# Subject

#### Interactive induction of decision trees with SIPINA.

Various functionalities of SIPINA are not documented. In this tutorial, we show how to explore nodes of a decision tree, in order to obtain a better understanding of the characteristics of the subpopulation on a node. This is an important task, for instance when we want to validate the rules with an expert domain.

## Dataset

We use the <u>BLOOD\_PRESSURE\_LEVELS.XLS</u> dataset, available on-line<sup>1</sup>. There are 399 examples. We want to predict the high blood pressure of a patient from their characteristics (smoke, sex, overweight, etc.). Here is the description of the attributes.

Attribute	Category	Informations
bpress_level	Discrete	2 values
gender	Discrete	2 values
smoke	Discrete	2 values
exercise	Continue	-
overweight	Continue	-
alcohol	Continue	-
stress	Continue	-
salt	Continue	-
income	Continue	-
education	Continue	-

We adopt a descriptive framework. We do not want to obtain the most accurate classifier, but rather to characterize the (groups of) individuals. In this point of view, the assessment relies mainly on the expertise. This is the physician who may confirm us if rules proposed by the tree are in adequacy with the domain knowledge or not. For this reason, we use all the dataset for the tree induction. The accuracy rate (or error rate) is not really a pertinent rating in this context.

# Induction of a decision tree using SIPINA

#### Data importation

The SIPINA add-in (SIPINA.XLA) is automatically installed on the hard disk. The simplest way to send the dataset from EXCEL to SIPINA is to use this add-in (see « SIPINA ADD-IN FOR EXCEL SPREADSHEET » on the website <a href="http://eric.univ-lyon2.fr/~ricco/sipina\_download.html">http://eric.univ-lyon2.fr/~ricco/sipina\_download.html</a>).

We select the range of cells, we activate the SIPINA / EXECUTE menu which is now available in EXCEL. A dialog box appears, we check if the selection is right, then we confirm by clicking on OK. Note, the first row of the cells range must correspond to the name of the attributes.

<sup>&</sup>lt;sup>1</sup> <u>http://eric.univ-lyon2.fr/~ricco/dataset/blood\_pressure\_levels.xls</u> ; the original source of this dataset is <u>http://www.math.yorku.ca/Who/Faculty/Ng/ssc2003/BPMain.htm</u>



SIPINA is started automatically. The data are transferred through the clipboard. The number of variables and the number of observations are displayed in the status bar (10 variables and 399 observations). Note: SIPINA does not handle missing data, we must treat them before.

S File Edit Data Statistics Indu	iction metho	d Analysis	View Windo	w Help					- 6	1 ×
	KI F	line a	-	I	I	1	Linder	Station.	1.000	100
Attribute selection		bpress_ie	vel gender	smoke	exercise 4.00	overweight	alcohol	stress	sait	- ^
	1	nign	temale	no	1.00	1.00	1.00	2.00	1.00	-
	2	normal	female	no	3.00	1.00	1.00	3.00	2.00	
	3	normal	Temale	yes	2.00	1.00	2.00	2.00	2.00	
	4	normal	formele	no	2.00	3.00	2.00	3.00	4.00	
	5	normal	female	yes	3.00	1.00	2.00	2.00	1.00	
	b	norma	female	yes	2.00	1.00	3.00	1.00	2.00	
	/	nign	temale		2.00	3.00	3.00	1.00	1.00	
	8	normai	temale	no	3.00	1.00	2.00	1.00	3.00	
1	9	normai	temale	no	2.00	1.00	1.00	3.00	1.00	
	10	high	temale	no	1.00	1.00	3.00	3.00	2.00	
	11	normai	male	no	3.00	1.00	1.00	3.00	2.00	
	12	high	male	yes	1.00	1.00	3.00	2.00	3.00	
	13	high	female	yes	1.00	3.00	3.00	1.00	2.00	
1	14	high	male	no	3.00	1.00	3.00	3.00	3.00	
	15	high	female	no	1.00	3.00	3.00	1.00	1.00	
1	16	high	female	yes	1.00	1.00	2.00	2.00	1.00	
	17	normal	female	no	1.00	1.00	1.00	2.00	3.00	1
1	18	high	male	no	1.00	2.00	1.00	2.00	1.00	
	19	high	female	yes	1.00	2.00	3.00	2.00	1.00	
1	20	normal	male	no	2.00	1.00	1.00	2.00	3.00	
l	21	high	male	yes	3.00	3.00	1.00	3.00	2.00	
MathadNamo-Improved ChâID (Tac A	22	normal	male	yes	3.00	1.00	2.00	3.00	3.00	
MethodClassName=TArbreDecision	23	high	female	yes	1.00	1.00	2.00	2.00	3.00	
Hdl=8	24	normal	male	no	1.00	1.00	3.00	3.00	1.00	
Merge=0.05	25	normal	female	yes	3.00	2.00	1.00	1.00	3.00	
/Split=0.001 TupeBonferroni=1	26	normal	male	no	1.00	1.00	2.00	1.00	2.00	
ValueBonferroni=1	27	high	male	no	1.00	1.00	2.00	1.00	2.00	
Sampling=0	28	high	female	no	1.00	2.00	2.00	1.00	1.00	
Examples selection	29	normal	male	no	3.00	1.00	1.00	2.00	2.00	
399 examples selected 0 examples idle	30	normal	female	no	3.00	1.00	1.00	1.00	2.00	1
	Editing	NEW.FDM				Attribu	tes : 10	Examples :	399	Ľ

### Choosing the learning method

The first step is to select the induction algorithm. We activate the INDUCTION METHOD / STANDARD ALGORITHM menu. A dialog box appears with the list of available methods.

