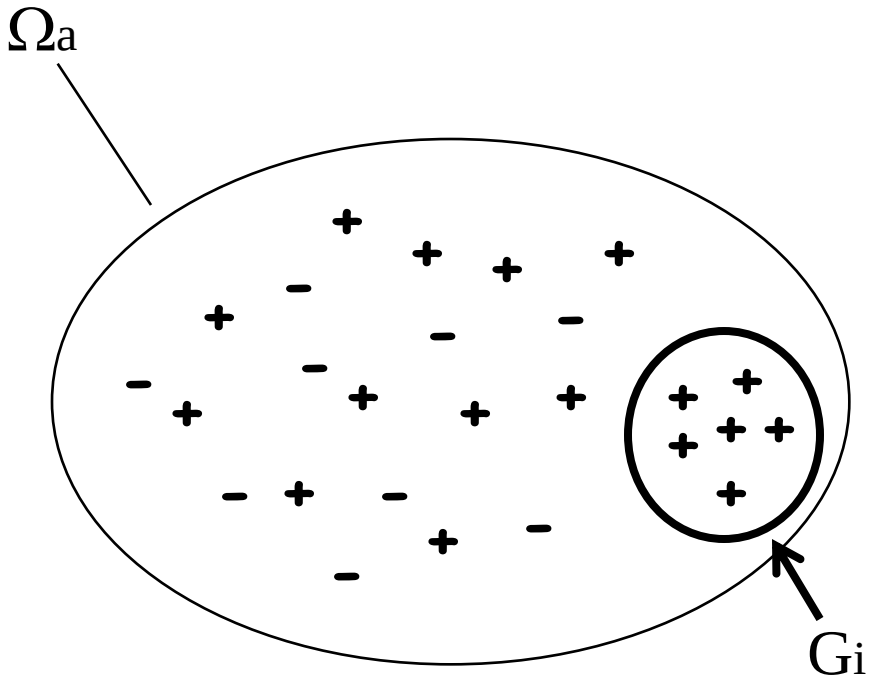
Arbres de Décision

Ricco RAKOTOMALALA
ricco.rakotomalala@univ-lyon2.fr

Construction d'un arbre de décision

Arbres de décision – Apprentissage par partitionnement

Objectif : on veut construire des sous-groupes les plus « homogènes » du point de vue de la variable à prédire



La variable qualitative Y prend ses valeurs dans {+,-}

Le sous-groupe Gi est complètement pur du point de vue de Y, il ne possède que des individus portant la valeur + de Y

$$si(\omega \in G_i)\ alors\ (Y = +)$$

L'idée est de trouver le plus rapidement Possible (avec le moins de variables) des groupes où P(Y=+) # 1

La description des sous-groupes repose sur :
- ☞ la fonction f et ses paramètres éventuels α
- ☞ les variables exogènes Xi

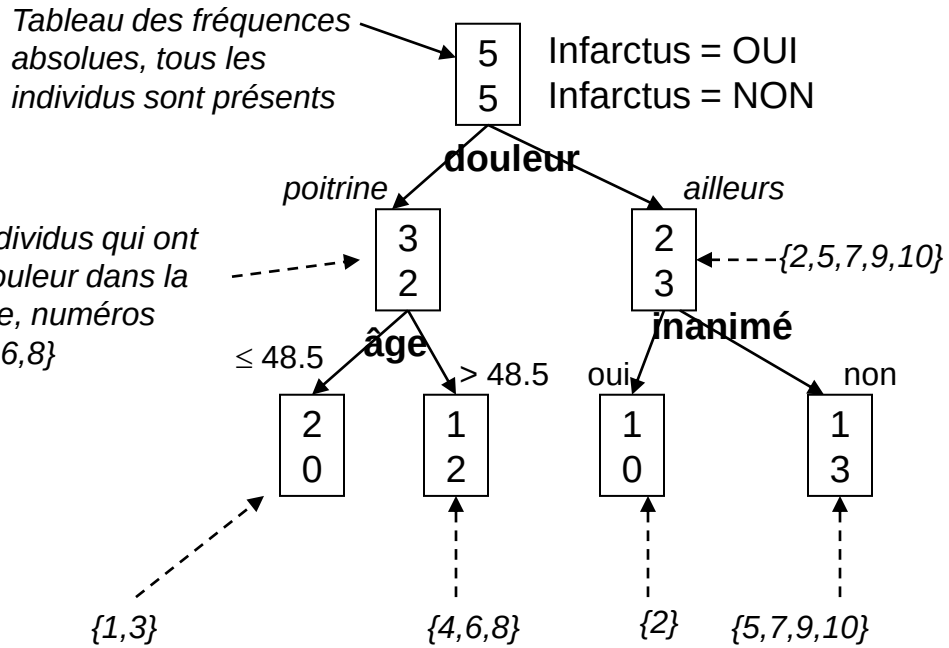
Arbres de décision – Un exemple

Numéro	Infarctus	Douleur	Age	Inanimé
1	oui	poitrine	45	oui
2	oui	ailleurs	25	oui
3	oui	poitrine	35	non
4	oui	poitrine	70	oui
5	oui	ailleurs	34	non
6	non	poitrine	60	non
7	non	ailleurs	67	non
8	non	poitrine	52	oui
9	non	ailleurs	58	non
10	non	ailleurs	34	non

Y X

A résoudre :

- choix de la variable de segmentation
- traitement des variables continues
- règle d'arrêt dans la construction
- décision sur une feuille



Premier sous-groupe, complètement homogène du point de vue de la variable à prédire : il est constitué exclusivement d 'individus qui ont un infarctus

Arbres de décision – Choix de la variable de segmentation

On choisit la variable X^* telle qu'elle est la plus liée (corrélée) avec Y

☞ on utilise la quantité du χ^2 calculée sur le tableau de contingence (croisement de Y avec Xi) pour quantifier cette liaison

	$x_{i,1}$	…	x_{i,L_i}
Y_1			
⋮		$n_{k,l} = card(\{\omega \in \Omega_a / Y(\omega) = Y_k \text{ et } X_i(\omega) = X_{i,l}\})$	
Y_K			

$$X^* = \arg \max_{i=1,\ldots,p} \chi^2_{Y,X_i}$$

Amélioration : la mesure du χ^2 augmente avec

☞ n, l 'effectif sur le nœud à segmenter

☞ le nombre de lignes

☞ le nombre de colonnes

Ces valeurs sont constantes dans les comparaisons deux à deux du χ^2

Les variables qui ont beaucoup de modalités (et ainsi induisent beaucoup de colonnes dans le tableau de contingence) sont avantagés

$$t^2_{Y,X_i} = \frac{\chi^2_{Y,X_i}}{n\sqrt{(K-1)(L_i-1)}}$$

(le t de Tschuprow varie entre 0 et 1)

