# 1 Subject

SIPINA proposes some descriptive statistics functionalities.

In itself, the information is not really exceptional; there is a large number of freeware which do that. It becomes more interesting when we combine these tools with the decision tree. The exploratory phase is improved. Indeed, every node of the tree corresponds to a subpopulation. The variables which do not appear in the tree are not necessarily irrelevant. Perhaps, some of them were hided during the tree learning which selects the "best" variables. By computing contextual descriptive statistics, in connection with the each node, we better understand the prediction rules highlighted during the induction process.

## 2 Dataset

We use the HEART\_DISEASE\_MALE.XLS<sup>1</sup> dataset. We want to predict the DISEASE from patient's characteristics (AGE, SUGAR in the blood, etc.). There are 209 examples.

# **3** Descriptive statistics

## 3.1 Data importation

The easiest way to import the dataset is to download the file into the EXCEL spreadsheet (see <a href="http://eric.univ-lyon2.fr/~ricco/doc/sipina\_xla\_installation.htm">http://eric.univ-lyon2.fr/~ricco/doc/sipina\_xla\_installation.htm</a> for the installation of the SIPINA.XLA add-in). Then we select the cells and activate the SIPINA / EXECUTE SIPINA menu (see <a href="http://eric.univ-lyon2.fr/~ricco/doc/sipina\_xla\_processing.htm">http://eric.univ-lyon2.fr/~ricco/doc/sipina\_xla\_installation.htm</a> for the installation of the SIPINA.XLA add-in).

<b>N</b>	licrosoft	Exce	l - heart_disea	ise_male.xls						
<b>B</b> 1	Fichier F	dition	Affichage Ir		Outils Donné	es Fenêtre	7 Tapagra	Sipipa	Tapez une questi	
								Evenute Sinin	-	
	🗁 📙 d	3 2	1 🖨 🖸 🟹	¥ 🖻 🖫 •	🗞   10 + 01 -	- 🔚 Σ - 🛱	+ &+   📖 -	Execute Sipin	· · · ·	
. ←	→ ⊗	<b>(</b>	🚮 🔕 <u>F</u> avoris	• Aller <u>à</u> • 🔽	D:\_Travaux\ur	niversity\SiteWeb	Labo\Site_Eric_F	Ricco\public_html\d	• •	
Bite	tream Vera	Sans	- 8 - 6	7 c = 3		₩ 000 <b>*</b> ,0		- A - A		. 🔦
Dics		Juna				· /o UUU ,00 +	,0   == ==			
	H210	•	<i>†</i> ≁ neg:	ative	_	_	_			
	A		В	С	D	E	F	G	Н	
1	age		chest_pain	rest_bpress	blood_sugar	rest_electro	max_heart_ra	texercice_angir	nå disease	
2		43	asympt	140	t	normal	13	35 yes	positive	
3		39	atyp_angina	120	t	normal	16	0 yes	negative	
4		39	non_anginal	160	t	normal	16	ou no	negative	
5		42	non_anginal	160	t r	Execute Sipina	14			
7		49	asympt	140	T c					
6		50	asympt	140	T	Dataset rang	e (including the name	e of the attributes firs	row):	
		59	asympt	140	t r	\$A\$1:\$H\$	210	_		
9		54	asympt	200	T r			~~~~~	C	
11		59	asympt	130	T 4			UK 54/3	Cancer	
12		50	asympt	1/0	f I	of thus abr			DOG DENO	
12		52	non_anginai	140	۲ ۲	st_t_wave_abr	1/	o no	negative	
14		50	asympt	100	+	normal	14	2 10	positive	
14		55	atyp_angina	140	+	normal	14	io yes	positive	
16		20	acyp_angina	140	f f	normal	14	6 no	negative	
17		60	asympt pop anginal	120	f	left vent byne	10	25 00	positive	
18		55	at/n_angina	140	f	normal	13	0 00	negative	
19		50	asympt	140	f	st t wave abr	1.	l0 ves	nositive	
20		48	asympt	106	+	normal	11	0 00	positive	
20		39	atvn angina	190	f	normal	10	)6 no	negative	
22		66	asympt	140	f	normal	10	4 ves	positive	
23		56	asympt	140	t	normal	19	0 ves	positive	
24		44	asympt	135	f	normal	13	35 no	positive	
25		43	asympt	120	f	normal	12	0 ves	positive	
14 4	► H\h	eart	dataset /	110						
Des	sin • 🗟	Forme	, es automatiques •	NND0	) 🔮 📣 🔅 🛛	2 🔜 💩 - ,	<u>/</u> - <u>A</u> - =		0.	
Prêt								Somme=66713		NUM

<sup>&</sup>lt;sup>1</sup> <u>http://eric.univ-lyon2.fr/~ricco/dataset/heart\_disease\_male.xls</u>

SIPINA is automatically started. The data were transferred through the clipboard. The data file contains 209 individuals and 8 variables.

**Note:** We can save the dataset in the SIPINA binary file format (\*.FDM) by clicking the FILE /SAVE AS menu. The format is useful when we handle a large dataset. During the transfer, numeric columns are encoded as continuous attributes, the other ones as discrete attributes. The first row is always the variable names.

🕱 Sipina Research Version - [Learning set editor]											
🔉 File Edit Data Statistics Induc	tion method	Analysis Vie	w Window	Help				-	. 8 ×		
<u> 🔁</u> 📑 ங 🖻											
×		age	chest_pain	rest_bpress	blood_sugar	rest_electro	max_heart_r	; exercice_an	dise: 🔨		
Attribute selection	1	43.00	asympt	140.00	f	normal	135.00	yes	posit		
	2	39.00	atyp_angina	120.00	f	normal	160.00	yes	nega		
	3	39.00	non_anginal	160.00	t	normal	160.00	no	nega		
	4	42.00	non_anginal	160.00	f	normal	146.00	no	nega		
	5	49.00	asympt	140.00	f	normal	130.00	no	nega		
	6	50.00	asympt	140.00	f	normal	135.00	no	nega		
	7	59.00	asympt	140.00	t	left_vent_hyp	119.00	yes	posit		
	8	54.00	asympt	200.00	f	normal	142.00	yes	posit		
	9	59.00	asympt	130.00	f	normal	125.00	no	posit		
	10	56.00	asympt	170.00	f	st_t_wave_a	122.00	yes	posit		
	11	52.00	non_anginal	140.00	f	st_t_wave_a	170.00	no	nega		
1	12	60.00	asympt	100.00	f	normal	125.00	no	posit		
Learning method	13	55.00	atyp_angina	160.00	t	normal	143.00	yes	posit		
MethodName=Improved ChAID (Tsc 🔨	14	57.00	atyp_angina	140.00	t	normal	140.00	no	nega		
Hdl=8	15	38.00	asympt	110.00	f	normal	166.00	no	posit		
Merge=0.05	16	60.00	non_anginal	120.00	f	left_vent_hy;	135.00	no	nega		
Split=0.001	17	55.00	atyp_angina	140.00	f	normal	150.00	no	nega		
I ypeBonterroni=1 ValueBonterroni=1	18	50.00	asympt	140.00	f	st_t_wave_a	140.00	yes	posit		
Sampling=0	19	48.00	asympt	106.00	t	normal	110.00	no	posit		
Examples selection	20	39.00	atyp_angina	190.00	f	normal	106.00	no	nega		
209 examples selected	21	66.00	asympt	140.00	f	normal	94.00	yes	posit 🗸		
0 examples idle									>		
J	Editing M	EW.FDM			(	Attribut	es : 8	Examples : 20	19 🟒		
Improved ChAID (Tschuprow Goodness o	f Split)										

## 3.2 Univariate statistics

Descriptive statistics commands are available through the **STATISTICS** menu.

