La méthode SIPINA

1 Objectif

Décrire la méthode SIPINA à travers une étude de cas.

SIPINA est un logiciel. Mais c'est aussi une méthode d'apprentissage. Elle généralise les arbres en introduisant une opération supplémentaire, la fusion, lors de l'induction du modèle de prédiction. On parle de « Graphes d'Induction »[1].

L'idée de fusion des sommets existe déjà dans des méthodes telles que CART ou CHAID. Mais dans ce cas, il s'agit de procéder au regroupement des feuilles issues du même nœud père lors d'une segmentation. Pour une variable explicative discrète comportant K modalités, CART effectue des regroupements de manière à proposer 2 super modalités, l'arbre est binaire ; CHAID effectue un regroupement sélectif en comparant les profils des distributions, il y a bien regroupement mais l'arbre n'est pas forcément binaire. SIPINA généralise cette idée en permettant le regroupement de 2 feuilles quelconques de la structure. La fusion peut donc s'appliquer à deux feuilles géographiquement éloignées dans le graphe.

Schématiquement, à chaque étape du processus de construction du graphe, la méthode évalue et met en compétition la segmentation d'un nœud et la fusion de deux nœuds. Elle choisit l'opération qui améliore la mesure d'évaluation globale de la partition. Cela est possible car le critère pénalise les nœuds à faibles effectifs. Dans certaines situations, il peut être avantageux de fusionner des sommets avant de segmenter à nouveau. L'objectif est d'explorer plus finement des sous-groupes d'individus, sans tomber dans un des inconvénients récurrents des arbres de décision, la tendance au sur-apprentissage consécutive à l'éparpillement excessif des observations.

La méthode SIPINA n'est disponible que dans la version 2.5 du logiciel. Ce dernier concentre bien des défauts[2]. Mais c'est néanmoins le seul logiciel à proposer la méthode SIPINA telle qu'elle est décrite dans la littérature (voir les références en section 7.3, page 16). C'est la raison pour laquelle je le mets encore en ligne d'ailleurs. Sinon, si l'on veut utiliser d'autres algorithmes d'induction d'arbres (C4.5, CHAID, etc.), il est préférable de se tourner vers la version « Recherche »[3], nettement plus performante et fiable.

Dans ce didacticiel, nous montrons la mise en œuvre de la méthode SIPINA dans le logiciel éponyme, version 2.5. Le problème traité est l'explication du faible poids de certains bébés à la naissance à partir des caractéristiques de la mère. L'interprétation des résultats est anecdotique dans notre contexte. On cherche surtout (1) à montrer la

[1] D. Zighed et R. Rakotomalala, « Graphes d'Induction : Apprentissage et Data Mining », Hermès, 2000.

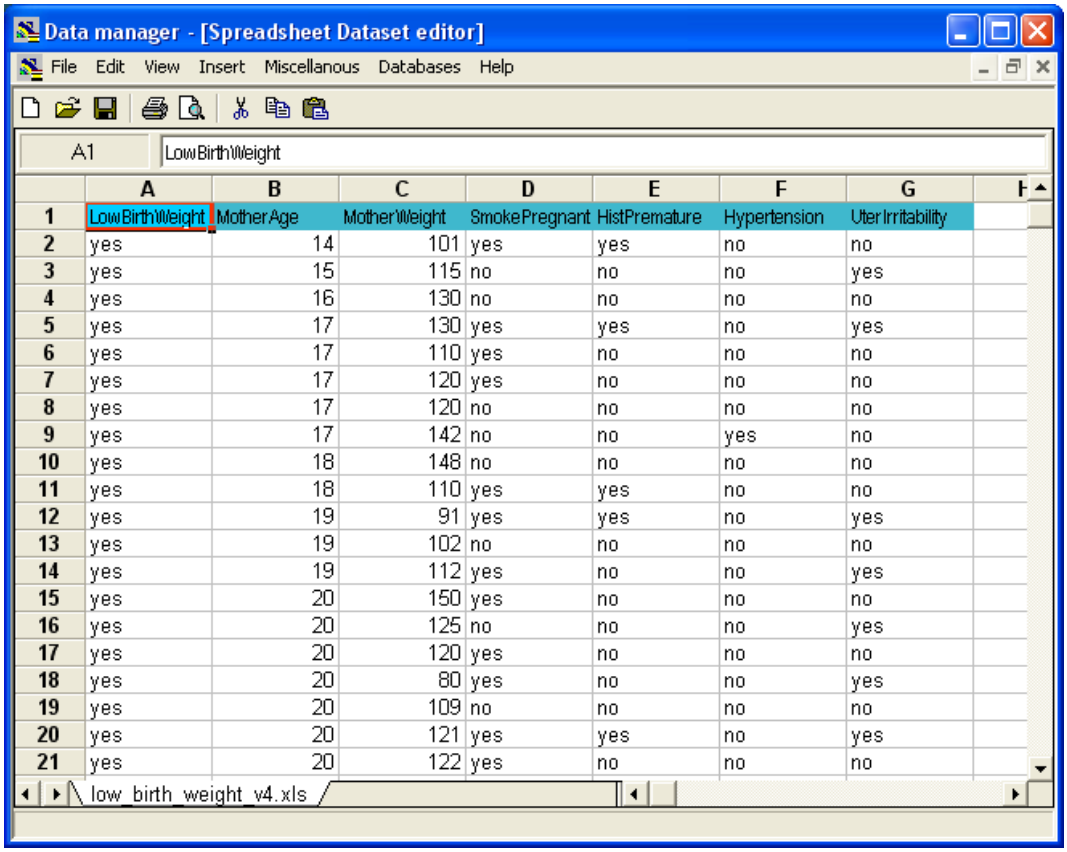
[2] Voir http://sipina.over-blog.fr/pages/Sipina_version_25-361234.html

[3] http://sipina.over-blog.fr/ - Cliquer sur le lien « **Télécharger Sipina** »

prise en main de cette version du logiciel qui est très peu documentée, (2) à mettre en avant les avantages de la méthode lorsque l'on traite des fichiers comportant peu d'observations.

2 Données

Le fichier LOW_BIRTH_WEIGHT_V4.XLS comporte 113 observations et 7 variables. On cherche à prédire le faible poids des bébés à la naissance LOWBIRTHWEIGHT (YES ou NO) à partir de 6 descripteurs relatifs à la mère, à savoir : l'âge, le poids, le fait de fumer pendant la grossesse, le fait d'avoir déjà accouché de prématurés ou non, la présence d'hypertension et l'irritabilité utérine (Figure 1).



	A	B	C	D	E	F	G
1	LowBirthWeight	MotherAge	MotherWeight	SmokePregnant	HistPremature	Hypertension	UterIrritability
2	yes	14	101	yes	yes	no	no
3	yes	15	115	no	no	no	yes
4	yes	16	130	no	no	no	no
5	yes	17	130	yes	yes	no	yes
6	yes	17	110	yes	no	no	no
7	yes	17	120	yes	no	no	no
8	yes	17	120	no	no	no	no
9	yes	17	142	no	no	yes	no
10	yes	18	148	no	no	no	no
11	yes	18	110	yes	yes	no	no
12	yes	19	91	yes	yes	no	yes
13	yes	19	102	no	no	no	no
14	yes	19	112	yes	no	no	yes
15	yes	20	150	yes	no	no	no
16	yes	20	125	no	no	no	yes
17	yes	20	120	yes	no	no	no
18	yes	20	80	yes	no	no	yes
19	yes	20	109	no	no	no	no
20	yes	20	121	yes	yes	no	yes
21	yes	20	122	yes	no	no	no

Figure 1 - Les 20 premières observations du fichier LOW_BIRTH_WEIGHT_V4.XLS

3 Installation de la version 2.5

SIPINA version 2.5 est accessible sur deux sites[4], mais l'URL du fichier d'installation à télécharger est le même http://eric.univ-lyon2.fr/~ricco/softs/Setup_Sipina_V25.exe

[4] http://sipina.over-blog.fr/pages/Sipina_version_25-361234.html

http://eric.univ-lyon2.fr/~ricco/sipina.html (voir Figure 2)

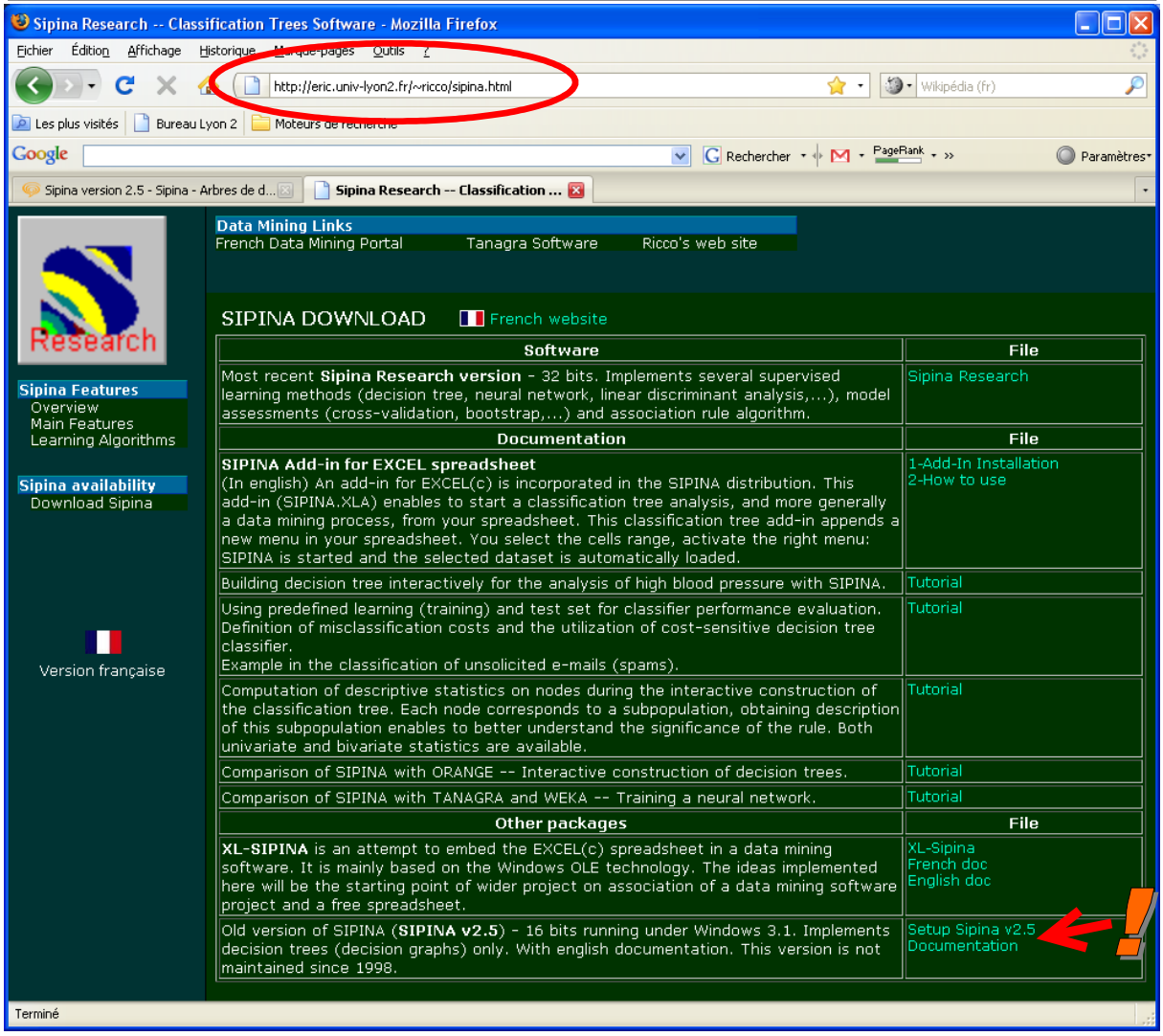
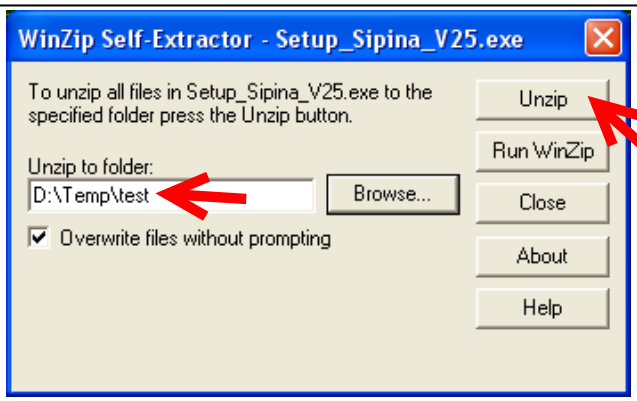


