## **1** Introduction

# Applying a predictive model on new unlabeled sample. Estimation of the generalization error rate by resampling approach.

**Model deployment.** Model deployment is the last step of the Data Mining process. In its simplest form in a supervised learning task, it consists in to apply a predictive model on unlabeled cases. Most of the tools have this kind of capability. The other alternatives are programming a specific tool (in Java for instance, but it can be very costly if we must repeatedly implement models in various contexts), using SQL queries (especially useful when the model can be expressed in a set of rules), or using the PMML framework<sup>1</sup>.

Sipina can apply a learned classifier on a data file with unlabeled cases. There are some constraints however: the data file must be in the Sipina binary format (FDM<sup>2</sup>), we must then convert the data file before; we must apply the model (the most often a decision tree but it can be any classifier available into Sipina) into the Sipina environment, we cannot export the model.

**Generalization error rate estimation**. Applying the model on unseen cases is a very useful functionality. But it would be even more interesting if we could announce its accuracy. Indeed, a misclassification can have dramatic consequences. We must measure the risk we take when we make decisions from a predictive model. An indication about the performance of a classifier is important when we decide or not to deploy it.

The usual approach to assess the accuracy of a classifier is to subdivide the dataset into train and test samples. The first is used for the learning phase; the second is used for the construction of the confusion matrix comparing the observed values of the class attribute and the predicted values of the model. We then obtain an unbiased estimation of the indicators (error rate, recall, precision, etc.). But this approach is not judicious if we have a moderate size dataset. Indeed, on the one hand, we penalize the construction of the classifier by using a reduced size sample during the learning phase. On the other hand, we have often not enough instances into the test sample in order to obtain an accurate estimation of the error rate.

Thus, in the moderate size dataset context, it is more suited to use a resampling approach such as bootstrap to evaluate the generalization error rate. The aim is to obtain an estimation of the error rate of the classifier learned on the whole dataset. During the resampling process, we can learn many models. Studying these individual models has not an intrinsic interest. They contribute only for the error rate evaluation. This is the reason for which they are not displayed in the majority of the Data Mining tools.

<sup>&</sup>lt;sup>1</sup><u>http://www.dmg.org/</u>

<sup>&</sup>lt;sup>2</sup> See <u>http://data-mining-tutorials.blogspot.com/2009/11/sipina-supported-file-format.html</u> about the supported file formats in Sipina.

#### SIPINA

**Organization of this tutorial.** In this tutorial, we show how to learn a decision tree from a dataset, how to apply the classifier on unlabeled cases, and how to evaluate the generalization error rate using the bootstrap approach. The samples are gathered into Excel workbook. There are several sheets: (1) the dataset used for the learning process and the bootstrap error rate evaluation; (2) the unlabeled instances, only the descriptors are available; (3) the labels of the instances from the second sheet, we can evaluate the true classifier accuracy, we will compare this last one with the value estimated by the bootstrap approach.

## 2 Dataset

We use the WINE<sup>3</sup> (<u>wine\_deployment.xls</u><sup>4</sup>) dataset. We want to classify alcohols from their chemical properties. The class attribute has 3 values. There are 12 predictive continuous descriptors.

We have the labeled dataset into the first sheet (Data.Learning). We use this sample (100 instances) for the training task and the bootstrap error evaluation. The unlabeled cases (78 instances) are into the second sheet (Data.Deployment). This is the generalization sample. Only the predictive descriptors are available. An identification column ID is added. It allows to control the consistency of this sample with the values into the third sheet. Last, the third sheet (Data.Deployment.Class.Value) contains the true label of the instances into the second sheet. We use this dataset for computing the true accuracy of the classifier.

<b>N</b>	Microsoft	Excel	- wine_dep	oloyment	.xls							
•	Eichier	<u>E</u> dition	<u>A</u> ffichage	Insertion	Forma <u>t</u>	<u>O</u> utils	<u>D</u> onnée	s Fe <u>n</u> être <u>?</u>	Tanagra Sipi	na Tapez un	e question	×
D	<b>I</b> 0	- 🛍	nial		-	8 -	G Z	S = = = =	i 🗟 💱 %	000 58 498 1	= 💷 + 🖇	» - <u>A</u>
	A1	-	f⊷ Ty	уре								
	A		В		С	D		E	F	G	Н	
1	Тур	e i	Alcohol	Mali	_Acid	Asł	1	Ash_Alcalinity	Magnesium	Total_Phenols	Flavanoids	nflavanoid
2	в		12	.33	0.99		1.95	14.8	136	1.9	1.85	
3	А		13	.56	1.73		2.46	20.5	116	2.96	2.78	
4	B			12	0.92		2	19	86	2.42	2.26	
5	С		12	.51	1.24		2.25	17.5	85	2	0.58	
6	в		12	.21	1.19		1.75	16.8	151	1.85	1.28	
7	A		14	.22	3.99		2.51	13.2	128	3	3.04	
8	с		12	.82	3.37		2.3	19.5	88	1.48	0.66	
9	B		12	.47	1.52		2.2	19	162	2.5	2.27	
10	в		12	.37	1.63		2.3	24.5	88	2.22	2.45	
11	C		12	.81	2.31		2.4	24	98	1.15	1.09	
12	A		13	.73	1.5		2.7	22.5	101	3	3.25	
13	C		13	.36	2.56		2.35	20	89	1.4	0.5	
14	A		13	.50	1.66		2.36	19.1	106	2.00	3.19	
16	0		12	.45	2.4		2.42	20	30	2.3	2.73	
17	A		13	.23	3.3		2.20	17.3	30	2.45	2.03	
18	8		13	.03	1.05		2.0	17.2	34	2.45	2.33	
19	8		11	84	2.89		2.23	18.5	112	1.72	1.04	
20	B		12	42	4.43		2.73	26.5	102	22	2.13	
21	A .		1.	4.1	2.02		2.10	18.8	102	2.2	2.10	
22	Δ.		1	42	1.76		2.45	15.2	112	3.27	3.39	
23	B		12	.72	1.81		2.2	18.8	86	2.2	2.53	
24	c		13	.08	3.9		2.36	21.5	113	1.41	1.39	
25	B		11	.96	1.09		2.3	21	101	3.38	2.14	
26	в		12	.07	2.16		2.17	21	85	2.6	2.65	
27	в		11	.65	1.67		2.62	26	88	1.92	1.61	
28	B		12	.64	1.36		2.02	16.8	100	2.02	1.41	
29	С		13	.27	4.28		2.26	20	120	1.59	0.69	
30	С		13	.78	2.76		2.3	22	90	1.35	0.68	
31	С		1:	3.4	4.6		2.86	25	112	1.98	0.96	
32	в		12	.51	1.73		1.98	20.5	85	2.2	1.92	
33	A		13	.71	1.86		2.36	16.6	101	2.61	2.88	
34	A		13	.29	1.97		2.68	16.8	102	3	3.23	
35	A		1-	4.3	1.92		2.72	20	120	2.8	3.14	
36	B		11	.81	2.12		2.74	21.5	134	1.6	0.99	
37	B		13	.34	0.94		2.36	17	110	2.53	1.3	
38	A		13	.74	1.67		2.25	16.4	118	2.6	2.9	
14 -	• • • N	Data.Le	earning / C	ata.Deplo	yment /	Data.D	eploym	ent.Class.Value	•			
De	ssin + 🗟		automatique:			) 🖾 🚄		🗟 🔷 - 🏄	( • 🛕 • 🔳 🛛	= = • 🖌 .		
Prêt						<ul> <li></li> </ul>	$\overline{}$				NUM	

<sup>3</sup> UCI Machine Learning Repository - <u>http://www.ics.uci.edu/~mlearn/MLSummary.html</u> or <u>ftp://ftp.ics.uci.edu/pub/machine-learning-databases/wine/</u> <sup>4</sup> <u>http://eric.univ-lyon2.fr/~ricco/dataset/wine\_deployment.xls</u>

## 3 Learning and deployment

#### 3.1 Preparing the unlabeled sample

We can launch Sipina from Excel using an add-in<sup>5</sup>. Before to apply a model on a new sample, we must convert it into the native Sipina file format (\*.fdm).

