

Orange

February 2024

Tutorial

KHIOPS 10.2

KHIOPS & KHIOPS VISUALIZATION

KHIOPS COCLUSTERING & KHIOPS COVISUALIZATION

MULTI-TABLE FUNCTIONALITIES

Khiops

2



- **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

60

Khiops & Khiops Visualization

3



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



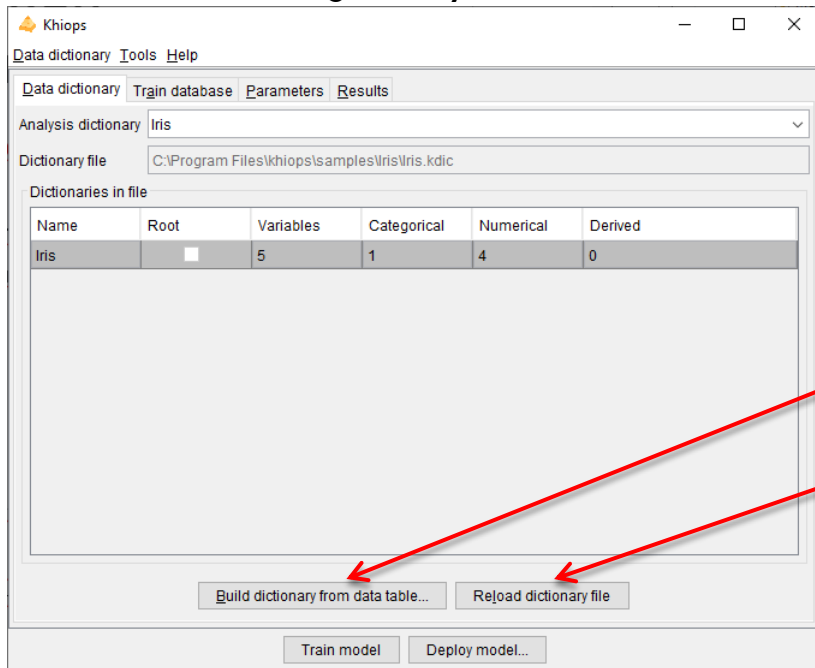
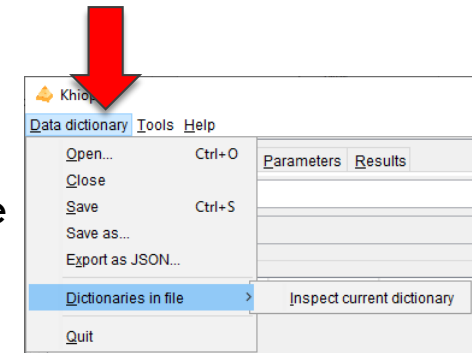
- **Khiops Visualization**
 - Analysis of Khiops results using an interactive visualization tool

Supervised classification

4

- **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



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

5

- **Step 1, bis : Build a new dictionary from a data table**
(If no available dictionary)

• 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

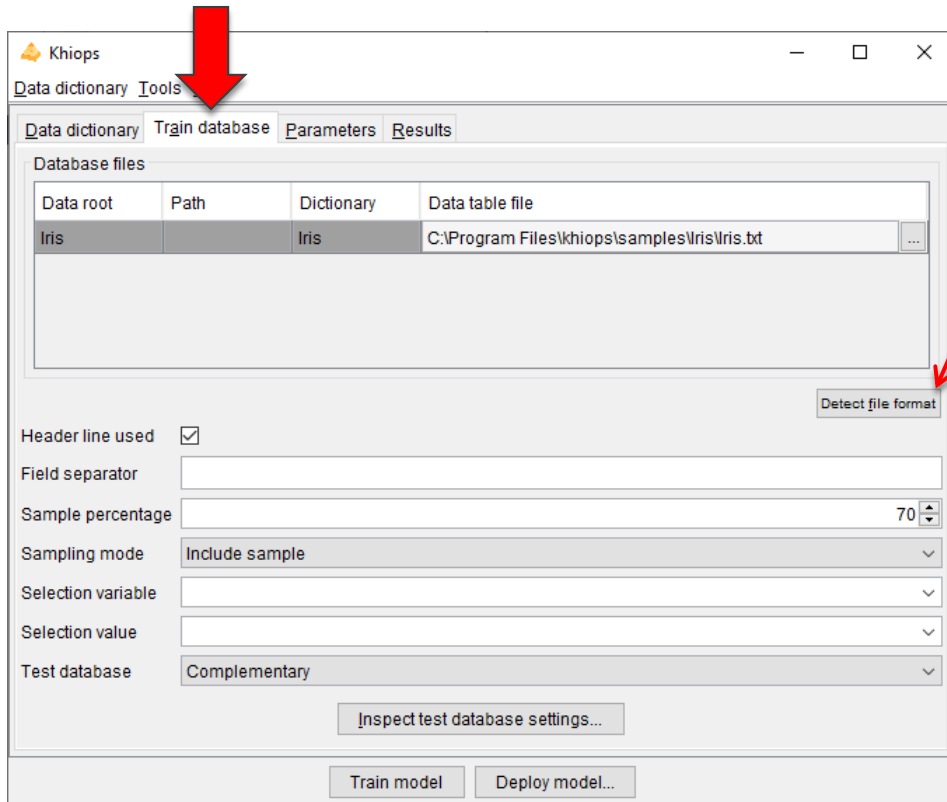
• 4. Close to save dictionary file

• 5. In pane **Data dictionary**
• Inspect the built dictionary
• Check variables types
• Select used variables

Supervised classification

6

- **Step 2 : Specify train 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

Sampling mode

Selection variable

Selection value

Test database

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

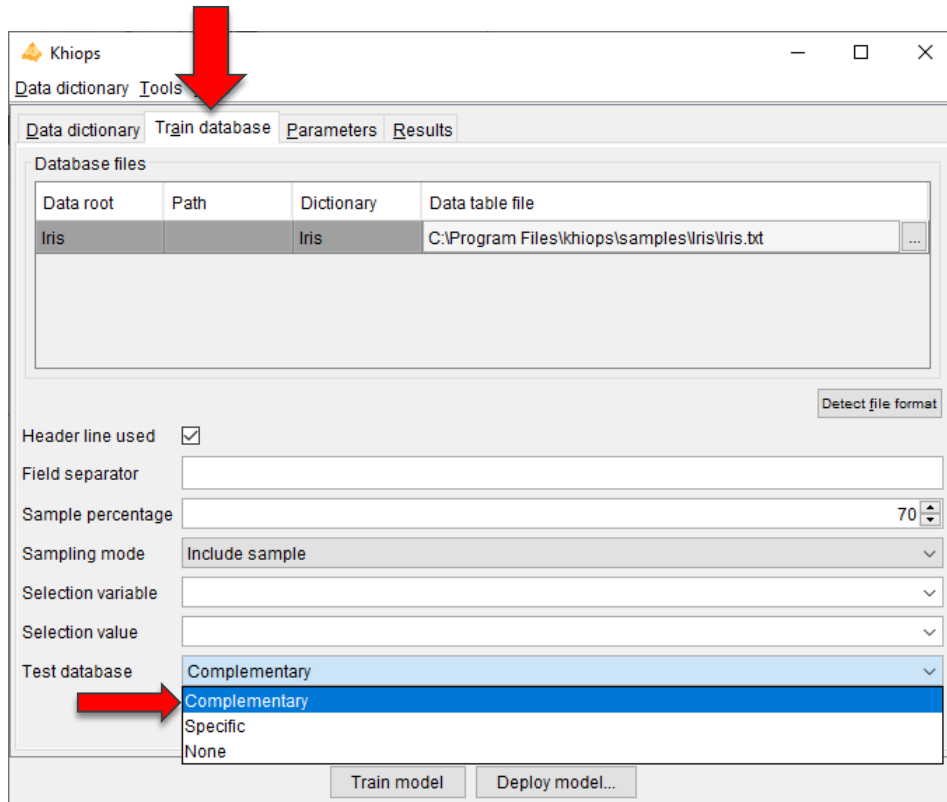
Sample percentage : default 70%

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

Supervised classification

7

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



Three possibilities :

Complementary

Specific

None

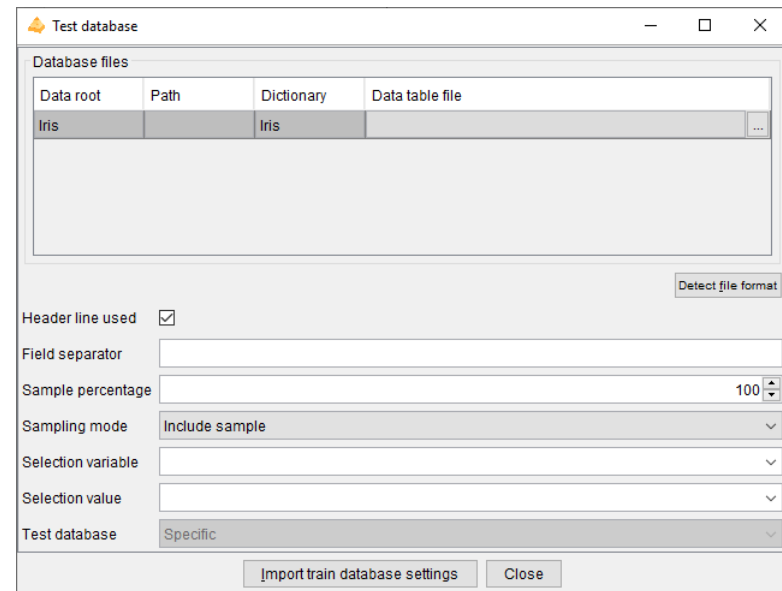
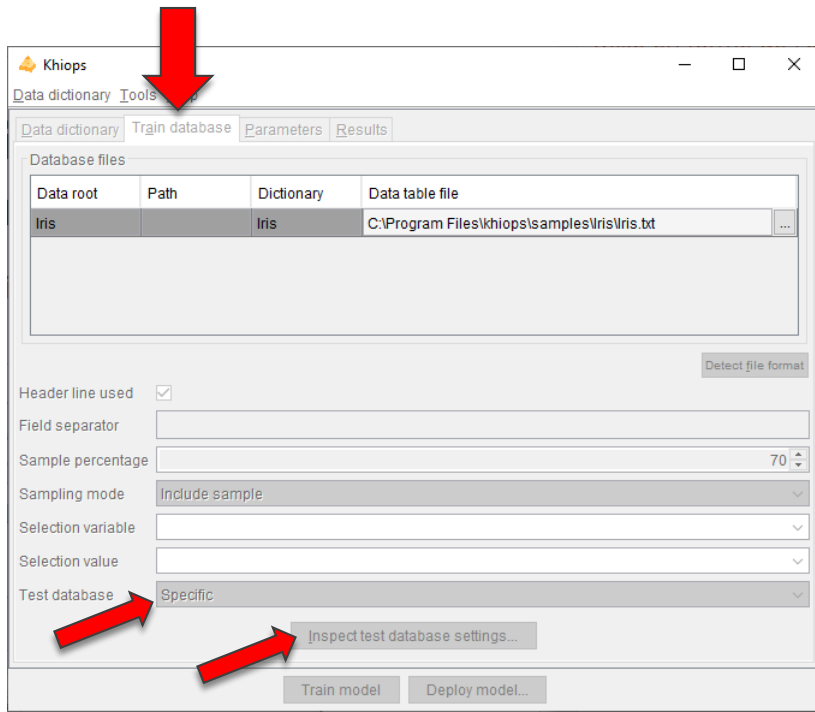
Complementary (default)

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

Supervised classification

8

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

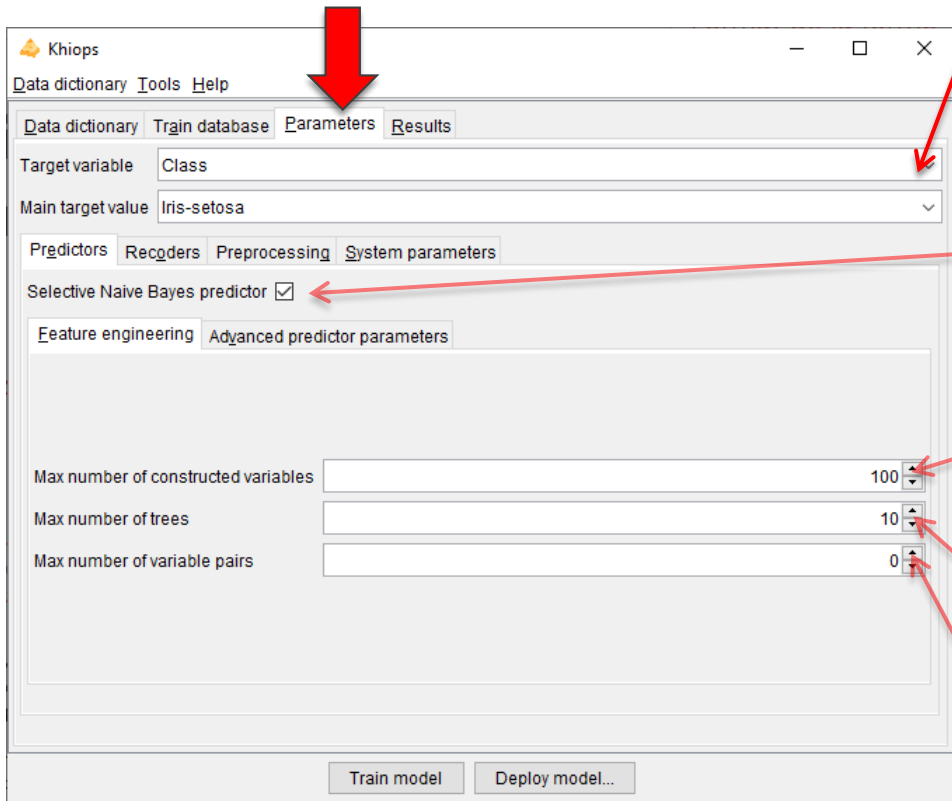


Specific

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

Supervised classification

- **Step 3 : Parameters**



Type of selected target variable implies type of analysis

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

Selective Naive 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)

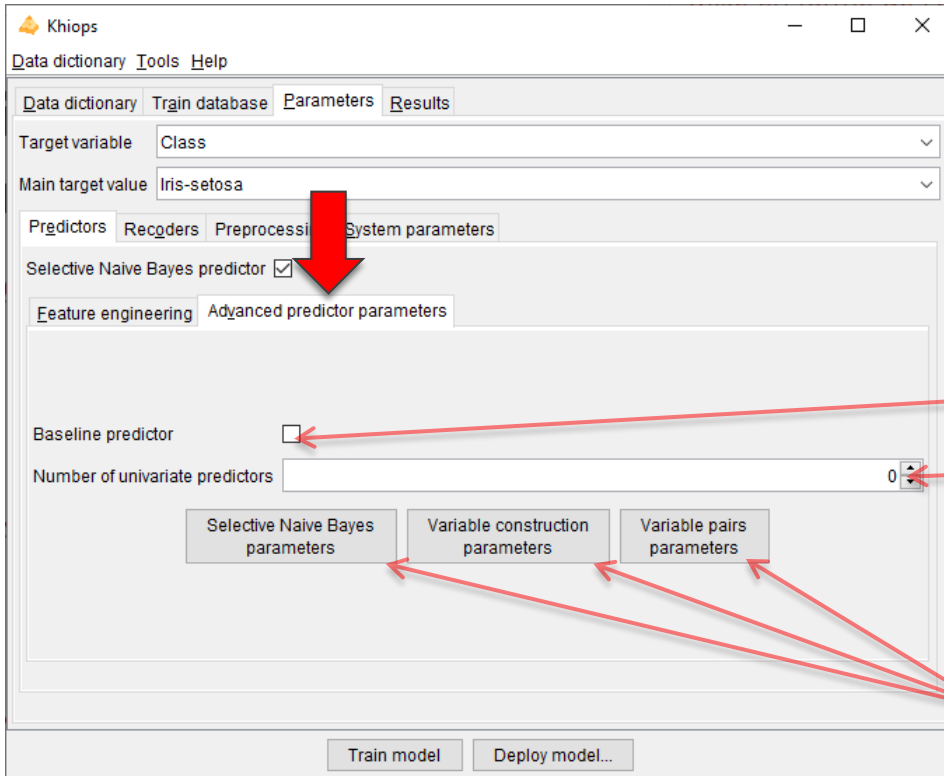
The constructed trees allow to combine variables, either native or constructed (default 10)

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

Supervised classification

10

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



Khiops

Data dictionary Tools Help

Data dictionary Train database Parameters Results

Target variable Class

Main target value Iris-setosa

Predictors Recorders 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...

Two other optional predictors
(only in the supervised case)

- **Baseline** : prediction of the majority class
(default false)

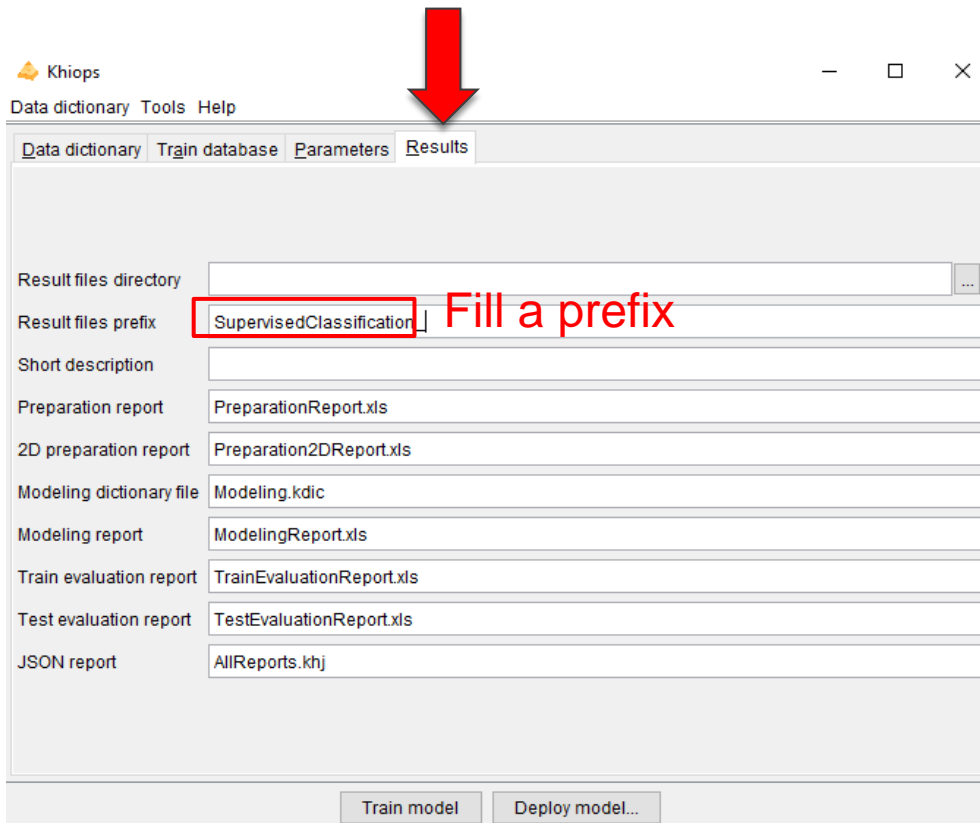
- **Univariate**: predictors exploiting one single variable
(default none)

Advanced parameters to inspect

Supervised classification

11

- **Step 4 : Results**



Wheniops

Data dictionary Tools Help

Data dictionary Train database Parameters Results

Result files directory

Result files prefix **Fill a prefix**

Short description

Preparation report

2D preparation report

Modeling dictionary file

Modeling report

Train evaluation report

Test evaluation report

JSON report

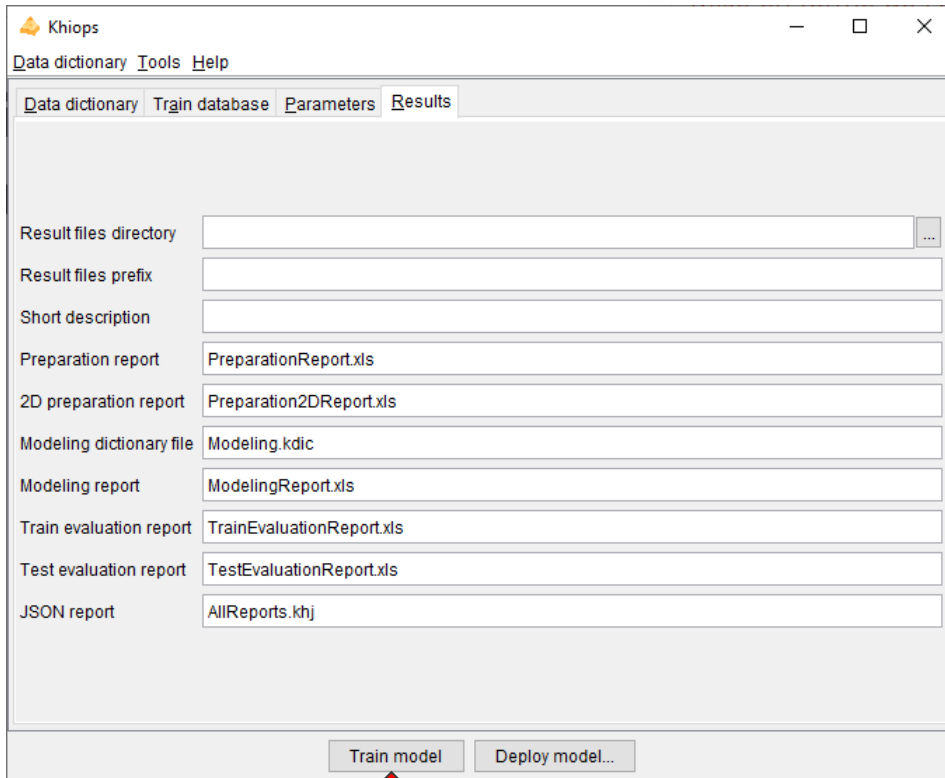
Train model Deploy model...