Figure 2 - Un des sites de téléchargement de la version 2.5 de Sipina

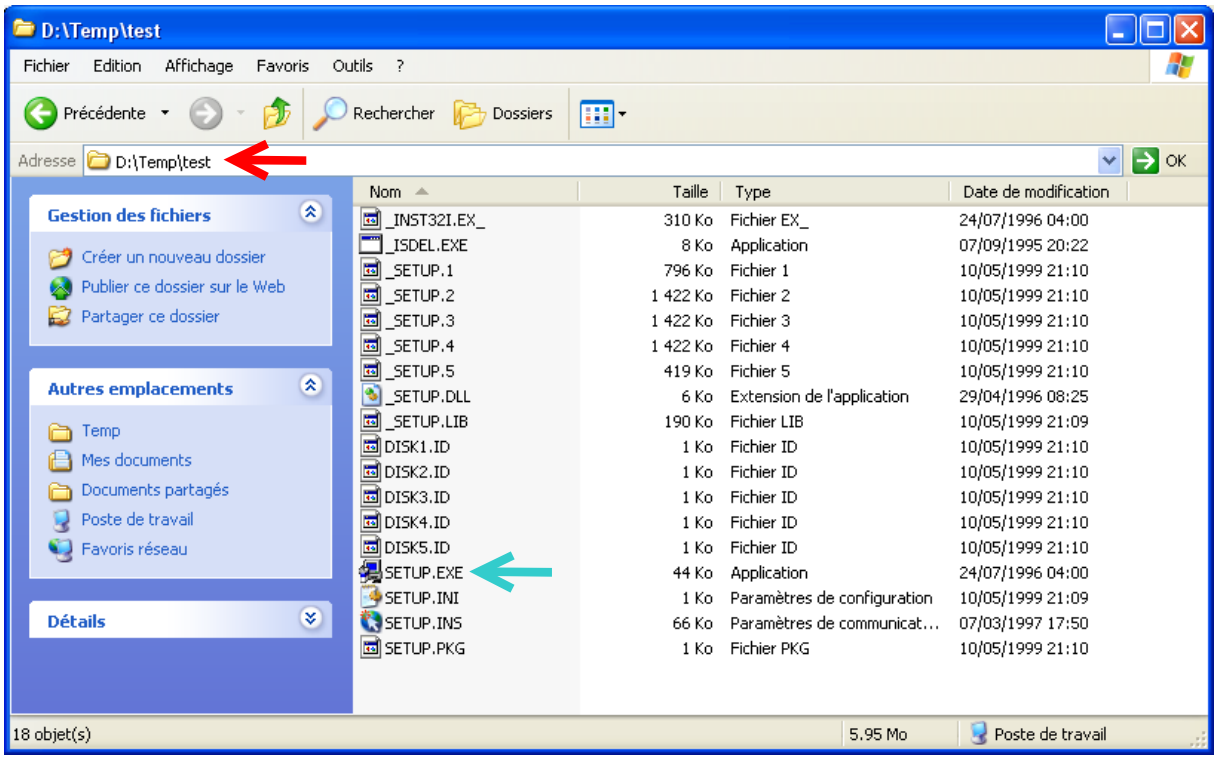
Une documentation en anglais est disponible[5]. Elle est un peu obsolète par rapport à la version en ligne. En effet, elle décrit la version 1.1 alors que le logiciel en est à la 2.5. Néanmoins, elle résume bien les principales fonctionnalités du logiciel.

Après avoir téléchargé le SETUP_SIPINA_V25.EXE, nous l'exécutons. Il demande à copier un certain nombre de fichiers dans un répertoire temporaire. Les fichiers d'installation sont copiés dans ce répertoire en cliquant sur le bouton UNZIP.

[5] http://eric.univ-lyon2.fr/~ricco/softs/EnglishDocSipinaV25.pdf



La liste des fichiers copiés est la suivante



Nous double-cliquons sur le fichier SETUP.EXE pour lancer l'installation proprement dite. La démarche à suivre est conforme à ce que l'on peut attendre de tout script d'installation sous WINDOWS. Le programme peut être installé sur n'importe quel disque de la machine. Les composants additionnels (DLL et OCX) sont copiés dans les sous répertoires de l'exécutable.

Remarque : Les tests ont été effectués sur les versions de WINDOWS allant de 95 à XP. Je ne sais pas comment se comporte le SETUP et SIPINA v2.5 lui même sous VISTA.

Une fois l'installation réalisée. Le programme est accessible via le menu DEMARRER / PROGRAMMES de WINDOWS. Le nom du dossier, s'il n'a pas été explicitement modifié, est **SIPINA v2.5**.

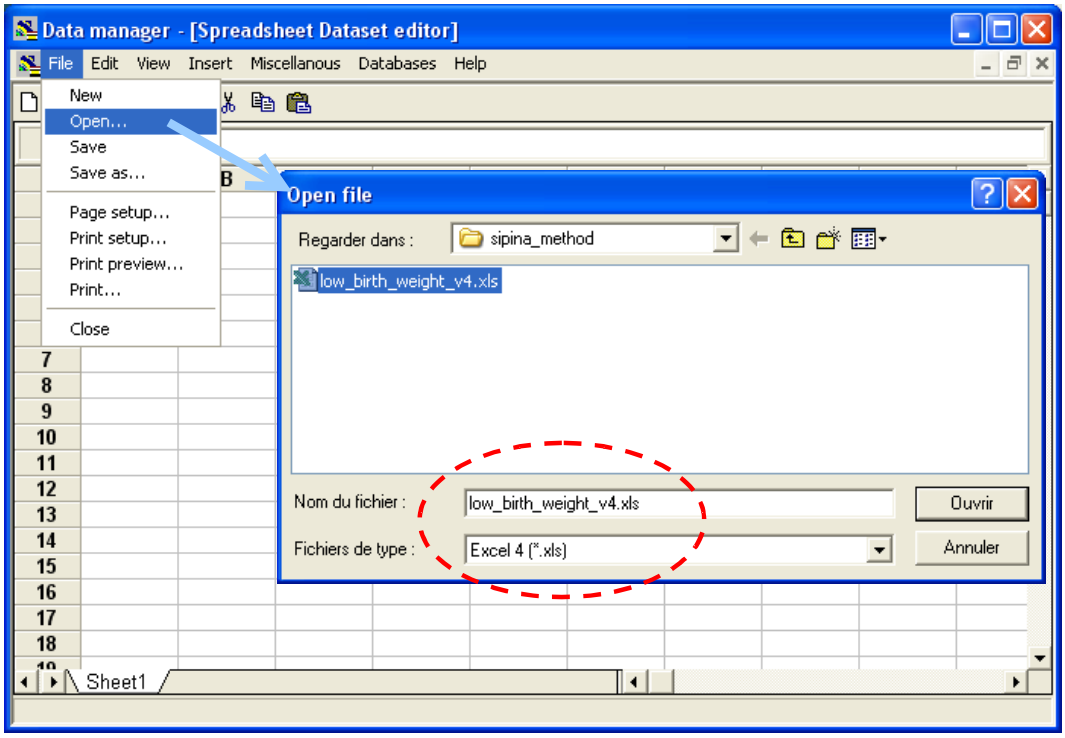
4 Préparation des données avec DATAMANAGER

Cette version de SIPINA utilise un format propriétaire. Les données sont décomposées en 2 fichiers : la partie « PAR » correspond au dictionnaire des données ; la partie « DAT » contient les valeurs. La préparation de ces fichiers est pour le moins ardue. C'est pour cette raison que nous avons adjoint à SIPINA un outil spécifique, DATAMANAGER, dédié à la manipulation et la préparation des fichiers PAR et DAT.

4.1 Lecture des données au format EXCEL 4.0

DATAMANAGER peut lire directement les fichiers de la version 4.0 et 5.0 d'EXCEL[6]. Le plus simple est d'exporter le fichier de données à partir d'un tableur, que ce soit OPENOFFICE ou une version plus récente d'EXCEL, en précisant un de ces formats.

Ceci étant réalisé, nous pouvons alors ouvrir le fichier dans SIPINA DATAMANAGER, accessible dans le dossier DEMARRER / PROGRAMMES / SIPINA v2.5. Nous chargeons notre fichier de données LOW_BIRTH_WEIGHT_V4.XLS, au format EXCEL 4.0, à l'aide du menu FILE/OPEN.



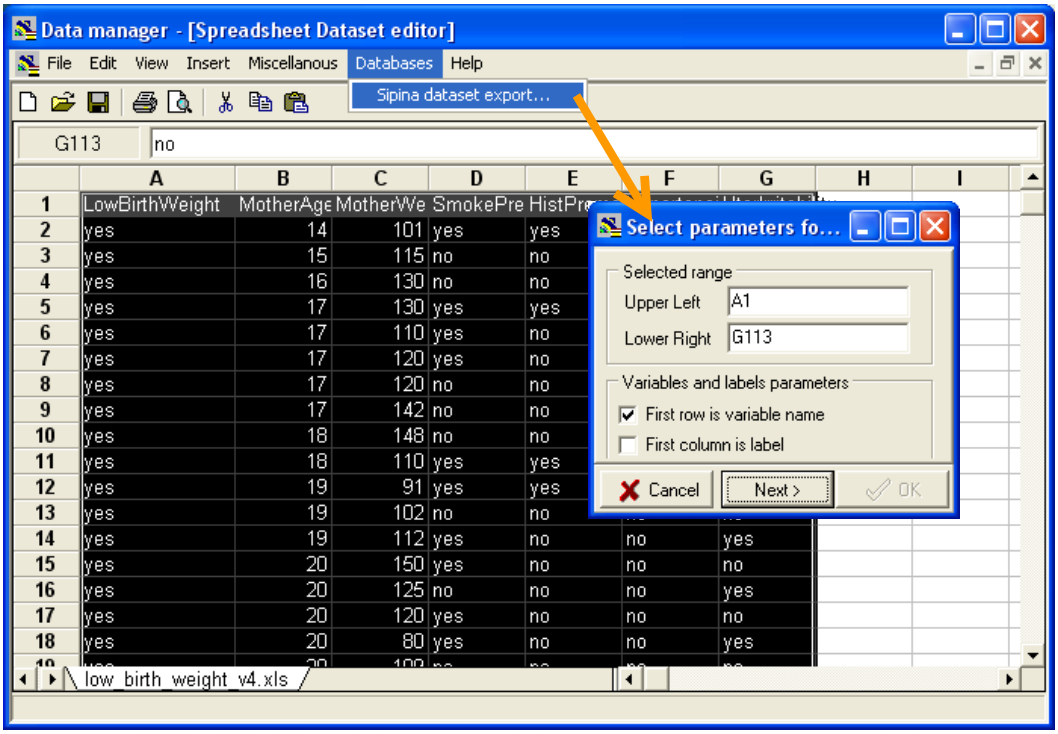
DATAMANAGER possède les fonctionnalités d'un tableur. Il nous est possible d'ajouter de nouvelles colonnes, de rentrer des valeurs, de définir des cellules calculées. La seule restriction est qu'il est limité à 16384 lignes, comme les anciens tableurs. C'est néanmoins amplement suffisant pour des applications académiques.