**Note:** This **menu is only visible if the data grid is selected**. In the other situation i.e. another window is selected, this menu is hidden. Among the various ways to select the data grid, we can use the WINDOW / LEARNING SET EDITOR menu.

#### **3.2.1 Continuous variables**

We select the STATISTICS / DESCRIPTIVE STATISTICS / UNIVARIATE menu in order to compute the descriptive statistics for continuous variables. In the dialog box which appears, we activate the CONTINUOUS VARIABLES tab. Then, we select the two following variables: REST\_BPRESS and MAX\_HEART\_RATE.

🔉 Sipina Research Version - [Learning set editor]											
💦 File Edit Data Statistics Induc	tion method:	Analysis Vie	w Window	Help					- 8 ×		
🏋 🎬 📴 👘 📴 Descriptive st.	atistics 🔸	Univariate	_								
Transform	• _	Bivariate	• ·	rest hnress	bloor	sugar rest electro	may heart	réevercice			
Attribute selection Missing data	•	Conditional		140.00	f	normal	135.00	ves	posit		
	2	Contingency	tables	120.00	f	normal	160.00	yes	nega		
	3	39.00	non_anginal	160.00	t	normal	160.00	no	nega		
	4	Descriptive	etatistics onti	ops	V						
	5	Descriptive									
	6			dbles			1				
	7	Statistics age									
		max, heart, rate									
	10	Variance					- Exan	nples			
	11	M Standard deviation									
<u> </u>	12	Kurtosis									
Learning method	13	🔽 Skewne	88					le examples			
MethodName=Improved ChAID (Tsc 🔨	14	🗌 🗌 Median						overed example			
Hdl=8	15							OB	Annuler		
Merge=0.05	16	60.00	non_anginal	120.00	t	left_vent_hy	¢135.00	no	nega		
Split=0.001	17	55.00	atyp_angina	140.00	f	normal	150.00	no	nega		
ValueBonferroni=1	18	50.00	asympt	140.00	1	st_t_wave_s	a140.00	yes	posit		
Sampling=0 💌	19	40.00	asympt atum angina	106.00	1 4	normal	110.00	no	posit		
Examples selection	20	00.88	asympt	140.00	f	normal	94.00	ves	nege		
209 examples selected O examples idle			ore y mpr	0.00	ľ	norma	01.00	700			
	Editing	NEW EDM				<b>Attribut</b>	tec : 8	Examples	209		
, Improved ChAID (Tschuprow Goodness o	proved ChAID (Tschuprow Goodness of Split)										

We note that the statistical indicators can be computed only on the active (selected) examples. This is useful for instance if we have partitioned the dataset into learning set and test set.

The results are displayed in a new window. Each column corresponds to a statistical indicator; each row to a variable. We can copy the values in the clipboard or modify the numerical precision using the contextual menu.

🛿 Sipina Research Version - [Descriptive statistics results on continuous variables]												
🔉 Induction method Analysis View	Window He	lp						-	. 🗗 🗙			
<u> 🕀</u> 📑 뉅 🖻												
Attribute selection												
		Examples	Min	Max	Mean	Std	Mean/Std	Skewness	Sig. S			
	rest bpre	209	92.0000	200.0000	133.6603	17.4330	7.6671	0.6640	0.0001			
	max heart	209	82.0000	188.0000	137.5742	23.8768	5.7618	-0.0784	0.6434			
Learning method MethodName=Improved ChAID (Tsc MethodClassName=TArbreDecision Hdl=8 Merge=0.05 Split=0.001 TypeBonferroni=1 ValueBonferroni=1 ValueBonferroni=1 Sampling=0 Examples selection 209 examples selected 0 examples selected 0 examples idle												
	<								>			
Improved ChAID (Tschuprow Goodness o	f Split)								11			

#### **3.2.2 Discrete variables**

We follow the same way for the discrete variables. We activate before the data grid, by clicking the window or by clicking the WINDOW / LEARNING SET EDITOR menu.

🔉 Sipina Research Version - [De	escriptive st	atistics re	sults on co	ontinuous v	ariables]				
🔉 Induction method Analysis View	Window He	lp							. a ×
Attribute selection	Cascade Tile Arrange ic Close All 1 Learning 2 Analysis	Cascade Tile Arrange icons Close All  1 Learning set editor 2 Analysis Report 3 Descriptive statistics results on continuous variables							
	3 Descript	ve statistics	The second state of	(m)	[a				
		Examples	nin	nax	nean	5ta	nean/std	Skewness	51g. 2
	rest bpre	209	92.0000	200.0000	133.6603	3 17.4330	7.6671	0.6640	0.0001
Learning method MethodName=Improved ChAID (Tsc MethodClassName=TArbreDecision  Hdl=8 Merge=0.05 Split=0.001 TypeBonferroni=1 ValueBonferroni=1 ValueBonferroni=1 Sampling=0 Examples selection 209 examples selected 0 examples selected									
									>
Improved ChAID (Tschuprow Goodness of	f Split)								

Then, we select again the STATISTICS / DESCRIPTIVE STATISTICS / UNIVARIATE menu. In the dialog box, we choose the DISCRETE VARIABLES tab. We want to compute statistical indicators about CHEST\_PAIN and EXERCICE\_ANGINA.

🔉 Sipina Research Version - [L	earning set	editor]									
💦 File Edit Data Statistics Indu	ction method	Analysis Vie	w Window	Help					- 8 ×		
🎦 🗃 📴 💏 📴 Descriptive st	atistics 🔸	Univariate	_								
Transform	•	Bivariate	►	rest boress	blood sugar	rest electro	max heart r	exercice an	ndise: 🔨		
Attribute selection Missing data	•	Conditional		140.00	f	normal	135.00	ves	posit		
	2	Contingency	ables	120.00	f	normal	160.00	yes	nega		
	3	39.00	non_anginal	160.00	t	normal	160.00	no	nega		
	4	42.00	non_anginal	160.00	Y	normal	146.00	no	nega		
	5 Descri	iptive statisti	cs options						egs 🔽		
	6 Contin	uous variables	Discrete variable	88					ege		
	7 Stati	stics		abaat poin			1		osit		
		Inde		blood_sugar					psit		
		nebelik, decite		rest_electro exercice angin	а				hsit		
	11		lons	disease			Examples =	-11	eqa		
<u> </u>	12						whole a	ataset	psit		
Learning method	13						C Idla outre	kampies soloo	psit		
MethodName=Improved ChAID (Tsc 🔨	14						C Covered	evamples	ege		
Hdl=8	15						1000000		psit		
Merge=0.05	16						🗸 0 <del>. R</del> 2	Annule 🗙	er ege		
Split=0.001	17	55.00	atyp_angina	140.00	T	normal	150.00	no	nèga		
ValueBonferroni=1	18	50.00	asympt	140.00	T +	st_t_wave_a	140.00	yes	posit		
Sampling=0 💌	20	40.00	asymp. atum angina	100.00	4	normal	106.00	10	posit		
Examples selection	20	66.00	asymp_angina	140.00	f	normal	94.00	ves	nega posit 👽		
209 examples selected O examples idle			ore j mpt	. 10.00		rise risel	01.00	,			
	Editing	DI NEW EDM						Evennles : 21	na		
, Improved ChAID (Tschuprow Goodpess (	porroved ChAID (Tschunzow Goodness of Shift)										
ing the second provide decarloss (									14		

