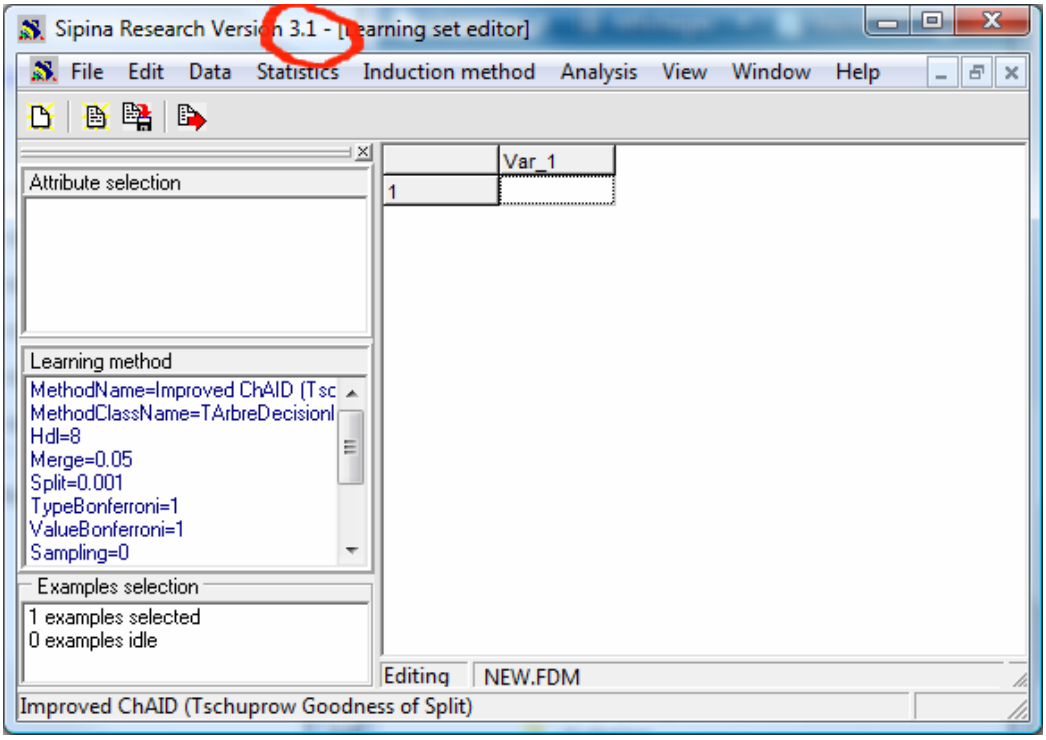
1 Introduction

Gestion des données manquantes dans SIPINA.

Attention, pour reproduire pleinement les opérations décrites dans ce didacticiel, assurez vous de disposer de la version 3.1 de la version recherche de Sipina[1]. Vous pouvez vérifier cela en observant le numéro de version dans la barre de titre du logiciel lorsqu'il est démarré.



L'appréhension des données manquantes est un problème difficile. La gestion informatique en elle-même ne pose pas de problème, il suffit de signaler la valeur manquante par un code spécifique. En revanche, son traitement avant ou durant l'analyse des données est très compliqué. Il faut prendre en considération deux aspects[2] :

- La nature de la valeur manquante. Est-elle complètement aléatoire c.-à-d. toutes les valeurs ont la même probabilité d'être manquante (ex. les personnes qui omettent d'indiquer leur revenu dans une fiche signalétique ne répondent pas à des caractéristiques particulières). Est-elle conditionnellement aléatoire c.-à-d. dans certaines conditions, l'occurrence d'une valeur manquante suit un processus aléatoire (par exemple, il n'y a pas de données manquantes, sauf parmi les cadres

[1] La page de téléchargement du logiciel est http://eric.univ-lyon2.fr/~ricco/sipina_download.html ; chargez et installez la SIPINA RESEARCH VERSION.

[2] Ce petit fascicule est très intéressant pour comprendre les tenants et aboutissants du problème des données manquantes : P.D. Allison, « Missing Data », in Quantitative Applications in the Social Sciences Series n°136, Sage University Paper, 2002.

où l'absence de réponse est distribuée aléatoirement). Est-elle non aléatoire (par exemple, le nombre de cases pour inscrire les chiffres est limité à 4, toutes les personnes qui ont un salaire supérieur à 9999 euros mensuel ne peuvent pas inscrire leur salaire[3]). On considère généralement que nous nous inscrivons dans la première situation pour pouvoir travailler, mais rien n'est moins sûr dans les études réelles.

- La technique statistique que nous mettons en œuvre par la suite. En effet, certaines méthodes de traitement des données manquantes sont plus ou moins adaptées selon les techniques statistiques que nous utilisons.

Prenons l'exemple de la suppression des lignes du tableau des données. Nous supprimons du fichier toutes les observations comportant au moins une valeur manquante (*listwise deletion* ou *casewise deletion* en anglais). L'approche paraît primaire, voire brutale. Nous pouvons réduire considérablement la taille du fichier ainsi. Pourtant, on montre qu'elle est plus robuste que les méthodes sophistiquées (maximum de vraisemblance, imputation multiple), en termes de biais et variance des estimations, lorsque la formation des valeurs manquantes s'écarte du processus complètement aléatoire et que nous implémentons une régression linéaire ou logistique (Allison, 2001 ; pages 84-85).

Prenons un autre exemple, nous utilisons les informations fournies par les autres descripteurs pour « deviner » les valeurs manquantes des variables (ex. un arbre de décision, une régression linéaire ; on parle d'imputation déterministe). Ce faisant, nous renforçons artificiellement le lien entre les variables. Toutes les méthodes statistiques qui s'appuient sur la matrice des corrélations sont faussées, les écarts-type des coefficients sont sous-estimées dans la régression (Allison, 2001 ; pages 11-12).

TANAGRA ayant une vocation essentiellement pédagogique, je ne voulais pas introduire des outils automatisés de gestion des données manquantes. Il ne me paraissait pas souhaitable que l'étudiant puisse cliquer sur un bouton et évacuer ce problème négligemment. Il doit préparer ses données en ayant pleinement conscience de ce qu'il fait avant de pouvoir lancer un traitement statistique dans de bonnes conditions.

Je n'avais pas ce type de scrupule du temps de SIPINA, qui a été pour moi un véritable laboratoire à idées. Plusieurs techniques ont été implémentées, je les redécouvre moi-même aujourd'hui. L'objectif de ce tutoriel est de montrer leur mise en œuvre et les conséquences des choix sur l'induction des arbres de décision avec la méthode C4.5 (Quinlan, 1993).

2 Données

Notre fichier provient du site de Gilles Hunault de l'Université d'Angers[4]. On veut prédire le ronflement chez des individus à partir de leurs caractéristiques (âge, poids, taille, etc.). Nous en avons extrait 30

[3] Il paraît qu'il y en a. Ils ne sont pas enseignants-chercheurs en tous cas.

[4] http://www.info.univ-angers.fr/~gh/Datasets/datasets.htm

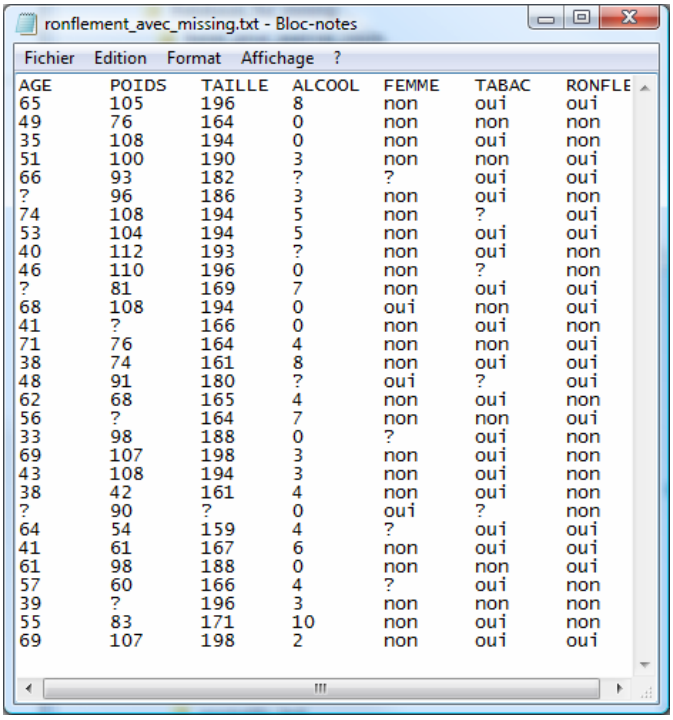
observations, puis nous avons supprimé quelques valeurs totalement au hasard.

Nous manipulons plusieurs fichiers dans ce didacticiel :

- RONFLEMENT_ALL.FDM est le fichier complet au format SIPINA, sans données manquantes. Nous nous en servirons pour élaborer l'arbre de décision de référence.
- RONFLEMENT_WITH_MISSING.FDM est le fichier avec données manquantes. C'est la traduction au format SIPINA du fichier texte ci-dessous. Nous l'utiliserons pour montrer les différentes stratégies de traitement des données manquantes.