[6] La version 4.0 d'EXCEL ne gère que les classeurs à une seule feuille de calcul.

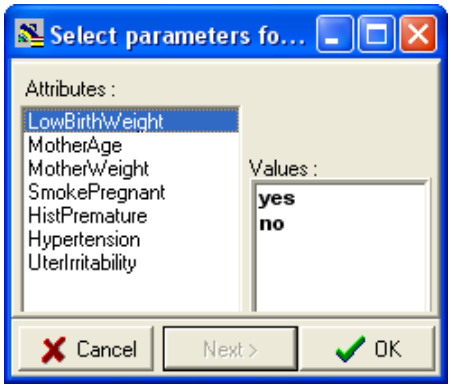
4.2 Elaboration des fichiers DAT et PAR pour SIPINA

Avant d'initier une analyse, nous devons produire les fichiers DAT et PAR à destination de SIPINA v2.5. Nous sélectionnons la plage de donnée, y compris la ligne des noms de variables, puis nous activons le menu DATABASES / SIPINA DATASET EXPORT.

Une boîte de dialogue apparaît, elle précise les coordonnées de la plage de cellules et confirme que la première ligne correspond bien aux noms de variables[7].

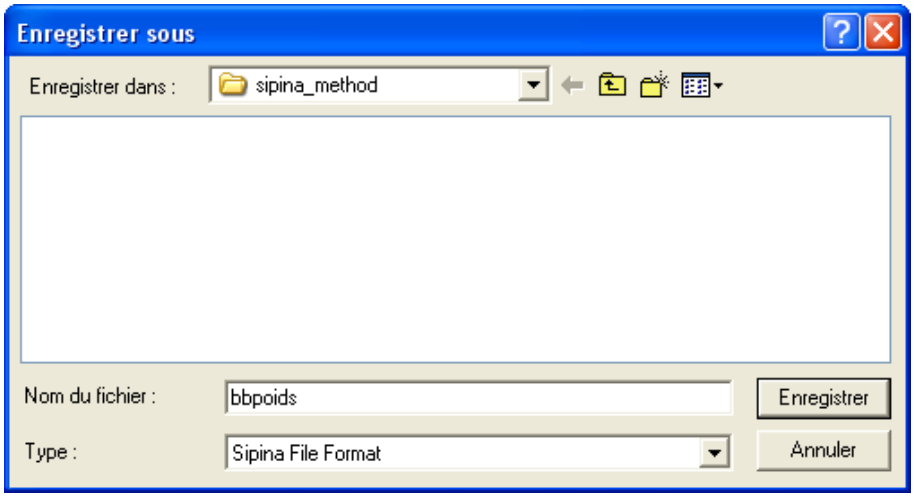


Nous cliquons sur le bouton NEXT. Dans la page suivante, nous devons préciser la variable à prédire, en l'occurrence LOWBIRTHWEIGHT. La liste des modalités apparaît automatiquement. Nous validons en cliquant sur OK.

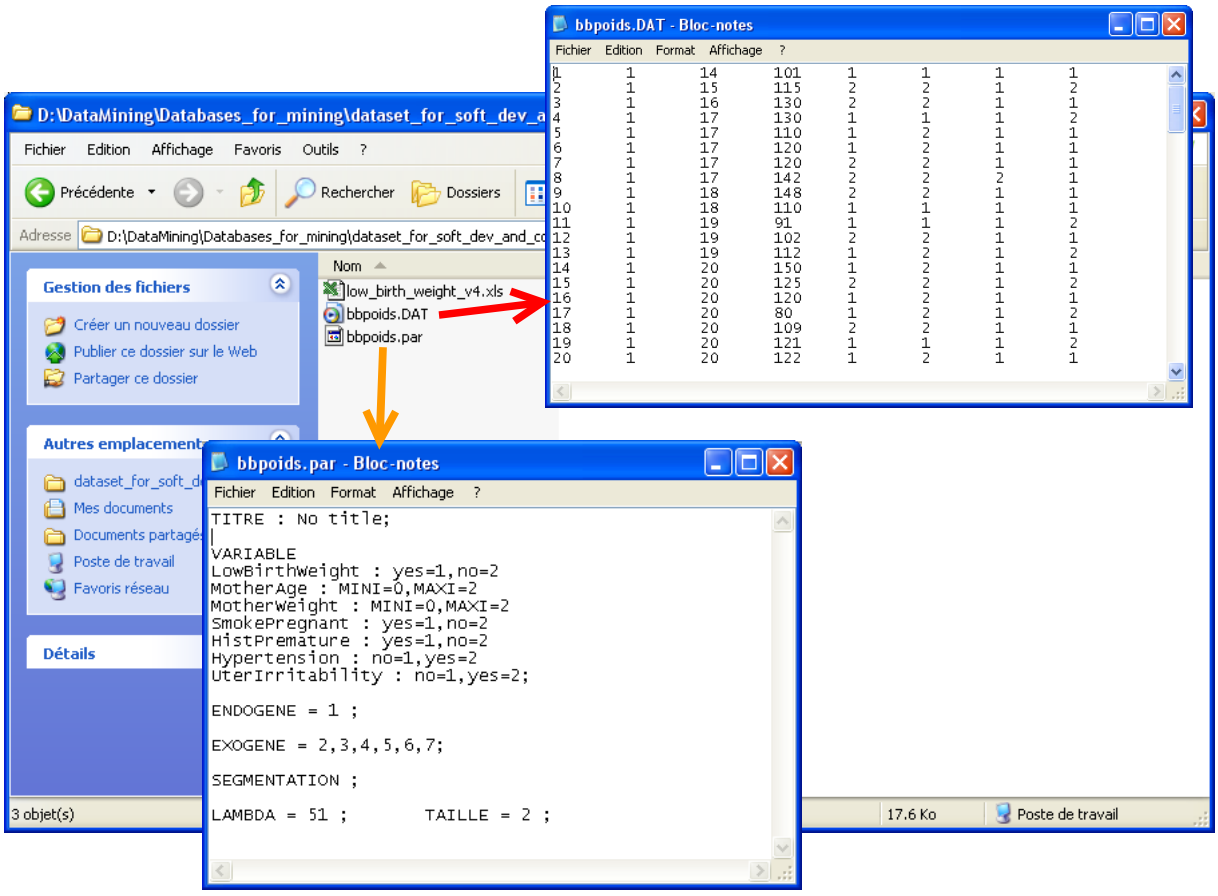


[7] Si ce n'est pas le cas, nous devons désactiver l'option. SIPINA attribuera automatiquement des noms standardisés (VAR1, VAR2, etc.). Autre point important, si la première colonne correspond à des étiquettes (ex. le nom des patients), nous pouvons le préciser dans cette boîte de dialogue.

Une nouvelle boîte de dialogue surgit, nous invitant à spécifier le nom du fichier SIPINA, nous introduisons BBPOIDS.



Un petit détour par l'explorateur WINDOWS nous permet de constater que les fichiers DAT et PAR ont bien été générés. Si nous ouvrons le fichier BBPOIDS.PAR dans le bloc-notes, nous obtenons le dictionnaire de données. Dans le cas du fichier DAT, nous avons la liste des valeurs. La première colonne correspond au numéro des observations.

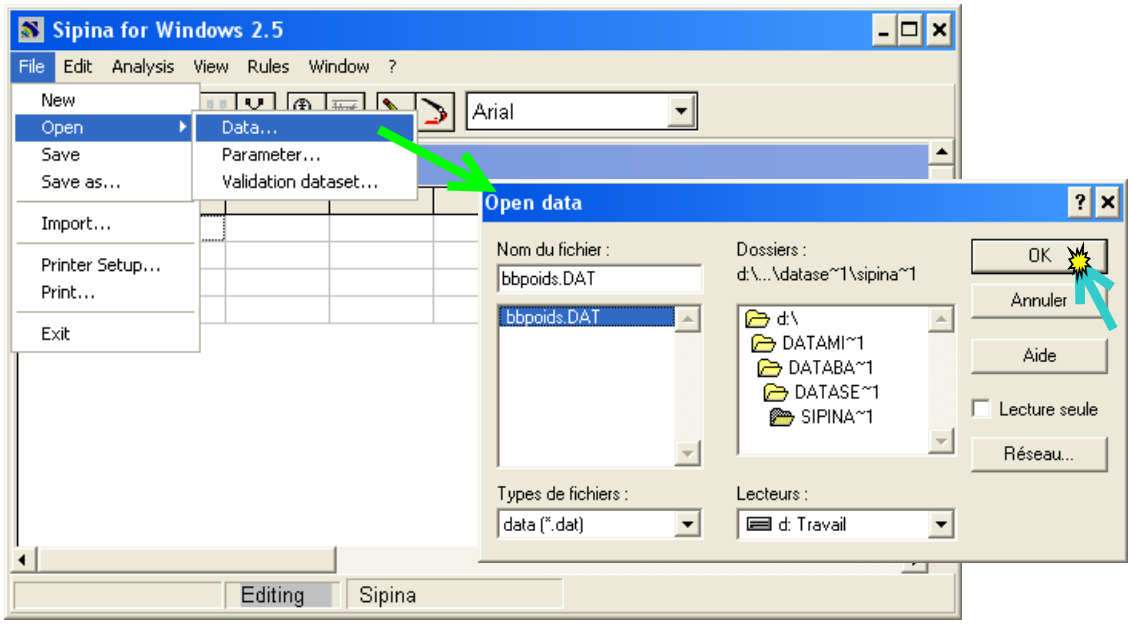


5 Traitements avec le logiciel SIPINA v2.5

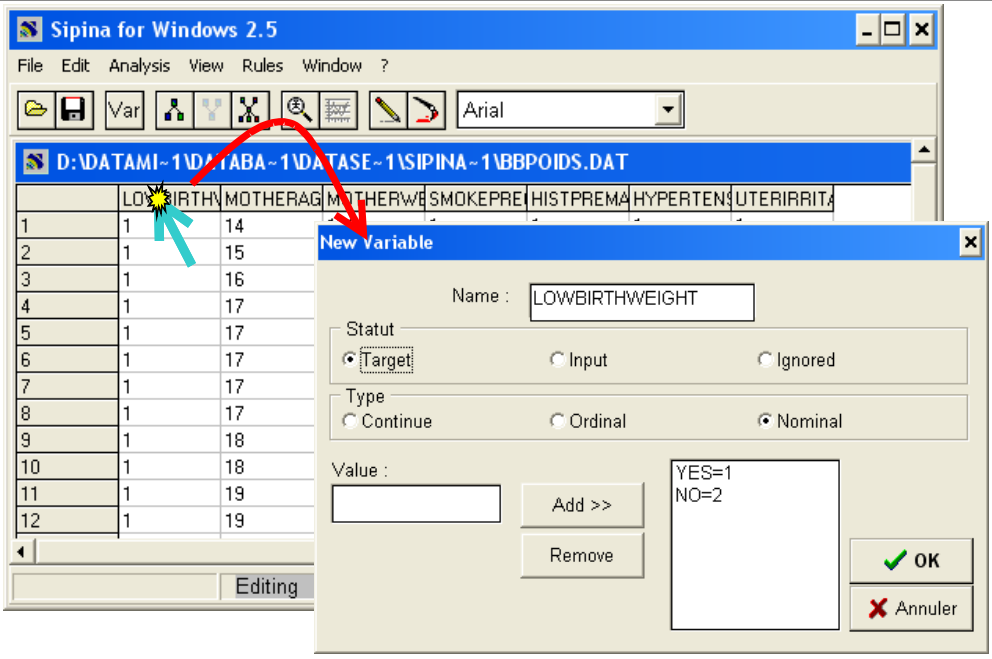
5.1 Chargement des données

Nous pouvons maintenant lancer le logiciel SIPINA version 2.5. Nous activons pour cela le menu DEMARRER / PROGRAMMES / SIPINA V2.5 / SIPINA FOR WINDOWS. Le logiciel est démarré, nous retrouvons l'interface de type tableur, habituelle pour les outils de Data Mining.