- 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









12

- **Step 5 : Start the analysis**



1 – Train model

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

Nom	Modifié
 TrainEvaluationReport.xls	09/06/21
 TestEvaluationReport.xls	09/06/21
 PreparationReport.xls	09/06/21
 ModelingReport.xls	09/06/21
 Modeling.kdic	09/06/21
 Iris.txt	12/06/21
 Iris.kdic	25/04/21
 AllReports.khj	09/06/21



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



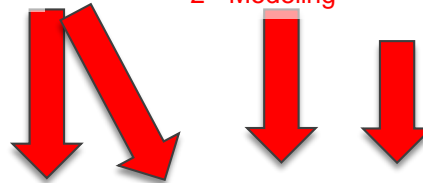
Exploratory of classification results using KHIOPS Visualization

13

1 - Preparation

2 - Modeling

3 - Evaluation



KHIOPS Visualization

File View Help Report a bug

Project Preparation Tree preparation Modeling 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

Target variable stats

Iris-setosa Iris-versicolor Iris-virginica

Species	Count
Iris-setosa	36
Iris-versicolor	20
Iris-virginica	30

4 Variables

Level distribution

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

PetalLength

Internal Coverage

Coverage

Bin	Coverage
[1,2,4]	30
[2,4,4.95]	36
[4.95,6.9]	30

Target distribution

Iris-setosa Iris-versicolor Iris-virginica

Values Probabilities

Bin	Iris-setosa	Iris-versicolor	Iris-virginica
[1,2,4]	36	0	0
[2,4,4.95]	0	20	0
[4.95,6.9]	0	0	30



Exploratory of classification results using KHIOPS Visualization

Preparation pane

Coverage of the selected variable

Zoom

Variables

Target distribution of the selected variable

The screenshot displays the KHIOPS Visualization software interface. The main window is titled 'KHIOPS Visualization' and has a menu bar with 'File', 'View', 'Help', and 'Report a bug'. Below the menu bar is a navigation bar with 'Project', 'Preparation', 'Tree preparation', 'Modeling', and 'Evaluation'. The 'Preparation' tab is active. On the left, there is a 'Summary' panel with a 'Dictionary: Iris' section. Below it, a '4 Variables' table is shown. The main area is divided into several sections: 'Target variable stats' with a bar chart showing the distribution of 'Iris-setosa', 'Iris-versicolor', and 'Iris-virginica'; 'Internal Coverage' with a bar chart showing coverage for three intervals of 'PetalLength'; and 'Target distribution' with a bar chart showing the distribution of 'Iris-setosa', 'Iris-versicolor', and 'Iris-virginica'. A 'Zoom' button is visible in the top right corner. A 'Scale chart' is also present in the bottom right corner.

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

Target variable stats

Internal Coverage

Target distribution



Exploratory of classification results using KHIOPS Visualization

15

Tree preparation pane

The screenshot displays the KHIOPS Visualization software interface. The main window is titled "KHIOPS Visualization" and has a menu bar with "File", "View", "Help", and "Report a bug". The navigation bar includes "Project", "Preparation", "Tree preparation" (highlighted), "Modeling", and "Evaluation".

Summary Panel:

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

Target variable stats: A horizontal bar chart showing the distribution of the target variable across three classes: Iris-setosa (dark blue), Iris-versicolor (light blue), and Iris-virginica (pink).

6 Variables Table:

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 Panel:

- Internal Coverage:** A bar chart showing coverage for nodes L4,]L5,L3], and L6.
- Target distribution:** A bar chart showing the distribution of the target variable for nodes L4,]L5,L3], and L6.

Decision tree Panel:

- Selection details:** L5, L3
- Leaf infos:** A table showing the target distribution for leaf nodes L5 and L3.

Hyper tree visualization Panel:

- Buttons: Values, Display purity by opacity, Display leaf sizes by population
- A large circular area containing a tree diagram with nodes L0, L1, L2, L3, L4, L5, and L6.



Exploratory of classification results using KHIOPS Visualization

Tree preparation pane

KHIOPS Visualization

Project Preparation **Tree preparation** Modeling Evaluation

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

Target variable stats
 Iris-setosa Iris-versicolor Iris-virginica

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

Internal Coverage
 Coverage

Target distribution
 Iris-setosa Iris-versicolor Iris-virginica

Decision tree
 L0 SepalLength
 L1 SepalWidth
 L3
 L4
 L2 PetalWidth
 L5
 L6

Selection details: L5, L3
 L5 ["Iris-setosa","Iris-versicolor","Iris-virginica"]
 L3 ["Iris-setosa","Iris-versicolor","Iris-virginica"]

Leaf infos
 Target distribution
 Probabilities

Hyper tree visualization
 Values
 Display purity by opacity
 Display leaf sizes by population

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



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

The screenshot displays the KHIOPS Visualization interface with the following components:

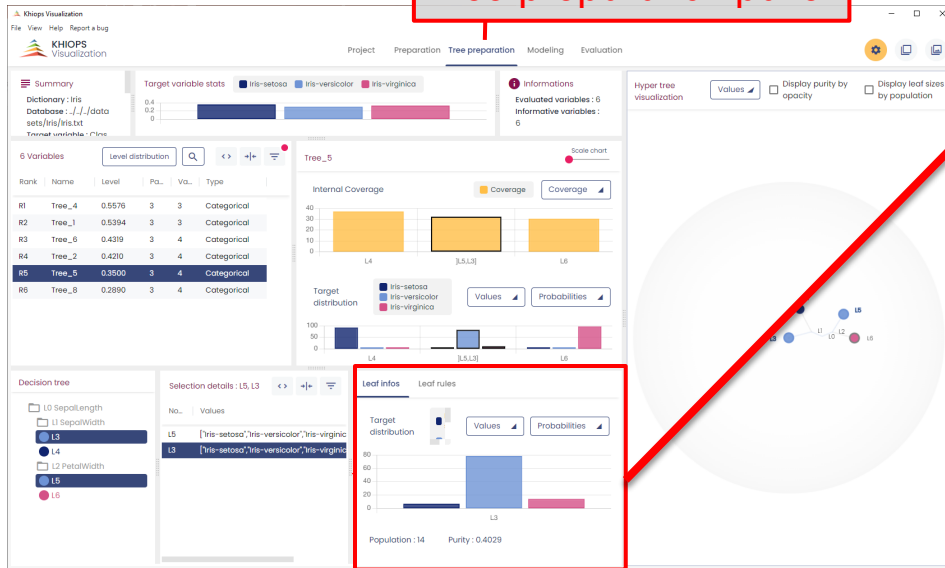
- Summary Panel:** Shows target variable stats for Iris-setosa, Iris-versicolor, and Iris-virginica. Includes a dictionary, database path, and target variable.
- 6 Variables Table:**

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 Panel:** Shows internal coverage and target distribution for node]L5,L3]. Includes a scale chart and target distribution bar chart.
- Decision tree Panel:** Shows a tree structure with nodes L0 through L6. Node L3 is selected.
- Leaf infos Panel:** Shows leaf rules for node L3: ["Iris-setosa","Iris-versicolor","Iris-virginica"]. Includes target distribution and population/purity statistics.
- Hyper tree visualization Panel:** Shows a tree diagram with nodes L0 through L6. A red arrow points to node L3, labeled "User click".



Exploratory of classification results using KHIOPS Visualization

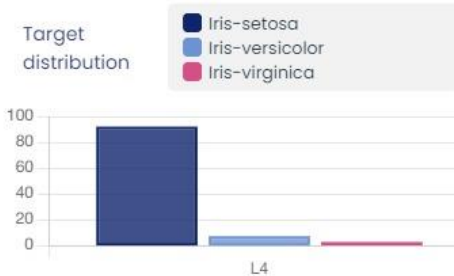
Tree preparation pane



Information on selected leaf

Leaf infos

Target distribution in leaf



Leaf rules

Sequence of tree 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

19

The screenshot displays the KHIOPS Visualization software interface. At the top, the 'Modeling pane' is highlighted with a red box. The interface includes a navigation menu with 'Project', 'Preparation', 'Tree preparation', 'Modeling', and 'Evaluation'. The 'Modeling' pane shows a 'Trained predictors' section with 'Selective Naive Bayes : 4 Variables'. Below this, a table lists the variables used in the predictor:

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

The 'PetalLength' variable is highlighted in blue, and a red box labeled 'Variables of the predictor' points to it. A red box labeled 'Selected predictor' points to the 'Selective Naive Bayes' dropdown menu. The interface also features a 'Target variable stats' bar chart showing the distribution of the target variable across three classes: Iris-setosa, Iris-versicolor, and Iris-virginica. Additionally, there is a 'PetalLength' section with an 'Internal Coverage' bar chart and a 'Target distribution' bar chart. The 'Current interval' is shown as [1,2.4].



Exploratory of classification results using KHIOPS Visualization

20

KHIOPS Visualization

File View Help Report a bug

Project Preparation Tree preparation Modeling Evaluation

Evaluation type

Type	Dictionary	Instances
Train	Iris	99
Test	Iris	51

Evaluation list

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 the selected predictor

Confusion matrix of Selective Naive Bayes

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

Cumulative gain chart of Iris-setosa

Cumulative gain charts of each predictor




21

Exercises A and B ...

A : Perform a supervised classification on sample database Iris

B : Perform a Supervised classification on sample database Adult

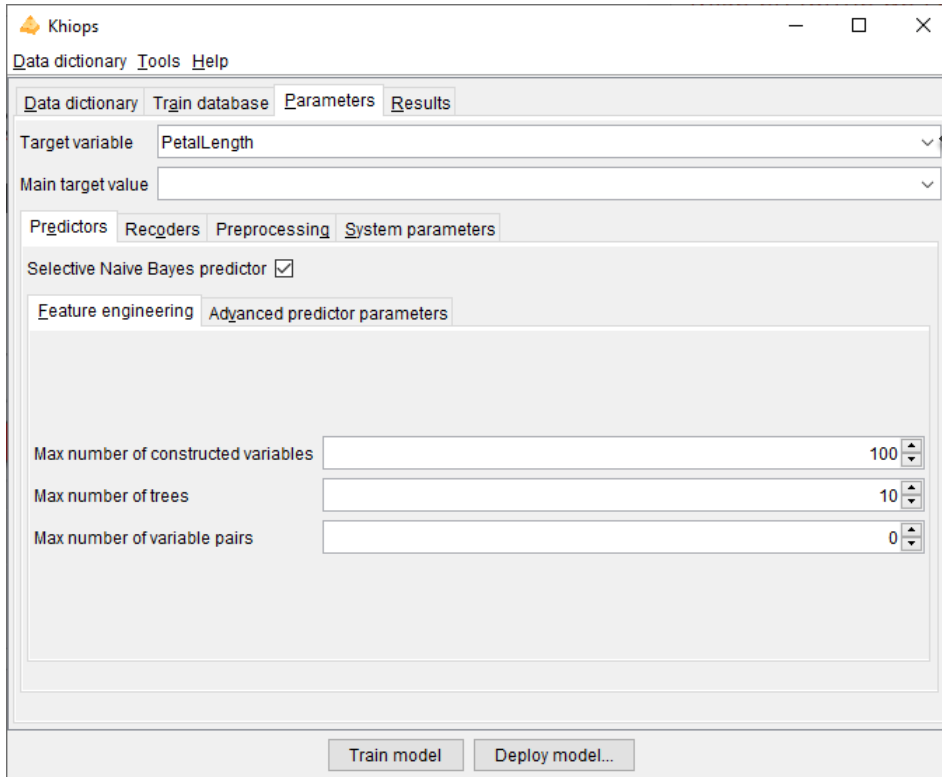
 Interpret the analysis results



Regression

(supervised)

- Same as classification
with a numerical target variable



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



Exploratory of regression results using KHIOPS Visualization

KHIOPS Visualization

File View Help Report a bug

Project Preparation Modeling Evaluation

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

4 Variables

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	3	3	22	Numerical

Class

Internal Coverage

Target values: 1 (Class, PetalLength)

Standard Frequency

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

Derivation rule

Name: Class

Values of Class | Frequency | Interval of PetalLength



24

Exercise C ...

C : Perform a regression of variable PetalLength of Iris

➔ Interpret the analysis results



Correlation analysis

(unsupervised, bivariate)

25

- **Train a correlation model between two variables**
(*categorical, numerical, both*)

Khiops

Data dictionary Tools Help

Data dictionary Train database Parameters Results

Target variable

Main target value

Predictors Recorders 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 5

Train model Deploy model...

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)

26

- **Train a correlation model** : advanced parameters

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

First name	Second name
SepalLength	SepalWidth
PetalLength	PetalWidth



Exploratory of correlation results using KHIOPS Visualization

KHIOPS Visualization

File View Help Report a bug

KHIOPS Visualization

Project Preparation Preparation 2D

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	PetalLength	0.1416	2	3	3	5
R3	PetalLength	PetalWidth	0.0823	2	3	3	4
R4	Class	SepalLength	0.0584	2	3	3	8
R5	PetalLength	SepalLength	0.0498	2	4	4	10

Level distribution

Matrix Cells

Standard Frequency

Co-occurrence I (Class, PetalWidth)

Contrast

0.3662

-0.3662

Class

Values of Class | Frequency | Interval of PetalWidth

Preparation 2D pane

Co-occurrence matrix of the selected variable pair

Variables pairs



28


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*

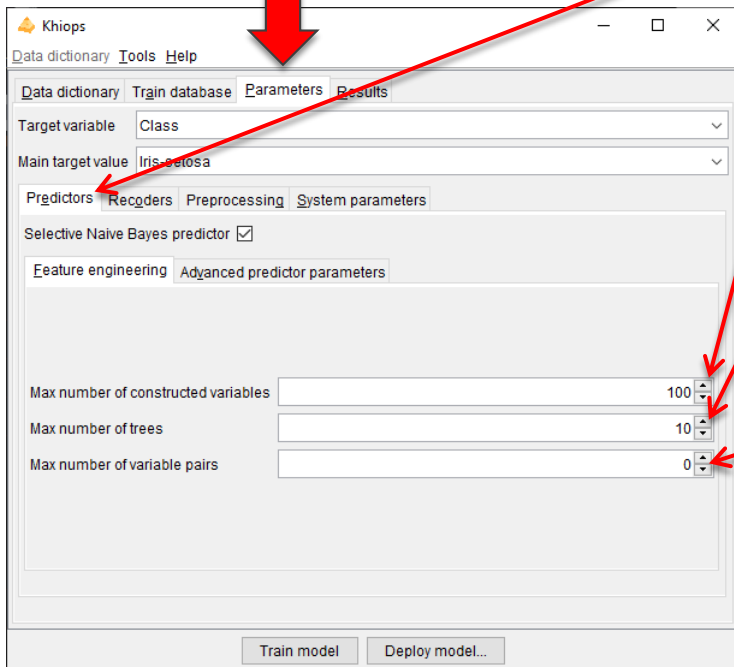
 Interpret the analysis results



Variable construction

29

- Parameters



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



30

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

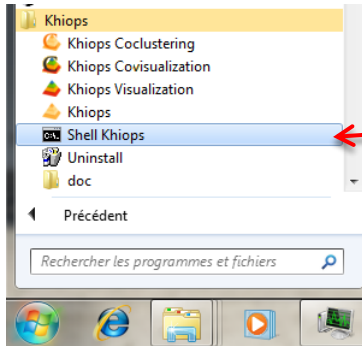
Integration in information systems

31

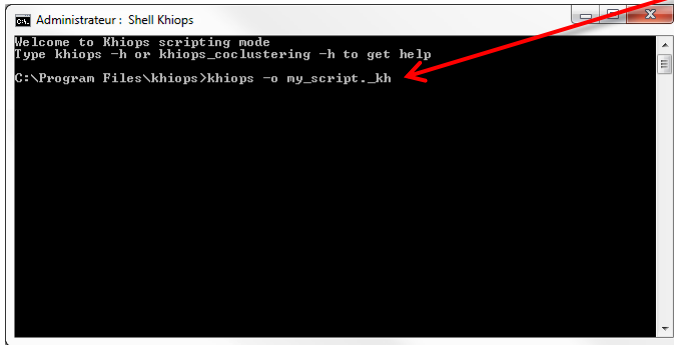
- Batch mode
 - to record and replay Khiops scripts
 - to perform any Khiops task from any programming language
 - see next slide
- Khiops Native Interface (KNI)
 - dynamic link library (DLL) for online deployment of Khiops models
 - package to download from <https://khiops.org>
- Python Khiops Library (pykhiops)
 - to perform any Khiops task from python
 - to inspect any Khiops analysis results from python
 - python package available from <https://khiops.org>
- 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



Record a script « automatically » using Khiops user interface

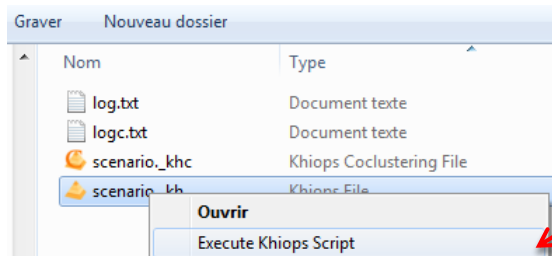
```
khiops -o my_script._kh
```

o = output

- Replay a script from the shell

```
khiops -i my_script._kh
```

i = input



Replay a script from Windows Explorer
right click on script file



33

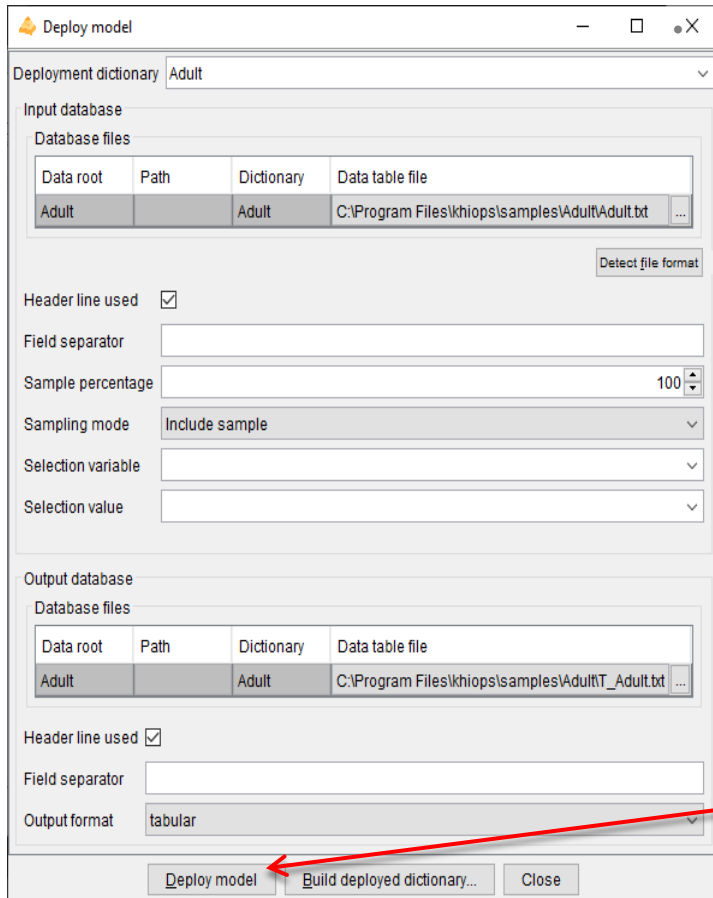
Exercise I ...

