1. Objectif

Installation de l'add-on SIPINA pour Open Office et Libre Office Calc.

Je ne me lasserai jamais de le dire, la connexion entre les logiciels de data mining et un tableur est primordial pour la popularité des premiers. Lorsqu'il s'agit de manipuler des bases de taille « raisonnable », avec plusieurs milliers d'observations et quelques dizaines de variables, le tableur est très pratique pour la gestion et le prétraitement des données (transformation, recodage, etc.). A l'issue de l'analyse, il constitue également un outil privilégié pour la mise en forme des résultats. Il n'est pas surprenant dès lors que des éditeurs de logiciels proposent des solutions de couplage fort sous forme de macro complémentaire pour Excel (ex. XLMiner). Particulièrement édifiant, des éditeurs tels que SAS s'y sont mis également[1]. Notons enfin que Microsoft propose son propre add-in pour Excel basé sur le moteur « SQL Server Analysis Services »[2].

Tout ça est très bien. On notera simplement que si les solutions commerciales sont assez répandues pour Excel, les équivalents gratuits sont plutôt rares. Il y a bien sûr SIPINA et TANAGRA dont la macro complémentaire date de 2006 ; il y a RExcel qui permet d'établir connexion entre Excel et R[3] ; à force de chercher sur le net, j'ai réussi à en dénicher d'autres : XL-Statistics ; XL Toolbox ; etc.

Mais Excel lui-même n'est pas gratuit. Heureusement, il existe des alternatives crédibles avec le tableur des suites bureautiques gratuites Open Office et Libre Office. Véritable signe des temps, je constate qu'une bonne partie de mes étudiants préfèrent utiliser ces logiciels plutôt que de s'embarquer dans des copies plus ou moins piratées de la suite MS Office. Ce qui constitue une véritable avancée. D'où la question suivante : existe-t-il des add-on dédiés au calcul statistique qui s'intégreraient dans le tableur libre Calc ? Après quelques recherches, j'ai découvert, entres autres, quelques produits intéressants tels que Statistical Data Analyser for OOCalc, R4Calc. Nous les étudierons de manière approfondie dans un prochain tutoriel.

En ce qui nous concerne, l'add-on Tanagra pour Calc existe depuis 2006[4]. En revanche, je n'ai jamais pris le temps de transposer l'idée à SIPINA alors que, par ailleurs, la macro-

[1] http://support.sas.com/documentation/onlinedoc/addin/index.html ; en lisant les tutoriels, on se rend compte que la très grande majorité des outils adoptent un mode opératoire similaire : http://support.sas.com/documentation/cdl/en/amogs/64464/PDF/default/amogs.pdf

[2] Voir les tutoriels sur ce site : http://www.slideshare.net/dataminingtools/excel-datamining-addin-beginner

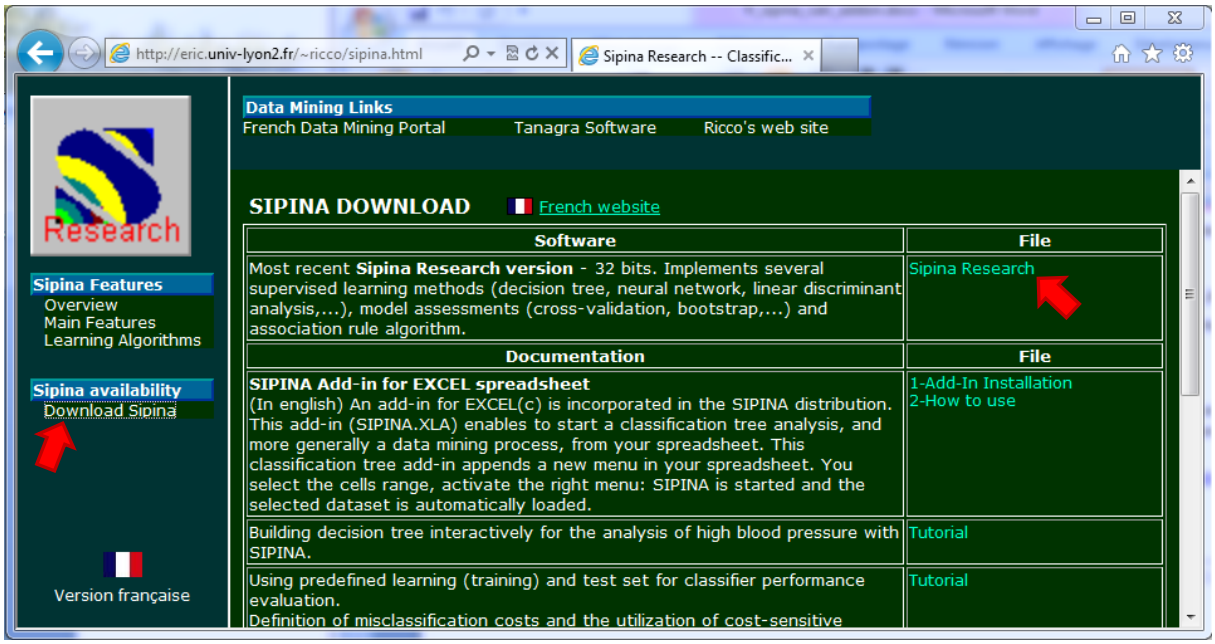
[3] http://tutoriels-data-mining.blogspot.fr/2011/12/connexion-entre-r-et-excel-via-rexcel.html