First, we select the data range into the Data.Deployment sheet. We click on the SIPINA / EXECUTE SIPINA. We check the selection and we validate.



Sipina is automatically launched. We obtain 14 attributes and 78 instances. In this data preparation phase, we save only this sample in the FDM file format. We click on the FILE / SAVE AS menu. We set "wine\_data\_deployment.fdm" as file name.

<sup>&</sup>lt;sup>5</sup> See <u>http://eric.univ-lyon2.fr/~ricco/doc/sipina\_xla\_installation.htm</u> for the installation of the add-in; <u>http://eric.univ-lyon2.fr/~ricco/doc/sipina\_xla\_processing.htm</u>, for its utilization.

Sipina Research Version - [Le	earning se	t editor]						_	
💦 File Edit Data Statistics Induc	tion method:	Analysis Vi	ew Window	Help					- 8 ×
New New									
Open		lin	lashal	Malia Aaid	10 mb	Lash Alaslini	Maanaalum	Total Dhar	
Att Save		1.00	Alconor	A CA	Asri	AST_AICAILTI	oz oo	D 40	
Save as		Enregistrer sou					22	4 70	2.1
	2	E		de chilere becc		1. (b) and (0)		0.60	1.0
Access foreign databases 🔸		Enregistrer dans :	D-measure for	decision tree	-	- E C. E.	·	2.00	26
Subsample management 🕨		2	ini wine_data_de	aloyment.fdm				2.00	2.0
Exit		Mes documents						2.45	2.0
	7	récents						2.00	2.6
	8							2.53	2.6
	9	Bureau						3.10	3.6
	10							1.45	1.2
	11							2.40	21
	12	mes documents						2.00	1.5
	13							2.86	3.0
	14	Poste de travail						1.55	3.0
	15	~						2.65	2.3
	16							2.95	3.4
	17	Favoris réseau	Nom du fichier :	wine_data_d	leployment.fdm	<u>└</u> , ⊡	Enregistrer	2.42	2.E
	18		Туре:	Data Manag	er File (".FDM)		Annuler	2.74	3.1
Learning method	19	19.00	13.05	1.77	2.10	17.00	107.00	3.00	3.C
MethodName=Improved ChAID (Tsc 👗	20	20.00	12.88	2.99	2.40	20.00	104.00	1.30	1.2
MethodClassName=TArbreDecisionI	21	21.00	11.79	2.13	2.78	28.50	92.00	2.13	2.2
Hdl=8 Merce=0.05	22	22.00	13.72	1.43	2.50	16.70	108.00	3.40	3.E
Split=0.001	23	23.00	12.25	1.73	2.12	19.00	80.00	1.65	2.0
TypeBonferroni=1	24	24.00	11.61	1.35	2.70	20.00	94.00	2.74	2.9
ValueBonterroni=1	25	25.00	14.39	1.87	2.45	14.60	96.00	2.50	2.5
Consulta a la dian	26	26.00	12.52	2.43	2.17	21.00	88.00	2.55	2.2
Examples selection	27	27.00	13.87	1.90	2.80	19.40	107.00	2.95	2.5 🤜
0 examples idle	<	1		1				- · · · ·	>
	Editing	NEW EDM			-	Attribut	es · 14	Examples : 1	78
, Improved ChAID (Tschuprow Goodpess o	f Solit)	NEWS DPI			•	Accribac	03.14	examples :	- 11.
Improved Chinas (rachapi ow doodness o	- opicy								11

We can close Sipina now. We are back into the Excel spreadsheet. We select the first sheet.

#### 3.2 Importing the learning sample into SIPINA

We select the data range into the "Data.Learning" sheet. Again, we click on the SIPINA / EXECUTE SIPINA menu. We check the range selection and we validate.



Sipina is automatically launched. We have 100 instances and 15 variables (14 predictive attributes and the class attribute) now.

#### 3.3 Decision tree learning

The default selected method is IMPROVED CHAID. It is a variant of CHAID learning algorithm (Kass, 1980). It is suitable for a first exploration of a dataset. Of course, we can select another decision tree induction method (C4.5, etc.) by clicking on the INDUCTION METHOD / STANDARD ALGORITHM menu. But here, we use the default approach. At the same time, all the instances are automatically selected for the learning process (we can modify this by clicking on the ANALYSIS / SELECT ACTIVE EXAMPLES menu). All these information are described in the left part of the main window.

💦 Sipina Research Version - [Le	earning set	editor]							
🔉 File Edit Data Statistics Induc	tion method	Analysis Vie	ew Window	Help					_ 8 ×
🛅 🛅 📸 🛤 🛤									
× 1		Туре	Alcohol	Malic_Acid	Ash	Ash_Alcalin	il Magnesium	Total_Phene	Flava 🔨
Attribute selection	1	в	12.33	0.99	1.95	14.80	136.00	1.90	1.85 🔳
	2	A	13.56	1.73	2.46	20.50	116.00	2.96	2.78
	3	в	12.00	0.92	2.00	19.00	86.00	2.42	2.26
	4	с	12.51	1.24	2.25	17.50	85.00	2.00	0.58
	5	в	12.21	1.19	1.75	16.80	151.00	1.85	1.28
	6	A	14.22	3.99	2.51	13.20	128.00	3.00	3.04
	7	с	12.82	3.37	2.30	19.50	88.00	1.48	0.66
	8	в	12.47	1.52	2.20	19.00	162.00	2.50	2.27
	9	в	12.37	1.63	2.30	24.50	88.00	2.22	2.45
	10	с	12.81	2.31	2.40	24.00	98.00	1.15	1.09
	11	A	13.73	1.50	2.70	22.50	101.00	3.00	3.25
	12	с	13.36	2.56	2.35	20.00	89.00	1.40	0.50
	13	A	13.58	1.66	2.36	19.10	106.00	2.86	3.19
	14	в	11.45	2.40	2.42	20.00	96.00	2.90	2.79
	15	с	13.23	3.30	2.28	18.50	98.00	1.80	0.83
	16	A	13.83	1.65	2.60	17.20	94.00	2.45	2.99
	17	в	11.82	1.72	1.88	19.50	86.00	2.50	1.64
	18	в	11.84	2.89	2.23	18.00	112.00	1.72	1.32
	19	в	12.42	4.43	2.73	26.50	102.00	2.20	2.13
	20	A	14.10	2.02	2.40	18.80	103.00	2.75	2.92
Learning method	21	A	14.20	1.76	2.45	15.20	112.00	3.27	3.39
MethodName=Improved ChAID (Tsc A	22	в	12.72	1.81	2.20	18.80	86.00	2.20	2.53
MethodClassName=TArbreDecisionI	23	с	13.08	3.90	2.36	21.50	113.00	1.41	1.39
Hdl=8	24		11.96	1.09	2.30	21.00	101.00	3.38	2.14
Split=0.001	1		12.07	2.16	2.17	21.00	85.00	2.60	2.65
TypeBonferroni=1	26	в 💋	11.65	1.67	2.62	26.00	88.00	1.92	1.61
ValueBonferroni=1	27	в	12.64	1.36	2.02	16.80	100.00	2.02	1.41
j sampling=u 💌	28	с	13.27	4.28	2.26	20.00	120.00	1.59	0.69
Examples selection	29		13.78	2.76	2.30	22.00	90.00	1.35	0.68 👡
0 examples idle		Ъ <u>/</u>	1	1	1	1	1	1	>
	Editing	IEW.FDM				Attribu	tes : 14	Examples : 1	00 /
Improved ChAID (Tschuprow Goodness o	of Split)								

In the next step, we want to specify the target attribute (TYPE) and the input ones (ALCOHOL to PROLINE). We click on the ANALYSIS / DEFINE CLASS ATTRIBUTE menu.