🔉 File Edit Dat	a Statistics Ind	auccion meche	e mayors m						
0 🛯 🖳 🐴		Standard alg	orithm	_					
		×	bpress_leve	gender	smoke	exercise	overweight	alcohol	st ^
Attribute selection		_ 1	high	female	no	1.00	1.00	1.00	2.
		2	normal	female	no	3.00	1.00	1.00	3.
		3	normal	female	yes	2.00	1.00	2.00	2.
	Colored and the days	den and be d	la sur st	hard a	12143	la col	1.00	200	3.
	Select an induc	tion method				1 1	<b>,</b>		2.
	Induction Graph   F	Rule Induction   1	Neural network   Dis	oriminant analys	is   Decision list	Other	1		1.
	A limited search ind ID3-IV (Quinlan - 19	uction tree algorit 186)	hm (Catlett - 1991)			-	/		1.
	GID3 (Cheng, Fayy	ad, Irani & Qian -	1988) o t Protko - 1996)						1.
	ASSISTANT 86 (Cestnik, Kononenko & Bratko - 1986) ChAID (Kass - 1980)							3.	
	LINAID (Kass - 1980	0							_
	C4.5 (Quinlan - 1980 Improved C4.5 (Rai	l) 3) «otomalala & Lalli	ch - 1999)	6-					3.
	ChAID (Kass - 1980 C4.5 (Quinlan - 1990 Improved C4.5 (Rail Improved ChAID (1) Cost sensitive C4.5	l) 3) kotomalala & Lalliv schuprow Goodn (Bakotomalala &	ch - 1999) ess of Split) Chauchat - 2001)	<u></u> -					3. 3.
	ChaiD (Kass - 1980 C4.5 (Quinlan - 199 Improved C4.5 (Rai Improved ChAID (1 Cost sensitive C4.5 One-Vs-All Decision	l) 3) kotomalala & Lalliu <del>schuprow Goodn</del> (Rakotomalala & 1 Tree	ch - 1999) ess of Split) Chauchat - 2001)	<u></u>					3. 3. 2.
	ChaiD (Kass - 1980 C4.5 (Quinlan - 1993 Improved C4.5 (Rai Improved ChAID (T Cost sensitive C4.5 One-Vs-All Decision	l) 3) kotomalala & Lalli schuprow Goodn (Rakotomalala & 1 Tree	ch - 1999) ess of Split) Chauchat - 2001)	<u></u>					3. 3. 2. 1.
earning method	ChaiD (Kass - 1980 C4.5 (Quinlan - 199 Improved C4.5 (Rai Improved ChAID (T Cost sensitive C4.5 One-Vs-All Decision	)) 3) schuprow Goodn Schuprow Goodn (Rakotomalala & Tree	ch - 1999) ess of Split) Chauchat - 2001)	£-					3. 3. 2. 1. 3.
Learning method	CA5 (Quinlan - 1980) Improved CA5 (Rai Improved CA5 (Rai Improved CA5 (Rai Cost sensitive CA5 One-Vs-All Decision	)) 3) cotomalala & Lalli <del>schuprow Goodr</del> (Rakotomalala & Tree	ch - 1999) ess of Split) Chauchat - 2001) Improved ChAID	(Tschuprow	Goodness of S	plit)			3. 3. 2. 1. 3. 1.
Learning method MethodName=Impro MethodClassName=	Chalu (Kass - 1980) C4.5 (Quinan - 199 Improved C4.5 [Rah Improved ChAlD (T Cost sensitive C4.5 One-Vs-All Decision	i) 3) sotomalala & Lalli sohuprow.Goodr (Rakotomalala & Tree	ch - 1999) ess of Split) Chauchat - 2001) Improved ChAID	(Tschuprow	Goodness of S	:plit)	akre X Ar	nnuler	3. 3. 2. 1. 3. 1. 2.
Learning method MethodName=Impro MethodClassName= HdI=8 Metro=0.05	Chall (Kass - 1984) C4.5 (Quintan - 199 Improved C4.5 (Ra) Improved C4.5 (Ra) Cost sensitive C4.5 One-Vs-All Decision	)) 3) cotomalala & Lallii schuprow Goodr (Rakotomalala & Tree	ch - 1999) ess of Split) Chauchat - 2001) Improved ChAID	(Tschuprow	Goodness of S	iplit)	QK.4. X Ar	muler	3. 3. 2. 1. 3. 1. 2. 2. 2. 2.
Learning method MethodName=Impro MethodClassName= HdI=8 Merge=0.05 Split=0.001	Chall (Kass - 1984) C4.5 (Quintan - 199 Improved C4.5 (Ra) Improved C4.5 (Ra) Cost sensitive C4.5 One-Vs-All Decision	) 3) cotomalala & Lalli cotomalala & Lalli cotomalala & Tree 18 19	ch - 1999) ess of Split Chauchat - 2001) Improved ChAID	(Tschuprow ( male female	Goodness of S	pik)	2.00 2.00	nuler	3. 3. 2. 1. 3. 1. 2. 2. 2. 2. 2.
Learning method MethodName=Impr MethodClassName= Hdl=8 Merge=0.05 Split=0.001 TypeBonferroni=1	Chalu (Kass - 1984) C4.5 (Quintan - 199 Improved C4.5 (Fab Improved C4.5 (Fab Cost sensitive C4.5 Dne-Vs-All Decision	) 3) cotomalala & Lalli cotomalala & Lalli cotomalala & Tree 18 19 20	ch + 1999) ess of Spill Chauchat + 2001) Improved ChAID high high normal	(Tschuprow ( male female male	Goodness of S no yes no	pik)	2.00 1.00	nnuler 1.00 3.00 1.00	3. 3. 1. 3. 1. 2. 2. 2. 2. 2. 2. 2.
Learning method MethodName=Impro MethodClassName= Hdl=8 Werge=0.05 Split=0.001 TypeBonferroni=1 ValueBonferroni=1 ValueBonferroni=1	Linau (Lass - 1980) (Improved C.4.5 (Quinter Improved C.4.5 (Quinter Improved C.4.5 (Quinter Cast sensitive C.4.5 One-Vs-AI Decision	)) 3) cotomalala & Lalii Soluurow Goodr (Rakotomalala & Tree 18 19 20 21	ch - 1999) ess of Split Chauchat - 2001) Improved ChAID Nigh high normal high	(Tschuprow ( male female male male	Goodness of S no yes no yes	pik)	2.00 1.00 3.00	nuler 1.00 3.00 1.00	3. 3. 1. 3. 1. 2. 2. 2. 2. 2. 2. 3.
Learning method MethodName=Impro MethodClassName= HdI=8 Merge=0.05 Split=0.001 TypeBonferroni=1 ValueBonferroni=1 Sampling=0 Fixamples selection	Linki (Likas - 184) (Lak (Quinien - 193) Improved C4 S (Rich Improved C4 S (Rich Improved C4 S (Rich Cot ternitive C4 S Cot ternitive C4 S C C4	)) 3) cotomalala & Lalii Schuprov Goodr (Rakotomalala & 1 Tree	ch - 1999) ess of Spill Chauchat - 2001) Improved ChAlD high high normal	(Tschuprow ) male female male male male	Goodness of S no yes yes yes	plit)	2.00 1.00 3.00	nuler 1.00 3.00 1.00 1.00 2.00	3. 3. 2. 1. 3. 1. 2. 2. 2. 2. 2. 2. 3. 3.
Learning method MethodName=Impro MethodClassName= HdI=8 Merge=0.05 Split=0.001 TypeBonferroni=1 ValueBonferroni=1 Sampling=0 Examples selection 399 examples selection	Linki (Liss: 193) (Improved C4 5 (Faine) Improved C4 5 (Faine) Improved C4 5 (Faine) (Cost sensitive C4 5 (Cost sensitive C4 5 (Cost sensitive C4 5) (Cost	)) sotomalala & Lalii Schustow Goodr (Rakotomalala & Tree 18 19 20 21 22 23	ch + 1999) ess of Spint Cheuchat - 2001) Improved ChAlD ingh high high normal high high	(Tschuprow I male female male male female female	Goodness of S no yes no yes yes yes	pik) 1.00 1.00 2.00 3.00 3.00 1.00	2 00 1.00 1.00 1.00	nuler 1.00 3.00 1.00 2.00 2.00	3. 3. 2. 1. 3. 1. 2. 2. 2. 2. 2. 2. 3. 3. 3. 2. ▼
Learning method MethodName=Impro MethodClassName= HdI=8 Merge=0.05 Spilt=0.001 TypeBonferroni=1 ValueBonferroni=1 Sampleng=0 Examples selection 399 examples selection o examples idle	Linki (Lass - 184) CA5 (Quinien - 189 Improved CA5 (Filt Improved CA5 (Filt Improved CA5 (Filt Improved CA5 (Filt Improved CA5 (Filt Cost sensitive CA5 Decision	) cotomalala & Lalia Engrarow Ecologia (Rakotomalala & Tree 18 19 20 21 22 23 C m	ch - 1939) ess of Split Chauchat - 2001) Improved ChAID ingh high normal high normal high	Tschuprow ( Tschuprow ( female male male female female	no yes no yes yes yes yes	1.00 2.00 3.00 1.00 1.00	2.00 1.00 1.00 1.00	nuler 1.00 3.00 1.00 2.00 2.00	3. 3. 2. 1. 3. 1. 2. 2. 2. 2. 2. 2. 3. 3. 3. 3. 2. ▼

