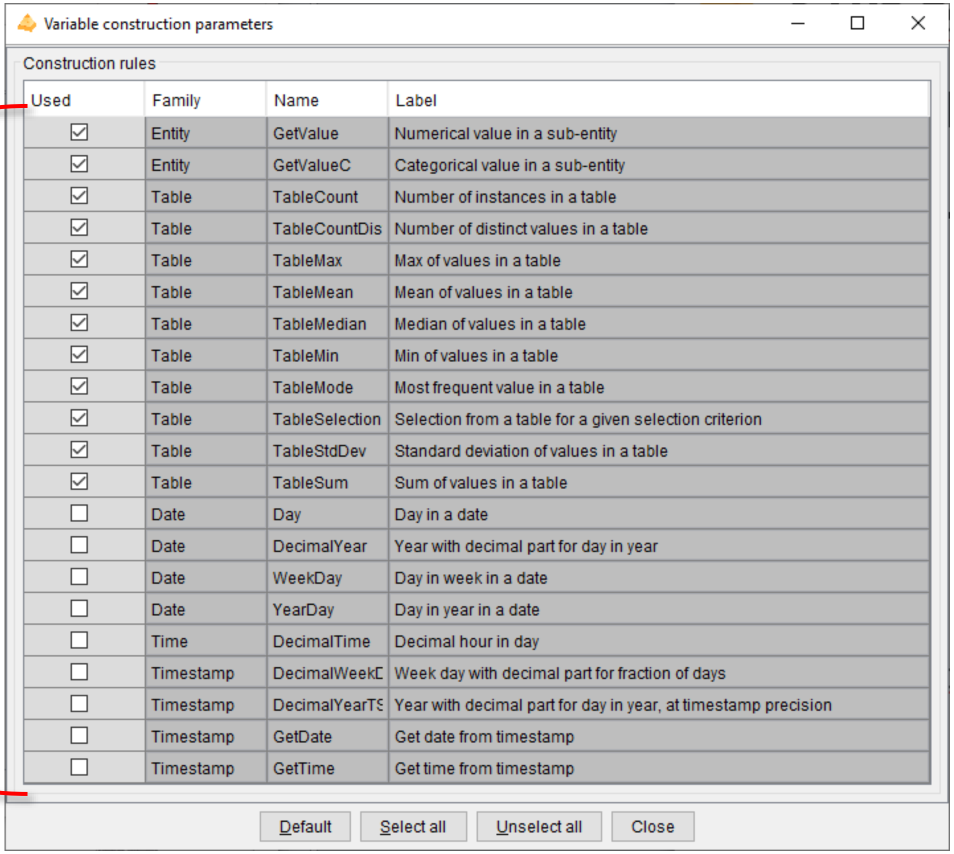
Selective Naive Bayes parameters
Variable construction parameters
Variable pairs parameters
Train model
Deploy model...

Optional

Choice of construction rules

Supervised classification

- **Step 4 :** Variable construction parameters



Variable construction parameters

Construction rules

Used	Family	Name	Label
☑	Entity	GetValue	Numerical value in a sub-entity
☑	Entity	GetValueC	Categorical value in a sub-entity
☑	Table	TableCount	Number of instances in a table
☑	Table	TableCountDis	Number of distinct values in a table
☑	Table	TableMax	Max of values in a table
☑	Table	TableMean	Mean of values in a table
☑	Table	TableMedian	Median of values in a table
☑	Table	TableMin	Min of values in a table
☑	Table	TableMode	Most frequent value in a table
☑	Table	TableSelection	Selection from a table for a given selection criterion
☑	Table	TableStdDev	Standard deviation of values in a table
☑	Table	TableSum	Sum of values in a table
□	Date	Day	Day in a date
□	Date	DecimalYear	Year with decimal part for day in year
□	Date	WeekDay	Day in week in a date
□	Date	YearDay	Day in year in a date
□	Time	DecimalTime	Decimal hour in day
□	Timestamp	DecimalWeekD	Week day with decimal part for fraction of days
□	Timestamp	DecimalYearTS	Year with decimal part for day in year, at timestamp precision
□	Timestamp	GetDate	Get date from timestamp
□	Timestamp	GetTime	Get time from timestamp

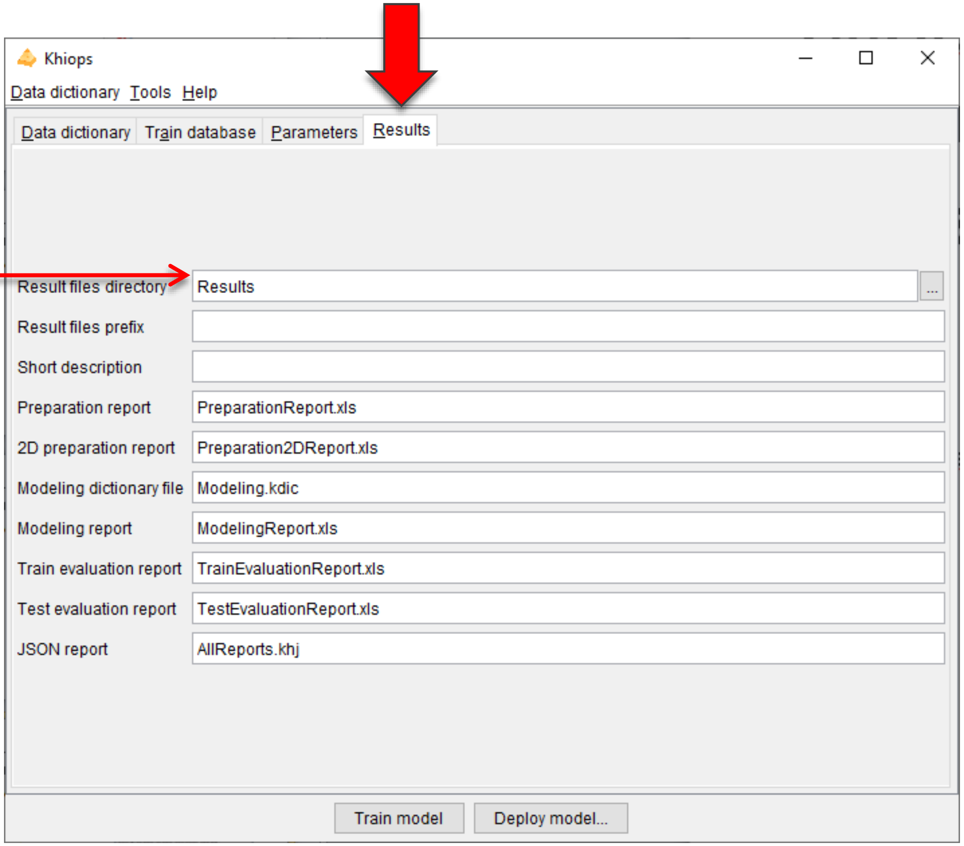
Optional

Choice of construction rules

Supervised classification

- **Step 5 :** Analysis results

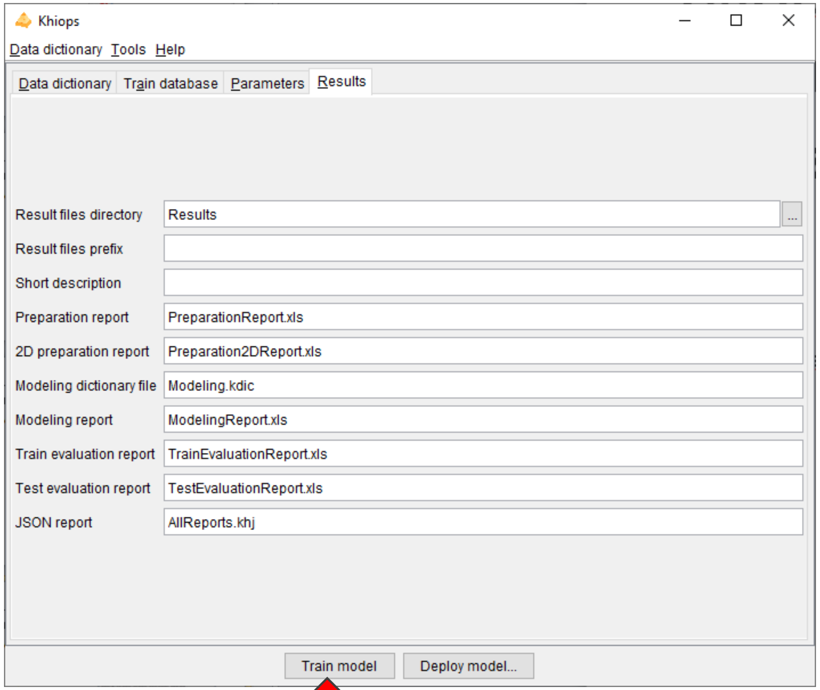
Results files directory


Khiops
Data dictionary Tools Help
Data dictionary Train database Parameters Results
Train model
Deploy model...

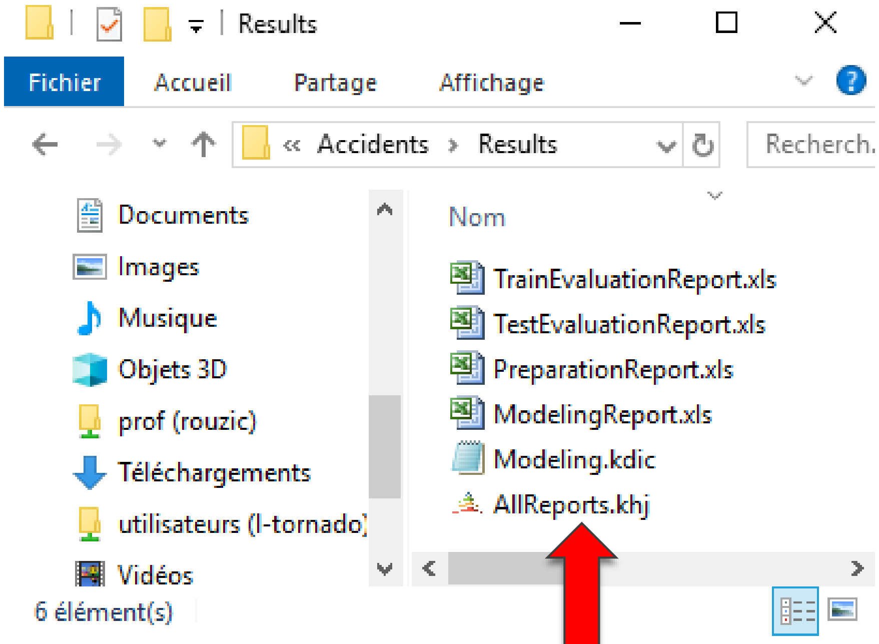
Result files directory	Results
Result files prefix	
Short description	
Preparation report	PreparationReport.xls
2D preparation report	Preparation2DReport.xls
Modeling dictionary file	Modeling.kdic
Modeling report	ModelingReport.xls
Train evaluation report	TrainEvaluationReport.xls
Test evaluation report	TestEvaluationReport.xls
JSON report	AllReports.khj

Supervised classification

- **Step 6 : Start the analysis**


Khiops
Data dictionary Tools Help
Data dictionary Train database Parameters Results
Train model
Deploy model...