[4] http://chirouble.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_OOoCalc_Addon.pdf

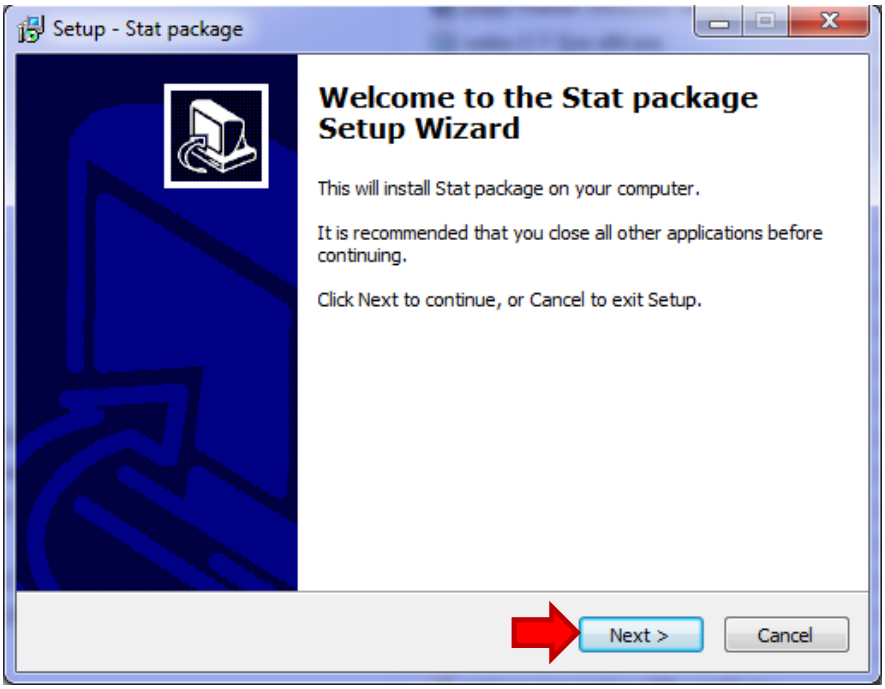
complémentaire « sipina.xla » pour Excel existe depuis plusieurs années. Cet oubli est réparé avec la version 3.9 de SIPINA (du 22 mars 2012). Nous montrons dans ce tutoriel l'installation et la mise en œuvre de l'add-on pour **Open Office Calc 3.3.0**. La transposition à **Libre Office 3.5.1** est immédiate.

2. Installation de Sipina

Le setup du logiciel SIPINA est accessible en ligne http://eric.univ-lyon2.fr/~ricco/sipina.html