Various supervised learning algorithms are available. The most interesting ones are the decision tree methods. SIPINA can learn automatically the tree from the dataset. But it is also possible to interactively build the tree, guided by the domain knowledge. For most complete software, with other learning algorithms than trees, try **TANAGRA** which is also freely available (<u>http://eric.univ-lyon2.fr/~ricco/tanagra/</u>).

We select the IMROVED CHAID method. It is rather simple; it builds a short tree, useful in a first data exploration. We click on OK. A new dialog box appears, it enables us to set the parameters of the algorithm. We validate the default parameters.

Improved ChAID parameters	
Parameters Sampling Priors	
p-level for merging nodes : 0.05 for splitting nodes : 0.001 Bonferroni adjustments C Automatic Manual	<b>÷</b>
Other pruning parameters	
Max. depth : 🗧 🗲	
Min size of node to split : 20 🚖	
Min size of leaves : 10 🚖	
tum?	

The user's choice (method and parameters) are displayed in the middle part of the project explorer.

🕅 File Edit Data Statistics Indu 🗅 🗈 📴 🐂 🍋 🗈	ction metho	id Analysis Vie	nalysis View Window Help _ & >							
	1	bpress_level	gender	smoke	exercise	overweight	al			
Attribute selection	1	high	female	no	1.00	1.00	1			
	2	normal	female	no	3.00	1.00	1			
	3	normal	female	yes	2.00	1.00	2			
	4	normal	male	no	1.00	3.00	2			
	5	normal	female	yes	3.00	1.00	2			
	6	normal	female	yes	2.00	1.00	3			
	7	high	female	no	2.00	3.00	3			
	8	normal	female	no	3.00	1.00	2			
	9	normal	female	no	2.00	1.00	1			
	10	high	female	no	1.00	1.00	3			
	11	normal	male	no	3.00	1.00	1			
earning method	12	high	male	yes	1.00	1.00	з			
tethodName=Improved ChAID (Tsc 🔊	13	high	female	yes	1.00	3.00	3			
1ethodClassName=TArbreDecision	14	high	male	no	3.00	1.00	3			
Idl=8	15	high	female	no	1.00	3.00	3			
	16	high	female	yes	1.00	1.00	2			
ypeBonferroni=1	17	normal	female	no	1.00	1.00	1			
(alueBonferroni=1 🛛 📕 🌉	18	high	male	no	1.00	2.00	1			
ampling=U	19	high	female	yes	1.00	2.00	3			
Examples selection	20	normal	male	no	2.00	1.00	1			
99 examples selected examples idle	š	haran		1	0.00	0.00	>			

#### Class attribute and predictive attributes

In the next step, we must specify the class attribute (the attribute that we want to predict) and the predictive ones (the descriptors).

We activate the ANALYSIS/DEFINE CLASS ATTRIBUTE menu. A dialog box appears. Using drag and drop principle, we set BPRESS LEVEL as CLASS (TARGET), and the other attributes as ATTRIBUTES (INPUT).

Sipina Research Version - [Learn S File Edit Data Statistics Induction r	ing set editor] method Analysis View Window Help		×	
Attribute selection	Define class attribute Select active examples	exercise	overweight al	
	Set positive class value	1.00 3.00 2.00	1.00 1 1.00 1 1.00 2	
Learning method MethodName=Improved ChAID [Tsc MethodClassName=TArbreDecision] Hdl=8 Metge=0.05 Split=0.001 TypeBonferroni=1 YalueBonferroni=1 YalueBonferroni=1 YalueBonferroni=1	ass rress_level tributes ender recise rervise rervise rervise ress st come lucation		Variables bpress_level gender smoke exercise overweight alcohol stress salt income education	
Examples selection 399 examples selected 0 examples idle Improved ChAID (Tschuprow Goodness	0 Only discrete 0 Only continuous 8 Both			
			🗸 OK	🗱 🗶 Annuler

We validate the selection with OK. The user's choices are displayed in the top part of the project explorer.

🕻 File Edit Data Statistics Induct	tion method	Analysis Vie	w Windo	w Help		-	- 8 ×	
N 19 19 19 19 19 19 19 19 19 19 19 19 19	-	bpress_leve	gender	smoke	exercise	overweight	al	
attribute selection	1	high	female	no	1.00	1.00	1	
🗆 🗊 Class attribute	2	normal	female	no	3.00	1.00	1	
<ul> <li>D bpress_level</li> <li>□ D predictive attributes</li> </ul>	3	normal	female	yes	2.00	1.00	2	
	4	normal	male	no	1.00	3.00	2	
gender	5	normal	female	yes	3.00	1.00	2	
Smoke	6	normal	female	yes	2.00	1.00	3	
exercise	7	high	female	no	2.00	3.00	3	
alcohol	8	normal	female	no	3.00	1.00	2	
stress	9	normal	female	no	2.00	1.00	1	
salt	10	high	female	no	1.00	1.00	3	
income	11	normal	male	no	3.00	1.00	1	
education	12	high	male	yes	1.00	1.00	3	
	13	high	female	yes	1.00	3.00	3	
earning method	14	high	male	no	3.00	1.00	3	
tethodName=Improved ChAID [1 sc A	15	high	female	no	1.00	3.00	з	
dl=8	16	high	female	yes	1.00	1.00	2	
lerge=0.05	17	normal	female	no	1.00	1.00	1	
plit=0.001	18	high	male	no	1.00	2.00	1	
Abeneurieurieurieurieurieurieurieurieurieuri	19	high	female	yes	1.00	2.00	3	
Examples selection	20	normal	male	no	2.00	1.00	1	
39 examples selected examples idle	×	heine		1.1.2	0.00	0.00	>	

Some of the input attributes are maybe irrelevant. Induction tree algorithm can highlight automatically the most relevant predictive attributes. It is one of its key points.