💦 Sipina Research Version - [Le	arning set	editor]								
🔉 File Edit Data Statistics Induc	tion method:	Analysis Vi	ew Windo	w Help						- 8 ×
15 16 De les		Define cla	iss attribute			-				
		Select act	ive exampl	es		$\rightarrow$		,		
Attribute selection		Sot woigh	r Gold		Acid	Ash	Ash_Alcali	nil Magnesium	Total_Ph	enol;Flava 📥
Addibute selection	1	Set priors	ic field			1.95	14.80	136.00	1.90	1.85
	2	Set costs				2.46	20.50	116.00	2.96	2.78
	3	Set positi	ve clace val	10		2.00	19.00	86.00	2.42	2.26
	4		10 0035 10	00111		2.25	17.50	85.00	2.00	0.58
	5	Learning.				1.75	16.80	151.00	1.85	1.28
	<u>ь</u>	Stop and	usia.			2.51	13.20	128.00	3.00	3.04
	/	Attribut	e selectio	n						
	<u> </u>	Class -		_				ariables		
	3	— Туре						ype		
	10	· · · · ·					<u> </u>	alic_Acid		
	12	Attribut	es				A	sh T		
	13	- Alcoho	I			~	M	sn_Aicalinity agnesium		
	14	Malic_/	Acid				1	otal_Phenols		
	14	Ash_A	calinity				E N	avanoids onflavanoid Phe	nols	
	16	LA Total I	sium Phenols		~			roanthocyanins		
	17	Flavan	oids	<	-			olor_Intensity		
	18	- Nonflar B Proant	vanoid_Pher	ols	1			D280/0D315		
	10	Color_I	ntensity					roline		
	13	P Hue								
Learning contract	20									
MethodNamo-Improved ChAID (Tec -	22		nly discrete	2						
MethodClassName=TArbreDecision	23		nly continu	ous						
Hdl=8	24	E © B	oth							
Merge=0.05	24									
TupeBonferroni=1	20							🔹 🖌 🗸 🕻	23	🗙 Annuler
ValueBonferroni=1	20		13.64	1.26		2.02	16.90	100.00	12.0	1.41
Sampling=0 💌	28		13.04	1.30		2.02	20.00	120.00	1.50	0.69
Examples selection	20	Lo Contraction	13.27	7.20		2.20	22.00	90.00	1.35	0.68
100 examples selected	23	10	15.70	2.70		2.50	22.00	30.00	1.55	0.00
0 examples idle										>
1	Editing N	IEW.FDM					Attrib	utes : 14	Examples	:100 //
Improved ChAID (Tschuprow Goodness o	f Split)									

The variable selection is summarized in the left part of the main window.

Then, we can launch the learning process by clicking on the ANALYSIS / LEARNING menu.

🔉 Sipina Research Version - [Le	arning set	editor]								
🔉 File Edit Data Statistics Induc	tion method	Analysis	view Windov	/ Help						_ 8 ×
<u> 🖹 📑</u> 🐴 🖻	-	Define o Select a	lass attribute. ctive example:	 						
×					Acid	Ash	Ash_Alcal	inil Magnesium	Total_Pher	iol Flava 🔨
	1	Set weig	pht held			1.95	14.80	136.00	1.90	1.85 📃
	2	Set prio	rs			2.46	20.50	116.00	2.96	2.78
	3	Set cost	s			2.00	19.00	86.00	2.42	2.26
	4	Set posi	tive class valu	3		2.25	17.50	85.00	2.00	0.58
Malio Aoid	5	Learning	1	2 3		1.75	16.80	151.00	1.85	1.28
Ash	6	Stop an	alvsis	-74	~	2.51	13.20	128.00	3.00	3.04
Ash Alcalinitu	7					2.30	19.50	88.00	1.48	0.66
Magnesium	8	Classific	ation			2.20	19.00	162.00	2.50	2.27
	9	Test				2.30	24.50	88.00	2.22	2.45
- Flavanoids	10	LIFT P				2.40	24.00	98.00	1.15	1.09
Nonflavanoid Phenols	11					2.70	22.50	101.00	3.00	3.25
Proanthocyanins	12	Error measurements				2.35	20.00	89.00	1.40	0.50
Color_Intensity	13	E				2.36	19.10	106.00	2.86	3.19
Hue -	14	- Feature	selection			2.42	20.00	96.00	2.90	2.79
- C 0D280/0D315	15	Personn	al tests	•		2.28	18.50	98.00	1.80	0.83
- C Proline	16	A	13.83	1.65		2.60	17.20	94.00	2.45	2.99
J	17	в	11.82	1.72		1.88	19.50	86.00	2.50	1.64
Learning method	18	в	11.84	2.89		2.23	18.00	112.00	1.72	1.32
MethodName=Improved ChAID (Tsc 🔨	19	в	12.42	4.43		2.73	26.50	102.00	2.20	2.13
MethodClassName=TArbreDecisionI	20	A	14.10	2.02		2.40	18.80	103.00	2.75	2.92
Hdl=8	21	A	14.20	1.76		2.45	15.20	112.00	3.27	3.39
Split=0.001	22	в	12.72	1.81		2.20	18.80	86.00	2.20	2.53
TypeBonferroni=1	23	c	13.08	3.90		2.36	21.50	113.00	1.41	1.39
ValueBonferroni=1	24	B	11.96	1.09		2.30	21.00	101.00	3.38	2.14
	25	B	12.07	2.16		217	21.00	85.00	2.60	2.65
Examples selection	26	B	11.65	1.67		2.62	26.00	88.00	1.92	1.61
100 examples selected		-12	11.00	1.01		2.02	20.00	50.00	1.002	
o examples idle				_	_					2
J]	Editing	NEW.FDM					Attrib	utes : 14	Examples :	100
Improved ChAID (Tschuprow Goodness o	f Split)									1

The obtained decision tree is rather simple.



#### 3.4 Applying the decision tree on the unlabeled sample

In order to apply the learned tree on the unlabeled sample, the data file must be in FDM format. Sipina needs only the variables which are selected into the decision tree. It makes the correspondence by using the names of the variables (case sensitive). We click on the ANALYSIS / CLASSIFICATION / **ON OTHER DATASET** menu, we select wine\_data\_deployment.fdm.



The dataset is loaded into a new visualization grid, and the column with the predicted values is added (C\_TYPE in our example).

