Orange
July 2024

Khiops Visualization Guide 10.2

- **Khiops**
  - Optimal data preparation based on discretization and value grouping
  - Scoring models for classification and regression
  - Correlation analysis between pairs of variables
  - Automatic variable construction for multi-table relational mining
- **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
  - Prediction models for cluster assignment
- Khiops Covisualization
  - Exploratory analysis of Khiops Coclustering results using an interactive visualization tool

This guide is about the Khiops Visualization component (version 11.0.5)

Abstract

Khiops Visualization is a visualization plug-in of the data preparation and scoring tool Khiops. Khiops Visualization allows visualizing all analysis results in an intuitive way, providing a quick and easy interpretation of the discovered patterns.

Summary

1. Presentation

Khiops Visualization allows to visualize and navigate in an intuitive way in the result reports of the Khiops software. This document presents only general features. It is strongly recommended to read the Khiops tutorial which presents all types of analysis and the associated visualization.

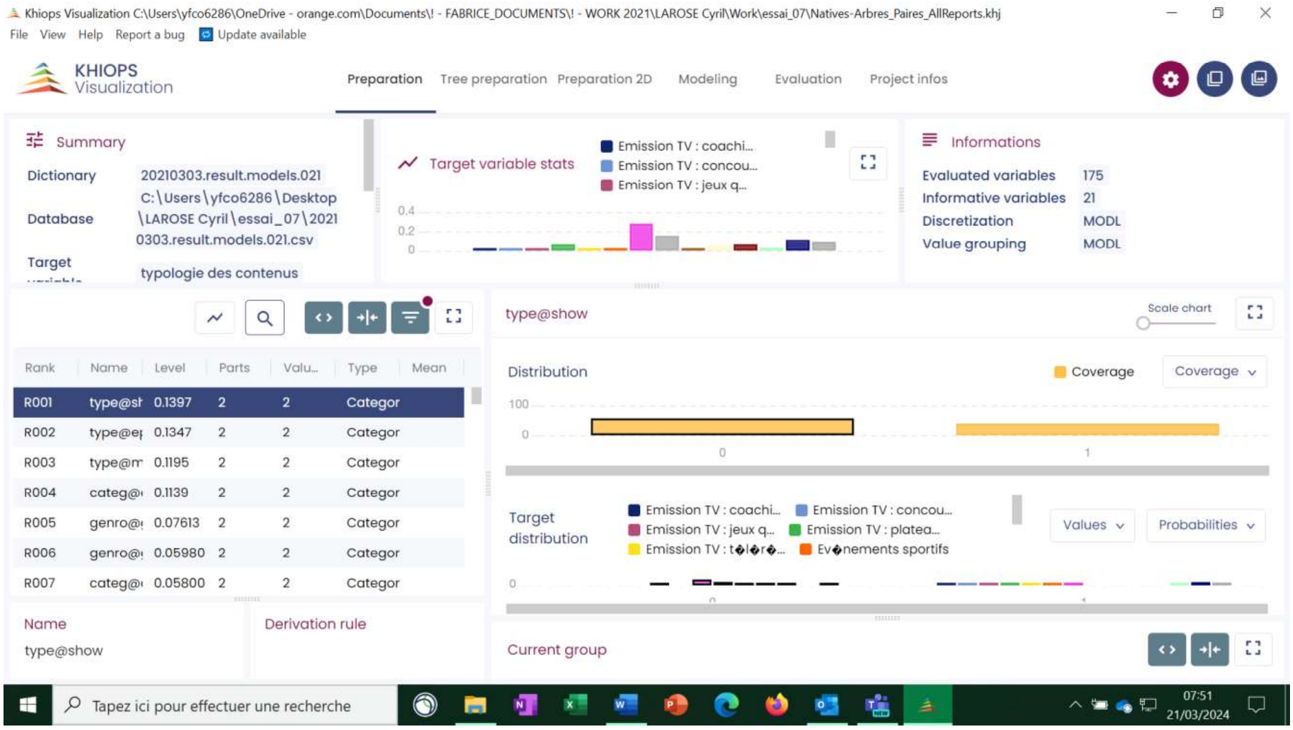
See the Khiops guide for more information on Khiops.

2. Installation

The installation of Khiops Visualization is supported by the installation of the ***Khiops*** tool available on the website https://khiops.org.

3. First step

The entry point of Khiops Visualization is the khj file generated by Khiops at the end of the analysis. A double-click on this file opens Khiops Visualization.



Summary

Dictionary	20210303.result.models.021
Database	C:\Users\yfco6286\Desktop\LAROSE Cyril\essai_07\20210303.result.models.021.csv
Target variable	typologie des contenus

Informations

Evaluated variables	175
Informative variables	21
Discretization	MODL
Value grouping	MODL

Rank	Name	Level	Parts	Valu...	Type	Mean
R001	type@sł	0.1397	2	2	Categor	
R002	type@e|	0.1347	2	2	Categor	
R003	type@m	0.1195	2	2	Categor	
R004	categ@	0.1139	2	2	Categor	
R005	genro@	0.07613	2	2	Categor	
R006	genro@	0.05980	2	2	Categor	
R007	categ@	0.05800	2	2	Categor	

Name	Derivation rule
type@show	

Khiops Visualization is composed of several tabbed panes.

Depending on the analysis type, the panes and their contents are not the same. Khiops Visualization can be composed with 5 panels:

- Preparation: displays the Preparation report.
- Tree preparation : displays the preparation report for tree variables
- Preparation 2D: displays the 2D preparation report.
- Modeling: displays the modeling report
- Evaluation: displays on one panel the test, train and evaluation reports.
- Project Infos : displays the report file and database locations plus some short comments on the analysis

The description of all these reports can be found in section 3.3. of Khiops Guide.

As supervised and non-supervised (hereafter descriptive) analysis have different goals and produce different kinds of results, the visualization interface and the content of the tabs are slightly different for both analysis. This is described in the following sections

4. Descriptive analysis visualization

In case of non-supervised analysis, the report is organized in the following tabs :

- Preparation
- Preparation 2D (only in case of constructed bivariate variables)
- Project information

For the figures below, unless otherwise specified, we used the Adult database.

4.1. Preparation tab

The general layout of this tab is as follows (see the two figures below) : Panel A shows some information related to the data and the analysis. Panel B stores the list of the variables as a table together with some descriptive statistics information.

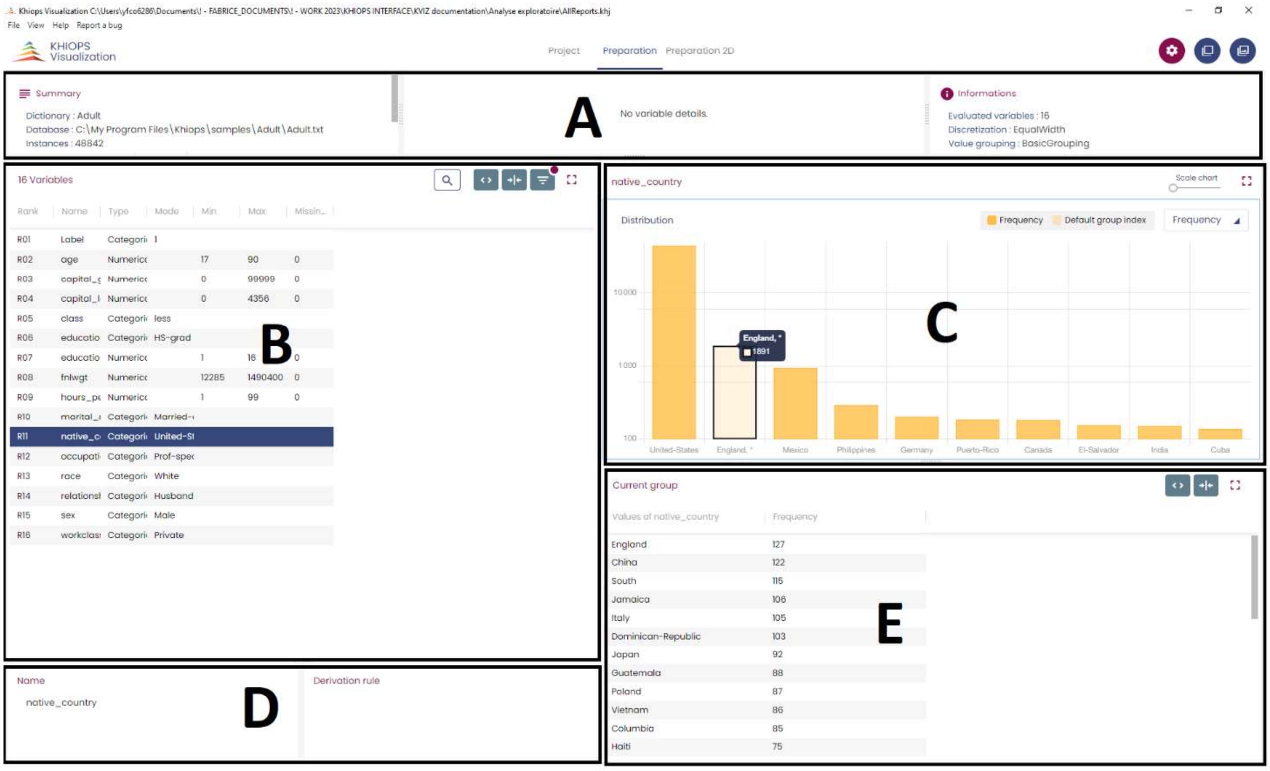
Clicking on a line of the table in panel B displays the histogram of the corresponding variable in panel C and information about the variable (Name and derivation rule (if any)) in panel D. Clicking on a bin of the histogram displays information about this bin in panel E.

Categorical and numerical variables have different histograms.

For categorical variables, the number of the bins is automatically set to its optimal value from an MDL approach.

In the case of a variable with many categories, a specific bin ("Default group index") may gather many modalities and is shown with a different color. This specific bin is given the name of its most populated modality and clicking on the bin makes its content is available in panel E. As it may gather many small modalities, it may appear as the dominant bin of the distribution and the name might be misleading, therefore the different color.

The Coverage/Frequency toggle allows to display either the coverage (in %, linear scale) or the populations (in log scale).