### Selection of the learning sample

We use mainly an interactive approach in this tutorial. We use all the dataset for the tree induction. It is the default selection. We see the dataset selection in the bottom part of the project explorer.

🖇 File Edit Data Statistics Indu 🥆 🤮 📴 🍬 🖎	ction method	l Analysis Vie	w Windov	v Help		-	đΧ
		bpress level	aender	smoke	exercise	overweight	al 🔨
Attribute selection	1	high	female	no	1.00	1.00	1
🖃 🗊 Class attribute	2	normal	female	no	3.00	1.00	1
De bpress_level     Predictive attributes     gender     smoke     swarie	3	normal	female	yes	2.00	1.00	2
	4	normal	male	no	1.00	3.00	2
	5	normal	female	yes	3.00	1.00	2
smoke	6	normal	female	yes	2.00	1.00	3
exercise	7	high	female	no	2.00	3.00	3
C overweight C alcohol C stress	8	normal	female	no	3.00	1.00	2
	9	normal	female	no	2.00	1.00	1
salt	10	high	female	no	1.00	1.00	3
- C income	11	normal	male	no	3.00	1.00	1
education	12	high	male	yes	1.00	1.00	3
	13	high	female	yes	1.00	3.00	з
Learning method	14	high	male	no	3.00	1.00	3
MethodName=Improved UhAID [Isc 🔨	15	high	female	no	1.00	3.00	3
Hdl=8	16	high	female	yes	1.00	1.00	2
Merge=0.05	17	normal	female	no	1.00	1.00	1
Split=0.001 TurseBenterreni=1	18	high	male	no	1.00	2.00	1
	19	high	female	yes	1.00	2.00	3
Examples selection	20	normal	male	no	2.00	1.00	1
399 examples selected	3	lateria		1	2.00	2.00	>

Note on subsample definition: If we want to subdivide the dataset into learning and test sample. We use the ANALYSIS / SELECT ACTIVE EXAMPLES menu. We can subdivide randomly the dataset, we can also use a rule based selection, etc.

#### Start problem analysis

We want to build the tree using the selected method. We activate the ANALYSIS / LEARNING menu. From the decision tree (Figure 1), we extract the following rules.

Rule (If Premise Then Conclusion)	Confidence	Lift	Support
If overweight $>= 2.5$ Then Blood pressure = high	71%	1.25	40% <sup>2</sup>
If overweight < 2.5 ET Exercise < 1.5 Then Blood Pressure = high	62%	1.08	24%
If overweight < 2.5 ET Exercise >= 1.5 Then Blood pressure = low	61%	1.40	36%

A rule is relevant if it has a high confidence and a high support.



Figure 1 – Decision tree on the "blood pressure" dataset

We can evaluate the information provided by a rule by computing the ratio between the proportion of class in the whole dataset and the proportion of the class in the rule covered examples: it is the LIFT value. If the LIFT is upper than 1, the rule is interesting. In the first rule, the LIFT is 1.40 = 71%/57%.

#### Interactive exploration of the tree

Now begins really the work of domain experts. The goal is to obtain a better characterization of rules associated to the leaves of the tree. For that purpose, we have to determine the role of the variables which do not apparently appears in the tree. Are they really irrelevant or are they masked by the selected variables?

<sup>&</sup>lt;sup>2</sup> The definition of the support for the predictive rules (supervised learning) is different from that used in the association rules induction. Here, it is the number of examples covered by the premise divided by the whole dataset size i.e. 40% = (113+46)/399.

#### Alternative splitting variables

On the root of the tree, the first split uses "overweight". Is it the only relevant variable? What about the other predictive variables?

Indeed, the technique chooses simply the best variable in the sense of a given criterion. The other variables may be masked, even if they have almost an equivalent quality. This is not absolutely harmless. If we choose to split with another variable on the place of that automatically detected by the method, it is possible that the other variables occurring in the low parts of the tree are completely different. We obtain very different rules<sup>3</sup>.

We are going to study the alternative solutions of the variable "overweight" during the segmentation of the root of the tree. For that purpose, we make a click with the right button of the mouse on this node. In the contextual menu which appears, we activate the option NODE INFORMATION ...



A new window appears. We observe the list of predictive variables and their respective contributions if we use them. This method uses the TSCHUPROW's T in order to characterize a split.

<sup>&</sup>lt;sup>3</sup> But, if the rules seem different, they classify the examples in the same way. In this point of view, the trees are similar.

Informations or	i:Leve	l 1, Node 1						×
IF								
Characterization	Desci	ptors' importan	ce					
Select an attribu Double-clik to sp	te to vie Ilit with t	w the suggest he selected at	ed split tribute					
	Go	odness of spli	t Corre	ation			Accept or Rejec	:t
overweight	0.0	5246551	0.052	5	}			
alcohol	0.0	1769547	0.017	7	١			
exercise	0.0	1762666	0.017	6	L			
smoke	0.0	1308122	0.013	1	ſ			
stress	0.0	1059854	0.010	6	J			
income	0.0	0551802	0.005	5	١			
salt	0.0	0034315	0.000	3	L			
education	0.0	0005825	0.000	1	ſ			
gender	0.0	0000000	0.000	D	J			
Split suggestion								
< 1.	50	>=1.50						
high 76		152						
normal 54		117						
	39	9 examples (1)	00.00% of	the le	ean	ning se	t)	

We observe mainly three groups of variables:

- 1. OVERWEIGHT is really the most interesting variable for splitting the root node. In the low part of the window appears the associated partitioning. It is the same that we observe in the graphical representation of the tree.
- 2. ALCOHOL, EXERCISE, SMOKE and STRESS are the next. They bring less information than OVERWEIGHT, but they still provide information about BLOOD PRESSURE discrimination problem.
- 3. Then, INCOME, SALT, EDUCATION and GENDER seem not relevant, at this step, for the detection of individuals with high BLOOD PRESSURE.

We wonder what happens if we choose another splitting variable, ALCOHOL for instance. Because the expert points out that it is important. In that case, criteria which are not numeric come into play in the study.

In a first step, we do not want to modify the tree, we want to see only the resulting partition if we use ALCOHOL. For that, we click on GOODNESS OF SPLIT value for each variable. In the case of ALCOHOL, we activate the corresponding box; the bottom part of the window reflects the associated segmentation.

Informations o	on : Lev	vel 1, Node 1			X
Characterization	n Des	ciptors' importanc	e		
Select an attrib Double-clik to s	ute to v plit with:	iew the suggeste h the selected attr	d split ibute		
	G	oodness of split	Correlation	Accept or Reject	T
overweight	0	.05246551	0.0525		
alcohol	Ö	.01769547 🐝	0.0177		
exercise	0	.01762666	0.0176		
smoke	0	.01308122	0.0131		
stress	0	.01059854	0.0106		
income	0	.00551802	0.0058		
salt	0	.00034315	0.0003		
education	0	.00005825	0.0001		
gender	0	.00000000	0.0000		
		1			
Split suggestion		k			
< 2	2.50	>=2.50			_
high 14	2	86			
normal 12	8	43			
		· · · · · · · · · · · · · · · · · · ·			
	3	399 examples (10	0.00% of the learning	ng set)	

The expert, according the domain knowledge, may say us if this solution is really irrelevant or not.