😹 Sipina Research Version													
Induction method Analysis View Wir	ndow	Help											
🖸 🗎 🐂 🐂 🐚												/	
Attribute selection	8	Learnin	g set edi	tor							B		^
Class attribute			Type	Alcoho		Malic	Acid	Ash	1	Ash Alca	Magne	sium Total R	henol:Elay
D Type	1		1300	40.99	~	0.00	Hold	4.05		14 90	126.00	1.00	1.85
🖻 🗊 Predictive attributes													2.76
C Alcohol		🔊 An	other dat	aset :	<u>_/Ja</u>	taMin	ing₩	ataba	ises_fo	or_min	/···· [=		2.26
Malic_Acid	4		-	nensi	ue		OD280.	/OD31[1	Proline	C_Typ	e	^	0.58
Ash	5	1	1.90	1	1.71		2.87		407.00	в			1.26
Ash_Alcalinity	6	2	2.45	1	1.33		2.26		495.00	в			3.04
Total Phenole	7	3	11.75	5 (	).57		1.78		620.00	С			0.66
Elavanoids	8	4	4.80	0	).92		3.22		1195.00	) A			2.27
Nonflavanoid Phenols	9	5	4.28	0	).91		3.00		1035.00	) A			2.45
Proanthocyanins	10	6	8.90	1	i.12		3.10		1260.00	) A			1.05
Color_Intensity	11	7	3.84	1	.23		2.87		990.00	A			3.25
- C Hue	12	8	3.52	1	.12		3.82		845.00	A			0.50
- C 0D280/0D315	13	9	5.40	1	.25		2.73		1150.00	) A			3.19
Proline	14	10	3.60	1	.05		2.65		450.00	в			2.79
	15	11	3.95	1	.02		2.77		1285.00	) A			0.83
Learning method	16	12	2.20	1	.31		2.72		630.00	В			2.95
MethodName=Improved ChAID (Tsc 🔨	17	13	2.80		).75		3.64		380.00	С			1.64
MethodClassName=TArbreDecisionI	18	14	8.66	0	).74		1.80		750.00	С			1.32
Merge=0.05	19	15	4.70	1	.04		3.59		1035.00	) A			2.13
Split=0.001	20	16	6.60	1	.13		2.57		1130.00	) A			2.92
TypeBonferroni=1	21	17	4.60	1	.19		2.30		678.00	в		-	3.35
Sampling=0	22	18	3.94	0	1.69		2.84		352.00	C			2.53
Examples selection	23				_	_		_	_		_		1.39
100 examples selected	24		в	11.96		1.09		2.30		21.00	101.00	3.38	2.14
0 examples idle	25		в	12.07		2.16		2.17		21.00	85.00	2.60	2.65 🤜
	<												>
Improved ChAID (Tschuprow Goodness of	Split)												Time : 94 //

The predicted column is defined as the class attribute, it is a discrete attribute with the same values.

### 3.5 Retrieving the predicted column

We want to retrieve the predicted column and copying it into the third sheet of the Excel workbook. The aim is to compare latter the observed values and the predicted values of the target attribute. To do this, we activate the contextual menu by (right) clicking on the header of the column. We select the COPY SELECTION menu.

😹 Sipina Research Version										
Induction method Analysis View Win	idow He	lp								
15 🚯 📴 🦡 🖪										
	r									
Attribute selection	💦 Lea									<u></u>
🖃 🗊 Class attribute		Type		Noohol	Malic Acid	Ash	Ash Al	alini Magne	sium Total P	henol
<b>D</b> Туре	1	_ 0		1.22	0.00	1.05	44.90	126.00	1.00	1.85
Predictive attributes	2	Another	datas	ot · D·\Da	laMining\[	atabaso	e for mir	vina)		2.76
Alcohol Malia Asid	3	Another	uatas	st . D. Da	amingu	/atabase	s_101_1111		<u>مالات</u>	2.26
Ash	4		Color_Int	tensilHue	OD280	0/OD31 Pro	line C	Туре	=ile	▶ 0.58
Ash Alcalinity	5 1		1.90	1.71	2.87	407	.00 8		Cut	1.26
C Magnesium	6 4		2.43	0.57	1.78	480			Copy selection	3.04
C Total_Phenols	$\frac{7}{4}$		4.80	0.92	3.22	119	1500 A		Paste	
Flavanoids	<u> </u>		4.28	0.91	3.00	103	5.00 A		Replace	2.45
Nonflavanoid_Phenois	10 6		8.90	1.12	3.10	126	0.00 A		Fyamples	► 1.05
Color Intensity	11 7		3.84	1.23	2.87	990	).00 A		variables	3.25
- C Hue	12 8		3.52	1.12	3.82	845	5.00 A			0.50
	13 9		5.40	1.25	2.73	115	60.00 A			3.19
Proline	14 1	0	3.60	1.05	2.65	450	).00 🛛 🖪			2.75
<u> </u>	<u>15</u>	1	3.95	1.02	2.77	128	85.00 A			0.85
Learning method	<u>16</u> 1	2	2.20	1.31	2.12	290				2.95
MethodName=Improved ChAID (Tsc A	17 1	4	2.00	0.73	1.80	750				1.64
Hdl=8	18 1	5	4.70	1.04	3.59	103	5.00 A			1.32
Merge=0.05	20 1	- Б	6.60	1.13	2.57	113	0.00 A			2.10
TypeBonferroni=1	21 1	7	4.60	1.19	2.30	678	3.00 B			3.35
ValueBonferroni=1	22 1	в	3.94	0.69	2.84	352	2.00 C		~	2.53
Samping=U	23									1.39
Examples selection	24	в	1	1.96	1.09	2.30	21.00	101.00	3.38	2.14
0 examples idle	25	в	1	2.07	2.16	2.17	21.00	85.00	2.60	2.65 🤜
	<									>
Improved ChAID (Tschuprow Goodness of	Split)									Time : 94 //

Into Excel, we select the third sheet (Data.Deployment.Class.Value) and we paste the dataset.

<b>N</b>	Aicrosoft Exce	el - wine_dep	oloyment.xls	7				
8	<u>F</u> ichier <u>E</u> ditio	n <u>A</u> ffichage	Insern 🔔	na <u>t O</u> utils <u>D</u>	onnées Fe <u>n</u> êt	re <u>?</u> Tanag	ra Sipina	- 8 ×
D	🚔 🔲 🔒 🎙	5 🗠 🖏	2   <b>)</b>	- 🚿 🗠 -	CH + 🍓 Σ	- 21 21 🛍	100%	- 🛛 🗸
Aria	al	- 8 -	6 S 🗏		S % 000	*28 208 €≣ €	😑 🛛 🗸 🕭	- A -
-	C1 -	f≱ C				100 110 1.		<b>—</b> •
	A	B		D	E	F	G	Н
1	ID	Туре	C_Type					
2	1	в	в					
3	2	B	B					
4	3	С	С					
5	4	A	A					
<u> </u>	5	A	A					
	6 7	A .	A					
- ä		A A	A A					
10	9	A A	A A					
11	10	B	B					
12	11	A	A					
13	12	в	в					
14	13	в	с					
15	14	С	С					
16	15	A	A					
17	16	A	A					
18	17	в	в					
19	18	в	с					
20	19	A	A					
21	20	C	C					
22	21	8	8					
23	22	B	A A					
25	23	8	8					
26	24	A	A					
27	26	B	B					
14	► ► ∠ Data	.Deployment	Data.Deploy	ment.Class.V	alue			
Des	sin 🕶 🔓 🛛 Forn	nes automatique	s • 🔪 📜	○ 🖾 ᆀ :		- <u>- A</u>	• = = =	
Prêt	_						NUM	

We utilize these values below in order to evaluate the "true" model accuracy.

## 4 Resampling error rate evaluation

We want to estimate the generalization error rate using the labeled sample only (Data.Learning). Because the size of our dataset is rather small, we cannot reasonably subdivide it into train and test samples for learning and assessing a model. Thus, we use the bootstrap resampling approach.

#### 4.1 Bootstrap

To launch the bootstrap process with Sipina, we must stop the current analysis before. We click on the ANALYSIS / STOP ANALYSIS menu.