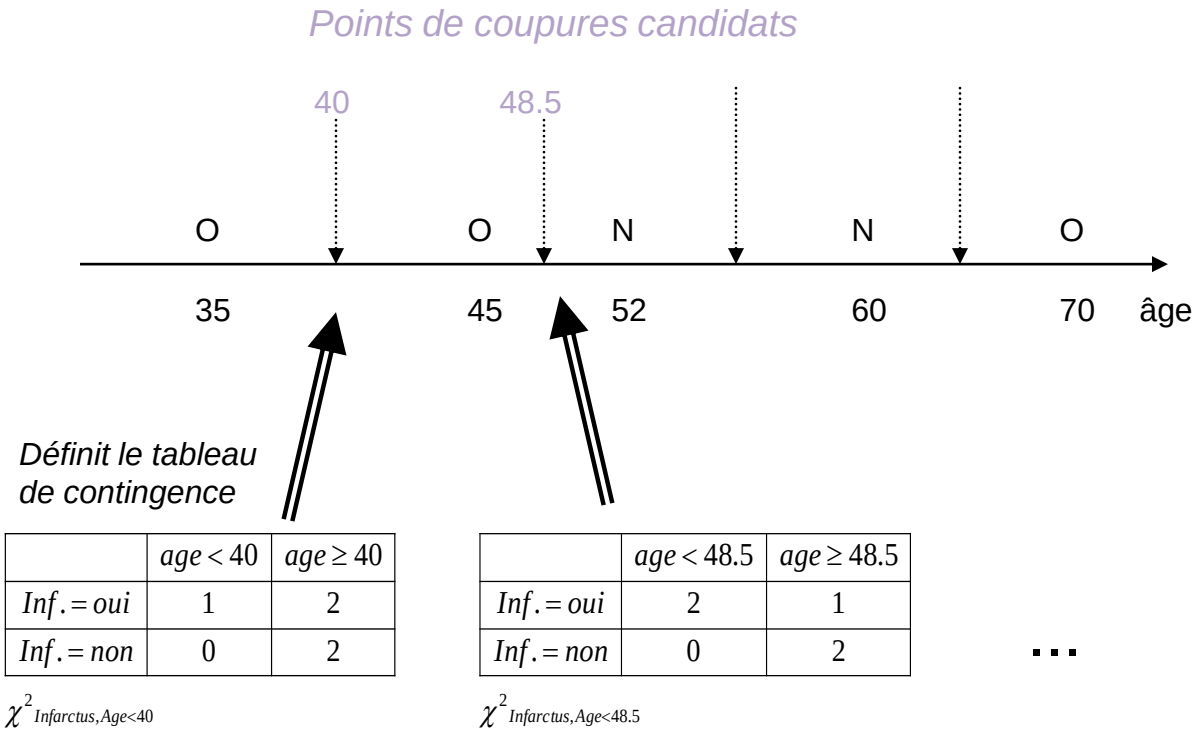
Arbres de décision – Traitement des variables continues

Comment est réalisé le choix du point de coupure
(ex: d 'où vient la valeur 48.5 de découpage de l 'âge dans l 'arbre exemple)

Point de coupure : borne de discrétisation

☞ il doit toujours être situé entre deux points consécutifs sur l 'axe de la variable quantitative

☞ il permet de définir un tableau de contingence



	$age < 40$	$age \geq 40$
$Inf. = oui$	1	2
$Inf. = non$	0	2

$\chi^2_{Infarctus, Age<40}$

	$age < 48.5$	$age \geq 48.5$
$Inf. = oui$	2	1
$Inf. = non$	0	2

$\chi^2_{Infarctus, Age<48.5}$

...

Arbres de décision – Règle d'arrêt

Quand décider qu'un sommet devient une feuille ?

Homogénéité des groupes : critère de précision (confiance)

Pureté d'un sommet

Seuil de spécialisation (ex. si une classe est représentée à 98% -> stop)

Effectif des groupes : critère de support

Taille minimale pour segmenter (ex. en dessous de 10 obs, on ne segmente plus)

Effectif d'admissibilité (ex. si un des sommets produit couvre moins de 2 obs. -> refus)

Test d'indépendance du CHI-2 : démarche statistique

$$\begin{cases} H_0 : Y \; et \; X^* \text{ indépendants} \\ H_1 : Y \text{ est lié avec } X^* \end{cases}$$

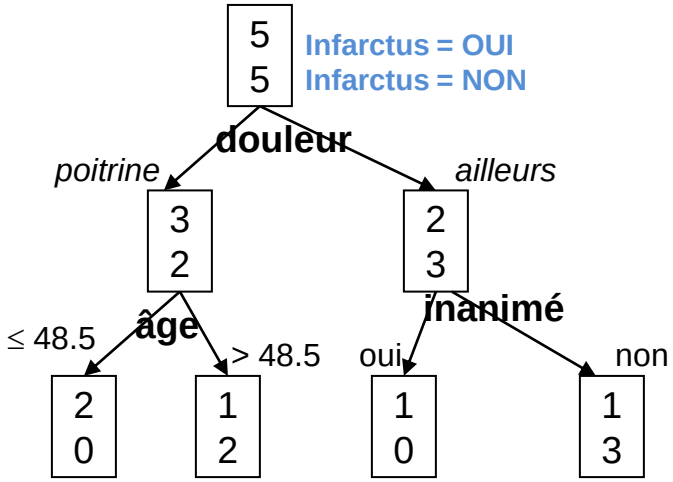
Comment fixer le risque du test ?

L'idée est surtout de contrôler la profondeur de l'arbre !