#### Description of a node (subpopulation)

Now, we want to study the node at the following level, the group of the people "OVERWEIGHT > 2.5". We select the node. It is not necessary to close the previous window. The values are automatically updated.

nduction method Analysis Tree man	agement '	View Window	Help								
5 🖪 🖳 🐴 🖻											
ittibute extension	Decision	tree						X			
titibute selection	<u>; 1,50</u> 59 (( 36 ()	< 2.50 115 ( 125 ( exercis 521) 521)	228 0vervs 484) 522) ►=1.5( 56 89	(57%) (43%) eiglit = 1 4 (39%) (61%)	hig nor 13 (714) 6 (294)	h mal	Informations on IF overweight >=2 7 Characterization Select an attribut Double-clik to spi alcohol education satt income stress exercise	ss         sait           a         sait           a         a           a         a           a         a           a         a           a         a           a         a           a         a           a         a           a         a           b         a           b         a           b         a           b         a           b         a           b         a           c         a           c         a           c         a           c         a           c         a           c         a           c         a           c         a           c         a           c         a           c         a           c         a	income           1.00           3.00           4.00           3.00           4.00           6           ed split           mbute           6           0.0214           0.0214           0.0137           0.0116           0.0107           0.0004	Accept or R	Reject
eaming method							smoke	0.00000000	0.0000		
1ethodName=Improved ChAID (Tsc 🗡	21	high	male	Ves	3.00	3.00	genaer	10.0000000	0.0000		
dehodLlassName=IArbreDecisioni	22	normal	male	ves	3.00	1.00					
erge=0.05	23	high	female	Vac	1.00	1.00	Split suggestion				
plit=0.001	24	ngr	molo	700	1.00	1.00	< 2.5	50 >=2.50			
ypeBonferroni=1	24	Hormal	male	110	0.00	0.00	high 80	33			
alueBonterroni=1	25	normal	remale	yes	3.00	2.00	normal 39				
amping-o 💽	26	normal	male	no	1.00	1.00					
Examples selection	27	high	male	no	1.00	1.00					
99 examples selected	28	high	female	no	1.00	2.00					
		NOTES SECTION	male		2.00	4.00					

This node might be split using the variables ALCOHOL or EDUCATION. The operation was not carried out because it seems not to be numerically relevant.

The second important matter legitimates the interactive analysis in the induction trees: it is understood that the people associated with this node are overweight and hypertensive, but what about other variables, what are the other characteristics of these individuals? This is one of the drawbacks of the decision trees. The method proposes the relevant variables. But we have not visibility on the variables which were not integrated in the model. However, they can help us to deeply characterize the rules.

The interactive functionalities of SIPINA enable us to answer this requirement. The CHARACTERIZATION tab describes the groups associated with each node of the tree. SIPINA computes comparative statistics between the root node, representing the whole population, and the current node, representing the subpopulation defined by the rule.

To evaluate the importance of the difference, the value test (strength) which is a test statistic: a comparison of mean when the variable is continuous (t-test), a comparison of proportion when the variable is discrete. It is not strictly speaking a statistical test since the samples are not independent, but its interest and its flexibility are undeniable in practice.

In the "Continuous" tab, we have the values of mean on the root node (Global Avg.) and on the current node (Local Avg.).

Informations on : Leve	formations on : Level 2, Node 2										
IF overweight >=2.50											
Characterization Desci	tors' importance										
Continuous attributes	Discrete attributes										
Attribute	Strength	Local Avg	Global Avg								
overweight	18.24	3.0000	1.9925								
exercise	0.41	1.9811	1.9599								
stress	0.18	2.0314	2.0226								
income	0.05	1.9497	1.9474								
salt	-0.52	2.0063	2.0326								
education	-0.54	1.9748	2.0025								
alcohol	-1.91	1.9057	2.0000								
JL											
15	i9 examples (39.85%	of the learning s	et)								

The OVERWEIGHT is really high on this subpopulation, 3.0 vs. 1.9925 on the whole population. The importance of the difference is materialized with a STRENGTH = +18.24. This result is obvious because OVERWEIGHT takes part in the decision tree construction.

But another issue, which is not obvious, is that ALCOHOL consumption seems significantly low in this subpopulation: STRENGTH = -1.91.

It is not really possible to give a threshold which enables us to decide if the difference is significant or not. Because the samples are not independent, and the group is designed in order to optimize a purity criterion. But, we can however distinguish an abnormal deviation according to the comparison based on the other variables. In this subgroup, it seems that OVERWEIGHT and ALCOHOL are interesting. For the "Discrete" variables, we obtain the following results:

nto IF o	rmations on : Level 2, Node 2 overweight >=2.50									
CL.	K									
ιn	aracterization   Desci	otors' importar	nce							
С	ontinuous attributes	Discrete at	tributes							
Г		harasa	laural ( 0 0220			_				
	Values	Strength	Local Dist.	Global Dist.	Recall					
	high	4.57	113 (71%)	228 (57%)	50%					
	normal	-4.57	46 (29%)	171 (43%)	27%					
		smok	e f 0.0065 1							
	Values	Strength	Local Dist.	Global Dist.	Recall					
	no	-2.43	65 (41%)	193 (48%)	34%					
	yes	2.43	94 (59%)	206 (52%)	46%					
		gend	er ( 0.0000 )							
	Values	Strength	Local Dist.	Global Dist.	Recall					
	female	0.11	87 (55%)	217 (54%)	40%					
	male	-0.11	72 (45%)	182 (46%)	40%					
L										
_				·		-				

Now, STRENGTH is the statistical test of the comparison of proportions (frequencies). We observe that there is over representation of BPRESS = HIGH. But the proportion of SMOKE = YES is also high in this subgroup, compared with the whole population. This last characteristic is not visible in the decision tree.

#### User-driven induction tree

The possibility for the user of guiding the exploration is certainly one of the most desirable aspects of the decision trees. Let us take again the variable SMOKE which seems very important finally. We decide to insert it as first split variable of the root node. We select the root of the tree. In order to prune it manually, we carry out a click with the right button of the mouse, in the contextual menu we select the CUT option.