I : Record a script file, then replay it ...



Deploy a model

34



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



35

Exercise J ...

J: Deploy a classifier on database Iris

Khiops Coclustering & Khiops Covisualization

36



- **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

37



- **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

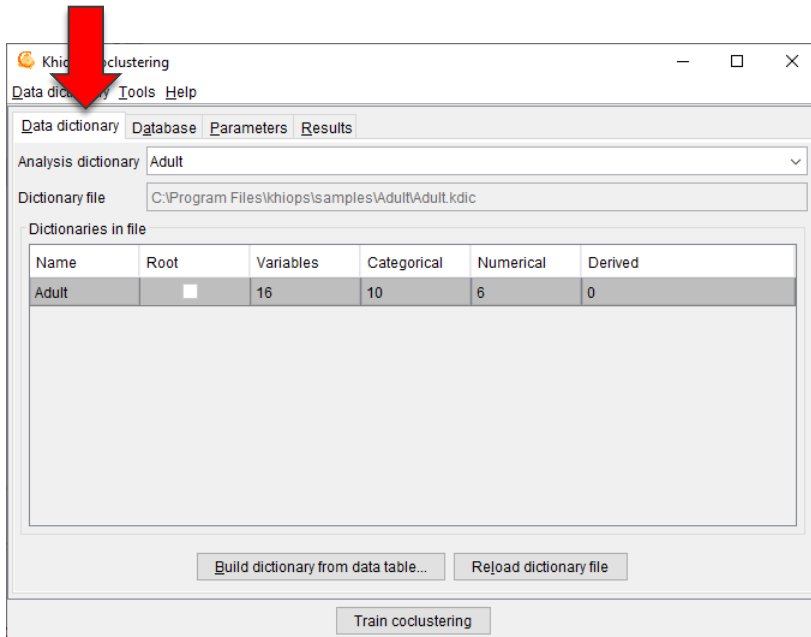




Train a coclustering model

38

- **Step 1 : Open an existing dictionary**
(ex: sample Adult.kdic)
- Description of variables to use during analysis



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  fnlwtgt;
    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

39

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

Khiops Coclustering

Data dictionary Tools

Data dictionary Database Parameters Results

Database files

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

Detect file format

Header line used

Field separator

Sample percentage

Sampling mode

Selection variable

Selection value

Train coclustering

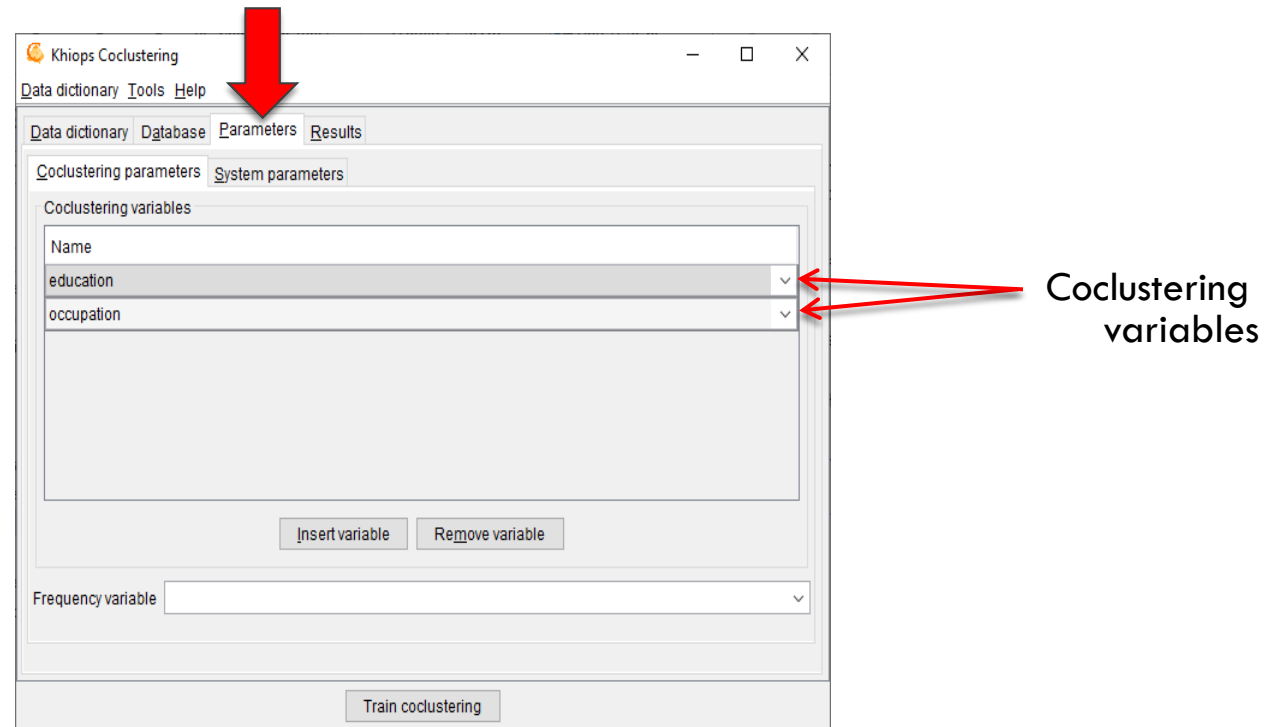
Detect file format : heuristic help that scans the first few lines to guess the file format. The header line and field separator are updated on success, with a warning or an error in the log window only if necessary.



Build a coclustering model

40

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





Build a coclustering model

41

- **Step 4 : Results**



Khiops Coclustering

Data dictionary Tools Help

Data dictionary Database Parameters Results

Result files directory CorrelationEducationOccupation

Result files prefix

Short description

Coclustering report Coclustering.khc

Export JSON

Train coclustering

- 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

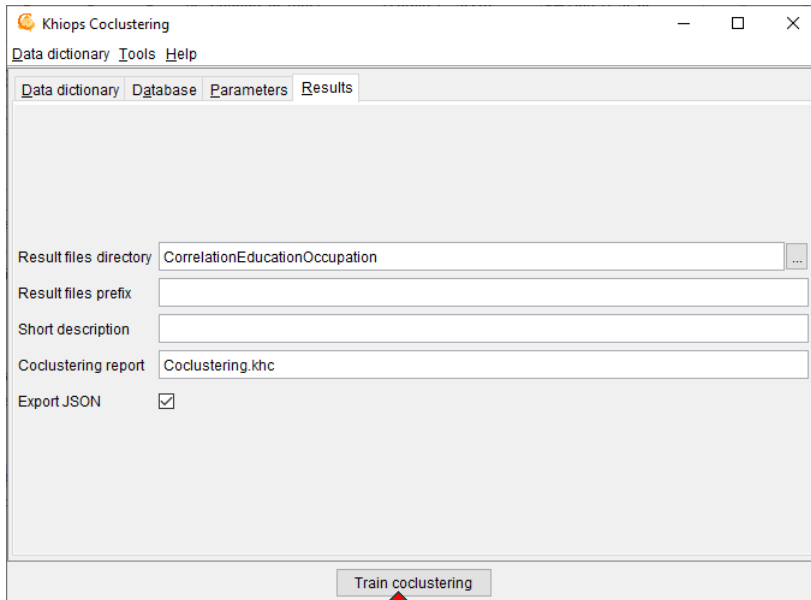




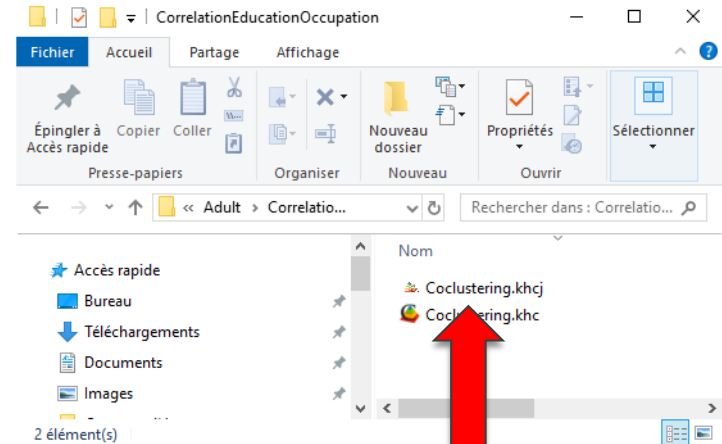
Build a coclustering model

42

- **Step 5 : Start the analysis**



1 – Train the coclustering



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

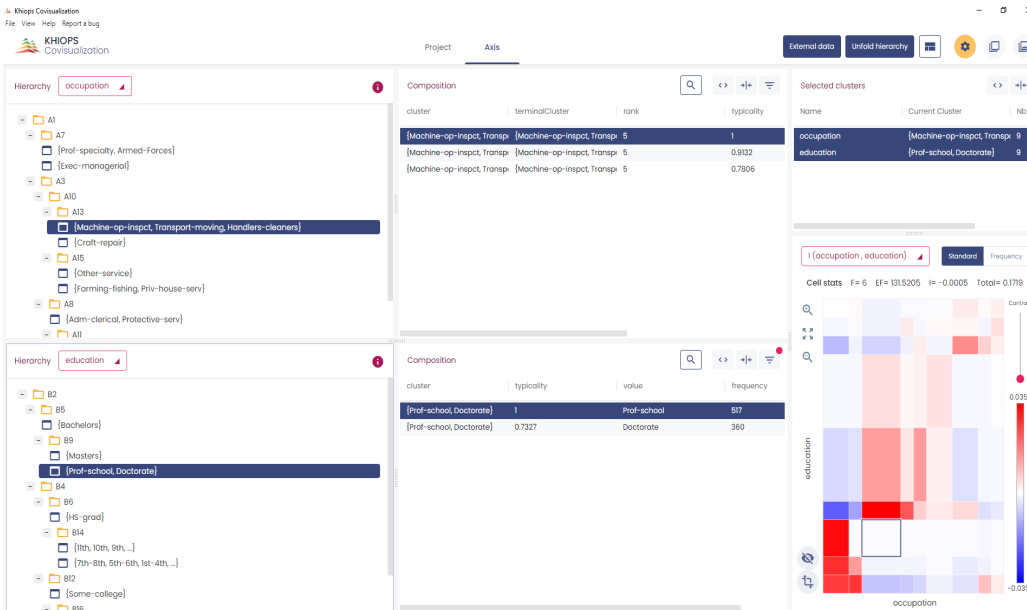
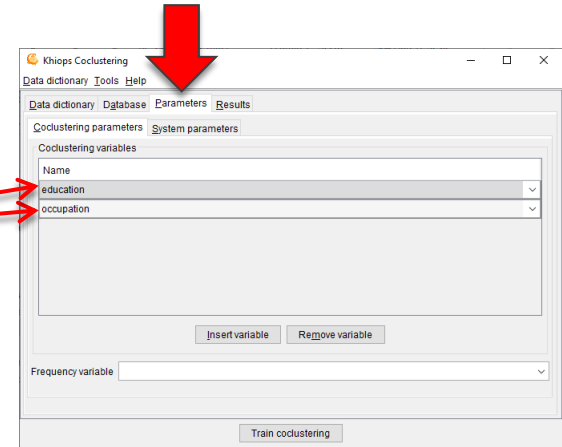


Example: base Adult education*occupation

43



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



- With Khiops Covisualization
 - Exploratory analysis of the results



44

Exercise J ...

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

➔ Explore the analysis results

Khiops Covisualization

45

Interactive hierarchy of variable #1

Hierarchy: occupation

- A1
 - A7
 - {Prof-specialty, Armed-Forces}
 - {Exec-managerial}
 - A3
 - A10
 - A13
 - {Machine-op-inspct, Transport-moving, Handlers-cleaners}
 - {Craft-repair}
 - A15
 - A8
 - A9
 - All

Interactive hierarchy of variable #2

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

List of variables

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

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

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

Heatmap: Contrast scale from -0.0357 to 0.0357. X-axis: occupation, Y-axis: education.

Khiops Covisualization

46

The screenshot displays the Khiops Covisualization interface with two main panels. The top panel is for the 'occupation' variable, and the bottom panel is for the 'education' variable. Both panels show a hierarchy tree on the left, a composition table in the center, and a heatmap on the right.

Top Panel: occupation

Composition Table:

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

Heatmap: A heatmap showing the relationship between 'occupation' and 'education'. The color scale ranges from -0.0357 (blue) to 0.0357 (red). The selected cluster for 'occupation' is highlighted in red.

Bottom Panel: education

Composition Table:

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

Heatmap: A heatmap showing the relationship between 'education' and 'occupation'. The color scale ranges from -0.0357 (blue) to 0.0357 (red). The selected cluster for 'education' is highlighted in red.

Annotations: Two red boxes with arrows point to the selected clusters in the composition tables. The top box contains the text "Composition of selected cluster of variable #1" and the bottom box contains "Composition of selected cluster of variable #2".

Khiops Covisualization

47

The screenshot displays the Khiops Covisualization interface. On the left, there are two hierarchical trees: 'occupation' and 'education'. The 'occupation' tree shows clusters like A1, A7, A3, A10, A13, A15, A8, and A11. The 'education' tree shows clusters like B2, B5, B9, B4, B6, B14, B12, and B16. The main area shows a 'Composition' table with columns for 'cluster', 'terminalCluster', 'rank', and 'typicality'. Below this is a 'frequency' table with columns for 'cluster', 'terminalCluster', and 'frequency'. A red-bordered box highlights a section of the interface with the following text:

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

On the right side of the interface, there is a 'Selected clusters' table and a 'Co-occurrence matrix' heatmap. The 'Selected clusters' table shows the following data:

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

The heatmap shows the relationship between 'occupation' (x-axis) and 'education' (y-axis). A color scale on the right indicates the contrast, ranging from -0.0357 (blue) to 0.0357 (red). The heatmap shows a strong positive correlation (red) between the selected clusters in both partitions.

Khiops Covisualization

48

The screenshot displays the Khiops Covisualization interface with two main panels. The top panel is for the 'occupation' hierarchy, and the bottom panel is for the 'education' hierarchy. Each panel includes a tree view on the left, a 'Composition' table in the center, and a heatmap on the right. A red callout box highlights the 'Unfold hierarchy' button and the 'Selected clusters' table in the top panel.

Top Panel: occupation

Hierarchy

- 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, Tra	{Machine-op-inspct, Transp	5	