Result files directory	Results
Result files prefix	
Short description	
Preparation report	PreparationReport.xls
2D preparation report	Preparation2DReport.xls
Modeling dictionary file	Modeling.kdic
Modeling report	ModelingReport.xls
Train evaluation report	TrainEvaluationReport.xls
Test evaluation report	TestEvaluationReport.xls
JSON report	AllReports.khj

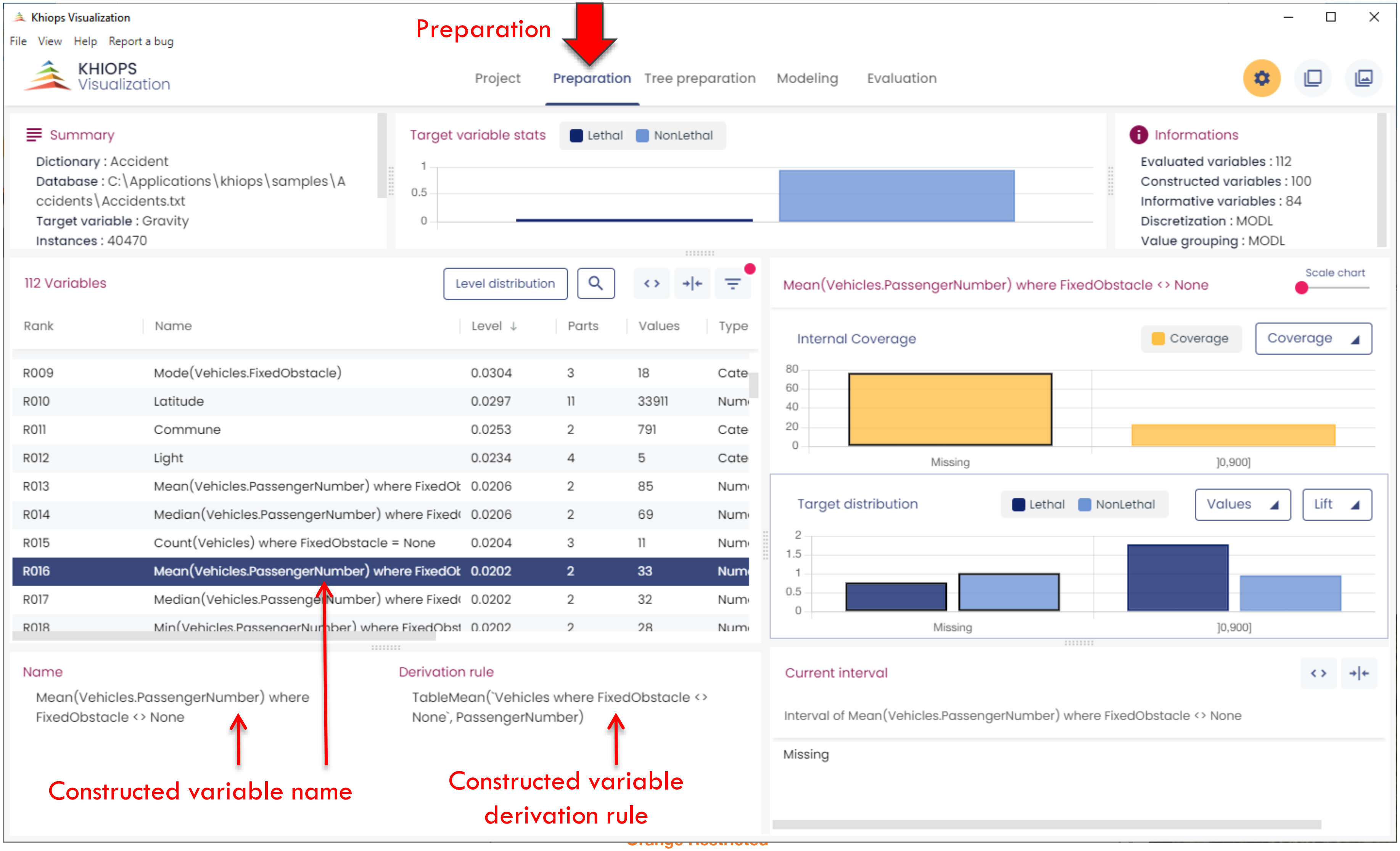


- TrainEvaluationReport.xls
- TestEvaluationReport.xls
- PreparationReport.xls
- ModelingReport.xls
- Modeling.kdic
- AllReports.khj

1 - Train model

2 - Inspect the results using Khiops Visualization (double-click on *.khj* file)

Exploratory of classification results using Khiops Visualization



Khiops Visualization

Preparation

Summary

Dictionary : Accident
Database : C:\Applications\khiops\samples\Accidents\Accidents.txt
Target variable : Gravity
Instances : 40470

Target variable stats

Informations

Evaluated variables : 112
Constructed variables : 100
Informative variables : 84
Discretization : MODL
Value grouping : MODL

112 Variables

Rank	Name	Level ↓	Parts	Values	Type
R009	Mode(Vehicles.FixedObstacle)	0.0304	3	18	Cate
R010	Latitude	0.0297	11	33911	Num
R011	Commune	0.0253	2	791	Cate
R012	Light	0.0234	4	5	Cate
R013	Mean(Vehicles.PassengerNumber) where FixedOt	0.0206	2	85	Num
R014	Median(Vehicles.PassengerNumber) where Fixed	0.0206	2	69	Num
R015	Count(Vehicles) where FixedObstacle = None	0.0204	3	11	Num
R016	Mean(Vehicles.PassengerNumber) where FixedOt	0.0202	2	33	Num
R017	Median(Vehicles.PassengerNumber) where Fixed	0.0202	2	32	Num
R018	Min(Vehicles.PassengerNumber) where FixedObst	0.0202	2	28	Num

Mean(Vehicles.PassengerNumber) where FixedObstacle <> None

Name

Mean(Vehicles.PassengerNumber) where FixedObstacle <> None

Derivation rule

TableMean(`Vehicles where FixedObstacle <> None`, PassengerNumber)

Current interval

Interval of Mean(Vehicles.PassengerNumber) where FixedObstacle <> None

Missing

Constructed variable name

Constructed variable derivation rule

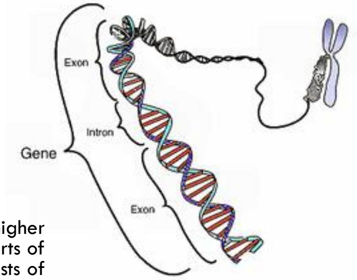
Khiops multi-table

- Khiops can deal with multi-table databases
  - star schema: one root entity and several 0-1 or 0-n secondary entities
  - snowflake schemas and beyond

- Impact on Khiops Coclustering
  - Deployment of coclustering models
    - Given a text*word coclustering model, assign new texts (with their words) to their closest cluster
    - Given a cookie*page coclustering model, assign new cookies (with their pages) to their closest cluster
    - Given a curve*X*Y triclustering model, assign new curves (with their X*Y points) to their closest cluster

- In this tutorial
  - Build a triclustering model on the SpliceJunctionDNA data table
    - Clusters of sequence samples
    - Intervals of positions in the sequences
    - Clusters of DNA chars
  - Prepare a deployment model
    - Build a deployment dictionary
  - Deploy the model on the multi-table SpliceJunction database
    - Assign new DNA sequences to trained clusters of sequences

Splice junction multi-table database



- Molecular Biology (Splice-junction Gene Sequences)
  - Objective:
    - Recognition of boundaries between exons and introns in DNA sequences
    - Splice junctions are points on a DNA sequence at which `superfluous' DNA is removed during the process of protein creation in higher organisms. The problem posed in this dataset is to recognize, given a sequence of DNA, the boundaries between exons (the parts of the DNA sequence retained after splicing) and introns (the parts of the DNA sequence that are spliced out). This problem consists of two subtasks: recognizing exon/intron boundaries (referred to as EI sites), and recognizing intron/exon boundaries (IE sites). (In the biological community, IE borders are referred to as ``acceptors'' while EI borders are referred to as ``donors''.)

- Database dictionary
  - Root entity: splice junction
    - SampleId
    - Class (EI, IE, NEG)
    - Sequence of DNA
  - Secondary entity: DNA
    - SampleId:
    - Pos: position in the sequence
    - Char (A, C, G, T)

- Database files
  - SpliceJunction.txt
  - SpliceJunctionDNA.txt

SpliceJunction.txt

SampleId	Class
AGMKPNRSB-NEG-1	N
AGMORS12A-NEG-181	N
AGMORS9A-NEG-481	N
AGMRSKPNI-NEG-1141	N
ATRINS-ACCEPTOR-1678	IE
ATRINS-ACCEPTOR-701	IE
ATRINS-DONOR-521	EI
ATRINS-DONOR-905	EI
…	

SpliceJunctionDNA.txt

SampleId	Pos	Char
AGMKPNRSB-NEG-1	1	C
AGMKPNRSB-NEG-1	2	A
…		
AGMKPNRSB-NEG-1	58	A
AGMKPNRSB-NEG-1	59	C
AGMKPNRSB-NEG-1	60	A
AGMORS12A-NEG-181	1	A
AGMORS12A-NEG-181	2	G
…		
AGMORS12A-NEG-181	59	G
AGMORS12A-NEG-181	60	G
AGMORS9A-NEG-481	1	T
AGMORS9A-NEG-481	2	G
AGMORS9A-NEG-481	3	G
…		

Exploratory analysis of DNA sequences:

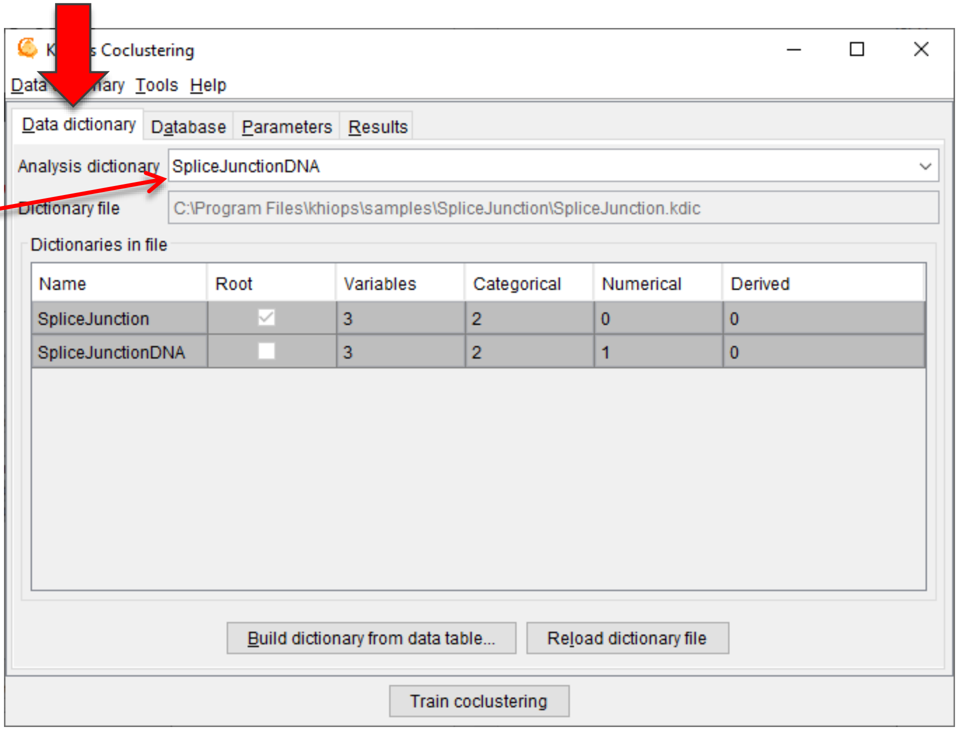
. find clusters of similar DNA sequences

. using a triclustering SampleId x Pos x Char

Train a triclustering model

- **Step 1 :** Open an existing dictionary

  (ex: sample SpliceJunction.kdic)



Analysis dictionary: SpliceJunctionDNA

Dictionary file: C:\Program Files\khiops\samples\SpliceJunction\SpliceJunction.kdic

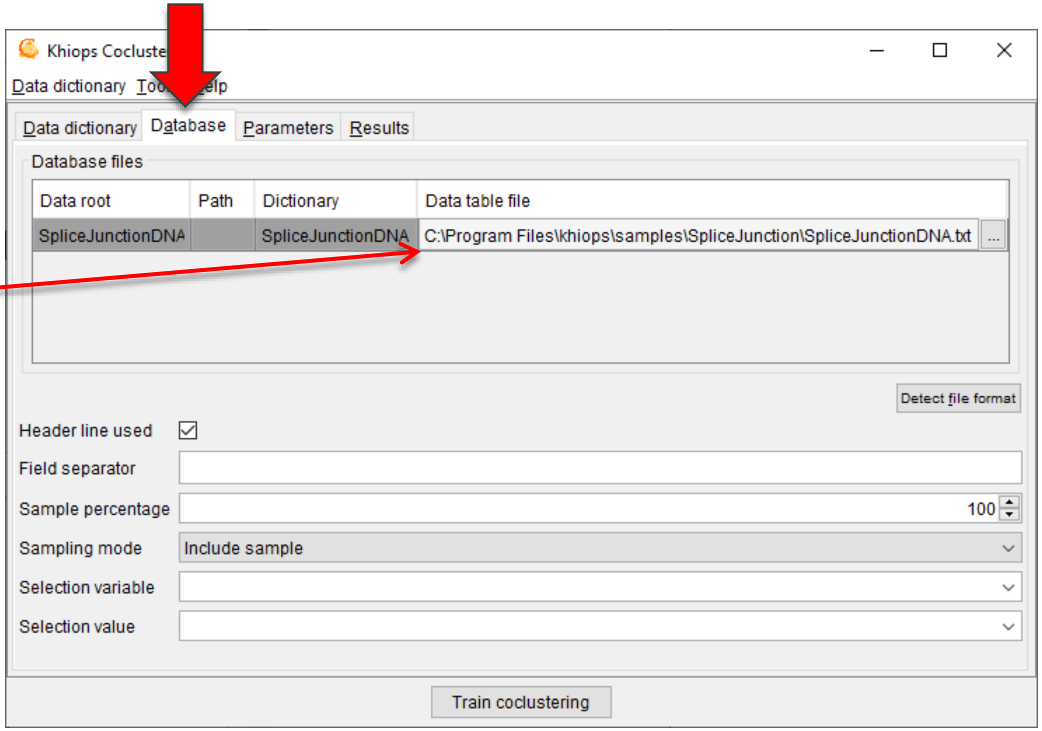
Dictionaries in file

Name	Root	Variables	Categorical	Numerical	Derived
SpliceJunction	☑	3	2	0	0
SpliceJunctionDNA	□	3	2	1	0

Analysis dictionary
(secondary entity)

Train a triclustering model

- **Step 2 :** Specification of used database



Khiops Cocluste

Database files

Data root	Path	Dictionary	Data table file
SpliceJunctionDNA		SpliceJunctionDNA	C:\Program Files\khiops\samples\SpliceJunction\SpliceJunctionDNA.txt

Header line used ☑

Field separator

Sample percentage 100

Sampling mode Include sample

Selection variable

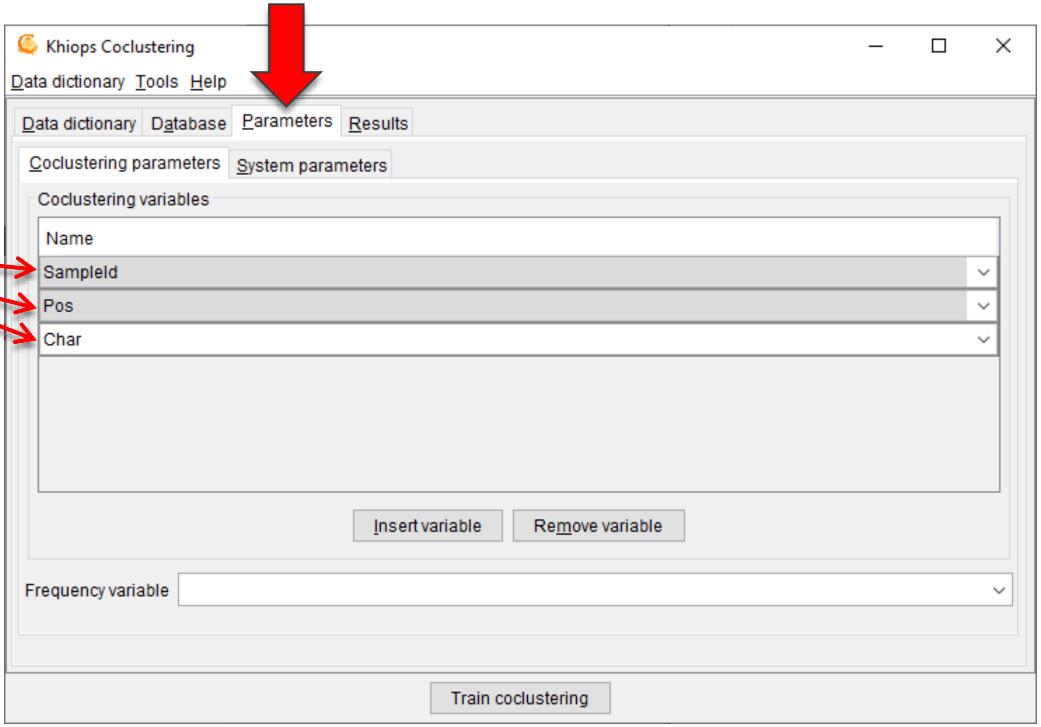
Selection value

Data table file

(one single file for analysis of the secondary entity)

Train a triclustering model

- **Step 3 : Specification of triclustering variables**



Coclustering variables

Name
SampleId
Pos
Char

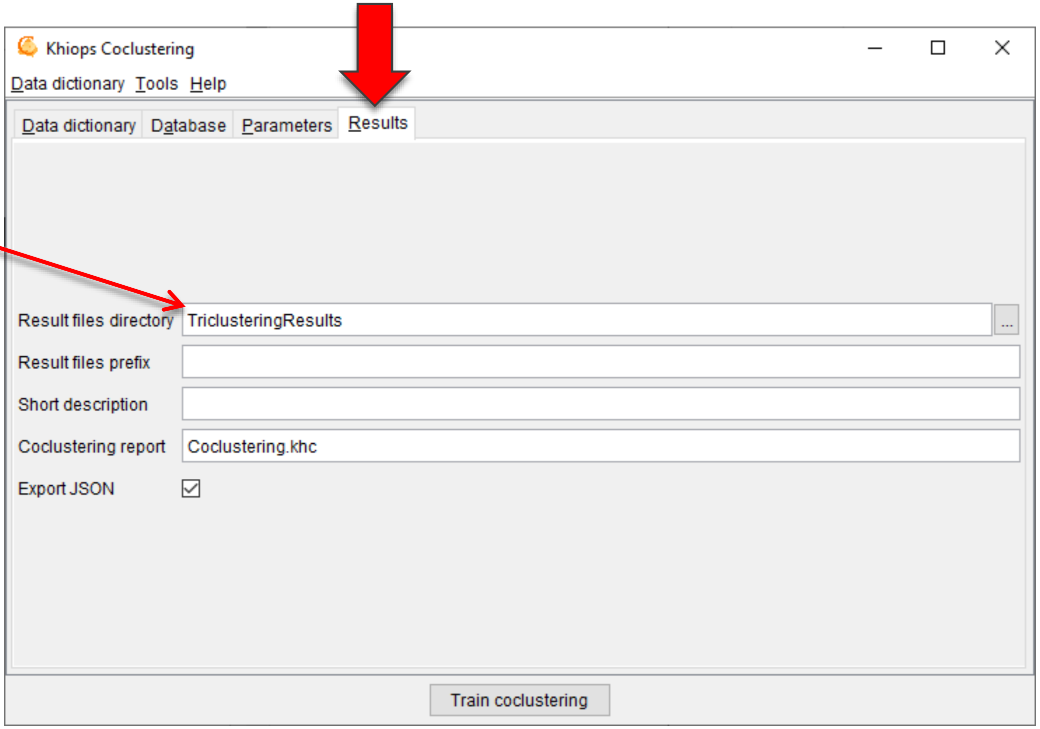
Frequency variable

Triclustering variables

Train a triclustering model

- **Step 4 :** Analysis results

Result files directory


Khiops Coclustering
Data dictionary Tools Help
Data dictionary Database Parameters Results
...
Train coclustering