🔉 Sipina Reseau	rch Version									ſ	
Induction method	Analysis View Wir	ndow Hel	n								
	Define class attrib	uto									
山山中	Select active exam	noles									
	Soloce dealto oxali	piositi	-								^
Attribute selection	Set weight field		nng s	eteano	or						
E 🗊 Ulass attrib	Set priors		Тур	e	Alcohol	Malic_Acid	Ash	Ash_Alcal	inil Magnesiu	m Total_Ph	nenol:Flav
Predictive	Set costs				40.92	0.00	14.05	14.00	126.00	1 00	1.85
	Set positive class	value	_\noth	er data	iset : D:\Da	ntaMining\I	Databases,	_for_minir	ıg\ 🗖		2.78
Malic_	Learning			Color	IntensilHue	0028	0/OD31 Prolin	е Сту	me		2.26
C Ash	Stop analysis	2	3	1.90	1.71	2.87	407.0	)0 B			1.56
C Ash_A	Classification	- 71	N	2.45	1.33	2.26	495.0	DO B			3.04
- C Magne	Test			11.75	0.57	1.78	620.0	00 C			0.66
I otal_I			-	4.80	0.92	3.22	1195	.00 A			2.27
Nopfla	LIFT ROC curve			4.28	0.91	3.00	1035	.00 A			2.45
Proant	Error measuremen	its I		8.90	1.12	3.10	1260	.00 A			1.05
Color_1	En alema and address			3.84	1.23	2.87	990.0	00 A			3.25
C Hue	Feature selection			3.52	1.12	3.82	845.0	00 A			0.50
- C 0D280	Personnal tests	1	•	5.40	1.25	2.73	1150	.00 A			3.15
🔤 🚺 Proline		14 14		3.60	1.05	2.65	450.0				2.75
ļ		15 11		3.95	1.02	2.77	1285	.UU A			0.83
Learning method		16 12		2.20	0.75	2.12	200.0				2.95
MethodName=Impro	oved ChAID (Tisc 🔺	17 13		8.66	0.75	1.80	750.0	0 C			1.64
Hdl=8	- Albiebecision	18 15		4 70	1.04	3.59	1035				1.32
Merge=0.05	=	20 16		6.60	1.13	2.57	1130	.00 A			2.10
TypeBonferroni=1		20 17		4.60	1.19	2.30	678.0	00 B			3.30
ValueBonferroni=1	_	22 18		3.94	0.69	2.84	352.0	00 C		🗸 🕹	2.55
Sampling=0	<b>×</b>	23									1.35
<ul> <li>Examples selection</li> </ul>	·	24	в		11.96	1.09	2.30	21.00	101.00	3.38	2.14
100 examples select	ted	25	в		12.07	2.16	2.17	21.00	85.00	2.60	2.65 🥃
o champios idio		<	-						i	i	>
, Improved ChâiD (Tsr	chuprow Goodpess of	Solit)									Time · 94
Improved Crimito (Tst	chaprow aboundss of	Spiic)								1	11110 . 97 //

Then, we select the ANALYSIS / ERROR MEASUREMENTS / .632 BOOTSTRAP menu<sup>6</sup>. A dialog box appears. We ask 20 replications. We click on the OK button.

<sup>&</sup>lt;sup>6</sup> See <u>http://eric.univ-lyon2.fr/~ricco/cours/slides/resampling\_evaluation.pdf</u> ; and <u>http://bioinformatics.oxfordjournals.org/cgi/content/full/21/15/3301</u>

🔉 Sipina Research Version											
File Edit Data Statistics Induction	method	Analysis	View	Window H	Help	_					
沓 🐴 🛤 🐚		Define	e class a	ttribute							
×		Select	active	examples							
Attribute selection	👗 Le	Set we	eight fie	ld							<u> </u>
🖃 🗊 Class attribute		Set pri	iors			/alic Acid	Ash	Ash Alcali	nil Maanesium	Total Phe	nol:Fla
D Type	1	Set co	sts			).99	1.95	14.80	136.00	1.90	1.8
🖻 📭 Predictive attributes	2	Set po	sitive d	lass value		.73	2.46	20.50	116.00	2.96	2.7
	3					).92	2.00	19.00	86.00	2.42	2.2
Malic_Acid	4	Learni	ng			.24	2.25	17.50	85.00	2.00	0.5
Ash	5	Scop a	Inalysis			.19	1.75	16.80	151.00	1.85	1.2
Asn_Alcalinity	6	Classif	ication		×	3.99	2.51	13.20	128.00	3.00	3.0
Total Phenols	7	Test				3.37	2.30	19.50	88.00	1.48	0.E
Flavanoids	8					.52	2.20	19.00	162.00	2.50	2.2
Nonflavanoid Phenols	9		- ROC C	urve		.63	2.30	24.50	88.00	2.22	2.4
Proanthocyanins	10	Error r	measure	ements	►	Cross-Va	lidation		98.00	1.15	1.0
Color_Intensity	11	Easter				.632 Boo	tstrap		101,00	3.00	3.2
	12	Featur	re selec	tion	*	Multi-L.Z		d-M	89.0	1.40	0.5
- C 0D280/0D315	13	Persor	nnal tes	ts	►	Multiple C	ross-vall	dation	106.00	2.86	3.1
Proline	14	В		11.45		.032+ DU	ocstrap.		96.00	2.90	2.7
Learning method	15	С		13.23		Stratified			11 <b>V</b>	1	
MethodName-Improved ChAID (Tac	16	A		13.83		Clustered	.632	l bootstrap	parameter		$\mathbf{X}$
MethodClassName=TArbreDecisionI	17	в		11.82		Die e Heui	- Ber	plication			
Hdl=8	<u>18</u>	В		11.84		DIdS-VdFk	110			•	
Merge=0.05	<u>19</u>	в		12.42		4.43		20		-	1
TypeBonferroni=1	20	A		14.10		2.02					
ValueBonferroni=1	21	A		14.20		1.76			/ nxm2	Y Annul	
Sampling=0 💌	22	в		12.72		1.81			W	Annui	
Examples selection	23	С		13.08		3.90	2.36	21.50	113.00	1.41	1.3
100 examples selected	24	в		11.96		1.09	2.30	21.00	101.00	3.38	2.1
u examples idle	25	IR		12.07		216	2.17	21.00	85.00	2.60	2 6
					_						2
Improved ChAID (Tschuprow Goodness o									Tir	ne:94 /	

The process is rather quick. The decision tree learning algorithm is in general fast and we use a small dataset. We obtain a confusion matrix and the related error rate. We have also the confusion matrix for each trial. But they are not really useful.