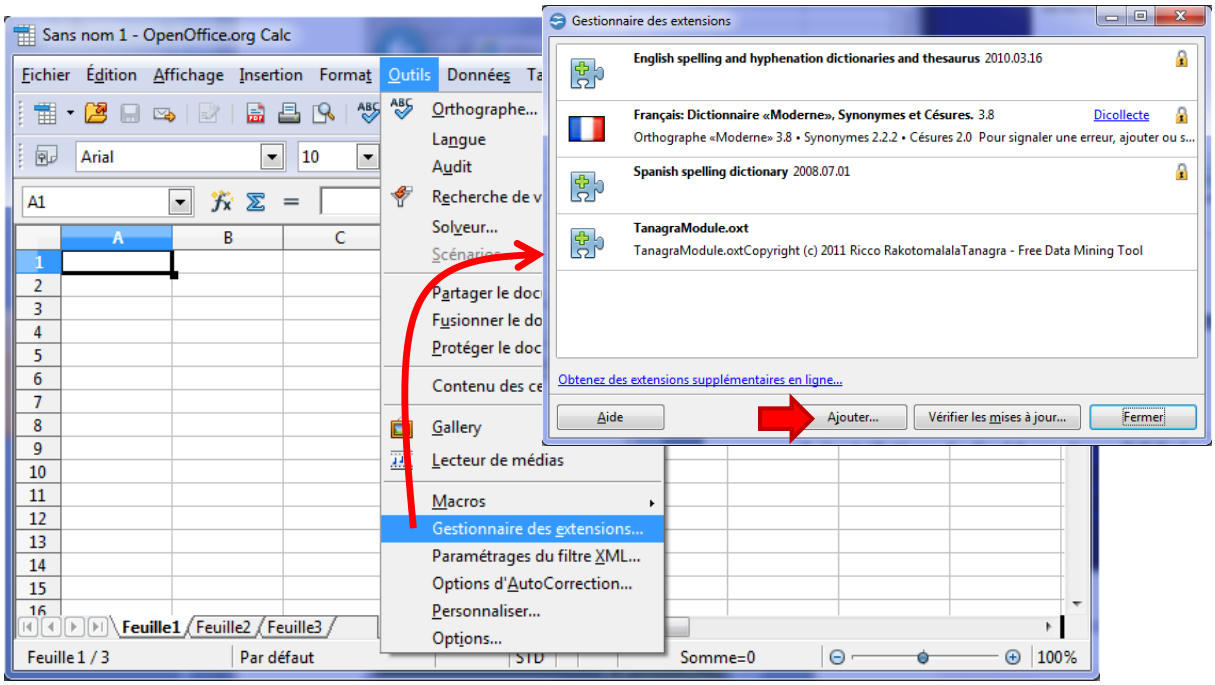


Après avoir téléchargé le fichier, nous démarrons l'installation. Elle est complètement standardisée. On peut se contenter de cliquer sur NEXT tout au long du processus.



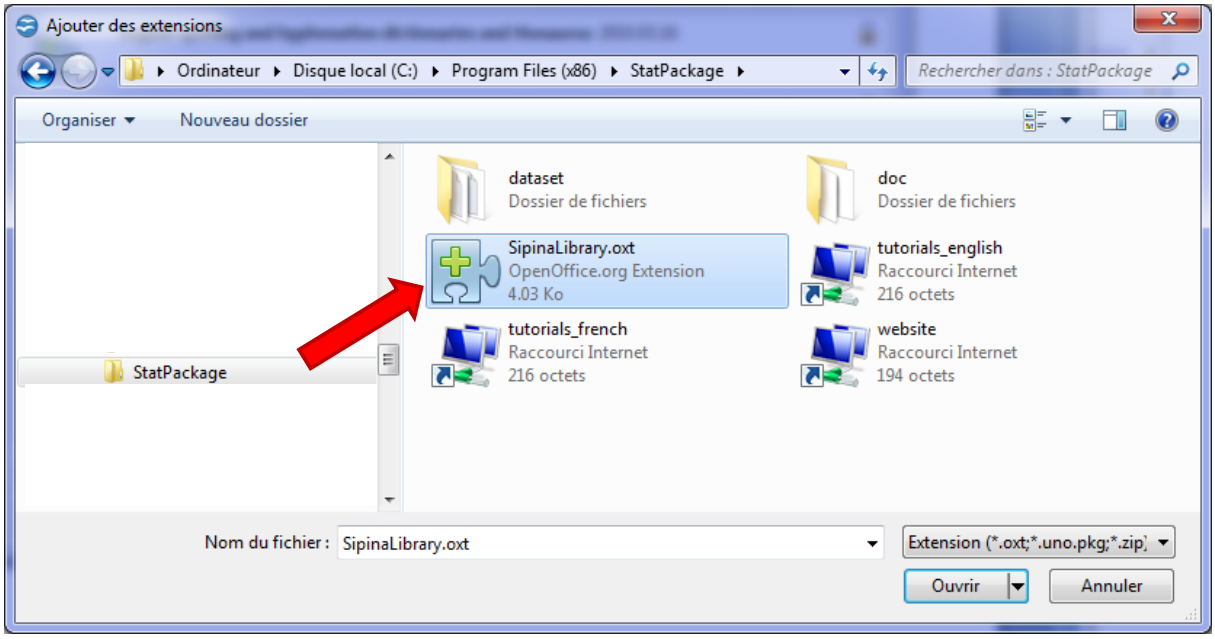
3. Installation de l'add-on dans OOCalc

SIPINA étant installé, nous passons à son intégration dans OOCalc. Nous lançons ce dernier. Dans le menu OUTILS, nous actionnons l'item GESTIONNAIRE DES EXTENSIONS.