Result files directory: TriclusteringResults

Result files prefix:

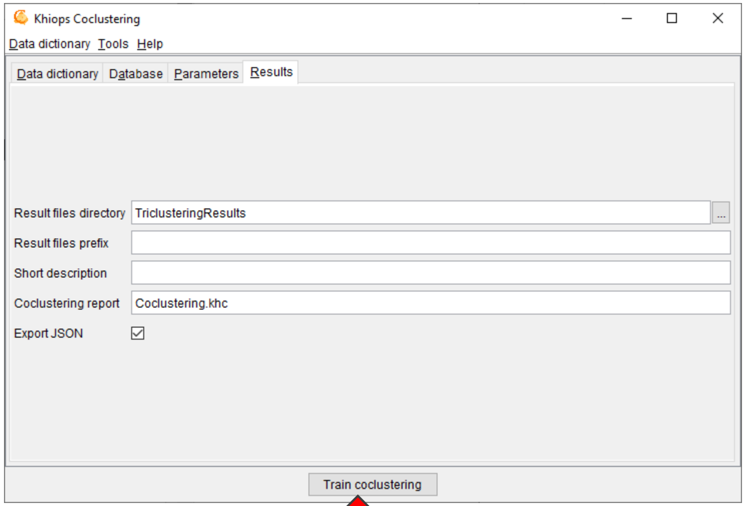
Short description:

Coclustering report: Coclustering.khc

Export JSON ☑

Train a triclustering model

- **Step 5 : Start the analysis**


Khiops Coclustering
Data dictionary Tools Help
Data dictionary Database Parameters Results
Train coclustering

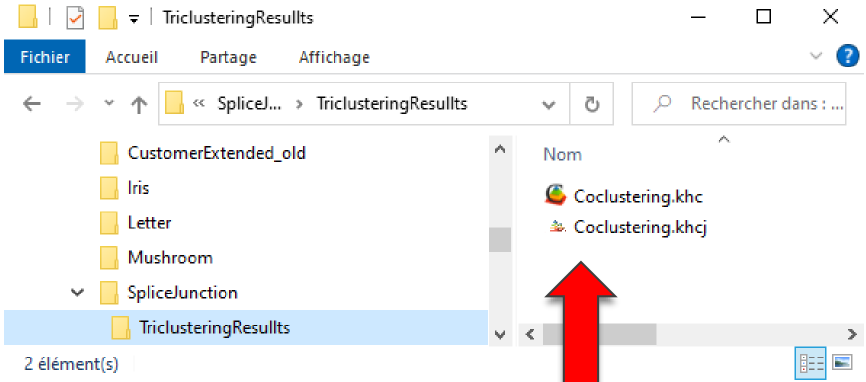
Result files directory: TriclusteringResults

Result files prefix:

Short description:

Coclustering report: Coclustering.khc

Export JSON: ☑

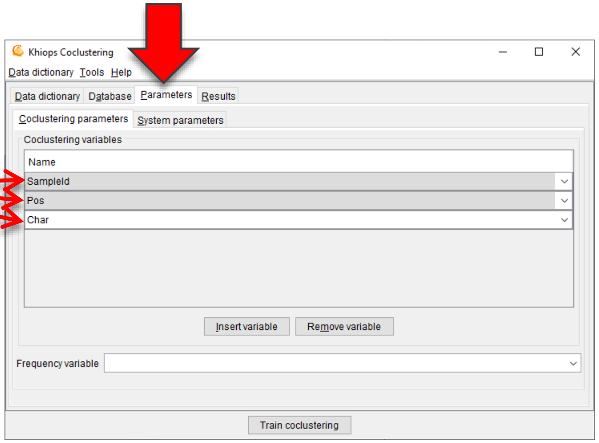


1 – Start the analysis

2 - Inspect the results using Khiops Covisualization
(double-click on *.khcj* file)

Khiops covisualisation: base SpliceJunctionDNA

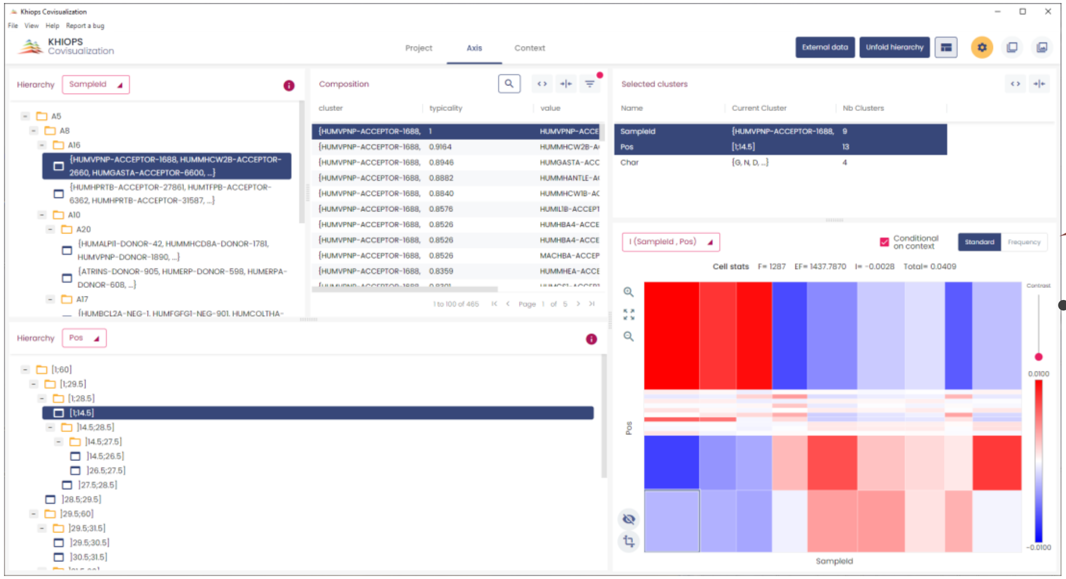
- With Khiops Coclustering
  - Analysis of correlation between variables SampleId*Pos*Char



Coclustering variables

Name
SampleId
Pos
Char

Frequency variable:



Hierarchy: SampleId

```
- A5
  - A8
    - A16
      {HUMVPNP-ACCEPTOR-1688, HUMMHCW2B-ACCEPTOR-2660, HUMGASTA-ACCEPTOR-6600, ...}
      {HUMHPRTB-ACCEPTOR-27861, HUMTFPB-ACCEPTOR-6362, HUMHPRTB-ACCEPTOR-31587, ...}
    - A10
      - A20
        {HUMALPI1-DONOR-42, HUMMHCD8A-DONOR-1781, HUMVPNP-DONOR-1890, ...}
        {ATRINS-DONOR-905, HUMERP-DONOR-598, HUMERPA-DONOR-608, ...}
      - A17
        {HUMBCL2A-NEG-1. HUMFGFG1-NEG-901. HUMCOLTHA-
```

Composition

cluster	typicality	value
{HUMVPNP-ACCEPTOR-1688,	1	HUMVPNP-ACCE
{HUMVPNP-ACCEPTOR-1688,	0.9164	HUMMHCW2B-A
{HUMVPNP-ACCEPTOR-1688,	0.8946	HUMGASTA-ACC
{HUMVPNP-ACCEPTOR-1688,	0.8882	HUMMHANTLE-A
{HUMVPNP-ACCEPTOR-1688,	0.8840	HUMMHCW1B-AC
{HUMVPNP-ACCEPTOR-1688,	0.8576	HUMIL1B-ACCEPT
{HUMVPNP-ACCEPTOR-1688,	0.8526	HUMHBA4-ACCE
{HUMVPNP-ACCEPTOR-1688,	0.8526	HUMHBA4-ACCE
{HUMVPNP-ACCEPTOR-1688,	0.8526	MACHBA-ACCEP
{HUMVPNP-ACCEPTOR-1688,	0.8359	HUMMHEA-ACCE

1 to 100 of 465 Page 1 of 5

Selected clusters

Name	Current Cluster	Nb Clusters
SampleId	{HUMVPNP-ACCEPTOR-1688,	9
Pos	[1;14.5]	13
Char	{G, N, D, ...}	4

I (SampleId , Pos) ☑ Conditional on context Standard Frequency

Cell stats F= 1287 EF= 1437.7870 I= -0.0028 Total= 0.0409

Hierarchy: Pos

```
- [1;60]
  - [1;29.5]
    - [1;28.5]
      [1;14.5]
      - ]14.5;28.5]
        - ]14.5;27.5]
          ]14.5;26.5]
          ]26.5;27.5]
        ]27.5;28.5]
    ]28.5;29.5]
  - ]29.5;60]
    - ]29.5;31.5]
      ]29.5;30.5]
      ]30.5;31.5]
```

- With Khiops Covisualization
  - Exploratory analysis of the results

Khiops multi-table

- Khiops can deal with multi-table databases
  - star schema: one root entity and several 0-1 or 0-n secondary entities
  - snowflake schemas and beyond