🔉 Sipina Research Version									
Induction method Analysis Tree mana	igement Vie	w Window	Help						
<mark>D 🛯 📲 🐴 🖒</mark>									
Attribute selection  Type Class attribute  Type Class attributes  Type class attributes  Type redictive attributes  Type redictiv	<ol> <li>Deci</li> <li>&lt; 1.</li> <li>59</li> <li>36</li> </ol>	sion tree.	2.50 1115 (488) 125 (524) ercise	228 (57 171 (43 0vertreigh -=1.50 56 (39) 89 (614	<ul> <li>3)</li> <li>4)</li> <li>5)</li> <li>113</li> <li>46</li> </ul>	high ✓ Level 1, N Node infor Cut Split node Explore th Other des Other SIP	ode 1 mations in node criptive statis INA session	tics )	<ul> <li>3</li> <li>3</li> <li>1.0</li> <li>2.0</li> <li>3.0</li> <li>1.0</li> <li>2.0</li> <li>3.0</li> <li>1.0</li> <li>2.0</li> <li>3.0</li>     &lt;</ul>
Learning method MethodName=Improved ChAID (Tsc ∧ MethodClassName=TArbreDecision) Hdl=8 Merge=0.05 Splite-0.001 TypeBonterroni=1 ValueBonterroni=1 Sampling=0 Examples selected 0 examples selected 0 examples selected	21 22 23 24 25	high normal high normal normal	male male female male female	yes yes yes no ves	3.00 3.00 1.00 1.00 3.00	3.00 1.00 1.00 1.00 2.00	1.00 2.00 2.00 3.00 1.00	3.00 3.00 2.00 3.00 1.00	3.0 1.0 1.0 3.0 1.0 3.0 2.0 3.0 3.0 3.0 3.0 3.0 3.0 3.0 3
	<								>
Improved ChAID (Tschuprow Goodness of	<sup>:</sup> Split)							Ţ	ime : 313 🏒

Always with the contextual menu, we activate the option SPLIT NODE, we obtain the same window as for NODE INFORMATION option. We select DESCRIPTORS' IMPORTANCE tab. In the list of the candidate descriptors, we click on SMOKE (Figure 2).

In order to perform a split, we double-click on the box containing the value of GOODNESS OF SPLIT for the selected variable. In this case, splitting is carried out even if the conditions of acceptance (significance test etc.) are not met (Figure 3).



Figure 2 - Splitting with SMOKE attribute

We note that there is indeed an over representation of hypertensive among smokers (63% compared with 57% in the global population). The proportions are balanced for non smokers.