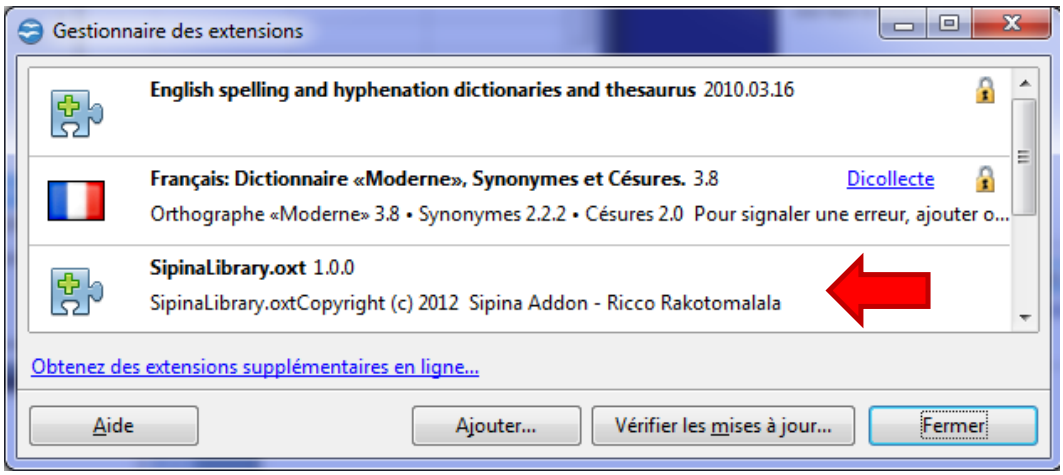


Nous ajoutons alors l'add-on SIPINA en cliquant sur le bouton AJOUTER et en allant chercher le fichier SIPINALIBRARY.OXT dans le répertoire d'installation de SIPINA. A priori,

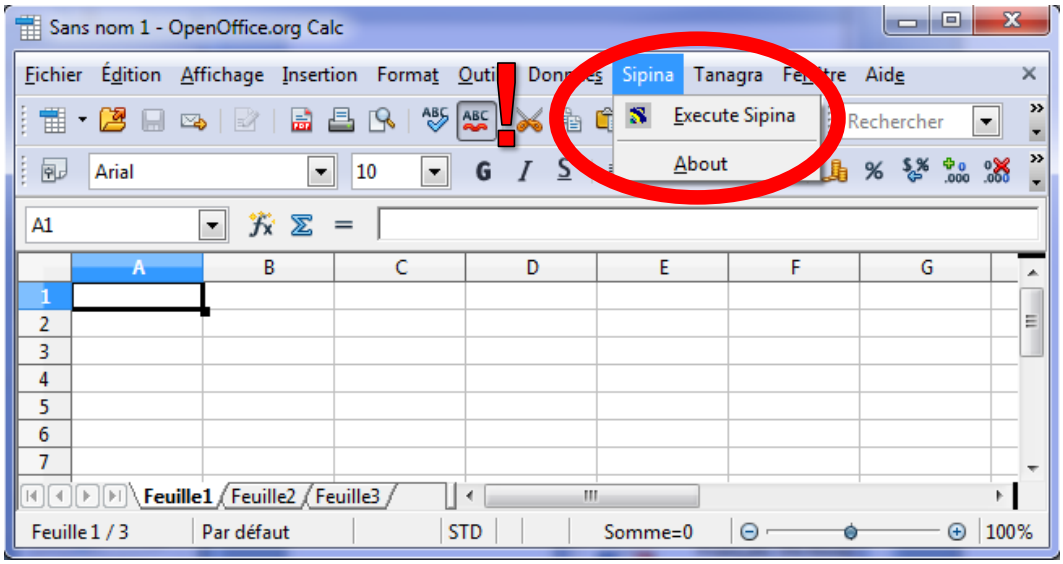
- « c:\Program Files\StatPackage » sous Windows 32 bits ;
- « c:\Program Files (x86)\StatPackage » sous Windows 64 bits).