Selected clusters

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

Bottom Panel: education

Hierarchy

- 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

Heatmap

education

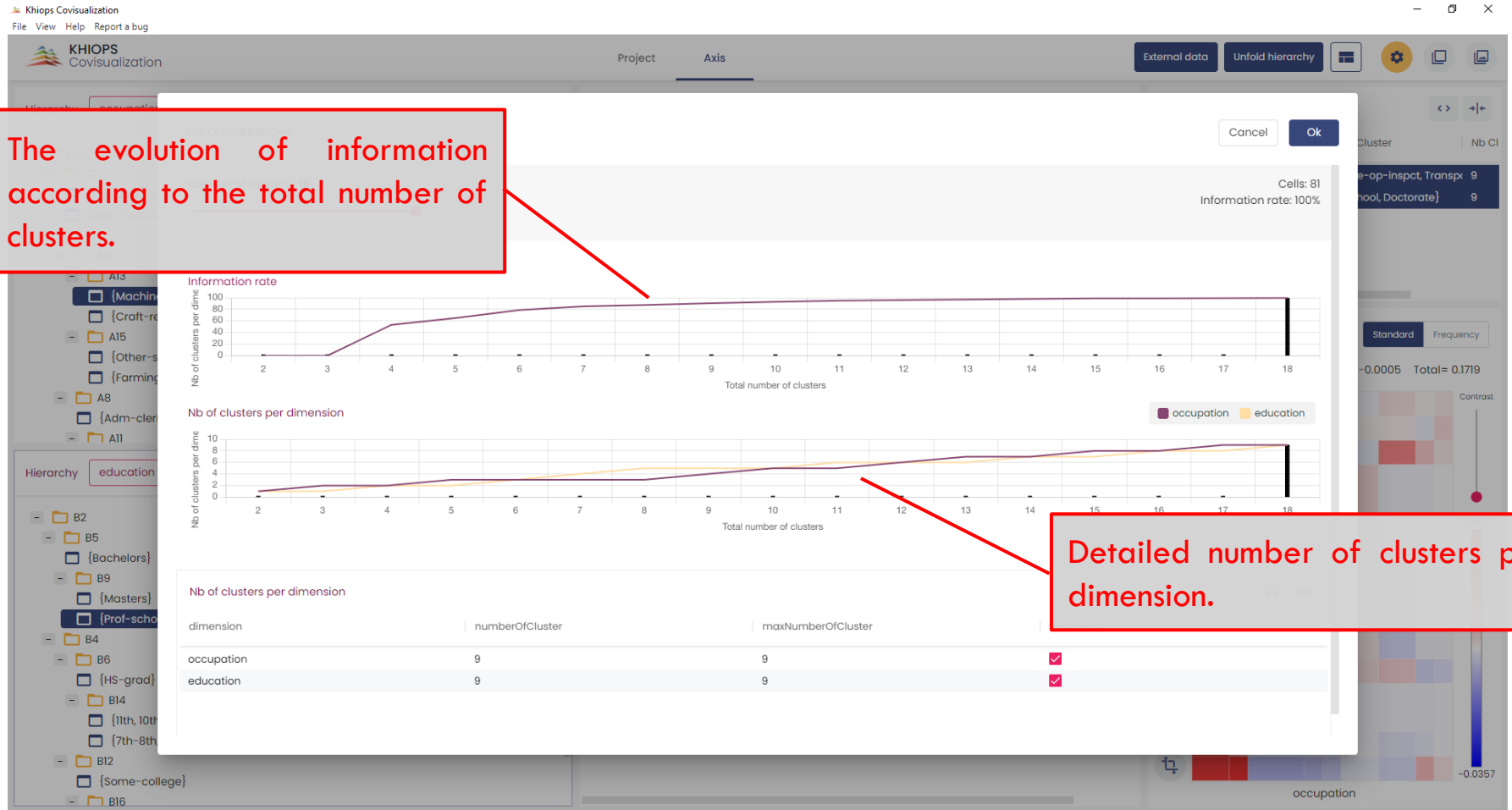
occupation

Contrast: 0.0357 to -0.0357

EF= 131.5205 I= -0.0005 Total= 0.1719

Khiops Covisualization

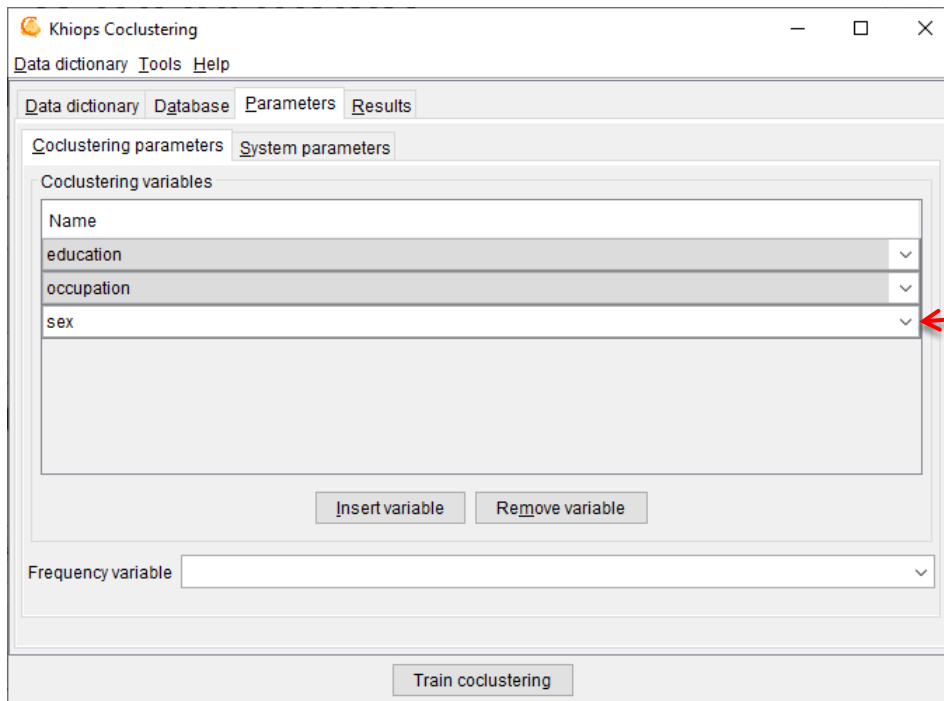
49



Training a triclustering

50

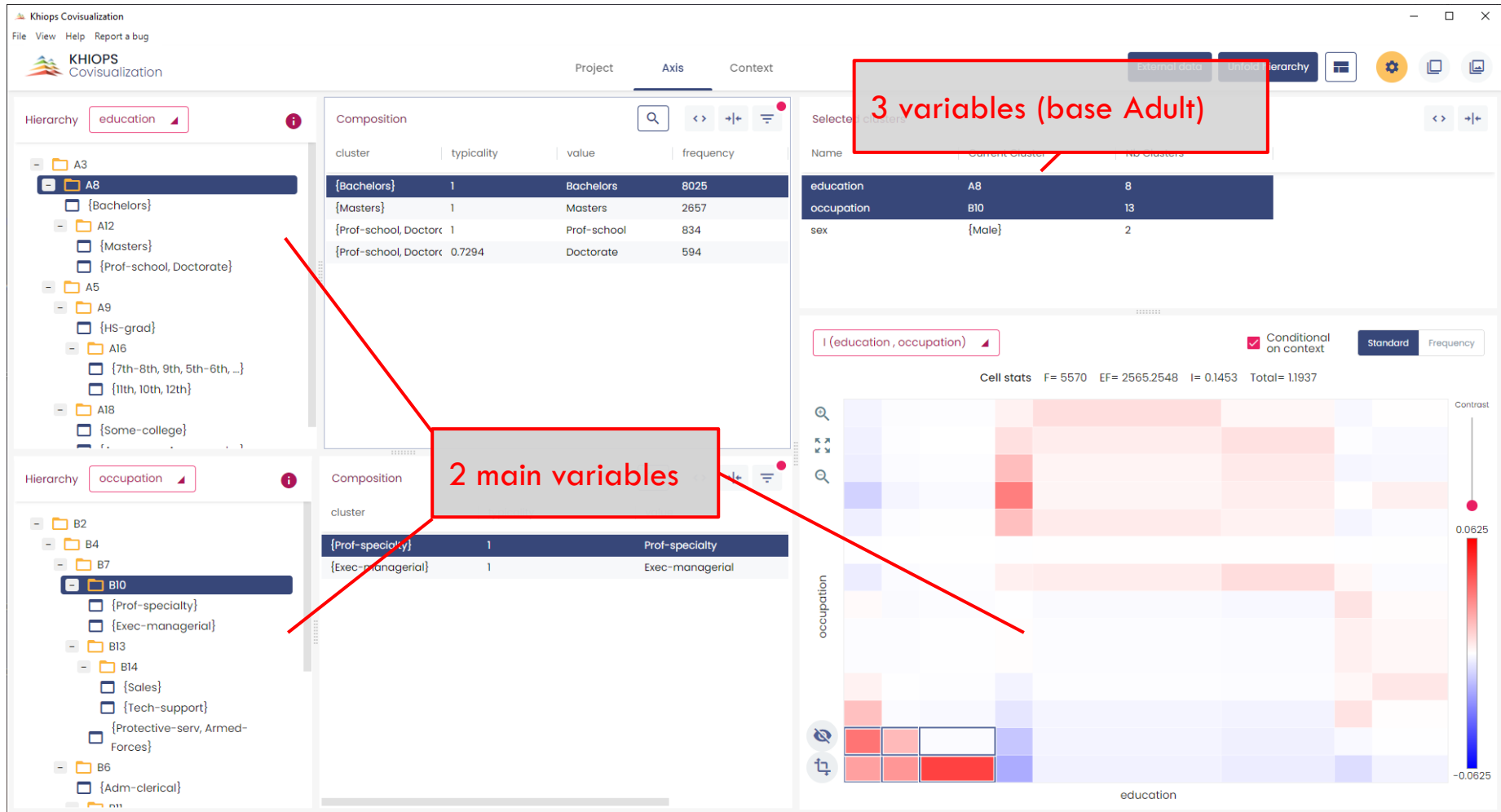
- Same as coclustering (Step 3) by inserting a third variable



The third variable

Exploring a triclustering

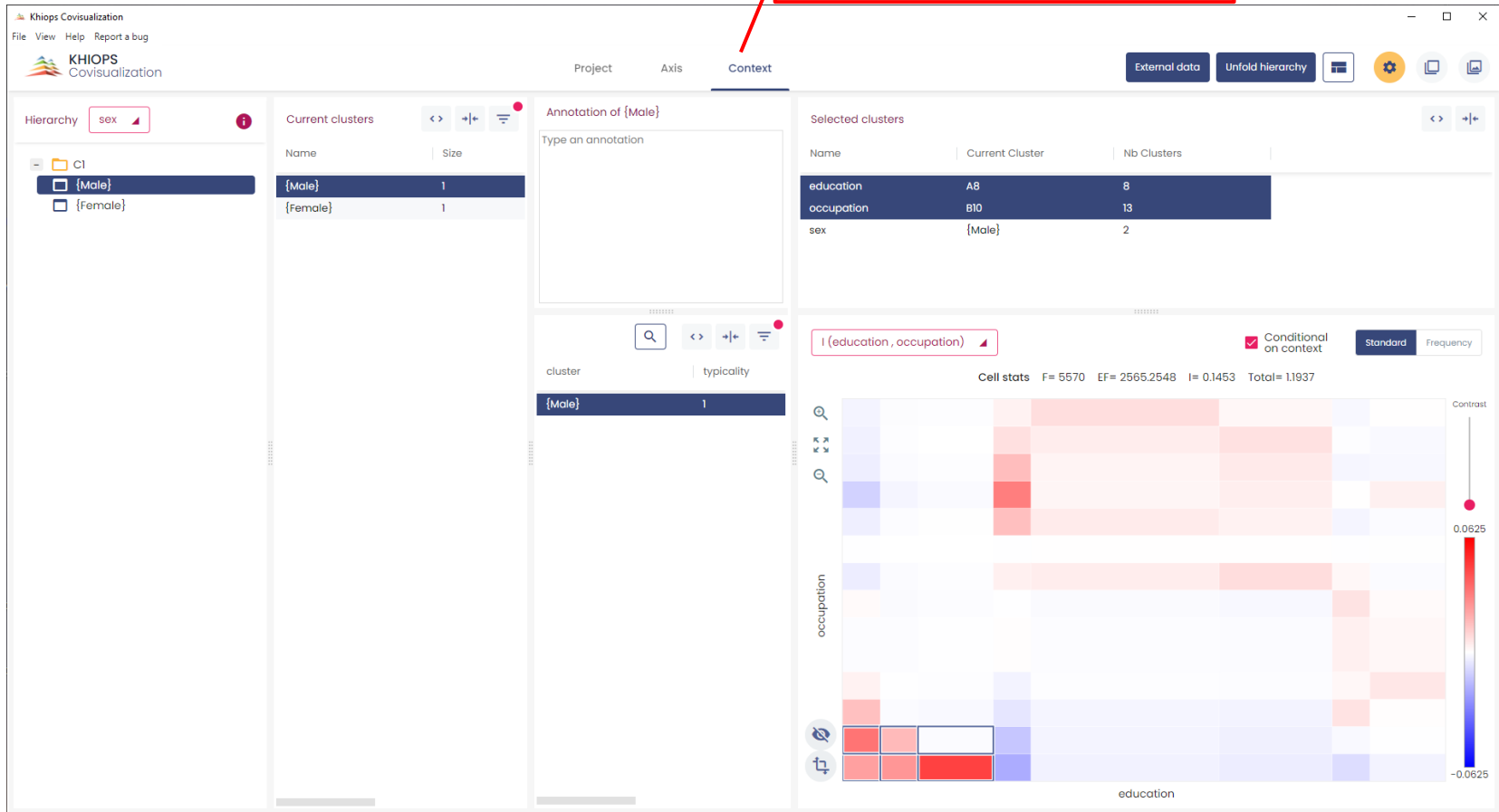
51



Exploring a triclustering

52

Context pane



Exploring a triclustering

53

UNFOLD HIERARCHY

Number of clusters : **8**

Information rate

Nb of clusters per dimension

dimension	numberOfCluster	maxNumberOfCluster	foldUnfold
education	3	8	<input checked="" type="checkbox"/>
occupation	3	13	<input checked="" type="checkbox"/>
sex	2	2	<input checked="" type="checkbox"/>

Cells: 18
Information rate: 69.2%

Standard Frequency

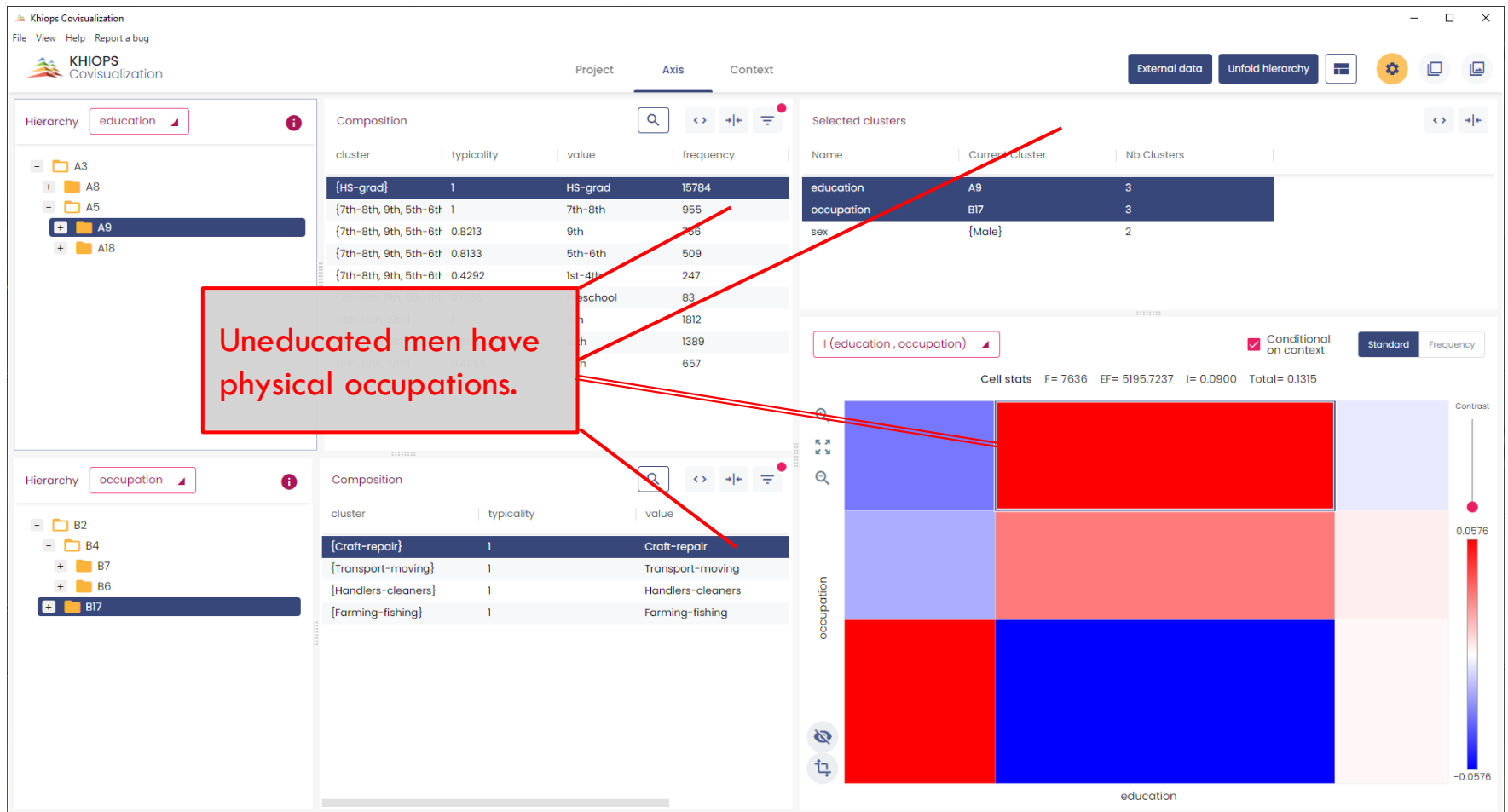
Contrast: 0.0625

education

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

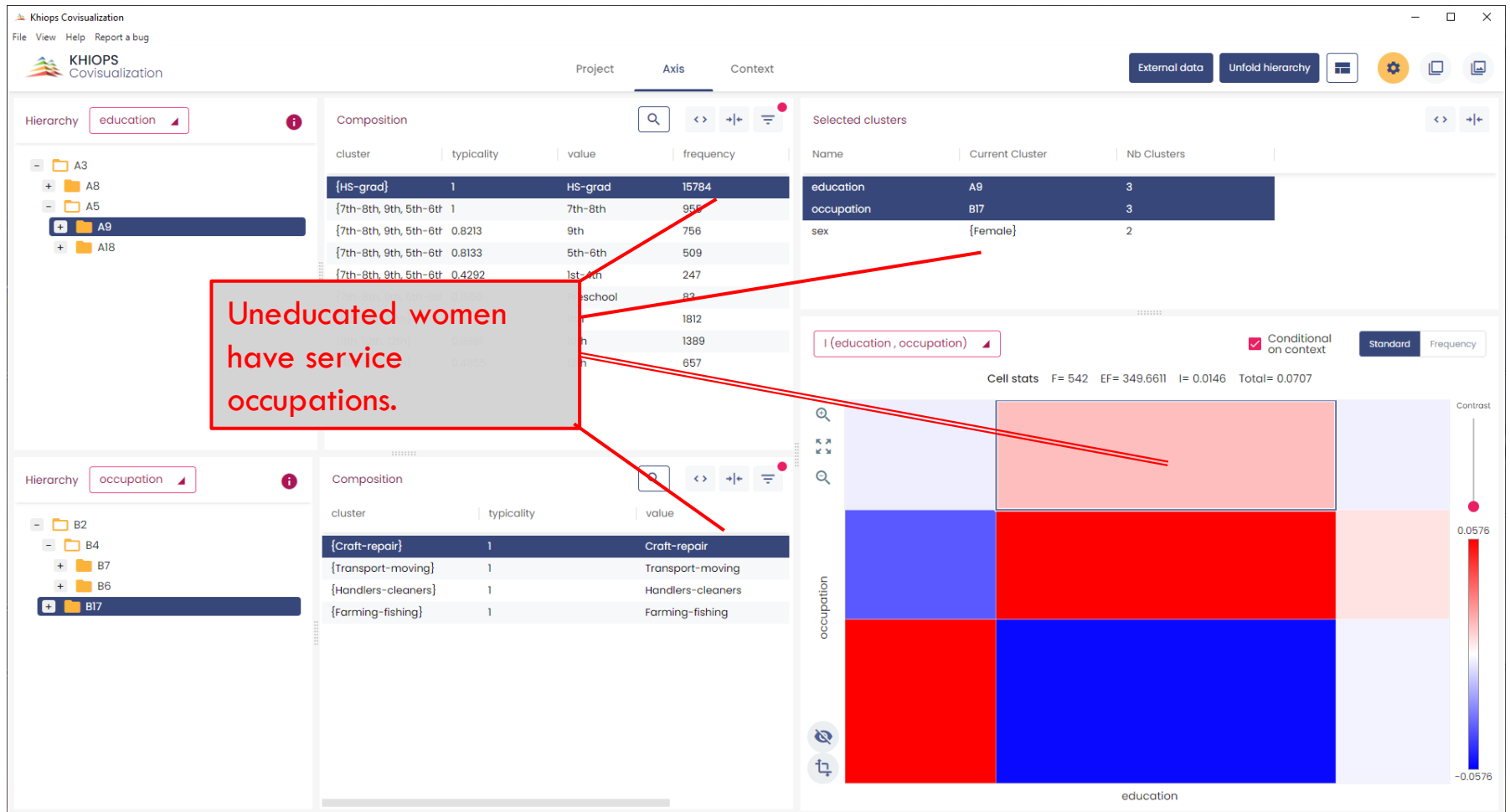
Exploring triclustering

54



Exploring a triclustering

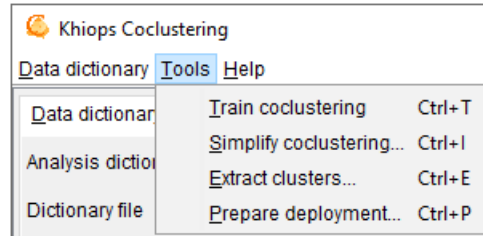
55





Exploiting a coclustering model

56



- **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

57

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

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)

Result files directory: C:\Program Files\khiops\samples\Adult\CorrelationE ...

Result files prefix:

Simplified coclustering report: SimplifiedCoclustering.khc

Export JSON:

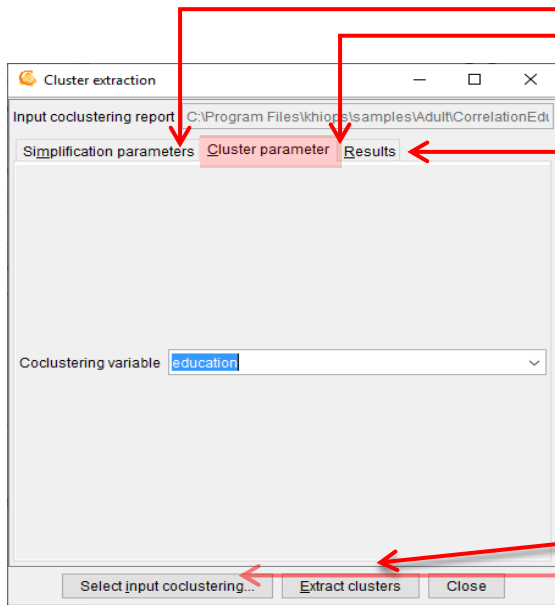
3- Select result files directory

4- Click on « Simplify coclustering »



Extracting clusters in a text file

58



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



59

Exercise K, L ...

K : Simplify previously built adult coclustering model
Keep 50% of the information in the model

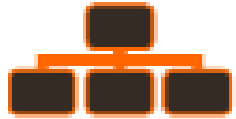
➔ Explore the simplified analysis results with Khiops covisualization

L : Extract clusters from variable education of adult coclustering model

➔ Inspect the cluster file with a text editor

Multi-table functionalities

60



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

Why extending to multi-table?

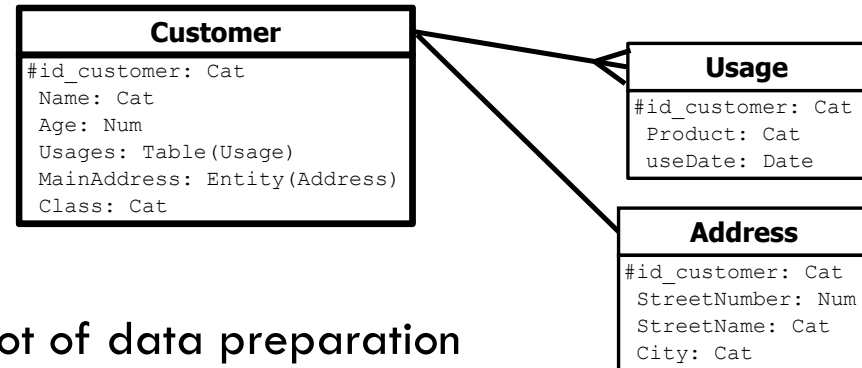
61

- 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
- ...