🕵 Sipina Research Version										1	
Induction method Analysis Tree ma	anagement '	View Window	Help								
D 🖪 🛤 🐂 🖡											
		44.4							<b>1</b>		1
Attribute selection	E RAL De	cision tree.	22 <sup>-1</sup>								
E 🗊 Class attribute								^	salt	income	educ
bpress_level									1.00	1.00	3.00
E IF Predictive attributes			228 (57%)	2000	high				2.00	3.00	1.00
smoke		>	171 (43%)	1.0	normal				2.00	1.00	2.00
evercice			ioke						3.00	3.00	1.00
	1 10	[no]		in lyes	CO11	_			1.00	1.00	1.00
alcohol		9 (SI%) 4 (49%)		77 (	3781				2.00	2.00	2.00
stress						_			1.00	3.00	3.00
salt									3.00	1.00	3.00
C income									1.00	3.00	3.00
C education									2.00	3.00	2.00
N. 199									2.00	2.00	3.00
•					100	23	-13 8-1-3		0.00	4.00	4.00
	<b>III</b> .				Informacions	s un aces	rei 1, noue 1				
	1	17 m	<b>.</b>		Characterizati	tion Desi	ciptors' importance	•			
	<				Characterizati Select an attr Double-clik to	tion Des tribute to v to split with	ciptors' importance iew the suggested the selected attrib	e   I split pute			
	21	high	male	yes	Characterizati Select an attr Double-clik to	tion Desitribute to v	ciptors' importance iew the suggested the selected attrib codness of split 05046551	s   I split bute Correlation	Acc	ept or Reject	
	21 22	high	male	yes yes	Characterizati Select an attr Double-clik to overweight	tion Desitivite to v	ciptors' importance iew the suggested the selected attrib oodness of split .05246551	split oute Correlation 0.0525 0.0177	Acc	ept or Reject	_
	21 22 23	high normal high	male male female	yes yes yes	Characterizati Select an attr Double-clik to overweight alcohol exercise	tion Desitivite to v	ciptors' importance iew the suggested the selected attrib oodness of split .05246551 .01769666	split oute Correlation 0.0525 0.0177 0.0176	Acc	ept or Reject	<b>-</b>
	21 22 23 24	high normal high normal	male male female male	yes yes yes no	Characterizati Select an attr Double-clik to overweight alcohol exercise smoke	tion Desitivite to v tribute to v to split with 0 0 0 0	ciptors' importance iew the suggested the selected attrib 05246551 01769547 01769646 01308122	s plit oute Correlation 0.0525 0.0177 0.0176 0.0131	Acc	ept or Reject	
	21 22 23 24 25	high normal high normal normal	male male female male female	yes yes yes no yes	Characterizati Select an attr Double-clik to overweight alcohol exercise smoke stress	tion Desitivity of the second	ciptors' importance iew the suggested the selected attrit oodness of split .05246551 .01769547 .01769547 .01308122 .01308122 .01059654	s plit oute Correlation 0.0525 0.0177 0.0176 0.0131 0.0131 0.0106	Acc	ept or Reject	
	21 22 23 24 25 26	high normal normal normal normal	male male female female female male	yes yes no yes no	Characterizati Select an attr Double-clik to overweight alcohol exercise smoke stress income	tion Desu tribute to v to split with 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	ciptors' importance iew the suggested the selected attrit 05246551 01769665 01308122 01059654 00150854	2 split oute Correlation 0.0525 0.0177 0.0176 0.0131 0.0131 0.0106 7.0055	Acc	ept or Reject	
	21 22 23 24 25 26 27	high normal high normal normal normal	male male female male female female male male	yes yes yes no yes no no	Characterizati Select an attr Double-clik to alcohol exercise smoke stress income salt	tion Desi tribute to v to split with 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	ciptors' importance iew the suggested the selected attri 007469551 01769665 01306122 01059854 000551802 00054180	3 split oute 0.0525 0.0177 0.0176 1.0.0131 0.0106 0.0003	Acc	ept or Reject	
Transition multiple	21 22 23 24 25 26 27 28	high normal high normal normal high high	male male female female female male male female	yes yes no yes no no no no	Characterizati Select an attr Double-clik to overweight alcohol exercise smoke stress income salt education	tion Dest tribute to v to split with 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	ciptors' importance iew the suggested the selected atrii 007695475 01769547 01769547 010309122 \$2 01059854 00551802 000054315 00005825	split oute Correlation 0.0525 0.0177 0.0176 0.0176 0.0131 0.0106 0.0055 0.0003 0.0001	Acc	ept or Reject	
Learning method	21 22 23 24 25 26 27 28 29	high normal normal normal normal high normal	male male female male female male female female male	yes yes yes no yes no no no no no	Characterizati Select an attr Double-clik to overweight alcohol exercise smoke stress income salt education gender	tion Deservations Deservations Deservations of the second	ciptors' importance iew the suggested the selected attrit oodness of split 01769665 01300122 01058654 01058654 000551802 000058130 00005825 00000000	s plit cute Correlation 0.0525 0.0177 0.0176 0.0131 0.0106 0.0005 0.0003 0.0000	Acc	ept or Reject.	
Learning method MethodName=Inproved ChAID [Tsc A MethodClassName=TAhroeDcsison]	21 22 23 24 25 26 27 28 29 30	high normal high normal normal high high normal	male male female female female male female female female female	yes yes no yes no no no no no no	Characterizati Select an attr Double-clik to over?weight alcohol exercise smoke stress income salt education gender	tion Deservations Deservations Deservations of the second	ciptors' importance iew the suggested the selected attrition codness of split .01769665 .01308122 .01059864 .000591802 .00054180 .00054180 .0005425 .0000000	split oute Correlation 0.0525 0.0177 0.0176 0.0131 0.0055 0.0003 0.0001 0.0000	Acc	ept or Reject	
Learning method MethodName=Improved ChAID (Tisc & MethodQlassName=TArbreDecisionI Hdi=8	21 22 23 24 25 26 27 28 29 30 31	high normal normal normal normal high high normal normal	male male female male male male female female female female male	yes yes no yes no no no no no no no no	Characterizati Select an attr Double-clik to over Weight alcohol exercise smoke stress income satt education gender Split suggestic	tion Desu tribute to v o split with 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	ciptors' importance iew the suggested the selected attrib .05246551 .01769642 .01769642 .01769642 .01059854 .0005415 .00004315 .00004315	s plit oute Correlation 0.0525 0.0177 0.0176 0.0131 0.0106 1.0055 0.0003 0.0001 0.0000	Acc	ept or Reject	
Learning method MethodName=Improved ChAID [Tsc A MethodDastAme=TAthreDecision] Hdi=8 Metge=0.05	21 22 23 24 25 26 27 28 29 30 31 32	high normal high normal normal high normal normal normal normal	male male female male female male female female female male female male	yes yes no no no no no no no no no no no no	Characterizati Select an attr Double-clik to overviveight alcohol exercise smoke stress income sait education gender Splt suggestic	tion Desit tribute to v o split with 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	ciptors' importance iew the suggested the selected attri 00246551 01769547 01769547 010306122 000551802 000541802 000034315 0000000	split oute 0.0525 0.0177 0.0176 0.0176 0.0176 0.0106 0.0003 0.0000 0.0000	Acc	ept or Reject	
Learning method MethodName=Improved Ch4ID (Tsc A MethodClassName=TA/breDecision Metge=005 Spil=0001 ToxeBorderoni=1	21 22 23 24 26 26 27 28 29 30 31 32 33	high normal high normal normal normal normal normal normal normal high	male male female male male male female male female male male male male	yes yes no yes no no no no no no no no no no no no no	Characterizeti Select an attr Double-clik to overviveght alcohol exercise smoke stress income sait education gender Split suggestic	tion Desit tribute to v o split with 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	ciptors' importance iew the suggested the selected attrit 05246551 01769547 01769647 01059654 000541802 00054315 00005825 00000000	split correlation 0.0525 0.0177 0.0176 0.0131 0.0106 0.0055 0.0003 0.0001 0.0000	Acc	ept or Reject	
Learning method MethodName=Improved ChAID [Tsc A MethodClassName=TArbreDecision] Hdl=8 Merge=0.05 Spli=0.001 TypeBorferroni=1 VaueBonferroni=1	21 22 23 24 25 26 27 27 28 29 31 31 32 33 34	high normal high normal normal high normal normal normal high high high high	male male female female male male female female male male male male male male male	yes yes yes no no no no no no no no no yes	Characterizeti Select an attr Double-clik to overversight alcohol exercise smoke stress income sait education gender Split suggestic inormal solit suggestic	tion Dest tribute to v o split with 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	ciptors' importance iew the suggested the selected attil 00769647 01769647 01769642 01769642 01769642 01769612 017697612 017697612 017697612 017697612 017697612 017697612 017697612 017697612 017697612 017697612 017697612 0176976100000000000000000000000000000000	split Correlation 0.0525 0.0176 0.0176 0.0176 0.0131 0.0055 0.0003 0.0000 0.0000		ept or Reject	
Learning method MethodName=Inproved Ch4ID (Tsc A MethodClassName=TAthreDecision) Hdl=8 Merge=0.05 Spli=0.001 TypeBorferoni=1 ValueBorferoni=1 ValueBorferoni=1 ValueBorferoni=1	21 22 23 24 25 26 27 28 28 30 31 32 33 34 35	high normal normal normal normal high normal normal normal high normal normal normal	male male female female male female female male male male male male male male	yes yes yes no no no no no no no yes yes	Characterizati Select an attr Double-clik to alcohol exercise smoke stress income salt education gender Split suggestic high a normal s	tion Dest tribute to v o split with 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	ciptors' importance iew the suggested the selected attri oodness of split 0.05246551 0.1769547 0.1769547 0.01508122 0.00541802 0.000541802 0.000541802 0.000541802 0.000541802 0.000541802 0.00008255 0.0000000	spik Jule Correlation 0.0525 0.0177 0.0176 0.0013 0.0016 0.0005 0.0000 0.0001 0.0000	Acc	ept or Reject	
Learning method MethodName=Inproved Ch4ID (Tsc ~ MethodClassName=TArbreDecisionI Hdi=8 Merge=0.05 Spil=0.001 TypeBoriteroni=1 ValueBoriteroni=1 SampIng=0 Examples selection	21 22 23 24 25 26 26 26 27 28 29 31 31 33 33 34 34 35 36	high normal high normal high normal normal normal high normal high normal high hormal high	male male female male female male female male male male male male male male	yes yes yes no no no no no no no yes yes	Characterizett Select an attr Double-clik to overweight alcohol exercise smoke stress income sait education gender Split suggestic inormal	tion Dest tribute to v o split with 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	ciptors' importance iew the suggested the selected attribucture 05248551 01769612 01769612 01059847 01059612 01059840 000551802 000058180 00000000 in [ves] 120 77	split           uste           0.0525           0.0177           0.0178           0.0173           0.0176           0.0000           0.0000	Acc	ept or Reject	
Learning method MethodName-Improved Ch4ID (Ts c MethodDlassName=TArbreDecisioni Hdin8 Merge=0.05 Spil=0.001 TypeBorferroni=1 Sampling=0 Examples selection S39 examples selected	21 22 23 24 25 26 26 27 27 29 30 31 32 33 34 35 35 35 37	high normal high normal normal high high normal normal high high high high	male male female male female male female male female male male male male female female	yes yes yes no no no no no no yes yes yes yes	Characterizeti Select an attr Double-olik te alcohol exercise smoke stress income salt education gender Split suggestic high i normal s	tion Dest tribute to v to split with 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	ciptors' importance iew the suggested the selected attri codness of split 0.5246551 0.1769547 0.1769547 0.0169854 0.00051802 0.00054315 0.00054315 0.000054315 0.0000000	spit Joure Correlation 0.0525 0.0177 0.0176 0.0106 0.0003 0.0001 0.0000	Acc	ept or Reject	
Learning method MethodName=Improved Ch4ID (Tsc × MethodClassName=TAbreDecision Merge=005 Spil=0.001 TypeBonteroni=1 ValueBonterroni=1 ValueBonterroni=1 Samples selection 399 examples selected 0 examples selected	24 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 37 36	high normal high normal normal high normal normal normal normal normal normal normal normal normal normal	male male female male female male female male male male male male male male	yes yes no no no no no no no yes yes yes yes yes	Characterizati Select an attr Double-clik to over weight alcohol exercise smoke stress income salt education gender Split suggestie high s	tion Dest tribute to v o split with 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	ciptors' importances iew the suggested ocdness of split 05246551 01769547 01769547 010309122 01059854 00034315 000034315 00000000 in [yes] 123 77	spik Jule Correlation 0.0525 0.0177 0.0176 0.0176 0.0035 0.0005 0.0005 0.0000 0.0000	Acc	ept or Reject	