🕱 Sipina Research Version - [Descriptive statistics results on categorical variables]												
🔉 Induction method Analysis View	Window Helj	p				_ 8 ×						
🖰 🖹 📲 h 🖒												
Attribute selection	Descriptiv Whole dat:	7e statisti Aset	ics on NE	J. FDM								
chest_pain   exercice_angina												
		Count	Frequency									
	asympt	102	0.4880									
	atyp_angina	65	0.3110									
	non_anginal	36	0.1722									
	typ_angina	ь 200	0.0287									
Learning method	Sum	203										
MethodClassName=TArbreDecision												
Hdl=8												
Split=0.001												
TypeBonferroni=1												
ValueBonterron=1												
Examples selection												
209 examples selected												
0 examples idle												
<u> </u>	<u> </u>											
Improved ChAID (Tschuprow Goodness of	f Split)					11						

The frequencies of values are displayed for each variable (one variable by tab).

## 3.3 Bivariate statistics

We can also compute bivariate statistics: combining two discrete variables (contingency table), two continuous variables (correlation) or mixed variables (comparison of populations).

We click on the WINDOW / LEARNING SET EDITOR menu in order to activate the data grid. The STATISTICS menu is now visible.

#### 3.3.1 Two continuous variables: scatter plot

The scatter plot provides useful information about the relation between two variables (kind of association, outliers, etc.).

This functionality is available with the STATISTICS / DESCRIPTIVE STATISTICS / BIVARIATE / SCATTERPLOT menu. A dialog box enables to select the variable on the horizontal axis (X: REST\_BPRESS) and the vertical axis (Y: MAX\_HEART\_RATE).

🔊 Sipina Research Version - [Learning set editor]											
🔉 File Edit Data Statistics Indu	ction method	Analysis Vie	w Window	Help					. 8 ×		
🏹 🏔 🅦 🛊 Descriptive st	atistics 🕨	Univariate									
Transform	, I	Bivariate	Þ	Scatterplot	t		hand a	.(	(		
Attribute selection Missing data	•	Conditional		Scatterplot	t with markers		135 00	r;exercice_an	nosit		
	2	Contingency t	ables	120.00	f	normal	160.00	ves	nega		
	3	39.00	non_anginal	160.00	t	normal	160.00	no	nega		
	4	42.00	non_anginal	160.00	f	normal	146.00	no	nega		
	5	49.00	asympt	140.00	f /	normal	130.00	no	nega		
	6	50.00 Set	ect attribute	s for scatterp	lot				nega		
	7	59.00	ttributes selection	ι <u> </u>					posit		
	8	54.00									
	10	59.00 r	6.00 rest_bpress rest_bpress Examples Examples Whole dataset								
	11	52.00									
	12	60.00		max_	_heart_rate	C Inactive exa	mples remoles		posit		
Learning method	13	55.00				C Covered e	kamples		posit		
MethodName=Improved ChAID (Tsc 🔨	14	57.00				<u> </u>			nega		
MethodClassName=TArbreDecision	15	38.00				VIAS	Anr	nuler	posit		
Merge=0.05	16	60.00	non_anginal	120.00	f	left_vent_hy	(¢135.00	no	nega		
Split=0.001	17	55.00	atyp_angina	140.00	f	normal	150.00	no	nega		
ValueBonferroni=1	18	50.00	asympt	140.00	f	st_t_wave_	a140.00	yes	posit		
Sampling=0 💌	19	48.00	asympt	106.00	t	normal	110.00	no	posit		
Examples selection	20	39.00	atyp_angina	190.00	1	normal	106.00	no	nega		
209 examples selected		00.00	asympt	140.00	1	normai	94.00	yes	posit		
u examples idle											
Incompared ChATD / Tachymany Condense of Sellity											
Improved CHAID (Ischuprow Goodness o	or Split)										

The display window is generated. In the same time, a new menu (GRAPH) is now available. Some options enable us to copy the graph in the clipboard, to print it, to modify the size of the points, etc.



#### 3.3.2 Conditional scatter plot

The scatter plot is all the more interesting when we can illustrate the relative situation of groups of individuals. In this case, we use a third variable in order to "colorize" the points. In our preceding example, we want to distinguish the people according the DISEASE.

We select again the data grid (WINDOW / LEARNING SET EDITOR menu). Then we click on the STATISTICS / DESCRIPTIVE STATISTICS / BIVARIATE / SCATTERPLOT WITH MARKERS menu. In the

dialog box, we set REST\_PRESS as horizontal axis, MAX\_HEART\_RATE as vertical axis, and DISEASE as marker.

🔊 Sipina Research Version - [l	earning se.	t editor]								
💦 File Edit Data Statistics Indi	uction method	Analysis Vi	ew Window	Help					- 8 ×	
🕰 📷 🛤   👞 🖻 🛛 Descriptive s	tatistics 🕨	Univariate								
		Bivariate	۱.	Scatterpl	ot			,		
Attribute selection		en data a d		Scatterple	ot with marke	rs ectro	o max_hear	t_r:exercice_a	ngdiser 🔨	
Attribute selection Missing data	,	Conditional	: 1	140.00	1	normai	135.00	yes	posit	
	2	Contingency	tables	120.00	f	normal	193.00	yes	nega	
	3	39.00	non_anginal	160.00	t	somal	160.00	no	nega	
	4	42.00	non_anginal	160.00	f	normal	146.00	no	nega	
	5	49.00	asympt	140.00	1	normal	130.00	no	nega	
	6				V				nega	
	7	Select attri	butes for so	atterplot				<u> </u>	posit	
	8	Attributes selection							posit	
9		age		∐ Axis		disease			posit	
	10	rest_bpress		rest bore	222				posit	
	11	max_heart_r	rate	lun ann		Examples			nega	
	12			<u>Y</u> Axis		Whole d	ataset		posit	
Learning method	13			max_hea	art_rate 🧲	C Active ex	kamples		posit	
MethodName=Improved ChAID (Tsc 🔨	14			,	1	C Inactive	examples		nega	
MethodClassName=TArbreDecision	15					C Covered	examples		posit	
Marae=0.05	16								nega	
Split=0.001	17					🔰 🗸 ок		🗙 Annuler	negs	
TypeBonferroni=1	18	50.00	asympt	140.00	f	st_t_wave	_a140.00	yes	posit	
Sampling=0	19	48.00	asympt	106.00	t	normal	110.00	no	posit	
Examples selection	20	39.00	atyp_angina	190.00	f	normal	106.00	no	nega	
209 examples selected	- 21	66.00	asympt	140.00	f	normal	94.00	yes	posit 🗸	
0 examples idle	<									
	Editing	NEW.FDM	VEW.FDM Attributes : 8					Examples : 2	209 //	
nproved ChAID (Tschuprow Goodness of Split)										

We obtain the same scatter plot than previously. The difference is we distinguish now the people DISEASE = YES.