- 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

62

- **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

63

- 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

64

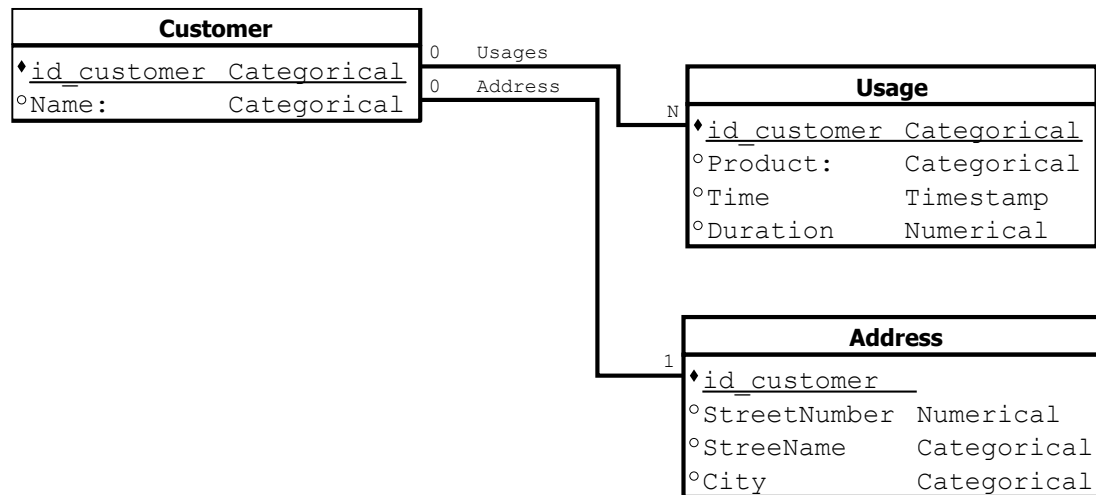
- Khiops can deal with multi-table databases
 - star schema: one root entity and several 0-1 or 0-n secondary entities
 - snowflake schemas and beyond
- Impact on Khiops
 - Multi-table dictionary
 - to describe star-schema input representation
 - Multi-table database
 - to store input data on multiple 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

65

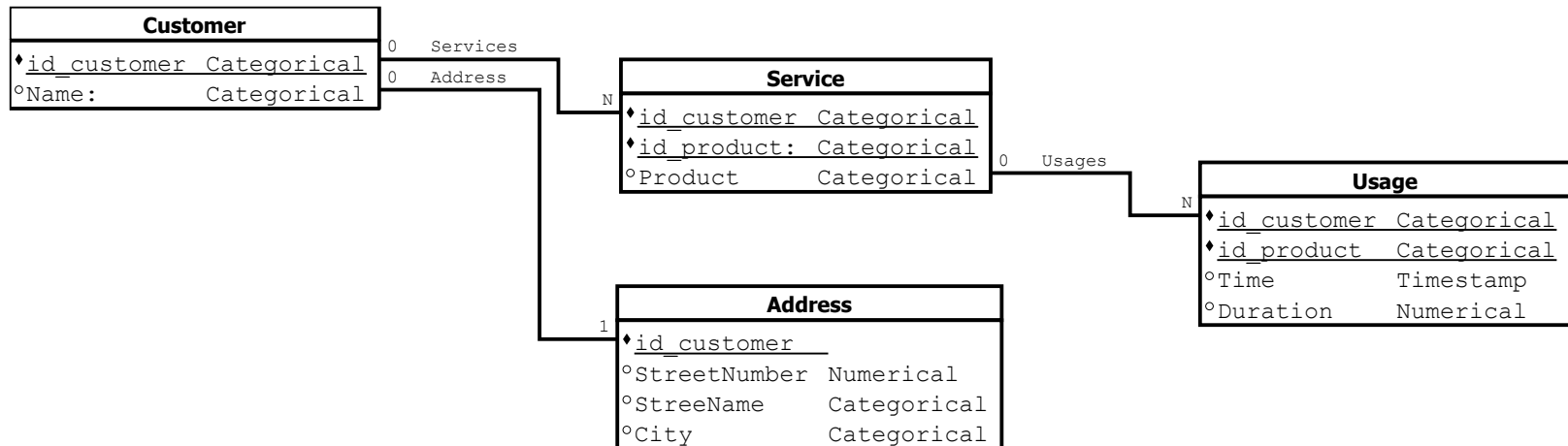
- One root entity
 - secondary tables in 0-1 relationship: Entity
 - secondary tables in 0-n relationship: Table



Snowflake schema

66

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

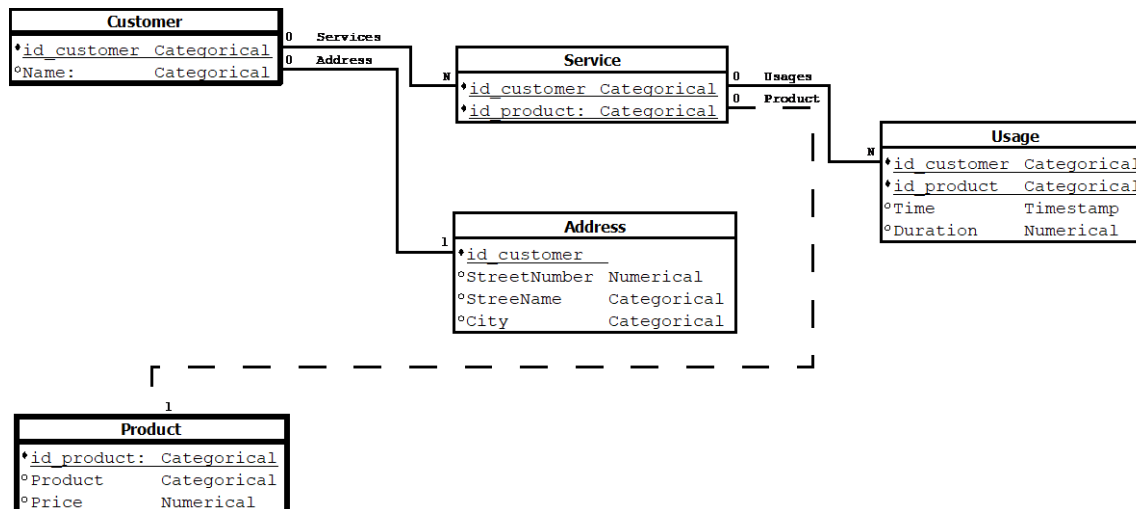


- Example in samples/Customer
 - detailed explanations in sample

External tables

67

- 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

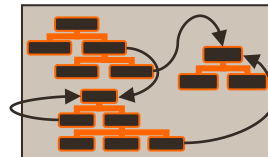
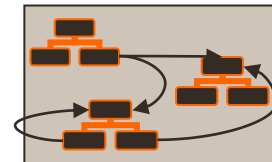
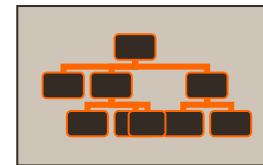
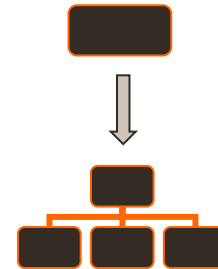


- Example in samples/CustomerExtended
 - detailed explanations in sample

Multi-table schemas: synthesis

68

- 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

69

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)

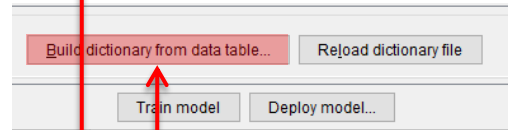
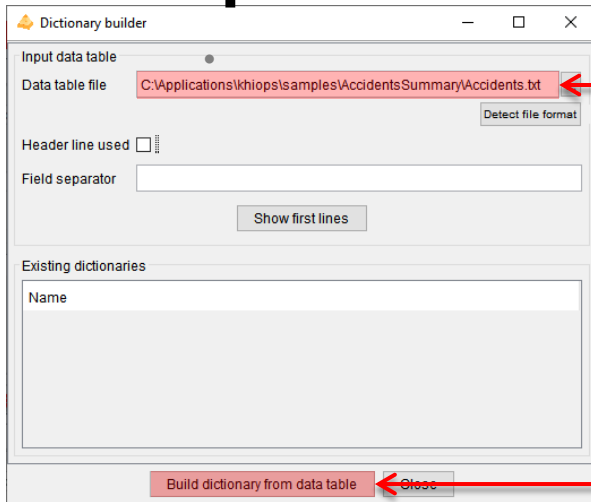


Build a multi-table dictionary

70

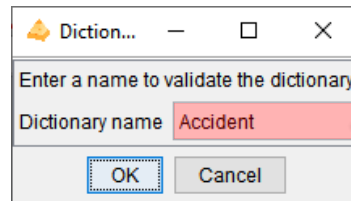
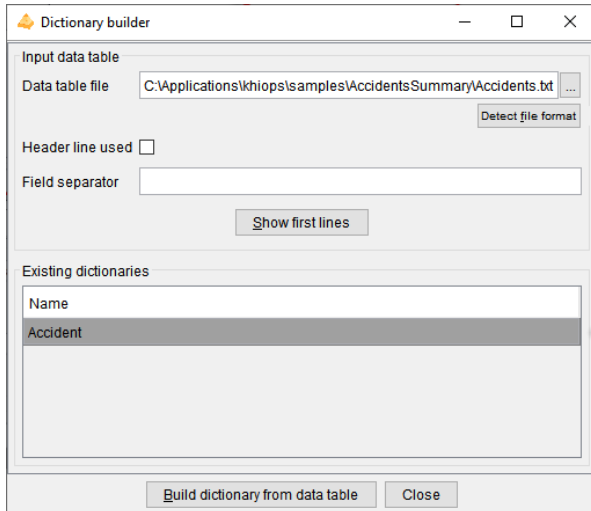
- Step 1: Build one dictionary per data table**

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`

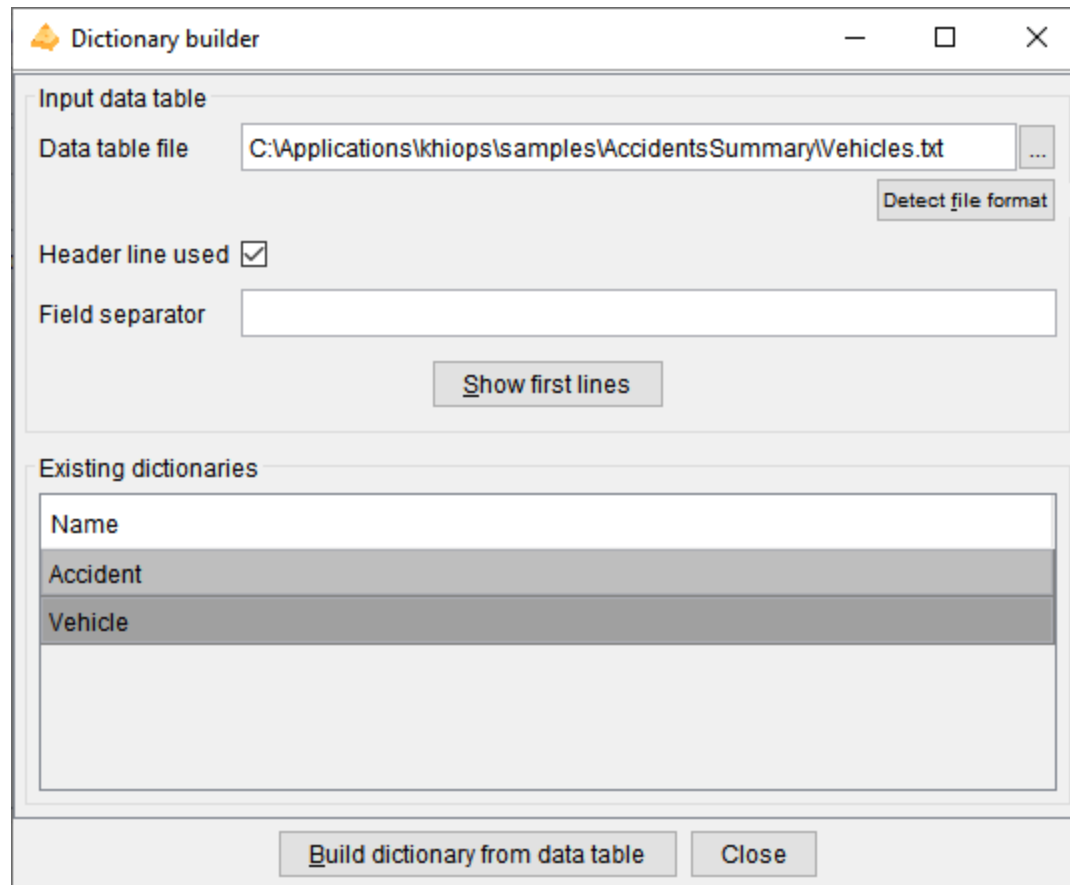


Build a multi-table dictionary

71

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

- b : Repeat for the `vehicles.txt` table



Dictionary builder

Input data table

Data table file ...

Detect file format

Header line used

Field separator

Show first lines

Existing dictionaries

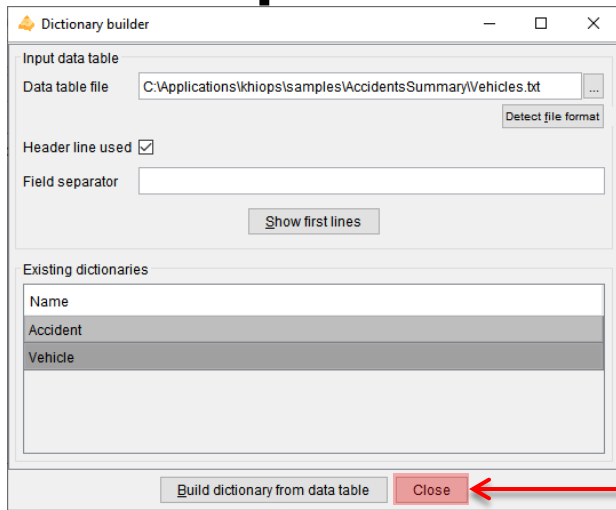
Name
Accident
Vehicle

Build dictionary from data table Close

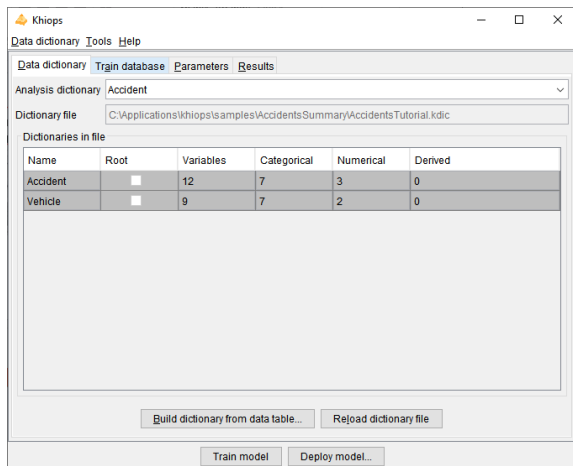
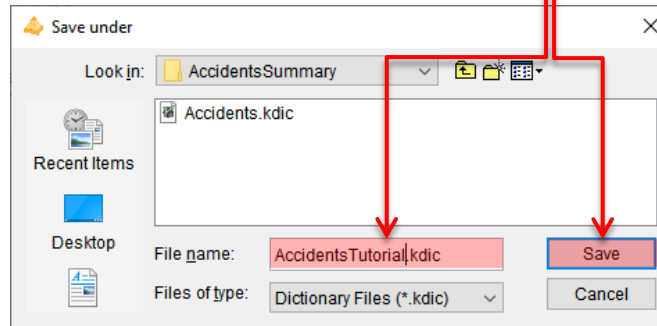


