Goal

Show how to import a spreadsheet file format dataset (EXCEL 97 & 2000).

The main advantage of this approach is that it is possible to modify on the fly the data source without having to rebuild the stream diagram (!).

Dataset

ADULT dataset with 48842 examples and 15 attributes. We want to characterize the people who have a high income (more than 50K\$).

Handle a spreadsheet EXCEL file

Import

The first step is to import the dataset in TANAGRA. Build a new stream diagram according the usual method (File / New).

TANAGRA 1.1.2				E
File Diagram Window Help				
D 📽 🖩 🔛				
Data mir	ning diagram (empty)			
	Choose your d	ataset and start download		
	Diagram t	itle :		
	Default tit	e		
	Data minir	ng diagram file name :		
	D:\DataM	ining\Databases_for_mining\b	enchmark_datasets\adult\adu	
	Dataset (te	ext file) :		
			OK Cancel	Help
		1		
		Components		
Data visualization	Descriptive stats	Instance selection	Feature construction	Feature selection
Regression	Factorial analysis	Clustering	Spv learning	Meta-spy learning
Spv learning assessment	Association			
Correlation scatterplot	👺 Export dataset	🛃 Scatterplot	🚉 View dataset	🔛 View multiple scatterplot

In the dialog box enabling to select the file, several file format is now available, especially the EXCEL file format. Select this one and download the ADULT.XLS dataset.

Cases studies



Some precautions are necessary to import the file:

(1) only EXCEL 97 & 2000 file format are accepted;

(2) if there are several sheets in the workbook, the dataset must be in the first one;

(3) the dataset must be aligned in the left corner of the sheet i.e. first value (name of the first attribute) must be in the A1 cell;

(4) the first row corresponds to the name of attributes;

(5) there is no examples label;

(6) there is no missing values or column/row empty in the dataset.

Maximum number of examples and attributes that we can handle is limited by EXCEL capacities: 65534 examples (first row contains attributes name) and 256 attributes.

Importation is rather fast (7 seconds on a P4 at 3Ghz).

Cases studies

Handle a spreadsheet EXCEL file

🌋 TANAGRA 1.1.2 - [Dat	aset (adult.xls)]						X
🏆 File Diagram Componen	t Window Help					- 6	5 ×
D 🛩 🖪 🔛							
	Default title				Dataset ()	adult vis)	~
Dataset (adult.xls)					Param	neters	
			Database : D:\Da	taMining\Da	atabases_for_mini	ng\benchmark_datasets\adult\adult.xls	
					Res	utts	
			15 attribute(s) 48842 example(s	;)			
			Attribute	Category	Informations		
			age	Continue	-		
			workclass	Discrete	8 values		
			fnlwgt	Continue	-		
			education	Discrete	16 values		
			education_num	Continue	-		~
			Component	s			
Data visualization	Descriptive stats	Instar	nce selection	Featur	e construction	Feature selection	
Regression	Factorial analysis	C	lustering	Sp	ov learning	Meta-spv learning	
Spv learning assessment	Association						
+Correlation scatterplot	🛱 Export dataset	🚀 Sc	atterplot	🔛 \	/iew dataset	É.: View multiple scatterplot	

Characterization of the attribute CLASS

We use the GROUP CHARACTERIZATION component to describe CLASS. Before, insert the DEFINE STATUS in the diagram and set CLASS to TARGET, and the other one to INPUT. **Save the diagram**.