L'ensemble des fichiers sont réunis dans une archive accessible en ligne[5].

Nous montrons ici le fichier au format texte, les valeurs manquantes sont symbolisées par le caractère « ? »



AGE	POIDS	TAILLE	ALCOOL	FEMME	TABAC	RONFLE
65	105	196	8	non	oui	oui
49	76	164	0	non	non	non
35	108	194	0	non	oui	non
51	100	190	3	non	non	oui
66	93	182	?	?	oui	oui
?	96	186	3	non	oui	non
74	108	194	5	non	?	oui
53	104	194	5	non	oui	oui
40	112	193	?	non	oui	non
46	110	196	0	non	?	non
?	81	169	7	non	oui	oui
68	108	194	0	oui	non	oui
41	?	166	0	non	oui	non
71	76	164	4	non	non	oui
38	74	161	8	non	oui	oui
48	91	180	?	oui	?	oui
62	68	165	4	non	oui	non
56	?	164	7	non	non	oui
33	98	188	0	?	oui	non
69	107	198	3	non	oui	non
43	108	194	3	non	oui	non
38	42	161	4	non	oui	non
?	90	?	0	oui	?	non
64	54	159	4	?	oui	oui
41	61	167	6	non	oui	oui
61	98	188	0	non	non	oui
57	60	166	4	?	oui	non
39	?	196	3	non	non	non
55	83	171	10	non	oui	non
69	107	198	2	non	oui	oui

Nous avons préparé les fichiers pour faciliter les manipulations dans ce didacticiel. Mais notons que SIPINA sait importer des fichiers de données au format texte avec séparateur tabulation comportant des données manquantes. Il suffit de les symboliser par le caractère « ? » ou de laisser l'emplacement vide.

De même nous pouvons utiliser la macro complémentaire pour envoyer les données d'Excel vers SIPINA[6]. Il faut simplement laisser la cellule vide lorsque l'on a une valeur manquante. Un classeur au format XLS accompagne les données pour que l'utilisateur puisse réaliser lui même les tests.

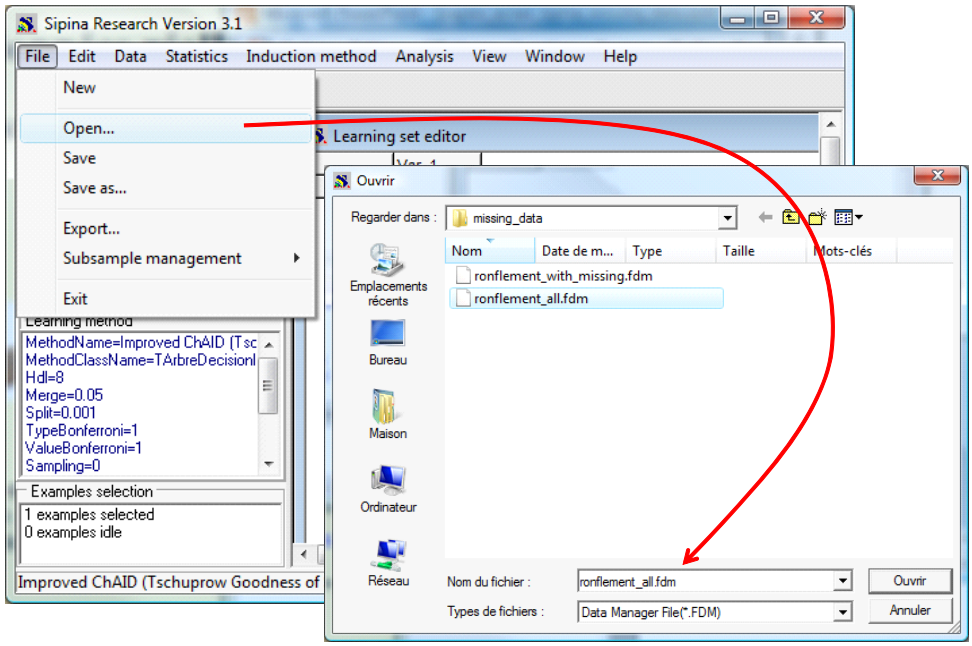
[5] http://eric.univ-lyon2.fr/~ricco/dataset/ronflement_missing_data.zip

[6] http://tutoriels-data-mining.blogspot.com/2008/03/connexion-excel-sipina.html

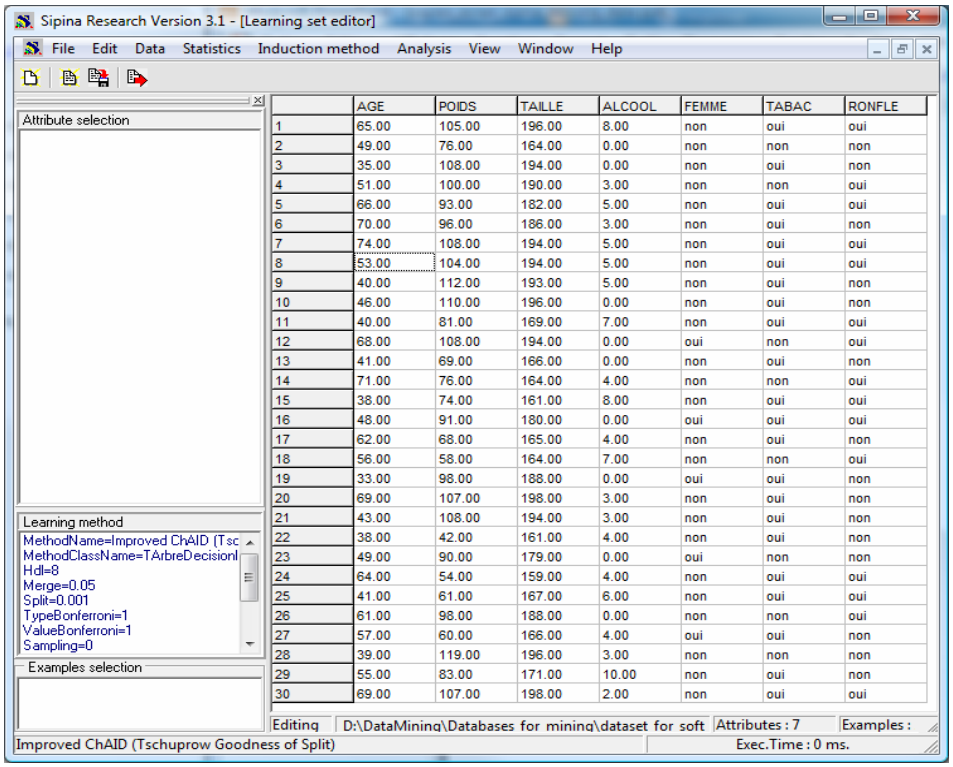
3 Traitement de la base complète

3.1 Chargement des données

Après avoir démarré SIPINA, nous chargeons le fichier complet RONFLEMENT_ALL.FDM en actionnant le menu FILE / OPEN.



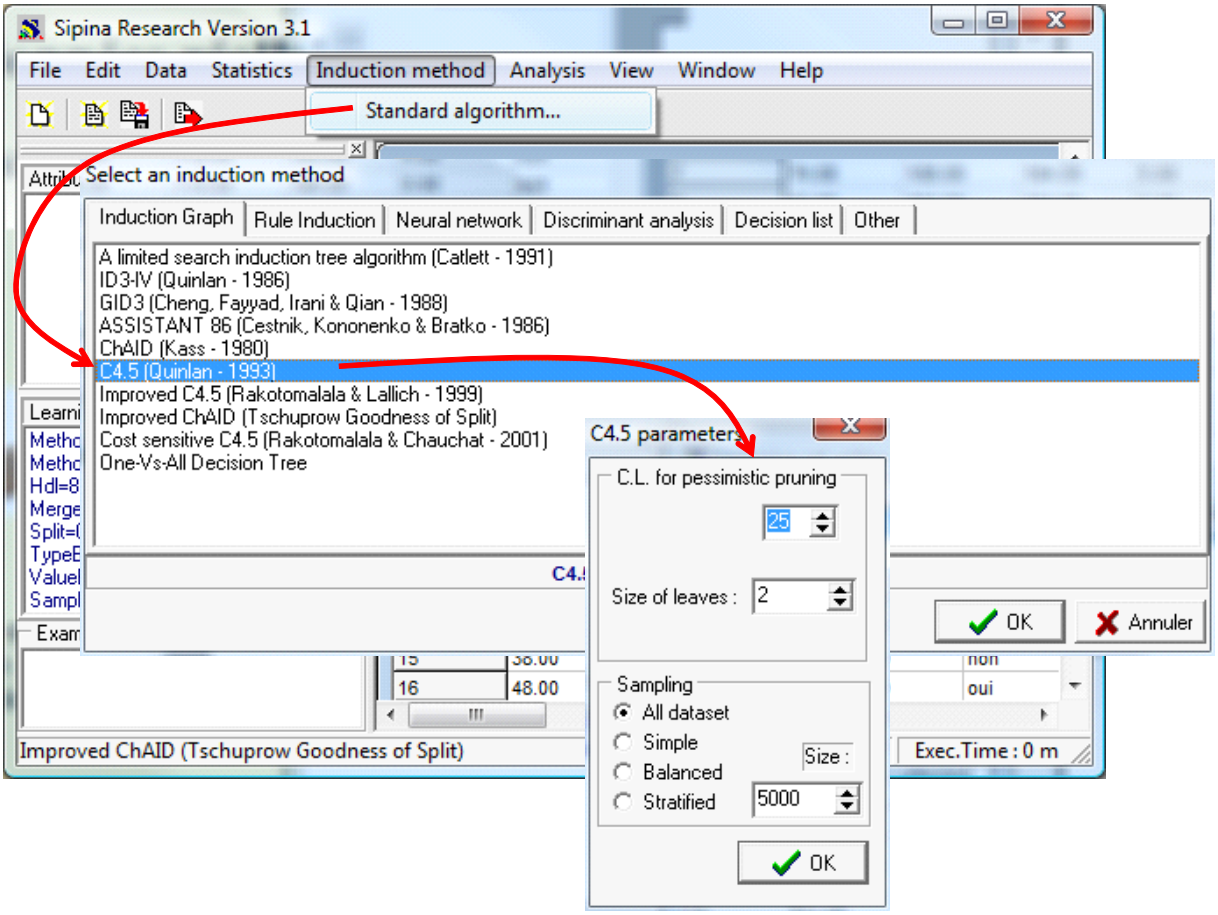
Les 30 observations sont affichées dans la grille de données.