Khiops Visualization C:\Users\yfco6286\Documents\! - FABRICE_DOCUMENTS\! - WORK 2023\KHIOPS INTERFACE\KVIZ documentation\Analyse exploratoire\AllReports.khj
File View Help Report a bug
KHIOPS Visualization
Project
Preparation
Preparation 2D
A
B
C
D
E

Summary

Dictionary : Adult
Database : C:\My Program Files\Khiops\samples\Adult\Adult.txt
Instances : 48842

No variable details.

Informations

Evaluated variables : 16
Discretization : EqualWidth
Value grouping : BasicGrouping

16 Variables

Rank	Name	Type	Mode	Min	Max	Missin...
R01	Label	Categori	1			
R02	age	Numericc		17	90	0
R03	capital_ç	Numericc		0	99999	0
R04	capital_l	Numericc		0	4356	0
R05	class	Categori	less			
R06	educatio	Categori	HS-grad			
R07	educatio	Numericc		1	16	0
R08	fnlwgt	Numericc		12285	1490400	0
R09	hours_pe	Numericc		1	99	0
R10	marital_	Categori	Married-(			
R11	native_c	Categori	United-St			
R12	occupati	Categori	Prof-spec			
R13	race	Categori	White			
R14	relationsl	Categori	Husband			
R15	sex	Categori	Male			
R16	workclas	Categori	Private			

native_country

Current group

Values of native_country	Frequency
England	127
China	122
South	115
Jamaica	106
Italy	105
Dominican-Republic	103
Japan	92
Guatemala	88
Poland	87
Vietnam	86
Columbia	85
Haiti	75

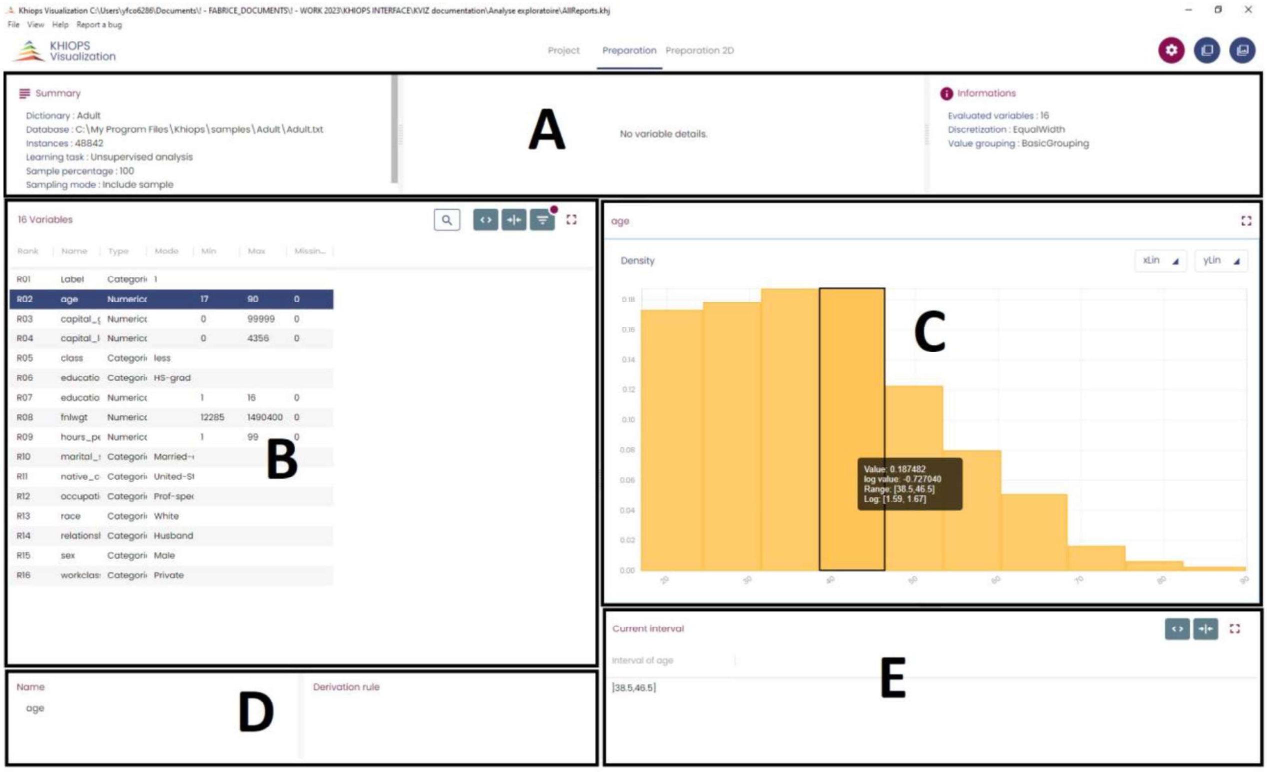
Name

native_country

Derivation rule

The figure above shows the histogram of the variable "native_country" which has 41 modalities. The default group is selected, and its content is displayed in panel E.

For numerical variables, the bounds of the bins is also automatically set to its optimal value from an MDL approach.



Khiops Visualization C:\Users\yfco6286\Documents\! - FABRICE_DOCUMENTS\! - WORK 2023\KHIOPS INTERFACE\KVIZ documentation\Analyse exploratoire\AllReports.khj

Summary

Dictionary : Adult
Database : C:\My Program Files\Khiops\samples\Adult\Adult.txt
Instances : 48842
Learning task : Unsupervised analysis
Sample percentage : 100
Sampling mode : Include sample

No variable details.

Informations

Evaluated variables : 16
Discretization : EqualWidth
Value grouping : BasicGrouping

16 Variables

Rank	Name	Type	Mode	Min	Max	Missin...
R01	Label	Categori	1			
R02	age	Numericc		17	90	0
R03	capital_(	Numericc		0	99999	0
R04	capital_l	Numericc		0	4356	0
R05	class	Categori	less			
R06	educatio	Categori	HS-grad			
R07	educatio	Numericc		1	16	0
R08	fnlwgt	Numericc		12285	1490400	0
R09	hours_pe	Numericc		1	99	0
R10	marital_s	Categori	Married-c			
R11	native_c	Categori	United-St			
R12	occupati	Categori	Prof-spec			
R13	race	Categori	White			
R14	relationsl	Categori	Husband			
R15	sex	Categori	Male			
R16	workclas:	Categori	Private			

Value: 0.187482
log value: -0.727040
Range: [38.5,46.5]
Log: [1.59, 1.67]

Name	Derivation rule
age	

Current interval

Interval of age
]38.5,46.5]

Different representations of the histogram are available. The y-axis shows the density and can be plotted in linear or log scale.

The x-axis can also be plotted in linear or log scale ; for the log scale representation, the variable range is split in (up to) three areas :

- the bins in the negative range, excluding any bin containing zero (even as a bound)
- the bins containing zero (one or two bins, if zero is a bound)
- the bins in the positive range, excluding any bin containing zero (even as a bound)

The log scale for the negative range is log10(-x), running to infinity toward the left.

The log scale foe the positive range is log10(x), running to infinity towards the right.

The bins containing zero are given an arbitrary width, set to 1/10 of the figure.

This allows to plot informative histograms of challenging heavy-tailed distributions such as Cauchy distribution. The figure below shows the histogram of this distribution in log-log scale.

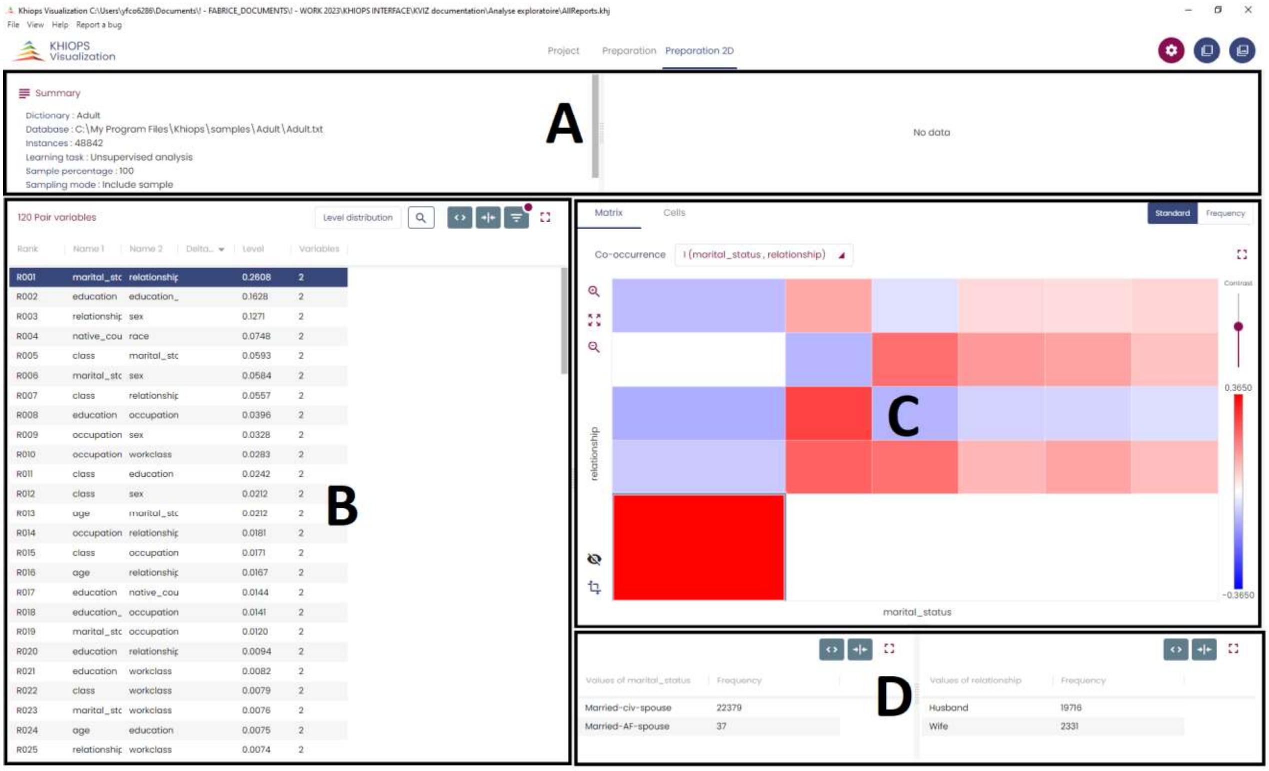
<<insert Numerical Cauchy histogram figure here>>

4.2. Preparation 2D

The general layout is similar to the Preparation tab, except for panel C which shows the bivariate discretization of the pair selected in panel B.

