

1 Objectif

Les Arbres de décision interactifs (IDT) avec le logiciel SPAD.

Dans le domaine du Data Mining, les logiciels libres et commerciaux ne s'adressent pas au même public. Ils ne répondent pas aux mêmes besoins. Les premiers sont plutôt destinés aux étudiants et aux chercheurs. Leur but est de mettre à leur disposition un grand nombre de méthodes, à des fins pédagogiques, ou à des fins d'expérimentation. L'utilisateur doit pouvoir monter simplement des comparaisons à grande échelle, pour comprendre le comportement des méthodes, pour évaluer leurs performances, etc. Le logiciel R (<http://www.r-project.org/>) en est certainement le meilleur représentant. Avec le système des packages, il est extensible à l'infini. Le dispositif est maintenant bien accepté, un grand nombre de chercheurs viennent enrichir la bibliothèque de calcul au fil du temps, signe que le mécanisme a été très bien conçu.

Les outils commerciaux s'adressent plutôt aux praticiens du Data Mining, y compris les chercheurs d'autres domaines. Leur objectif est de pouvoir mener à bien une étude intégrant le cycle complet de la fouille de données, partant de l'accès aux fichiers jusqu'au déploiement et la production de rapports. Dans ce cas, l'outil doit surtout leur faciliter le travail en prenant en charge, le plus simplement possible, un grand nombre de tâches répétitives et fastidieuses, comme l'accès aux données, leur préparation, la production de tableaux et graphiques pour les rapports, l'industrialisation des résultats, etc.

Bien entendu, la frontière n'est pas aussi tranchée. Bien d'outils issus du monde universitaire tentent de franchir le Rubicon en intégrant des fonctionnalités qui intéresseraient plutôt le monde industriel (ex. déploiement des modèles avec PMML - <http://www.dmg.org/>). A l'inverse, des logiciels commerciaux s'approprient les formidables bibliothèques de calculs que proposent les outils libres, notamment ceux de R (ex. [SAS / IML Studio](#), [SPSS PASW](#) ou [SPAD](http://www.spad.eu/) <http://www.spad.eu/>).

Dans ce didacticiel, nous montrons la mise en œuvre des Arbres de Décision Interactifs (IDT – Interactive Decision Tree) de **SPAD 7.0** sur un jeu de données constitué d'un classeur Excel décomposé en 3 feuilles : (1) on doit construire un arbre de décision à partir des données d'apprentissage ; (2) appliquer le modèle sur les données de la seconde feuille, nous adjoignons ainsi une nouvelle colonne « prédiction » aux données ; (3) vérifier la qualité de la prédiction en la confrontant à la vraie valeur de la variable cible située dans la troisième feuille du classeur.

Bien sûr, toutes ces opérations sont réalisables avec la grande majorité des logiciels libres. Un utilisateur un tant soit peu habile vous programme cela en trois coups de cuiller à pots sous R. Nous y reviendrons dans la section 4. L'intérêt ici est de montrer qu'un utilisateur novice, réfractaire à l'informatique, peut les enchaîner très facilement avec ce type d'outil, en prenant comme source de données un classeur Excel.

2 Données

Notre fichier « PIMA-ARBRE-SPAD.XLS¹ » est une version du fichier PIMA en provenance du serveur UCI². L'objectif est diagnostiquer la présence du diabète chez des patientes à partir de leurs caractéristiques (âge, pression sanguine, etc.).

	A	B	C	D	E	F
1	pregnant	plasma	bodymass	pedigree	age	diabete
2	0	138	36.3	0.933	25	positive
3	4	142	44	0.645	22	positive
4	3	142	32.4	0.2	63	negative
5	3	113	29.5	0.626	25	negative
6	5	88	27.6	0.258	37	negative
7	2	110	32.4	0.698	27	negative
8	2	129	28	0.284	27	negative
9	9	57	32.8	0.096	41	negative
10	1	79	43.5	0.678	23	negative
11	2	99	20.4	0.235	27	negative
12	5	158	39.4	0.395	29	positive
13	2	175	22.9	0.326	22	negative
14	14	100	36.6	0.412	46	positive
15	2	106	29	0.426	22	negative
16	1	97	27.2	1.095	22	negative
17	10	115	35.3	0.134	29	negative
18	0	104	33.6	0.51	22	positive
19	2	89	33.5	0.292	42	negative
20	3	106	30.9	0.292	24	negative

Notre classeur comporte 3 feuilles : (1) « **apprentissage** » correspond à l'échantillon de données étiquetées (568 observations), nous nous en servons pour construire le modèle de prédiction ; (2) « **à classer** » correspond aux observations pour lesquelles nous devons effectuer la prédiction, seules les descripteurs sont disponibles sur cette partie des données, la colonne « diabète » n'est pas renseignée, c'est la situation usuelle à laquelle nous sommes confrontés lorsque nous souhaitons déployer un modèle dans la population ; (3) enfin, « **étiquette** » ne contient que la colonne « diabète » des observations de la feuille précédente, cette partie des données n'est pas accessible dans la pratique usuelle du Data Mining, nous l'utilisons uniquement à des fins pédagogiques dans ce didacticiel.

3 Traitements sous SPAD

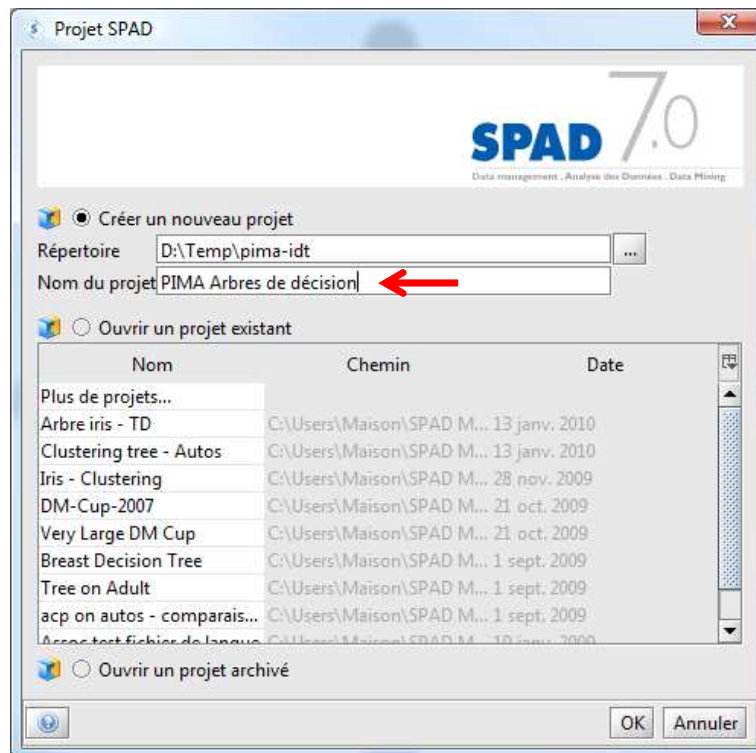
3.1 Création et archivage d'un modèle de prédiction

3.1.1 Création d'un projet

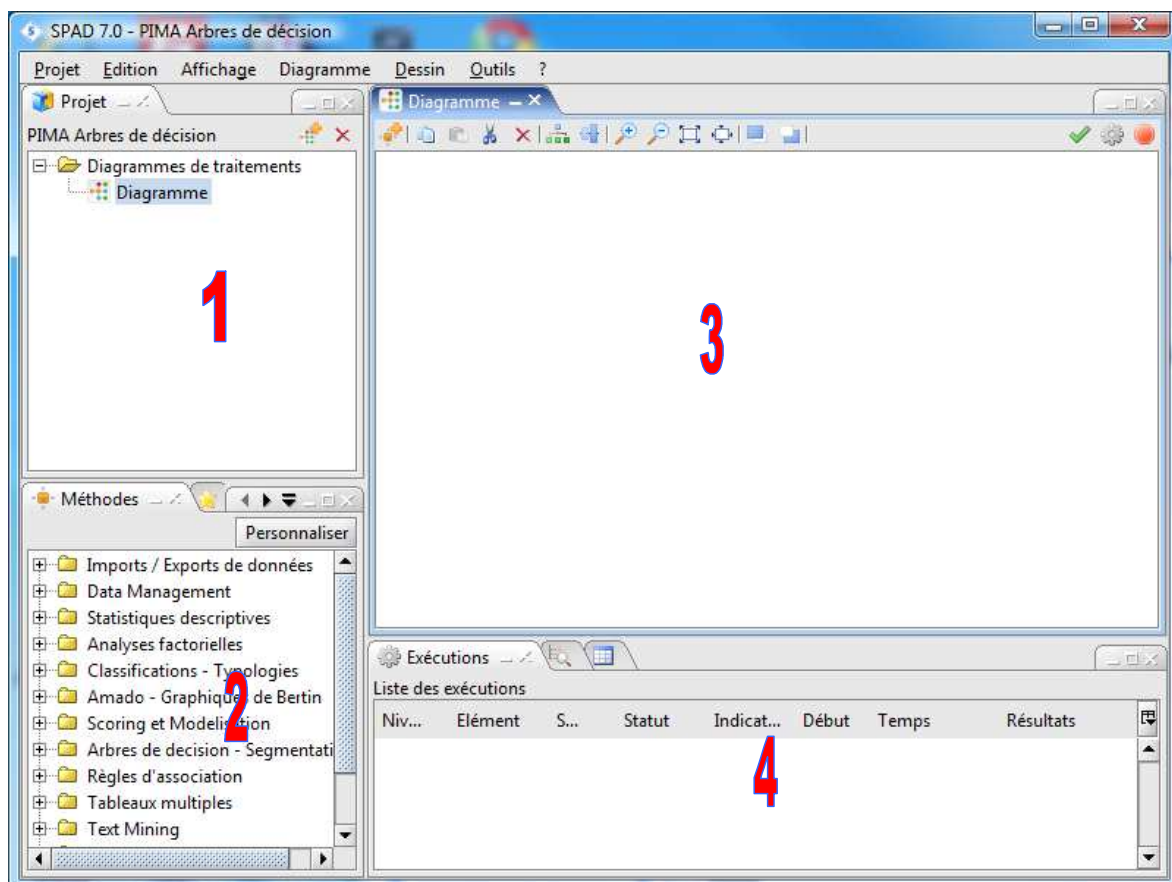
Au démarrage de SPAD, nous avons la possibilité d'ouvrir un projet existant ou d'en définir un nouveau. Nous créons le projet « PIMA Arbres de décision ».

¹ <http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/pima-arbre-spad.zip>

² <http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes>



Un nouveau diagramme de traitements est créé. L'interface de SPAD est conforme aux standards du domaine, la fenêtre principale est subdivisée en plusieurs parties :

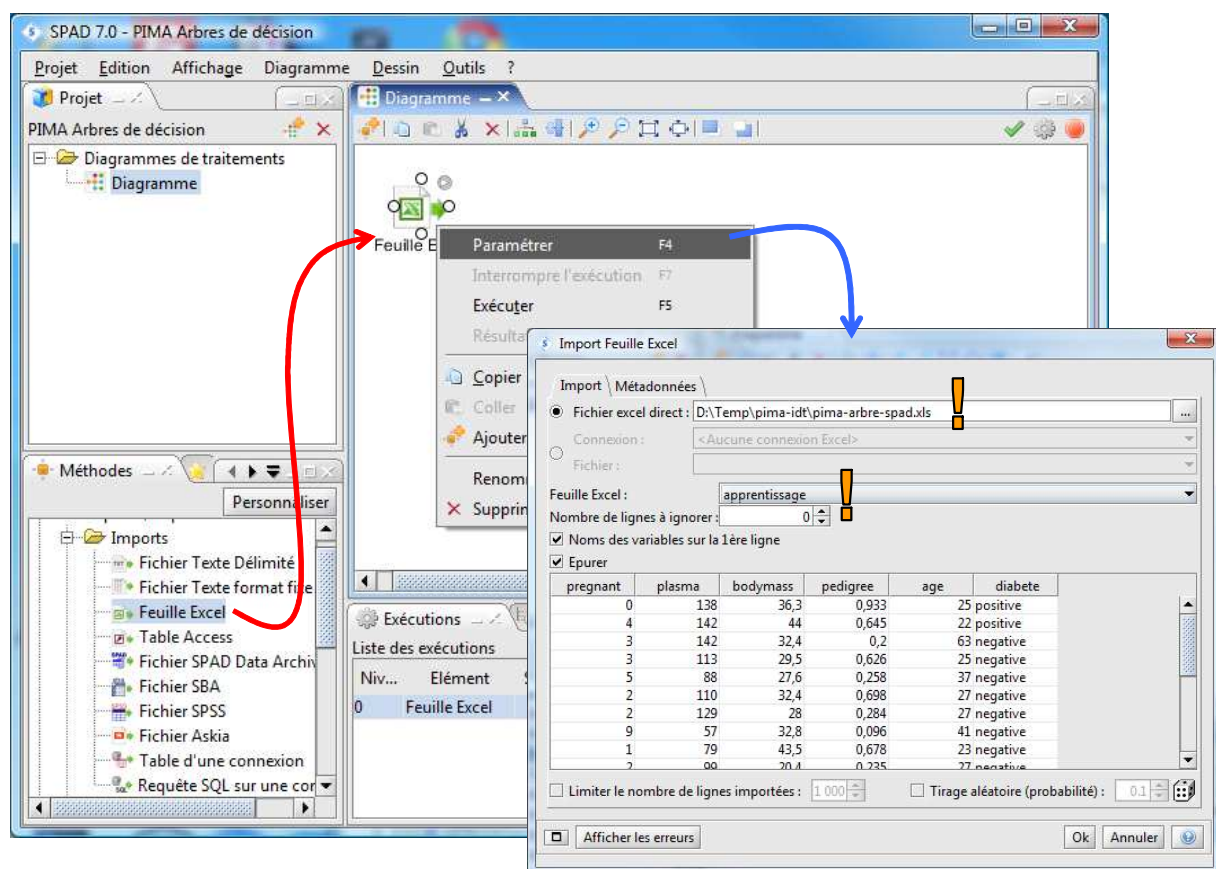


1. Le gestionnaire de projets.
2. La bibliothèque de méthodes, nous y puiserons les outils qui serviront au traitement des données.
3. L'espace de travail « Diagramme » dans lequel nous définirons l'enchaînement des traitements.
4. Et la fenêtre « Exécution » qui sert au suivi des traitements.