TANAGRA 1.1.2 - [Group caracterization 1]									
File Diagram Component Window Help									- é
📽 🔲 👫									
Default title				Grour	a caracterization 1				
Dataset (adult.xls)					Parameters				
😑 🏠 Define status 1	(
Group caracterization 1					Results				
				Descriptio	n of "class"				
	class	=less			class=n	nore			
	Examples			37155	Examples			11687	
	Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral	
	Continuous attributes				Continuous attributes				
	fnlwgt	1.4	190039.57	189664.13	education_num	73.5	11.60	10.08	
	capital_loss	-32.6	54.15	87.50	age	50.9	44.28	38.64	
	capital_gain	-49.3	147.01	1079.07	hours_per_week	50.3	45.45	40.42	
	hours_per_week	-50.3	38.84	40.42	capital_gain	49.3	4042.24	1079.07	
	age	-50.9	36.87	38.64	capital_loss	32.6	193.53	87.50	
	education_num	-73.5	9.60	10.08	fnlwgt	-1.4	188470.57	189664.13	
	Discrete attributes				Discrete attributes				
	marital_status=Never-married	70.5	41.40%	33.00%	marital_status=Married-civ-spouse	98.5	85.43%	45.82%	
	relationship=Own-child	49.9	20.10%	15.52%	relationship=Husband	89.2	75.69%	40.37%	
	sex=Female	47.4	38.82%	33.15%	sex=Male	47.4	84.86%	66.85%	
	relationship=Not-in-family	42.1	30.43%	25.76%	occupation=Exec-managerial	46.6	24.88%	12.46%	
	occupation=Other-service	34.3	12.70%	10.08%	education=Bachelors	39.9	28.35%	16.43%	
		С	omponents						
Data visualization Descriptive stats	Instance selection	Feature c	onstruction	n F	eature selection Regr	ression	1	Factorial analys	is
Clustering Spv learning	Meta-spv learning	Spv learning	g assessme	nt	Association				
Cross Tabulation Group caracterization Group caracterization	on di Non-parametrical cont stat di One-way ANOVA	Test (KW)	g assessme ⊈Univariat <u>U</u> Univariat	e continuo e discrete	Association us stat stat				

We see that the people with high income are married men, rather old, with a high level of education. These persons have a high capital gain (ok) and high capital loss (???).

While returning on the initial file (ADULT.XLS), we note that these two variables gather two types of information: if the person does not place are money, the profit (the loss) is equal to zero; and if they place their money, the variable indicates the amount of the profit (of the loss). Moreover, one individual having carried out a profit cannot carry out loss, and conversely.

	licrosoft Exce	el - adult.xls														×
	<u>F</u> ichier <u>E</u> dition	<u>A</u> ffichage <u>I</u> nse	ertion Forma <u>t</u> (<u>O</u> utils <u>D</u> onnées	Fe <u>n</u> être	2									- 8	×
	🖻 🔒 🚔	a 🔍 💞 🛛	X 🗈 🛍 ダ	10 × 01 ×		f _*	€ 2↓	Z↓		45 😨	» ▼	G	;,08	,00 ,00	8	•
-	L12 -	= 0							-			1		-		
	Н		J	K	L			M		N			0			
1	relationship	race	sex	capital gain	capital	oss	hours	per	wen	ative cou	untr	class	-			
2	Not-in-family	White	Male	2174		0			40 U	Jnited-Sta	ates	less				
3	Husband	White	Male	0		0			13 U	Jnited-Sta	ates	less				
4	Not-in-family	White	Male	0		0			40 U	Jnited-Sta	ates	less				
5	Husband	Black	Male	0		0			40 U	Jnited-Sta	ates	less				
6	Wife	Black	Female	0		0			40 C	≿uba		less				
7	Wife	White	Female	0		0			40 U	Jnited-Sta	ates	less				
8	Not-in-family	Black	Female	0		0			16 J	amaica		less				
9	Husband	White	Male	0		0			45 U	Inited-Sta	ates	more				
10	Not-in-family	White	Female	14084		0			50 U	Inited-Sta	ates	more				
11	Husband	White	Male	5178		0			40 U	Inited-Sta	ates	more				
12	Husband	Black	Male	0		0		{	30 U	Inited-Sta	ates	more				
13	Husband	Asian-Pac-Isl	Male	0		0			40 Ir	ndia		more				
14	Own-child	White	Female	0		0			30 U	Inited-Sta	ates	less				
15	Not-in-family	Black	Male	0		0		!	50 U	Inited-Sta	ates	less				
16	Husband	Asian-Pac-Isl	Male	0		0			40 U	Inited-Sta	ates	more				
17	Husband	Amer-Indian-B	Male	0		0			45 N	/lexico		less				
18	Own-child	White	Male	0		0			35 U	Inited-Sta	ates	less				
19	Unmarried	White	Male	0		0			40 U	Inited-Sta	ates	less				
20	Husband	White	Male	0		0		(50 L	Inited-Sta	ates	less				
21	Unmarried	White	Female	0		0			45 U	Inited-Sta	ates	more				
22	Husband	White	Male	0		0		f	60 L	Inited-Sta	ates	more				
23	Unmarried	Black	Female	0		0			20 U	Inited-Sta	ates	less				
24	Husband	Black	Male	0		0			40 U	Inited-Sta	ates	less				-
• •	🕨 🕨 🔪 adult	/					•								•	
Deg	sin + 🗟 🍈	Formes automa	tiques 👻 🔨 🍾		4 🗕	<u></u> -	- 🚄 -	<u>A</u> -	- =	i≡ ≓		1	-			
Prê	t											NUM				/