	AGE	POIDS	TAILLE	ALCOOL	FEMME	TABAC	RONFLE
1	65.00	105.00	196.00	8.00	non	oui	oui
2	49.00	76.00	164.00	0.00	non	non	non
3	35.00	108.00	194.00	0.00	non	oui	non
4	51.00	100.00	190.00	3.00	non	non	oui
5	66.00	93.00	182.00	5.00	non	oui	oui
6	70.00	96.00	186.00	3.00	non	oui	non
7	74.00	108.00	194.00	5.00	non	oui	oui
8	53.00	104.00	194.00	5.00	non	oui	oui
9	40.00	112.00	193.00	5.00	non	oui	non
10	46.00	110.00	196.00	0.00	non	oui	non
11	40.00	81.00	169.00	7.00	non	oui	oui
12	68.00	108.00	194.00	0.00	oui	non	oui
13	41.00	69.00	166.00	0.00	non	oui	non
14	71.00	76.00	164.00	4.00	non	non	oui
15	38.00	74.00	161.00	8.00	non	oui	oui
16	48.00	91.00	180.00	0.00	oui	oui	oui
17	62.00	68.00	165.00	4.00	non	oui	non
18	56.00	58.00	164.00	7.00	non	non	oui
19	33.00	98.00	188.00	0.00	oui	oui	non
20	69.00	107.00	198.00	3.00	non	oui	non
21	43.00	108.00	194.00	3.00	non	oui	non
22	38.00	42.00	161.00	4.00	non	oui	non
23	49.00	90.00	179.00	0.00	oui	non	non
24	64.00	54.00	159.00	4.00	non	oui	oui
25	41.00	61.00	167.00	6.00	non	oui	oui
26	61.00	98.00	188.00	0.00	non	non	oui
27	57.00	60.00	166.00	4.00	oui	oui	non
28	39.00	119.00	196.00	3.00	non	non	non
29	55.00	83.00	171.00	10.00	non	oui	non
30	69.00	107.00	198.00	2.00	non	oui	oui

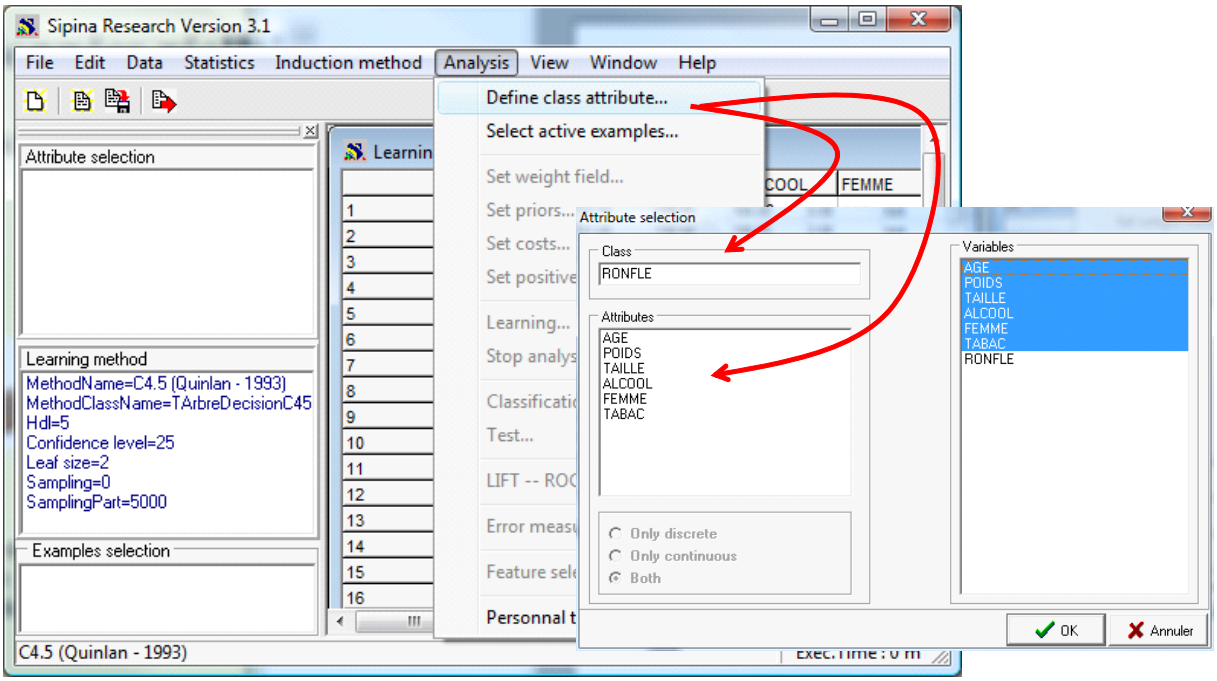
3.2 Choix de l'algorithme de traitement

L'étape suivante consiste à choisir l'algorithme d'apprentissage. Nous actionnons le menu INDUCTION METHOD / STANDARD ALGORITHM, nous choisissons la méthode C4.5 (Quinlan, 1993) et nous validons les paramètres par défaut. Nous noterons principalement que la méthode ne segmente pas un nœud si les feuilles subséquentes contiennent moins de 2 observations. Nous nous en souviendrons plus loin.



3.3 Définition du problème à traiter

Nous devons maintenant définir la variable à prédire et les variables explicatives. Nous actionnons le menu ANALYSIS / DEFINE CLASS ATTRIBUTES. Par glisser déposer, nous plaçons RONFLE en CLASS, les autres variables en ATTRIBUTES. Nous validons la sélection.



3.4 Arbre de décision

Il ne reste plus qu'à lancer les traitements. Nous actionnons le menu ANALYSIS / LEARNING. Nous obtenons l'arbre de décision.