3.1.2 Importation des données

La première étape consiste à charger la feuille « apprentissage » de notre classeur Excel. Nous introduisons par glisser déposer le composant **Feuille Excel**. Nous actionnons le menu « Paramétrer (F4) » pour configurer le traitement.

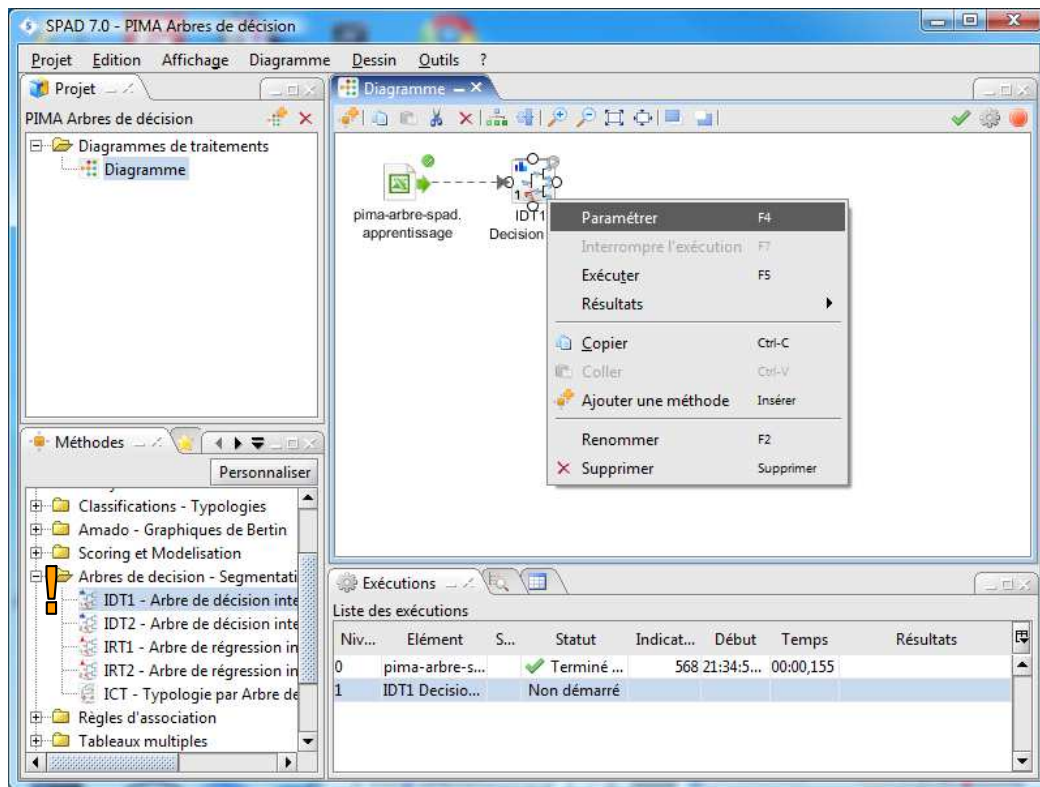
Dans la boîte de dialogue, nous devons surtout spécifier le nom du classeur et le nom de la feuille. Un aperçu des premières lignes est proposé à des fins de vérifications.



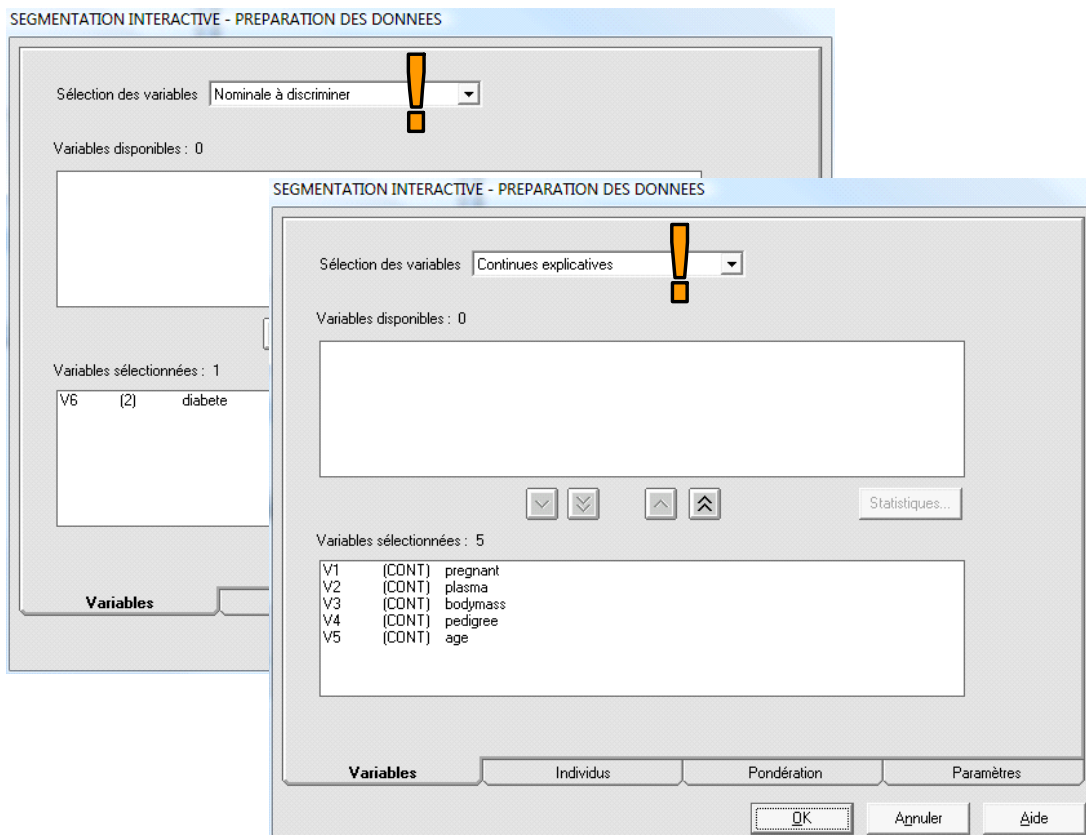
Il ne reste plus qu'à valider la configuration. Le fichier est automatiquement chargé. Les éventuelles erreurs sont signalées dans la fenêtre « Exécutions ». Dans notre cas, tout va bien.

3.1.3 Définition du problème à traiter

Nous devons maintenant indiquer au logiciel la variable cible (diabète) et les variables prédictives (les autres). Pour ce faire, nous introduisons le composant **IDT1**. Puis nous lui relions le composant d'accès aux données. Enfin, nous le paramétrons.



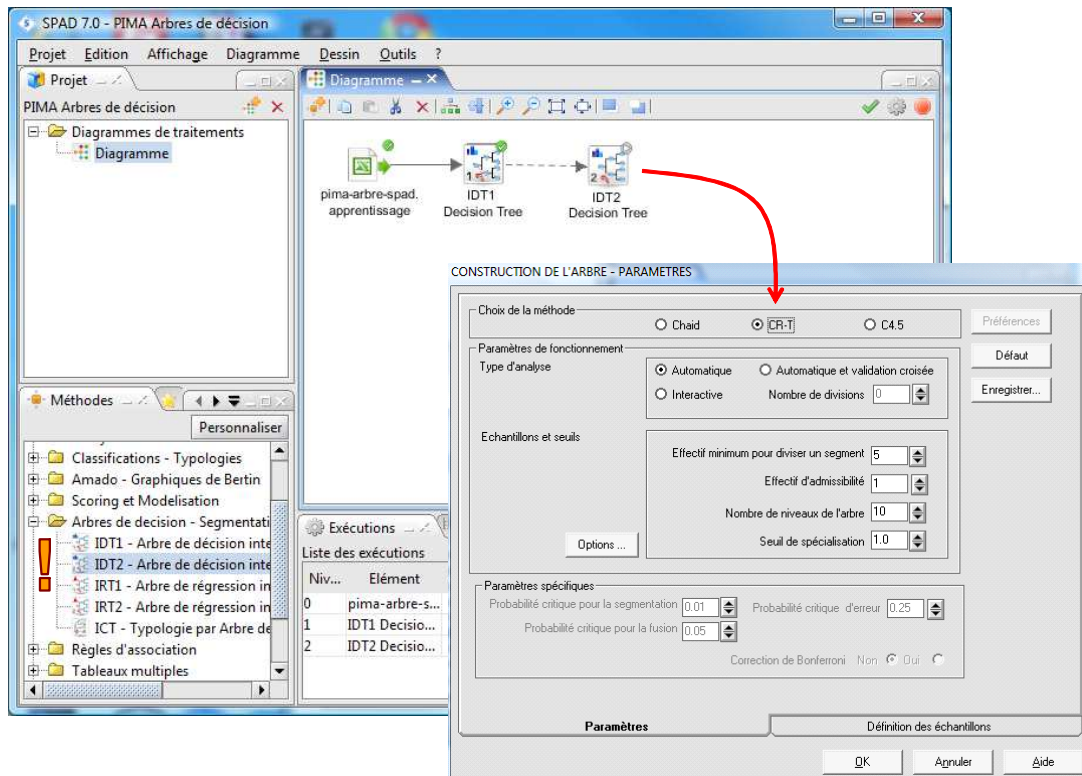
La nominale à discriminer est « diabète », les continues explicatives sont les autres.



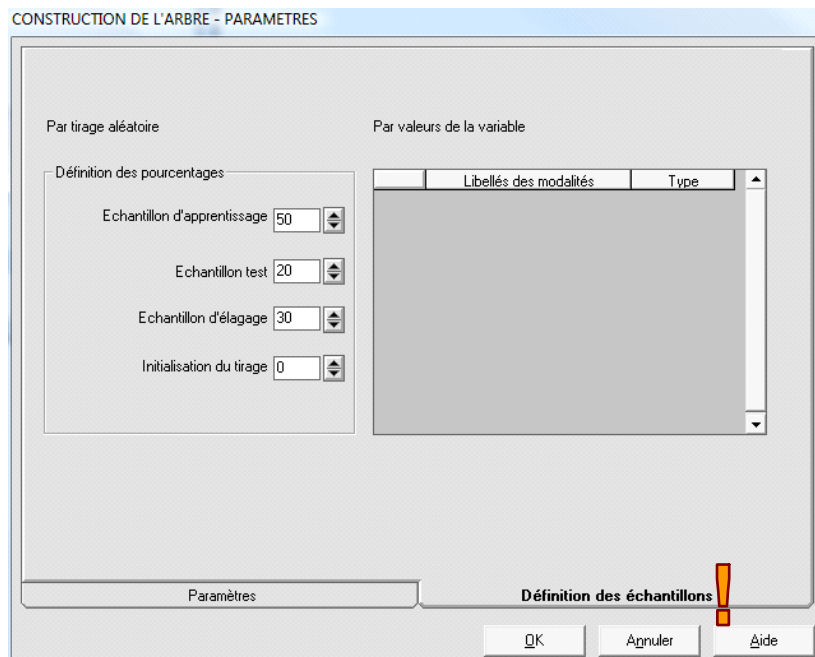
Remarque : Ce composant sert aussi à spécifier d'autres paramètres telles que la pondération des individus, le mode de gestion des données manquantes, etc.