We want to create a new attribute which points out the fact of investing or not its money.

Cases studies

Handle a spreadsheet EXCEL file

	🖾 Microsoft Excel - adult.xls 📰 🗖 🔀										
	📳 Eichier Edition Affichage Insertion Format Qutils Données Fenêtre ?										
	□ 🚅 🗑 🗿 🕼 🖤 🐰 🛍 🛍 🚿 い・○→ 🍓 Σ 🕫 🖓 ટੈ↓ 🕌 🚛 ?? » G % 4% 🦄 ->										
1	M2 -	= =	SI(OU(K2>0:L2	2>0):"ves":"no			formul	2			
	Н	1	J	K	Ĺ	M	ioimuia	1 0	N.S.Mart	•	
1	relationship	race	sex	capital gain	capital loss	capital inve	sthours per we	native countr	class		
2	Not-in-family	White	Male	2174	0	yes	40	United-States	less		
3	Husband	White	Male	0	0	no	13	United-States	less		
4	Not-in-family	White	Male	0	0	no	40	United-States	less		
5	Husband	Black	Male	0	0	no	40	United-States	less		
6	Wife	Black	Female	0	0	no	40	Cuba	less		
7	Wife	White	Female	0	0	no	40	United-States	less		
8	Not-in-family	Black	Female	0	0	no	16	Jamaica	less		
9	Husband	White	Male	0	0	no	45	United-States	more		
10	Not-in-family	White	Female	14084	0	yes	50	United-States	more		
11	Husband	White	Male	5178	0	yes	40	United-States	more		
12	Husband	Black	Male	0	0	no	80	United-States	more		
13	Husband	Asian-Pac-Isl	Male	0	0	no	40	India	more		
14	Own-child	White	Female	0	0	no	30	United-States	less		
15	Not-in-family	Black	Male	0	0	no	50	United-States	less		
16	Husband	Asian-Pac-Isl	Male	0	0	no	40	United-States	more		
17	Husband	Amer-Indian-B	Male	0	0	no	45	Mexico	less		
18	Own-child	White	Male	0	0	no	35	United-States	less		
19	Unmarried	White	Male	0	0	no	40	United-States	less		
20	Husband	White	Male	0	0	no	50	United-States	less		
21	Unmarried	White	Female	0	0	no	45	United-States	more		
22	Husband	White	Male	0	0	no	60	United-States	more		
23	Unmarried	Black	Female	0	0	no	20	United-States	less		
24	Husband	Black	Male	0	0	no	40	United-States	less	-	
K	I I I Adult					4			•		
Deg	sin • 🗟 🙆	Formes automa	tiques 🕶 🔨 🦄		4 🗕 🖄	• 🔟 • 🛕 •		.			
Prê	t							NUM		1	

After having saved the workbook and closed EXCEL, close and open the diagram, the new version of the dataset is parsed and downloaded.

TANAGRA 1.1.2						
File Diagram Window Help						
🗅 📽 🔲 👪 🛛 Tanagra					? 🔀	
Data mir Regarder dans	: 🚞 adult		g 🕜 💆	•111 😒		
Mes documents récents Bureau Mes documents	adult.tdm					
Data visualizati Poste de travail						
Feature select						
Spv learning	Nom du fichier :	adult.tdm		1	Ouvrir	
E Cross Tabulatio	Fichiers de type :	Text data mining diagram (*.	.tdm)	v	Annuler	nt stat
]					>

R.R.