🔉 Sipina Research Version									
Induction method Analysis View Wir	ndow Help								
🖰 🗄 🛤 🐂									
X	<u> </u>								
Attribute selection	🔊 Lear	ning set edit							
🖃 🗊 Class attribute		Туре	Alcohol	Malic Acid	Ash	Ash Alcalini	Magnesium	Total Pheno	l:Fla
🛛 🕖 Туре	1	В	12.33	0.99	1.95	14.80	136.00	1.90	1.8
Predictive attributes	2	A	13.56	1.73	2.46	20.50	116.00	2.96	2.7
Alcohol	3	B	12.00	0.92	2.00	19.00	86.00	2.42	2.2
Malic_Acid	4	S. 20-boots	trap : NEW	.FDM				2.00	0.5
Ash	5	Tupe						1.85	1.2
Ash_Alcalinity	6	Operrall and	0			aan ang Émaa		3.00	3.0
Total Phenols	7		ODEITAL SCG	11101 1-1				1.48	0.6
Flavanoide	8		B.	A	C			2.50	2.2
Nonflavanoid Phenols	9	B	11.50	1.60	1.95			2.22	2.4
Proanthocyanins	10	C.	0.10	10.25	7.75			1.15	1.0
Color Intensity	11	<u> </u>	0.50	0.00				3.00	3.2
- Hue	12							1.40	0.5
C 0D280/0D315	13							2.86	3.1
C Proline	14							2.90	2.7
	15							1.80	0.8
Learning method	16	<u> </u>						2.45	2.9
MethodName=Improved UhAID (1 sc A MethodClassName=TArbreDecision	17	Cost : 0.1422						2.50	1.6
Hdl=8	18	в	11.84	2.89	2.23	18.00	112.00	1.72	1.3
Merge=0.05	19	в	12.42	4.43	2.73	26.50	102.00	2.20	2.1
Split=0.001	20	A	14.10	2.02	2.40	18.80	103.00	2.75	2.9
ValueBonferroni=1	21	A	14.20	1.76	2.45	15.20	112.00	3.27	3.3
Sampling=0 💌	22	в	12.72	1.81	2.20	18.80	86.00	2.20	2.5
Examples selection	23	С	13.08	3.90	2.36	21.50	113.00	1.41	1.3
100 examples selected	24	в	11.96	1.09	2.30	21.00	101.00	3.38	2.1
0 examples idle	25	в	12.07	2.16	2.17	21.00	85.00	2.60	2.6 💙
	<								>
Improved ChAID (Tschuprow Goodness o	Split)							Time :	125 r //

The estimated generalization error rate using the bootstrap procedure is 0.1422 i.e. when we apply this classifier on an unlabeled case, the misclassification probability is 14.22%.

#### 4.2 Checking out on the generalization sample

In a normal situation, the labels of the instances in the generalization sample are not available. One of the aim of our experiments is to compare the bootstrap (more generally resampling scheme) error rate estimation with the "true" error rate when we deploy the model.

We come back to our Excel workbook. We use the pivot table<sup>7</sup> in order to cross-tabulate the observed labels and the predicted labels into the third sheet (DATA.DEPLOYMENT.CLASS.VALUE).

Nombre de Type	C_Type				
Туре	А	В	}	С	Total
A		28	1		29
В		2	23	4	29
С			4	16	20
Total		30	28	20	78

The measured error rate is  $\varepsilon = \frac{1+2+4+4}{78} = 14.10\%$ 

The error rate obtained by bootstrap on the learning sample seems to be a good estimation of the "true" error rate measured on the generalization sample. However, the precision of the estimation is not always as well as on our dataset.

## 5 About the other learning methods

Sipina is especially intended to decision tree. But other supervised learning algorithms are also available. We can use the same process in order to create a classifier and apply it on unlabeled instances.

Let's take a Linear Discriminant Analysis<sup>8</sup> (LDA) for instance.

We stop the current analysis by clicking on the WINDOW / CLOSE ALL menu. Then we click on the INDUCTION METHOD / STANDARD ALGORITHM menu. A dialog box appears, we select the DISCRIMINANT ANALYSIS tab. We select the LINEAR DISCRIMINANT ANALYSIS method.

<sup>&</sup>lt;sup>7</sup> http://en.wikipedia.org/wiki/Pivot\_table

<sup>&</sup>lt;sup>8</sup> <u>http://en.wikipedia.org/wiki/Linear\_discriminant\_analysis</u>



Another dialog box appears. We validate the default settings.

Parameters for linear Discrimina	ant Analysis 🛛 🔀
Priors <u>U</u> ser defined <u>Same of all classes      Upconditional Class Distribution </u>	Costs C User defined C Symmetrical costs
	V 012443

We launch the learning process by clicking again on the ANALYSIS / LEARNING menu.

The classifier is described in a new window. There are: (1) the conditional mean of each descriptor according to the values of the target attribute (A, B and C); (2) the correlation matrix; (3) the classification functions which are used to assign the predicted value when we deploy the model.

Certainly, some important indications are missing, such as global significance of the model or individual significance of the descriptors. It is one of the reason for which I advise to use Tanagra (<u>http://eric.univ-lyon2.fr/~ricco/tanagra/</u>) for the other supervised learning methods than the decision tree learning<sup>9</sup>.

<sup>&</sup>lt;sup>9</sup> In effect, Tanagra implements also decision tree learning algorithms (CART, C<sub>4.5</sub>, ID<sub>3</sub>, etc.). But it does not supply, currently, the same interactive capabilities as Sipina when we explore a tree (choosing manually the split attribute, extracting the examples or computing descriptive statistics on each node, etc.).