La taille de l'arbre influe fortement sur ses performances

Les arbres de décision dans la pratique
Post-traitement, analyse des résultats

Lecture des règles (1)



Chaque chemin partant de la racine vers les feuilles constitue une règle → 4 règles ici

A. Les règles sont mutuellement exclusives c.-à-d. un individu ne peut déclencher qu'une et une seule règle.

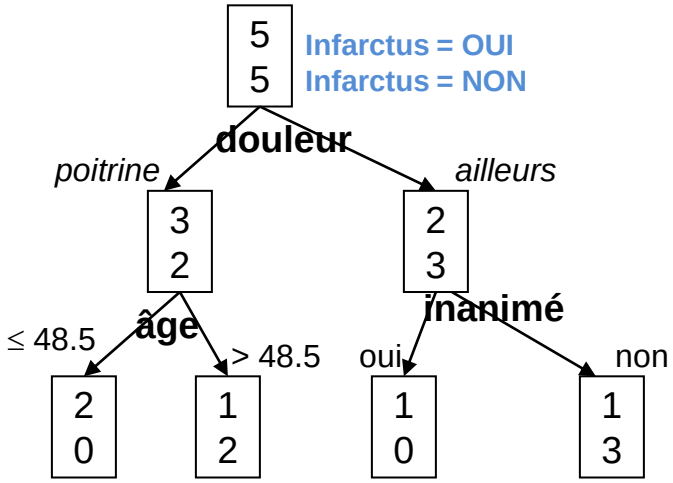
B. L'ensemble de règles couvre tout l'espace des valeurs possibles c.-à-d. tout individu à classer va déclencher une des règles.

1. **SI** douleur = poitrine **ET** âge ≤ 48.5 **ALORS** infarctus = oui
2. **SI** douleur = poitrine **ET** âge > 48.5 **ALORS** infarctus = non
3. **SI** douleur = ailleurs **ET** inanimé = oui **ALORS** infarctus = oui
4. **SI** douleur = ailleurs **ET** inanimé = non **ALORS** infarctus = non

⇨ **SI** prémisse **ALORS** conclusion

Conjonction de propositions de type « attribut .comparaison. valeur »

Lecture des règles (2)



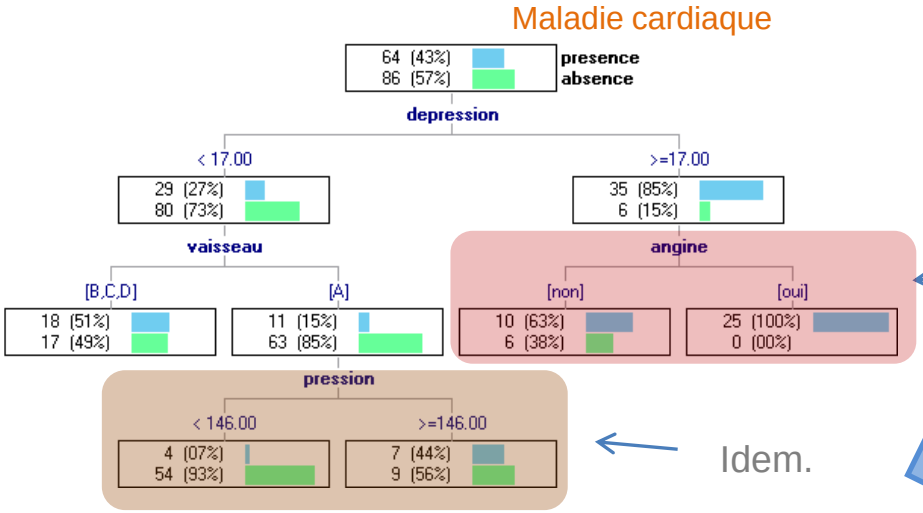
L'arbre peut se lire également comme une cascade de règles (propositions) imbriquées.

SINON signifie que « douleur ≠ poitrine » c.-à-d. « douleur = ailleurs » dans notre fichier de données.

SI poitrine = douleur
ALORS
SI âge ≤ 48.5
ALORS infarctus = oui
SINON infarctus = non
SINON
SI inanimé = oui
ALORS infarctus = oui
SINON infarctus = non

Mode de lecture intéressant si l'on souhaite traduire les règles en fonction SI(…) sous Excel.

Post-élagage manuel – Suppression des feuilles non pertinentes

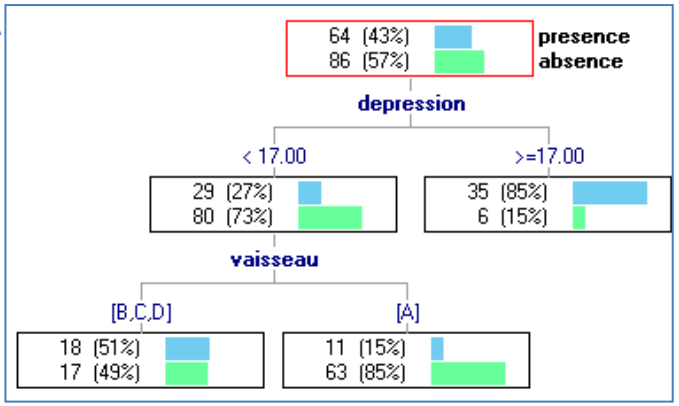


A priori, on aurait 5 feuilles = 5 règles.

Est-ce que ces 2 feuilles sont pertinentes ? Est-ce qu'il est vraiment nécessaire d'induire deux règles à partir de ces feuilles ?

Idem.

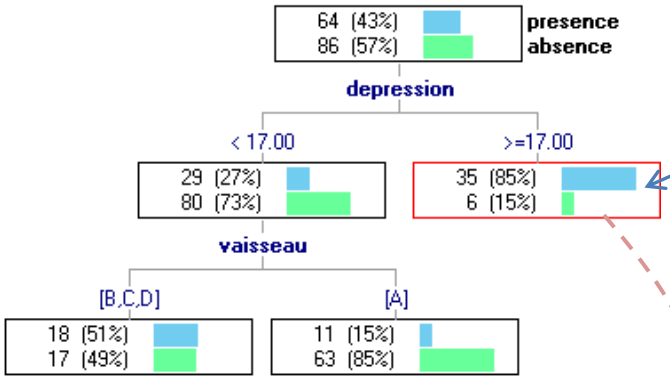
Post-élagage. Suppression des feuilles sœurs (issues du même père) portant des conclusions identiques. Suppression de proche en proche c.-à-d. on part des feuilles et on remonte jusqu'à la racine. On arrête lorsque la suppression n'est plus possible.



Arbre à 3 règles. Totalement équivalent du point de vue du classement.