This discretization is the cartesian product of the parts of each variable. By default, the information displayed on the subtab "Matrix" is the mutual information between both groupings, allowing to analyze the correlations between the variables of the pair. Other information such as the frequencies in each cell can be selected from the menu.

The subtab "Cells" displays the information about the cell statistics as a table.


Khiops Visualization C:\Users\yfco6286\Documents\! - FABRICE_DOCUMENTS\! - WORK 2023\KHIOPS INTERFACE\KVIZ documentation\Analyse exploratoire\AllReports.khj
File View Help Report a bug
KHIOPS Visualization
Project
Preparation
Preparation 2D
A
No data
120 Pair variables
Level distribution
B
Matrix
Cells
Standard
Frequency
Co-occurrence
1 (marital_status , relationship)
C
relationship
marital_status
Contrast
0.3650
-0.3650
D

Summary

Dictionary : Adult
Database : C:\My Program Files\Khiops\samples\Adult\Adult.txt
Instances : 48842
Learning task : Unsupervised analysis
Sample percentage : 100
Sampling mode : Include sample

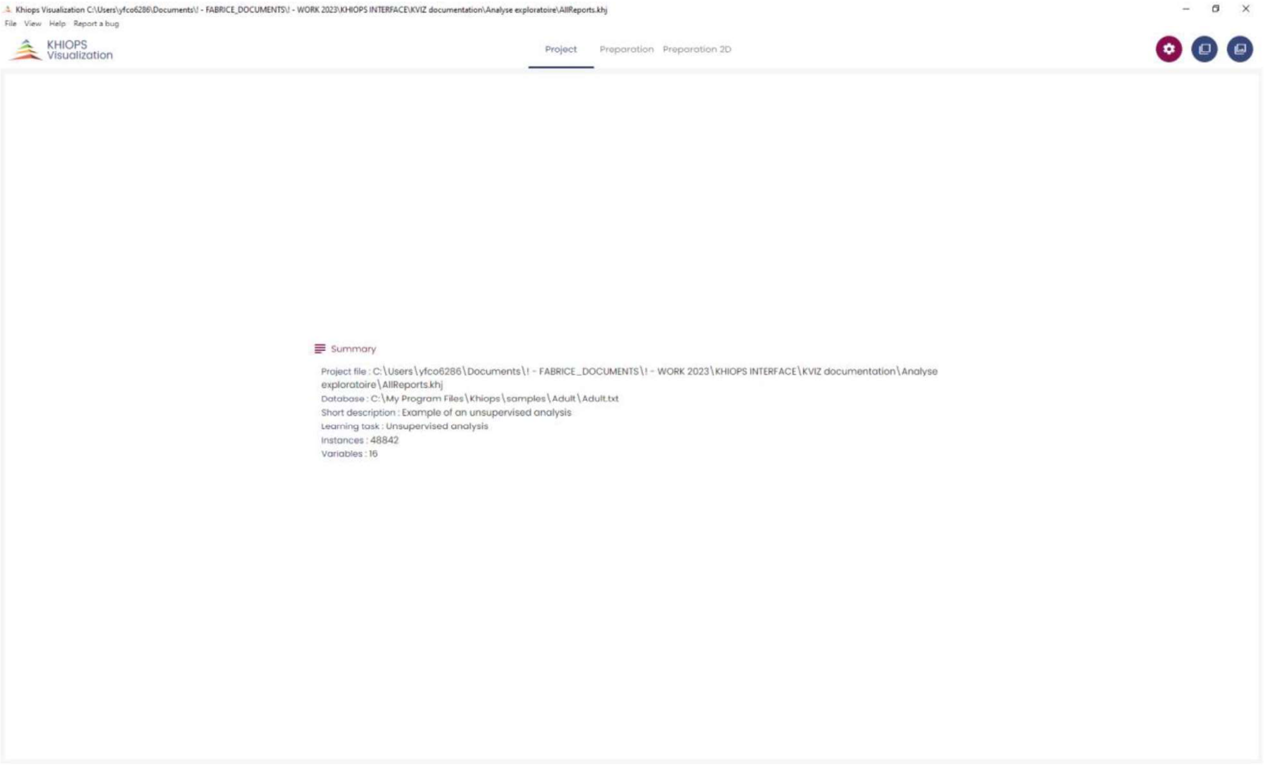
Rank	Name 1	Name 2	Delta...	Level	Variables
R001	marital_stc	relationshir		0.2608	2
R002	education	education_		0.1628	2
R003	relationshir	sex		0.1271	2
R004	native_cou	race		0.0748	2
R005	class	marital_stc		0.0593	2
R006	marital_stc	sex		0.0584	2
R007	class	relationshir		0.0557	2
R008	education	occupation		0.0396	2
R009	occupation	sex		0.0328	2
R010	occupation	workclass		0.0283	2
R011	class	education		0.0242	2
R012	class	sex		0.0212	2
R013	age	marital_stc		0.0212	2
R014	occupation	relationshir		0.0181	2
R015	class	occupation		0.0171	2
R016	age	relationshir		0.0167	2
R017	education	native_cou		0.0144	2
R018	education_	occupation		0.0141	2
R019	marital_stc	occupation		0.0120	2
R020	education	relationshir		0.0094	2
R021	education	workclass		0.0082	2
R022	class	workclass		0.0079	2
R023	marital_stc	workclass		0.0076	2
R024	age	education		0.0075	2
R025	relationshir	workclass		0.0074	2

Values of marital_status	Frequency
Married-civ-spouse	22379
Married-AF-spouse	37

Values of relationship	Frequency
Husband	19716
Wife	2331

4.3. Project information

This tab shows information related to the project, in particular, the content of the "Short description" field.


Khiops Visualization C:\Users\yfco6286\Documents\! - FABRICE_DOCUMENTS\! - WORK 2023\KHIOPS INTERFACE\KVIZ documentation\Analyse exploratoire\AllReports.khj
File View Help Report a bug
KHIOPS Visualization
Project
Preparation
Preparation 2D

Summary

Project file : C:\Users\yfco6286\Documents\! - FABRICE_DOCUMENTS\! - WORK 2023\KHIOPS INTERFACE\KVIZ documentation\Analyse exploratoire\AllReports.khj
Database : C:\My Program Files\Khiops\samples\Adult\Adult.txt
Short description : Example of an unsupervised analysis
Learning task : Unsupervised analysis
Instances : 48842
Variables : 16

5. Supervised analysis visualization

There are two different reports depending on the categorical or numerical nature of the target variable.

5.1. Classification

The target variable is categorical.

The report is organized in the following tabs :
- Preparation
- Tree Preparation (only in case of constructed decision tree variables)
- Preparation 2D (only in case of constructed bivariate variables)
- Modeling
- Evaluation
- Project information

For the figures below, we used the Adult database, with "education" as a target variable (excluding "education num" from the predictors) and randomly selecting 70% of the data for the train set and the remaining 30% for the test set.

5.1.1. Preparation

Panel A shows some information related to the data and the analysis (left) and the histogram of the target modalities (right).

Panel B stores the list of the variables as a table together with some descriptive statistics information.

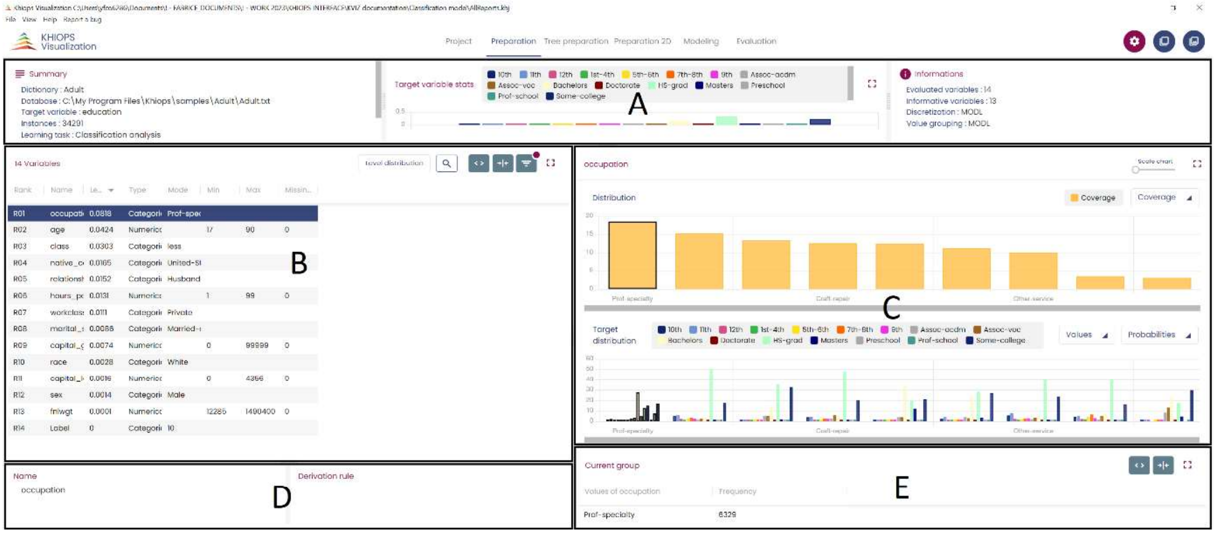
Clicking on a line of the table in panel B displays the information about the parts obtained from the optimal supervised grouping the selected variable selected in panel C and information about the variable (Name and derivation rule(if any)) in panel D.

Panel C shows the histogram of the group populations (top) and for each group, the distribution of the target modalities conditional to the group (bottom).

Group population histogram can be displayed in linear or log scale from the menu above the figure.

For the target modalities conditional to the group histogram, displayed target modalities can be selected from the "Values" menu and histogram can be displayed as probabilities or lift ("Probabilities/Lift" toggle).