3.1.4 Choix de la méthode d'apprentissage

Nous pouvons introduire le composant d'induction des arbres de décision **IDT2**. Nous lui relierons IDT1 puis nous le paramétrons. Nous choisissons la méthode C-RT de Brieman et al. (1984).



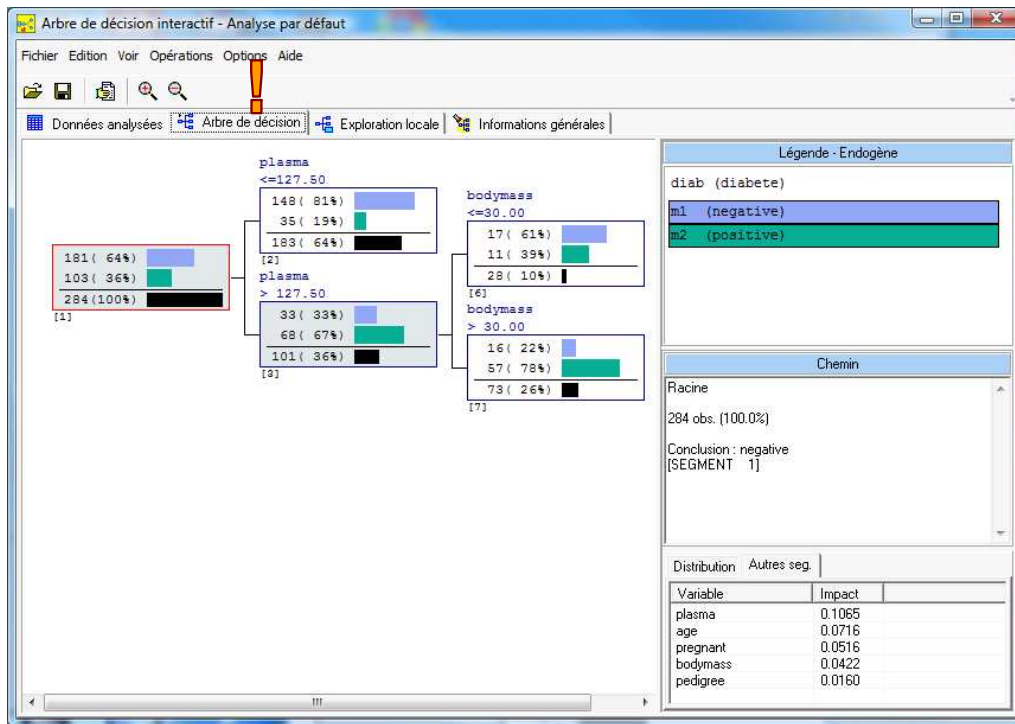
Le second onglet « Définition des échantillons » doit retenir notre attention.



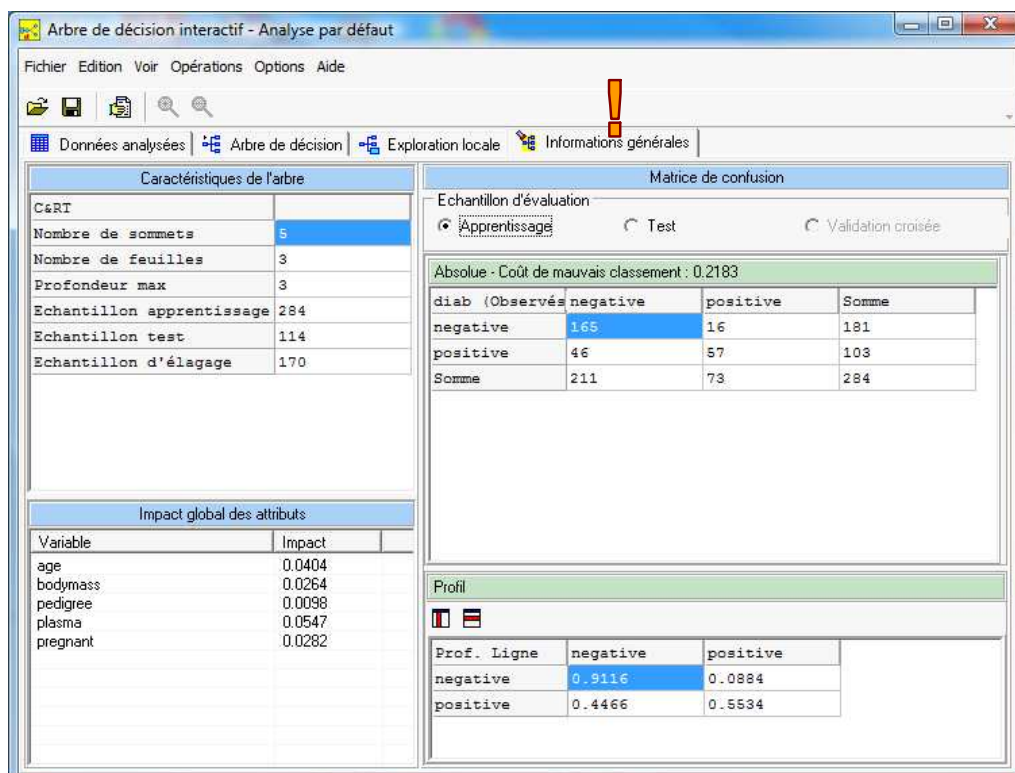
Y est spécifié la subdivision automatique des observations en échantillon d'apprentissage (50% des 568 observations), destiné à la phase d'expansion de l'arbre ; l'échantillon d'élagage (30%) réservé au post élagage ; l'échantillon de test (20%), qui sert à l'évaluation du modèle.

3.1.5 Etude globale de l'arbre

Après validation des paramètres, les calculs sont automatiquement démarrés. Pour visualiser l'arbre, nous actionnons le menu RESULTATS / INTERACTIVE DECISION TREE (d'autres manières sont également possibles, par exemple en cliquant sur l'icône adéquate dans la fenêtre « Exécutions »).

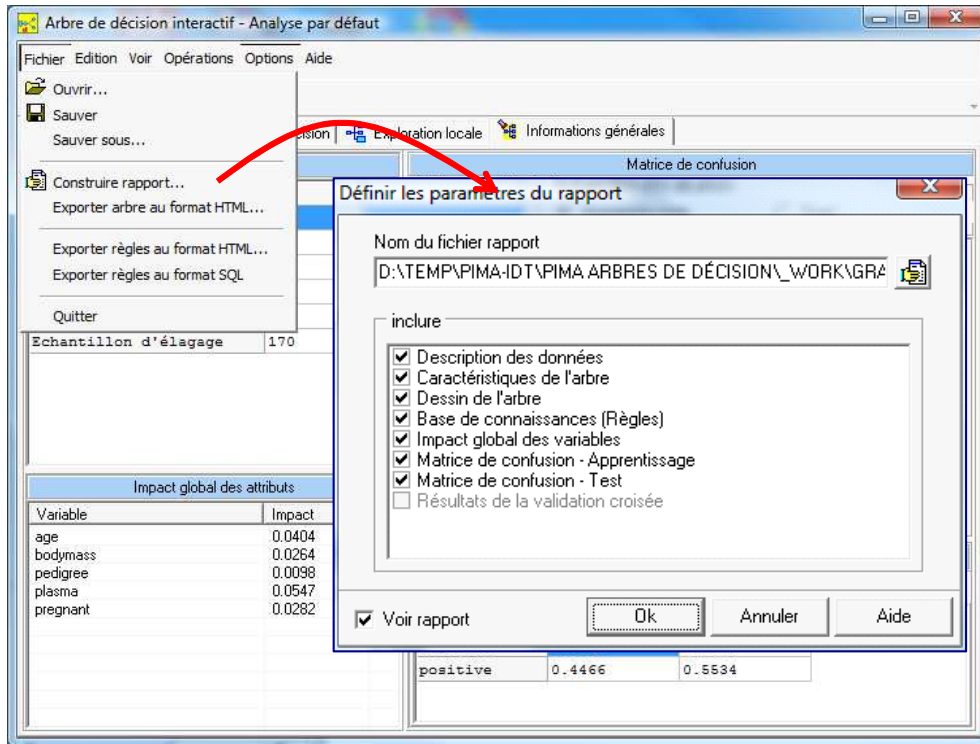


Dans l'onglet « Informations générales », nous observons la matrice de confusion calculée sur les échantillons d'apprentissage et de test.

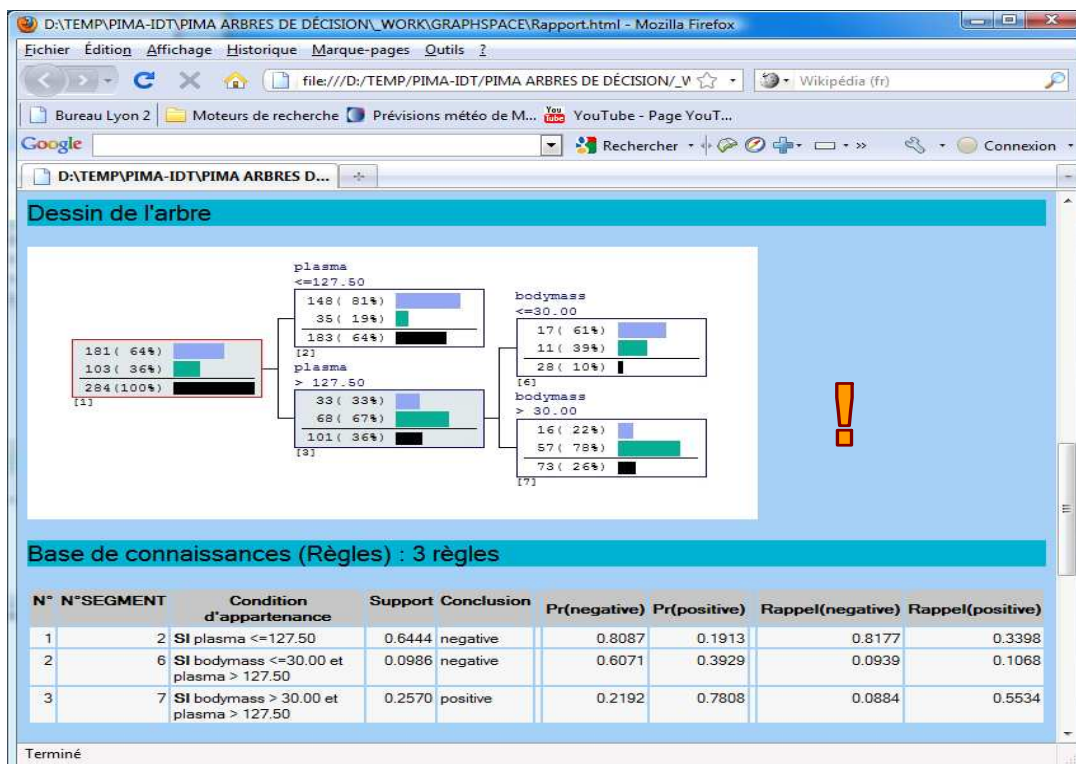


IDT se démarque réellement des outils libres à partir de ce stade...