Certaines méthodes de construction d'arbres intègrent une procédure automatisée de post-élagage (ex. C4.5, CART)

Expertise des règles – Rôle des autres variables quantitatives



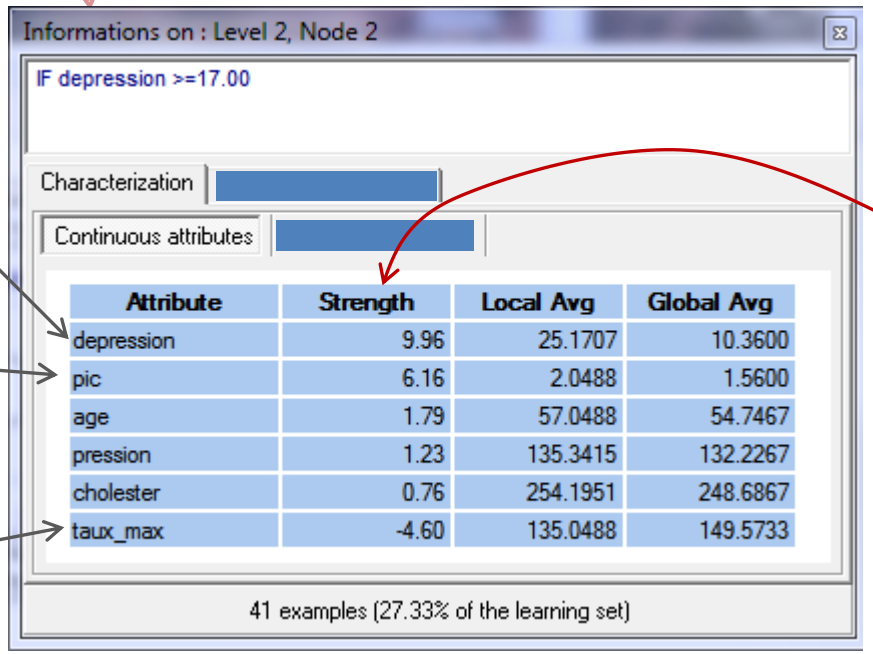
Une règle délimite une sous-population, décrite par la succession de propositions partant de la racine à la feuille.

La sous-population est caractérisée par les variables apparaissant sur le chemin. Est-ce que les autres variables (non-sélectionnées sur le chemin) nous permettent de mieux comprendre la formation du groupe ?

Ces personnes ont une « depression » plus élevée que les autres. Prévisible oui/non au vu de l'arbre ?

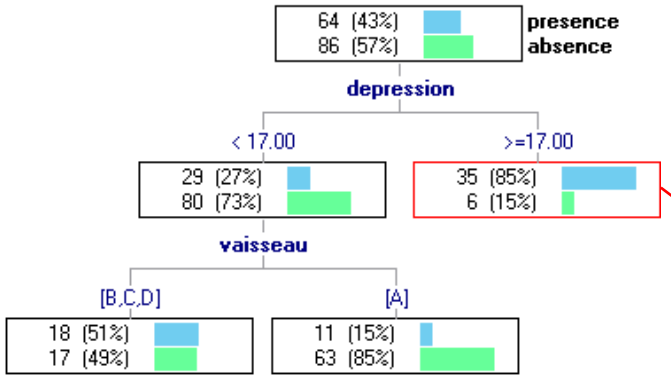
Ces personnes ont un « pic » plus élevé. On pouvait le deviner ?

Elles ont un « taux max » significativement plus bas. On pouvait le deviner ?



VALEUR TEST (≈ t de Student du test de comparaison de moyennes)

Expertise des règles – Rôle des autres variables qualitatives



170 individus en tout dans le fichier. 95 hommes et 55 femmes.

J-Measure. Pour distinguer les variables les plus caractérisantes.

Valeur-Test (≈ comparaison de proportions [80% vs. 63%]) pour distinguer les modalités les plus caractérisantes.

Informations on : Level 2, Node 2

IF depression >=17.00

Characterization | Continuous attributes

coeur (0.1547)

Values	Strength	Local Dist.	Global Dist.	Recall
presence	6.46	35 (85%)	64 (43%)	55%
absence	-6.46	6 (15%)	86 (57%)	7%

vaisseau (0.0673)

Values	Strength	Local Dist.	Global Dist.	Recall
B	-0.09	8 (20%)	30 (20%)	27%
A	-3.73	14 (34%)	88 (59%)	16%
C	3.82	13 (32%)	21 (14%)	62%
D	2.10	6 (15%)	11 (7%)	55%

angine (0.0506)

Values	Strength	Local Dist.	Global Dist.	Recall
non	-3.90	16 (39%)	96 (64%)	17%
oui	3.90	25 (61%)	54 (36%)	46%

sexe (0.0275)

Values	Strength	Local Dist.	Global Dist.	Recall
masculin	2.67	33 (80%)	95 (63%)	35%
feminin	-2.67	8 (20%)	55 (37%)	15%

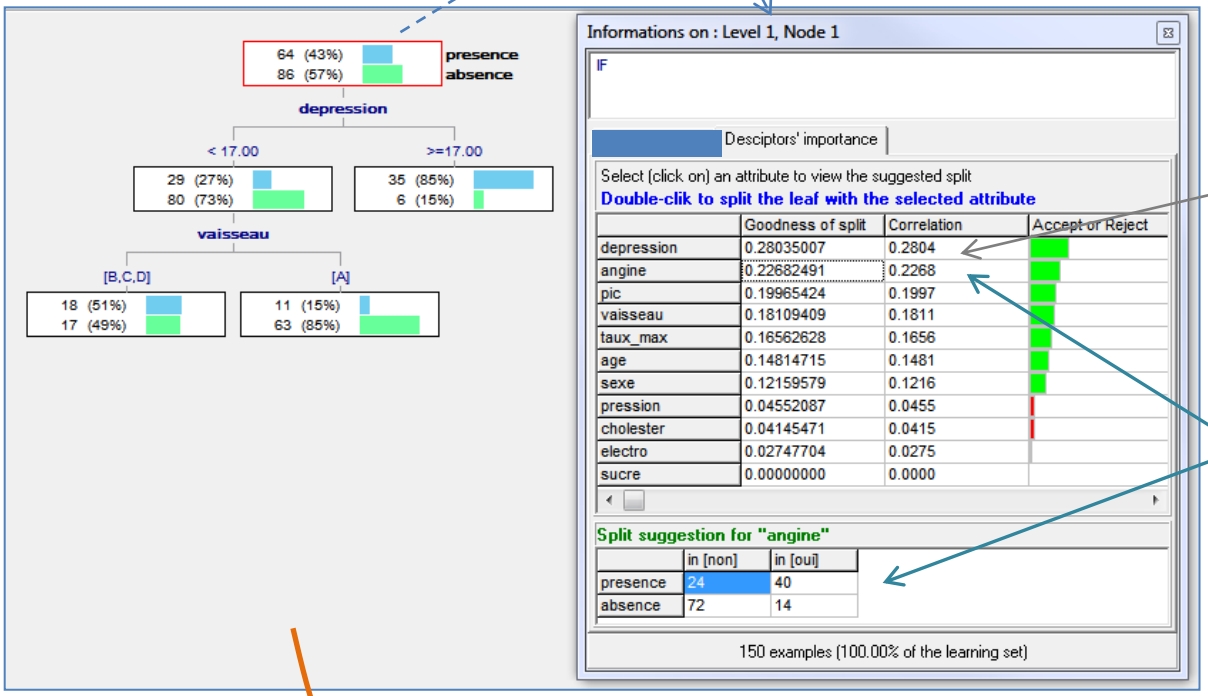
41 examples (27.33% of the learning set)