Clicking on a bin of the population histogram or on the corresponding target conditional histogram displays information about the selected group in panel E.



Summary
Dictionary : Adult
Database : C:\My Program Files\Khiops\samples\Adult\Adult.txt
Target variable : education
Instances : 34291
Learning task : Classification analysis

Informations
Evaluated variables : 14
Informative variables : 13
Discretization : MODL
Value grouping : MODL

14 Variables

Rank	Name	Le...	Type	Mode	Min	Max	Missin...
R01	occupati	0.0818	Categori	Prof-spec			
R02	age	0.0424	Numeric		17	90	0
R03	class	0.0303	Categori	less			
R04	native_c	0.0165	Categori	United-St			
R05	relationsh	0.0152	Categori	Husband			
R06	hours_pe	0.0131	Numeric		1	99	0
R07	workclas	0.0111	Categori	Private			
R08	marital_s	0.0088	Categori	Married-c			
R09	capital_g	0.0074	Numeric		0	99999	0
R10	race	0.0028	Categori	White			
R11	capital_l	0.0016	Numeric		0	4356	0
R12	sex	0.0014	Categori	Male			
R13	fnlwgt	0.0001	Numeric		12285	1490400	0
R14	Label	0	Categori	10			

Name: occupation
Derivation rule

Current group

Values of occupation	Frequency
Prof-specialty	6329

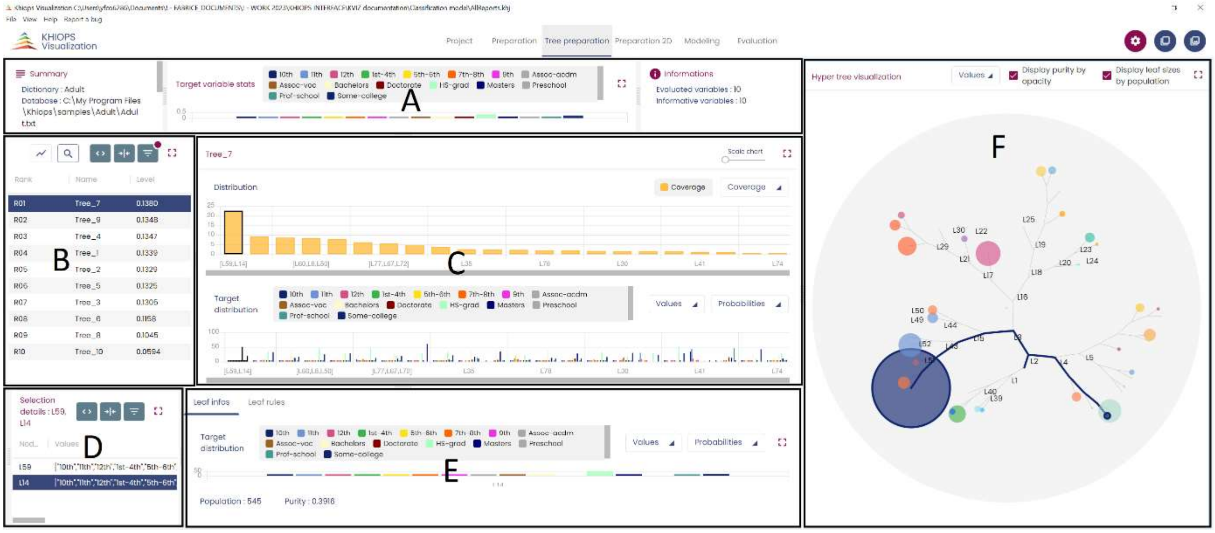
5.1.2. Tree Preparation

Decision trees are treated as categorical variables with as many modalities as terminal leaves.

The general layout of panels A, B, C and the information they display are similar to the Preparation tab.

Panel D shows the content of a selected group of leaves as a table of the leaves in the group and their target modalities statistics. Clicking on a leaf displays information about this leaf in panel E : subtab "Leaf infos" shows the histogram of the target modalities conditional to this leaf, together with the leaf population and leaf purity ; subtab "Leaf rules" gives the list of tests leading to this leaf, starting from the root node.

The left panel (panel F) is a hypertree visualization of the selected decision tree. Only terminal leaves are clickable ; clicking on a leaf dynamically points to the relevant leaf and group of leaves information in the other panels. The population and purity of the leaves can be visualized by their size and color density. This can be toggled on and off and the "values" menu allows to select some of the target modalities only.



Summary
Dictionary : Adult
Database : C:\My Program Files\Khiops\samples\Adult\Adult.txt

Informations
Evaluated variables : 10
Informative variables : 10

Rank	Name	Level
R01	Tree_7	0.1380
R02	Tree_9	0.1348
R03	Tree_4	0.1347
R04	Tree_1	0.1339
R05	Tree_2	0.1329
R06	Tree_5	0.1325
R07	Tree_3	0.1305
R08	Tree_6	0.1158
R09	Tree_8	0.1045
R10	Tree_10	0.0594

Selection details : L59, L14

Nod...	Values
L59	["10th","11th","12th","1st-4th","5th-6th"
L14	["10th","11th","12th","1st-4th","5th-6th"

Population : 545 Purity : 0.3916

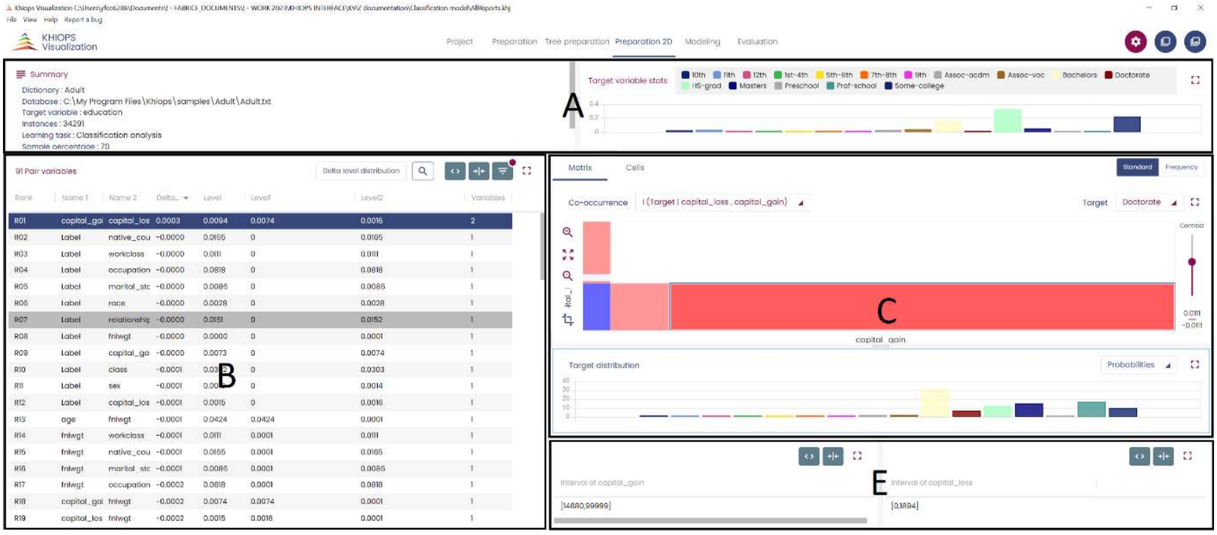
5.1.3. Preparation 2D

The general layout is similar to the Preparation tab, except for panel C which shows the bivariate supervised discretization of the pair selected in panel B.

This discretization is the cartesian product of the parts for each variable. By default, the information displayed on the subtab "Matrix" is the mutual information between a cell of the cartesian product and a modality of the target variable (selected from the "Target" menu), allowing to analyze the information brought by the variables as a pair to the target variable. Other information such as the frequencies in each cell can be selected from the menu.

The subtab "Cells" displays the information about the cell statistics as a table.

Clicking on a cell displays the target modalities histogram conditional to this cell (bottom of panel C) and the information about the parts of the selected cell in panel E.


Khiops Visualization C:\Users\yfco6286\Documents\! - FABRICE_DOCUMENTS\! - WORK 2023\KHIOPS INTERFACE\KVIZ documentation\Classification model\AllReports.khj
File View Help Report a bug
KHIOPS Visualization
Project
Preparation
Tree preparation
Preparation 2D
Modeling
Evaluation
A
B
C
E
Target variable stats
10th 11th 12th 1st-4th 5th-6th 7th-8th 9th Assoc-acdm Assoc-voc Bachelors Doctorate HS-grad Masters Preschool Prof-school Some-college
Matrix
Cells
Standard
Frequency
Co-occurrence
I (Target | capital_loss , capital_gain)
Target
Doctorate
Contrast
capital_gain
Target distribution
Probabilities

Summary
Dictionary : Adult
Database : C:\My Program Files\Khiops\samples\Adult\Adult.txt
Target variable : education
Instances : 34291
Learning task : Classification analysis
Sample percentage : 70