Pour charger les données, nous activons le menu FILE / OPEN / DATA. Une boîte de dialogue WINDOWS de saisie de noms de fichier apparaît. Attention, elle est à la norme Windows 16 bits, les noms de dossiers et de répertoire est limité au format 8.3. Cela n'étonnera pas les moins jeunes, c'est peut être un peu plus déroutant pour les autres. Nous validons la sélection en cliquant sur OK.

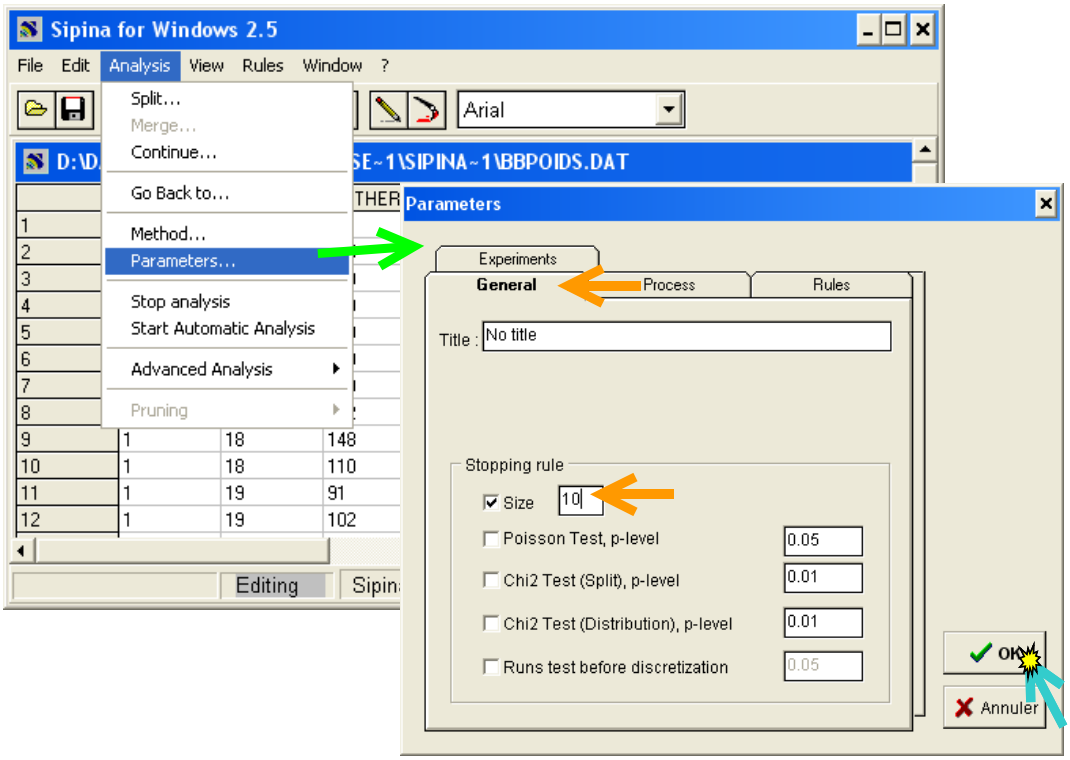


Les données sont automatiquement chargées. Nous avons désigné le fichier DAT, SIPINA a automatiquement cherché le fichier PAR. Ainsi, même si l'affichage est restreint aux codes des modalités pour les variables discrètes, le logiciel sait associer les noms de modalités à leurs codes. Pour s'en convaincre, nous double-cliquons sur l'en-tête de la colonne LOWBIRTHWEIGHT, une boîte de description apparaît. Nous retrouvons les couples codes – description pour cette variable nominale (TYPE = NOMINAL) que nous souhaitons prédire (STATUT = TARGET).



5.2 Paramétrage de la méthode

La méthode SIPINA est sélectionnée automatiquement. Pour modifier ses paramètres, nous activons le menu ANALYSIS / PARAMETERS. La boîte de dialogue adéquate apparaît, nous sélectionnons l'onglet GENERAL.

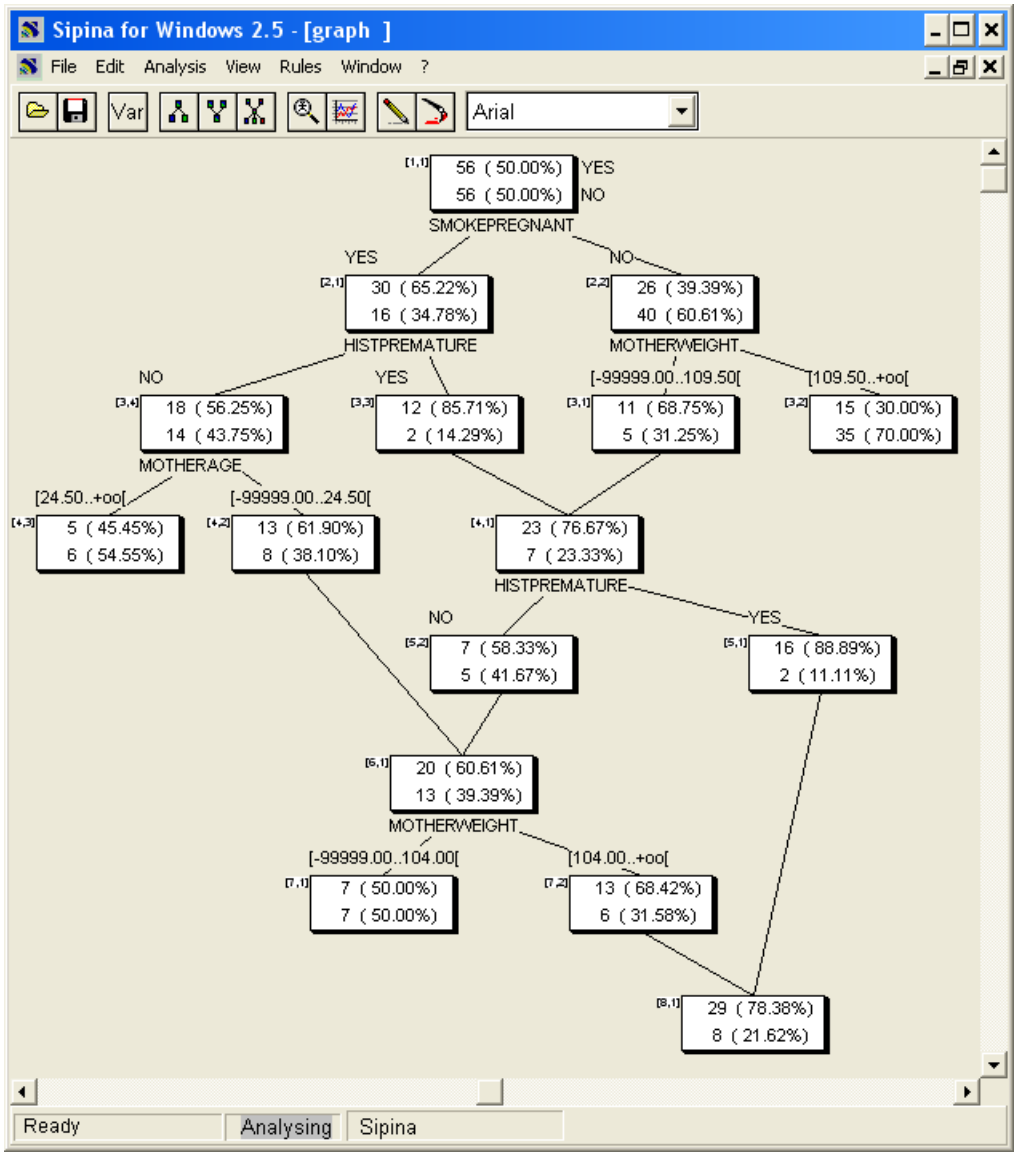


Il y a pléthore de paramètres plus ou moins pertinents. Certains sont un peu folkloriques. Nous restons sur des principes simples dans ce didacticiel. Nous passons le paramètre

SIZE à 10 c.-à-d. nous ne souhaitons pas voir apparaître des sommets avec moins de 10 observations dans notre graphe. Nous validons en cliquant sur OK.

5.3 Lancement des calculs

Nous pouvons lancer les calculs en activant le menu ANALYSIS / START AUTOMATIC ANALYSIS. Le graphe est construit. La structure est un peu emmêlée, nous pouvons améliorer l'affichage en déplaçant manuellement les nœuds. Nous obtenons ainsi la représentation suivante.

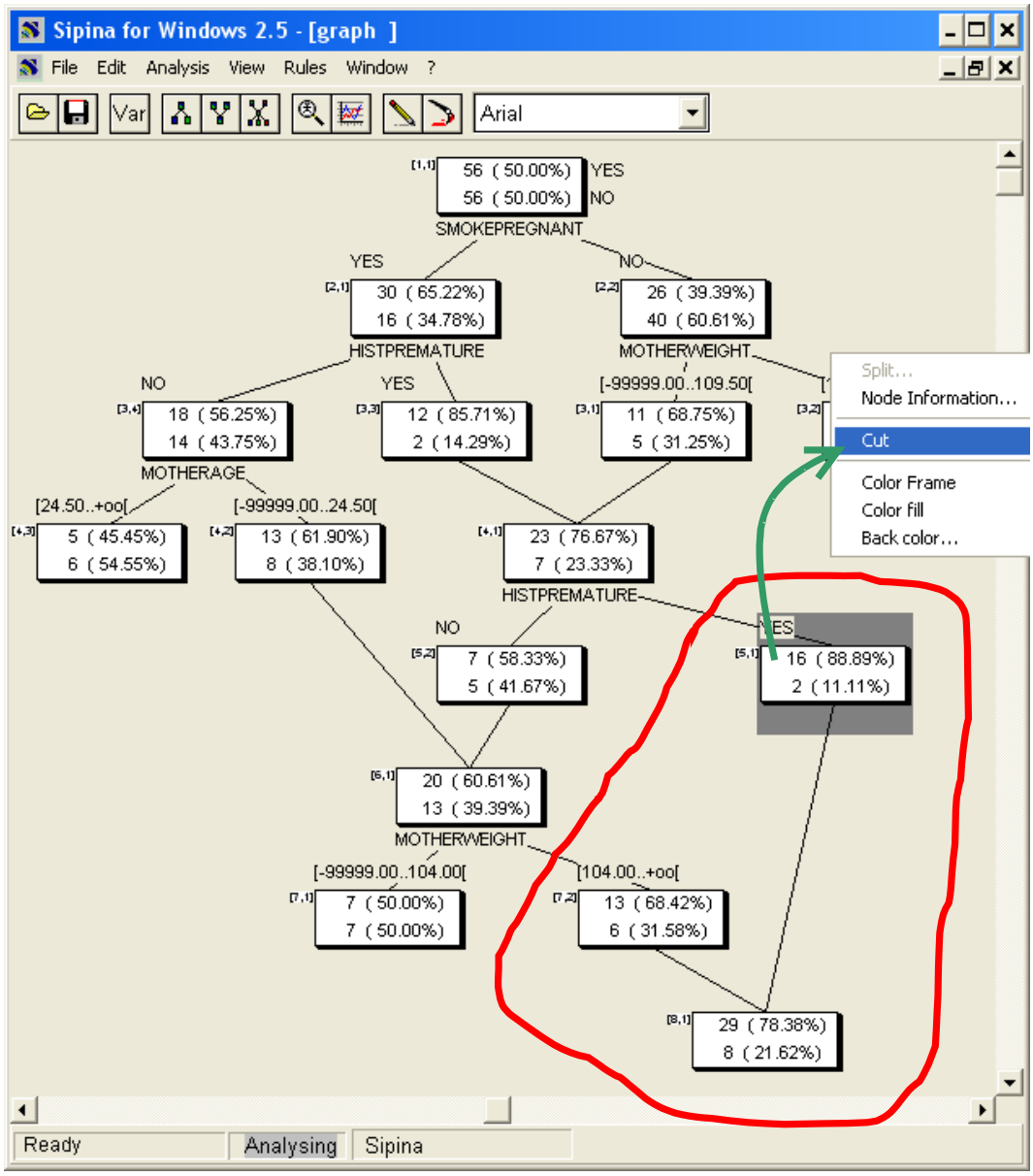


Nous constatons que le graphe de décision est constitué de successions de fusions et segmentations de sommets.