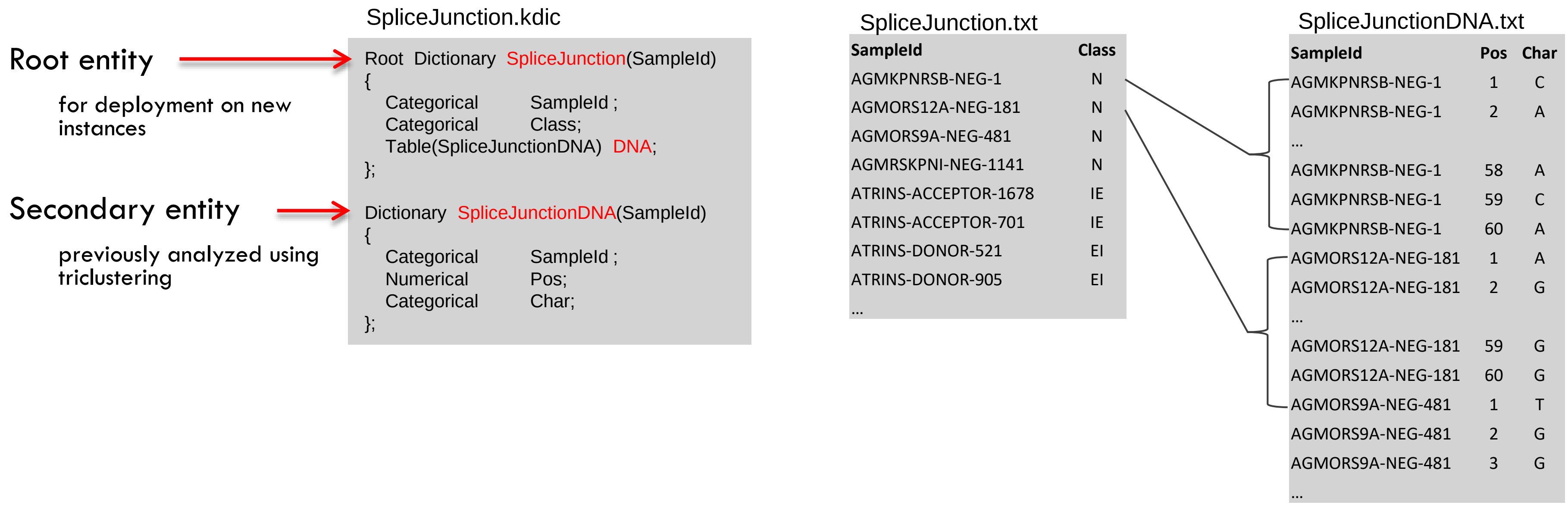
- Impact on Khiops Coclustering
  - Deployment of coclustering models
    - Given a text*word coclustering model, assign new texts (with their words) to their closest cluster
    - Given a cookie*page coclustering model, assign new cookies (with their pages) to their closest cluster
    - Given a curve*X*Y triclustering model, assign new curves (with their X*Y points) to their closest cluster

- In this tutorial
  - Train a triclustering model on the SpliceJunctionDNA data table
    - Clusters of sequence samples
    - Intervals of positions in the sequences
    - Clusters of DNA chars
  - Prepare a deployment model
    - Build a deployment dictionary
  - Deploy the model on the multi-table SpliceJunction database
    - Assign new DNA sequences to trained clusters of sequences

Prepare a deployment model

- **Prerequisite :** a multi-table database
  - dictionary file
  - data files

    (ex: sample SpliceJunction)



Root entity

for deployment on new instances

Secondary entity

previously analyzed using triclustering

SpliceJunction.kdic

```
Root  Dictionary  SpliceJunction(SampleId)
{
    Categorical          SampleId ;
    Categorical          Class;
    Table(SpliceJunctionDNA)  DNA;
};

Dictionary  SpliceJunctionDNA(SampleId)
{
    Categorical          SampleId ;
    Numerical            Pos;
    Categorical          Char;
};
```

SpliceJunction.txt

SampleId	Class
AGMKPNRSB-NEG-1	N
AGMORS12A-NEG-181	N
AGMORS9A-NEG-481	N
AGMRSKPNI-NEG-1141	N
ATRINS-ACCEPTOR-1678	IE
ATRINS-ACCEPTOR-701	IE
ATRINS-DONOR-521	EI
ATRINS-DONOR-905	EI
…	

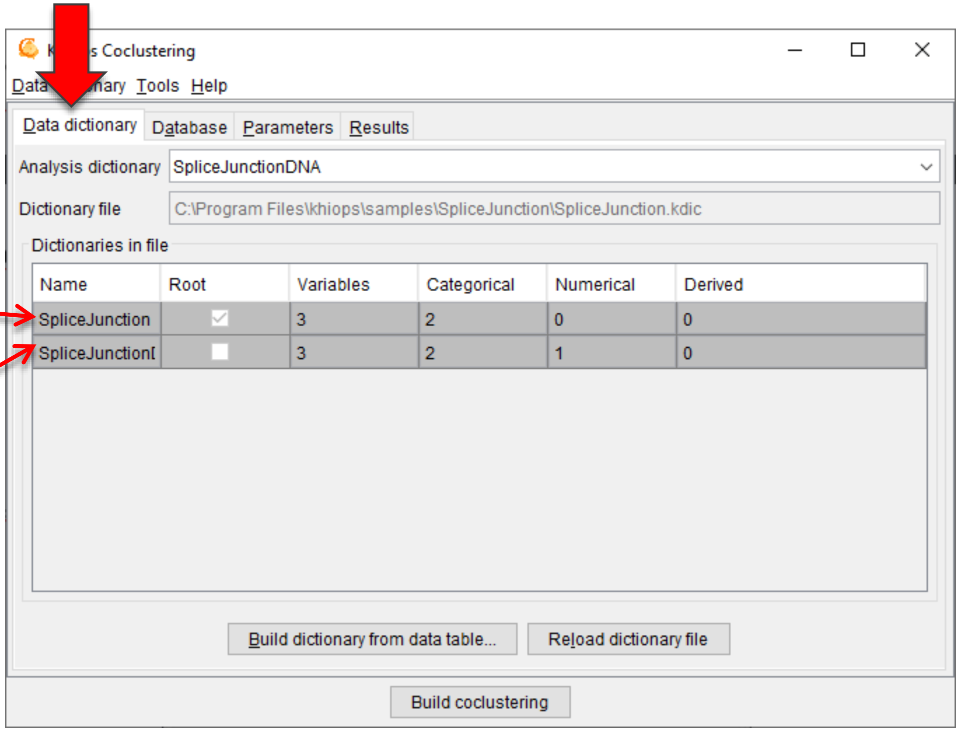
SpliceJunctionDNA.txt

SampleId	Pos	Char
AGMKPNRSB-NEG-1	1	C
AGMKPNRSB-NEG-1	2	A
…		
AGMKPNRSB-NEG-1	58	A
AGMKPNRSB-NEG-1	59	C
AGMKPNRSB-NEG-1	60	A
AGMORS12A-NEG-181	1	A
AGMORS12A-NEG-181	2	G
…		
AGMORS12A-NEG-181	59	G
AGMORS12A-NEG-181	60	G
AGMORS9A-NEG-481	1	T
AGMORS9A-NEG-481	2	G
AGMORS9A-NEG-481	3	G
…		

Prepare a deployment model

- **Step 1 :** Open an existing dictionary

  (ex: sample SpliceJunction.kdic)



Analysis dictionary: SpliceJunctionDNA

Dictionary file: C:\Program Files\khiops\samples\SpliceJunction\SpliceJunction.kdic

Dictionaries in file

Name	Root	Variables	Categorical	Numerical	Derived
SpliceJunction	☑	3	2	0	0
SpliceJunction[	□	3	2	1	0

Root entity

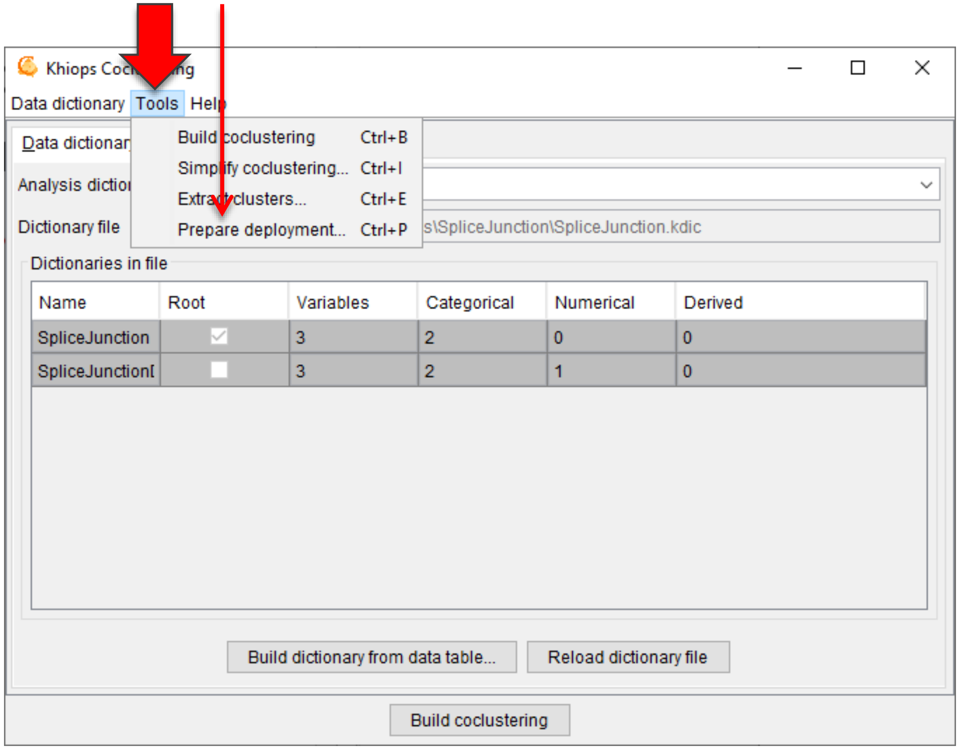
for deployment on new instances

Secondary entity

previously analyzed using triclustering

Prepare a deployment model

- **Step 2 :** *Start « Tools – Prepare deployment »*


Khiops Coc ng
Data dictionary
Tools
Help
Build coclustering Ctrl+B
Simplify coclustering... Ctrl+I
Extract clusters... Ctrl+E
Prepare deployment... Ctrl+P
Data dictionar
Build dictionary from data table...
Reload dictionary file
Build coclustering

Analysis dictio

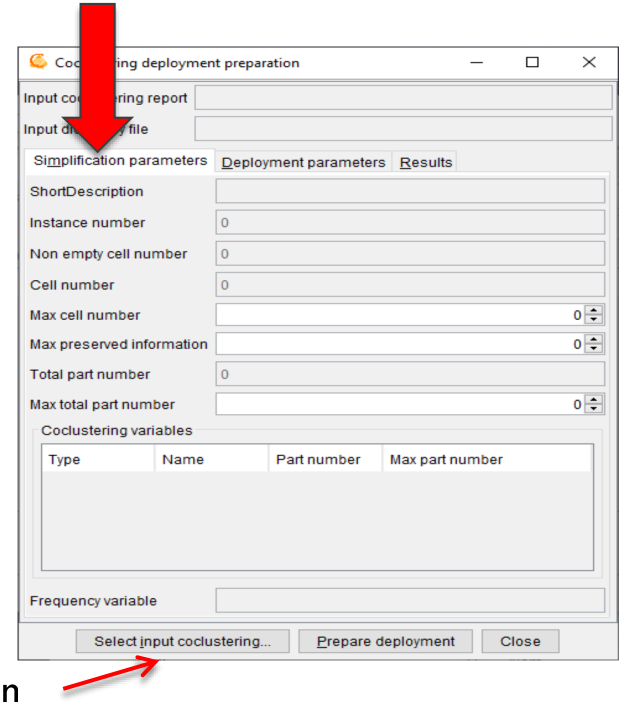
Dictionary file s\SpliceJunction\SpliceJunction.kdic