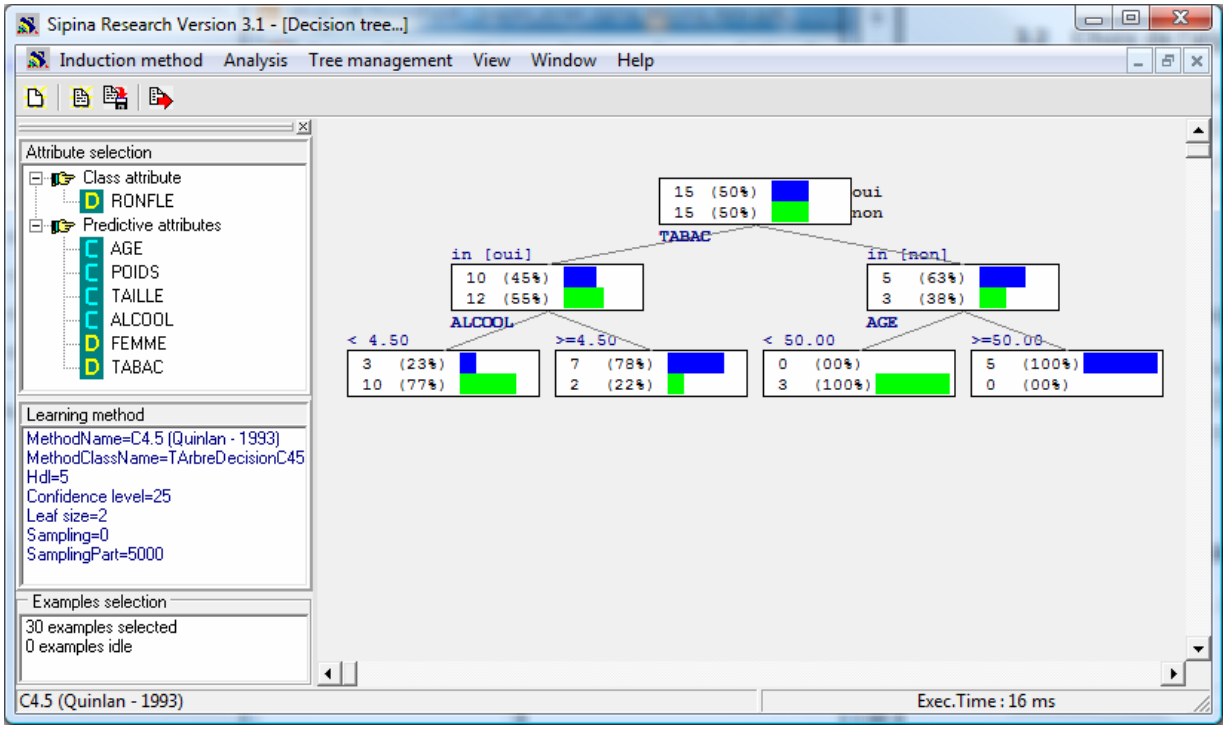


Figure 1 - Arbre sur le fichier sans valeurs manquantes

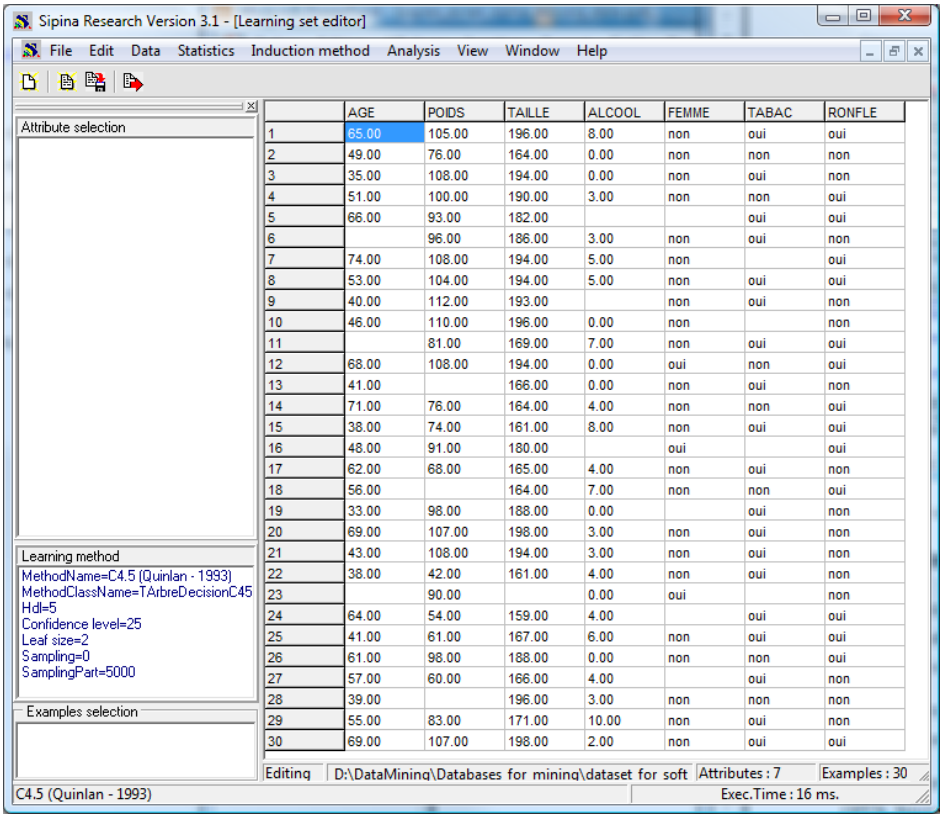
Nous avons un arbre à 3 niveaux. Les variables déterminantes sont le TABAC, l'ALCOOL et l'AGE.

4 Traitement des valeurs manquantes

L'idée maintenant est de travailler à partir du fichier comportant des valeurs manquantes. Nous souhaitons étudier dans quelle mesure les options proposées par SIPINA pour les traiter nous éloignent de l'arbre ci-dessus (Figure 1) lorsque nous lançons la méthode C4.5.

Nous devons stopper l'analyse courante en actionnant le menu ANALYSIS / STOP ANALYSIS. Puis vider la grille de données en cliquant sur le menu FILE / NEW.

Nous pouvons charger le fichier comportant les valeurs manquantes en cliquant sur le menu FILE / OPEN. Nous sélectionnons cette fois-ci le fichier RONFLEMENT_WITH_MISSING.FDM.



	AGE	POIDS	TAILLE	ALCOOL	FEMME	TABAC	RONFLE
1	65.00	105.00	196.00	8.00	non	oui	oui
2	49.00	76.00	164.00	0.00	non	non	non
3	35.00	108.00	194.00	0.00	non	oui	non
4	51.00	100.00	190.00	3.00	non	non	oui
5	66.00	93.00	182.00			oui	oui
6		96.00	186.00	3.00	non	oui	non
7	74.00	108.00	194.00	5.00	non		oui
8	53.00	104.00	194.00	5.00	non	oui	oui
9	40.00	112.00	193.00		non	oui	non
10	46.00	110.00	196.00	0.00	non		non
11		81.00	169.00	7.00	non	oui	oui
12	68.00	108.00	194.00	0.00	oui	non	oui
13	41.00		166.00	0.00	non	oui	non
14	71.00	76.00	164.00	4.00	non	non	oui
15	38.00	74.00	161.00	8.00	non	oui	oui
16	48.00	91.00	180.00		oui		oui
17	62.00	68.00	165.00	4.00	non	oui	non
18	56.00		164.00	7.00	non	non	oui
19	33.00	98.00	188.00	0.00		oui	non
20	69.00	107.00	198.00	3.00	non	oui	non
21	43.00	108.00	194.00	3.00	non	oui	non
22	38.00	42.00	161.00	4.00	non	oui	non
23		90.00		0.00	oui		non
24	64.00	54.00	159.00	4.00		oui	oui
25	41.00	61.00	167.00	6.00	non	oui	oui
26	61.00	98.00	188.00	0.00	non	non	oui
27	57.00	60.00	166.00	4.00		oui	non
28	39.00		196.00	3.00	non	non	non
29	55.00	83.00	171.00	10.00	non	oui	non
30	69.00	107.00	198.00	2.00	non	oui	oui

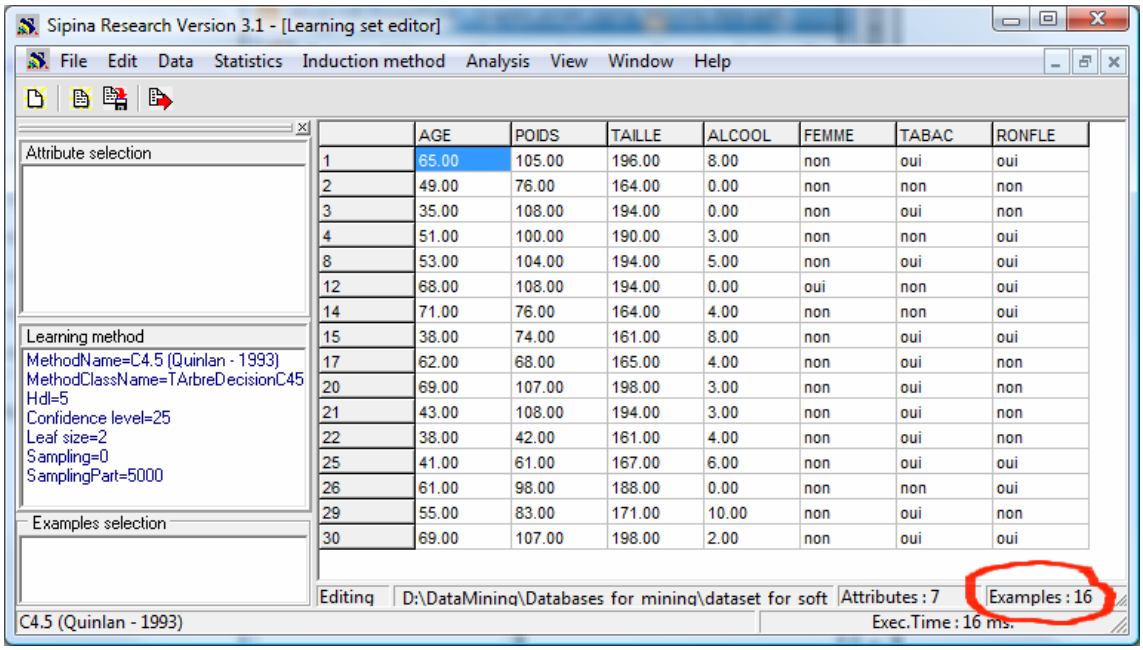
Figure 2 - Grille de données avec valeurs manquantes

Nous observons les cases vides dans la grille de données.