Build a multi-table dictionary

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



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



Build a multi-table dictionary

73

• **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

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

```
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

6. Save the dictionary file

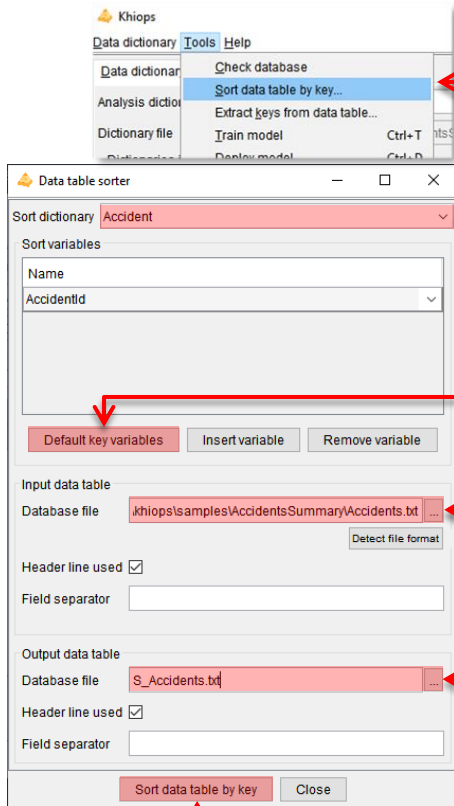


Sort data table files (if necessary)

74

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



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

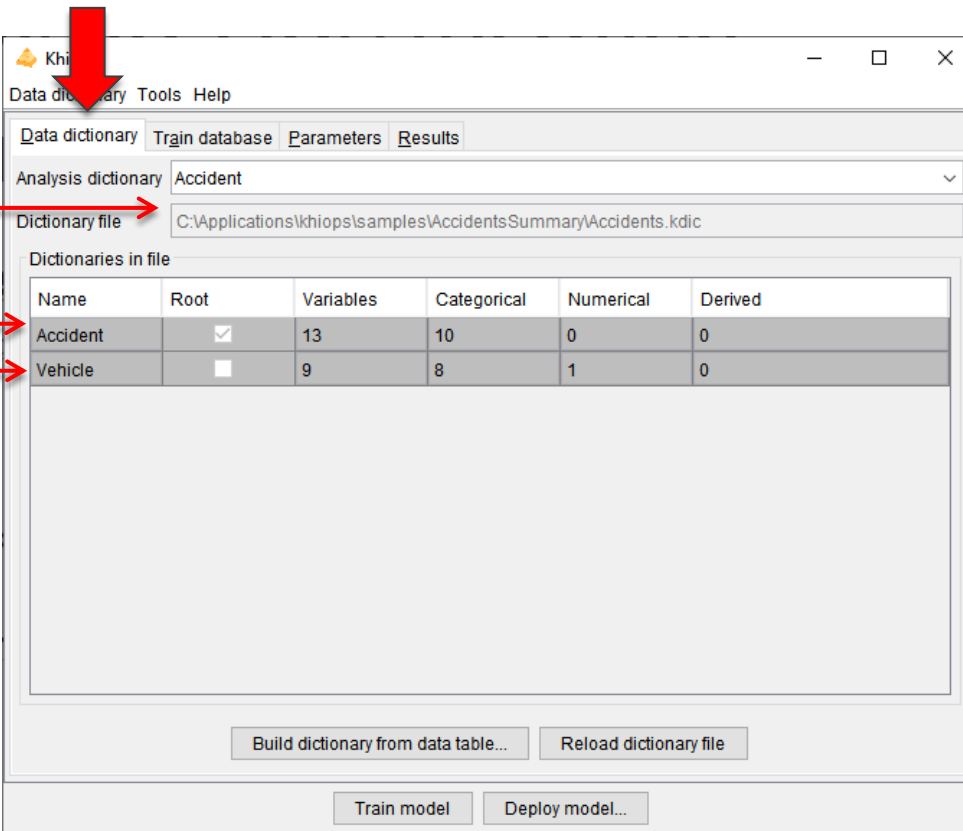
- The output file is sorted by key
- All native variables are kept (used or not in the dictionary)
- Derived variables are ignored



Supervised classification

75

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



Analysis dictionary

Root entity

Secondary entity

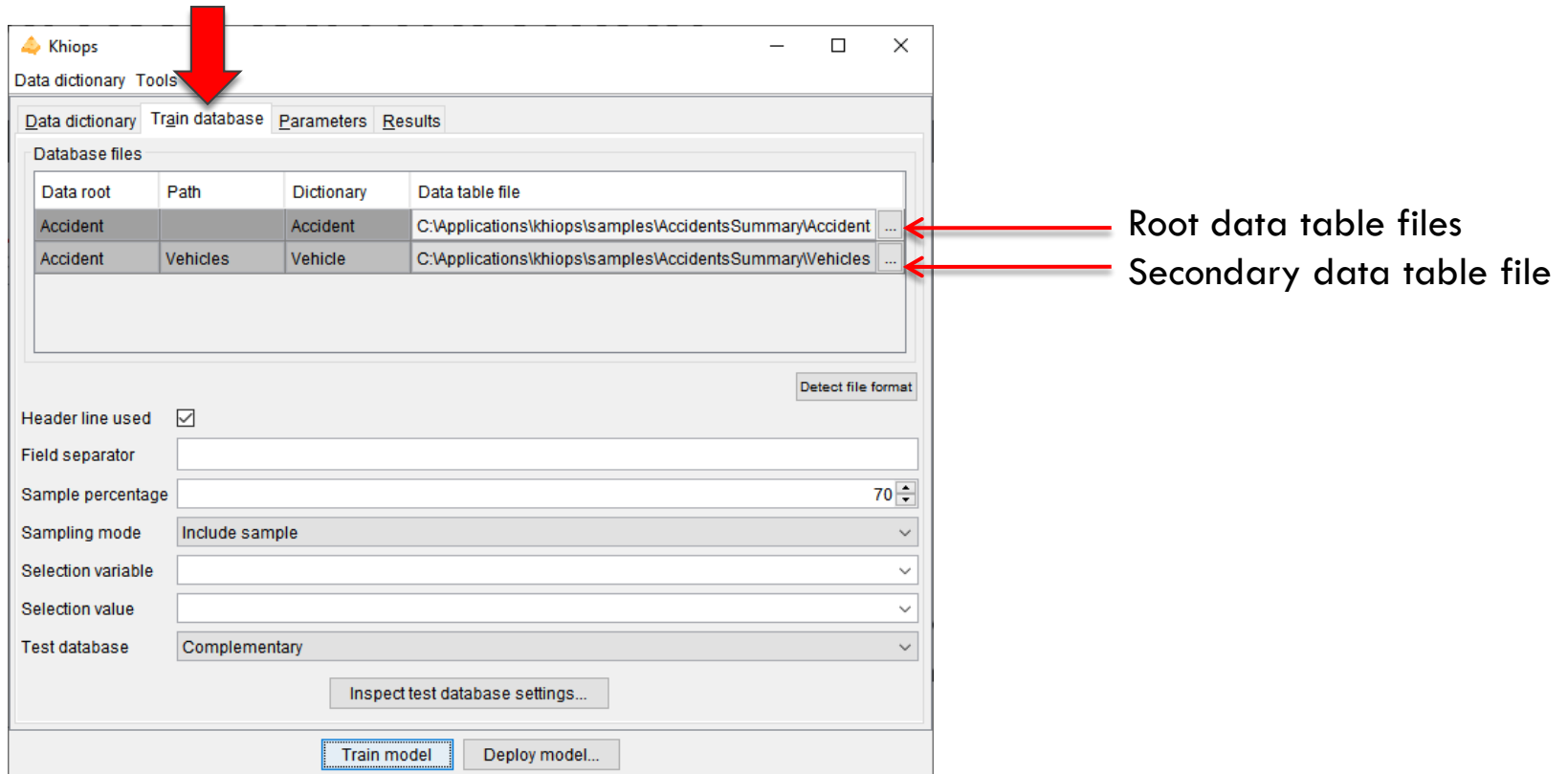
Name	Root	Variables	Categorical	Numerical	Derived
Accident	<input checked="" type="checkbox"/>	13	10	0	0
Vehicle	<input type="checkbox"/>	9	8	1	0



Supervised classification

76

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



The screenshot shows the 'Train database' dialog box in the Khiops software. The 'Train database' tab is selected. The 'Database files' section contains a table with the following data:

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

Below the table, there are several options: 'Header line used' (checked), 'Field separator' (empty), 'Sample percentage' (70), 'Sampling mode' (Include sample), 'Selection variable' (empty), 'Selection value' (empty), and 'Test database' (Complementary). At the bottom, there are buttons for 'Inspect test database settings...', 'Train model', and 'Deploy model...'.

Annotations:

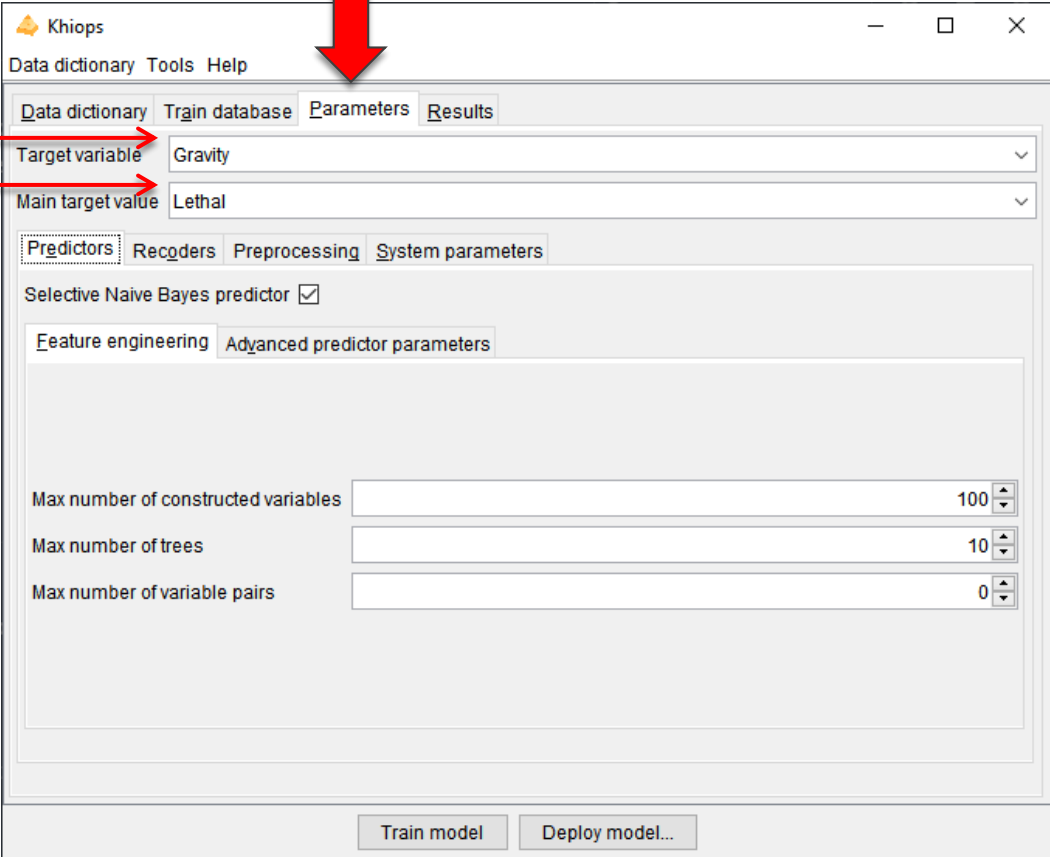
- A red arrow points to the 'Train database' tab.
- Two red arrows point to the 'Data table file' column of the table, with labels 'Root data table files' and 'Secondary data table file'.



Supervised classification

77

- **Step 3 : Parameters**



Target variable →

Main target value →

↓

Khiops

Data dictionary Tools Help

Data dictionary Train database Parameters Results

Target variable Gravity

Main target value Lethal

Predictors Recorders 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...



Supervised classification

78

- **Step 4 : Variable construction parameters**

Optional

Choice of construction rules

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

Default Select all Unselect all Close

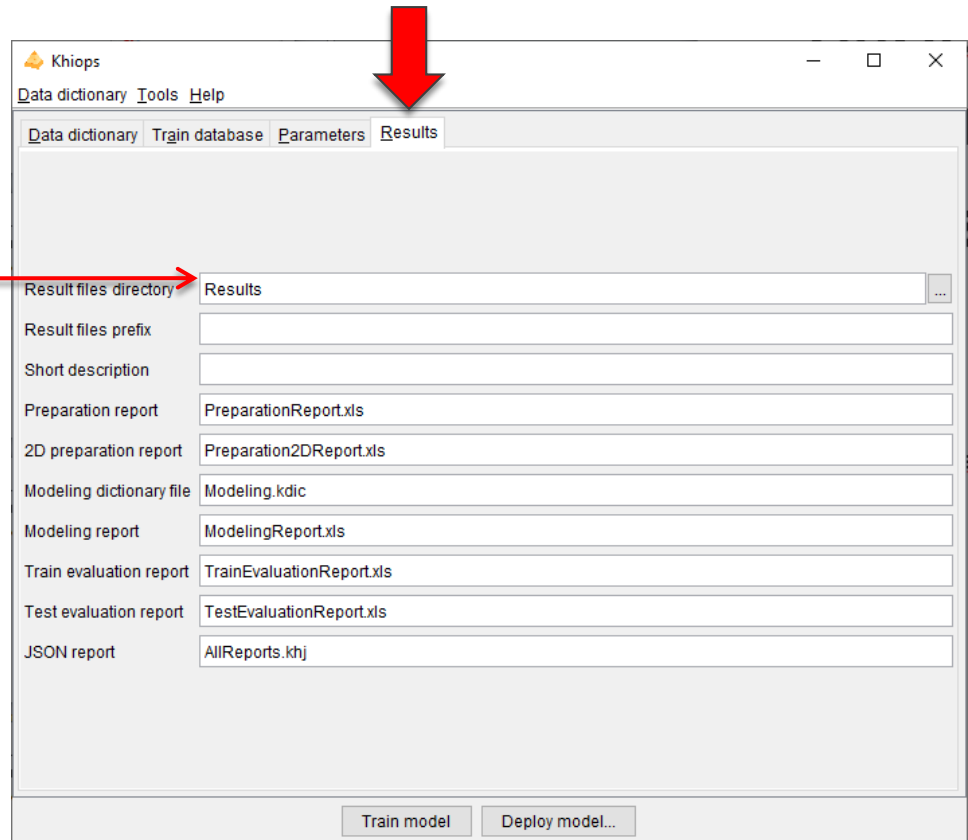


Supervised classification

79

- **Step 5 : Analysis results**

Results files directory



The screenshot shows the Khiops software interface with the 'Results' tab selected. The 'Result files directory' field is highlighted with a red arrow pointing to it from the text 'Results files directory' on the left. A large red arrow points down to the 'Results' tab in the top navigation bar. The interface includes a menu bar with 'Data dictionary', 'Tools', and 'Help'. The 'Results' tab contains several input fields for configuring report outputs:

Field Name	Value
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

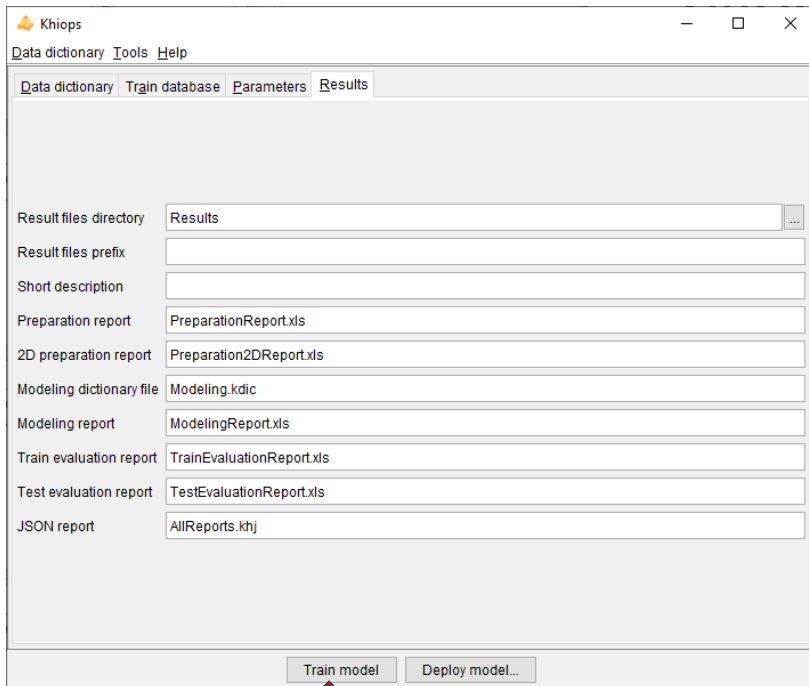
At the bottom of the window, there are two buttons: 'Train model' and 'Deploy model...'.



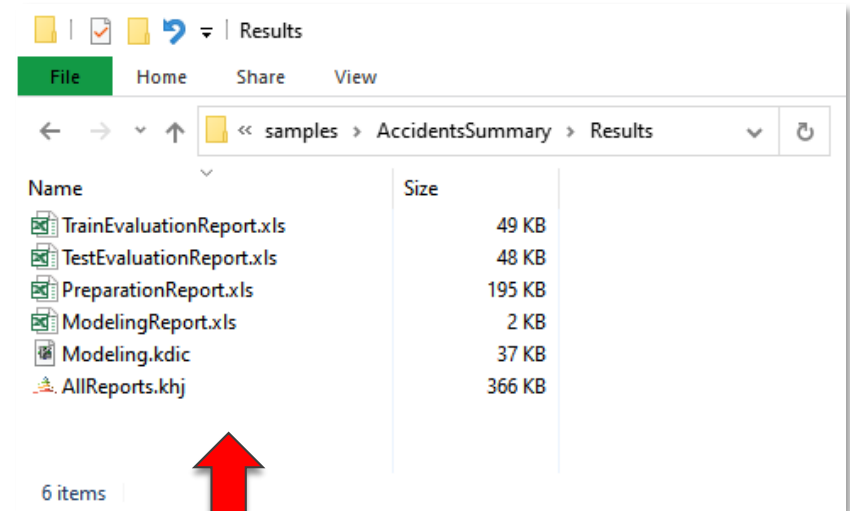
Supervised classification

80

- Step 6 : Start the analysis**



1 - Train model



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





Exploratory of classification results using KHIOPS Visualization

81

Preparation ↓

KHIOPS Visualization
File View Help Report a bug
Project Preparation Tree preparation Modeling Evaluation

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

Target variable stats Lethal NonLethal

109 Variables Level distribution

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	23	Numerical

Count(Vehicles) where FixedObstacle <> None Scale chart

Internal Coverage Coverage Coverage

Target distribution Lethal NonLethal Values Probabilities

Current interval <> +|-

Interval of Count(Vehicles... | [0,0.5]

Constructed variable name ↑

Constructed variable derivation rule ↑

Name: Count(Vehicles) where FixedObstacle <> None
Derivation rule: TableCount('Vehicles where FixedObstacle <> None')



Example of a complex multi-table database

82

French road accidents database (full version)

This is 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.

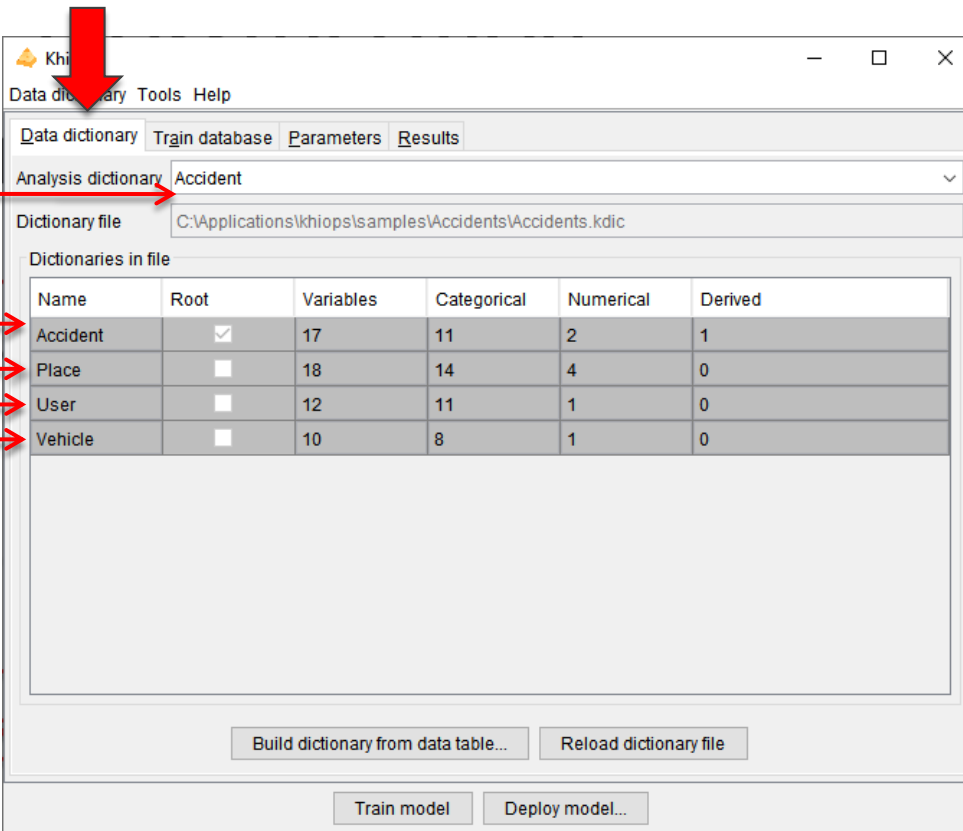
Objective: predict fatal traffic accidents (target variable: Gravity field of Accident table)



Supervised classification

83

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



Analysis dictionary

Root entity

Secondary entities

Name	Root	Variables	Categorical	Numerical	Derived
Accident	<input checked="" type="checkbox"/>	17	11	2	1
Place	<input type="checkbox"/>	18	14	4	0
User	<input type="checkbox"/>	12	11	1	0
Vehicle	<input type="checkbox"/>	10	8	1	0

Build dictionary from data table... Reload dictionary file

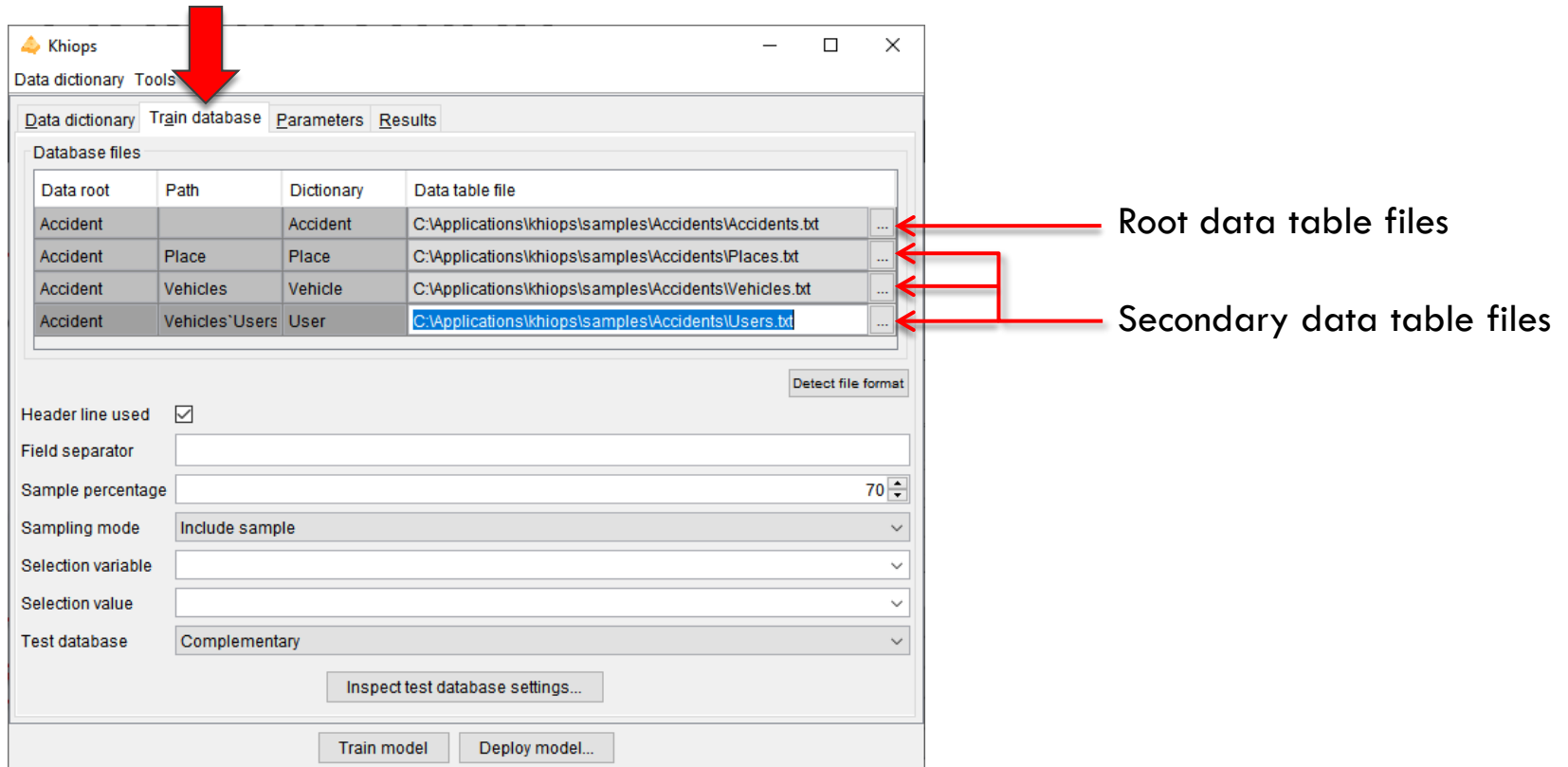
Train model Deploy model...



Supervised classification

84

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



The screenshot shows the 'Train database' tab in the Khlops software. The 'Database files' table is as follows:

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

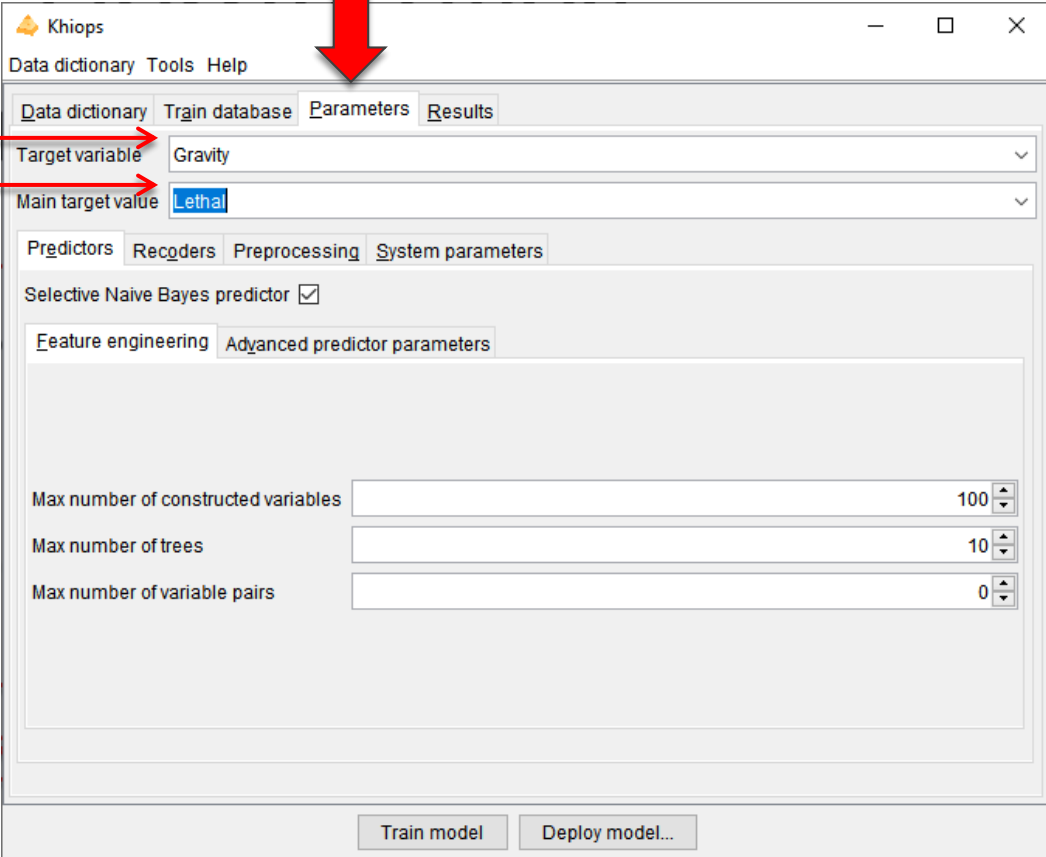
Red arrows point to the file paths in the 'Data table file' column. The label 'Root data table files' points to the first three rows, and 'Secondary data table files' points to the last row.



Supervised classification

85

- **Step 3 : Parameters**



Target variable → Gravity

Main target value → Lethal

Max number of constructed variables: 100

Max number of trees: 10

Max number of variable pairs: 0

Train model Deploy model...

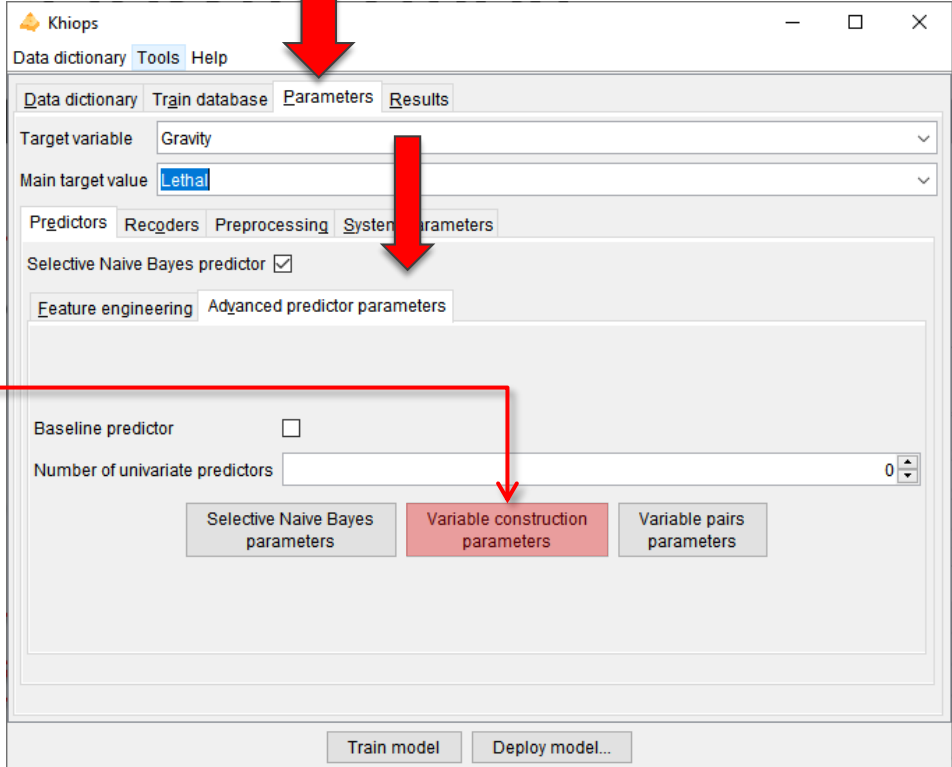


Supervised classification

86

- **Step 4 : Variable construction parameters**

Optional
Choice of construction rules



The screenshot shows the Khlops software interface with the 'Parameters' tab selected. The 'Main target value' is set to 'Lethal'. The 'Variable construction parameters' button is highlighted in red. A red arrow points to the 'Parameters' tab, another red arrow points to the 'Main target value' dropdown, and a third red arrow points to the 'Variable construction parameters' button. The 'Optional Choice of construction rules' text is connected to the 'Variable construction parameters' button by a red line.



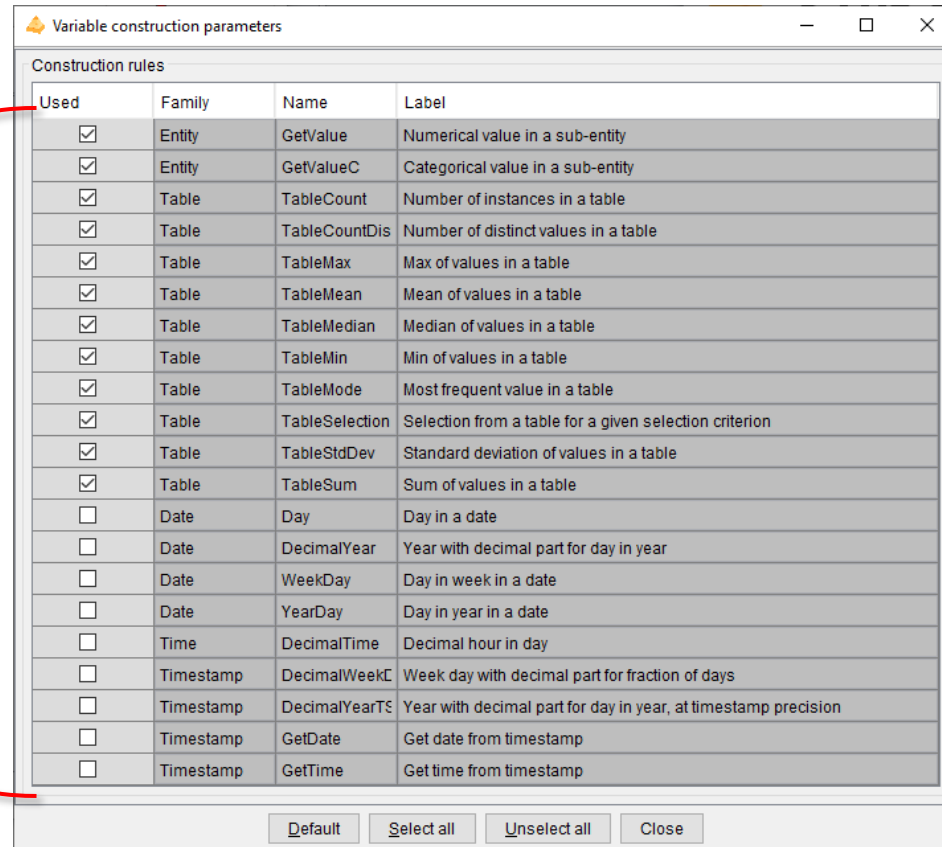
Supervised classification

87

- **Step 4 : Variable construction parameters**

Optional

Choice of construction rules