91 Pair variables — Delta level distribution

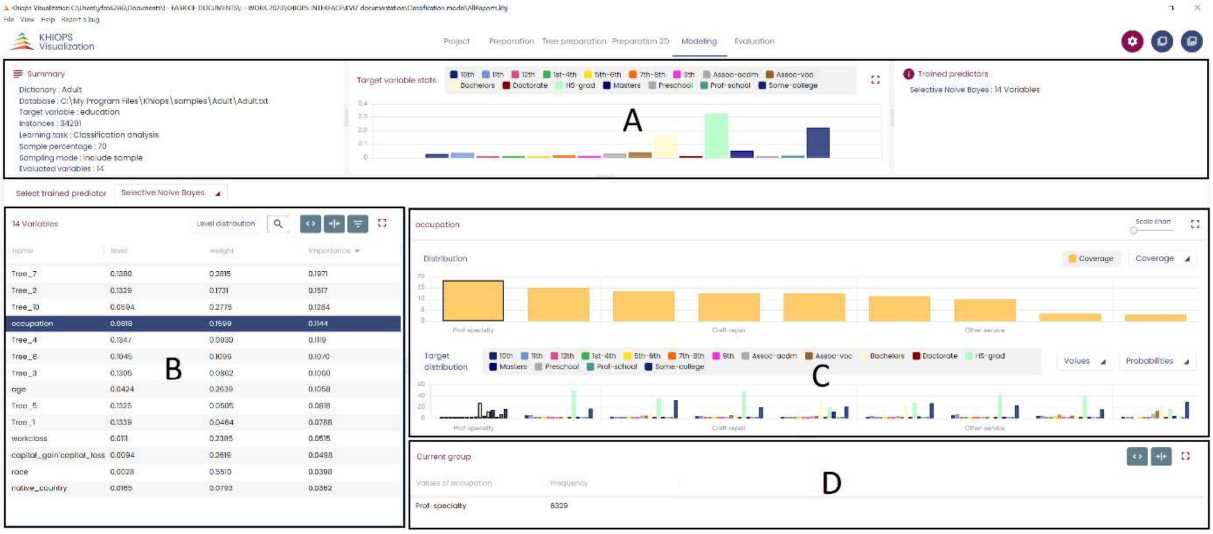
Rank	Name 1	Name 2	Delta...	Level	Level1	Level2	Variables
R01	capital_gai	capital_los	0.0003	0.0094	0.0074	0.0016	2
R02	Label	native_cou	-0.0000	0.0165	0	0.0165	1
R03	Label	workclass	-0.0000	0.0111	0	0.0111	1
R04	Label	occupation	-0.0000	0.0818	0	0.0818	1
R05	Label	marital_sta	-0.0000	0.0086	0	0.0086	1
R06	Label	race	-0.0000	0.0028	0	0.0028	1
R07	Label	relationshi	-0.0000	0.0151	0	0.0152	1
R08	Label	fnlwgt	-0.0000	0.0000	0	0.0001	1
R09	Label	capital_ga	-0.0000	0.0073	0	0.0074	1
R10	Label	class	-0.0001	[illegible]	0	0.0303	1
R11	Label	sex	-0.0001	[illegible]	0	0.0014	1
R12	Label	capital_los	-0.0001	0.0015	0	0.0016	1
R13	age	fnlwgt	-0.0001	0.0424	0.0424	0.0001	1
R14	fnlwgt	workclass	-0.0001	0.0111	0.0001	0.0111	1
R15	fnlwgt	native_cou	-0.0001	0.0165	0.0001	0.0165	1
R16	fnlwgt	marital_sta	-0.0001	0.0086	0.0001	0.0086	1
R17	fnlwgt	occupation	-0.0002	0.0818	0.0001	0.0818	1
R18	capital_gai	fnlwgt	-0.0002	0.0074	0.0074	0.0001	1
R19	capital_los	fnlwgt	-0.0002	0.0015	0.0016	0.0001	1

Interval of capital_gain:]14680,99999]

Interval of capital_loss:]0,1894]

5.1.4. Modeling

The general layout is similar to the Preparation tab. Panel B shows information related to those variables used in the model only. The table shows for each variable its level, weight (in the weighted naive Bayes model) and its importance (computed as the geometric mean of the level and the weight).


Khiops Visualization C:\Users\yfco6286\Documents\! - FABRICE_DOCUMENTS\! - WORK 2023\KHIOPS INTERFACE\KVIZ documentation\Classification model\AllReports.khj
File View Help Report a bug
KHIOPS Visualization
Project
Preparation
Tree preparation
Preparation 2D
Modeling
Evaluation
A
B
C
D
Target variable stats
10th 11th 12th 1st-4th 5th-6th 7th-8th 9th Assoc-acdm Assoc-voc Bachelors Doctorate HS-grad Masters Preschool Prof-school Some-college
Scale chart
Distribution
Coverage
Target distribution
Values
Probabilities
Prof-specialty
Craft-repair
Other-service

Summary
Dictionary : Adult
Database : C:\My Program Files\Khiops\samples\Adult\Adult.txt
Target variable : education
Instances : 34291
Learning task : Classification analysis
Sample percentage : 70
Sampling mode : Include sample
Evaluated variables : 14

Trained predictors
Selective Naive Bayes : 14 Variables

Select trained predictor: Selective Naive Bayes

14 Variables — Level distribution

name	level	weight	importance
Tree_7	0.1380	0.2815	0.1971
Tree_2	0.1329	0.1731	0.1517
Tree_10	0.0594	0.2776	0.1284
occupation	0.0818	0.1599	0.1144
Tree_4	0.1347	0.0930	0.1119
Tree_8	0.1045	0.1096	0.1070
Tree_3	0.1305	0.0862	0.1060
age	0.0424	0.2639	0.1058
Tree_5	0.1325	0.0505	0.0818
Tree_1	0.1339	0.0464	0.0788
workclass	0.0111	0.2385	0.0515
capital_gain`capital_loss	0.0094	0.2619	0.0498
race	0.0028	0.5510	0.0398
native_country	0.0165	0.0793	0.0362

occupation

Current group

Values of occupation	Frequency
Prof-specialty	6329

5.1.5. Evaluation

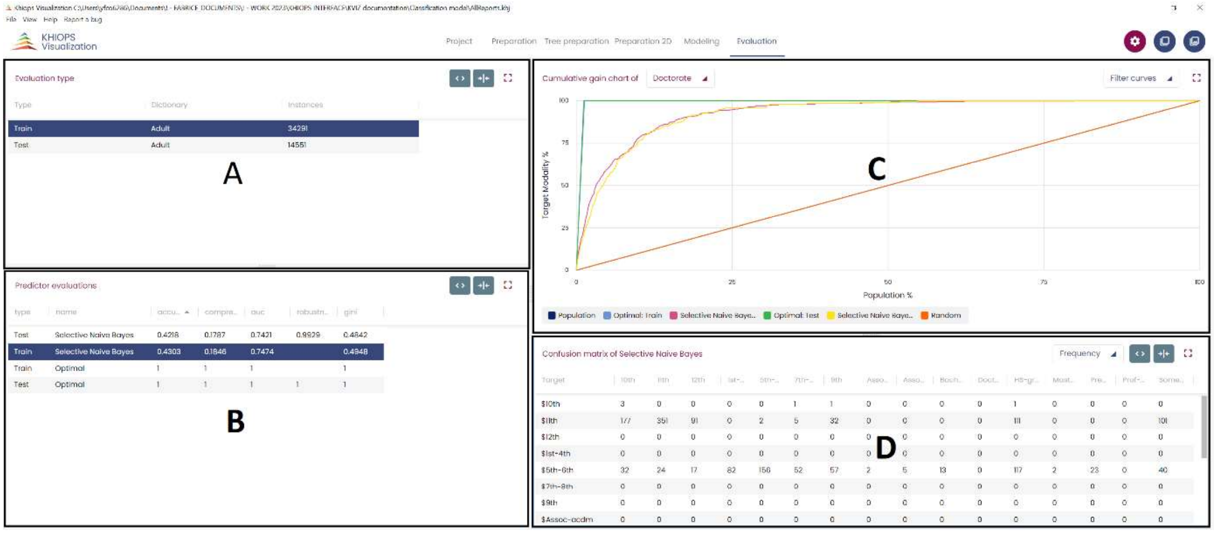
This tab gathers information about the performance of the model.

Panel A shows which datasets are concerned by the analysis.

Panel B gives some performance indicators for the model : accuracy, compression rate, area under the lift curve (auc) and gini coefficient.

Panel C shows the lift curves : the target modality can be chosen from the menu at the top and the "Filter curves" menu allows to choose which curves are displayed. The lift curve plots the target modality coverage as a function of the population selected according to the model score.

Panel D shows a confusion matrix with the predicted values (prefixed by $) and the actual values.


Khiops Visualization C:\Users\yfco6286\Documents\! - FABRICE_DOCUMENTS\! - WORK 2023\KHIOPS INTERFACE\KVIZ documentation\Classification model\AllReports.khj
File View Help Report a bug
KHIOPS Visualization
Project Preparation Tree preparation Preparation 2D Modeling Evaluation
A
B
C
D
Cumulative gain chart of Doctorate
Filter curves
Target Modality %
100
75
50
25
0
0
25
50
75
100
Population %
Population
Optimal: Train
Selective Naive Baye...
Optimal: Test
Selective Naive Baye...
Random
Frequency

Evaluation type

Type	Dictionary	Instances
Train	Adult	34291
Test	Adult	14551

Predictor evaluations

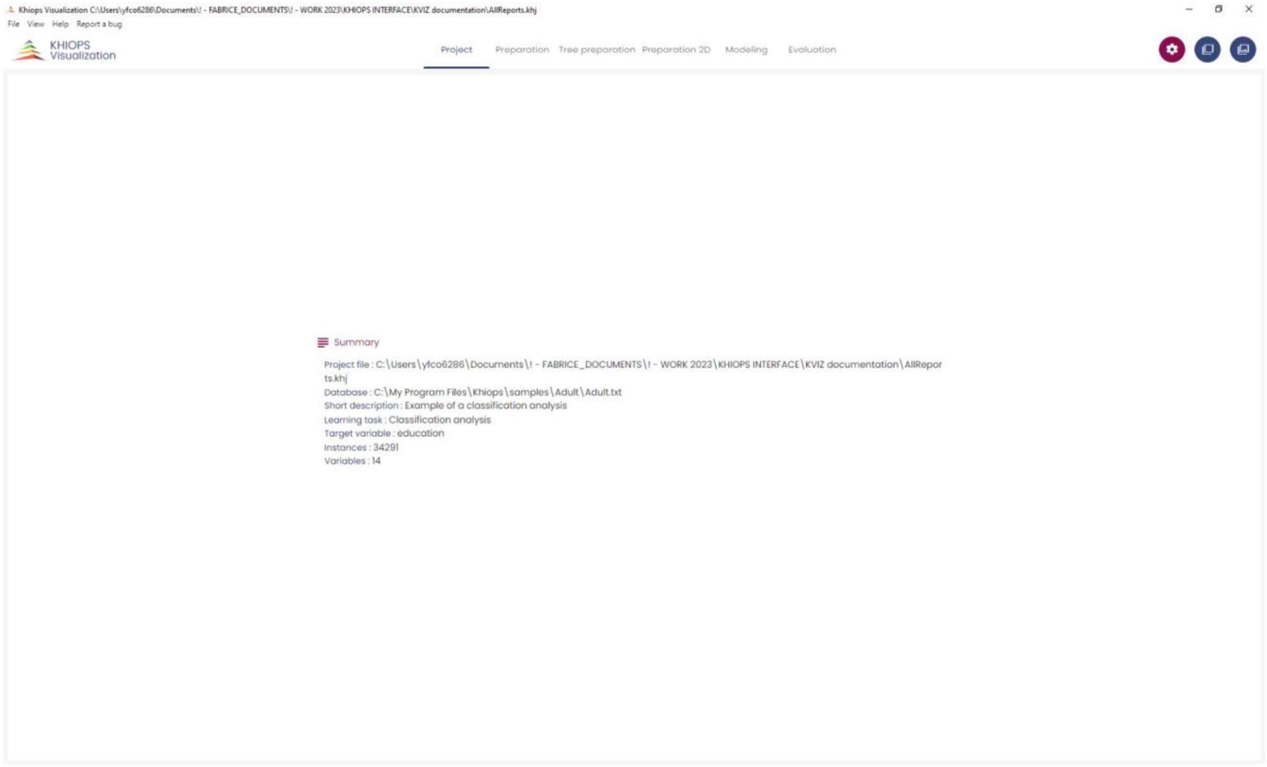
type	name	accu...	compre...	auc	robustn...	gini
Test	Selective Naive Bayes	0.4218	0.1787	0.7421	0.9929	0.4842
Train	Selective Naive Bayes	0.4303	0.1846	0.7474		0.4948
Train	Optimal	1	1	1		1
Test	Optimal	1	1	1	1	1