We note that:

- (1) diagram is preserved;
- (2) the new attribute was added in the dataset.

Pile Diagram Component: Window Help Hesuits Default bile Default bile Default bile Particular (adult.xts) Particul	🂯 TANAGRA 1.1.2 - [Dataset (adult.xls)]					
Default bile Results Image: Define status 1 16 attribute(s) Image: Define status 1 Image: Continue Image: Continue - Image: Continue	Tile Diagram Component Window Help					- 🗗 🗙
Default title Results Image: Define status 1 16 attribute(s) 4842 example(s) Attribute Category Informations age age Continue workclass Discrete 8 values milling Continue - workclass Discrete 16 values education Discrete 16 values education Discrete 7 values occupation Discrete 7 values occupation Discrete 7 values occupation Discrete 7 values occupation Discrete 5 values sex Discrete 2 values capital_gan Continue - native_country Discrete 2 values capital_lows Discrete 2 values capital_gan Continue - native_country Discrete 2 values class Discrete 2 values class Discrete 2 values class Discrete 2 values class Discrete 2 values </td <td>D 📽 🖶 👫</td> <td></td> <td></td> <td></td> <td></td> <td></td>	D 📽 🖶 👫					
Image: Second	Default title			Results		~
Group caracterization 1 Attribute Category Informations age Continue - workclass Discrete 8 values finiwgt Continue - education Discrete 16 values education_num Continue - marital_status Discrete 7 values occupation Discrete 14 values relationship Discrete 5 values sex Discrete 2 values capital_gain Continue - capital_gain Continue - capital_loss Continue - capital_loss Continue - capital_loss Continue - capital_gain Continue - capital_gain Continue - capital_gain Continue - capital_loss Continue - capital_gain Continue - capital_loss Continue - capital_loss Continue - capital_loss Continue - capital_loss Continue - capital_gain Continue - capital_gain Continue - capital_loss Continue - capital_loss Continue - capital_loss Continue - capital_loss Continue - capital_gain Continue - capital_loss Continue - capital_gain Continue - capital_gain Continue - capital_loss Continue - capital_loss Continue - capital_gain Continue - capital_loss Continue - capital_gain Continue - capital_loss Continue - capital_gain Continue - capit	🖃 🏢 Dataset (adult.xls) 🖶 🎇 Define status 1	16 attribute(s) 48842 example(s))			
age Continue - workclass Discrete 8 values fnlwgt Continue - education Discrete 16 values education_num Continue - marital_status Discrete 7 values occupation Discrete 7 values occupation Discrete 5 values race Discrete 5 values capital_loss Continue - capital_loss Continue - native_country Discrete 2 values Components - - Data visualization Descriptive stats Instance selection Feature construction Feature selection Regression Factorial analysis Clustering Spv learning Meta-spv learning Spv learning assessment Association	Group caracterization 1	Attribute	Category	Informations		
workclass Discrete 8 values fnlwgt Continue - education Discrete 16 values education_num Continue - marital_status Discrete 7 values occupation Discrete 7 values occupation Discrete 5 values race Discrete 5 values capital_loss Continue - capital_loss Continue - native_country Discrete 2 values cass Discrete 2 values Components - - Peature selection Regression Factorial analysis Spv learning Meta-spv learning Spv learning assessment Association Components - Cross Tabulation Group caracterization Chinear correlation		age	Continue	-		
fillingt Continue - education Discrete 16 values education_num Continue - marital_status Discrete 7 values occupation Discrete 6 values relationship Discrete 5 values sex Discrete 2 values capital_gain Continue - capital_loss Continue - native_country Discrete 2 values Discrete 2 values - Data visualization Descriptive stats </td <td></td> <td>workclass</td> <td>Discrete</td> <td>8 values</td> <td></td> <td></td>		workclass	Discrete	8 values		
education Discrete 16 values education_num Continue - marital_status Discrete 7 values occupation Discrete 14 values reac Discrete 5 values sex Discrete 2 values capital_gain Continue - capital_loss Continue - capital_loss Continue - capital_loss Continue - capital_loss Continue - native_country Discrete 2 values class Discrete 2 value		fnlwgt	Continue	-		
education_num Continue - marital_status Discrete 7 values occupation Discrete 14 values relationship Discrete 6 values race Discrete 5 values sex Discrete 2 values capital_gain Continue - capital_invest Discrete 2 values class Discrete 2 values Spv learning Meta-spv learning Spv learning assessment		education	Discrete	16 values		
marital_status Discrete 7 values occupation Discrete 14 values relationship Discrete 6 values race Discrete 5 values sex Discrete 2 values capital_loss Continue - capital_loss Continue - capital_loss Continue - native_country Discrete 2 values class Discrete 2 values capital_invest Discrete 2 values class Discrete 2 values spy learning Meta-spy learning Spy learning </td <td></td> <td>education_num</td> <td>Continue</td> <td>-</td> <td></td> <td></td>		education_num	Continue	-		