#### 3.3.3 Continuous variable vs. Discrete variable

This functionality enables, among other, to compare the characteristics of subpopulations. Especially, it allows comparing the conditional distribution of a continuous attribute according the value of a discrete variable. For our dataset, we want to study the distribution of MAX\_HEART\_RATE for each subpopulation corresponding to the DISEASE values.

We select the data grid (WINDOWS / LEARNING SET EDITOR menu). Then, we click on the STATISTICS / DESCRIPTIVE STATISTICS / CONDITIONNAL menu. In the dialog box, we select MAX\_HEART\_RATE and DISEASE.

Sipina Research	Version - IL	earning set	editor]								
🔉 File Edit Data	Statistics Indu	iction method	Analysis Vie	ew Window	Help					- 8 ×	
 15 16 18 18 18	Descriptive s	tatistics 🔸	Univariate								
	Transform	•	Bivariate		rest boress	blood sugar	rest electro	may beart	rleveroice, er	diaa 👗	
Attribute selection	Missing data	•	Conditional	_	140.00	Dioou_sugar	normal	135.00	ves	posit	
		2	Contingency	tables	120.00	f	normal	160.00	yes	nega	
		3	39.00	non_anginal	160.00		normal	160.00	no	nega	
		4	42.00	non_anginal	160.00	f	normal	146.00	no	nega	
		5	49.00	asympt	140.00	f	normal	130.00	no	nega	
		Conditionn	al descriptive	e statistics					X	nega	
Continuous variables Discrete attribute :								posit			
	🖵 Statistic						[ ] []	sease	<u> </u>	posit	
		Mean		age	horoco				······································	posit	
				max	heart rate					posit	
			e.			N	E	xamples		nega	
Learning method		: 🔽 Standa	rd deviation					Whole datas	et	posit	
MethodName=Improved	d ChAID (Tisc 🔨	🔽 🔽 Kurtosi	s				0	Active examp	bles	neas	
MethodClassName=TAr	rbreDecisionI	🔽 Skewn	ess				0	Idle example:		posit	
Marae=0.05		🗆 Median						Covered exa	mples	nega	
Split=0.001								V ORV	🗙 Annuler	nega	
ValueBonterroni=1		10	100.00	asympt	140.00	<u>1</u>	SI_I_WAVE_0	1140.00	700	posit	
Sampling=0	~	19	48.00	asympt	106.00	t	normal	110.00	no	posit	
Examples selection		20	39.00	atyp_angina	190.00	f	normal	106.00	no	nega	
209 examples selected		21	66.00	asympt	140.00	f	normal	94.00	yes	posit 🗸	
0 examples idle										>	
1		Editing	VEW.FDM				Attribut	es : 8	Examples : 20	09 //	
Improved ChAID (Tschu	mproved ChAID (Tschuprow Goodness of Split)										

The result grid gives the descriptive statistics. We obtain the mean, the standard deviation, etc. We observe for our dataset that the average of MAX\_HEART\_RATE is lower for the people with DISEASE = POSITIVE (YES).

🔉 Sipina Research Version - [Conditional statistic descriptive result]											
🔉 Induction method Analysis View	Window Hel	0						-	u X		
<u>ð</u> 🗎 🕌 📦											
Attribute selection	Memo		1								
	disease vs m	ax_neart_rate									
Learning method MethodName=Improved ChAID (Tsc MethodClassName=TArbreDecisionI HdI=8 Merge=0.05 Split=0.001 TypeBonferroni=1 Sampling=0	positive negative	Examples 92 117	Min 82.0000 98.0000	Max 180.0000 188.0000	<u>Mean</u> 127.9022 145.1795	8td 22.6085 22.1073	Mean/Std 5.6573 6.5670	Skewness 0.1460 -0.2298	<u>Sig.</u> 0.56 0.31		
209 examples selected 0 examples idle									>		
Improved ChAID (Tschuprow Goodness o	f Split)								//		



#### 3.3.4 Two discrete variables: contingency table

This functionality enables to measure the association between two discrete variables through the contingency table. The chi-square statistic for independence test is computed. We activate the menu STATISTICS / DESCRIPTIVE STATISTICS / CONTINGENCY TABLES menu. In the dialog box, we select BLOOD\_SUGAR and DISEASE.

🔉 Sipina Research Version					
File Edit Data Statistics Induction metho	od Analysis View Wind	ow Help			
Descriptive statistics	Univariate Bivariate	•			
Attribute selection Mission data	Conditional				<u>^</u>
Philipsing data	Contingency tables	heres	blood sugar rest electr	ro may beart reverci	ce e
1	43.00 asv	mpt 140.00	f normal	135.00 ves	<u>cc_u</u>
2	3 Select variable	es 🚺		×	
3	3 – Discrete attribute	BS			
4	4 Row	Column	<u></u>		
5	4 blood sugar	blood_s	ain ugar		
6	5 rest_electro exercice_angin	a exercice	ctro e_angina		
<mark>/</mark>	5 disease	disease	Exa	amples	
Learning method	5		• \	Whole dataset	
MethodName=Improved ChAID (Tsc A	5		04	Active examples	
Hdl=811	5			die examples	
Merge=0.05	6			Lovered examples	
TypeBonferroni=1	5	J		🖊 🖓 🦉 🗶 Annuler	
ValueBonferroni=1	57.00 aty;	p_angina 140.00	t normal	140.00 no	
Sampling=U 15	38.00 asy	mpt 110.00	f normal	166.00 no	_
Examples selection 16	60.00 non	_anginal 120.00	f left_vent_l	hy¢135.00 no	_
209 examples selected 17	55.00 aty;	p_angina 140.00	f normal	150.00 no	
18	ISU UU III III III III	mmm 11411111	IT IST T WAVE	e an40000 IVes	>
Improved ChAID (Tschuprow Goodness of Split	)				

We obtain the contingency table, the chi-square statistic, and the p-value of the test. We have also the contribution of each cell to the chi-square statistic.

