# Subject

Implementing RANDOM FOREST with TANAGRA.

RANDOM FOREST is a combination of an ensemble method (BAGGING) and a particular decision tree algorithm ("Random Tree" into TANAGRA).

### Dataset

We use the HEART dataset from UCI Repository (http://www.ics.uci.edu/~mlearn/MLRepository.html).

We aim to predict a heart disease from various descriptors such as the age of the patient, etc. We have already used this dataset in other tutorials (http://eric.univlyon2.fr/~ricco/tanagra/fichiers/dr\_comparer\_spv\_learning.pdf).

## **RANDOM FOREST**

#### Create a diagram and load the dataset

We open (FILE/OPEN) the file DR\_HEART.BDM. BDM is a binary format of TANAGRA.

TANAGRA 1.4.9					
File Diagram Window Help					
Data minir	ng diagram (empty)				
	lanagra			2 🛛	
	Regarder dans : 🛅 H	leart	🖌 🕑 🕲 🔽	· .	
	Mes documents récents	heart.bdm			
	<b>Direau</b>				
	Mes documents				
	Poste de travail				
	Nom o	du fichier: dr. heart.bdm		Ouvrir	
Data visualization	Equaio réconsul Fichia	re de ture : Disau data mining	diagram (* hdm)	Annular ruction	
Feature selection	ravoiis ieseau richie	Bihary data mining		ng	
Spv learning	Meta-spv learning	Spv learning assessment	Scoring	Association	
Correlation scatterplot	Export dataset	🔣 Scatterplot	🔛 View dataset	🦾 View multiple sca	otterplot
					.::

#### Define the task

In the next step, we use the DEFINE STATUS component in order to choose the target attribute (COEUR) and the input attributes (all the others).



#### C4.5 decision tree algorithm

We use C4.5 and cross-validation in order to evaluate the accuracy of a standard (individual) decision tree algorithm. The error rate is 29.7%

TANAGRA 1.4.9 - [Cross-validation 1]											
Tile Diagram Componen	t Window Help									- 8 ×	
🗅 🛩 🖬   👪											
	Default title									<b>^</b>	
🖃 🏢 Dataset (heart.txt)			Results								
😑 🚺 Define status 1			CV error rat	e							
📄 🕨 Supervised	Learning 1 (C4.5)		Range								
Cross-va	lidation 1		MIN 0.27	33							
			Trial Fer ca	te							
			1 0.31	11							
			2 0.27	04							
			3 0.27	78							
			4 0.29	26							
			5 0.33	33							
			Overall cros	s-valid	ation error ra	te					
			Error rate				0.2970				
			Values prediction			Confusio	Confusion matrix				
			Value	Recall	1-Precision	_	presence	absence	Sum		
			presence	0.6000	0.3226	presence	000	220	600		
			absence	0.7587	0.2788	absence	181	569	750		
						Sum	501	/04	1350	~	
Components											
Data visualization	Statistics	Nonparametric statistics		cs	Instance selection		Feature construction		ion		
Feature selection	Regression	Factorial analysis			PLS		Clustering				
Spv learning	Meta-spv learning	Spv learning assessment			Scoring		Association				
P Dias-variance decompos	? Cross-validation			P Test Train-test							
<								>			

#### **RANDOM FOREST**

We want to implement the Random Forest algorithm now. There are two steps in order to insert the Random Forest method in the diagram: (1) Insert the BAGGING ensemble method (META-SPV LEARNING tab)

💯 TANAGRA 1.4.9 - [Cro	oss-validation 1]					
🕎 File Diagram Componen	t Window Help					_ @ ×
D 📽 🖪  🎎						
Default title						~
🖃 🏢 Dataset (heart.txt)	🖃 🥅 Dataset (heart.txt)				Results	
🚊 🚰 Define status 1			C¥ eri	or rate		
🛓 🕟 Supervised	Learning 1 (C4.5)		R	ange		
Cross-va	Cross-validation 1			0.2704		=
🕑 🕑 Bagging 1	Bagging 1			0.3333		-
1			Trial	Err rate		
			1	0.3111		
				0.2704		
				0.2778		
			4	0.2926		
			<			>
		Co	mponent	s		
Data visualization	Data visualization Statistics Nonpara			atistics	Instance selection	
Feature construction	Feature selection	Regression		ı	Factorial analysis	
PLS	Clustering	Spv learning		g	Meta-spv learning	
Sov learning assessment	Scoring	Association				
🕨 Aggregating 🚺	• Supervised Learning					
Arcing [Arc-x4]						
Bagging						
Bagging : simple example	les resampling and simple classifie	r voting.				

(2) Embed in this component the RANDOM TREE algorithm (SPV LEARNING tab).

It is obvious that using this learning algorithm without an aggregating framework often gives poor performances.

💯 TANAGRA 1.4.9 - [Cro	ss-validation 1]						
👚 File Diagram Component	t Window Help					-	Ξ×
🗅 📽 🔚 👫							
	Default title						~
🖃 🎟 Dataset (beart.txt)					Results		
Define status 1			C∀er	ror rate			
Supervised	earning 1 (C4.5)		R	ange			
	lidation 1		MIN	0.2704			=
Bagging 1 (Rod Tree)			MAX	0.3333			-
Engling I (K	<b>A</b>		Trial	Err rate			
	<b>小</b>		4	0.0444			
			1	0.3111			
			2	0.2704			
			3	0.2778			
			4	0.2926			× ×
		Co	mponent	s			
Data visualization	Statistics	Nonpara	imetric s	tatistics	Instance selection		
Feature construction	Feature selection	Regression			Factorial analysis		
PLS	Clustering	Spv learning			Meta-spv learning		
Spv learning assessment	Scoring	A	ssociatio	n			
Hanary logistic regression							
A.S Hard C4.5			1 1+E	- ¤Log-Regʻ	TRIRLS 📩 Pi	rototype-NN	
🔀 C-PLS	🖧 ID3	🔁 Multilayer perceptron				adial basis function	
🥂 C-RT	📴 K-NN	1.2 Multinomat Logistic Regress on A Rnd Tree					
<							>

So, we evaluate the learning accuracy with a cross-validation component.



The reduction of the error rate is impressive, we obtain 22.9% now.