Confusion matrix of Selective Naive Bayes

Target	10th	11th	12th	1st-...	5th-...	7th-...	9th	Asso...	Asso...	Bach...	Doct...	HS-gr...	Mast...	Pre...	Prof-...	Some...
$10th	3	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0
$11th	177	351	91	0	2	5	32	0	0	0	0	111	0	0	0	101
$12th	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$1st-4th	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$5th-6th	32	24	17	82	156	52	57	2	5	13	0	117	2	23	0	40
$7th-8th	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$9th	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$Assoc-acdm	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

5.1.6. Project information

This tab shows information related to the project, in particular, the content of the "Short description" field.


Khiops Visualization C:\Users\yfco6286\Documents\! - FABRICE_DOCUMENTS\! - WORK 2023\KHIOPS INTERFACE\KVIZ documentation\AllReports.khj
File View Help Report a bug
KHIOPS Visualization
Project Preparation Tree preparation Preparation 2D Modeling Evaluation

Summary

Project file : C:\Users\yfco6286\Documents\! - FABRICE_DOCUMENTS\! - WORK 2023\KHIOPS INTERFACE\KVIZ documentation\AllReports.khj
Database : C:\My Program Files\Khiops\samples\Adult\Adult.txt
Short description : Example of a classification analysis
Learning task : Classification analysis
Target variable : education
Instances : 34291
Variables : 14

5.2. Regression

The target variable is numerical.

The report is organized in the following tabs :
- Preparation
- Modeling

- Evaluation
- Project information

For the figures below, we used the Adult database, with "education_num" as a target variable (excluding "education" from the predictors) and randomly selecting 70% of the data for the train set and the remaining 30% for the test set.

5.2.1. Preparation

Panel A shows some information related to the data and the analysis.

Panel B stores the list of the variables as a table together with some descriptive statistics information.

Clicking on a line of the table in panel B displays the information about the parts obtained from the optimal supervised grouping the selected variable selected in panel C and information about the variable (Name and derivation rule (if any)) in panel D.

The top of Panel C shows the histogram of the group populations obtained from the optimal supervised grouping for the predictive variable.

The bottom of panel C shows the optimal grouping of the (predictive variable, target variable) pair.

Note that the discretization of the target variable depends on the predictive variable, which is useful for interpretation purposes.

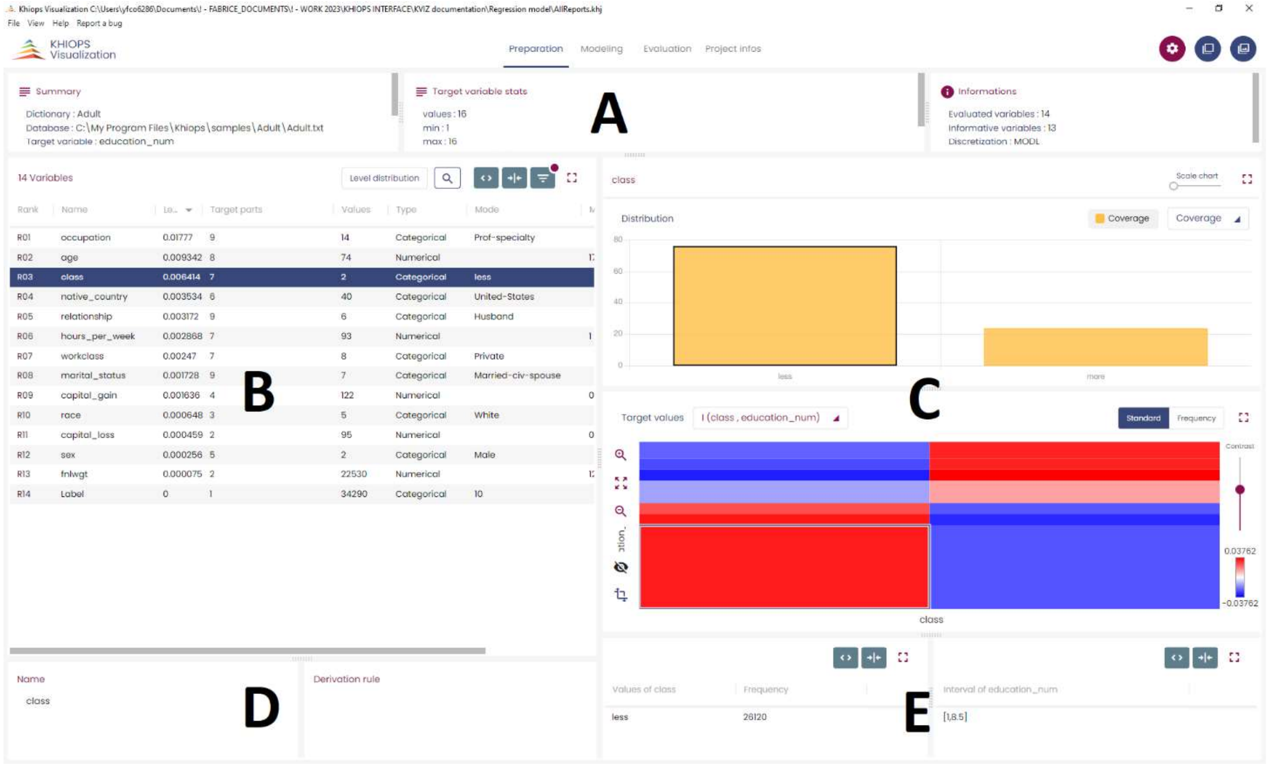
By default, the mutual information brought by each cell is given (red means positive association and blue negative association) ; other information can be selected from the menu.

The toggle Standard/Frequency selects a representation mode

- in "Standard" mode the span of group is proportional to the number of modalities in the group for categorical variables or constant for numerical variables

- in "Frequency" mode, the span of a group is proportional to its population

Clicking on a cell of the cartesian product shows information about the selected group in panel E.



Khiops Visualization C:\Users\yfco6286\Documents\! - FABRICE_DOCUMENTS\! - WORK 2023\KHIOPS INTERFACE\KVIZ documentation\Regression model\AllReports.khj

Summary

Dictionary : Adult
Database : C:\My Program Files\Khiops\samples\Adult\Adult.txt
Target variable : education_num

Target variable stats

values : 16
min : 1
max : 16

Informations

Evaluated variables : 14
Informative variables : 13
Discretization : MODL

14 Variables

Rank	Name	Le..	Target parts	Values	Type	Mode	M
R01	occupation	0.01777	9	14	Categorical	Prof-specialty	
R02	age	0.009342	8	74	Numerical		1
R03	class	0.006414	7	2	Categorical	less	
R04	native_country	0.003534	6	40	Categorical	United-States	
R05	relationship	0.003172	9	6	Categorical	Husband	
R06	hours_per_week	0.002868	7	93	Numerical		1
R07	workclass	0.00247	7	8	Categorical	Private	
R08	marital_status	0.001728	9	7	Categorical	Married-civ-spouse	
R09	capital_gain	0.001636	4	122	Numerical		0
R10	race	0.000648	3	5	Categorical	White	
R11	capital_loss	0.000459	2	95	Numerical		0
R12	sex	0.000256	5	2	Categorical	Male	
R13	fnlwgt	0.000075	2	22530	Numerical		1
R14	Label	0	1	34290	Categorical	10	

class

Target values I (class , education_num)