P(Homme / Sommet)= 80%

P(Sommet / Homme)= 35%

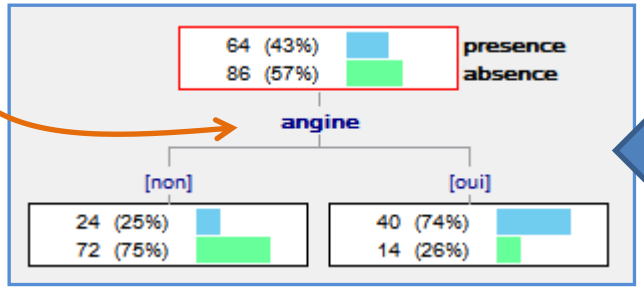
Sur le sommet (c.-à-d. individus avec « depression ≥ 17 »), il y 41 personnes, dont 33 hommes et 8 femmes.

Construction interactive des arbres



« depression » est la meilleure variable de segmentation avec un impact = 0.2804

« angine » est la 2ème meilleure variable avec impact = 0.2268. On obtiendrait le partitionnement ci-dessous.



On peut modifier interactivement l'arbre compte tenu de la qualité numérique (impact) et de la pertinence par rapport aux connaissances du domaine.

! On associe connaissances expertes et critères numériques pour construire un modèle efficace et interprétable.

! Très peu d'outils libres proposent la construction interactive des arbres.

Arbres de décision - Bilan

Arbres de décision
Une méthode très populaire

Kdnuggets

« Methods/algorithms used for data analysis/data mining in 2011 », Oct-Nov, 2011.

Algorithms for data analysis / data mining

Which methods/algorithms did you use for data analysis in 2011? [311 voters]

Decision Trees/Rules (186)	59.8 %
Regression (180)	57.9 %
Clustering (163)	52.4 %
Statistics (descriptive) (149)	47.9 %
Visualization (119)	38.3 %
Time series/Sequence analysis (92)	29.6 %
Support Vector (SVM) (89)	28.6 %
Association rules (89)	28.6 %
Ensemble methods (88)	28.3 %
Text Mining (86)	27.7 %
Neural Nets (84)	27.0 %
Boosting (73)	23.5 %
Bayesian (68)	21.9 %
Bagging (63)	20.3 %
Factor Analysis (58)	18.7 %
Anomaly/Deviation detection (51)	16.4 %
Social Network Analysis (44)	14.2 %
Survival Analysis (29)	9.32 %
Genetic algorithms (29)	9.32 %
Uplift modeling (15)	4.82 %

Site data mining et scoring

« Quelle méthode de data mining utilisez-vous le plus ? »

Résultats du sondage

"Quel logiciel de data mining utilisez-vous le plus ?"

Vous semblez avoir déjà voté !

Nombre de réponses "Analyse discriminante " : 188
Nombre de réponses "Arbres de décision " : 210
Nombre de réponses "Régression logistique " : 387
Nombre de réponses "Régression PLS " : 85
Nombre de réponses "Réseaux bayésiens " : 49
Nombre de réponses "Réseaux de neurones " : 105
Nombre de réponses "Support Vector Machines " : 51
Nombre de réponses "Autre " : 58

Nombre total de votes : 1133

Pourquoi cette popularité ? Avantages et inconvénients

Avantages :

• connaissances « intelligibles » -- validation d'expert (si arbre pas trop grand)
• traduction directe de l'arbre vers une base de règles
• sélection automatique des variables pertinentes

• non paramétrique
• traitement indifférencié selon le type des variables prédictives
• robuste face aux données aberrantes, solutions pour les données manquantes
• robuste face aux variables redondantes
• rapidité et capacité à traiter des très grandes bases