Dictionaries in file

Name	Root	Variables	Categorical	Numerical	Derived
SpliceJunction	☑	3	2	0	0
SpliceJunctionI	☐	3	2	1	0

Prepare a deployment model

- **Step 3 :** Select input coclustering file
  - (ex: previously trained triclustering model)



Coc ing deployment preparation

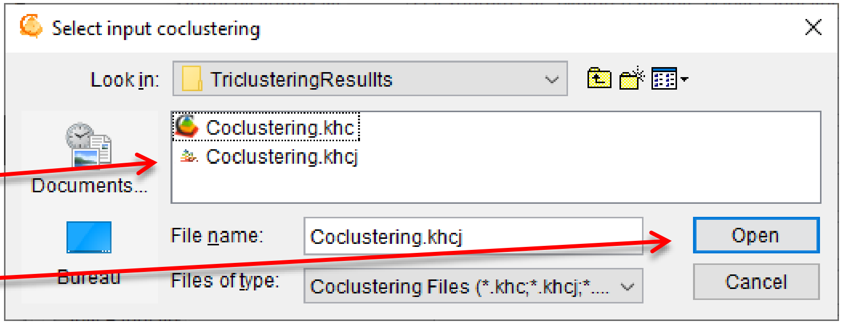
Input co ring report	
Input d y file	

ShortDescription	
Instance number	0
Non empty cell number	0
Cell number	0
Max cell number	0
Max preserved information	0
Total part number	0
Max total part number	0

Coclustering variables

Type	Name	Part number	Max part number

Frequency variable	



Select input coclustering

Look in: TriclusteringResullts

Coclustering.khc
Coclustering.khcj

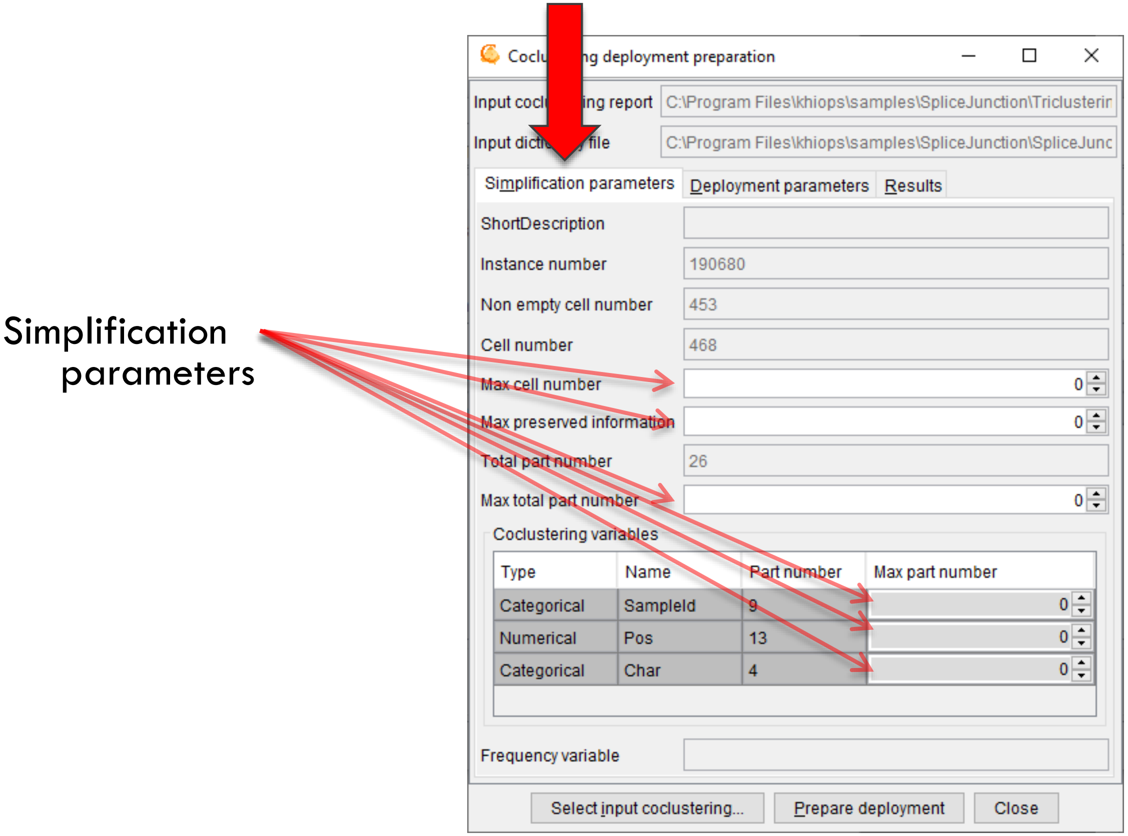
File name: Coclustering.khcj

Files of type: Coclustering Files (*.khc;*.khcj;*....

1. click on button
2. select a triclustering model file
3. open

Prepare a deployment model

- **Step 4 :** The triclustering model is summarized in the first pane
  - if necessary, specify simplification parameters



Simplification parameters
Coclu ng deployment preparation

Input cocl ing report: C:\Program Files\khiops\samples\SpliceJunction\Triclusterir

Input dicti y file: C:\Program Files\khiops\samples\SpliceJunction\SpliceJunc

Simplification parameters

ShortDescription	
Instance number	190680
Non empty cell number	453
Cell number	468
Max cell number	0
Max preserved information	0
Total part number	26
Max total part number	0

Coclustering variables

Type	Name	Part number	Max part number
Categorical	SampleId	9	0
Numerical	Pos	13	0
Categorical	Char	4	0

Frequency variable:

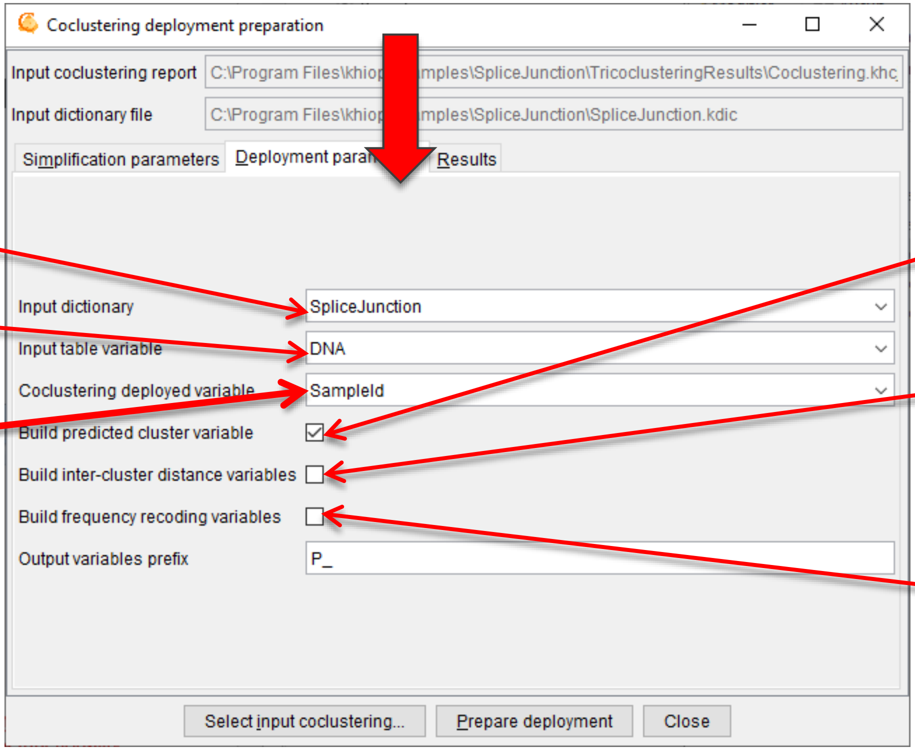
Prepare a deployment model

- **Step 5 : Specify deployment parameters**

Specification of input dictionary to enrich

```
Root  Dictionary  SpliceJunction(SampleId)
{
    Categorical         SampleId ;
    Categorical         Class;
    Table(SpliceJunctionDNA)  DNA;
};

Dictionary  SpliceJunctionDNA(SampleId)
{
    Categorical         SampleId ;
    Numerical           Pos;
    Categorical         Char;
};
```



Coclustering deployment preparation

Input coclustering report: C:\Program Files\khiop mples\SpliceJunction\TricoclusteringResults\Coclustering.khc

Input dictionary file: C:\Program Files\khiop mples\SpliceJunction\SpliceJunction.kdic

Input dictionary: SpliceJunction

Input table variable: DNA

Coclustering deployed variable: SampleId

Build predicted cluster variable ☑

Build inter-cluster distance variables □

Build frequency recoding variables □

Output variables prefix: P_

Specification of deployment variables to build

One variable to assign closest cluster

. closest cluster of *SampleId*

Several variables for distance to each cluster

. distance to each cluster of *SampleId*

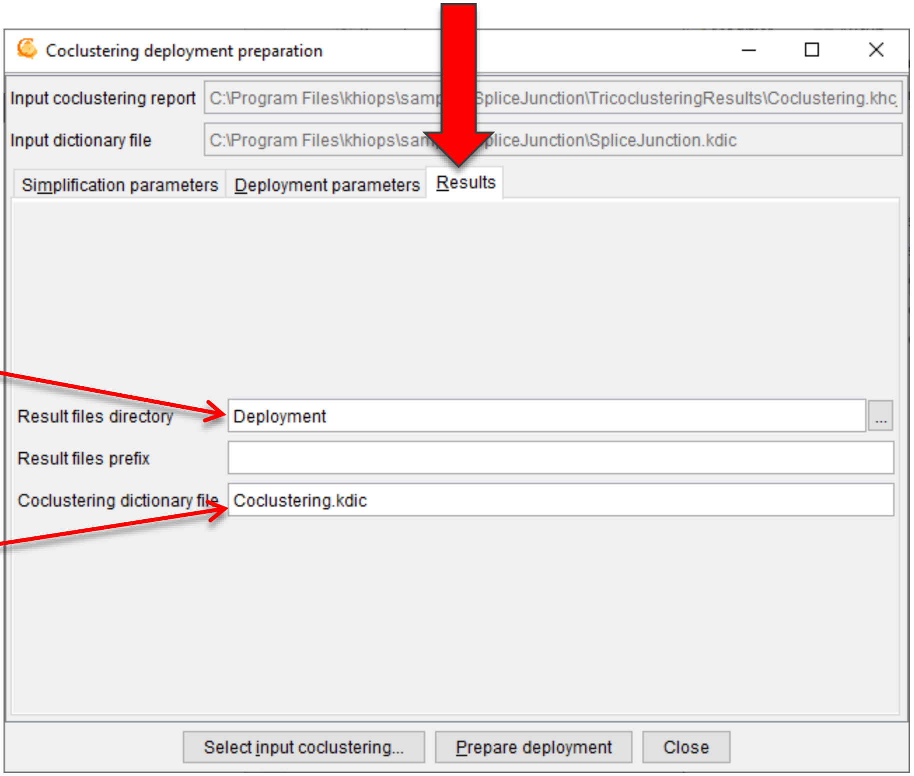
Several variables for secondary record number per interval/group of the other dimensions of the triclustering