IN 188 IN 18	2437203 - 33537233												
With the exclusion	8-												
- Pa Class attribute	🚆 🚨 Linear d	liscrimin	ant analysi	s results							- <u>()</u>		
	7		Conditional	average					Corre	elation matrix			
Predictive attributes	-	в	A	C			Alcohol	Malic_Acid	Ash	Ash_Alcalinity	y Magneolom	Total_Phenois	Flavanoids
Alcohol	Alcohol	12.2729	13.8627	13.1379		Alcohol	1.0000	0.0970	0.2108	-0.3018	0.2844	0.3236	0.2338
Malic Acid	Malic_Acid	1.8660	2.0887	3.3207		Malic_Acid	0.0970	1.0000	0.2885	0.2935	0.0108	-0.4086	-0.4499
Ash	4 Ash	2.2274 2.4600 2.4339		Ash	0.2108	0.2885	1.0000	0.4492	0.2755	0.1307	0.1082		
Ash_Alcalinity	5 Ash_Alcalinity	19.919D	16.8067	21.4107		Ash_Alcalinity	-0.3018	0.2935	0.4492	1.0000	-0.1487	-D.363D	-0.3612
- C Magnesium	6 Magnesium	96.2857	109.8333	100.3571		Magnesium	0.2844	0.0108	0.2755	-0.1487	1.0000	0.1827	0.1690
C Total_Phenols	Z Total_Phenois	2.2698	2.9203	1.6650		Total_Phenois	0.3236	-0.4086	0.1307	-0.3630	0.1827	1.0000	0.8603
E Flavanoids	8 Flavanoids	2.1214	3.0350	0.7886		Flavanoids	0.2338	-0.4499	0.1082	-0.3612	0.1690	0.8603	1.0000
C Nonflavanoid_Phenols	9 Nonflavanoid_	0.3621	0.2747	0.4325		Nonflavanoid	-0.1626	0.2929	0.1393	0.3291	-0.2974	-0.4094	-0.4913
C Proanthocyanins	1 Proanthocyanin	1.6040	1.9283	1.1379		Proanthocyani	0.2363	-0.3359	-0.0184	-0.2664	0.3472	0.6004	0.6363
Color_Intensity	1 Color_Intensity	3.2645	5.7170	7.3357		Color_Intensity	0.5754	0.3069	0.2905	0.0657	0.1797	-0.0372	-0.1642
- C Hue	1 Hue	1.0388	1.0423	0.6946		Hue	-0.0828	-0.5597	-0.2094	-0.3580	0.0620	0.4668	0.5753
C 0D280/0D315	1 0D280/0D315	2.8098	3.2330	1.6968		OD280/OD315	0.0857	-0.4165	-0.0217	-0.3472	0.0756	0.6829	0.7889
C Proline	1 Proline	524.0714	1121.5667	644.6429		Proline	0.6591	-0.1379	0.2657	-0.4494	0.4301	0.5079	0.4560
						Discrim	inant functi	ion					
		в	A	C									
	1 Alcohol	67.0456	75.8063	68.8171									
	1 Alcohol 2 Malic_Acid	67.0456 2.4433	75.8063 3.6496	68.8171 4.4530									
	1 Alcohol 2 Malic_Acid 2 Ash	67.0456 2.4433 51.4372	75.8063 3.6496 65.1975	68.8171 4.4530 56.5170									
	1 Alcohol 2 Malic_Acid 2 Ash 2 Ash_Alcalinity	67.0456 2.4433 51.4372 -0.3279	75.8063 3.6496 65.1975 -1.6040	68.8171 4.4530 56.5170 -0.2489									
	1 Alcohol 2 Malic_Acid 2 Ash 2 Ash_Alcalinity 2 Magnesium	67.0466 2.4433 51.4372 -0.3279 0.4379	75.8063 3.6496 65.1975 -1.6040 0.4176	68.8171 4.4530 56.5170 -0.2489 0.3992	<u>9</u>								
	1 Alcohol 2 Malic_Acid 2 Ash 2 Ash_Acalinity 2 Magnesium 2 Total_Phenols	67.0456 2.4433 51.4372 -0.3279 0.4379 -4.6791	75.8063 3.6496 65.1975 -1.6040 0.4176 -7.6797	68.8171 4.4530 56.5170 -0.2489 0.3992 -2.9283	3								
	1 Acohol 2 Malio_Acid 2 Ash 2 Ash_Acalinity 2 Magnesium 2 Tota_Phenols 5 Flavanoids	67.0466 2.4433 51.4372 -0.3279 0.4379 -4.0791 6.5810	75.8063 3.6496 65.1975 -1.6040 0.4176 -7.6797 10.7379	68.8171 4.4530 56.5170 -0.2489 0.3992 -2.9283 -0.7734	3								
saming method	1         Acohol           2         Malio_Acid           2         Ash           2         Ash_Alcalinity           2         Magnesium           2         Total_Phenols           2         Flavanoids	67.0466 2.4433 51.4372 -0.3279 0.4379 -4.6791 6.5810 5.3319	75.8063 3.6496 65.1975 -1.6040 0.4176 -7.6797 10.7379 -6.1810	68.8171 4.4530 56.5170 -0.2489 0.3992 -2.9283 -0.7734 -4.1768	3								
saming method ethodName=Linead discriminant anely: ethodNameLinearDia	1 Alcohol 2 Matic_Acid 2 Ash 2 Ash_Alcalinity 2 Magnesium 2 Total_Phenois 2 Flavanoids 2 Nonflavanoid_1 2 Poanthocyanir	67.0466 2.4433 51.4372 -0.3279 0.4379 -4.6791 6.5810 5.3319 -8.8987	75.8063 3.8496 65.1975 -1.6040 0.4176 -7.6797 10.7379 -6.1810 -11.1828	68.8171 4.4530 56.5170 -0.2489 0.3992 -2.9283 -0.7734 -4.1768 -9.8899	3								
aming method ethodName=Linear discriminant analys ethodDasName=TLinearDA l=11	1 Alcohol 2 Matic_Acid 2 Ash 2 Ash 2 Ash_Acalinity 2 Magnesium 2 Total_Phenols 2 Flavanoids 2 Nonflavanoid 2 Proanthocyanit Color_Intensity	67.0466 2.4433 51.4372 -0.3279 0.4379 -4.6791 6.5810 5.3319 -8.8987 -5.5660	75.8063 3.8496 65.1975 -1.6040 0.4176 -7.6797 10.7379 -6.1810 -11.1828 -5.9097	68.8171 4.4530 56.5170 -0.2489 0.3992 -2.9283 -0.7734 -4.1768 -9.8899 -3.1997	3								
earning method ethodName=Linear discriminant analy: ethodClassName=TLinearDA gi=11 size=2	1 Alcohol 2 Malic_Acid 2 Ash 2 Ash_Alcalinity 4 Magnesium 2 Total_Phenols 5 Flavanoids 9 Proanthocyanit 2 Color_Intensity 1 Hue 1 Hu	67.0466 2.4433 51.4372 -0.3279 0.4379 -4.6791 6.5810 5.3319 -8.8987 -5.5660 63.9350	75.8063 3.8496 65.1975 -1.6040 0.4176 -7.6797 10.7379 -8.1810 -11.1828 -5.9097 61.3630	68.8171 4.4530 56.5170 -0.2489 0.3992 -2.9283 -0.7734 -4.1768 -9.8999 -3.1997 -5.3.4904	3								
aming method sthodName=Linear discriminant analy: sthodElasName=TLinearDA j=11 ors=2 sts=1	1 Acohol 2 Malio_Acid 2 Ash 2 Ash_Atcalinity 4 Magnesium 2 Total_Phenols 7 Intal_Phenols 2 Rearthocyanic 2 Color_Intensity 1 Hue 0 D280/0 D315	67.0466 2.4433 51.4372 -0.3279 0.4379 -4.6791 6.5810 5.3319 -8.8987 -5.5660 63.9350 21.1008	75.8063 3.6496 65.1975 -1.6040 0.4176 -7.6797 10.7379 -6.1810 -11.828 -5.9097 61.3630 24.9293	68.8171 4.4630 56.5170 -0.2489 0.3992 -2.9283 -0.7734 -4,1768 -9.8999 -3.1997 53.4904 16.9103	3								
earning method ethodName=Linear discriminant analys ethodClassName=TLinearDA di=11 cors=2 sts=1	1         Acohol           2         Malic_Acid           2         Ach           2         Ach           2         Ach           2         Achenity           2         Magnesium           2         Total_Phenois           2         Flavanoids           1         Proanthocyanit           2         Color_Intensity           4         Hue           3         Polav070016	67.0456 2.4433 51.4372 -0.3279 0.4379 -4.6791 6.5810 5.3319 -8.8987 -5.5660 63.9350 21.1008 0.0077	75.8063 3.6406 65.1975 -1.6040 0.4178 -7.6797 10.7379 -6.1810 -11.1828 -5.9097 61.3630 24.9293 0.0299	68.8171 4.4530 66.5170 0.2489 0.3992 -2.9283 -0.7734 -4.1768 -9.8999 -3.1997 63.4604 16.9103 0.0111	<b>S</b>								
saming method ethodParse=Linear discriminant anely: ethodParsName=TLinearDA di=11 ios=2 atts=1	Acohol           Amio_Acid           Ash           Ash_Alcalinity           Arad_Phenols           Flavanoids           Flavanoids           Proathocyanit           Color_intensity           Hue           OD280/00316           Constant	67.0466 2.4433 51.4372 -0.3279 0.4379 4.6791 6.5910 5.3319 -8.8987 -5.5660 63.9350 21.1008 0.0077 -540.0945	75.8063 3.8496 65.1975 -1.8040 0.4176 -7.6797 10.7379 -6.1810 -11.1828 -5.9097 61.3630 24.9293 0.0299 -684.4073	68.8171 4.4530 56.5170 -0.2489 0.3992 -2.9283 -0.7734 -4.1768 -9.8999 -3.1997 53.4904 16.9103 0.0111 -681.0825	3								
saming method ethod Name-Linear discriminant analys ethod Lasst Name = LinearDA d=11 d=11 d=11 d=11 d=11 d=11 d=11 d=1	1 Alcohol 2 Malic_Aeid 2 Aeh, 4 Ach, 2 Aeh, 4 Calinity 2 Aeh, 4 Calinity 2 Magnesium 2 Total_Phenois 2 Flavanoid; 2 Proartheoyanic; 2 Proartheoyanic; 2 Proartheoyanic; 3 Proline 3 Color_intensity 4 Nue 4	67.0466 2.4433 51.4372 0.3279 0.4379 -4.8791 6.6510 6.3319 -8.8987 -5.5660 63.9360 21.1008 0.0077 -640.0945	75.8063 3.8496 65.1975 -1.8040 0.4176 -7.6797 10.7379 -6.1810 -11.1828 -5.9097 61.3630 24.9293 0.0299 -884.4973	88.8171 4.4530 56.5170 0.2488 0.3992 -2.0283 0.7734 4.1768 -9.8999 -3.1997 53.4904 16.9103 0.0111 -501.0825	3								
saming method ethodName=Linear discriminant analys ethodClassName=TLinearDA J=11 ors=2 sts=1 xamples selection Researche selection	1 Acohol 2 Malio_Acid 2 Ach, Acalinity 2 Ach, Acalinity 2 Magnesium 2 Tata_Mole 2 Flavanoids 2 Flavanoids 2 Flavanoids 2 Colo_Texeo 3 Constant 3 Constant	67.0466 2.4433 61.4372 -0.3279 0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.4379 -0.5560 -0.5560 -0.100 -0.1008 -0.1008 -0.1008 -0.1008 -0.1008 -0.1008 -0.1008 -0.4079 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -0.4009 -	75.8063 3.6466 65.1975 -1.6040 0.4176 -7.0797 10.7379 -6.1810 -11.1828 -5.9097 61.3630 24.9293 0.0299 -884.4073	88.8171 4.4530 65.5170 -0.2489 0.3992 -2.9283 -0.7734 -4.1768 -9.8999 -3.1897 -3.1897 -3.1897 -3.1897 -5.4804 18.9103 0.0111 -561.0825	3								
saming method ethod Name-Linear discriminant analys ethod Class Name-T LinearDA iora-2 zata=1 xamples selection 0 examples selected xamples file	1 Alcohol 2 Malic_Acid 2 Ash_Acalinty 2 Ash_Acalinty 4 Malg.esium 1 Tatal_Phenols 2 Flavanoida 2 Flavanoida 2 Flavanoida 2 Proamtooyan 9 Proamtooyan 9 On290/00316 9 Poline 3 Octostant	67.0466 2.4433 61.4372 -0.3279 0.4379 0.4379 0.4379 6.5810 6.5319 -8.8987 -5.5660 21.1008 0.0077 -5.40.0945	75.8063 3.8406 65.1975 -1.6040 0.4176 -7.6797 10.7379 -6.1810 -11.1828 -5.9097 61.3630 24.9293 0.0299 -884.4073	88.8171 4.4500 66.5170 -0.2488 0.3992 -2.9283 -0.7734 -4.1768 -9.8999 53.4604 16.9103 0.0111 -561.0825	3								