5.4 Elagage manuel des nœuds

Certaines fusions ne sont pas très intéressantes. C'est le cas lorsqu'elle n'est pas suivie d'une segmentation. Cela revient simplement à réunir deux règles de prédiction aboutissant à la même conclusion. Nous pouvons les éliminer.

Nous avons ce type de configuration dans partie sud-est de notre graphe. Pour la supprimer, nous sélectionnons un des sommets pères, nous faisons apparaître le menu contextuel[8] et nous activons le menu CUT.

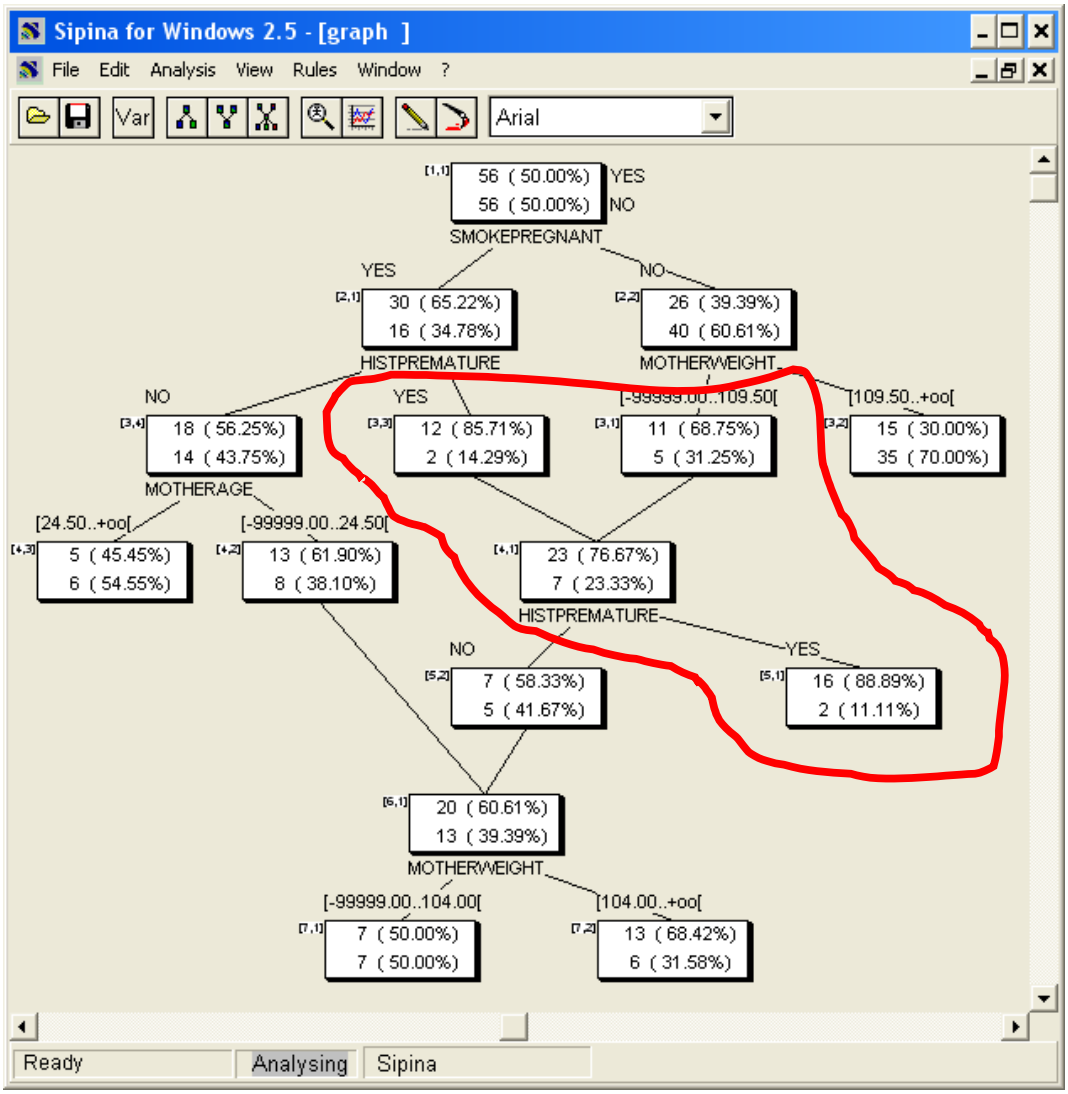


Le graphe est nettoyé.

[8] Le menu contextuel surgit un peu n'importe où. Je ne sais pas trop pourquoi. En tous les cas, nous avons l'assurance qu'il est bien relatif au sommet sélectionné, surlignée en gris.

6 Avantages (et inconvénients) de la méthode SIPINA

Concentrons nous sur une zone particulière du graphe pour comprendre ce qui fait le charme de cette méthode par rapport aux arbres de décision « classiques ».



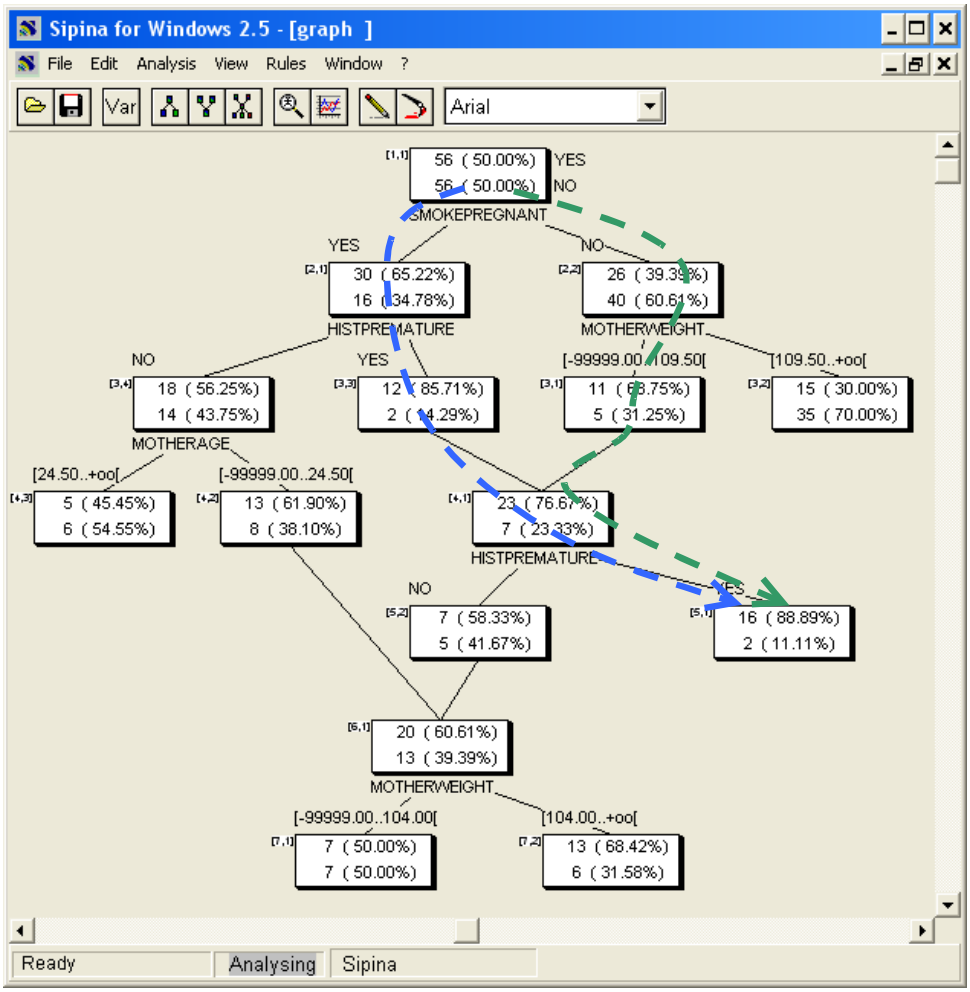
Une méthode d'induction d'arbres se serait arrêtée au niveau 3, avec les feuilles [3,3] et [3,1]. Nous aurions obtenu deux règles de prédiction menant à la même conclusion. La couverture[9] de la première [3, 3] est 14/112 = 12.50%, sa précision 12/14 = 85.71%. Pour la seconde, nous avons respectivement : couverture = 16/112 = 14.29%, et précision 11/16 = 68.75%.

SIPINA ne s'arrête pas à ce stade. Il introduit une fusion de manière à réunir les observations relatives à ces nœuds. Puis il introduit une nouvelle segmentation afin

[9] Dans la littérature « Machine Learning », nous aurions appelé cet indicateur « support ». Mais depuis l'avènement des règles d'association, l'acception usuelle de ce terme a été modifiée. Il représente la proportion d'individus positifs couverts par la règle. Dans notre cas, le support serait 12/112 = 10.71%. Pour éviter les confusions, je préfère utiliser le terme « Couverture ». Il représente la proportion d'individus présents sur le nœud, soit 14/112 = 12.5%.

d'analyser plus finement les disparités entre les individus. Nous pouvons produire ainsi, entre autres, la feuille [5,1] qui est une nouvelle règle de prédiction avec une couverture 18/112 = 16.07% et une précision 16/18=88.89% plus élevées, meilleures en tous les cas que celles des règles que l'on aurait pu associer aux sommets [3,3] et [3,1].