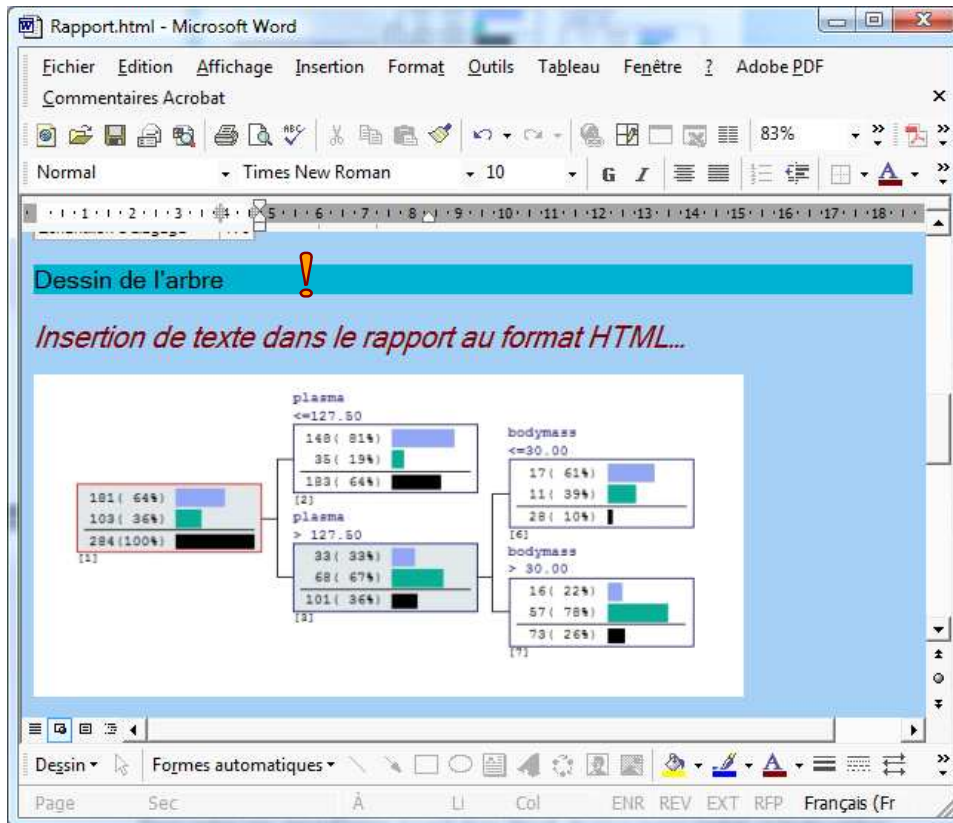
Rapport concernant l'analyse. Premier outil très important, il est possible de produire un rapport reprenant les principaux résultats. Il faut actionner le menu FICHIER / CONSTRUIRE RAPPORT. Une boîte de dialogue apparaît, elle nous permet d'en spécifier le contenu.



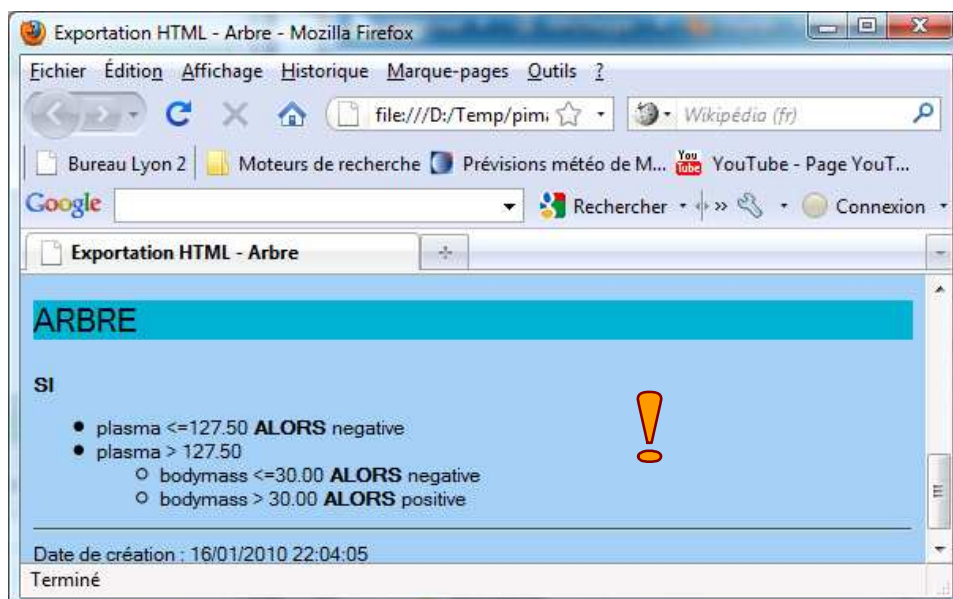
Nous laissons les choix par défaut et nous validons. Le navigateur de votre système est automatiquement démarré, un rapport au format HTML est affiché.



Il reprend la description des données, la description de la méthode, le dessin de l'arbre, la base de règle induite, les matrices de confusion, etc. Comme le rapport est au format HTML, nous pouvons l'ouvrir avec n'importe quel traitement de texte (Word, Open Office Writer) et l'éditer à notre convenance. Dans notre cas, nous l'avons ouvert dans Word, puis nous avons inséré un texte ad hoc.



Exportation de l'arbre au format HTML. D'autres rapports sont également disponibles. Par exemple, il est possible de produire une description textuelle, plus compacte, de l'arbre. Nous actionnons le menu FICHER / EXPORTER ARBRE AU FORMAT HTML. Après avoir spécifié le nom du fichier, nous disposons d'un document que l'on peut lire avec n'importe quel navigateur.



Exportation des règles au format HTML. Autre possibilité, puisque qu'un arbre peut être traduit en base de règles, nous pouvons aussi l'exporter au format HTML. Nous actionnons le menu FICHIER / EXPORTER REGLES AU FORMAT HTML, nous obtenons le document suivant.

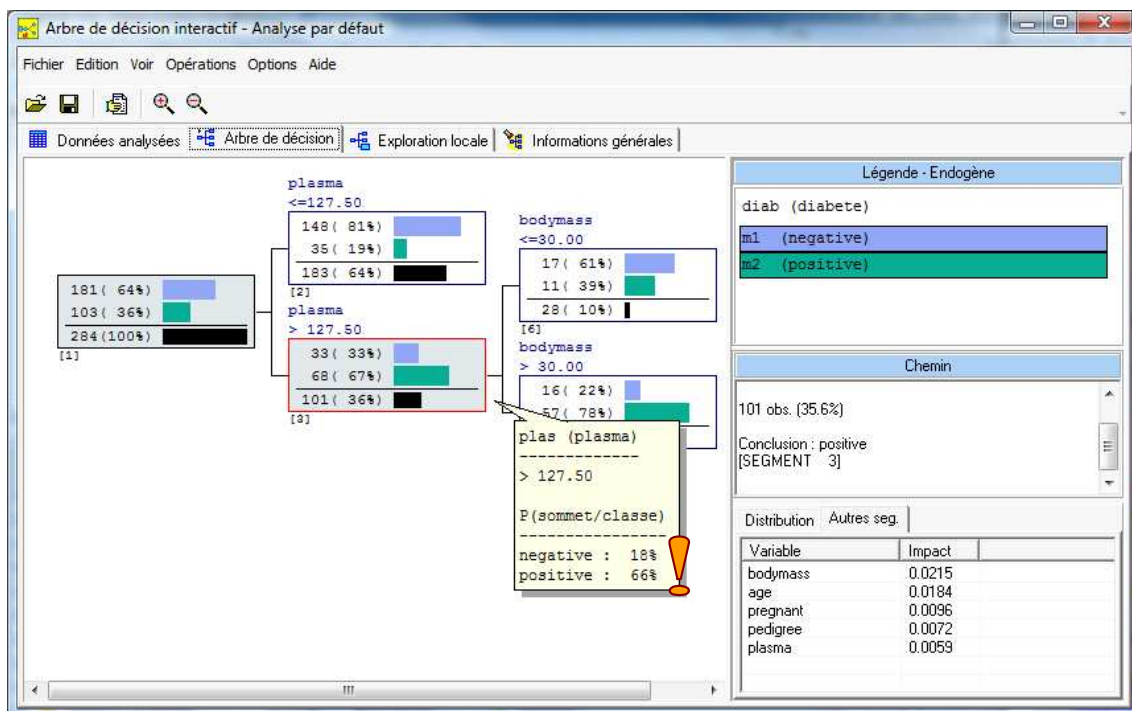
N°	N°SEGMENT	Condition d'appartenance	Support	Conclusion	Pr(negative)	Pr(positive)	Rappel(negative)	Rappel(positive)
1	2	SI plasma <=127.50	0.6444	negative	0.8087	0.1913	0.8177	0.3398
2	6	SI bodymass <=30.00 et plasma > 127.50	0.0986	negative	0.6071	0.3929	0.0939	0.1068
3	7	SI bodymass > 30.00 et plasma > 127.50	0.2570	positive	0.2192	0.7808	0.0884	0.5534

Date de création : 16/01/2010 22:07:49
Terminé

3.1.6 Exploration locale d'un sommet

IDT intègre les fonctionnalités interactives usuelles pour ce type de logiciel. Dans la fenêtre « Arbre de décision », nous disposons d'une série d'outils pour apprécier pleinement les informations que peuvent apporter la modélisation.

Par exemple, le sommet n°3 comporte 101 observations (36% de 284), dont 33 sont diabète = négatifs (33% = 33 / 101 et 68 sont positifs (67% = 68 / 101). En cliquant sur le sommet, dans une bulle apparaissent la description de la règle et d'autres probabilités conditionnelles : 18% des négatifs (18% = 33 / 181) et 66% des positifs (66% = 58 / 103) appartiennent au sommet.

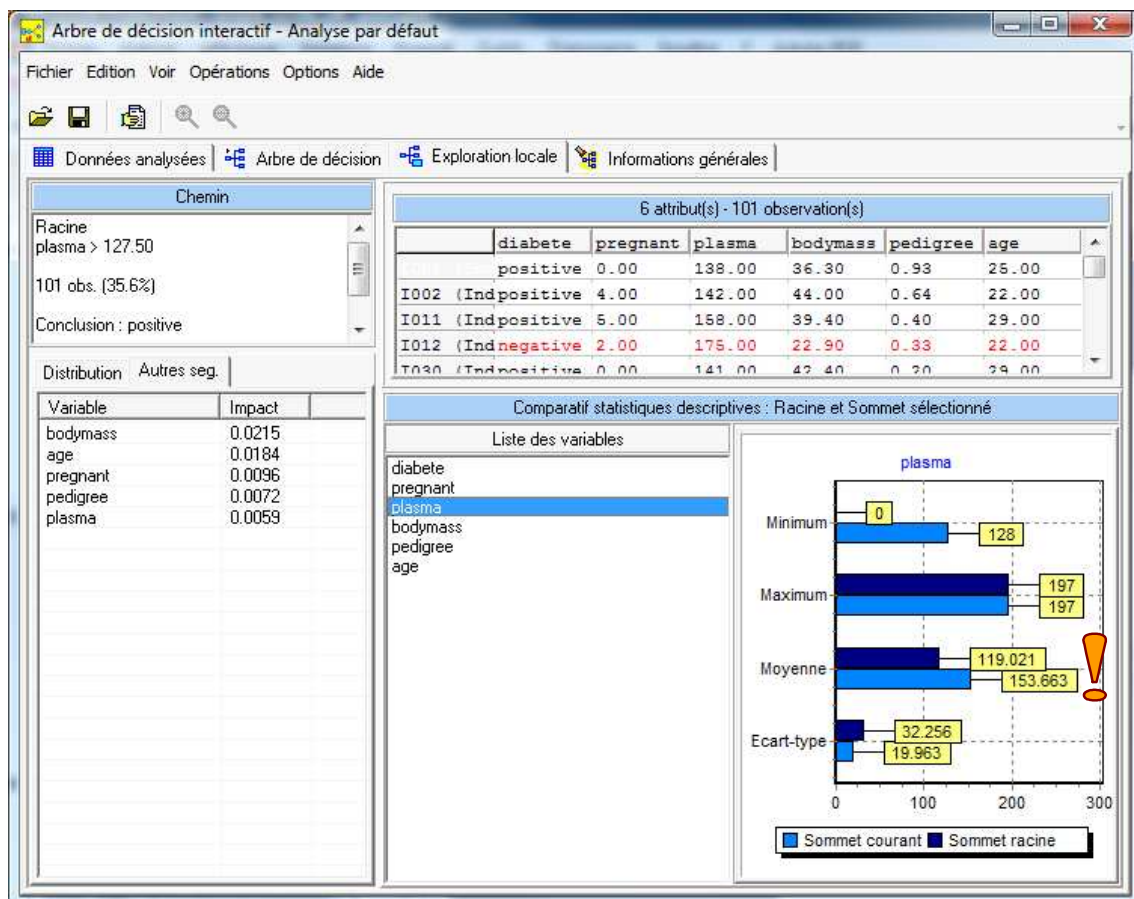


Dans la partie basse de la fenêtre, à droite, nous disposons de l'impact de chaque variable de segmentation (Gain de Gini). La variable la plus pertinente pour segmenter le sommet semble être BODYMASS avec un gain = 0.0215. C'est justement le découpage qui a été introduit pour aboutir aux sommets n°6 et n°7.