🔉 Sipina Research Version - [Co	ontingend	y table resu	lts]					
Induction method Analysis View	Window	Help				_ 8 ×		
🖰 👸 🞇 🐜 📬								
Attribute selection	Memo 							
	blood_su	gar vs disease						
Learning method		positive	negative	Sum				
	f	81	112	193				
	t	11	5	16				
	Sum	92	117	209				
MethodClassName=TArbreDecision								
Hdl=8 Merae=0.05		Contributions						
Split=0.001		positive	negative	Sum				
I ypeBonterroni=1 ValueBonterroni=1	f	-0.04	0.03	0.08				
Sampling=0	t	0.52	-0.41	0.92				
Examples selection	Sum	0.56	0.44	1.00				
209 examples selected 0 examples idle								
	Chi-2 statistic = 4.30 with significance 0.0381							
Improved ChAID (Tschuprow Goodness o	f Split)							

## 4 Descriptive statistics for a subpopulation

Each node of a classification tree corresponds to a subsample of the dataset. It will be very interesting to compare the characteristics of these groups using descriptive statistics. This functionality is very useful when we want to build interactively the tree.

### 4.1 Interactive tree induction

First of all, we must close all the windows in relation with the previous analysis. We click on the WINDOW / CLOSE ALL menu.

🔉 Sipina Research Version - [Co	ontingency	table resu	lts]			
Induction method Analysis View	Window He	lp				_ @ ×
L E E I I I I I I I I I I I I I I I I I	Cascade Tile Arrange ic Close All 1 Learning 2 Analysis 3 Descript 4 Descript 5 Scatterp 6 Scatterp	ons set editor Report ve statistics ve statistics lot lot				
Learning method MethodName=Improved ChAID (Tsc MethodClassName=TArbreDecision	9 Conditio 8 Conditio 9 Continge	nal statistic ( nal distributi ency table re				
Merge=0.05				Contribution	s	
Split=0.001 TypeBonferroni=1 ValueBonferroni=1 Sampling=0	f	positive -0.04 0.52	0.03 -0.41	Sum 0.08 0.92	_	
Examples selection 209 examples selected 0 examples idle	Sum Chi-2 statistic	0.56 c = 4.30 with	0.44 significance 0	1.00	]	
Improved ChAID (Tschuprow Goodness o	f Split)					

Only the main window, the data grid and the project explorer are visible.

🔉 Sipina Research Version - [Le	arning set	editor]						X
🔉 File Edit Data Statistics Induc	tion method	Analysis Vie	w Window	Help			- 1	3 ×
<u>r</u> 🖉 📲								
X		age	chest_pain	rest_bpress	blood_sugar	rest_electro	max_heart	r 🔨
Attribute selection	1	43.00	asympt	140.00	f	normal	135.00	
	2	39.00	atyp_angina	120.00	f	normal	160.00	
	3	39.00	non_anginal	160.00	t	normal	160.00	
	4	42.00	non_anginal	160.00	f	normal	146.00	
	5	49.00	asympt	140.00	f	normal	130.00	
	6	50.00	asympt	140.00	f	normal	135.00	
	7	59.00	asympt	140.00	t	left_vent_hy	119.00 x	
	8	54.00	asympt	200.00	f	normal	142.00	
1	9	59.00	asympt	130.00	f	normal	125.00	
Learning method	10	56.00	asympt	170.00	f	st_t_wave_a	a122.00	
MethodName=Improved ChAID (Tsc	11	52.00	non_anginal	140.00	f	st_t_wave_a	a170.00	
Hdl=8	12	60.00	asympt	100.00	f	normal	125.00	
Merge=0.05	13	55.00	atyp_angina	160.00	t	normal	143.00	
Split=0.001	14	57.00	atyp_angina	140.00	t	normal	140.00	
I ypeBonterroni=1 ValueBonterroni=1	15	38.00	asympt	110.00	f	normal	166.00	
Sampling=0	16	60.00	non_anginal	120.00	f	left_vent_hy	135.00	
Examples selection	17	55.00	atyp_angina	140.00	f	normal	150.00	
209 examples selected	18	50.00	asympt	140.00	f	st_t_wave_a	a140.00	~
0 examples idle	<							>
	Editina N	NEW.FDM Attribut						E
Improved ChAID (Tschuprow Goodness o	f Split)					,		

#### 4.1.1 Selecting the variables of the analysis

In order to defining the target and the input attributes, we select the ANALYSIS / DEFINE CLASS ATTRIBUTE menu. In the dialog box, we set DISEASE as CLASS (TARGET), the others as ATTRIBUTES (INPUT).

🔉 Sipina Research Version - [Le	arning set	editor]					
🔉 File Edit Data Statistics Induc	tion method	Analysis 🕚	View Window	Help			- 8 ×
🖸 🗎 🎥 🕨 🌬		Define c Select a	lass attribute ctive examples	ıpre	ss.blood_su	ugar rest_electro	max_heart_r
Attribute selection	1	Set weig Set prior	ght field rs		f f	normal	135.00
Learning method MethodName=Improved ChAID (Tsc MethodClassName=TArbreDecision HdI=8 Merge=0.05 Split=0.001 TypeBonferroni=1 ValueBonferroni=1 ValueBonferroni=1 Sampling=0 Examples selection 209 examples selected 0 examples idle	3     Attribute       4     Attribute       5     Class       6     disease       7     Attribute       8     Attribute       9     chest_p       11     rest_p       12     max_he       14     exercice       15     0       16     0       17     0       18     0       10     0       11     0       12     0       14     0	e selection e selection main ress ugar ctro art_rate a_angina hly discrete uly continuou th			Varia age ches rest max exerr dised	ibles t_pain bpress d_sugar electro heat_rate cice_angina ase	
Improved ChAID (Tschuprow Goodness o	S S						Annuler

The selection is now visible in the top part of the project explorer. The letter "C" pinpoints a continuous attribute, "D" a discrete variable. The target attribute must be discrete.

🔉 Sipina Research Version - [Le	earning set	editor]						×
🔉 File Edit Data Statistics Induc	tion method:	Analysis Vie	w Window	Help			- 6	'×
<u>Ŭ </u> 🐂 h 🖒								
×		age	chest_pain	rest_bpress	blood_sugar	rest_electro	max_heart_	r 🔺
Attribute selection	1	43.00	asympt	140.00	f	normal	135.00	
E 🗊 Class attribute	2	39.00	atyp_angina	120.00	f	normal	160.00	
disease	3	39.00	non_anginal	160.00	t	normal	160.00	
Predictive attributes	4	42.00	non_anginal	160.00	f	normal	146.00	
age Dispect pain	5	49.00	asympt	140.00	f	normal	130.00	
	<u> </u>	50.00	asympt	140.00	f	normal	135.00	
blood sugar	7	59.00	asympt	140.00	t	left_vent_hy	119.00	
	8	54.00	asympt	200.00	f	normal	142.00	
max heart rate	9	59.00	asympt	130.00	f	normal	125.00	
D exercice angina	10	56.00	asympt	170.00	f	st_t_wave_a	122.00	
	11	52.00	non_anginal	140.00	f	st_t_wave_s	170.00	
Learning method	12	60.00	asympt	100.00	f	normal	125.00	
MethodName=Improved ChAID (Tsc 🔨	13	55.00	atyp_angina	160.00	t	normal	143.00	
MethodClassName=TArbreDecisionI	14	57.00	atyp_angina	140.00	t	normal	140.00	
Hdl=8 Merce=0.05	15	38.00	asympt	110.00	f	normal	166.00	
Split=0.001	16	60.00	non_anginal	120.00	f	left_vent_hy	135.00	
Examples selection	17	55.00	atyp_angina	140.00	f	normal	150.00	
209 examples selected	18	50.00	asympt	140.00	f	st_t_wave_a	140.00	~
0 examples idle	<	•					>	
	Editing	NEW.FDM				Attribut	es:8	E
Improved ChAID (Tschuprow Goodness o	f Split)					,		

#### 4.1.2 Tree induction

We click on the ANALYSIS / LEARNING menu. The learning phase is finalized.