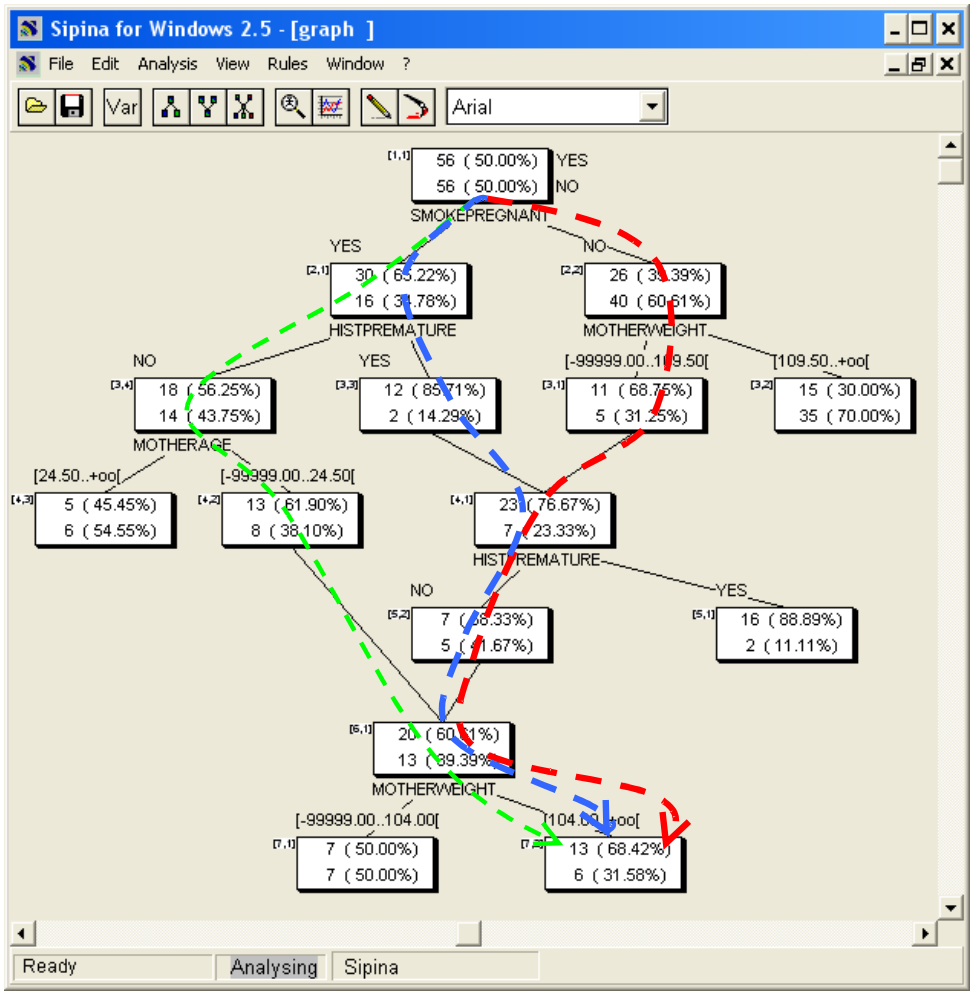
Cette capacité d'exploration supplémentaire liée à un biais de représentation plus puissant constitue le principal avantage de la méthode… mais aussi sa principale faiblesse. En effet, la règle de prédiction, formée de conjonctions (ET) et de disjonctions (OU) devient plus difficile à appréhender. Dans notre exemple, il y a deux chemins de la racine à la feuille. La règle se lit : « **si** (SOMKEPREGNANT = YES **et** HISTPREMATURE = YES **et** HISTPREMATURE = YES) **ou** (SMOKEPREGNANT = NO **et** MOTHERWEIGHT < 109.5 **et** HISTPREMATURE = YES) **alors** LOWERBIRTHWEIGHT = YES ».



Certes, il est parfois possible de procéder à des simplifications. Dans notre exemple, la règle peut être réduite en enlevant les propositions redondantes. Il reste que l'on perd (un peu) ce qui fait le charme des arbres de décision : proposer des règles d'affectation simples, faciles à interpréter et à manipuler.

Si l'on considère la règle associée au sommet [7,2] désignant également les bébés à faible poids, la couverture est 19/112 = 16.96% et la précision 13/19 = 68.42%. Il y a 3

chemins possibles de la racine à la feuille, la règle comporte donc 3 disjonctions, elle est encore un peu plus complexe que la précédente.

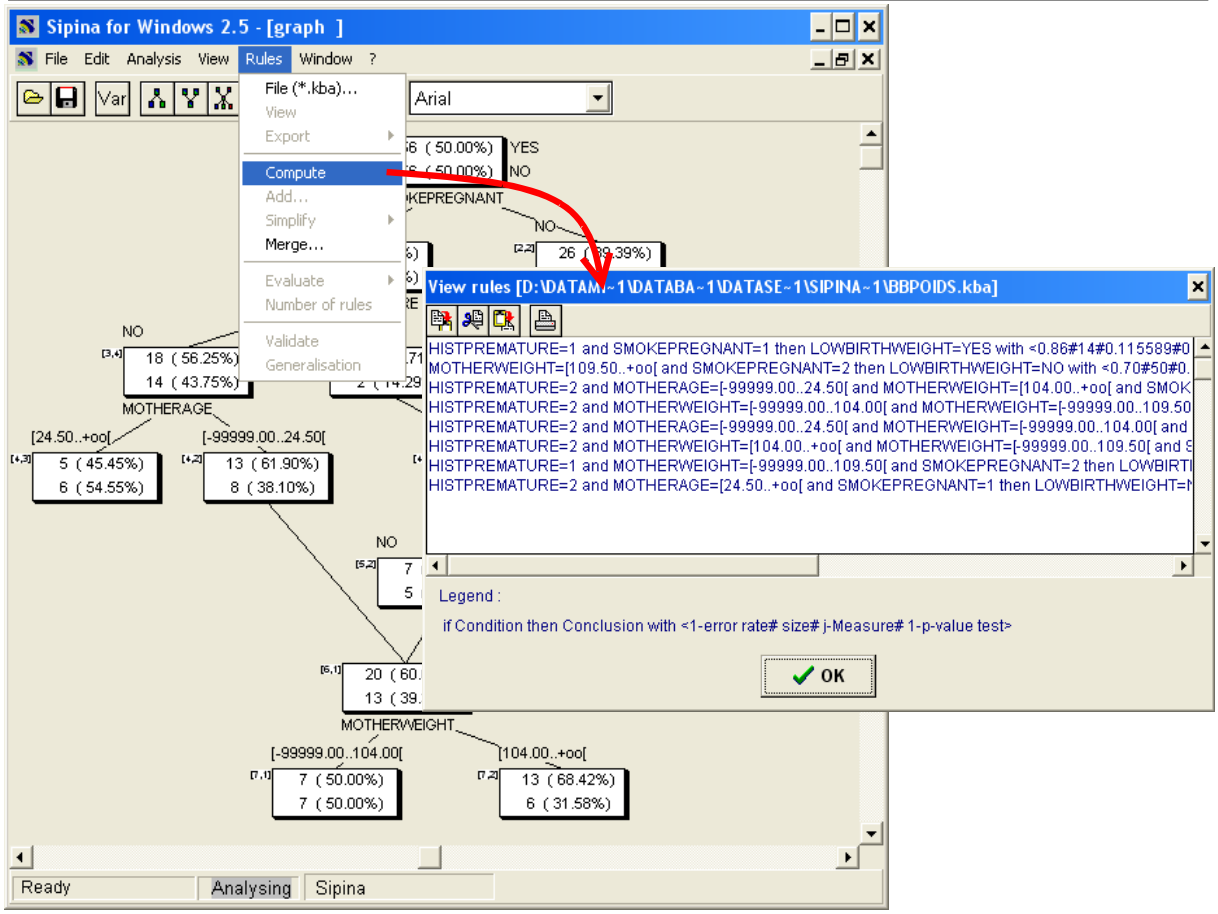


7 D'autres fonctionnalités de SIPINA v2.5

Cette version de SIPINA intègre d'autres outils assez originaux, faisons-en un tour d'horizon rapide.

7.1 Extraction des règles

La lecture des règles de prédiction reste un des écueils majeurs des graphes. Pour dépasser cela, SIPINA peut les produire automatiquement en parcourant tous les chemins de la structure. Pour ce faire, nous activons le menu RULES / COMPUTE. Une fenêtre énumérant les règles apparaît. Il y en a 8 dans notre exemple. Elles sont accompagnées de quelques indicateurs statistiques.



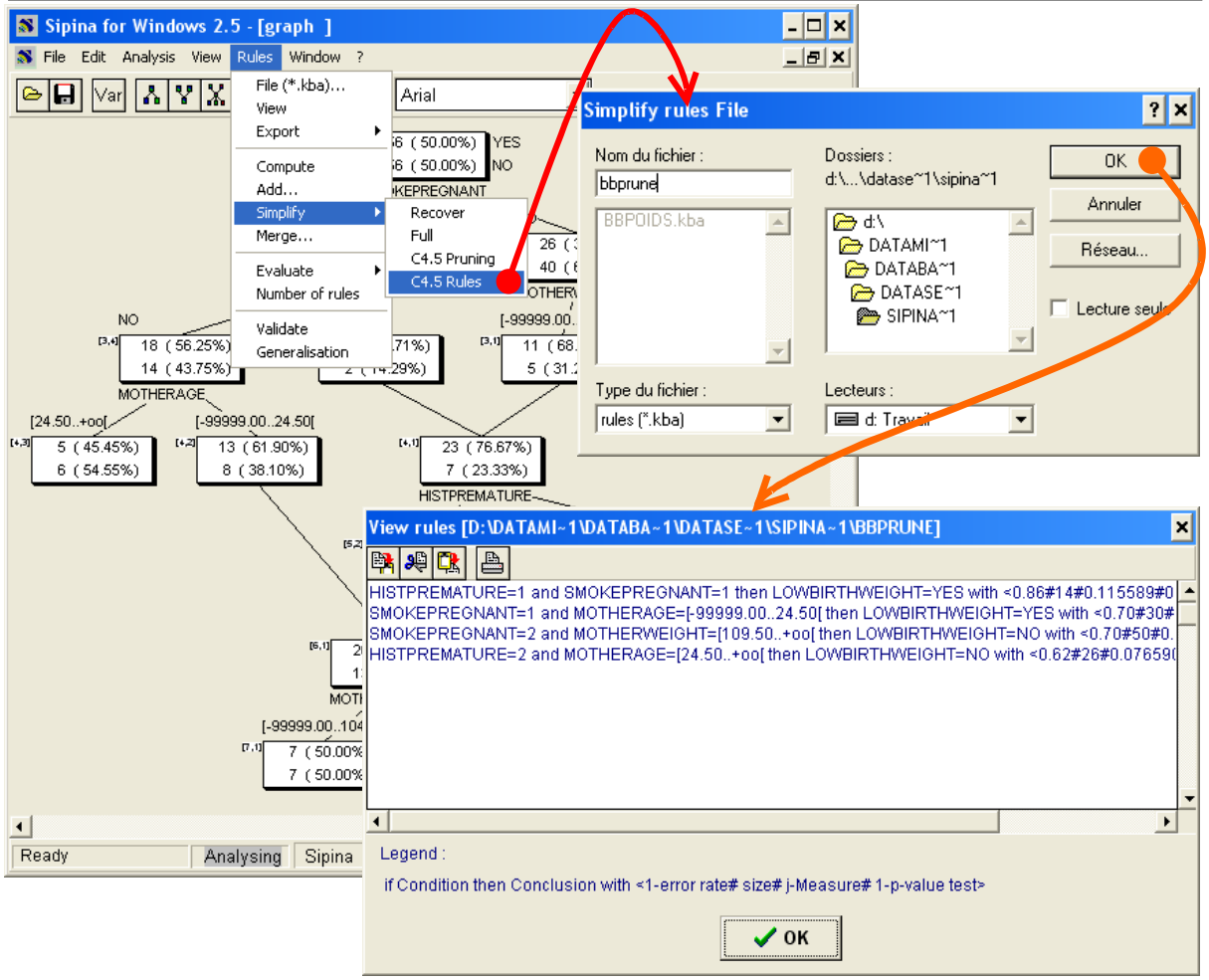
Notons qu'un post traitement a permis de produire des règles formées exclusivement de conjonctions. L'astuce consiste simplement à les associer non pas aux feuilles du graphe mais aux chemins menant de la racine aux feuilles.