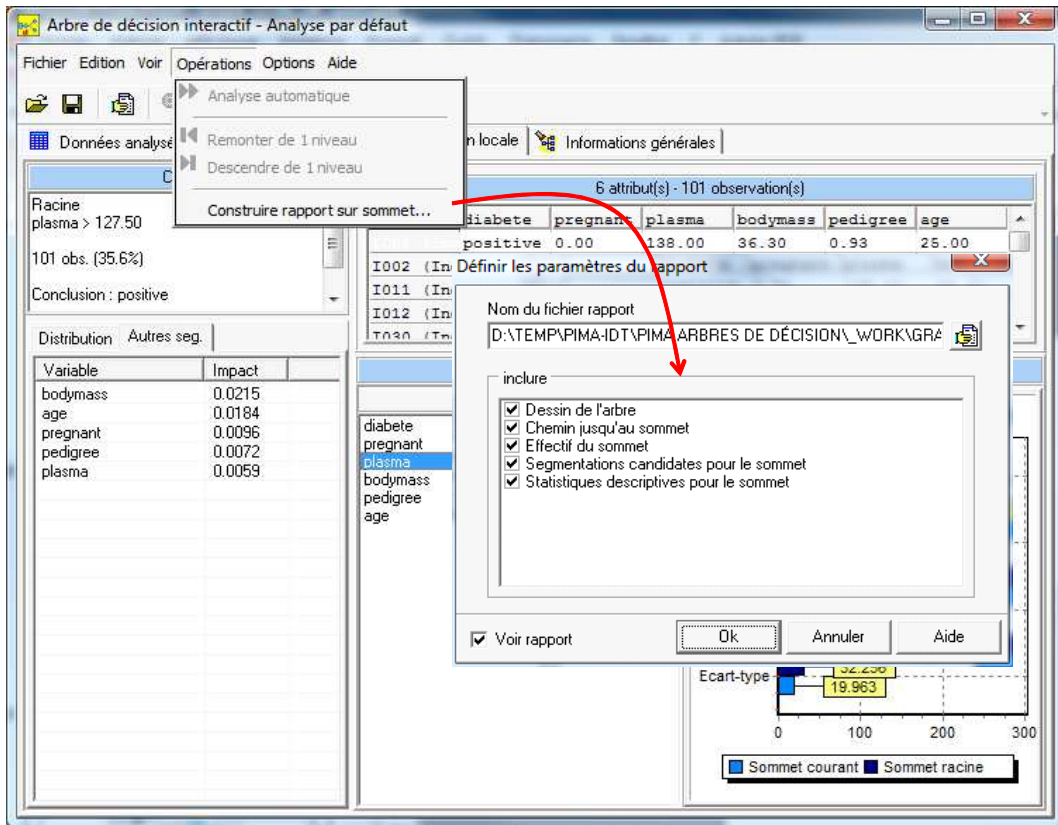
Pour explorer en détail ce sommet n°3, il faut activer l'onglet « Exploration locale » :

- Nous pouvons visualiser le tableau de données. La classe majoritaire est « positive », les individus négatifs sont donc signalés en rouge.
- Dans la partie basse de la fenêtre, nous avons les statistiques descriptives comparant la racine de l'arbre (la totalité de l'échantillon) avec le sommet sélectionné (le sous-groupe d'observations correspondant au sommet). Si nous sélectionnons la variable PLASMA par exemple, nous apprenons que globalement, sa moyenne est de 119.021 ; dans le groupe que nous étudions, il passe à 153.663. Les individus circonscrits par ce sommet ont, semble-t-il, un plasma plus élevé que la globalité de la population.

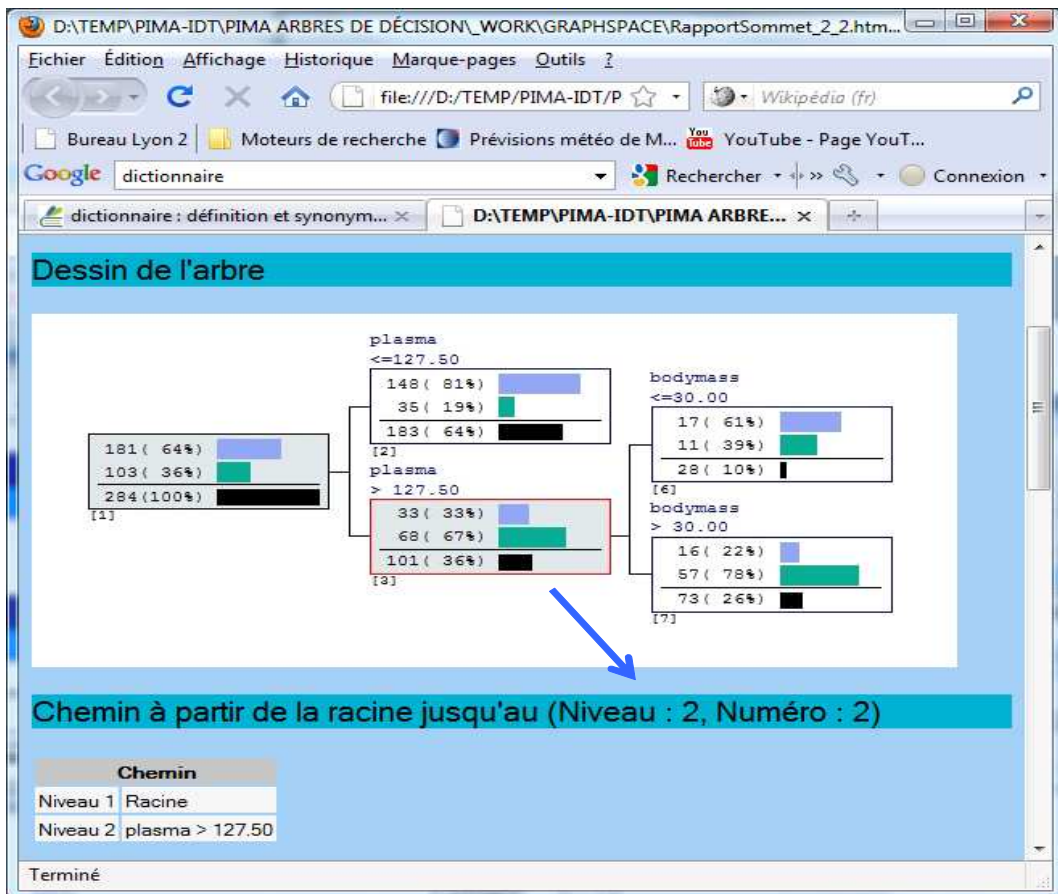
Nous pouvons procéder ainsi pour chaque variable de l'étude.



Rapport sur un sommet. Bien entendu, il est possible de produire un rapport personnalisé sur le sommet en cours d'analyse. Nous actionnons le menu « OPERATIONS / CONSTRUIRE RAPPORT SUR SOMMET ». Nous laissons la sélection par défaut dans la boîte de paramétrage.



Un rapport au format HTML est généré, il est chargé dans le navigateur par défaut de votre système.



Le sommet analysé est surligné en rouge dans le dessin de l'arbre. Nous disposons d'une série d'informations permettant de caractériser le groupe d'observations associé. A ce titre, dans la dernière partie du rapport, une confrontation systématique des statistiques descriptives pour la totalité des observations (le sommet de l'arbre) et celles couvertes par le sommet sélectionné est réalisé pour toutes les variables de l'analyse.

Variable(s)	Statistiques			
	Observations locales		Toutes les observations	
diabete	m1 (negative)	33%	m1 (negative)	64%
	m2 (positive)	67%	m2 (positive)	36%
pregnant	Min.	0.00	Min.	0.00
	Max.	15.00	Max.	15.00
	Moyenne	4.55	Moyenne	3.58
	Ecart-type	3.56	Ecart-type	3.18
plasma	Min.	128.00	Min.	0.00
	Max.	197.00	Max.	197.00
	Moyenne	153.66	Moyenne	119.02
	Ecart-type	19.96	Ecart-type	32.26
bodymass	Min.	0.00	Min.	0.00
	Max.	52.30	Max.	55.00
	Moyenne	33.80	Moyenne	31.92
	Ecart-type	7.57	Ecart-type	7.60
pedigree	Min.	0.12	Min.	0.09
	Max.	2.33	Max.	2.33
	Moyenne	0.54	Moyenne	0.48
	Ecart-type	0.39	Ecart-type	0.33
age	Min.	21.00	Min.	21.00
	Max.	81.00	Max.	81.00
	Moyenne	38.03	Moyenne	32.59
	Ecart-type	13.25	Ecart-type	11.64

Date de création : 17/01/2010 01:28:17
Terminé

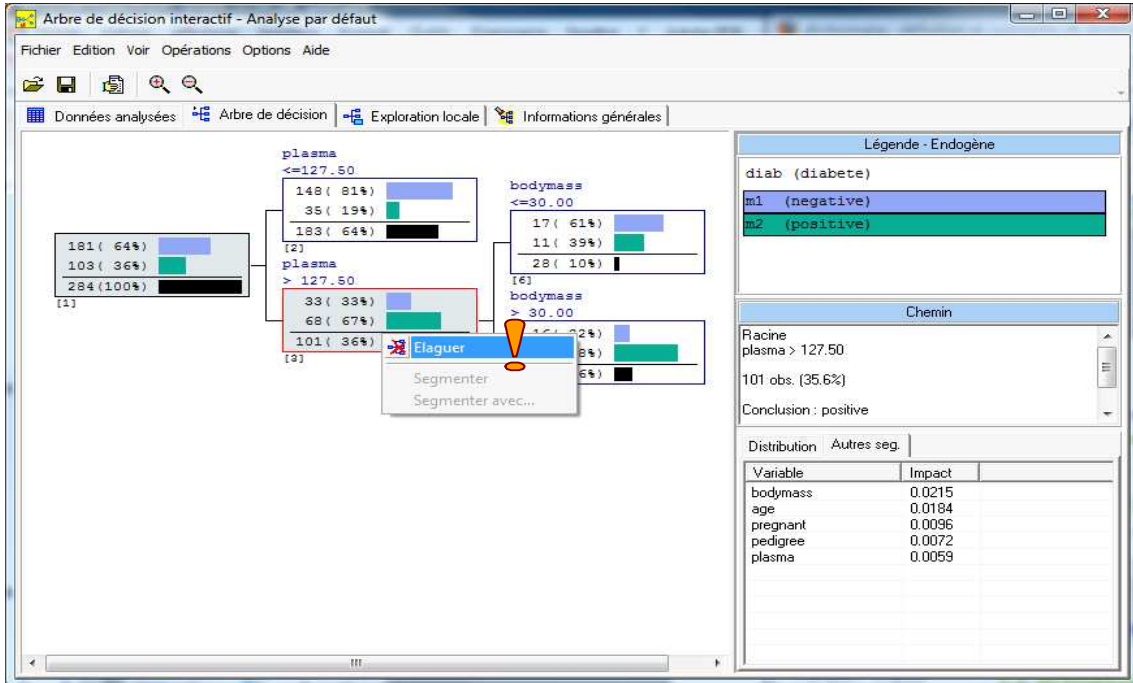
Nous retrouvons les valeurs lues dans la fenêtre d'exploration locale. Par exemple, la moyenne de PLASMA est de 119.02 dans la totalité de l'échantillon (par extension dans la population), elle est de 153.66 dans ce groupe (par extension, dans la sous population circonscrite par le sommet).

3.1.7 Construction interactive de l'arbre

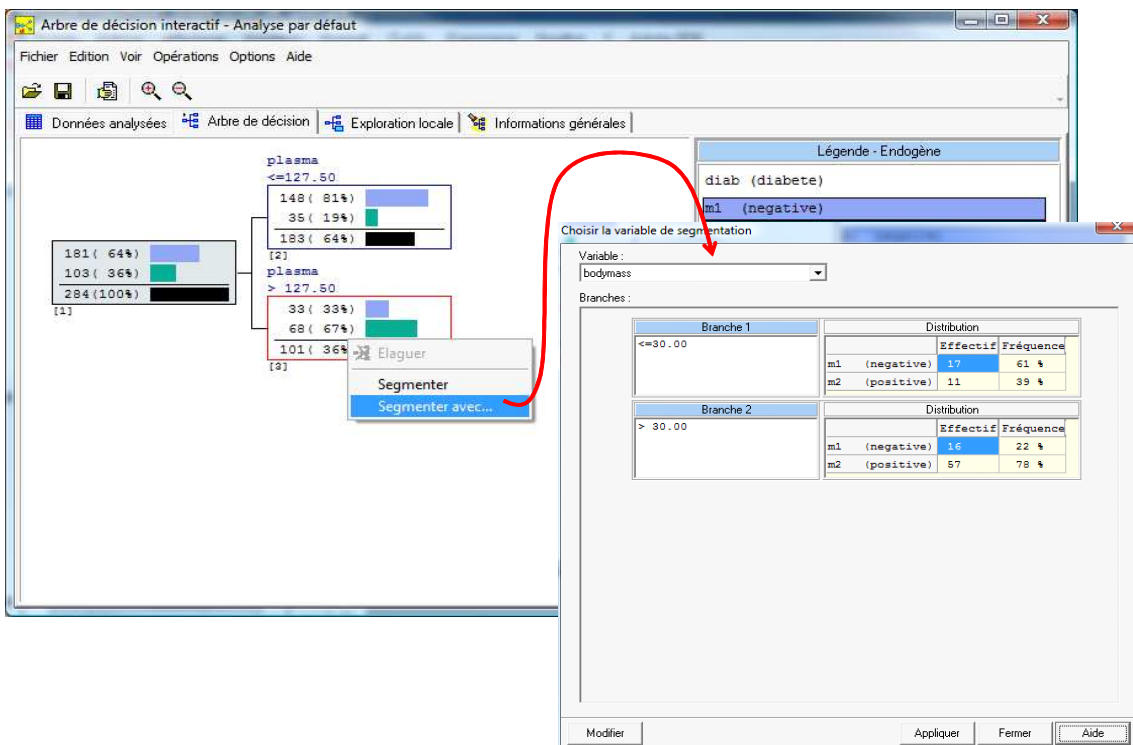
Elagage manuel et choix de la variable de segmentation. Revenons dans la fenêtre « Arbre de décision ». Nous sélectionnons toujours le sommet n°3, nous observons qu'il a été traité avec la variable BODYMASS, avec un « impact » (gain de Gini) de 0.0215. Nous observons également que la variable AGE est une alternative possible pour la segmentation. L'impact n'est guère plus faible avec

o.0184. Nous souhaitons remplacer BODYMASS par AGE pour la segmentation du sommet. L'opération est réalisée en plusieurs temps.