🔉 Sipina Research Version - [Le	earning set	editor]								
🔉 File Edit Data Statistics Indu	tion method:	Analysis Vie	w Window	Help				-	a ×	
<u>ð</u> 🖹 🐂 🖒		Define class attribute Select active examples								
Athlin de velection			· ·		press	blood_sugar	rest_electro	o max_hear	<u>_r[ ^</u>	
	1	Set weight	held		þ	f	normal	135.00		
	2	Set priors.			)	f	normal	160.00		
	3	Set costs		þ	t	normal	160.00			
Predictive attributes	Set positiv	e class value.		þ	f	normal	146.00			
age	5	Learning			)	f	normal	130.00		
	6	Stop apaly	-		)	f	normal	135.00		
	7					t	left_vent_h	yk119.00	_	
Diobo_sugar	8	Classificati	on	•	)	f	normal	142.00	_	
may heart rate	9	Test			)	f	normal	125.00	_	
	10	LIFT ROC curve			)	f	st t wave	a122.00	_	
	11				)	f	st t wave a170.00			
Learning method	12	Error meas	urements	•	5	f	normal	125.00	-	
MethodName=Improved ChAID (Tsc A	13		•		)	t	normal	143.00	-	
MethodClassName=TArbreDecision	14	Feature se	ection	<u> </u>	5	t	normal	140.00	-	
Hdl=8	15	Personnal	tests	•	6	f	normal	166.00	-	
Merge=0.05 Solit=0.001	16	160.00	non andinal	120.0	Γ Π	f	left vent h	vr135.00	-	
	17	55.00	atvo angina	140.0	- n	f	normal	150.00	-	
Examples selection	18	50.00	asymnt	140.0	0	f	st t wave	a140.00	-	
209 examples selected		100.00	ore y mpt	140.0	0		5 <u>_</u> ++4+6	_4140.00	<b>N</b>	
o examples luie										
1	Editing NEW.FDM Attributes : 8									
Improved ChAID (Tschuprow Goodness c	f Split)									

The tree is now displayed. The distribution of the values of DISEASE is available on each node.



We want to analysis the leaf at the last level of the tree. The corresponding prediction rule is « **IF** CHEST\_PAIN = (ASYMPT *OR* TYP\_ANGINA) **AND** EXERCICE\_ANGINA = YES **THEN** DISEASE = YES ». This group is defined by two variables. But, what about the other variables of the dataset? Are they really irrelevant for the characterization of this subpopulation? Computing descriptive statistics enables to answer to this question.

### 4.2 Node exploration - Elementary statistics

Each node of the tree matches to a subsample. The root node constitutes the whole dataset. In order to obtain descriptive statistics of the observations on a node, we select the node. Then we activate the contextual menu (right click). We select the NODE INFORMATIONS menu item. A new window appears. We observe the goodness-of-split for each predictive variable, the number of examples, some descriptive statistics, etc. **We select the CHARACTERIZATION tab**.



**CONTINUOUS ATTRIBUTES.** For each continuous attribute, we compare the local average (i.e. the mean of the variable for the subsample) and the global average (i.e. the mean of the variable for the whole dataset). In order to characterize the importance of the deviation, we compute also the t-test statistic (STRENGTH indicator) for a comparison of mean. It is not really a test because the samples are not independent. But it enables to order the variables according the importance of the difference. Indeed, the variables are not measured in the same unit and/or scale, the STRENGTH indicator can be understood as a normalized deviation. In this example, the mean of "age" for the whole dataset is 47.9. For the subgroup corresponding to the node, it is 50.76.

Informations on : Level 3, Node 1 IF chest_pain in [asymangina] and exercice_angina in [yes] Characterization Desciptors' importance Continuous attributes Attribute Strength Local Avg Global Avg age 3.19 50.7667 47.9665 rest_bpress 3.01 139.3833 133.6603 max_heart_rate -6.70 120.1333 137.5742
IF chest_pain in [asym
Attribute         Strength         Local Avg         Global Avg           age         3.19         50.7667         47.9665           rest_bpress         3.01         139.3833         133.6603           max_heart_rate         -6.70         120.1333         137.5742
Attribute         Strength         Local Avg         Global Avg           age         3.19         50.7667         47.9665           rest_bpress         3.01         139.3833         133.6603           max_heart_rate         -6.70         120.1333         137.5742
age         3.19         50.7667         47.9665           rest_bpress         3.01         139.3833         133.6603           max_heart_rate         -6.70         120.1333         137.5742
rest_bpress         3.01         139.3833         133.6603           max_heart_rate         -6.70         120.1333         137.5742
max_heart_rate -6.70 120.1333 137.5742
CO

**DISCRETE ATTRIBUTES.** We compute a statistical indicator for the comparison of proportion here.

Info	ormations on : Level 3	, Node 1							
IF (	F chest_pain in [asytyp_angina] and exercice_angina in [yes]								
UI	haracterization   Desciptors' importance								
0	Continuous attributes								
Г									
		exercice_	angina ( U.44	14 J					
	Values	Strength	Local Dist.	Global Dist.	Recall				
	yes	12.62	60 (100%)	72 (34%)	83%				
	no	-12.62	0 (0%)	137 (66%)	0%				
	chest_pain ( 0.2971 )								
	Values	Strength	Local Dist.	Global Dist.	Recall				
	asympt	9.37	60 (100%)	102 (49%)	59%				
	atyp_angina	-6.15	0 (0%)	65 (31%)	0%				
	non_anginal -4.17 0 (0%) 36 (17%) 0%								
	typ_angina -1.57 0 (0%) 6 (3%) 0%								
		disea	se ( 0.1953 )						
	Values	Strength	Local Dist.	Global Dist.	Recall				
	positive	8.48	54 (90%)	92 (44%)	59%				
	negative	-8.48	6 (10%)	117 (56%)	5%				
		rest ele	ctro ( 0.0119	1					
	Values	Strength	Local Dist.	Global Dist.	Recall				
	normal	-2.00	45 (75%)	173 (83%)	26%				
	left_vent_hyper	-0.44	1 (2%)	5 (2%)	20%				
	st_t_wave_abnormality	2.32	14 (23%)	30 (14%)	47%	T			
Ľ									
	60	examples (2	8.71% of the lea	arning set)					

The variable REST\_ELECTRO is really interesting. It is not visible in the tree. So it seems irrelevant. But when we compare the proportions, we observe that there is an over representation of the value ST\_T\_WAVE\_ABNORMALITY for this subgroup. In the whole dataset, 14% of the examples have this characteristic. They are 23% for the examples related to the node.

The RECALL indicator says that 47% of the examples "REST\_ELECTRO = ST\_T\_WAVE\_ABNORMALITY" are located on this subgroup.