4.1 Suppression des observations

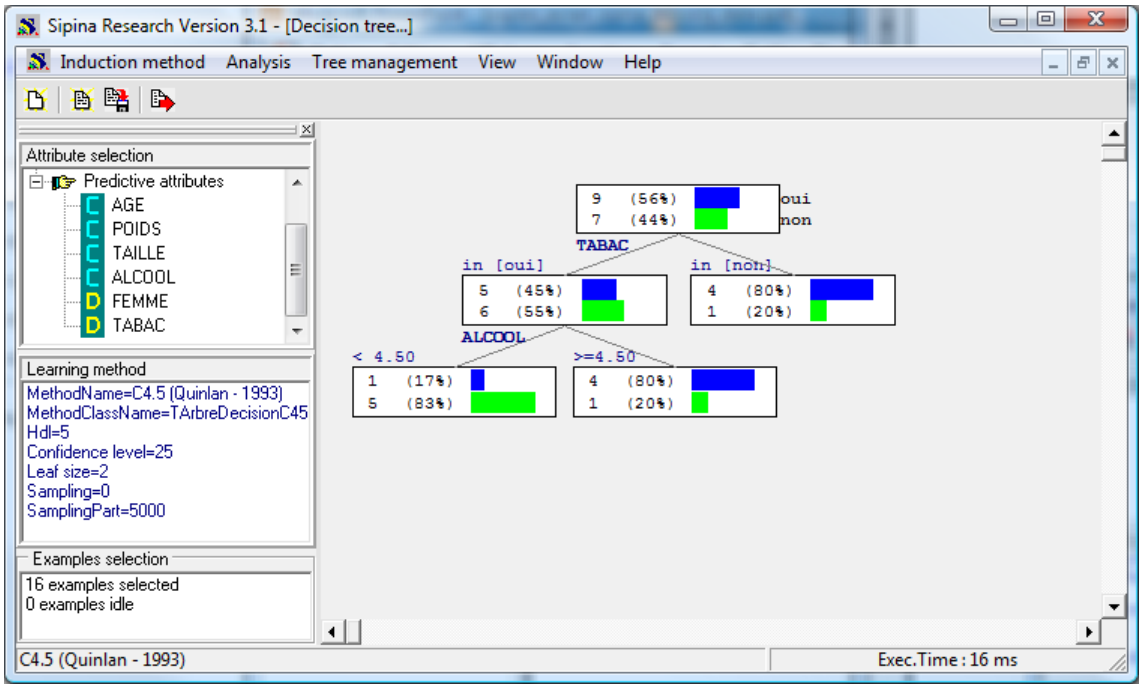
Une première stratégie, très simple, mais non dénuée d'intérêt malgré l'a priori négatif qui l'accompagne, est la suppression des observations comportant au moins une valeur manquante. Si les valeurs manquantes sont aléatoirement réparties avec une proportion assez faible, la stratégie est tout à fait viable. En revanche, si elles sont très nombreuses ou concentrées sur une des variables, l'approche a des conséquences catastrophiques, on peut littéralement « vider » le fichier de ses observations.

Nous actionnons le menu STATISTICS / MISSING DATA / DELETE EXAMPLES. Nous voyons immédiatement le tableau rétrécir, il ne reste plus que 16 observations, près de 50% de l'effectif initial.



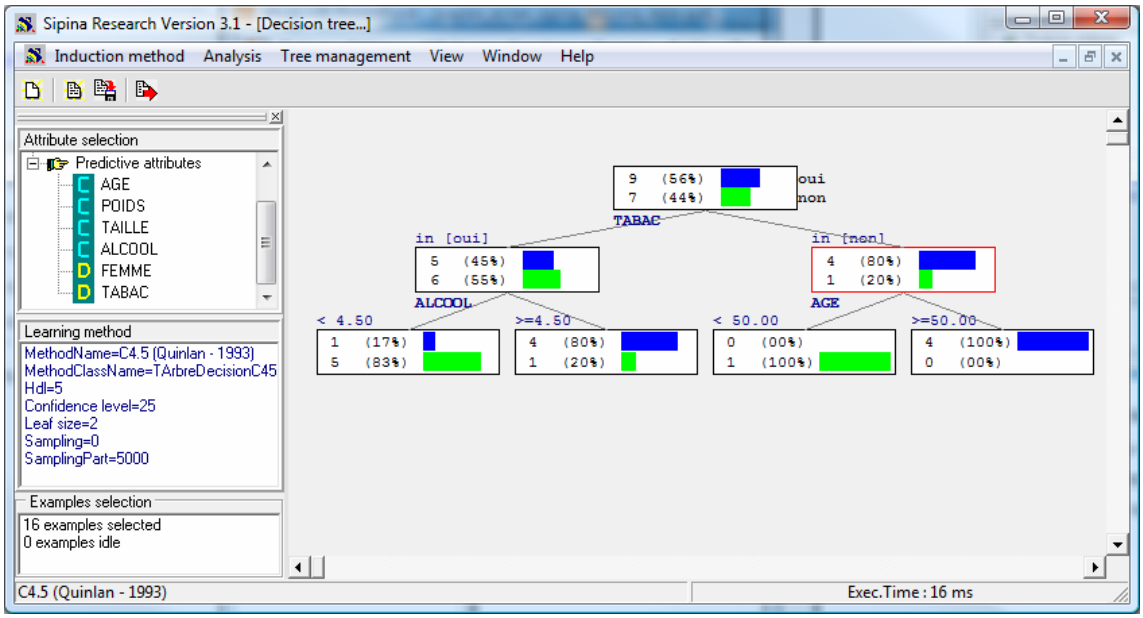
	AGE	POIDS	TAILLE	ALCOOL	FEMME	TABAC	RONFLE
1	65.00	105.00	196.00	8.00	non	oui	oui
2	49.00	76.00	164.00	0.00	non	non	non
3	35.00	108.00	194.00	0.00	non	oui	non
4	51.00	100.00	190.00	3.00	non	non	oui
8	53.00	104.00	194.00	5.00	non	oui	oui
12	68.00	108.00	194.00	0.00	oui	non	oui
14	71.00	76.00	164.00	4.00	non	non	oui
15	38.00	74.00	161.00	8.00	non	oui	oui
17	62.00	68.00	165.00	4.00	non	oui	non
20	69.00	107.00	198.00	3.00	non	oui	non
21	43.00	108.00	194.00	3.00	non	oui	non
22	38.00	42.00	161.00	4.00	non	oui	non
25	41.00	61.00	167.00	6.00	non	oui	oui
26	61.00	98.00	188.00	0.00	non	non	oui
29	55.00	83.00	171.00	10.00	non	oui	non
30	69.00	107.00	198.00	2.00	non	oui	oui

Qu'importe, nous réitérons l'analyse précédente. Nous définissons la variable à prédire et les variables prédictives (ANALYSIS / DEFINE CLASS ATTRIBUTES) puis nous lançons les traitements (ANALYSIS / LEARNING). Nous obtenons l'arbre suivant.



Finalement, malgré la réduction drastique du fichier de données, nous obtenons un arbre quasi-identique à l'arbre sur les données complètes (Figure 1). C4.5 n'a pas pu développer le nœud à droite parce qu'il dispose de trop peu d'observations, il ne peut pas produire des feuilles avec moins de 2

observations conformément au paramétrage par défaut de la méthode. Si nous forçons quand même la segmentation, nous obtiendrions l'arbre suivant.



C'est la copie conforme de l'arbre sur les données complètes. Seuls les effectifs sont différents, ce qui est tout à fait normal. Que les bornes de discrétisation soient les mêmes est un sacré coup de chance en revanche.

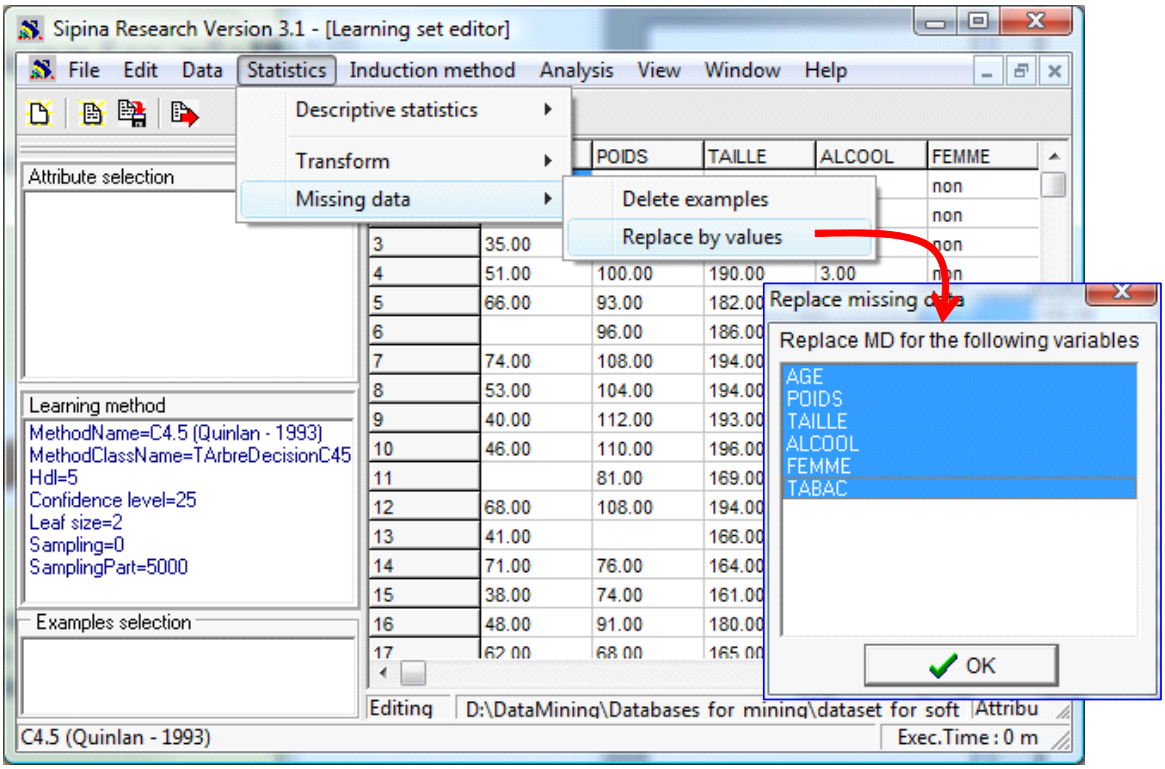
Lorsque nous utilisons la stratégie de suppression d'individus, nous constatons que les modifications sont minimes sur l'arbre produit dans la mesure où les valeurs manquantes sont aléatoirement répartis, et qu'ils sont proportionnellement faibles. En revanche, il faut modifier le paramétrage de la méthode pour l'adapter à la réduction des effectifs.