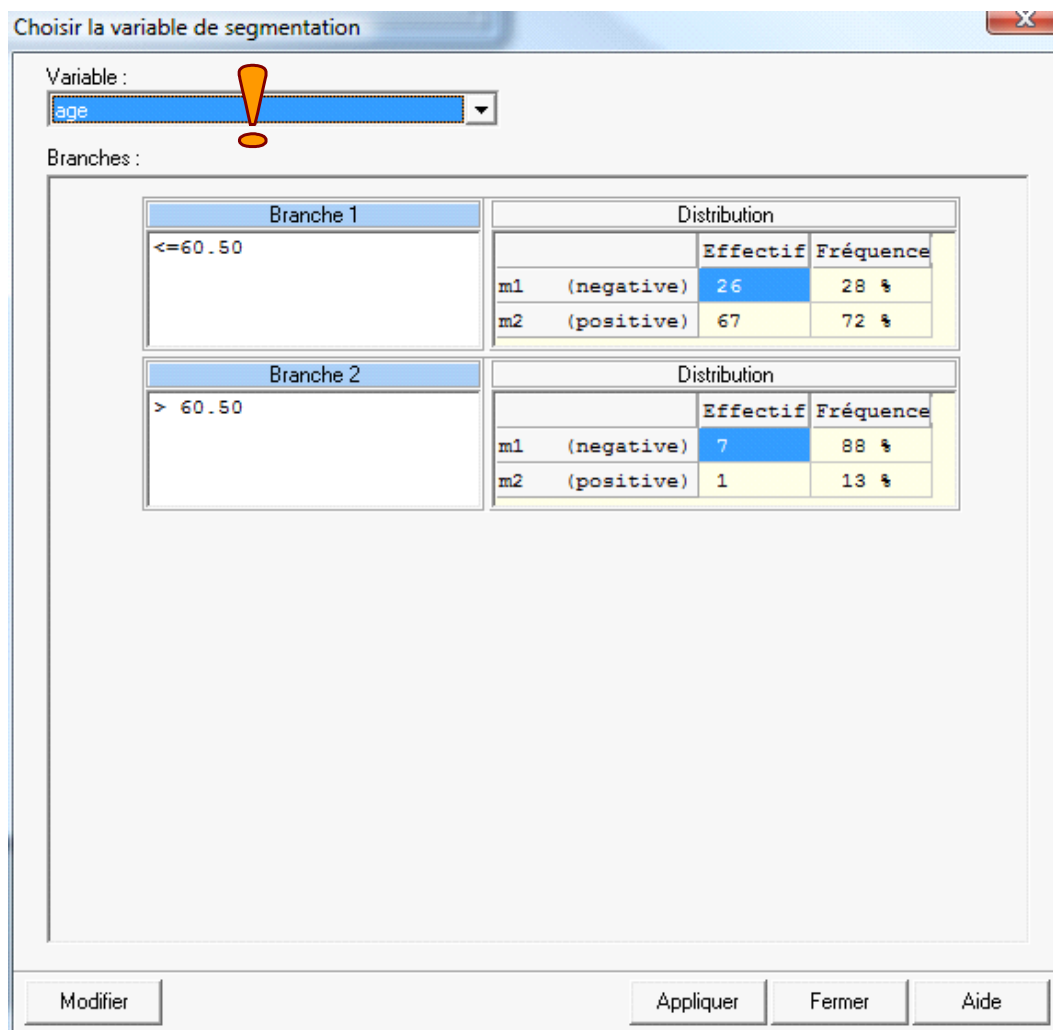
(1) Nous supprimons les branches situées en aval du sommet en actionnant le menu contextuel puis en cliquant sur l'item « Elaguer ».



(2) Toujours avec le menu contextuel, nous actionnons l'item « Segmenter avec... ». Une boîte de dialogue apparaît. Nous y observons la liste des variables potentielles de segmentation, triées selon leur impact. BODYMASS est la première puisqu'elle présente l'impact le plus élevé.

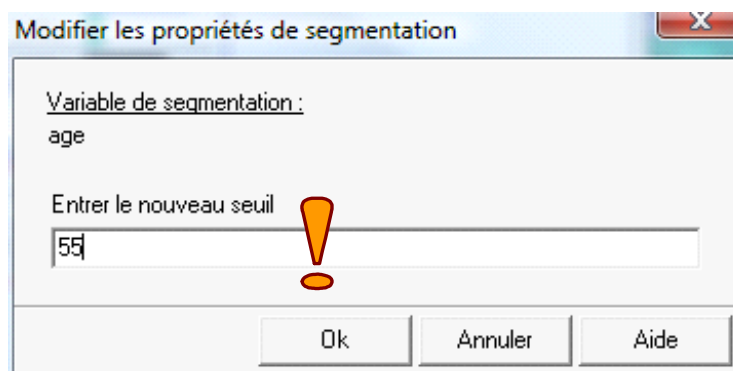


Mais, bien évidemment, nous pouvons sélectionner la variable de segmentation que l'on souhaite. Dans notre cas, nous choisissons AGE (qui est en deuxième position). Les distributions des classes dans les feuilles induites sont affichées.

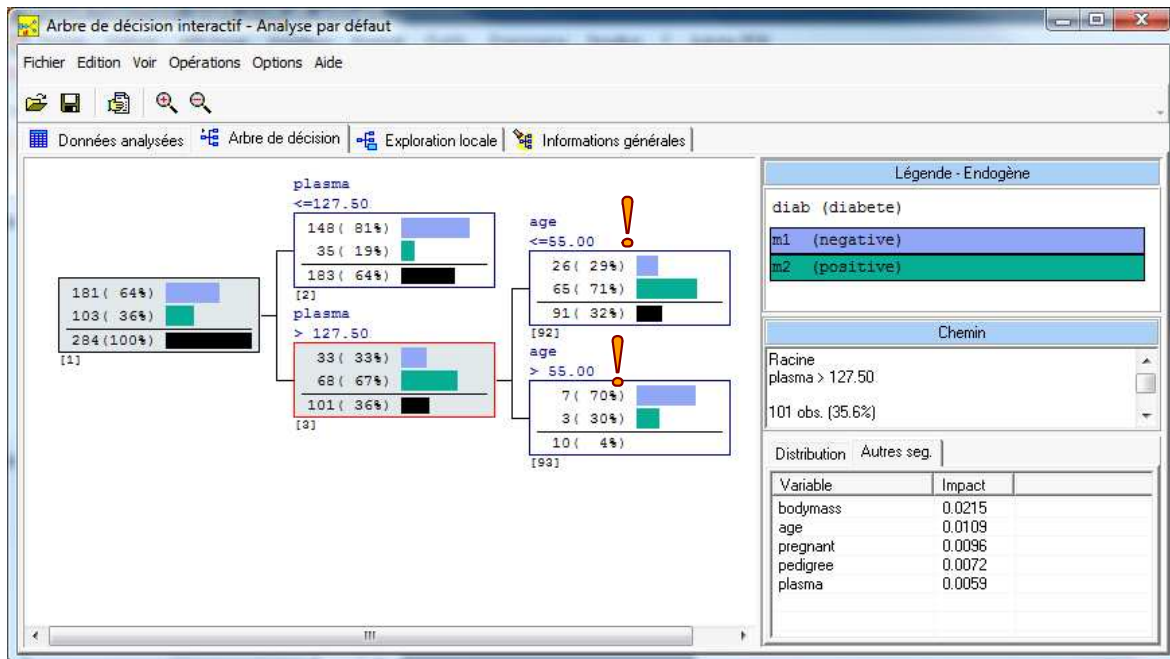


Modification des caractéristiques de la segmentation. Mettons, pour être pénible, que le seuil 60.50 déterminé automatiquement par calcul ne nous convienne pas. Nous souhaitons lui substituer la valeur 55, en accord avec les connaissances du domaine.

Il faut dans ce cas actionner le bouton **MODIFIER** situé dans la partie basse de la boîte de dialogue. Nous introduisons le nouveau seuil dans la boîte de saisie qui apparaît.

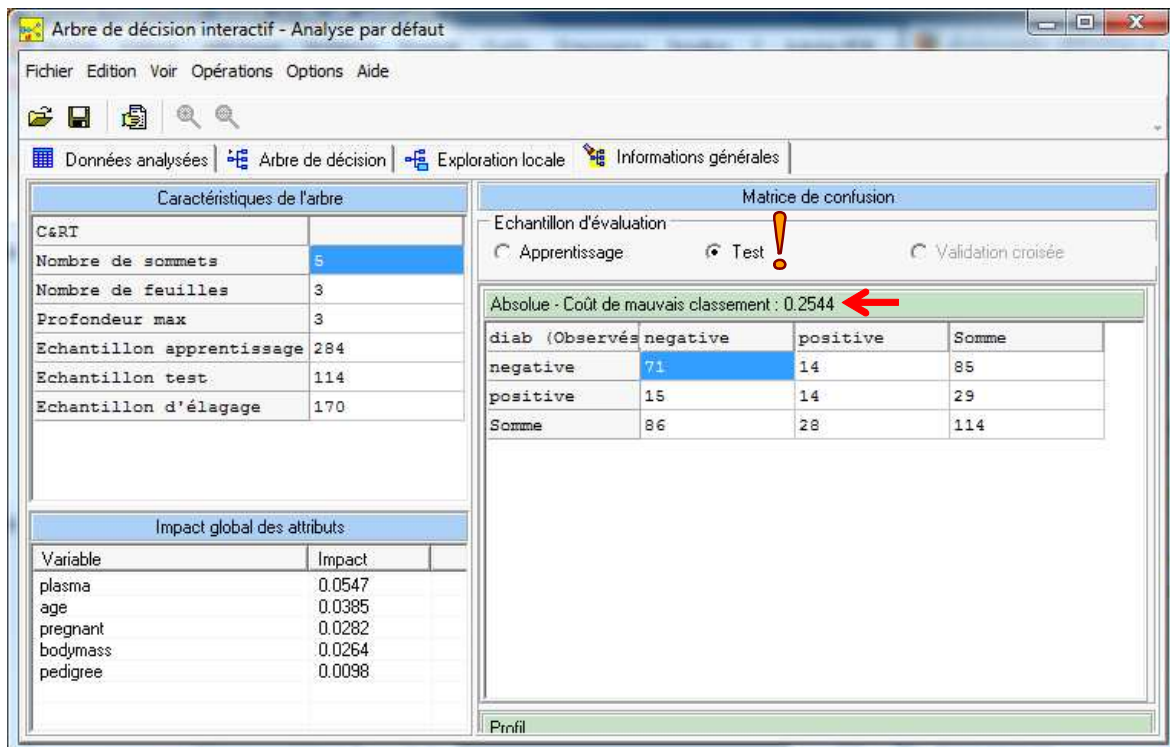


Il ne nous reste plus qu'à valider ces différentes modifications, nous obtenons une nouvelle version de l'arbre de décision.