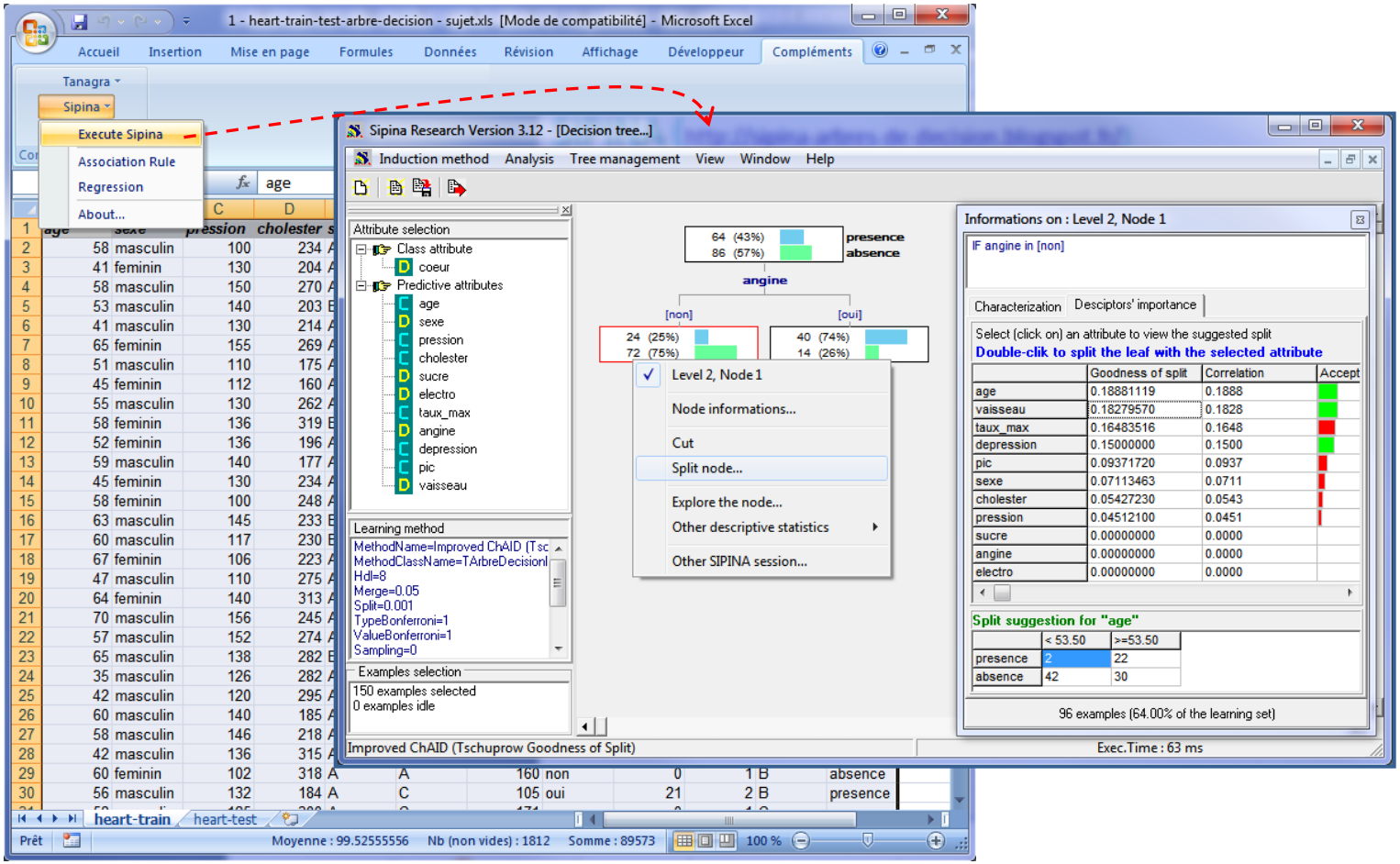
• enrichir l'interprétation des règles à l'aide des variables non sélectionnées
• possibilité pour le praticien d'intervenir dans la construction de l'arbre

Inconvénients :

• problème de stabilité sur les petites bases de données (feuilles à très petits effectifs)
• recherche « pas-à-pas » : difficulté à trouver certaines interactions (ex. xor)
• peu adapté au « scoring »
• performances moins bonnes en général par rapport aux autres méthodes (en réalité, performances fortement dépendantes de la taille de la base d'apprentissage)

Arbres de décision - Logiciels

SIPINA (http://sipina-arbres-de-decision.blogspot.fr/)



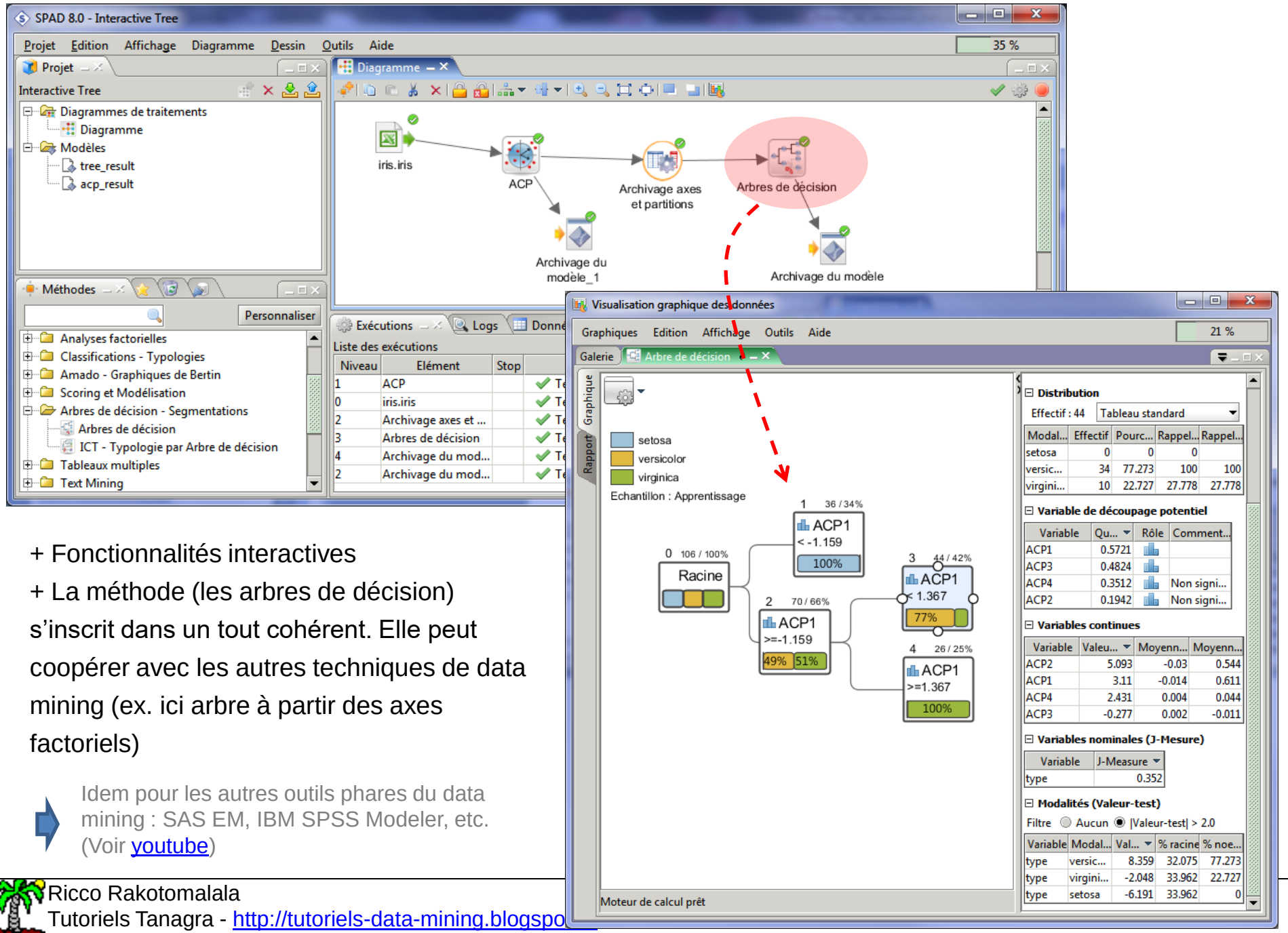
+ Association forte avec les tableurs (Excel, Libre / Open Office)