4.2 Remplacement des valeurs – A

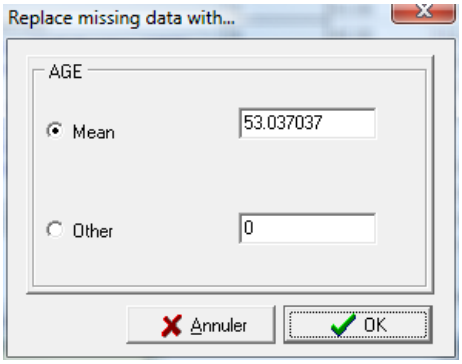
Voyons ce qu'il se passe si nous utilisons des stratégies de remplacement des valeurs manquantes : la moyenne pour les variables quantitatives, la valeur la plus fréquente (le mode) pour les qualitatives.

Nous stoppons l'analyse courante (ANALYSIS / STOP ANALYSIS), puis vidons la grille (FILE / NEW). Nous rechargeons le fichier comportant des valeurs manquantes (FILE / OPEN). Nous retrouvons notre grille de départ (Figure 2).

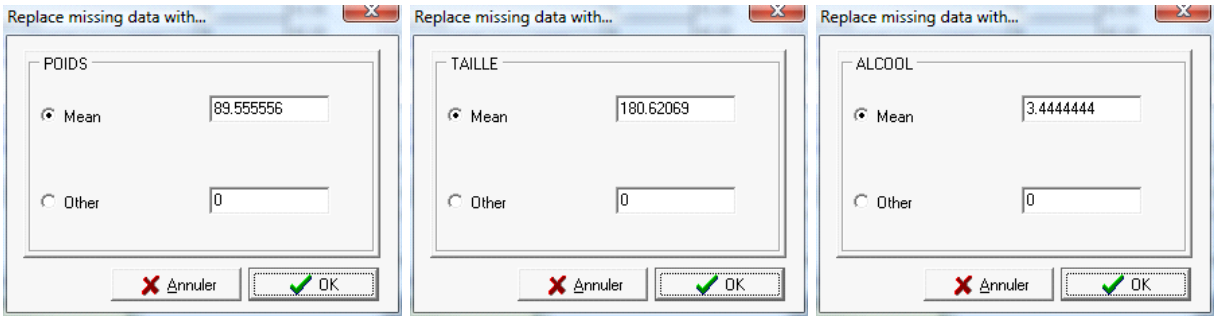
Nous actionnons maintenant le menu STATISTICS / MISSING DATA / REPLACE BY VALUES. Sipina affiche une boîte de dialogue listant les variables comportant au moins une valeur manquante. On notera ainsi que la variable à prédire RONFLE est complètement renseignée, elle est absente de la liste.



Nous sélectionnons toutes les variables puis nous cliquons sur OK. Pour chaque variable, Sipina propose une valeur de substitution. Pour les variables quantitatives, il propose la moyenne. Nous cliquons sur OK pour l'AGE.

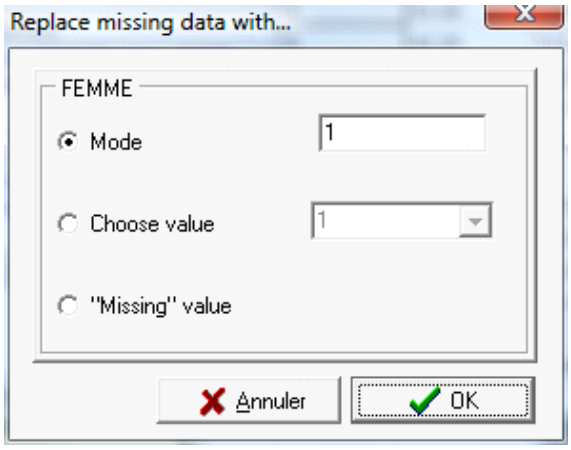


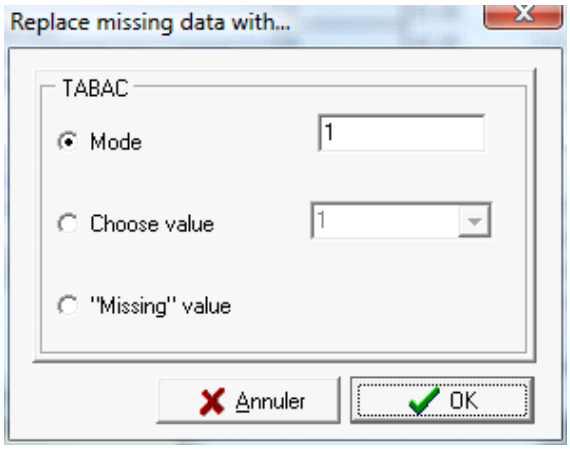
Idem pour le POIDS, la TAILLE et l'ALCOOL.



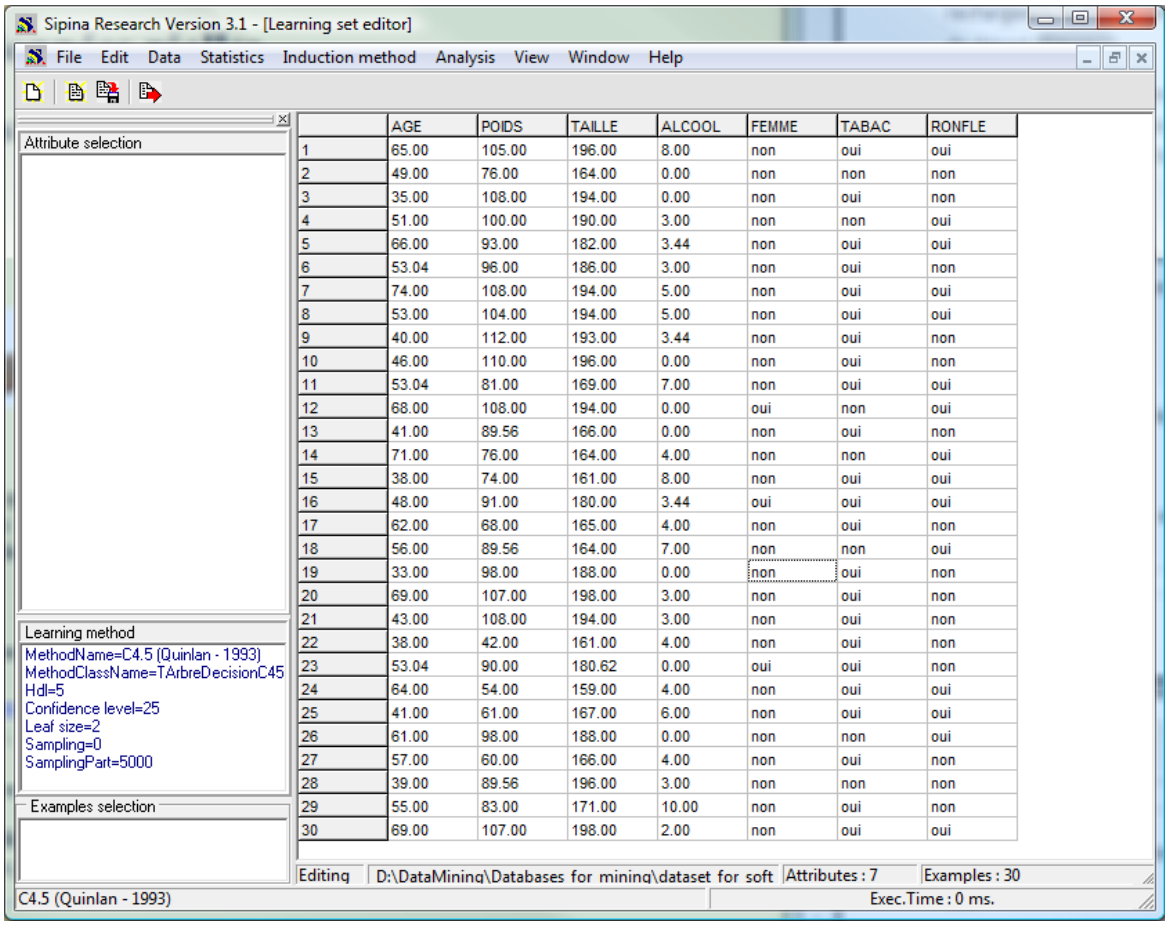
Lorsque nous arrivons au stade de la variable qualitative FEMME. Sipina propose par défaut de

remplacer la valeur manquante par le mode, en l'occurrence FEMME = 1. Nous validons. Idem pour TABAC.