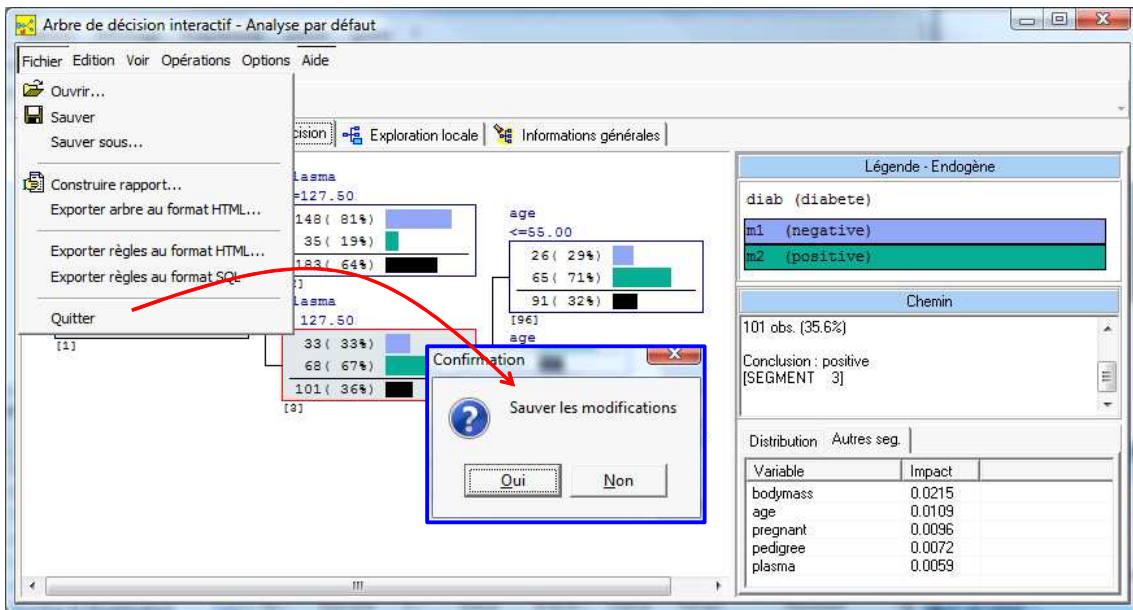
Bien entendu, comme nous avons modifié le seuil de discrétisation pour AGE, l'impact de la variable a été recalculé. Il est de 0.0109 maintenant.

Nous retournons dans la feuille « Informations Générales » pour obtenir une évaluation des performances de l'arbre ainsi défini. Le taux d'erreur en test est de 25.44%, il faudra s'en souvenir pour la suite.

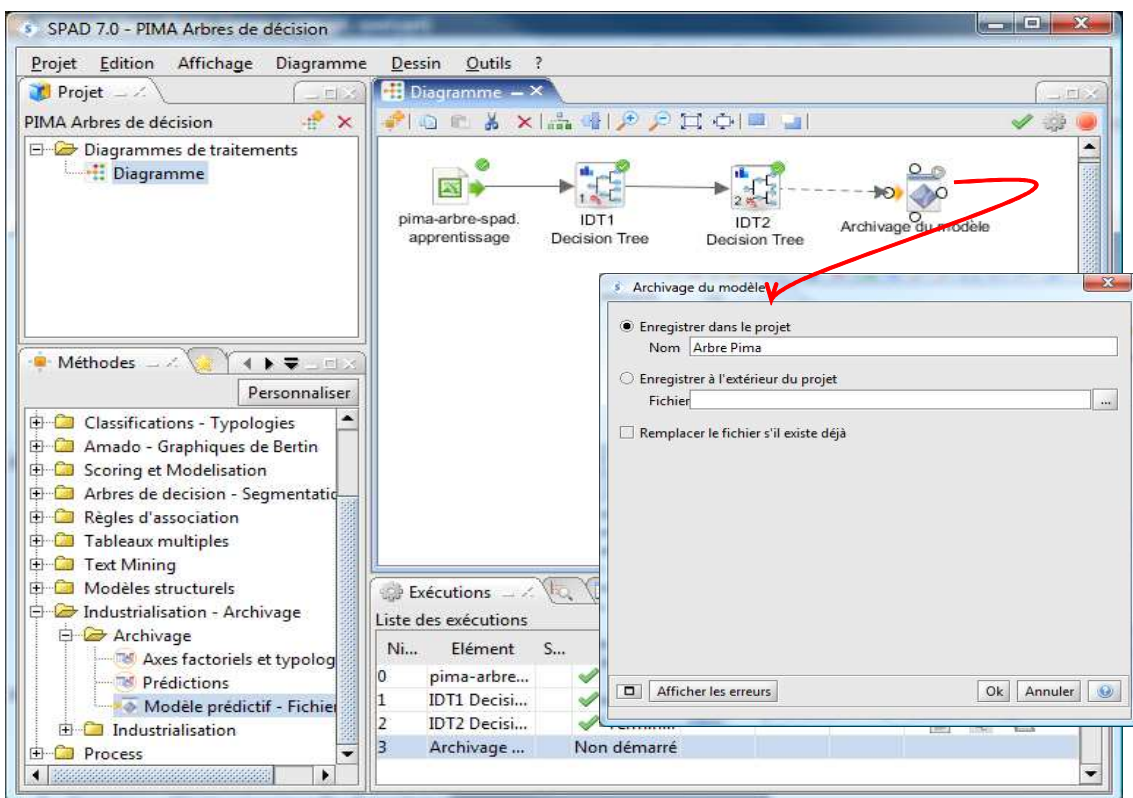


3.1.8 Archivage du modèle prédictif

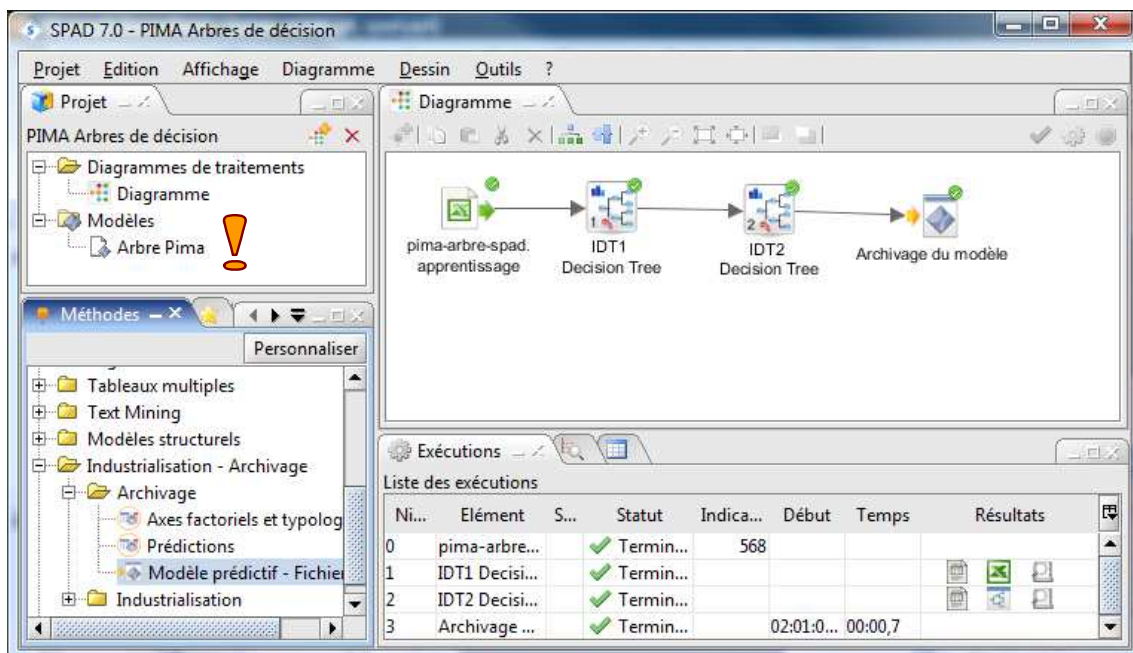
Nous refermons l'application « Arbre de décision interactif » (FICHIER / QUITTER). Nous sauvons la nouvelle version de l'arbre bien évidemment puisque nous avons modifié le modèle.



Nous devons archiver ce modèle, afin de pouvoir l'appliquer sur d'autres sources de données notamment. Nous introduisons le composant **Modèle prédictif – Fichier règles** dans l'espace de travail. Nous lui relions le composant IDT2. Nous le paramétrons (menu PARAMETRES). Nous affectons un nom, « Arbre Pima » par exemple.



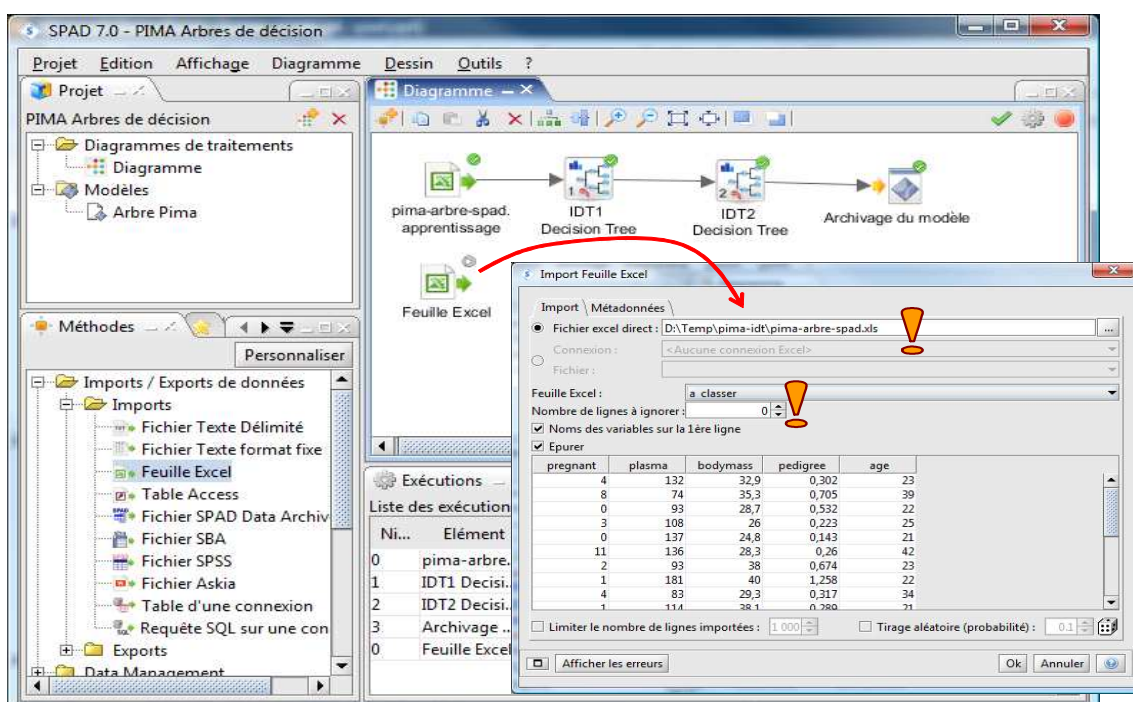
Après validation, nous remarquerons qu'un nouvel élément a été ajouté au gestionnaire de projet (à gauche). Dans une branche « Modèle » est maintenant intégré notre modèle prédictif.



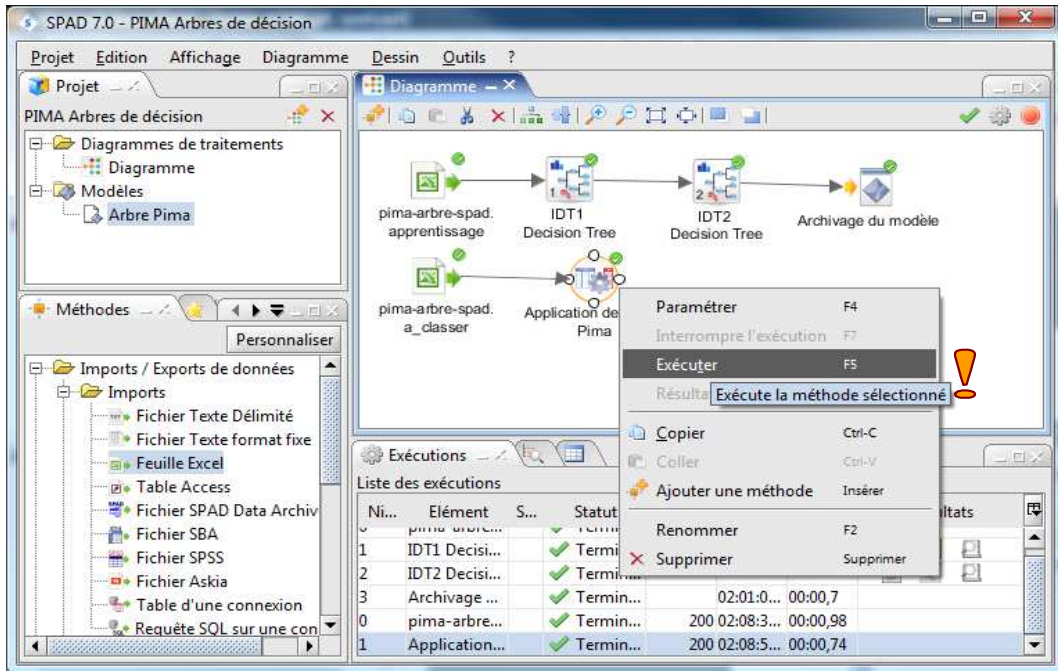
La première étape de notre étude est terminée. Nous avons construit un modèle prédictif, sous forme d'un arbre de décision dans ce tutoriel, mais ça pourrait être toute autre méthode d'apprentissage. Il est archivé, prêt à être utilisé en déploiement.

