1. Supports

[REF 1] Rakotomalala R., « Arbre de décision - Introduction » ;
http://eric.univ-lyon2.fr/~ricco/cours/slides/Arbres_de_decision_Introduction.pdf

[REF 2] Rakotomalala R., « Arbres de décision », Revue Modulad, n°33, pages 163-187, 2005 ;
https://www.rocq.inria.fr/axis/modulad/archives/numero-33/tutorial-rakotomalala-33/rakotomalala-33-tutorial.pdf

[REF 3] Denis F., Gilleron R., « Apprentissage automatique : les arbres de décision », Chapitre 2 in *« Apprentissage à partir d'exemples – Notes de cours »* ;
http://www.grappa.univ-lille3.fr/polys/apprentissage/sortie004.html

2. Outils – Excel + Sipina

1. Chargez et installez le logiciel Sipina sur votre ordinateur

http://sipina-arbres-de-decision.blogspot.fr/

2. Intégrez Sipina dans Excel en tant que macro complémentaire

http://tutoriels-data-mining.blogspot.fr/2014/08/ladd-in-sipina-pour-excel-2007-et-2010.html

3. Données

Nous utilisons les données « **heart_train_test.xlsx** » en provenance du serveur UCI (Statlog [Heart] Data Set). Il s'agit de prédire l'occurrence d'une maladie cardiaque (CŒUR) chez les individus à partir de leurs caractéristiques (AGE … VAISSEAU).

a. Le classeur Excel dispose de 2 feuilles : « Apprentissage » contient l'échantillon d'apprentissage ; « Test » correspond à l'échantillon test.
b. On considère que « **Présence** » de CŒUR (présence de malade cardiaque) est la modalité positive. Ce commentaire est important lorsqu'il s'agira de calculer les ratios d'évaluation à partir de la matrice de confusion.

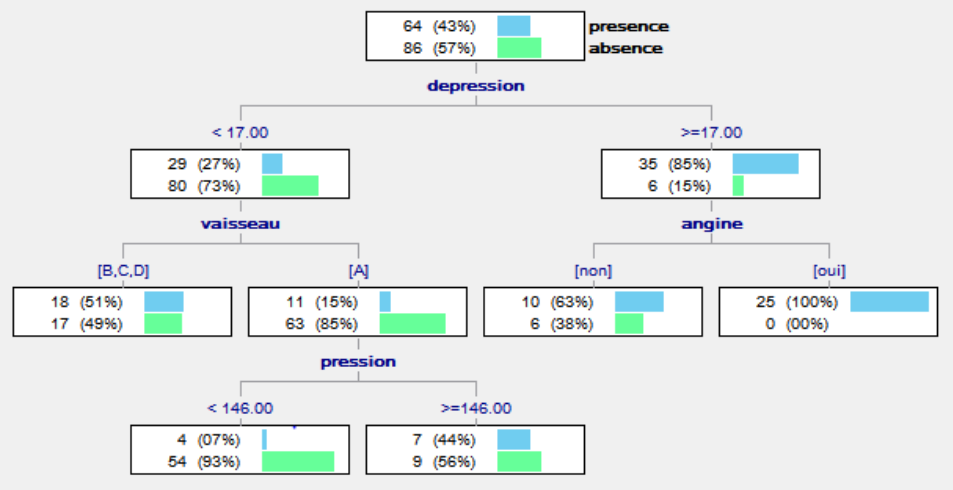
4. Modélisation, prédiction, évaluation

Appuyez-vous sur le tutoriel suivant pour la construction interactive des arbres sous Sipina : http://tutoriels-data-mining.blogspot.fr/2008/03/analyse-interactive-avec-sipina.html [TUTO].

A rendre : un fichier Excel avec des commentaires et le détail des calculs sur la feuille TEST du classeur.

1. Ouvrez le fichier « **heart_train_test.xlsx** » dans Excel. De combien d'observations disposons-nous dans l'échantillon d'apprentissage ? Dans l'échantillon test ?

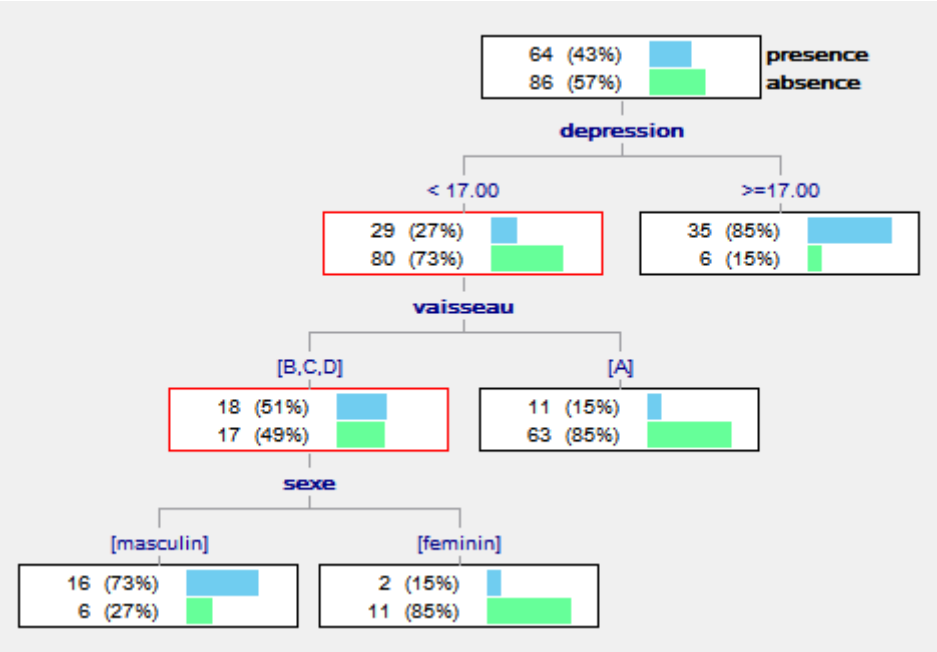
2. En suivant le tutoriel [TUTO, pages 1 à 6] : importez les **données d'apprentissage** dans Sipina, définissez les variables explicatives (**Attributes : AGE … VAISSEAU**) et cible (**Class = COEUR**) de l'étude, puis lancer la construction de l'arbre. Vous devriez obtenir l'arbre suivant :