Name
class

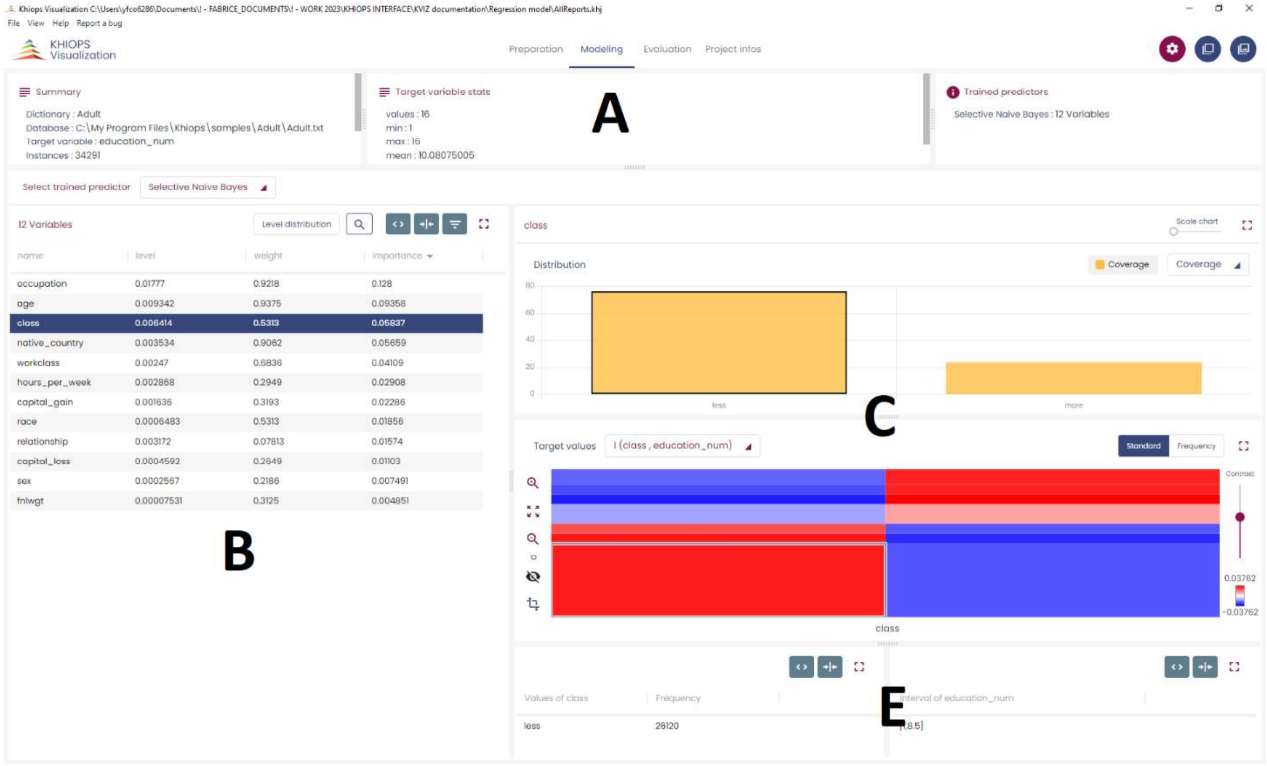
Derivation rule

Values of class	Frequency
less	26120

Interval of education_num
[1,8.5]

5.2.2. Modeling

The general layout is similar to the Preparation tab. Panel B shows information related to those variables used in the model only. The table shows for each variable its level, weight (in the weighted naive Bayes model) and its importance (computed as the geometric mean of the level and the weight).



Summary

Dictionary : Adult
Database : C:\My Program Files\Khiops\samples\Adult\Adult.txt
Target variable : education_num
Instances : 34291

Target variable stats

values : 16
min : 1
max : 16
mean : 10.08075005

Trained predictors

Selective Naive Bayes : 12 Variables

Select trained predictor: Selective Naive Bayes

12 Variables

name	level	weight	importance
occupation	0.01777	0.9218	0.128
age	0.009342	0.9375	0.09358
class	0.006414	0.5313	0.05837
native_country	0.003534	0.9062	0.05659
workclass	0.00247	0.6836	0.04109
hours_per_week	0.002868	0.2949	0.02908
capital_gain	0.001636	0.3193	0.02286
race	0.0006483	0.5313	0.01856
relationship	0.003172	0.07813	0.01574
capital_loss	0.0004592	0.2649	0.01103
sex	0.0002567	0.2186	0.007491
fnlwgt	0.00007531	0.3125	0.004851

class

Target values: I (class , education_num)

Values of class	Frequency
less	26120

Interval of education_num
[1,8.5]

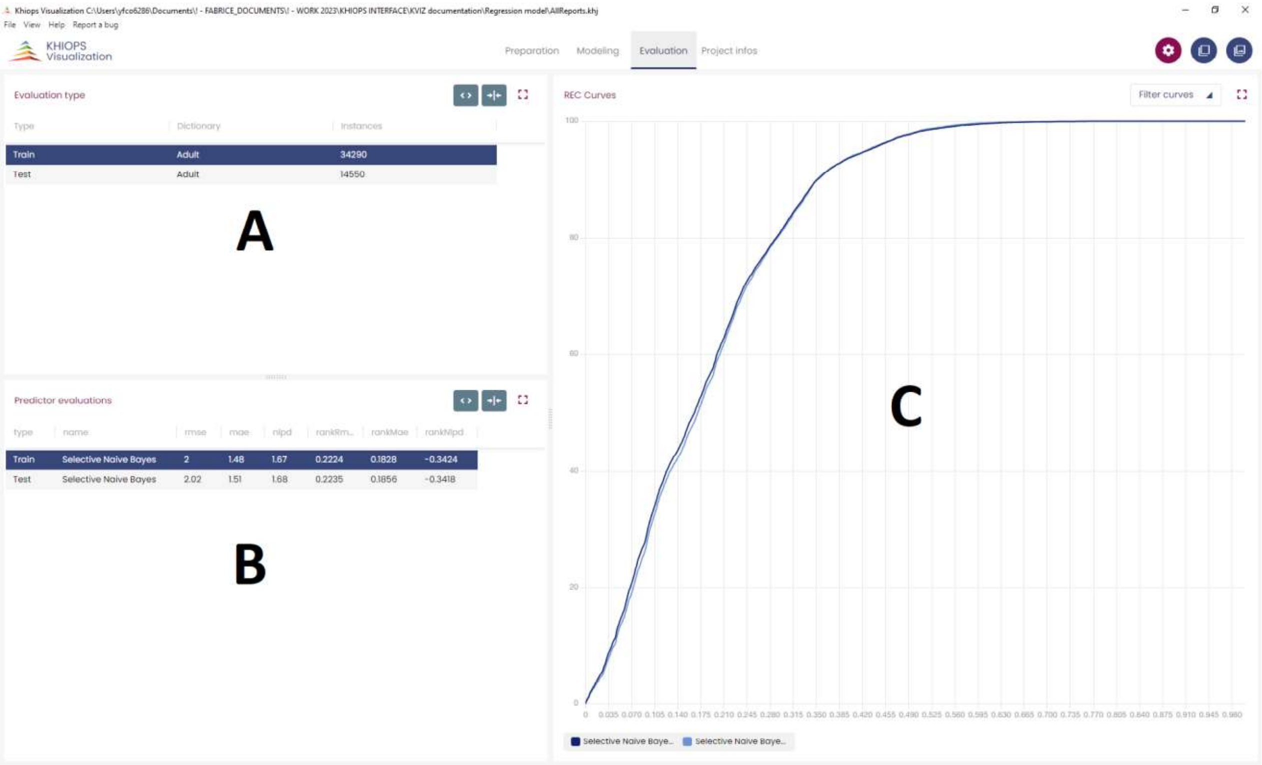
5.2.3. Evaluation

This tab gathers information about the performance of the model.

Panel A shows which datasets are concerned by the analysis.

Panel B gives some performance indicators for the model : rmse, mae, nlpd (negative log predictive density, see https://en.wikipedia.org/wiki/Negative_log_predictive_density for details) and their equivalents for rank regression (for the rank negative log predictive density, see https://www.jmlr.org/papers/volume8/hue07a/hue07a.pdf for details).

Panel C shows the REC curves (for details about REC curves, see https://dl.acm.org/doi/10.5555/3041838.3041844) : the target modality can be chosen from the menu at the top and the "Filter curves" menu allows to choose which curves are displayed.

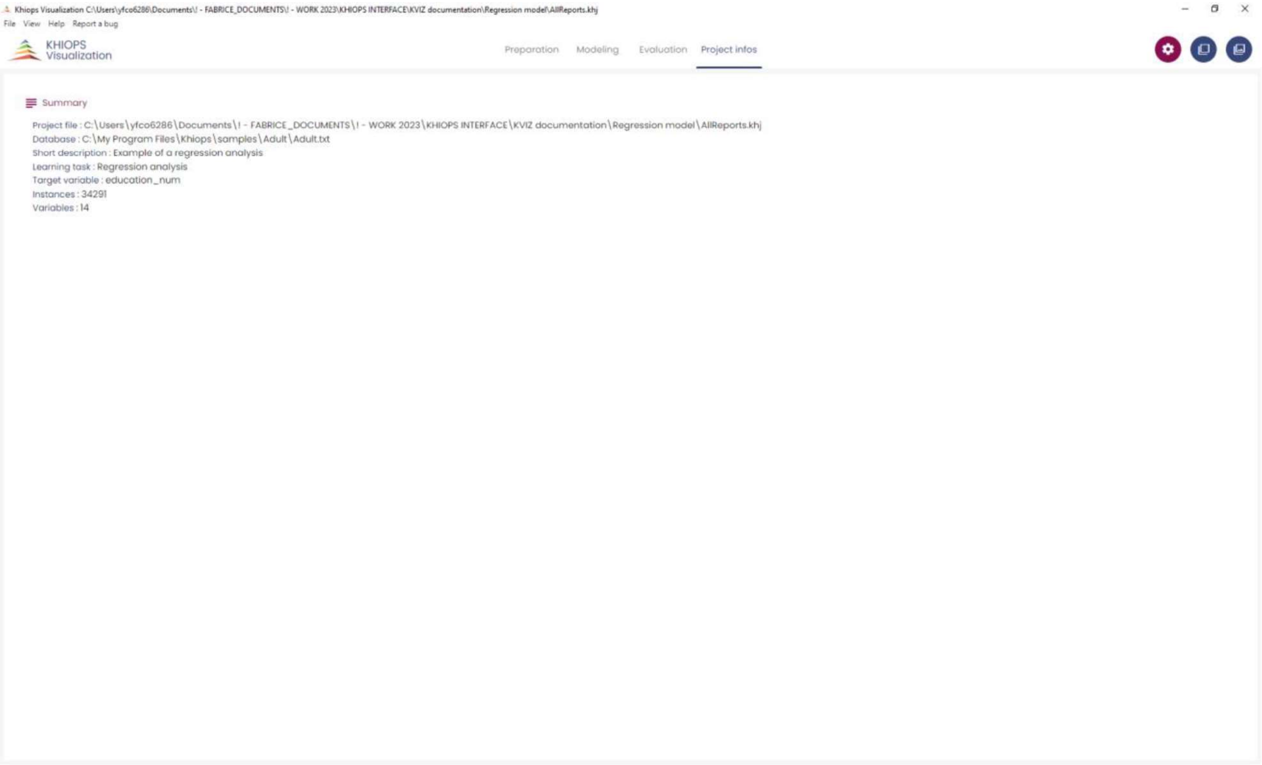


Type	Dictionary	Instances
Train	Adult	34290
Test	Adult	14550

type	name	rmse	mae	nlpd	rankRm...	rankMae	rankNlpd
Train	Selective Naive Bayes	2	1.48	1.67	0.2224	0.1828	-0.3424
Test	Selective Naive Bayes	2.02	1.51	1.68	0.2235	0.1856	-0.3418

5.2.4. Project information

This tab shows information related to the project, in particular, the content of the "Short description" field.