+ Le seul outil gratuit à proposer des fonctionnalités utilisatrices interactives de qualité

- Outil universitaire, peu d'outils pour l'exploitation des résultats et le reporting

- Stand alone, peu d'interaction avec les autres méthodes de data mining

SPAD (http://www.coheris.com/produits/analytics/logiciel-data-mining/)



+ Fonctionnalités interactives

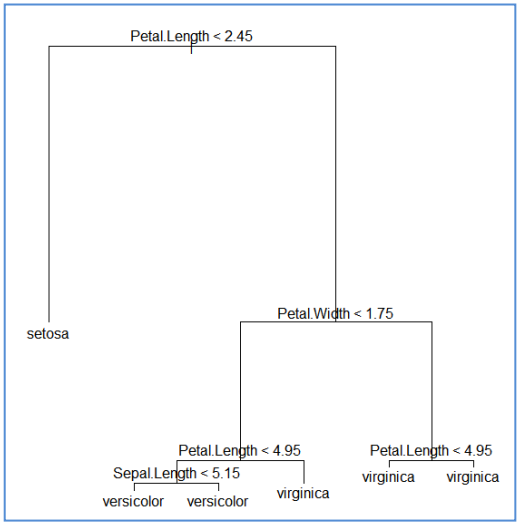
+ La méthode (les arbres de décision) s'inscrit dans un tout cohérent. Elle peut coopérer avec les autres techniques de data mining (ex. ici arbre à partir des axes factoriels)

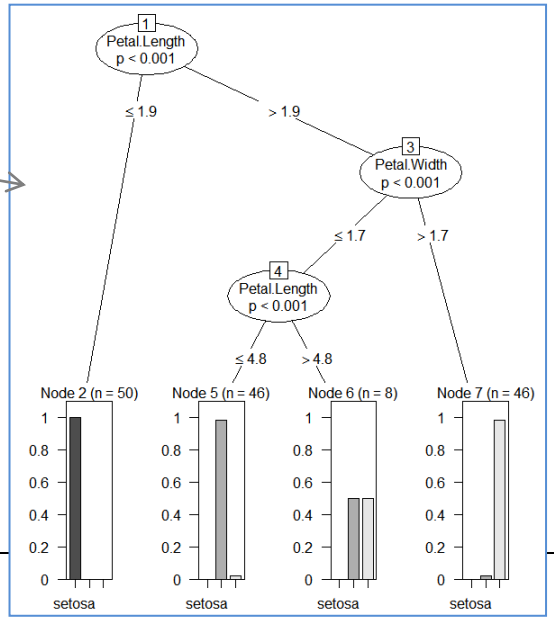
Idem pour les autres outils phares du data mining : SAS EM, IBM SPSS Modeler, etc. (Voir youtube)

R – Plusieurs packages spécialisés (tree, rpart, party, …)

```
> print(a1)
n= 150

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 150 100 setosa (0.33333333 0.33333333 0.33333333)
  2) Petal.Length< 2.45 50   0 setosa (1.00000000 0.00000000 0.00000000) *
  3) Petal.Length>=2.45 100  50 versicolor (0.00000000 0.50000000 0.50000000)
    6) Petal.Width< 1.75 54   5 versicolor (0.00000000 0.90740741 0.09259259) *
    7) Petal.Width>=1.75 46   1 virginica (0.00000000 0.02173913 0.97826087) *
```

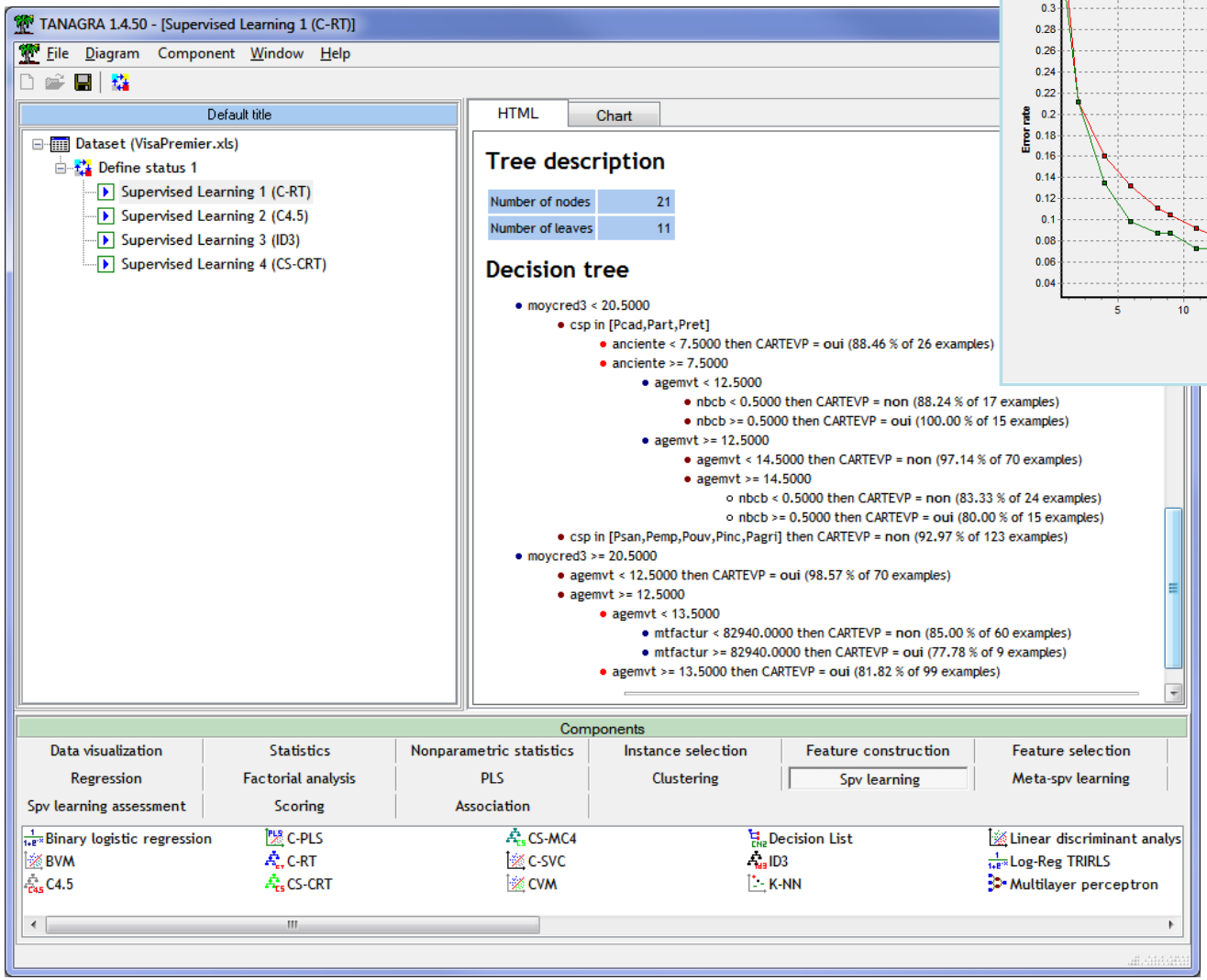
```
data(iris)

library(rpart)
a1 <- rpart(Species ~ ., data = iris)
print(a1)

library(tree)
a2 <- tree(Species ~ ., data = iris)
plot(a2)
text(a2)

library(party)
a3 <- ctree(Species ~ ., data = iris)
plot(a3)
```