7.2 Simplification des bases de règles

Plus le modèle de prédiction est complexe, avec de nombreuses règles, plus difficile sera son interprétation. SIPINA intègre plusieurs procédures de simplification des bases de règles. L'objectif est d'améliorer la lisibilité du modèle en raccourcissant les règles et en diminuant leur nombre, tout en préservant les performances en prédiction.

Nous souhaitons par exemple traiter la base précédente (8 règles) à l'aide de l'algorithme C4.5 RULES de Quinlan (1993). Nous activons le menu RULES / SIMPLIFY / C4.5 RULES. Une boîte de dialogue surgit, nous invitant à préciser le nom du fichier de règles. Nous introduisons BBPRUNE.KBA puis nous validons.

Une nouvelle fenêtre contenant la base réduite apparaît alors. Il n'y a pas plus que 4 règles, elles sont nettement plus concises.



Remarque : Notons que la méthode C4.5 RULES ne préserve pas les propriétés logiques de la base initiale. Il se peut qu'un individu soit couvert par 2 règles, il se peut aussi qu'un individu ne déclenche aucune règle. Quinlan (1993) propose des stratégies simples pour remédier à ces inconvénients.

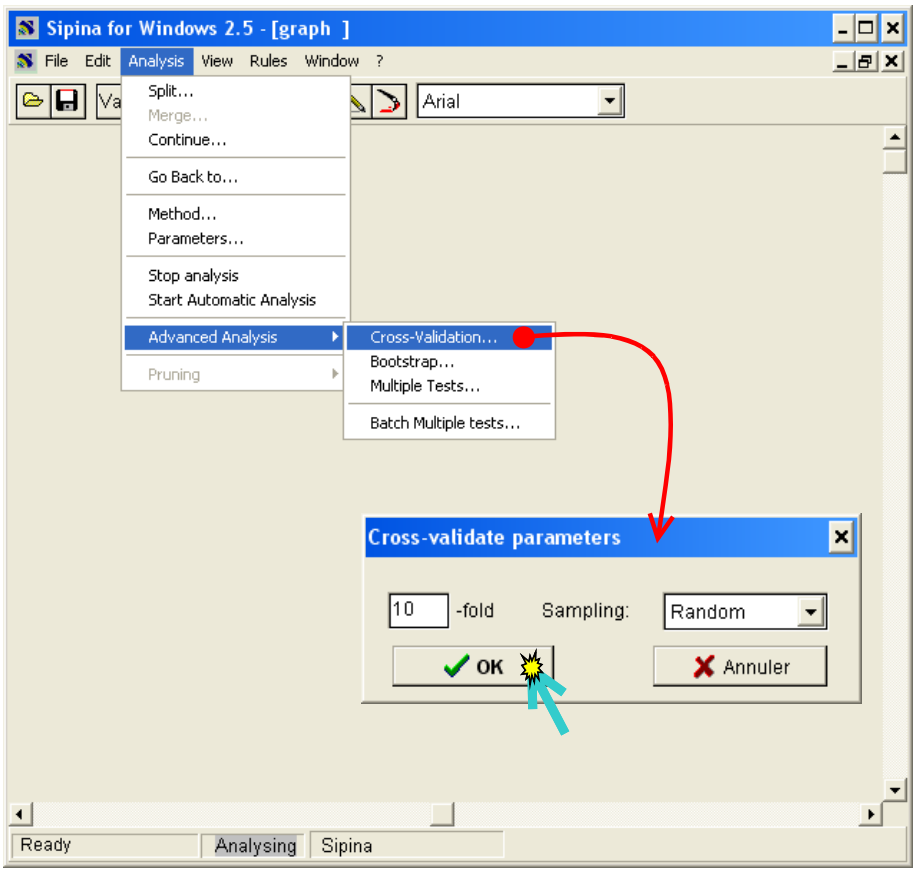
7.3 Mesurer les performances en validation croisée

Nous disposons d'un modèle de prédiction. Il est important d'en mesurer les performances en classement. Le taux de succès est un indicateur naturel pour cela. Il quantifie la probabilité de classer correctement un nouvel individu lorsque nous lui appliquons le graphe. Pour ne pas être biaisé, il doit être calculé soit sur un second échantillon qui n'a pas participé à l'élaboration du classifieur, lorsque nous disposons de suffisamment d'observations ; soit par ré échantillonnage, la validation croisée est certainement une des techniques les plus répandues[10].

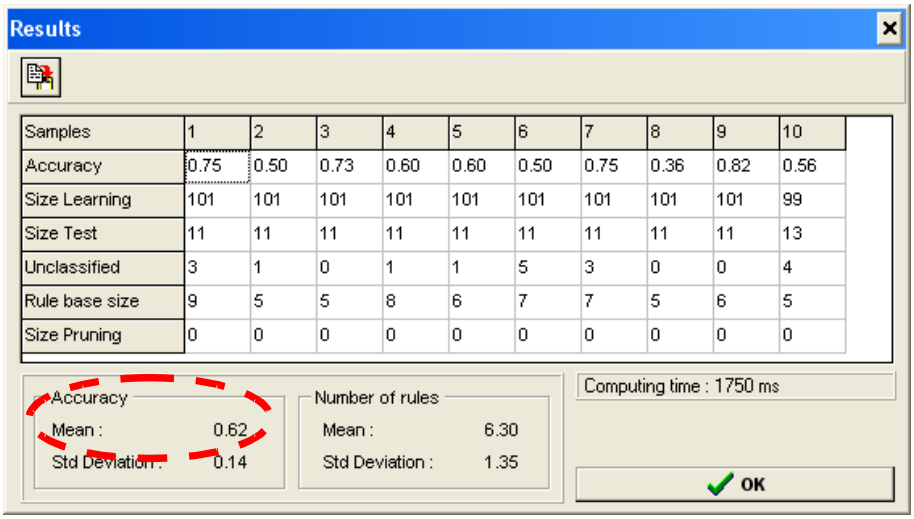
Notre effectif étant assez faible (n = 112 observations), il n'est pas question d'en sacrifier une partie pour l'évaluation, la validation croisée paraît plus indiquée. Nous interrompons l'analyse en cours en actionnant le menu ANALYS / STOP ANALYSIS.

[10] Voir http://eric.univ-lyon2.fr/~ricco/cours/slides/resampling_evaluation.pdf pour plus de détails.

Puis nous activons le menu ANALYSIS / ADVANCED ANALYSIS / CROSS-VALIDATION. Une boîte de paramétrage apparaît. Nous choisissons FOLD = 10 c.-à-d. la séquence apprentissage-test sera répété 10 fois, avec une subdivision 9/10 – 1/10. Nous validons.

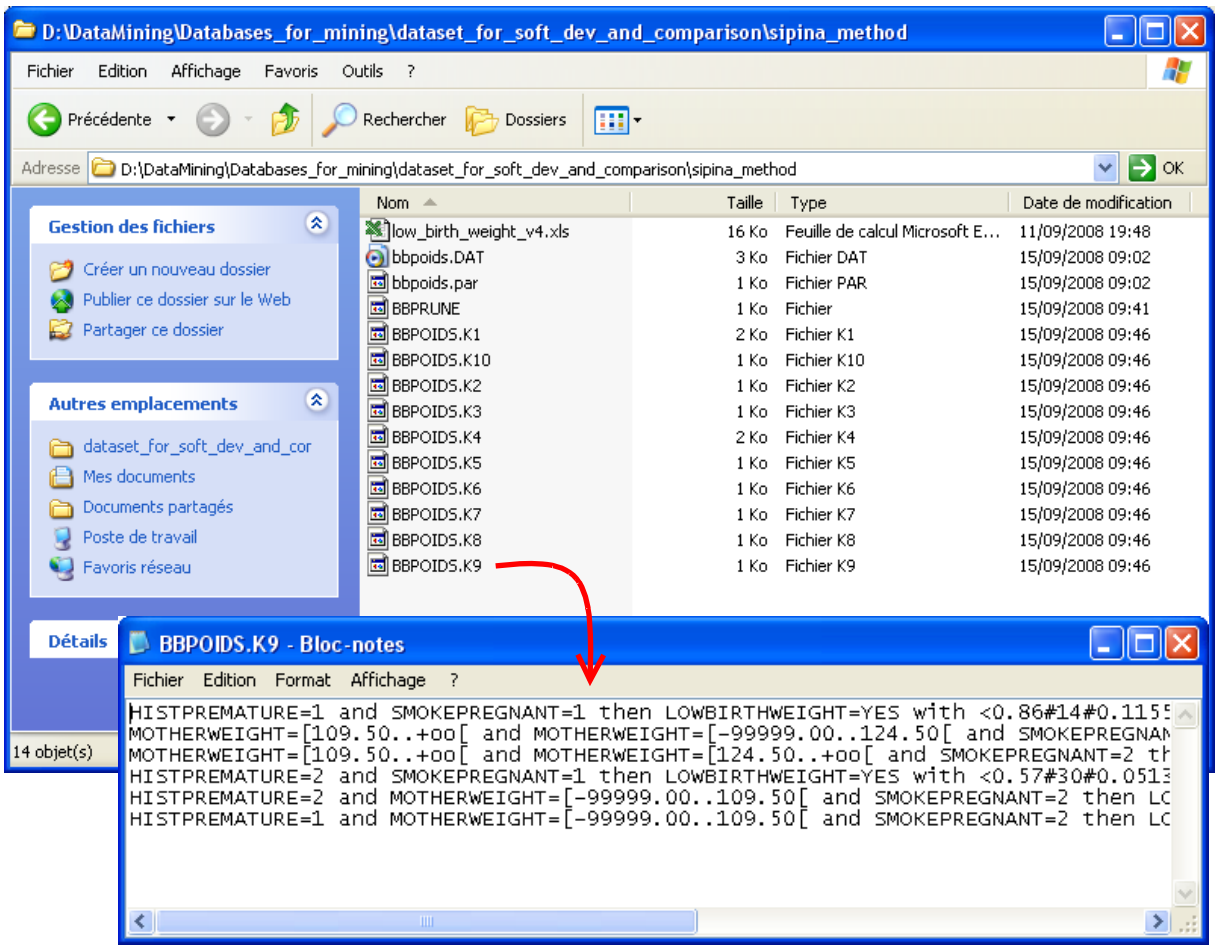


Les modèles intermédiaires s'affichent successivement. Une fenêtre récapitulative nous indique que la performance espérée de la méthode SIPINA sur ces données est 62%.