An additional indicator is used (J-MEASURE) in order to organize the variables. It has not really a valuable interpretation in our context.

### 4.3 Node exploration – Descriptive statistics

These comparative descriptive statistics are directly available. But they are mainly univariate. If we want to deeply analyze a subpopulation, it is (maybe) useful to compute the detailed descriptive statistics (univariate or bivariate) which were outlined previously (see section 3.2 and section 3.3).

SIPINA enables to compute the previous descriptive statistics on each node. Of course, the computation is restricted to the covered examples i.e. the subpopulation highlighted by the node.

Let us repeat the same analysis than previously (see section 3.2 and section 3.3). But the calculations are now restricted to the sample corresponding to the rule "CHEST\_PAIN = (ASYMPT  $OR TYP_ANGINA$ ) **AND** EXERCICE\_ANGINA = YES".

#### 4.3.1 Univariate statistics

Compared with the preceding tool (section 4.2), this functionality is not really useful for the univariate statistics. We obtain the same results.

**Continuous variables.** In order to obtain the descriptive statistics related to a node. We select first the node. Then we activate the contextual menu (right click). We select the OTHER DESCRIPTIVE STATISTICS / UNIVARIATE menu item. In the dialog box, we observe that "COVERED EXAMPLES" option is activated. Only the 60 examples related to the node are used for the statistical computation.



In the result window, we obtain all the descriptive indicators for each variable. We can compare these values with those computed for the whole dataset (see section 3.2.1).

🔉 Sipina Research Version										
Induction method Analysis View Wir	ndow Help									
📉 🖄 📴 🐜 🖿										
	N Decision to									
Attribute selection	KN Decision (	ee								
🖃 🗊 Class attribute						🔒 tro Im	ax heart réa	exercice anddi	isease	
D disease						- 1:	35.00	/es p	ositive	
🖻 📭 Predictive attributes				92 (4	14%)	16	60.00 y	/es n	egative	
age				chest na	in	10	60.00 r	no or	egative	
chest_pain		in [asympt]	typ_angina	1		14	46.00 r	no or	egative	
blood sugar		79 (73%	)			1:	30.00 r	no or	egative	
D rest electro		29 (278				1:	35.00 r	no n	egative	
max heart rate	in [yes]		in [no]		in [yes	1 _hyk1*	19.00 y	yes p	ositive	_
exercice_angina	54 (90%)		25 (52	*)	6 1	(SC 14	42.00 y	/es p	ositive	
_	6 (10*)		23 (48	*)	6	(30 🖌 1)	25.00 r	no p	ositive	
		🔉 Descrip	tive statis	tics results	on contin	uous varia	bles			
	11 5.	Descriptive s	tatistics on	NEW. FIM						1
	12 0	Only selected	i examples							
Learning method	10 5									
MethodName=Improved ChAID (Tsc 🔨	15 3									
MethodClassName=TArbreDecisionI	16 6									
Hdl=8 Merge=0.05	17 5:					[	1		[	
Split=0.001	18		Examples	Min	Max	Mean	Std	Mean/Std	Skewness	
Examples selection	19	rest bpre	ьU СО	92.0000	200.0000	139.3833	19.1622	7.2739	0.4264	
209 examples selected	20	max heart	60	82.0000	158.0000	120.1333	jio.5413	0.4792	-0.0580	
0 examples idle	21 6								>	~
	<									>
Improved ChAID (Tschuprow Goodness of	Split)								Time	78 /

**Discrete variables.** We follow the same approach for the discrete variables. We activate the OTHER DESCRIPTIVE STATISTICS / UNIVARIATE menu item in the contextual menu. We select the DISCRETE VARIABLE tab of course.

🔉 Sipina Research Version		
Induction method Analysis Tree man	nagement View Window Help	
<u>Ს </u> 👹 🖊 🖻		
Attribute selection	Solution tree      Tro max_heart_riexercice_an(disease     135.00 yes positive     135.00 yes positive     160.00 yes     positive     teo.00 yes     positive	
rest_bpress     blood_sugar     rest_electro     max_heart_rate     exercice_angina	In       12 symptor voyo       Continuous variables       Discrete variables         In       12 grade       (274)       Statistics         In       12 grade       Im       Mode         In       12 grade       Im       Im         S4       (204)       Im       Im         S4       (204)       Im       Im         S4       (204)       Im       Im         Mode       Im       Im       Im         S4       (204)       Im       Im	=s
Learning method MethodName=Improved ChAID (Tsc MethodClassName=TArbreDecisionI	10     Split node       14     Explore the node       15     Explore the node       00     Other description statistics	Annuler
Hdl=8 Merge=0.05 Split=0.001	Do         Contraction         Diversion           17         6         Other SIPINA session         Bivariate           18         5         rest bpre 60         92.0           20         3         nax heart 60         82.0   Conditional State Mean/State Stewness State Stewness Stewne	
U examples idle Improved ChAID (Tschuprow Goodness o	21         66         2	

We obtain the distribution of values for each variable. We can compare these results with those obtained for the whole dataset (see section 3.2.2).



In this case, because CHEST\_PAIN and EXERCICE\_ANGINA are involved in the tree, only the values used in the path appear in the distribution.

#### 4.3.2 Bivariate statistics

This tool is useful for the bivariate statistics. We can analyze the association between some variables in each subgroups related to the nodes of the tree. Then we can carefully characterize each subpopulation. This functionality is essential when the interpretation of the results is at least as significant as the accuracy of the rules.

**Scatter plot.** We want to repeat the previous analysis (see section 3.3.1). In the contextual menu, we select the OTHER DESCRIPTIVE STATISTICS / BIVARIATE / SCATTERPLOT option.



The scatter plot is restricted to the 60 examples related to the node. It is different from the previous result (see section 3.3.1). The variables are not any more correlated in this configuration.



**Other statistical indicators.** In the same way, it is possible to find the tools highlighted previously (see sections 3.3.2, 3.3.3 et 3.3.4). E.g. MAX\_HEART\_RATE distribution according the DISEASE (see section 3.3.3 for the result on the whole dataset).



**Note:** Actually, it is possible to make these operations on any statistical software. It is simply necessary to make a query and to compute the statistical indicators on the subpopulation. The main interest of SIPINA is to automate all intermediate operations which, when they are repetitive, can end up quickly boring. This shortcut is very useful in practice.

## 5 Subsample related to a node

When we wish to refine the results, it can be necessary to go back on the data, notably to analyze in a deepened way the subpopulations described by the nodes of the tree. When the observations are recognizable (each case is associated to a label), we can even distinguish each individual.

In order to obtain the detailed description of the dataset, we select the node; we activate the EXPLORE THE NODE option in the contextual menu. The subsample is displayed in a new window.