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

Default Select all Unselect all Close

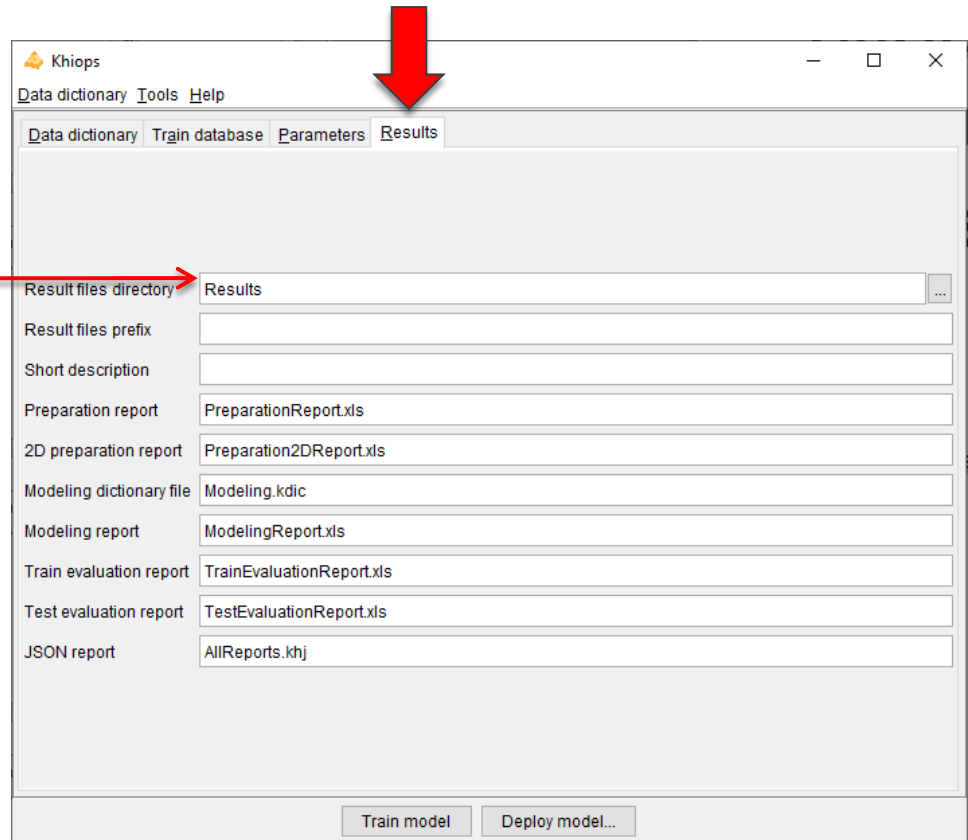


Supervised classification

88

- **Step 5 : Analysis results**

Results files directory



The screenshot shows the 'Khiops' application window with the 'Results' tab selected. The interface includes a menu bar with 'Data dictionary', 'Tools', and 'Help'. Below the menu bar are tabs for 'Data dictionary', 'Train database', 'Parameters', and 'Results'. The main area contains several input fields for configuring report outputs:

Field Name	Value
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

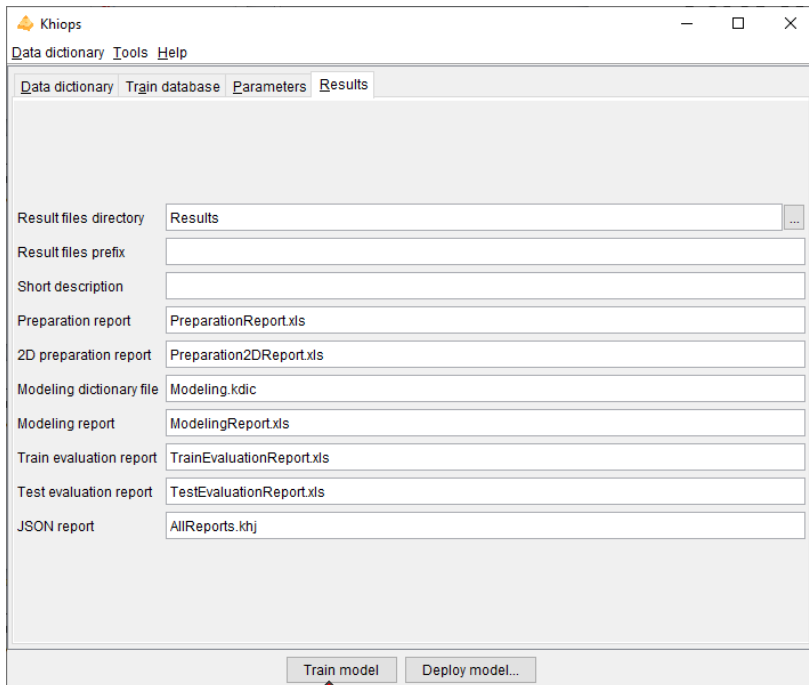
At the bottom of the window, there are two buttons: 'Train model' and 'Deploy model...'. A red arrow points to the 'Results' tab, and another red arrow points to the 'Result files directory' field.



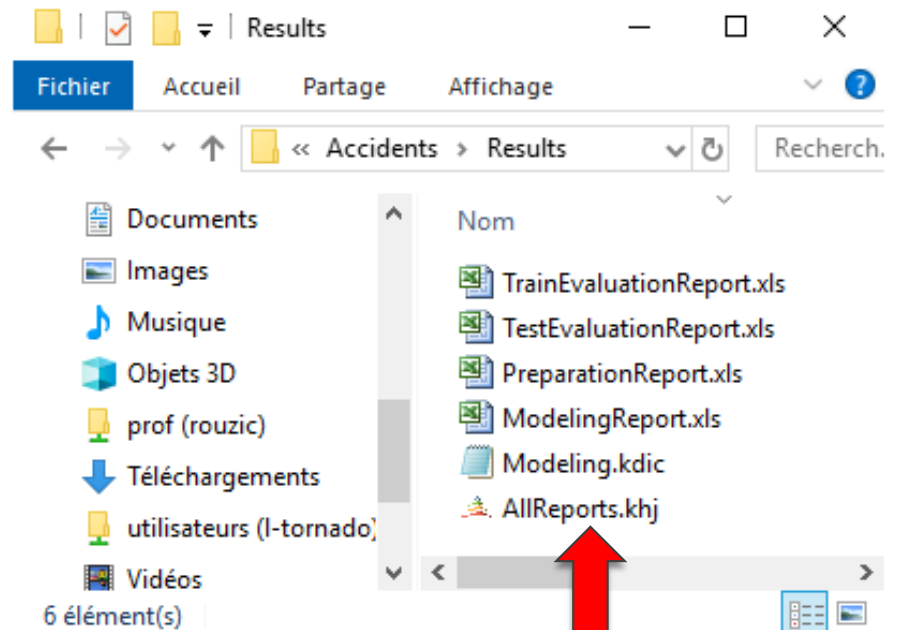
Supervised classification

89

- Step 6 : Start the analysis**



1 - Train model



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



Exploratory of classification results using Khiops Visualization

90

Preparation ↓

Khiops Visualization
File View Help Report a bug
KHIOPS Visualization

Project **Preparation** Tree preparation Modeling Evaluation

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

Target variable stats Lethal NonLethal

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 FixedOb	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 FixedOb	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

Level distribution

Mean(Vehicles.PassengerNumber) where FixedObstacle <> None

Internal Coverage

Target distribution

Current interval

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

Missing

Constructed variable name ↑

Constructed variable derivation rule ↑

Derivation rule
TableMean("Vehicles where FixedObstacle <> None", PassengerNumber)

Khiops multi-table

91

- 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

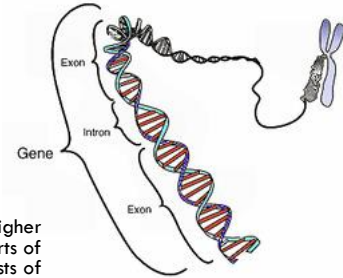


- Deploy the model on the multi-table SpliceJunction database

- 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)

SpliceJunction.txt		SpliceJunctionDNA.txt		
SampleId	Class	SampleId	Pos	Char
AGMKPNRSB-NEG-1	N	AGMKPNRSB-NEG-1	1	C
AGMORS12A-NEG-181	N	AGMKPNRSB-NEG-1	2	A
AGMORS9A-NEG-481	N	...		
AGMRSPNI-NEG-1141	N	AGMKPNRSB-NEG-1	58	A
ATRINS-ACCEPTOR-1678	IE	AGMKPNRSB-NEG-1	59	C
ATRINS-ACCEPTOR-701	IE	AGMKPNRSB-NEG-1	60	A
ATRINS-DONOR-521	EI	AGMORS12A-NEG-181	1	A
ATRINS-DONOR-905	EI	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

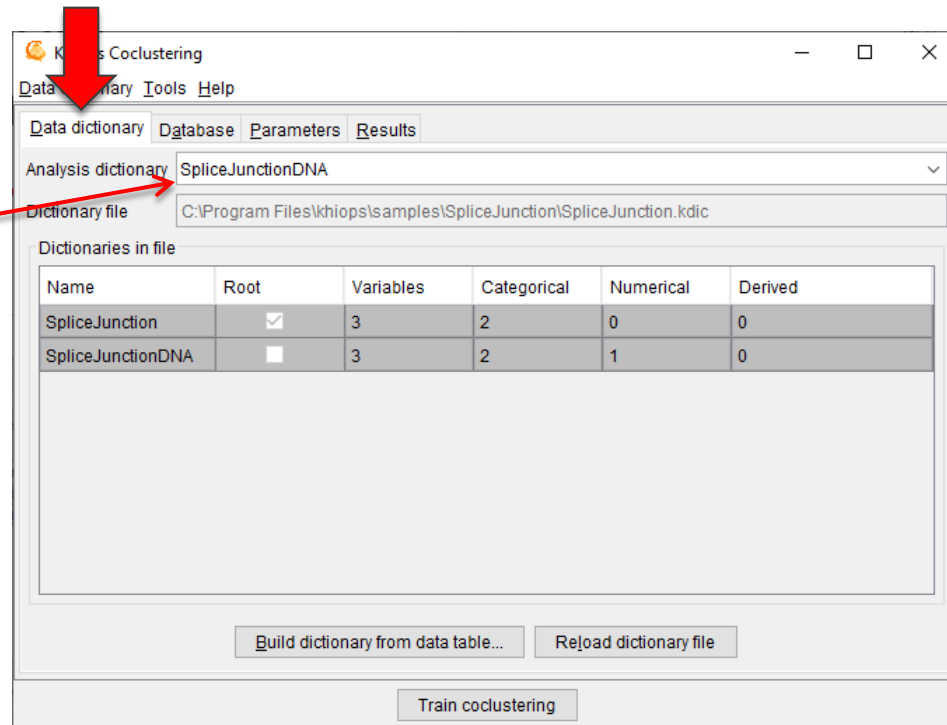


Train a triclustering model

93

- **Step 1** : **Open an existing dictionary**
(ex: sample SpliceJunction.kdic)

Analysis dictionary
(secondary entity)





Train a triclustering model

94

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

Data table file
(one single file for analysis of the secondary entity)

Khiops CoClustering

Data dictionary | **Database** | Parameters | Results

Database files

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

Detect file format

Header line used

Field separator

Sample percentage

Sampling mode

Selection variable

Selection value

Train coclustering



Train a triclustering model

95

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

Triclustering
variables

The screenshot shows the 'Khiops Coclustering' application window. The 'Parameters' tab is active, and the 'Coclustering parameters' sub-tab is selected. Under the 'Coclustering variables' section, a list of variables is displayed: 'Name', 'SampleId', 'Pos', and 'Char'. Each variable has a dropdown arrow on its right side. A red arrow points from the text 'Triclustering variables' to the 'SampleId', 'Pos', and 'Char' entries. Another red arrow points from the top of the window to the 'Parameters' tab. Below the list, there are 'Insert variable' and 'Remove variable' buttons. At the bottom, there is a 'Frequency variable' dropdown menu and a 'Train coclustering' button.



Train a triclustering model

96

- **Step 4 : Analysis results**

Result files directory

Khiops Coclustering

Data dictionary Tools Help

Data dictionary Database Parameters Results

Result files directory TriclusteringResults

Result files prefix

Short description

Coclustering report Coclustering.khc

Export JSON

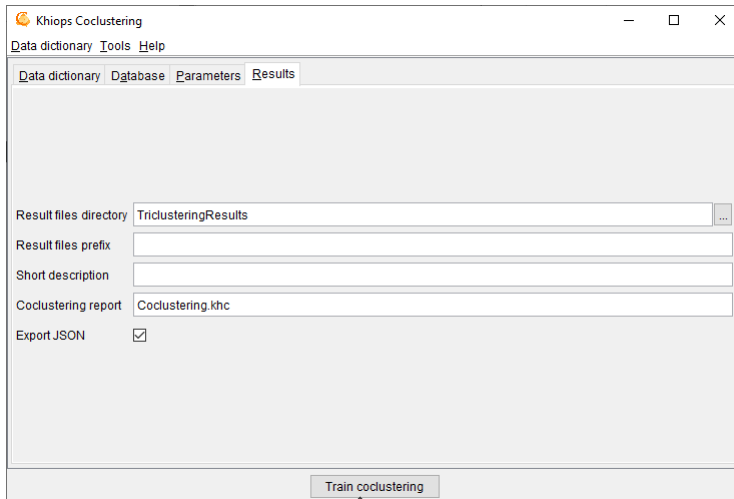
Train coclustering



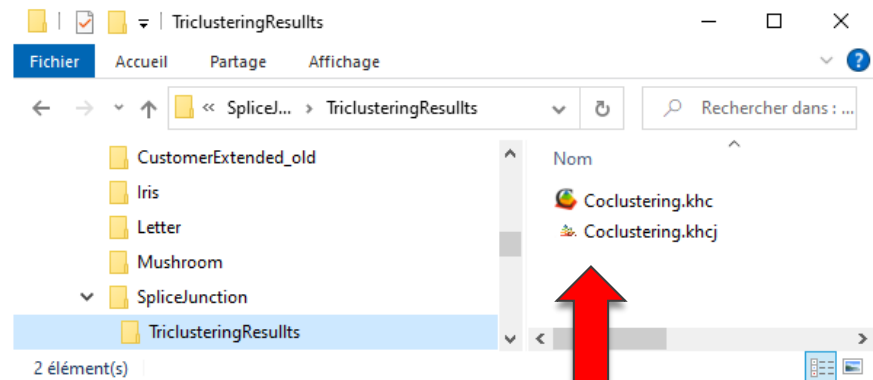
Train a triclustering model

97

- **Step 5 : Start the analysis**



1 – Start the analysis



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

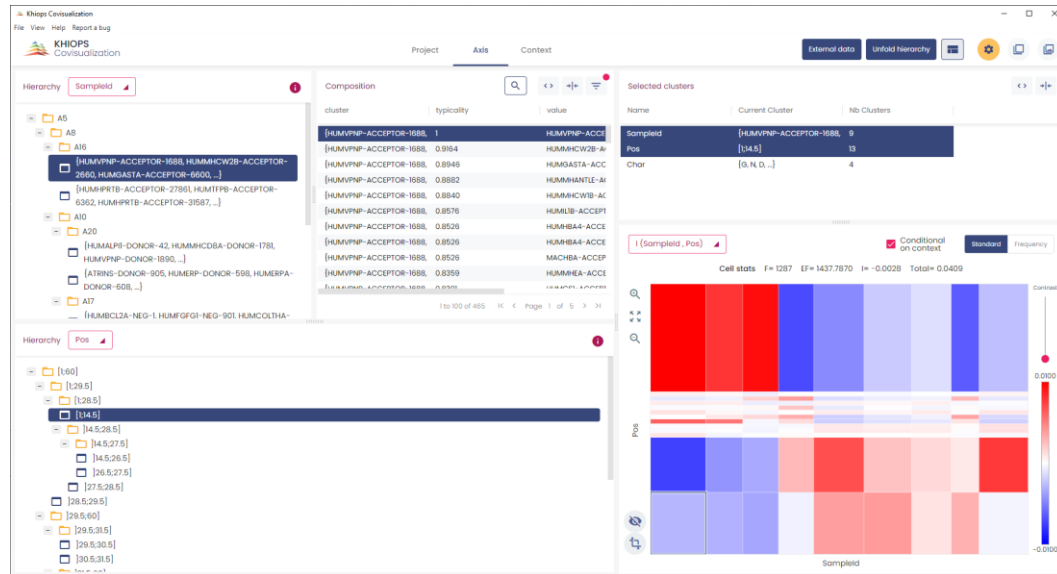
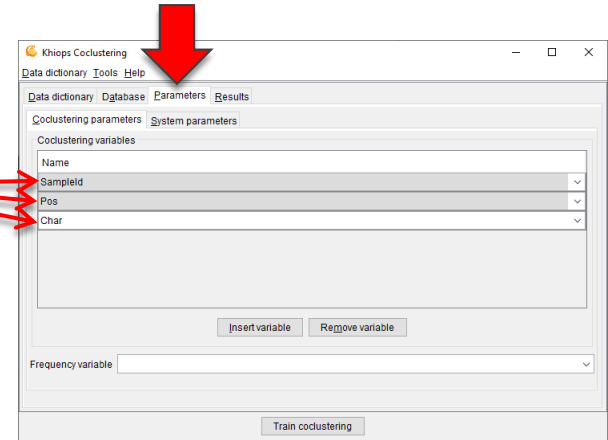




Khiops covisualisation: base SpliceJunctionDNA



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



With Khiops Covisualization

- Exploratory analysis of the results

Khiops multi-table

99

- 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

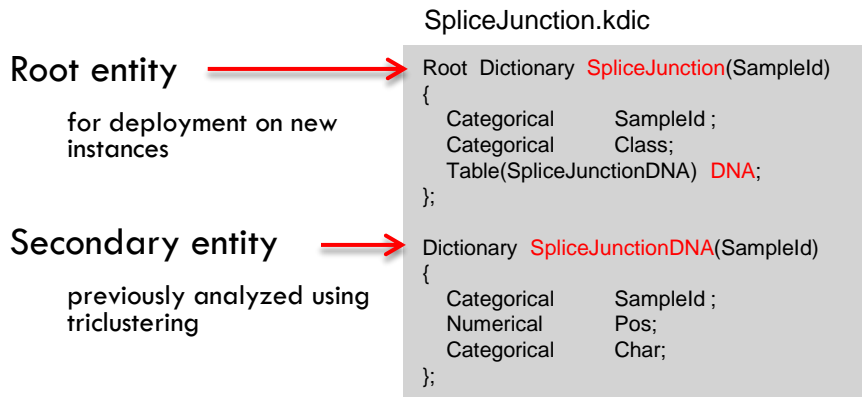
- Assign new DNA sequences to trained clusters of sequences



Prepare a deployment model

100

- **Prerequisite : a multi-table database**
 - **dictionary file**
 - **data files**
(ex: sample SpliceJunction)



SpliceJunction.txt

SampleId	Class
AGMKPNRSB-NEG-1	N
AGMORS12A-NEG-181	N
AGMORS9A-NEG-481	N
AGMRSPNI-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

101

- **Step 1 : Open an existing dictionary**
(ex: sample SpliceJunction.kdic)

Root entity
for deployment on new
instances

Secondary entity
previously analyzed using
triclustering

Name	Root	Variables	Categorical	Numerical	Derived
SpliceJunction	<input checked="" type="checkbox"/>	3	2	0	0
SpliceJunctionI	<input type="checkbox"/>	3	2	1	0

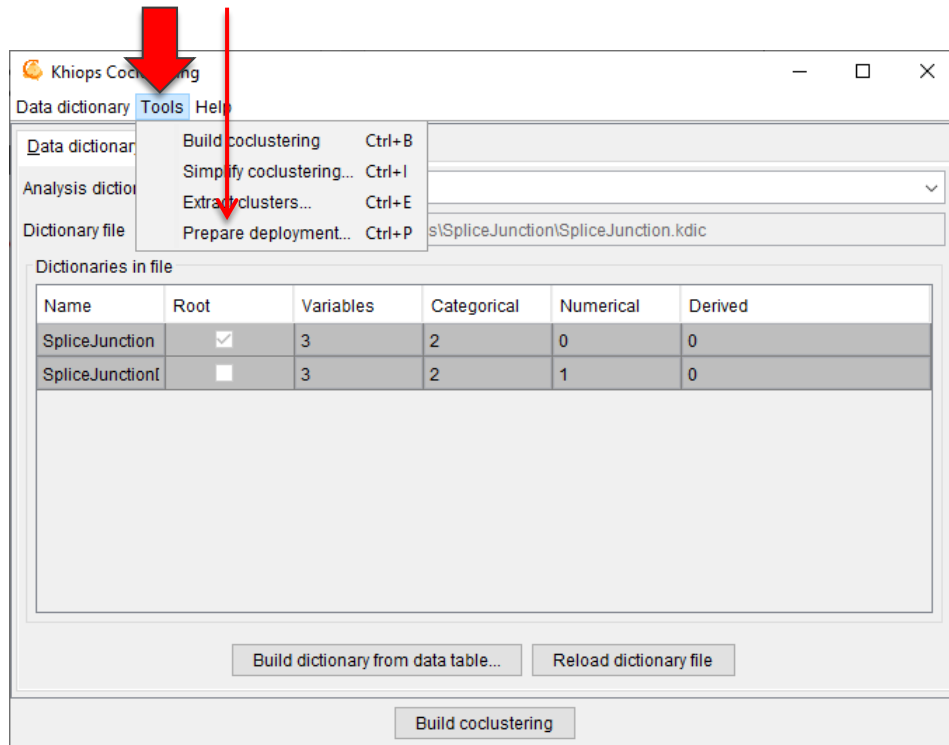
Buttons: Build dictionary from data table..., Reload dictionary file, Build coclustering



Prepare a deployment model

102

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

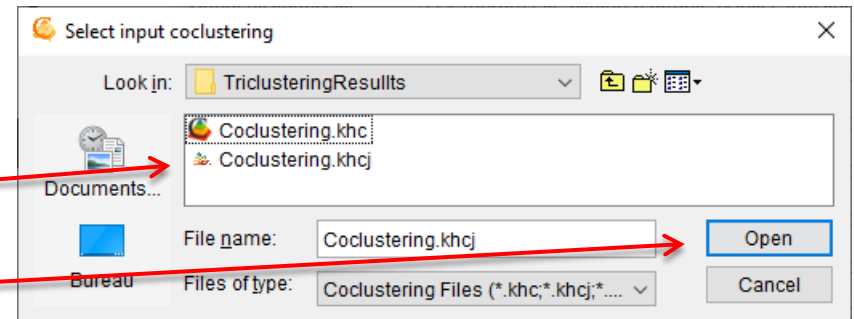
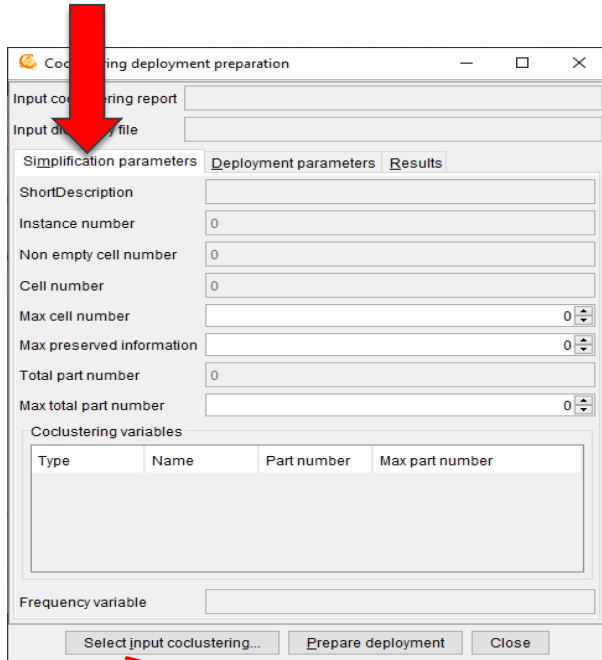




Prepare a deployment model

103

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



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



Prepare a deployment model

104

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

Simplification parameters

Coclustering deployment preparation

Input coclustering report C:\Program Files\khiops\samples\SpliceJunction\Triclusterir

Input dictionary file C:\Program Files\khiops\samples\SpliceJunction\SpliceJunc

Simplification parameters Deployment parameters Results

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

Select input coclustering... Prepare deployment Close



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;
};