Les graphes produisent en moyenne 6 règles. Il n'est pas possible d'accéder individuellement aux graphes calculés lors que la validation croisée. En revanche, les bases de règles associées sont conservées automatiquement dans le répertoire des

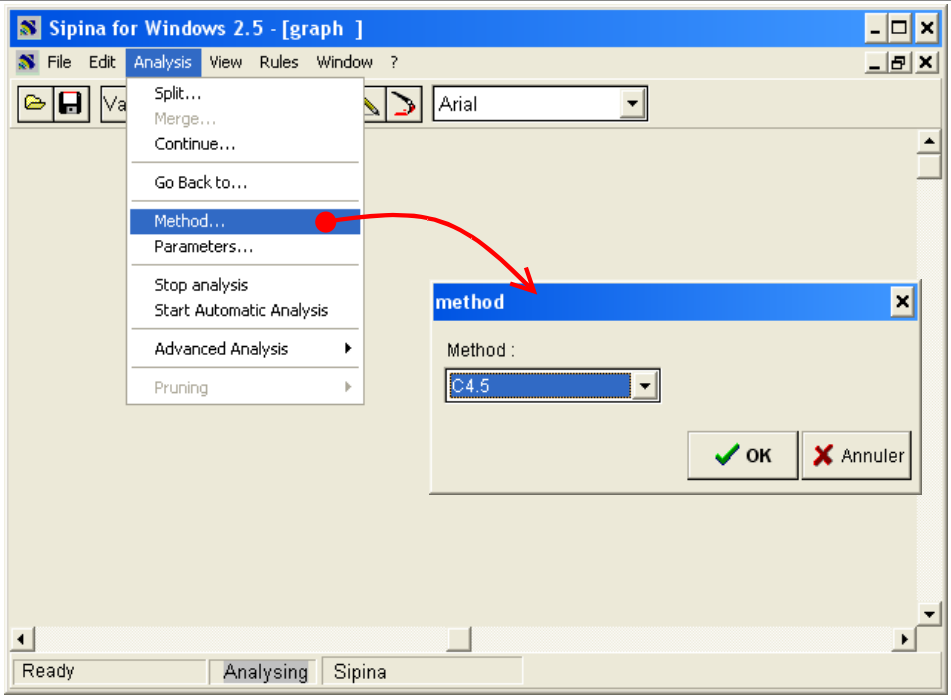
données. Ainsi, si l'en s'intéresse au graphe n°9 qui a été le plus performant, nous ouvrons dans le bloc-notes le fichier de règles « BBPOIDS.K9 » :



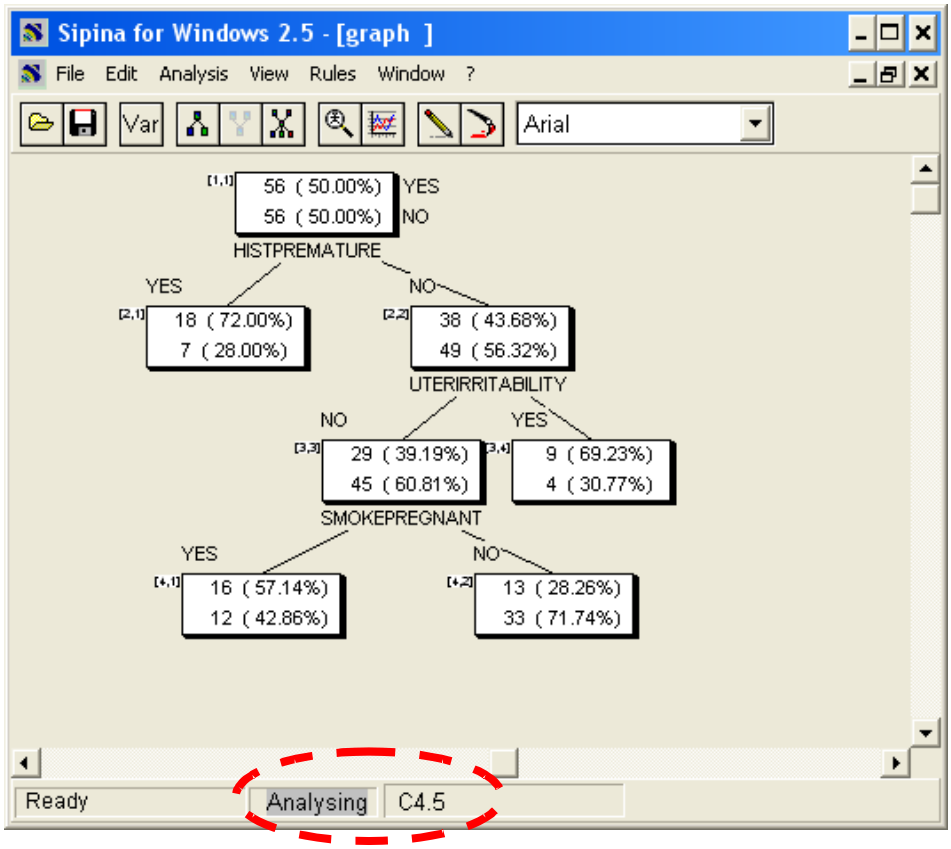
7.4 Utiliser un autre algorithme d'apprentissage

Tout comme la version Recherche, SIPINA v2.5 intègre d'autres techniques d'induction d'arbres. Elle propose, entre autres, la méthode C4.5 (Quinlan, 1993).

Pour modifier la méthode courante, nous activons le menu ANALYSIS / METHOD. Dans la boîte qui surgit, nous sélectionnons **C4.5**. Nous validons.



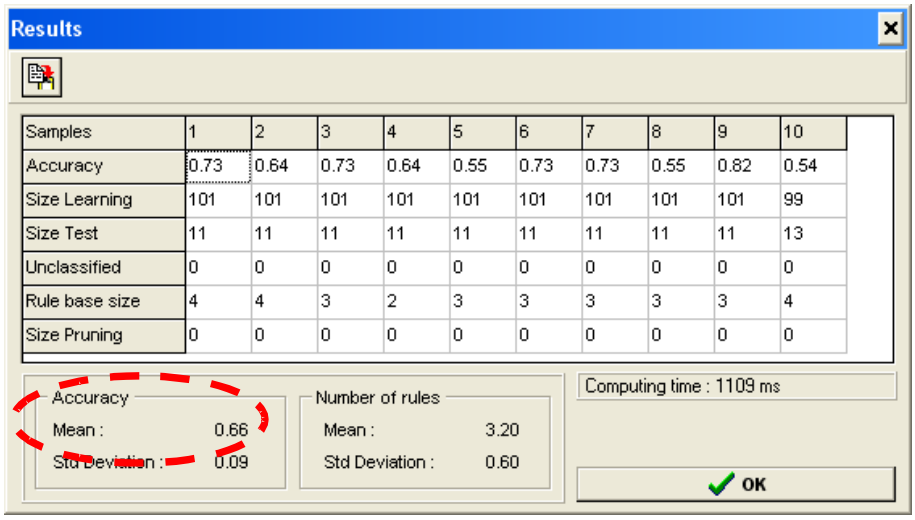
La méthode utilisée s'affiche dans la barre d'état. Pour lancer la construction de l'arbre, nous actionnons le menu ANALYSIS / START AUTOMATIC ANALYSIS.



La structure est relativement simple avec 4 feuilles. La base de règles comporte donc 4 règles, à comparer aux 8 règles produites par la méthode SIPINA précédemment.

Une question clé qui vient tout de suite est : « est-ce que cette simplicité se fait au détriment des performances en prédiction ? ». Pour le savoir, nous mesurons le taux de succès en validation croisée.

Nous stoppons l'analyse en cours en cliquant sur ANALYSIS / STOP ANALYSIS. Puis, nous activons le menu ANALYSIS / ADVANCED ANALYSIS / CROSS-VALIDATION. Nous conservons FOLD = 10. Le résultat montre une précision espérée de 66%.



Nous obtenons un classifieur plus précis avec moins de règles.

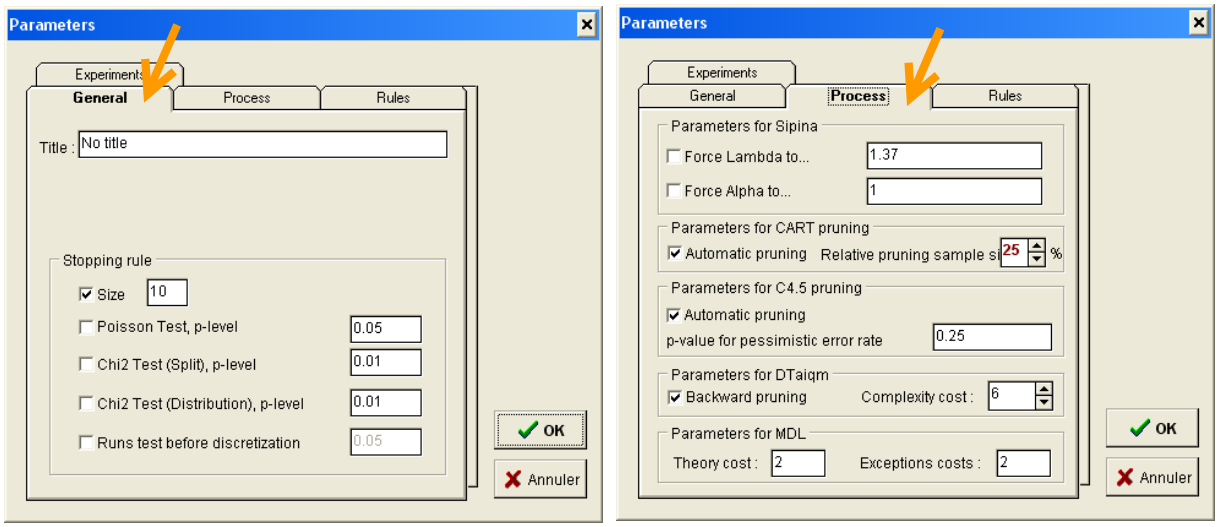
Bien entendu, ce type d'expérimentation n'a certainement pas valeur de preuve quant aux qualités générales des méthodes. L'idée est surtout de montrer comment comparer la méthode SIPINA avec une autre sur un fichier de données, en nous appuyant exclusivement sur un critère de performance.

8 Conclusion

La version 2.5 est le seul logiciel, à ma connaissance, à proposer la méthode SIPINA. C'est pour cette raison qu'elle est encore diffusée sur notre site alors que son développement a été stoppé il y a bien longtemps maintenant. De manière générale, cette version intègre des méthodes introuvables par ailleurs (Wehenkel, 1993 ; Quinlan et Rivest, 1989 ; etc.).

Concernant la méthode SIPINA, la supériorité théorique des graphes sur les arbres est attestée (Rakotomalala, 1997 ; chapitre 7). Mais elle ne se traduit pas en de meilleures performances dans les applications. Les innombrables expérimentations sur données réelles montrent que arbres et graphes présentent des performances similaires. Il semble que les graphes soient décisifs dans des configurations bien particulières, lors du traitement des bases de petite taille par exemple. Sa capacité à lutter contre la fragmentation des données est alors un atout. Dans le cas contraire, lors du traitement de grosses bases, fortement bruitées de surcroît, mieux vaut partir sur des techniques d'arbres telles que CART.

Enfin, à l'instar de toutes les méthodes d'apprentissage, le comportement de SIPINA peut être affiné à l'aide de paramètres accessibles via le menu ANALYSIS / PARAMETERS. Ils permettent d'orienter l'exploration des solutions en fonction des objectifs de l'étude et des caractéristiques des données (Voir Zighed et Rakotomalala, 2000 ; section 3.4). Il est même possible de fixer les paramètres de manière à ce que la méthode ne produise que des arbres.



9 Références

- D. Zighed, J.P. Auray, G. Duru, *SIPINA : Méthode et logiciel*, Lacassagne, 1992.
- J. Oliver, *Decision Graphs : An extension of Decision Trees*, in Proc. of Int. Conf. on Artificial Intelligence and Statistics, 1993.
- R. Quinlan, *C4.5 : Programs for Machine Learning*, Morgan Kaufmann, 1993.
- R. Rakotomalala, *Graphes d'induction*, Thèse de Doctorat, Université Lyon 1, 1997 (URL : http://eric.univ-lyon2.fr/~ricco/publications.html).
- D. Zighed, R. Rakotomalala, *Graphes d'induction : Apprentissage et Data Mining*, Hermès, 2000.
- M. Tenenhaus, *Statistique – Méthodes pour décrire, expliquer et prévoir*, Dunod, 2007 ; pages 540 à 545.