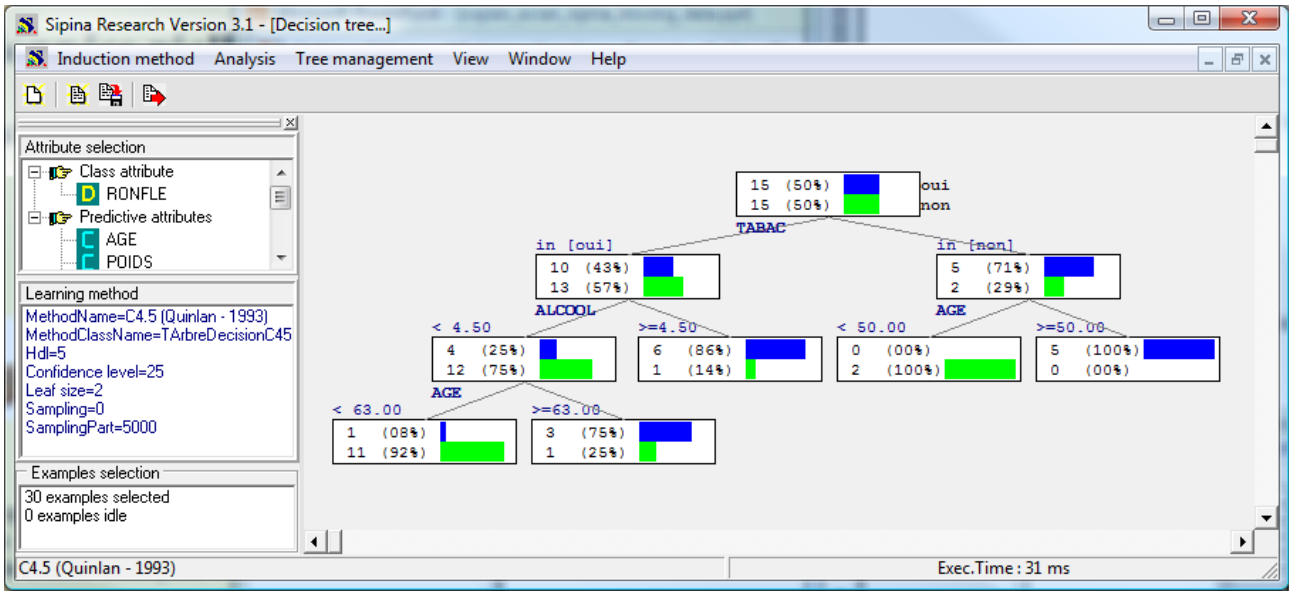


La grille de données est alors complétée.



	AGE	POIDS	TAILLE	ALCOOL	FEMME	TABAC	RONFLE
1	65.00	105.00	196.00	8.00	non	oui	oui
2	49.00	76.00	164.00	0.00	non	non	non
3	35.00	108.00	194.00	0.00	non	oui	non
4	51.00	100.00	190.00	3.00	non	non	oui
5	66.00	93.00	182.00	3.44	non	oui	oui
6	53.04	96.00	186.00	3.00	non	oui	non
7	74.00	108.00	194.00	5.00	non	oui	oui
8	53.00	104.00	194.00	5.00	non	oui	oui
9	40.00	112.00	193.00	3.44	non	oui	non
10	46.00	110.00	196.00	0.00	non	oui	non
11	53.04	81.00	169.00	7.00	non	oui	oui
12	68.00	108.00	194.00	0.00	oui	non	oui
13	41.00	89.56	166.00	0.00	non	oui	non
14	71.00	76.00	164.00	4.00	non	non	oui
15	38.00	74.00	161.00	8.00	non	oui	oui
16	48.00	91.00	180.00	3.44	oui	oui	oui
17	62.00	68.00	165.00	4.00	non	oui	non
18	56.00	89.56	164.00	7.00	non	non	oui
19	33.00	98.00	188.00	0.00	non	oui	non
20	69.00	107.00	198.00	3.00	non	oui	non
21	43.00	108.00	194.00	3.00	non	oui	non
22	38.00	42.00	161.00	4.00	non	oui	non
23	53.04	90.00	180.62	0.00	oui	oui	non
24	64.00	54.00	159.00	4.00	non	oui	oui
25	41.00	61.00	167.00	6.00	non	oui	oui
26	61.00	98.00	188.00	0.00	non	non	oui
27	57.00	60.00	166.00	4.00	non	oui	non
28	39.00	89.56	196.00	3.00	non	non	non
29	55.00	83.00	171.00	10.00	non	oui	non
30	69.00	107.00	198.00	2.00	non	oui	oui

Il ne nous reste plus qu'à lancer les traitements (choisir la variable à prédire et les prédictives, lancer l'analyse). Nous obtenons l'arbre suivant.

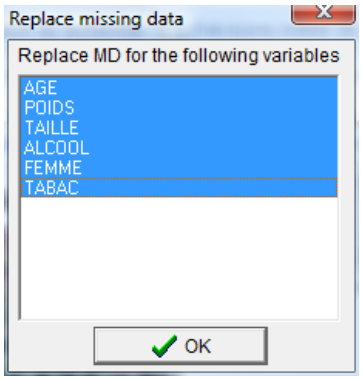


Les trois premiers niveaux sont identiques à notre arbre de référence (Figure 1). Les effectifs sont par contre différents. Les modifications introduites dans les données incitent C4.5 à introduire une segmentation supplémentaire dans la partie gauche de l'arbre. Après coup, ça paraît évident. Nous avons artificiellement favorisé TABAC = OUI (augmenté son poids, il s'y trouve un individu supplémentaire) en remplaçant les valeurs manquantes avec le mode. L'algorithme pense pouvoir trouver des informations intéressantes de ce côté.

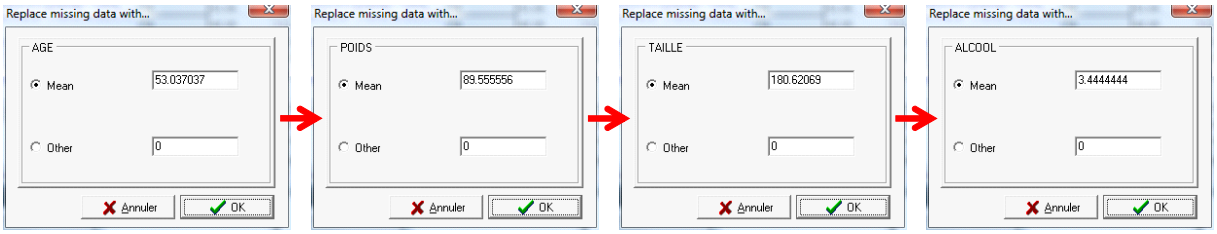
4.3 Remplacement des valeurs – B

Voyons maintenant une autre stratégie de remplacement. Nous utilisons la moyenne toujours pour les variables quantitatives, nous créons une nouvelle modalité « valeur manquante » (_MISSING_) pour les variables qualitatives.

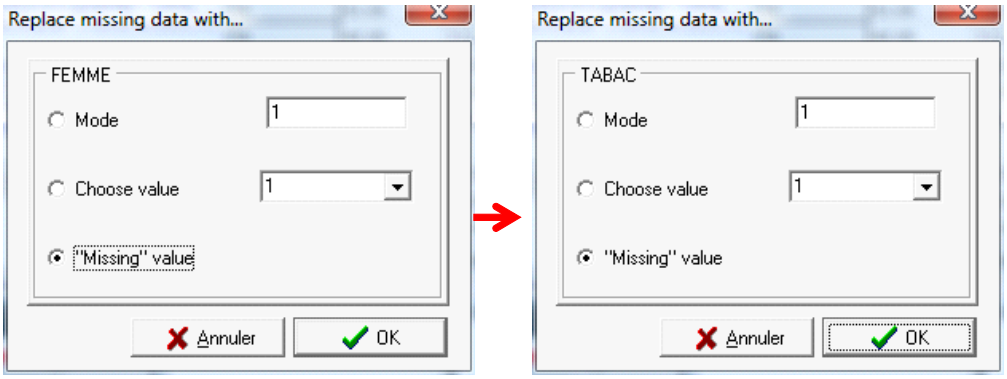
De nouveau, nous stoppons les traitements (ANALYSIS / STOP ANALYSIS) et vidons la grille (FILE / NEW). Nous rechargeons le fichier RONFLEMENT_WITH_MISSING.FDM pour revenir à notre point de départ (Figure 2). Nous cliquons sur STATISTICS / MISSING DATA / REPLACE BY VALUES. La boîte de dialogue indiquant les variables comportant des valeurs manquantes est affichée. Nous les sélectionnons toutes puis nous validons.