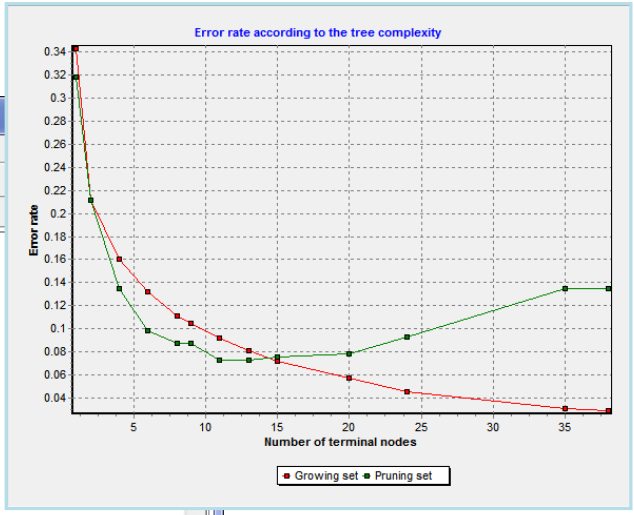




Aucune fonctionnalité interactive !

Tanagra – Plusieurs approches





Plusieurs approches, dont la méthode CART

Certaines peuvent prendre en compte les coûts de mauvais classement

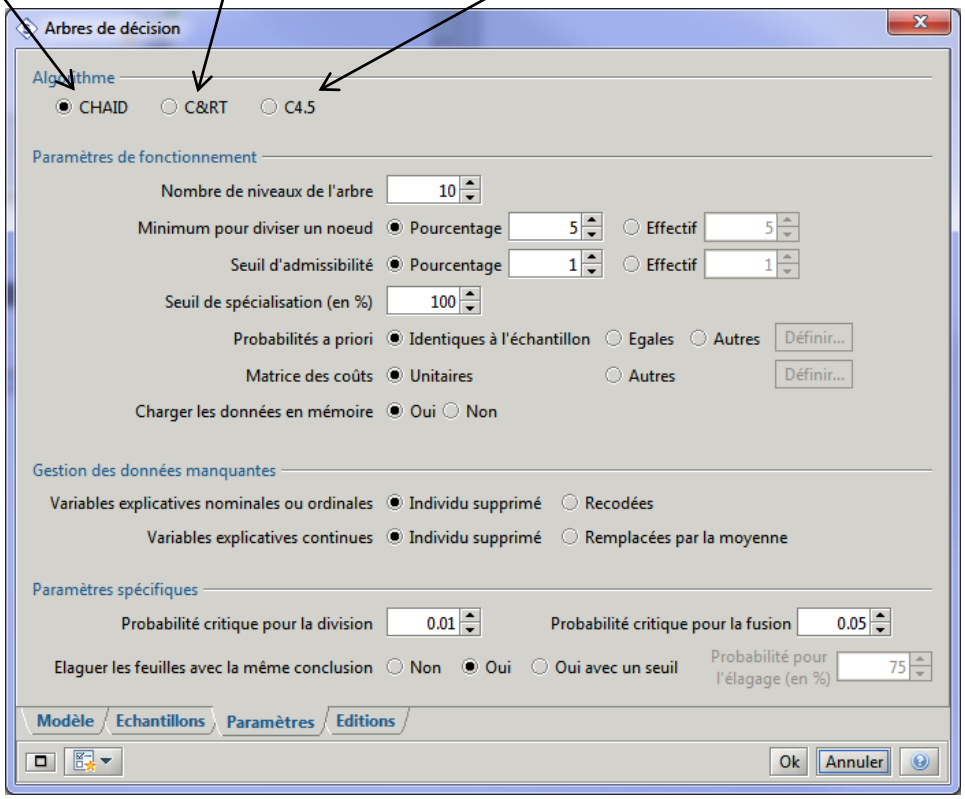
Pas de construction interactive des arbres

Méthodes phares

CART et apparentés (Breiman et al., 1984)

CHAID et apparentés (Kass, 1980)

C4.5 et apparentés (Quinlan, 1993)



Les méthodes se distinguent par : (1) les mesures d'évaluation des segmentations ; (2) les stratégies de regroupement durant la segmentation ; (3) la détermination de la taille « optimale » de l'arbre.

Bibliographie - Tutoriels

Tutoriels – http://tutoriels-data-mining.blogspot.fr/

« L'add-in Sipina pour Excel 2007 et 2010 », août 2014.

« Sipina add-on pour OOCalc », mars 2012.

« Apprentissage et test avec Sipina », mars 2008.

« Analyse interactive avec Sipina », mars 2008.

« Sipina – Traitement des très grands fichiers », octobre 2009.

« Le format PMML pour le déploiement de modèles », septembre 2010.

« Arbres de décision interactifs avec SPAD », janvier 2010.

« Nouveaux arbres interactifs dans SPAD 8 », août 2014.

« Introduction à R – Arbre de décision », mars 2012.

Etc., etc., et n'oublions pas les tutoriels sur youtube…

Bibliographie : arbres de décision

« Arbres de Décision », R. Rakotomalala, Revue MODULAD, 33:163-187, 2005

« Graphes d'induction », D. Zighed et R. Rakotomalala, Hermès, 2000.

« C4.5 – Programs for Machine Learning », Quinlan, 1993.

« Classification and Regression Tree », L. Breiman, J. Friedman, R. Olshen et C. Stone, 1984.

Autres références : http://sipina-arbres-de-decision.blogspot.fr/p/references.html