L'add-on est installé.



Pour l'activer, **nous devons fermer OOCalc, puis le relancer de nouveau**. Le menu SIPINA est maintenant visible dans la barre de menus.



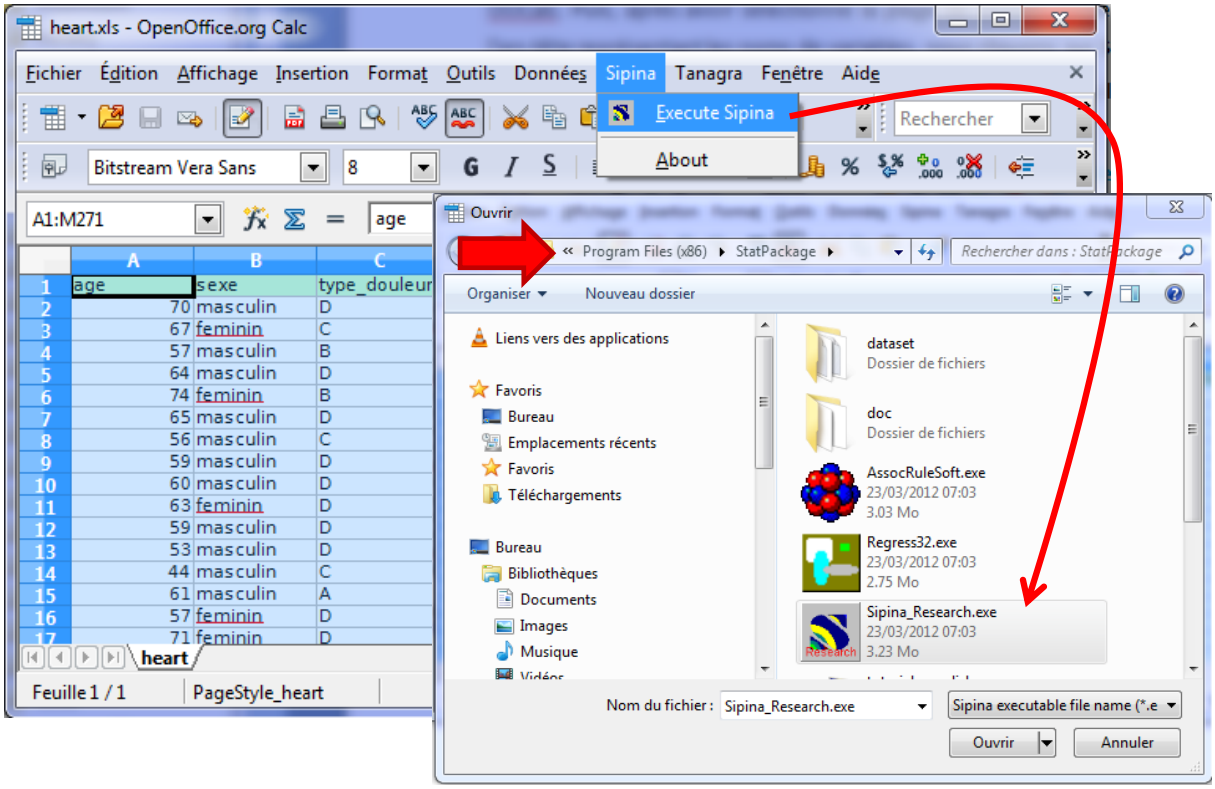
4. Utilisation de l'add-on

Pour utiliser l'add-on, nous devons tout d'abord charger le fichier à traiter (**heart.xls**[5]) dans OOCalc. Puis, après avoir sélectionné la plage de cellules contenant les données, y compris l'en-tête représentant les noms de variables, nous cliquons sur SIPINA / EXECUTE SIPINA.

Sur un système 32 bits, si nous avons installé le logiciel dans le répertoire usuel, l'add-on devrait automatiquement trouver l'exécutable et le démarrer.