Figure 3 – Tree with SMOKE as first splitting variable

From now, we can build the tree according to our domain knowledge. We select the group SMOKE = YES. We observe the Tschuprow's t of predictive variables. OVERWEIGHT seems again a relevant attribute. We double-click on the value in order to split the node.

🔉 Sipina Research Version										
Induction method Analysis Tree man	agement Vi	iew Window	Help							
🗅 🖻 陆 🍬 🖒										
	191 mm	cores mi				6				
Attribute selection	: 🔊 Dec	ision tree.				L	- [[] []			
🖃 🕼 Class attribute	1						^	salt	income	educ
bpress_level								1.00	1.00	3.00
Predictive attributes			228 (57%)		high			2.00	3.00	1.00
gender			171 (43%)	1. A. A.	normal			2.00	1.00	2.00
evercise		ST ST	oke					3.00	3.00	1.00
overweight	1n	Inol		in lyes				1.00	1.00	1.00
alcohol	94	(49%)		77 (0	37%)			2.00	2.00	2.00
stress				overwei	nt C			1.00	3.00	3.00
C salt	4	<	1.50	_	>=1.50			3.00	1.00	3.00
C income	4		40 (51%)		89 (70%)			1.00	3.00	3.00
education	4		38 (49%)		39 (30%)		N	2.00	3.00	2.00
	4						<u>``</u>	2.00	2.00	3.00
	4				Informations on :	Level 2. Node 2			4.00	100
	4				E omoleo in [uoo]	and straight				
					Select an attribute Double-clik to split	to view the suggested s with the selected attribu	split / ite/			
	24	biob	male	Ven		Goodness of split	Correlation	Acce	pt or Reject	
	21	normal	male	yes	overweight	0.03347432	0.0335			_
	23	high	female	ves	exercise	0.01548021	9.0155			-
	24	normal	male	00	income	0.01459941	0.0146			
	25	normal	female	ves	stress	0.00432071	0.0043			-
	26	normal	male	no	salt	0.00041953	0.0004			-
	27	high	male	no	education	0.00007935	0.0001			
	28	high	female	no	smoke	0.00000000	0.0000			
Learning method	29	normal	male	no	gender	0.00000000	0.0000			
MethodName=Improved ChAID (Tsc A	30	normal	female	no						
Hdl=8	31	normal	male	no	Solit suggestion					
Merge=0.05	32	high	male	no	<1.5	) >=1.50				
Split=0.001	33	high	male	no	high 40	89				
ValueBonferroni=1	34	normal	male	yes	normal 38	39				
Sampling=0 💌	35	normal	male	yes		1100				
Examples selection	36	high	female	yes						
399 examples selected	37	high	female	yes						
0 examples idle	38	normal	female	yes						Y
1	<					206 examples (51.)	53% of the lear	nina setì		
Improved ChAID (Tschuprow Goodness o	f Split)					See sumpto (or a		2		

BLOOD PRESSURE is high when OVERWEIGHT is high. But compared with the first tree (Figure 1), we note that the cut point is different for the SMOKE = YES group. A moderate OVERWEIGHT (> 1.5) is detrimental when the people are smoker. The threshold is higher (> 2.5) for the whole population. For non smokers, the cut point (> 2.5) is the same one.



When we continue the analysis, we may obtain (among various solutions) the following tree.

🔉 Sipina Research Version									
Induction method Analysis Tree mana	agement View	v Window	Help						_
									_
Attribute adaption	B Decis	ion tree.							ব ^
Class attribute								الم المالية الم	
boress level									<b></b>
E 📭 Predictive attributes					[	L			
gender	1				171 (433)	nign normal			
- P smoke					smoke				
overweight		in	[no]			in tyes!			
alcohol	4		9 (51%) 94 (49%)			77 (37%)			
C stress		ov	erweight	<u> </u>		overweight			
- 🖸 salt	< 2.5	50		>=2.50	< 1.50		>=1.50		
income	53	(41%)		46 (71%	) 40 (51)		89 (70%) 39 (30%)		
- Le education	-	10017		alcohol			00 (001)		
		<	2.50		>=2.50				
		3	32 (65%)		14 (88%)				
	17		.7 (334)		2 (13%)	_		_	
					Informations on : Level 4	I, Node 2		2	
					IF smoke in [no] and overw	eight >=2.50 and a	Icohol >=2.50	•	-
	-						•		
	<				Characterization Descipto	irs' importance			
	21	high	male	yes	Continuous attributes	Discrete attributes			
	22	normal	male	yes	V				
	23	high	female	yes	Attribute	Strength	Local Avg	Global Avg	
	24	normal	male	no	alcohol	5.07	3.0000	2.0000	
	25	normal	female	yes	overweight	4.58	3.0000	1.9925	
	26	normal	male	no	income	0.58	2.0625	1.9474	
	27	high	female	no	exercise	0.49	2.0625	1.9599	
	29	normal	male	00	stress	-0.11	2.0000	2.0226	
1	30	normal	female	no	education	-0.32	1.9375	2.0025	
Learning method	31	normal	male	no	salt	-1.39	1.7500	2.0326	
MethodName=Improved ChAID (Tsc 🔨	32	high	male	no					
MethodUlassName=IArbreDecisioni	33	high	male	no					
Merge=0.05	34	normal	male	yes					
Split=0.001	35	normal	male	yes					
ValueBonferroni=1	36	high	female	yes					
Sampling=0 💉	37	high	female	yes					
Examples selection	38	normal	female	yes	-1				
399 examples selected	38	normal	maie	no	-				
D outperplay idle	40	(bioth	famala	00					
0 examples idle	40	high	female	no					

Even if SIPINA is far from having all the functionalities of commercial software (SPAD Interactive Tree Decision, SPSS Answer Tree, SAS EM, etc), it nevertheless proposes some options intended to improve the presentation of the results. They are available in the TREE MANAGEMENT menu. We can, among other things, copy the graphical representation of the tree in the clipboard.



# Conclusion

The decision trees are a popular data mining method. This popularity relies partly on the flexibility of the software which gives to the user the possibility to guide the induction process according to the domain knowledge. In this tutorial, we present the functionalities of SIPINA for interactive exploration.