. frequency per interval of *Pos*

. frequency per group of *Char*

Prepare a deployment model

- **Step 6 :** Specify result parameters



Coclustering deployment preparation

Input coclustering report: C:\Program Files\khiops\samp...SpliceJunction\TricoclusteringResults\Coclustering.khc

Input dictionary file: C:\Program Files\khiops\sam...pliceJunction\SpliceJunction.kdic

Result files directory: Deployment

Result files prefix:

Coclustering dictionary file: Coclustering.kdic

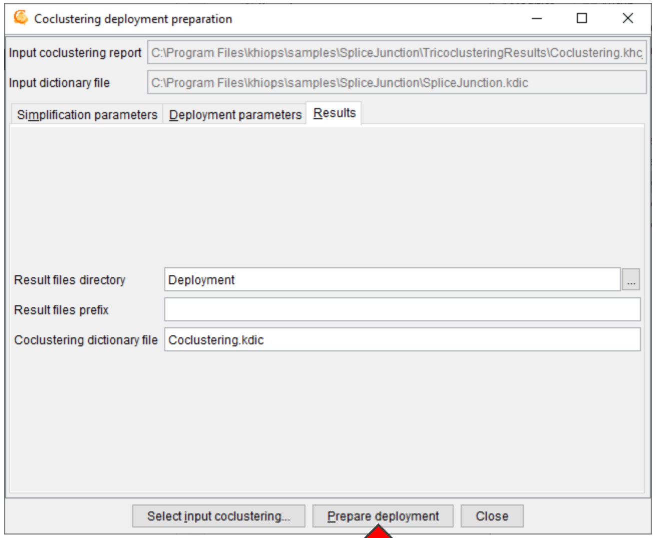
Result files directory

Deployment dictionary file

(to deploy cluster information on new data)

Prepare a deployment model

- **Step 7 :** Build the deployment dictionary



Coclustering deployment preparation

Input coclustering report: C:\Program Files\khiops\samples\SpliceJunction\TricoclusteringResults\Coclustering.khc

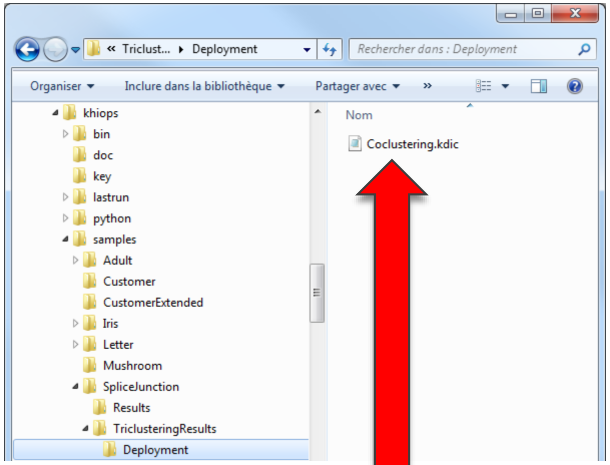
Input dictionary file: C:\Program Files\khiops\samples\SpliceJunction\SpliceJunction.kdic

Result files directory: Deployment

Result files prefix:

Coclustering dictionary file: Coclustering.kdic

1 – Build the deployment dictionary



2 – The deployment dictionary is ready for use with Khiops « *Transfer database* » functionality

Khiops multi-table

- Khiops can deal with multi-table databases
  - star schema: one root entity and several 0-1 or 0-n secondary entities
  - snowflake schemas and beyond

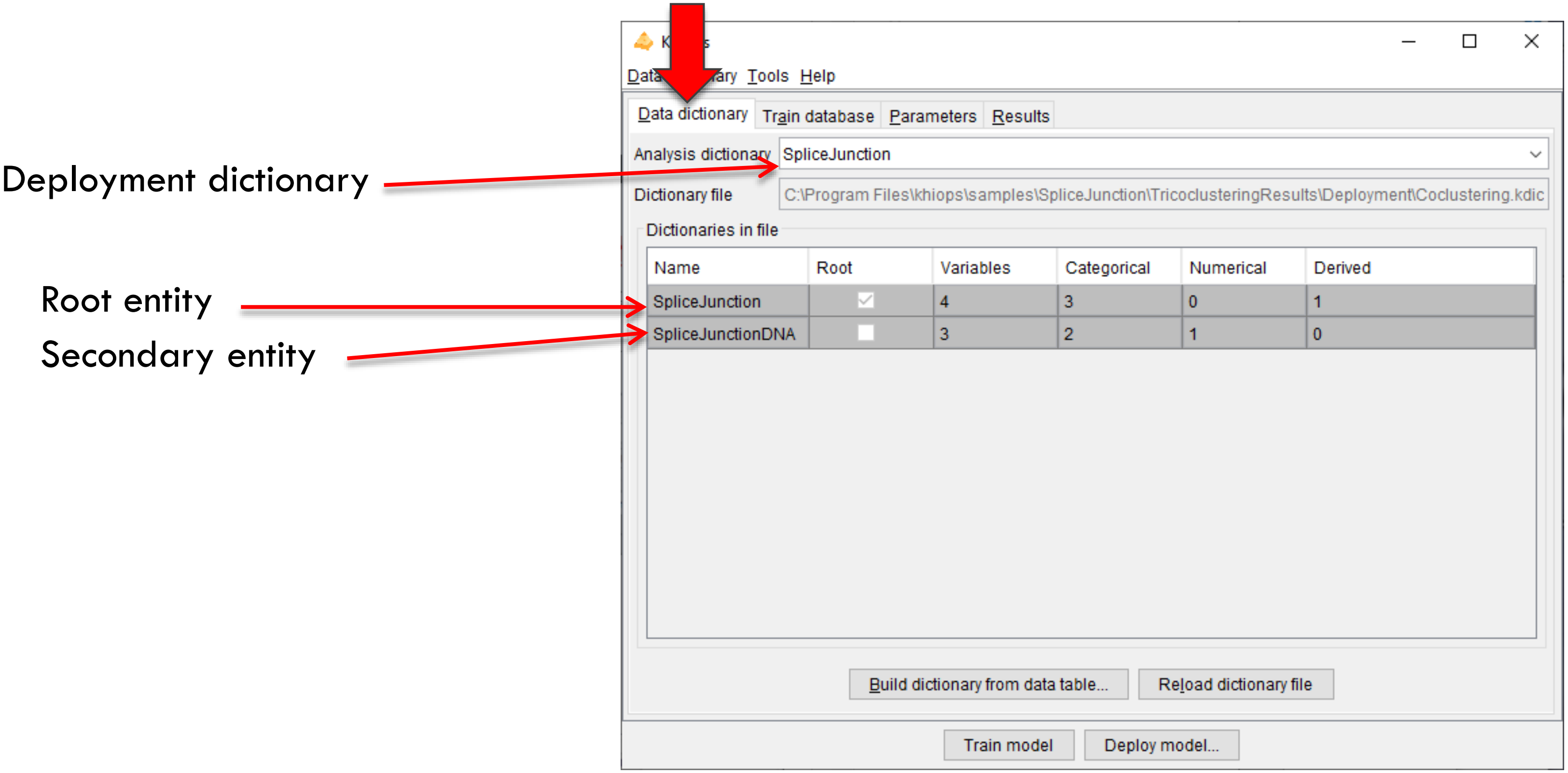
- Impact on Khiops Coclustering
  - Deployment of coclustering models
    - Given a text*word coclustering model, assign new texts (with their words) to their closest cluster
    - Given a cookie*page coclustering model, assign new cookies (with their pages) to their closest cluster
    - Given a curve*X*Y triclustering model, assign new curves (with their X*Y points) to their closest cluster

- In this tutorial
  - Train a triclustering model on the SpliceJunctionDNA data table
    - Clusters of sequence samples
    - Intervals of positions in the sequences
    - Clusters of DNA chars
  - Prepare a deployment model
    - Build a deployment dictionary
  - Deploy the model on the multi-table SpliceJunction database
    - Assign new DNA sequences to trained clusters of sequences

Deploy the model

- **Step 1 :** Open the deployment dictionary file with Khiops

(ex: Samples\SpliceJunction\TriclusteringResults\Deployment\Coclustering.kdic)



Deployment dictionary

Root entity

Secondary entity

Analysis dictionary: SpliceJunction

Dictionary file: C:\Program Files\khiops\samples\SpliceJunction\TricoclusteringResults\Deployment\Coclustering.kdic