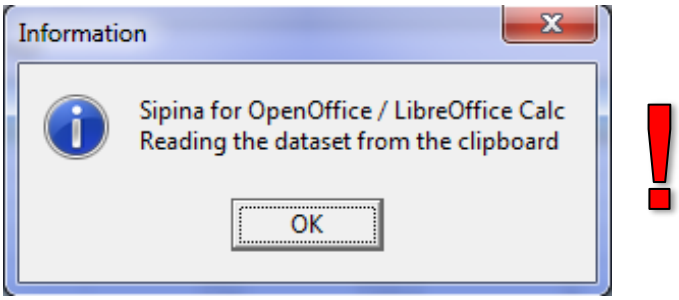
Sur un système 64 bits, ou si nous avons modifié le répertoire d'installation, une boîte de dialogue apparaît. Elle nous demande de spécifier le chemin et le nom de l'exécutable. Nous sélectionnons le fichier **SIPINA_RESEARCH.EXE**.

[5] OOCalc sait lire les fichiers Excel (XLS et XLSX). Mais il possède également un format natif « ODS ». Par rapport au XLSX, la taille des fichiers est un peu réduite et, surtout, OOCalc optimise les entrées / sorties.

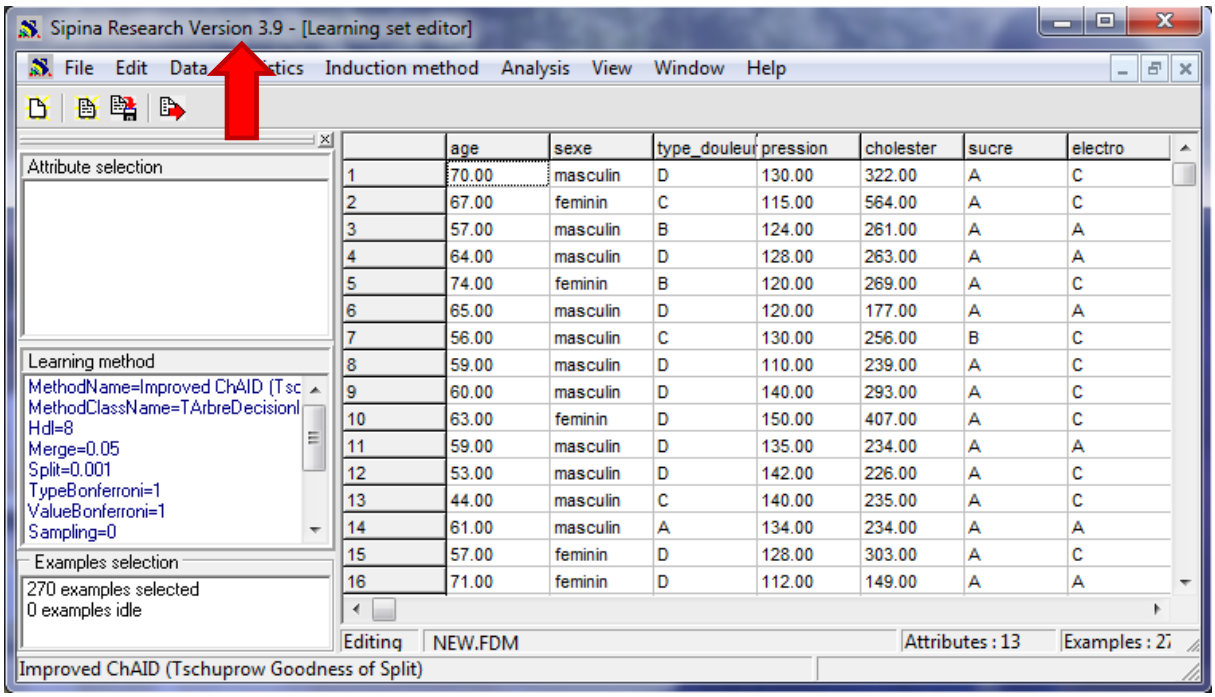


Remarque : *Nous aurions tout aussi bien sélectionné un autre programme du package SIPINA (REGRESS32 pour la régression, ASSOCRULESOFT pour les règles d'association).*

Une boîte de dialogue vient confirmer la bonne transmission des données.