3. Y a-t-il des feuilles que l'on pourrait élaguer dans cet arbre ? En partant des feuilles vers la racine, élaguer manuellement les branches qui n'apportent pas d'informations décisives dans la prédiction [cf. menu contextuel « CUT » : TUTO, page 12].
4. Lorsque vous êtes satisfaits de votre arbre, copiez-le (menu TREE MANAGEMENT / COPY) **dans la feuille test** du classeur Excel. Traduisez-le en règles prédictives [utilisez la fonction SI(… ; … ; …) d'Excel] pour définir la colonne Prédiction.1. Voici les premières lignes de la feuille de calcul que vous devriez obtenir :

age	sexe	pression	cholester	sucre	electro	taux_max	angine	depressio	pic	vaisseau	coeur	prediction.1
61	feminin	130	330	A	C	169	non	0	1	A	presence	absence
60	masculin	145	282	A	C	142	oui	28	2	C	presence	presence
66	masculin	160	228	A	C	138	non	23	1	A	absence	presence
61	feminin	145	307	A	C	146	oui	10	2	A	presence	absence

5. A partir des colonnes « Cœur » et « Prédiction.1 », construisez la matrice de confusion et calculez les différents indicateurs d'évaluation des classifieurs. Attention : « Présence » de la maladie correspond à la modalité positive de la variable cible.
6. Repartez de l'arbre précédent sous SIPINA. Essayez de le compléter en introduisant des segmentations supplémentaires de manière à obtenir des feuilles plus pures, avec des décisions plus tranchées [TUTO, page 12 à 15]. Lorsque vous pensez obtenir un arbre satisfaisant, copier l'arbre dans la feuille « TEST » dans Excel et construisez la colonne « **Prédiction.2** ». Votre arbre est-il plus performant que le précédent ? A titre d'exemple, voici une solution possible que j'ai pu développer sur ces données d'apprentissage, **il vous appartient d'explorer d'autres pistes**…



7. On vous demande d'utiliser en priorité les ANGINE, PIC, AGE et SEXE lors de la construction de l'arbre. Elaguez complètement votre arbre précédent (faites CUT à la racine). Puis construisez-le manuellement en privilégiant ces variables cette fois-ci. Lorsque vous pensez obtenir un arbre satisfaisant, portez le dans la feuille « HEART-TEST » en créant la colonne « **Prédiction.3** ». Qu'en est-il de cet arbre en termes de performances ?

5. Arbres sous R

Nous réalisons une analyse similaire sous R en utilisant la procédure **rpart()** du package éponyme.

Ce tutoriel [TUTO-R] devrait vous aider :

http://eric.univ-lyon2.fr/~ricco/cours/didacticiels/R/introduction_arbre_de_decision_avec_r.pdf

A rendre : un rapport .docx issu d'un projet R Markdown.

1. Importez la première feuille de « **heart_train_test.xlsx** » dans un premier data frame que vous nommerez DFApp.
2. Construire le modèle prédictif, « coeur » est la cible, les autres variables sont les explicatives. Utilisez la procédure rpart() du package RPART qui est installé par défaut (il n'est pas nécessaire de le charger sur le web) [TUTO-R, page 9].
3. Installez et chargez le package « rpart.plot ». Affichez l'arbre élaboré par rpart() avec la commande rpart.plot()
4. Importez la seconde feuille du classeur dans DFTest.
5. Réalisez la prédiction sur l'échantillon test (predict) [TUTO-R, page 10].
6. Pour construire la matrice de confusion, croisez les valeurs observées de la cible (coeur) avec celles prédites par le modèle (table) [TUTO-R, page 10].

7. Calculez alors les différents indicateurs de performance des modèles (taux d'erreur, sensibilité, précision) sachant que « cœur = présence » est la modalité positive.
8. Modifiez les paramètres de l'algorithme de manière à obtenir un arbre plus petit. L'arbre est-il meilleur au sens du taux d'erreur mesuré sur l'échantillon test ?
9. Essayez d'obtenir un arbre plus grand cette fois-ci. Les performances sont modifiées de quelle manière sur l'échantillon test ?