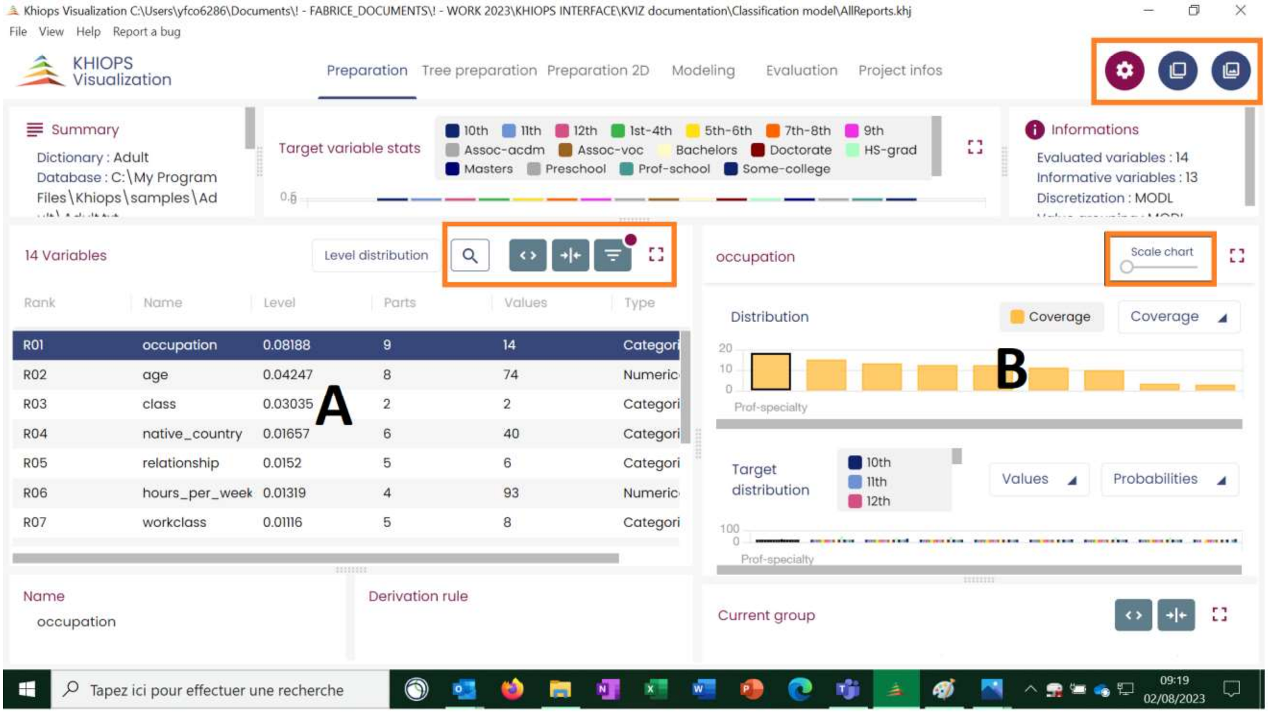

Khiops Visualization C:\Users\yfco6286\Documents\! - FABRICE_DOCUMENTS\! - WORK 2023\KHIOPS INTERFACE\KVIZ documentation\Regression model\AllReports.khj
File View Help Report a bug
KHIOPS Visualization
Preparation
Modeling
Evaluation
Project infos

Summary

Project file : C:\Users\yfco6286\Documents\! - FABRICE_DOCUMENTS\! - WORK 2023\KHIOPS INTERFACE\KVIZ documentation\Regression model\AllReports.khj
Database : C:\My Program Files\Khiops\samples\Adult\Adult.txt
Short description : Example of a regression analysis
Learning task : Regression analysis
Target variable : education_num
Instances : 34291
Variables : 14

6. Controls elements

Control elements appear at the top right of panels.



Khiops Visualization C:\Users\yfco6286\Documents\! - FABRICE_DOCUMENTS\! - WORK 2023\KHIOPS INTERFACE\KVIZ documentation\Classification model\AllReports.khj

Summary
Dictionary : Adult
Database : C:\My Program Files\Khiops\samples\Ad

Informations
Evaluated variables : 14
Informative variables : 13
Discretization : MODL

14 Variables

Rank	Name	Level	Parts	Values	Type
R01	occupation	0.08188	9	14	Categori
R02	age	0.04247	8	74	Numeric
R03	class	0.03035	2	2	Categori
R04	native_country	0.01657	6	40	Categori
R05	relationship	0.0152	5	6	Categori
R06	hours_per_week	0.01319	4	93	Numeric
R07	workclass	0.01116	5	8	Categori

Name
occupation

Derivation rule

At the top right of the screen,

from left to right :

- allows to set some environments variables such as the number of significant digits or the clear/dark mode. These settings are kept across sessions ;

- allows to copy the figure of a selected panel (select a panel by clicking in it ; a selected panel appears with a light blue border). This can be pasted in any other document ;

- allows to copy the data of a selected panel.

At the top right of panel A

Level distribution

from left to right

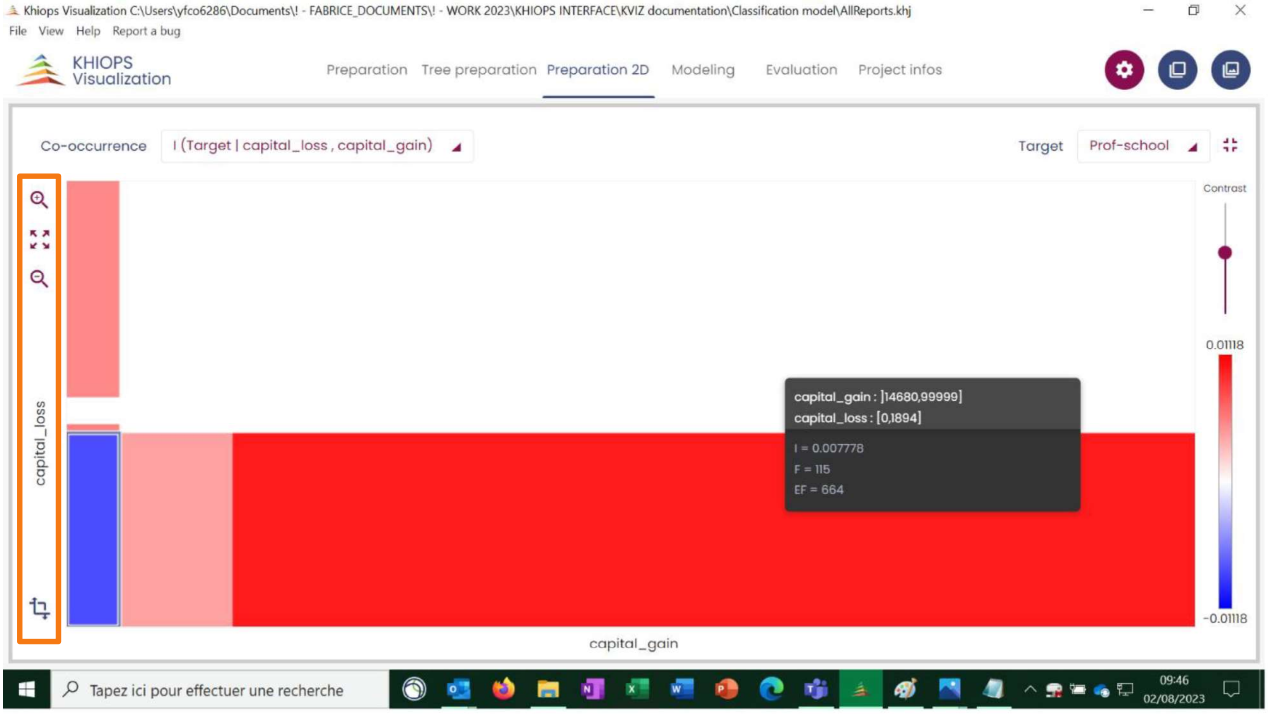
- allows to open a bar plot of the levels of the variables in decreasing order ;
- allows to open a search box on the variable names (fill at least two characters) ;
- allows to fit the table to the panel size ;
- allows to fit the columns of the table to their content ;
- allows to select/unselect columns to the table ;
- allows to open the panel in full screen. The control

at the top right of the full screen panel allows to go back to the multi-panel interface.

At the top right of panel B

allows to modifiy the horizontal scale of the histogram (by default, all groups appear in the figure, which may be messy when there are many groups).

Additional elements apply to bivariate cartesian product visualizations.


Khiops Visualization C:\Users\yfco6286\Documents\! - FABRICE_DOCUMENTS\! - WORK 2023\KHIOPS INTERFACE\KVIZ documentation\Classification model\AllReports.khj
File View Help Report a bug
KHIOPS Visualization
Preparation
Tree preparation
Preparation 2D
Modeling
Evaluation
Project infos
Co-occurrence
I (Target | capital_loss , capital_gain)
Target
Prof-school
Contrast
0.01118
-0.01118
capital_gain :]14680,99999]
capital_loss : [0,1894]
I = 0.007778
F = 115
EF = 664
capital_loss
capital_gain
Tapez ici pour effectuer une recherche
09:46
02/08/2023

On the left side of the panel (in full screen mode, here), from top to bottom
- allows to zoom in the figure (horizontal and vertical sliders allow to move in the figure) ;
- allows to reset the figure to its original scale ;
- allows to zoom out of the figure
- allows to transpose the figure.

7. Technical limits

In order to be responsive on very large files, some display limitations are taken into account in Khiops Visualization:

- Only the 10 000 most frequent values are displayed per group.