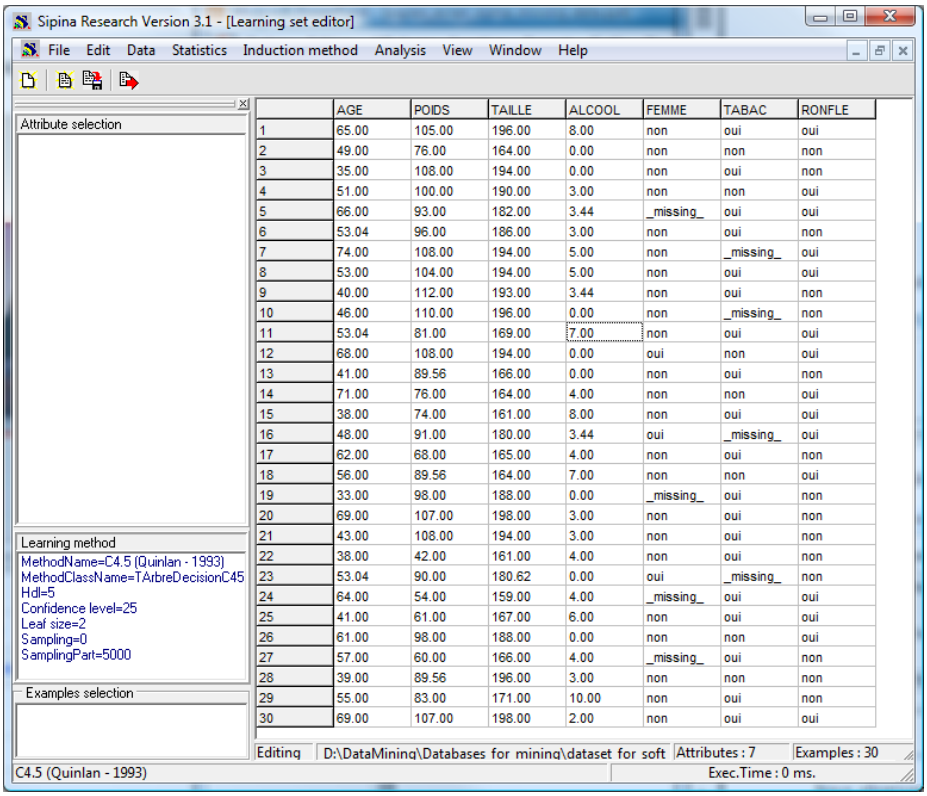
Encore une fois, nous utilisons la moyenne pour les variables continues AGE, POIDS, TAILLE et ALCOOL.



Pour les variables qualitatives, nous sélectionnons maintenant l'option « MISSING » VALUE.

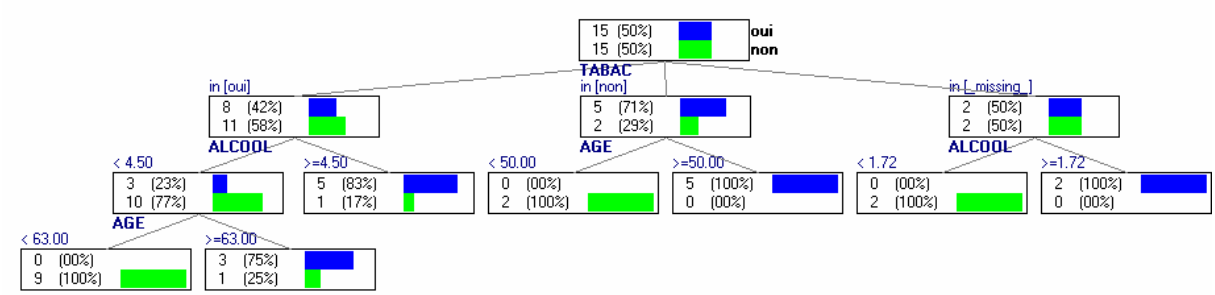


Nous observons une nouvelle grille de données. La modalité « _MISSING_ » saute aux yeux dans les colonnes des variables qualitatives. Le nombre total d'observations n'est pas modifié.



	AGE	POIDS	TAILLE	ALCOOL	FEMME	TABAC	RONFLE
1	65.00	105.00	196.00	8.00	non	oui	oui
2	49.00	76.00	164.00	0.00	non	non	non
3	35.00	108.00	194.00	0.00	non	oui	non
4	51.00	100.00	190.00	3.00	non	non	oui
5	66.00	93.00	182.00	3.44	_missing_	oui	oui
6	53.04	96.00	186.00	3.00	non	oui	non
7	74.00	108.00	194.00	5.00	non	_missing_	oui
8	53.00	104.00	194.00	5.00	non	oui	oui
9	40.00	112.00	193.00	3.44	non	oui	non
10	46.00	110.00	196.00	0.00	non	_missing_	non
11	53.04	81.00	169.00	7.00	non	oui	oui
12	68.00	108.00	194.00	0.00	oui	non	oui
13	41.00	89.56	166.00	0.00	non	oui	non
14	71.00	76.00	164.00	4.00	non	non	oui
15	38.00	74.00	161.00	8.00	non	oui	oui
16	48.00	91.00	180.00	3.44	oui	_missing_	oui
17	62.00	68.00	165.00	4.00	non	oui	non
18	56.00	89.56	164.00	7.00	non	non	oui
19	33.00	98.00	188.00	0.00	_missing_	oui	non
20	69.00	107.00	198.00	3.00	non	oui	non
21	43.00	108.00	194.00	3.00	non	oui	non
22	38.00	42.00	161.00	4.00	non	oui	non
23	53.04	90.00	180.62	0.00	oui	_missing_	non
24	64.00	54.00	159.00	4.00	_missing_	oui	oui
25	41.00	61.00	167.00	6.00	non	oui	oui
26	61.00	98.00	188.00	0.00	non	non	oui
27	57.00	60.00	166.00	4.00	_missing_	oui	non
28	39.00	89.56	196.00	3.00	non	non	non
29	55.00	83.00	171.00	10.00	non	oui	non
30	69.00	107.00	198.00	2.00	non	oui	oui

Il ne reste plus qu'à relancer les traitements (Définir la variable à prédire et les explicatives, puis lancer l'apprentissage). Nous obtenons un nouvel arbre.



L'arbre ressemble peu ou prou à celui construit sur la totalité des données. Nous notons qu'une nouvelle branche s'est formée sur la droite, composée à partir de la valeur _MISSING_ de TABAC. C4.5 en extrait une segmentation parfaite, avec des feuilles pures. C'est typiquement un artefact. Les valeurs manquantes ayant été introduites totalement au hasard, cette nouvelle règle n'est certainement pas reproductible sur un autre fichier.

5 Conclusion

N'allons surtout pas tirer des conclusions définitives à partir d'un petit exemple didactique. Bien sûr d'autres techniques existent. Certaines sont de bon sens : une colonne est quasi-vide, il faut virer la variable, une ligne est quasi-vide, il faut retirer les observations. D'autres sont plus techniques : la méthode du maximum de vraisemblance, l'imputation multiple (Allison, 2001 : chapitres 4 et 5).

Autre aspect important, nous avons privilégié l'analyse qualitative des résultats dans ce didacticiel, en comparant les arbres produits subséquemment (j'adore) au prétraitement des valeurs manquantes. Une approche plus mécanique est possible. On la retrouve souvent dans les publications. L'idée consiste à analyser les conséquences du traitement sur les performances du modèle de prédiction en généralisation. La démarche est schématiquement la suivante :

- Nous scindons une base sans valeurs manquantes en échantillon apprentissage (APP) et test (TEST).
- Nous créons un modèle à partir de APP, mesurons le taux d'erreur sur TEST. Ce sera la référence.
- Nous bruitons APP en retirant au hasard une certaine proportion de valeurs. Nous avons un fichier d'apprentissage avec données manquantes (APP-MD). On peut faire varier cette proportion.
- Nous appliquons sur APP-MD les différentes stratégies de traitement des données manquantes pour construire un modèle de prédiction dont nous mesurons les performances sur TEST. Nous avons ainsi une série de taux d'erreur que l'on peut comparer avec la référence. Nous écrivons un article pour montrer quelle est la meilleure méthode en traitant plusieurs bases UCI IRVINE.

C'est très bien. Il y a deux remarques à faire par rapport à ce schéma : on ne traite toujours que les valeurs manquantes totalement aléatoires dans ce cas ; l'autre piste à creuser est la gestion de données manquantes lorsque nous appliquons le modèle c.-à-d. lorsque les individus de l'échantillon fichier test eux-mêmes ne sont pas décrits complètement. Il existe des travaux très intéressants qui en parlent. L'idée mérite d'être creusée à mon avis. La gestion des données manquantes est au moins aussi importante lors de la phase de classement que lors de la phase d'apprentissage.