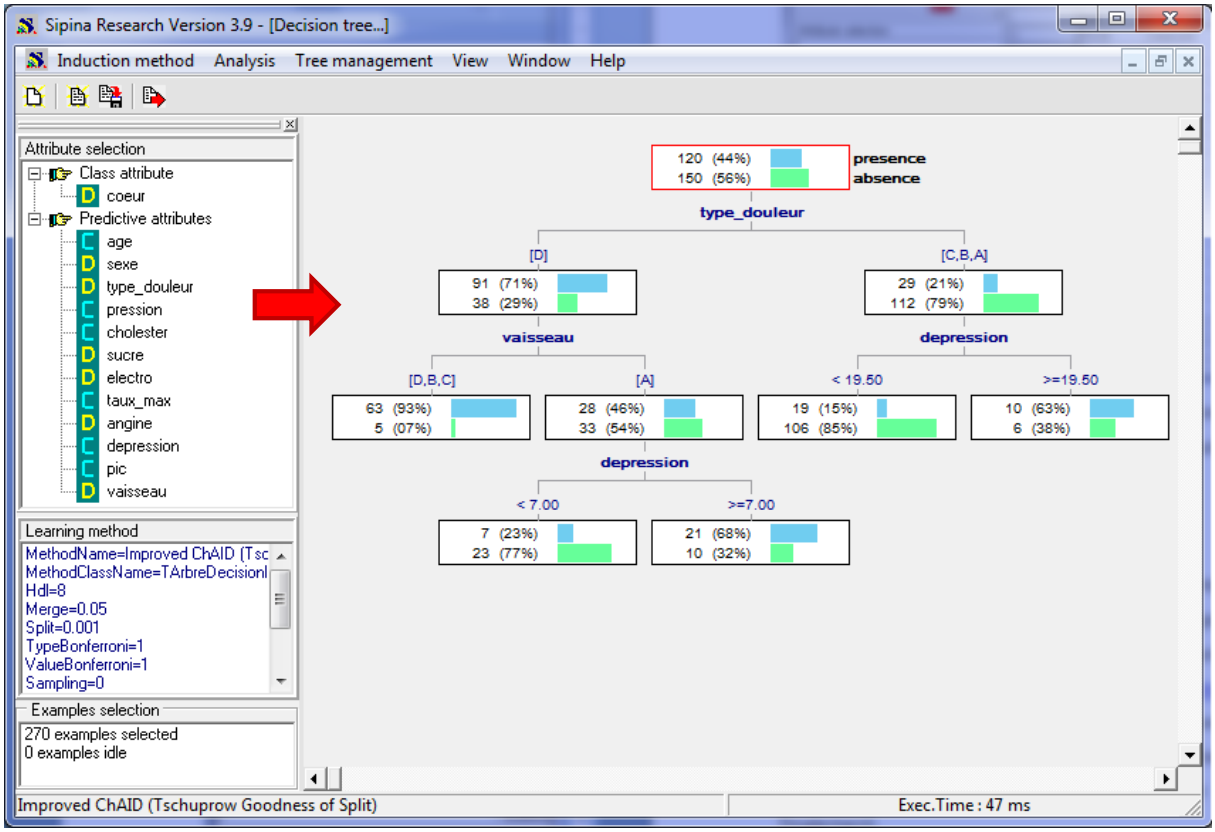


Au final, SIPINA apparaît. La base est visible dans la grille des données.