occupation Discrete 14 values relationship Discrete 6 values race Discrete 5 values sex Discrete 2 values capital_gain Continue - capital_loss Continue - capital_loss Continue - capital_invest Discrete 2 values hours_per_week Continue - native_country Discrete 2 values class Discrete 2 values Data visualization Descriptive stats Instance selection Feature selection Regression Factorial analysis Clustering Spv learning Meta-spv learning Spv learning assessment Association Image: Cross Tabulation Image: Correctization Image: Cross Tabulation Image: Correctization		marital_status	Discrete	7 values		
relationship Discrete 6 values race Discrete 5 values sex Discrete 2 values capital_gain Continue - capital_loss Continue - capital_invest Discrete 2 values capital_invest Discrete 2 values native_country Discrete 2 values loss Discrete 2 values capital_invest Discrete 2 values native_country Discrete 2 values loss Discrete 2 values class Components<		occupation	Discrete	14 values		
race Discrete 5 values sex Discrete 2 values capital_gain Continue - capital_loss Continue - capital_loss Continue - capital_loss Continue - capital_invest Discrete 2 values hours_per_week Continue - native_country Discrete 2 values class Discrete 2 values components Clustering Associatio		relationship	Discrete	6 values		1 2 .
sex Discrete 2 values capital_gain Continue - capital_loss Continue - capital_invest Discrete 2 values hours_per_week Continue - native_country Discrete 2 values class Instance selection Feature construction Spv learning Meta-spv le		race	Discrete	5 values		
capital_gain Continue - capital_loss Continue - capital_loss Continue - capital_invest Discrete 2 values hours_per_week Continue - native_country Discrete 41 values class Discrete 2 values Components - - Data visualization Descriptive stats Instance selection Feature construction Feature selection Regression Factorial analysis Clustering Spv learning Meta-spv learning Spv learning assessment Association		sex	Discrete	2 values		
capital_loss Continue - capital_invest Discrete 2 values hours_per_week Continue - native_country Discrete 41 values class Discrete 2 values Components - - Data visualization Descriptive stats Instance selection Feature construction Feature selection Regression Factorial analysis Clustering Spv learning Meta-spv learning Spv learning assessment Association		capital_gain	Continue	-		
capital_invest Discrete 2 values hours_per_week Continue - native_country Discrete 41 values class Discrete 2 values Components Components Data visualization Descriptive stats Instance selection Feature selection Regression Factorial analysis Spv learning Meta-spv learning Spv learning assessment Association Cross Tabulation Coup caracterization		capital_loss	Continue	-		
hours_per_week Continue native_country Discrete native_country Discrete class Discrete Data visualization Descriptive stats Instance selection Feature construction Feature selection Regression Spv learning Meta-spv learning Spv learning Group caracterization Cross Tabulation Group caracterization		capital_invest	Discrete	2 values	<u>← </u>	
native_country Discrete 41 values class Discrete 2 values Data visualization Descriptive stats Instance selection Feature selection Regression Factorial analysis Clustering Spv learning Meta-spv learning Spv learning assessment Association		hours_per_week	Continue	-	•	
class Discrete 2 values Components Instance selection Feature construction Feature selection Regression Factorial analysis Clustering Spv learning Meta-spv learning Spv learning assessment Association Cross Tabulation Image: Component selection Image: Component selection Image: Component selection		native_country	Discrete	41 values		
Components Data visualization Descriptive stats Instance selection Feature construction Feature selection Regression Factorial analysis Clustering Spv learning Meta-spv learning Spv learning assessment Association Cross Tabulation Group caracterization Image: Linear correlation Image: More Univariate cont state		class	Discrete	2 values		
Components Data visualization Descriptive stats Instance selection Feature construction Feature selection Regression Factorial analysis Clustering Spv learning Meta-spv learning Spv learning assessment Association Cross Tabulation Image: Cross Tabulation Image: Cross Tabulation Image: Cross Tabulation						<u> </u>
Data visualization Descriptive stats Instance selection Feature construction Feature selection Regression Factorial analysis Clustering Spv learning Meta-spv learning Spv learning assessment Association Image: Cross Tabulation I		Compone	nts			
Feature selection Regression Factorial analysis Clustering Spv learning Meta-spv learning Spv learning assessment Association Image: Cross Tabulation Image: Group caracterization Image: Cross Tabulation Image: Cross Tabulation	Data visualization Descriptive stats	Instan	ce selecti	ion Fe	ature construction	
Spv learning Meta-spv learning Spv learning assessment Association Image: Cross Tabulation Image: Cross Tabulation Image: Cross Tabulation Image: Cross Tabulation	Feature selection Regression	Facto	orial analys	is	Clustering	
Cross Tabulation III Group caracterization Linear correlation	Spv learning Meta-spv learning	g Spv learr	ning assess	ment	Association	
	Cross Tabulation II Group caracte	erization 🥻	Linear c	orrelation	🕂 More Univariate	cont stat
	< III					>