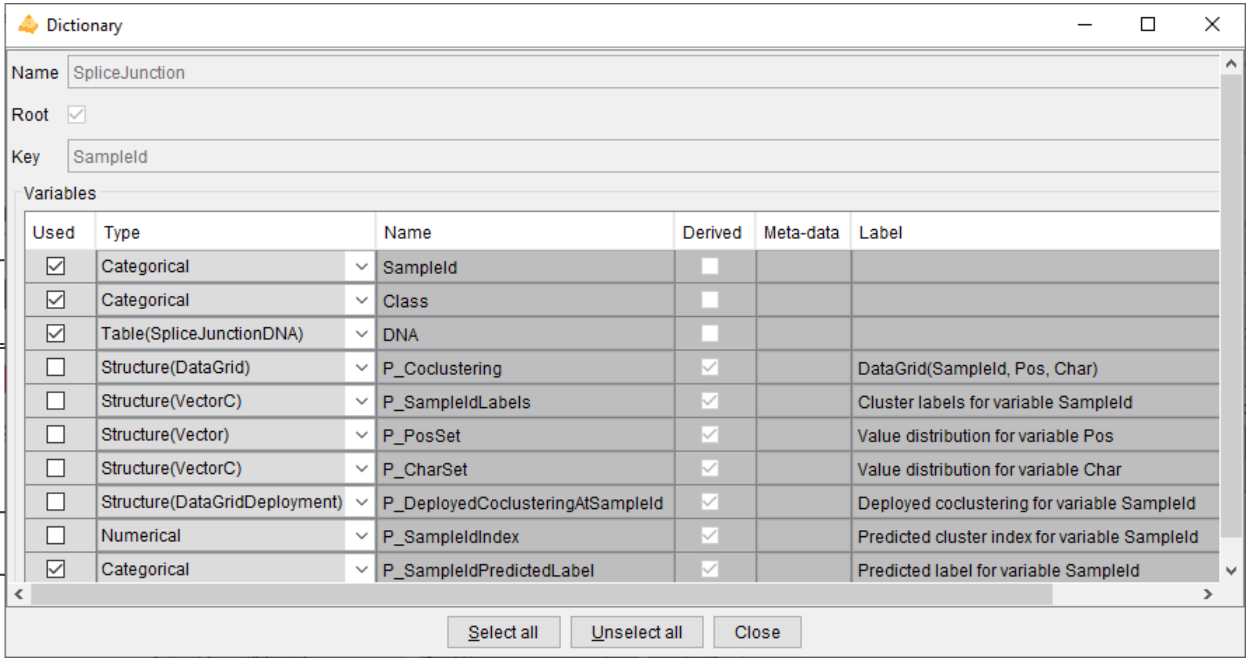
Dictionaries in file

Name	Root	Variables	Categorical	Numerical	Derived
SpliceJunction	☑	4	3	0	1
SpliceJunctionDNA	□	3	2	1	0

Deploy the model

- **Step 2 :** If necessary, select deployment variables

(use « *Inspect current dictionary* » by right-click on dictionary SpliceJunction)



Dictionary

Name: SpliceJunction

Root: ☑

Key: SampleId

Variables

Used	Type	Name	Derived	Meta-data	Label
☑	Categorical	SampleId	☐		
☑	Categorical	Class	☐		
☑	Table(SpliceJunctionDNA)	DNA	☐		
☐	Structure(DataGrid)	P_Coclustering	☑		DataGrid(SampleId, Pos, Char)
☐	Structure(VectorC)	P_SampleIdLabels	☑		Cluster labels for variable SampleId
☐	Structure(Vector)	P_PosSet	☑		Value distribution for variable Pos
☐	Structure(VectorC)	P_CharSet	☑		Value distribution for variable Char
☐	Structure(DataGridDeployment)	P_DeployedCoclusteringAtSampleId	☑		Deployed coclustering for variable SampleId
☐	Numerical	P_SampleIdIndex	☑		Predicted cluster index for variable SampleId
☑	Categorical	P_SampleIdPredictedLabel	☑		Predicted label for variable SampleId

Initial variables

Used by default

Model variables

(technical variables)

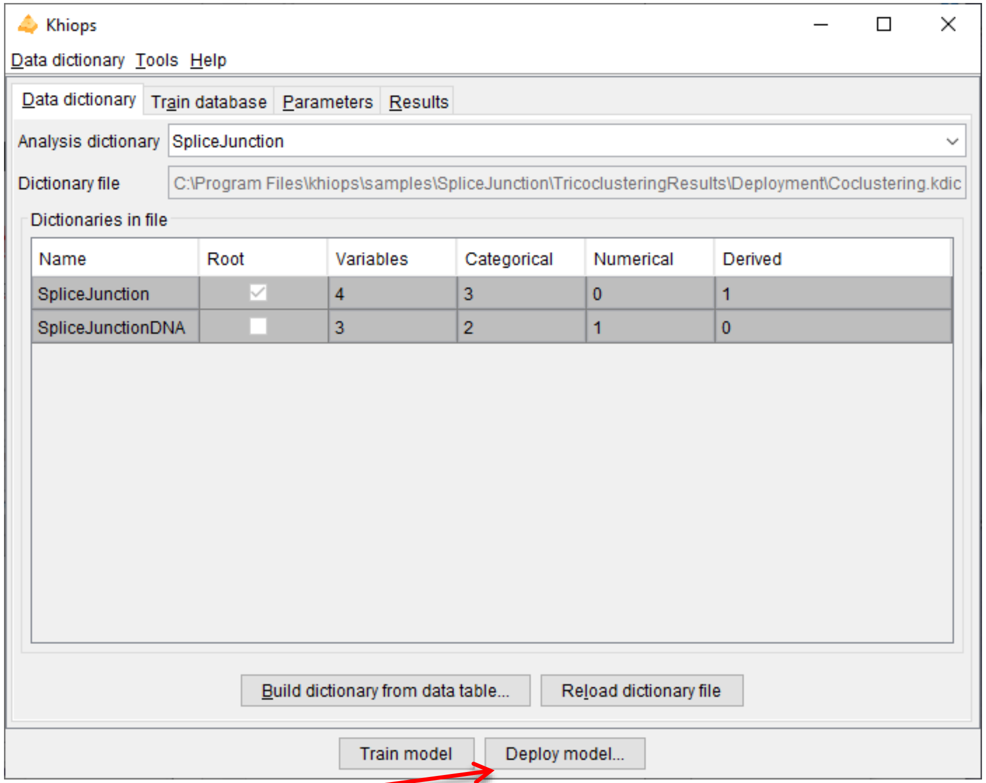
Deployment variables

Cluster index (unused by default)

Cluster label (used by default)

Deploy the model

- **Step 3 :** Open the « *Deploy model* » dialog box



Analysis dictionary: SpliceJunction

Dictionary file: C:\Program Files\khiops\samples\SpliceJunction\TricoclusteringResults\Deployment\Coclustering.kdic

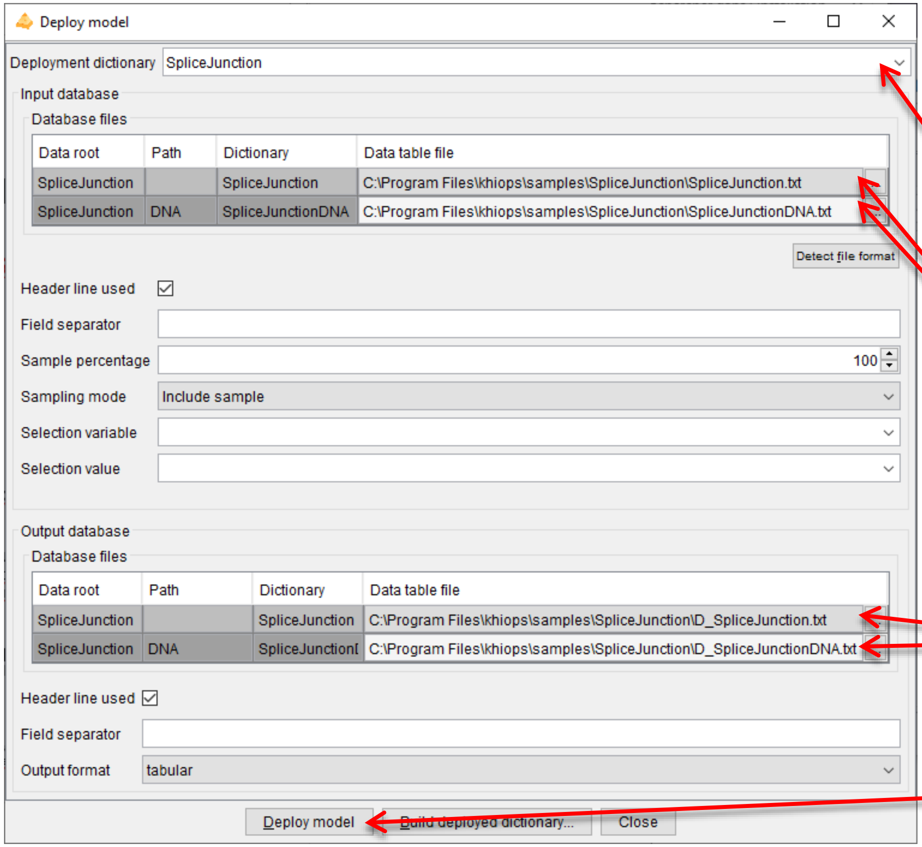
Dictionaries in file

Name	Root	Variables	Categorical	Numerical	Derived
SpliceJunction	☑	4	3	0	1
SpliceJunctionDNA	□	3	2	1	0

click on button

Deploy the model

- **Step 4 : Specify the file transfer parameters**



Deployment dictionary: SpliceJunction

Input database

Database files

Data root	Path	Dictionary	Data table file
SpliceJunction		SpliceJunction	C:\Program Files\khiops\samples\SpliceJunction\SpliceJunction.txt
SpliceJunction	DNA	SpliceJunctionDNA	C:\Program Files\khiops\samples\SpliceJunction\SpliceJunctionDNA.txt

Header line used ☑

Field separator:

Sample percentage: 100

Sampling mode: Include sample

Selection variable:

Selection value:

Output database

Database files

Data root	Path	Dictionary	Data table file
SpliceJunction		SpliceJunction	C:\Program Files\khiops\samples\SpliceJunction\D_SpliceJunction.txt
SpliceJunction	DNA	SpliceJunction[	C:\Program Files\khiops\samples\SpliceJunction\D_SpliceJunctionDNA.txt

Header line used ☑

Field separator:

Output format: tabular

1 Specify the deployment dictionary

2 Specify the input data table files

- splice junction samples with their DNA sequence
- all files are mandatory

3 Specify the output data table files

- secondary files are optional

4 Deploy

- The output files are enriched with new fields derived from the triclustering analysis

End of tutorial: summary

- ## Khiops
  - Optimal data preparation based on discretization and value grouping
  - Scoring models for classification and regression
  - Correlation analysis between pairs of variables
- ## Khiops Visualization
  - Analysis of Khiops results using an interactive visualization tool
- ## Khiops Coclustering
  - Correlation analysis of two or more variables using a hierarchical coclustering model
- ## Khiops Covisualization
  - Exploratory analysis of Khiops Coclustering results using an interactive visualization tool
- ## Multi-table functionalities
  - Multi-table database
  - Automatic feature construction
  - Multi-table functionalities in Khiops and Khiops Coclustering