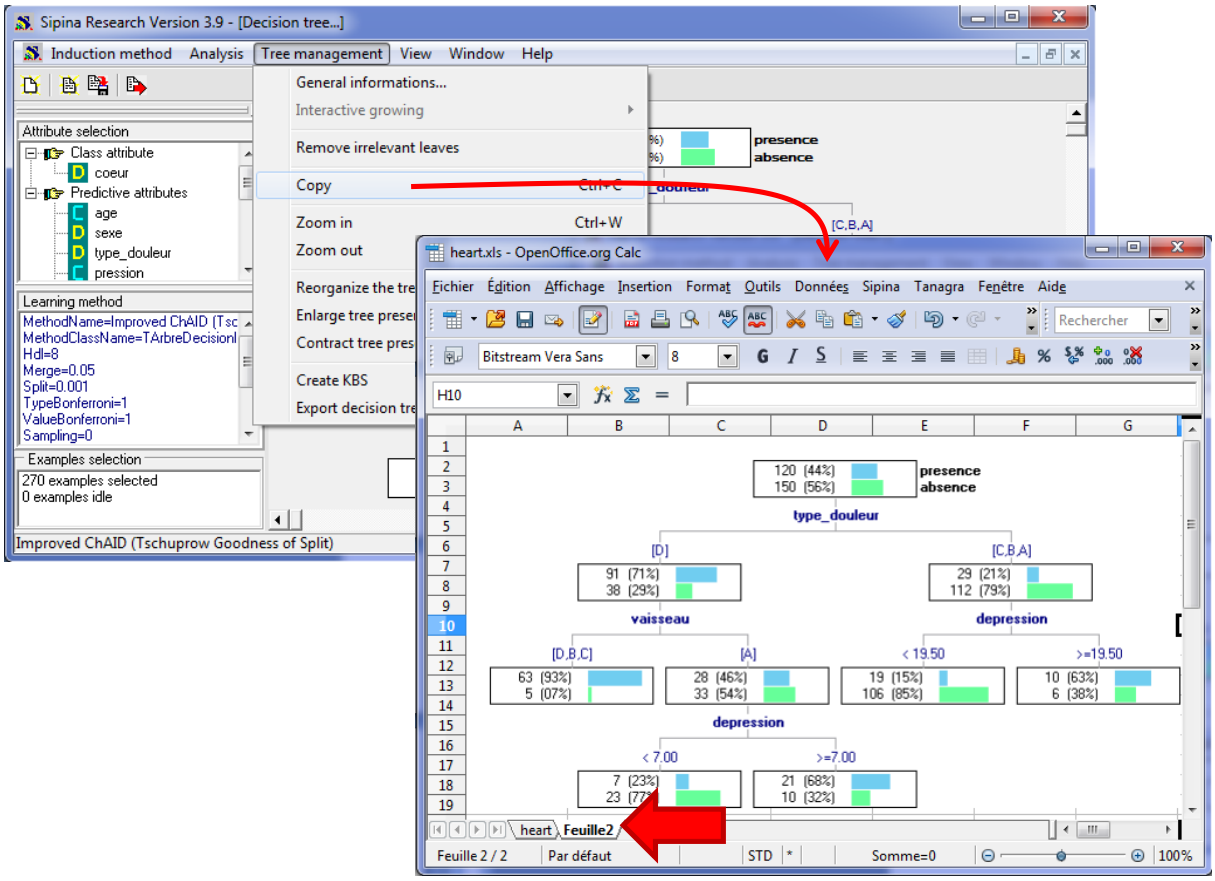
	age	sexe	type_douleur	pression	cholester	sucre	electro
1	70.00	masculin	D	130.00	322.00	A	C
2	67.00	feminin	C	115.00	564.00	A	C
3	57.00	masculin	B	124.00	261.00	A	A
4	64.00	masculin	D	128.00	263.00	A	A
5	74.00	feminin	B	120.00	269.00	A	C
6	65.00	masculin	D	120.00	177.00	A	A
7	56.00	masculin	C	130.00	256.00	B	C
8	59.00	masculin	D	110.00	239.00	A	C
9	60.00	masculin	D	140.00	293.00	A	C
10	63.00	feminin	D	150.00	407.00	A	C
11	59.00	masculin	D	135.00	234.00	A	A
12	53.00	masculin	D	142.00	226.00	A	C
13	44.00	masculin	C	140.00	235.00	A	C
14	61.00	masculin	A	134.00	234.00	A	A
15	57.00	feminin	D	128.00	303.00	A	C
16	71.00	feminin	D	112.00	149.00	A	A

L'utilisation de SIPINA est décrite dans de nombreux didacticiels[6]. Nous n'y reviendrons pas. Notons simplement que nous obtenons l'arbre suivant sur notre fichier de données.



[6] Cf. Les tutoriels décrivant SIPINA : http://tutoriels-data-mining.blogspot.fr/search/label/Sipina. Pour les non-spécialistes, celui-ci en particulier devrait vous mettre le pied à l'étrier : http://tutoriels-data-mining.blogspot.fr/2008/03/connexion-excel-sipina.html

Il est possible de copier l'arbre (menu TREE MANAGEMENT / COPY) et de le coller dans une nouvelle feuille du classeur (ou dans un traitement de texte, dans un diaporama, etc.). Cette fonctionnalité est très précieuse pour l'élaboration des rapports.



5. Conclusion

Moins connu qu'OOCalc, le tableur GNUMERIC représente une alternative intéressante. Il se démarque sur deux points : il s'agit d'un projet spécifique, non intégré dans une suite bureautique ; les fonctionnalités statistiques sont directement implémentées dans l'outil (http://projects.gnome.org/gnumeric/doc/chapter-stat-analysis.shtml) et non pas sous forme d'extensions. Pour ma part, je lui trouve plusieurs qualités, ne serait-ce que sa simplicité face aux mastodontes que représentent les tableurs des grandes suites bureautiques, et dont on n'exploite qu'une fraction très faible des fonctionnalités finalement.