We can apply this model on the unlabeled sample. We click on the ANALYSIS / CLASSIFICATION / ON OTHER DATASET menu. We select « wine\_data\_deployment.fdm ».



A data visualization grid appears. In addition to the descriptors, we have now a new column with the predicted values of the LDA classifier.

nduction method Analysis View Wir	ndow Help											
D 🖻 📲 📥 🛼												
And the second second	×											
Attribute selection	🚆 🔊 Linea											
- Uass attribute	Condition	al average			Correlation matrix							
Predictive attributes	1	в	A	lc			-	Alcohol	Malic Acid	Ash	Ash Alcalinity	Magnesiu
	2 Alcohol	12.2729	13.8627	13.1379	_		Alcohol	1.0000	0.0970	0.2108	-0.3018	0.2844
Malic Acid	3 Malic Acid	1.8660	2.0887	3.3207	_		Malic Acid	0.0970	1.0000	0.2885	0.2935	0.0108
Ash	4	2 2274	1800	2 4220			Arb	0.2109	2000 n	1.0000	0.4402	0.2766
Ash_Alcalinity	🔊 Anothe	er dataset	\DataMi	ning\Datal	oases_fo	_mining\	dataset_for	_soft_de	/_and_com	parison\f-m	easu	
Magnesium		Magnesium	Total_Pheno	Flavanoids	Nonflava	noic Proanth	ocya <sup>l</sup> Color_Int	ensilHue	OD280/0	D31 Proline	C_Type	1
Flaumeide	40	139.00	3.30	2.89	0.21	1.96	3.35	1.31	3.50	985.00	B	_
Nonflavanoid Phanola	41	122.00	1.51	1.25	0.21	0.94	4.10	0.76	1.29	630.00	С	
Proanthocuanins	42	95.00	2.20	2.43	0.26	1.57	5.00	1.17	2.82	1280.00	A	
Color Intensitu	43	86.00	1.98	1.60	0.30	1.53	1.95	0.95	3.33	495.00	B	
Hue	44	108.00	2.00	2.09	0.34	1.61	2.06	1.06	2.96	345.00	в	
OD280/0D315	45	91.00	1.68	0.70	0.44	1.24	9.70	0.62	1.71	660.00	С	
Proline	46	96.00	1.79	0.60	0.63	1.10	5.00	0.82	1.69	515.00	С	
	47	98.00	2.45	2.43	0.29	1.44	4.25	1.12	2.51	1105.00	A	
	48	124.00	2.63	2.68	0.47	1.92	3.58	1.13	3.20	830.00	A	
	49	103.00	1.48	0.58	0.53	1.40	7.60	0.58	1.55	640.00	С	
	50	86.00	2.95	2.86	0.21	1.87	3.38	1.36	3.16	410.00	в	
	51	97.00	1.55	0.52	0.50	0.55	4.35	0.89	2.06	520.00	С	
	52	105.00	2.95	3.32	0.22	2.38	5.75	1.25	3.17	1510.00	A	
	53	112.00	1.48	1.36	0.24	1.26	10.80	0.48	1.47	480.00	С	1
exprise method	54	80.00	0.98	0.34	0.40	0.68	4.90	0.58	1.33	415.00	С	
AethodName-Linear discriminant analys	55	92.00	2.72	3.27	0.17	2.91	7.20	1.12	2.91	1150.00	A	
tethodClassName=TLinearDA	56	81.00	1.60	1.50	0.52	1.64	2.40	1.08	2.27	480.00	B	
Idl=11	57	86.00	2.62	2.65	0.30	2.01	2.60	0.73	3.10	380.00	в	
Priors=2	58	94.00	2.10	1.79	0.32	0.73	3.80	1.23	2.46	630.00	в	
03(3=1	59	88.00	2.30	0.92	0.50	1.04	7.65	0.56	1.58	520.00	С	_
	60	120.00	1.65	0.68	0.53	1.46	9.30	0.60	1.62	840.00	с	
	<										1.	
Examples selection	3	-										
00 examples selected	Proline	0.0077	0.0299	0.0111	_							
J examples idle	Constant	-540.0945	-684,4973	-561.0825								

Again, we can estimate the generalization error rate by using the bootstrap method. We stop the current analysis by activating the ANALYSIS / STOP ANALYSIS menu. Afterwards, we click on the ANALYSIS / ERROR MEASUREMENTS / .6<sub>32</sub> BOOTSTRAP menu.

🔊 Sipina Research Version
Induction method Analysis View Window Help
🖸 🗃 👺 🍓 🖕
Attribute selection       X       Learning set editor         Image: transmitted selection       Type       Alcohol       Malic_Acid       Ash       Ash       Ash       Ash       Ash       Alcohol       190       182         Image: transmitted selection       Image: transmitted selectin       Image: transmitted selectin

The estimated generalization error rate is 2.23%. LDA outperforms definitely the decision tree (IMPROVED CHAID) on the Wine dataset.

## 6 Conclusion

In this tutorial, we show how to apply a classifier on unlabeled sample with Sipina. We show also how to estimate the generalization error rate using a resampling scheme such as bootstrap.

We were able to check out the credibility of this estimation because we have in reality the true labels of the observations of the generalization sample. We note that it is rather accurate.