Sipina Research Version								
Induction method Analysis Tree mana	igement View Window Help							
т 🔉 📴 🐜 陷								
	N Decision tree							
Attribute selection	Decision nee			<b>a</b>				
🖃 🗊 Class attribute								
D disease	92 (448)	_		$\mathbf{X}$				
Predictive attributes	117 (56%)	ate						
D chest_pain	in [asympt, typ angina]	~		f				
C rest_bpress	79 (73*)							
	29 (27%)	heart_rate given	i dis 💶 🗖	$\mathbf{X}$				
max heart rate	in [yes] in [no] in [yes]		_		_			
exercice_angina	54 (90*) 25 (52*) 5 5 (10*) 22 (48*) 5	🔝 Subsample on L	evel 3, Node 1	1				
	► (10*) ► Level 3, Node 1	Local examples						
	11 Node informations							
	12 110 disease Cut	- Attributes	IF chest_pain in	n (asympt,typ_angina) an	d exercice_ar	ngina in (yes)		
	105 - Split node	- age						
	15 100 posit Explore the node	✓ chest_pain ✓ rest_boress					1	
I aming well-ad	16 95 negat Other descriptive statistics	✓ blood_sugar	ag	je chest_pain	rest_bpress	blood_sugar	rest_electro	max_heart_
MethodName=Improved ChAID (Tsc	17 85 Other STPINA session	✓ rest_electro ✓ max heart rate	7 50	3.00 asympt	140.00		Inft went have	130.00
MethodClassName=TArbreDecisionI		exercice_angina	8 54	4.00 asympt 4.00 asympt	200.00	1 4	normal	142.00
Hd=8 Merge=0.05	19 100 2 20 Parl antign (99.29.92	- disease	10 50	6.00 asympt	170.00	f	d t wave abr	172.00
Split=0.001	20 Real position . (30. 30,02. 0		18 50	0.00 asympt	140.00	r f	st t wave abr	140.00
Examples selection	22 56.00 as		21 66	8.00 asympt	140.00	f	normal	94.00
209 examples selected	23 44.00 as	4	22 66	8.00 asympt	155.00	t	normal	150.00
U examples luie	24 42.00	-	24 43	3.00 asympt	120.00	f	normal	120.00
Improved ChAID (Tschuprow Goodness of	í Solit)	-	25 54	4.00 asympt	140.00	f	normal	118.00
Improved crime (recipion de sector	spicy	-	32 54	4.00 asympt	130.00	f	normal	91.00
			33 48	8.00 asympt	160.00	f	normal	92.00
			34 38	8.00 asympt	110.00	f	normal	150.00
			36 46	δ.00 asympt	120.00	f	normal	115.00
			40 49	asympt	140.00	f.	normal	140.00
			41 05	5.00 asympt	170.00	t	normal	112.00
			43 00	5.00 asympt 5.00 asympt	140.00	t 4	normal	130.00
			66 50	6.00 asympt	150.00	4 4	et t wave abr	124.00
			75 31	1.00 asympt	120.00	t.	normal	153,00
			79 42	3.00 asympt	150.00	t.	normal	130.00
			86 65	5.00 asympt	140.00	f	normal	110.00
			88 48	8.00 asympt	160.00	f	normal	103.00
			98 46	5.00 asympt	110.00	f	normal	150.00
			100 48	8.00 asympt	160.00	f	normal	102.00 🗸
			<					>
			Attributes : 8	Examples : 60				

The subsample can be saved in a new file (\*.fdm file format).

😹 Subsample on L	evel 3, Nod	e 1					×
Local examples							
Save covered exan	IF chest_pai	n in [asympt,t	yp_angina) an	id exercice_a	ngina in (yes)		
✓ rest_bpress		age	chest_pain	rest_bpress	blood_sugar	rest_electro	^
✓ blood_sugar ✓ rest_electro	1	43.00	asympt	140.00	f	normal	
max_heart_rate	7	59.00	asympt	140.00	t	left_vent_hype	
<ul> <li>exercice_angina</li> <li>disease</li> </ul>	8	54.00	asympt	200.00	f	normal	
	10	56.00	asympt	170.00	f	st_t_wave_abn	
	18	50.00	asympt	140.00	f	st_t_wave_abn	
	21	66.00	asympt	140.00	f	normal	
	22	56.00	asympt	155.00	t	normal	
	24	43.00	asympt	120.00	f	normal	
	25	54.00	asympt	140.00	f	normal	~
		•				>	
	Attributes : 8	Exam	ples : 60				

The data file is automatically named with the identifier of the node. It is placed in the same directory as the source dataset.

## 6 A new analysis of a subpopulation

In certain situations, we want to launch a new analysis on a subpopulation related to a node of the tree. For instance, we want to explain/predict the REST\_ELOECTRO variable for the subsample described by the rule "CHEST\_PAIN = (ASYMPT *OR* TYP\_ANGINA) **AND** EXERCICE\_ANGINA = YES".

With the contextual menu, we select the OTHER SIPINA SESSION option.



A new SIPINA session is launched. The subsample (8 variables and 60 examples) is automatically downloaded.

Sipina Research Version - [Learning set editor]									
💦 File Edit Data Statistics Induc	tion method	Analysis Vie	w Window	Help					- 8 ×
🖸 🗎 🎥   h 🖡									
X		age	chest_pain	rest_bpress	blood_sugar	rest_electro	max_heart_r	exercice_an	dises 🔨
Attribute selection	1	43.00	asympt	140.00	f	normal	135.00	yes	positi
Learning method MethodName=Improved ChAID (Tsc A MethodClassName=TArbreDecision HdI=8 Merge=0.05 Split=0.001 TypeBonferroni=1 ValueBonferroni=1 Sampling=0	7	59.00	asympt	140.00	t	left_vent_hy	119.00	yes	positi
	8	54.00	asympt	200.00	f	normal	142.00	yes	positi
	10	56.00	asympt	170.00	f	st_t_wave_a	122.00	yes	positi
	18	50.00	asympt	140.00	f	st_t_wave_a	140.00	yes	positi
	21	66.00	asympt	140.00	f	normal	94.00	yes	positi
	22	56.00	asympt	155.00	t	normal	150.00	yes	positi
	24	43.00	asympt	120.00	f	normal	120.00	yes	positi
	25	54.00	asympt	140.00	f	normal	118.00	yes	positi
	32	54.00	asympt	130.00	f	normal	91.00	yes	positi
	33	48.00	asympt	160.00	f	normal	92.00	yes	positi
	34	38.00	asympt	110.00	f	normal	150.00	yes	positi
	36	46.00	asympt	120.00	f	normal	115.00	yes	positi
	40	49.00	asympt	140.00	f	normal	140.00	yes	positi
	41	65.00	asympt	170.00	t	normal	112.00	yes	positi
Examples selection	43	65.00	asympt	140.00	t	normal	87.00	yes	positi
60 examples selected	49	55.00	asympt	140.00	f	normal	130.00	yes	positi 🞽
0 examples idle									
	Editing C:\Documents and Settings\Maison\Mes documents\NEW_No Attributes : 8 Examples : 60								
Improved ChAID (Tschuprow Goodness of Split)									

We define again the TARGET and the INPUT variables. Then we select the adequate parameters of the learning algorithm. We obtain, for instance, the following classification tree.



# 7 Conclusion

In this tutorial, we wanted to describe the descriptive statistics tools of SIPINA. These features are not really extraordinary. But combined with the interactive exploration of a tree of decision, they turn out very productive.