In the DEFINE STATUS component, add this new attribute among INPUT and run the diagram.

We note that CAPITAL rather indicates a behavior with respect to the saving: the people who have a higher income invest their money (12.9% in the whole dataset, 31% in this group).

TANAGRA 1.1.2 - [Group caracterizatio	n 1]											
File Diagram Component Window Help										- 8 >		
Default title		class=le:	55	~			class=mor	e		^		
🖃 🧰 Dataset (adult.xls)					37155	Examples				11687		
🖻 🎦 Define status 1	c		Test value	Group	Overral	Att - Des	c	Test value	Group	Overral		
	us attribu	tes				Continuo	us attributes	-				
			1.4	190039.57	189664.13	educatio	n_num	73.5	11.60	10.08		
	055		-32.6	54.15	87.50	age		50.9	44.28	38.64		
	ain		-49.3	147.01	1079.07	hours_pe	ar_week	50.3	45.45	40.42		
	r_week		-50.3	38.84	40.42	capital_g	ain 🦷	49.3	4042.24	1079.07		
	0		-50.9	36.87	38.64	capital_k	DSS	32.6	193.53	87.50		
	n_num		-73.5	9.60	10.08	fnlwgt		-1.4	188470.57	189664.13		
	attribute	5				Discrete	attributes					
	tatus=Ne	ver-married	70.5	41.40%	33.00%	marital_s	status=Married-civ-spouse	98.5	85.43%	45.82%		
	nvest=no		67.1	92.78%	87.07%	relations	hip=Husband	89.2	75.69%	40.37%		
	hip Own	*# #	10.0	20.40%	15.52	capital_i	nvest=yes	67.1	31,10%	12.93%		
•	ale		47.4	38.82%	33.15%	sex=Male		47.4	84.86%	66.85%		
	hip=Not-ir	i-family	42.1	30.43%	25.76%	occupati	on=Exec-managerial	46.6	24.88%	12.46%		
	on=Other	service	34.3	12.70%	10.08%	educatio	n=Bachelors	39.9	28.35%	16.43%		
	hip=Unma	ried	31.7	12.96%	10.49%	educatio	n=Masters	38.5	12,48%	5.44% ∨		
			Cor	nponents								
Data visualization Descriptive st	tats	Instance selec	tion	Feat	ure constru	uction	Feature selection		Regre	ssion		
Factorial analysis Clustering		Spv learnin	g	AA.	eta-spv lear	ning	Spv learning assessment		Association			
<	101	115			19 8 -					>		
1 Million Contraction Contraction												