```

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

106

- **Step 6** : Specify result parameters

Result files directory

Deployment dictionary file

(to deploy cluster information
on new data)

Coclustering deployment preparation

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

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

Simplification parameters | Deployment parameters | **Results**

Result files directory: Deployment

Result files prefix:

Coclustering dictionary file: Coclustering.kdic

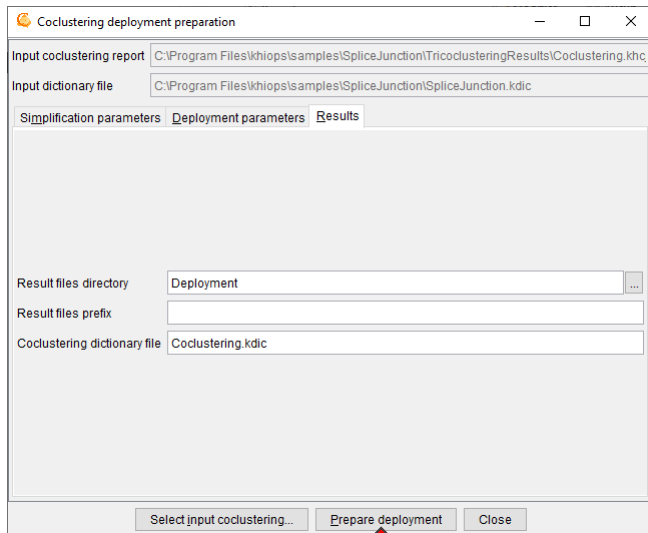
Select input coclustering... | Prepare deployment | Close



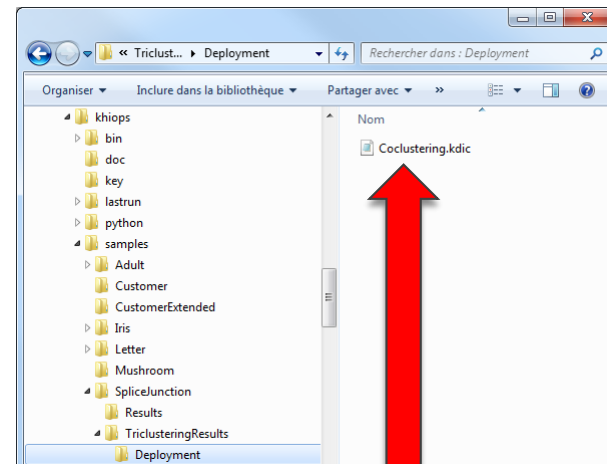
Prepare a deployment model

107

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



1 – Build the deployment dictionary



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

Khiops multi-table

108

- 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

109

- **Step 1** : **Open the deployment dictionary file with Khiops**
(ex: Samples\SpliceJunction\TriclusteringResults\Deployment\Coclustering.kdic)

Deployment dictionary

Root entity

Secondary entity

Name	Root	Variables	Categorical	Numerical	Derived
SpliceJunction	<input checked="" type="checkbox"/>	4	3	0	1
SpliceJunctionDNA	<input type="checkbox"/>	3	2	1	0



Deploy the model

110

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

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

Initial variables

Used by default

Model variables

(technical variables)

Deployment variables

Cluster index (unused by default)

Cluster label (used by default)

Dictionary

Name: SpliceJunction

Root:

Key: SampleId

Variables

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

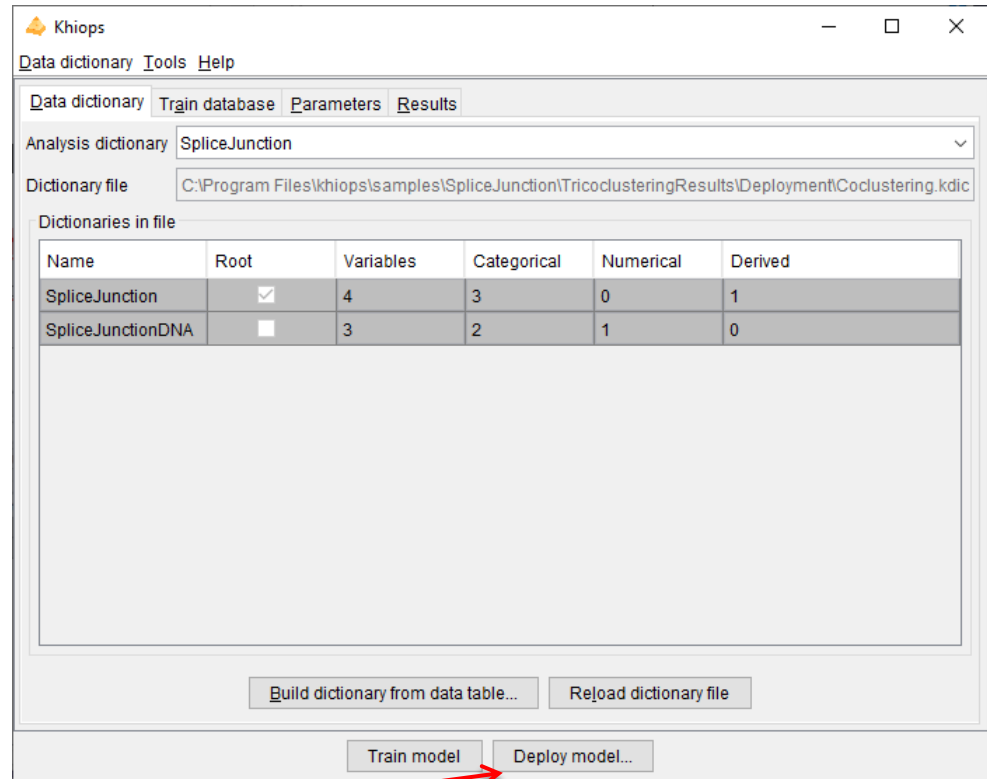
Select all Unselect all Close



Deploy the model

111

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



click on button



Deploy the model

112

- **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.bt
SpliceJunction	DNA	SpliceJunctionDNA	C:\Program Files\khiops\samples\SpliceJunction\SpliceJunctionDNA.bt

Detect file format

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\SpliceJunctionID_SpliceJunction.bt
SpliceJunction	DNA	SpliceJunctionDNA	C:\Program Files\khiops\samples\SpliceJunctionID_SpliceJunctionDNA.bt

Header line used

Field separator

Output format tabular

Deploy model Build deployed dictionary... Close

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

113



- **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