3.2 Déploiement du modèle sur les données « à classer »

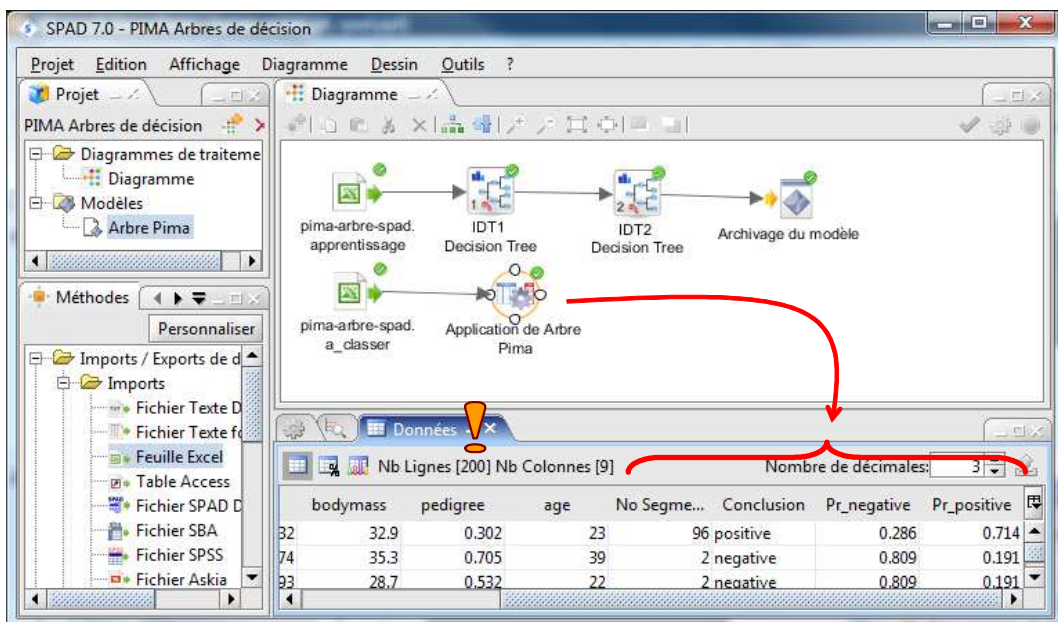
Pour appliquer le modèle sur de nouvelles données, nous devons charger les observations, en l'occurrence la seconde feuille « à classer » de notre fichier Excel. Nous introduisons un second composant **Feuille Excel**, nous le paramétrons de la manière suivante.



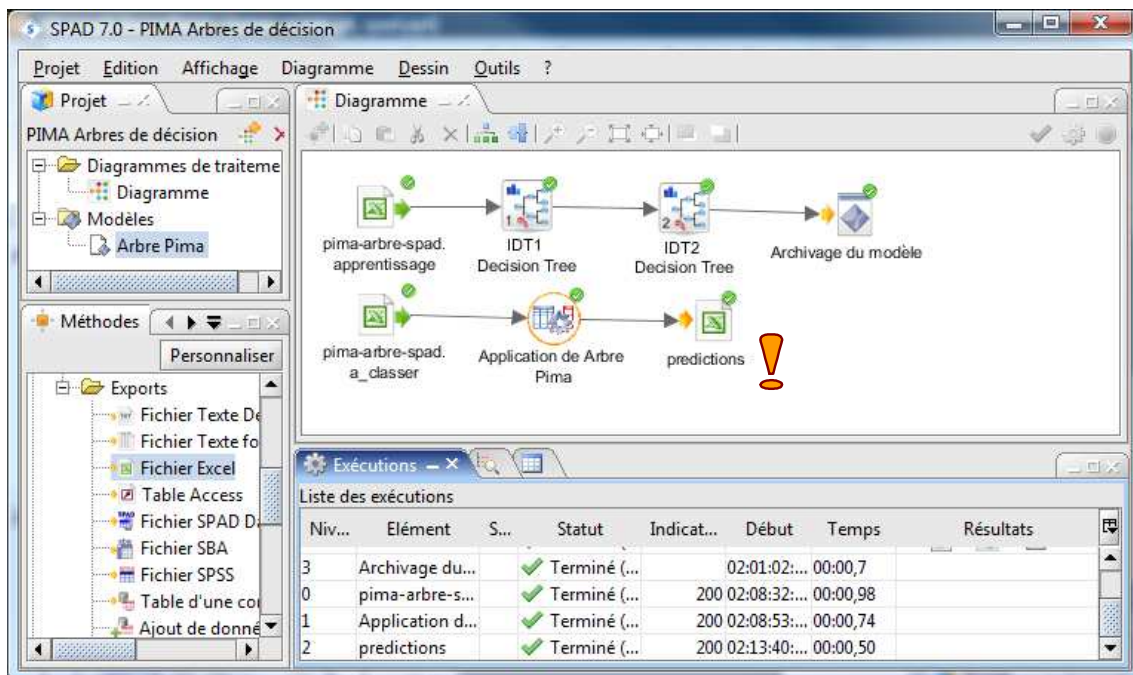
Nous ajoutons le modèle **Arbre Pima** dans l'espace de travail, nous lui relions la source de données « à classer ». Il n'y a pas de paramétrage à effectuer, nous actionnons directement le menu EXECUTER (F5). *Remarque : Dans certaines versions de SPAD, il semble nécessaire de paramétrer l'outil, de cliquer sur le bouton « Charger le Modèle », puis enfin de cocher l'option « Conclusion / Prédiction » pour que la conclusion du modèle soit additionnée aux données produites par le composant.*



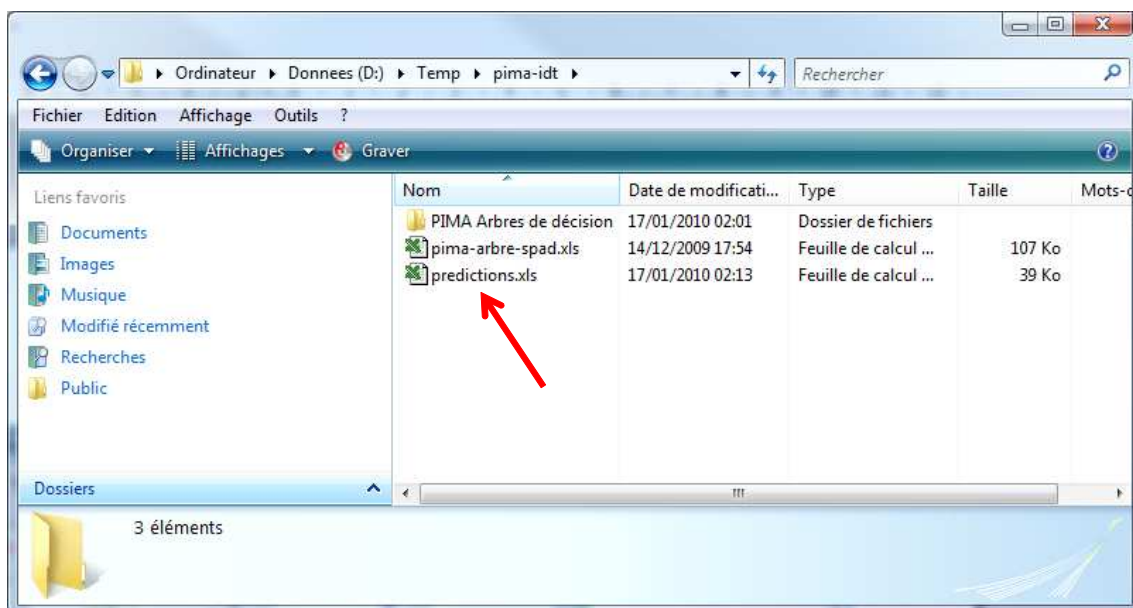
L'icône a été automatiquement baptisée «Application de Arbre Pima ». L'espace de suivi, dans la partie basse de la fenêtre principale, joue un rôle très important à ce stade. Nous sélectionnons l'onglet « Données », nous y observons : les données en provenance de la feuille Excel ; le numéro interne de sommet utilisé par IDT pour différencier les feuilles ; la prédiction du modèle (Conclusion); les probabilités conditionnelles d'appartenance aux classes (Pr_negative et Pr_positive) que l'on peut exploiter dans le cadre du scoring par exemple.



Il ne reste plus qu'à sauvegarder ces informations dans un nouveau fichier. Nous utilisons le composant **Fichier Excel** de la branche **Exports** de la palette des méthodes. Nous lui relions l'icône précédente, nous le paramétrons en spécifiant le nom du fichier à générer : « predictions.xls ».



Après validation, le fichier est créé, comme nous pouvons le constater dans l'explorateur Windows.



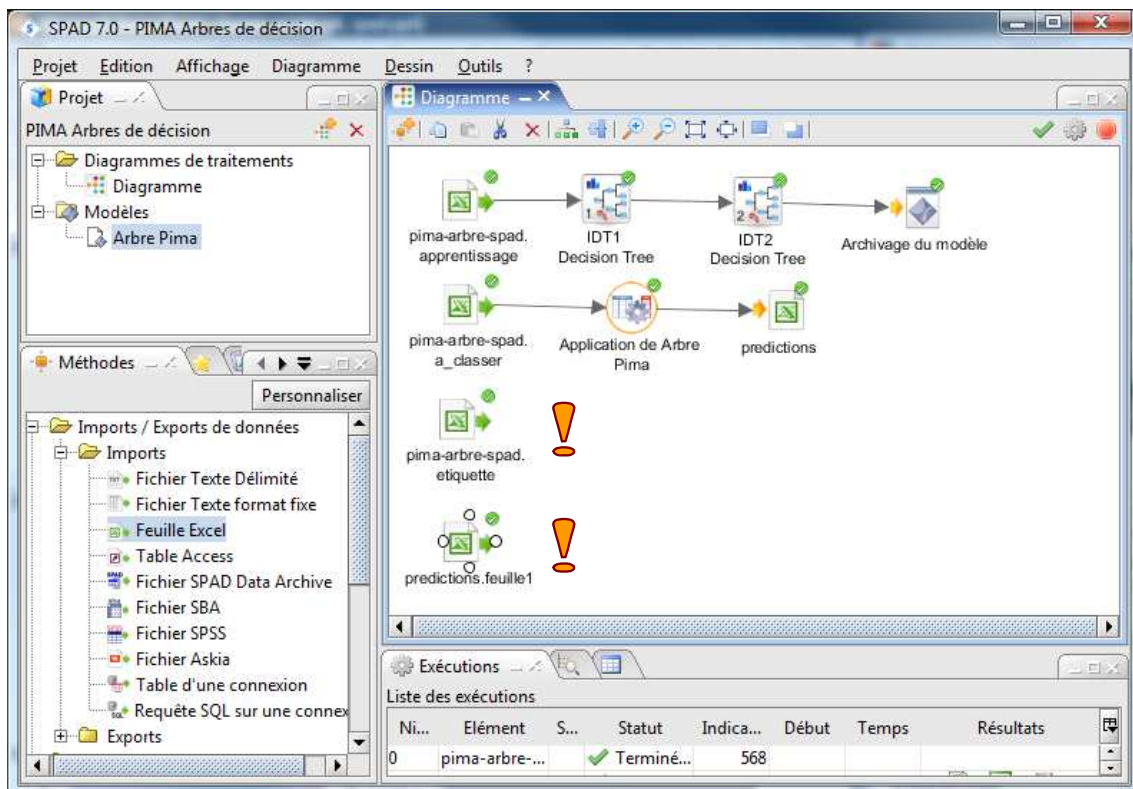
3.3 Vérification des prédictions – Confrontation avec les données « étiquettes »

En situation réelle, notre travail s'arrête là. Dans notre cas, nous avons la chance de disposer des étiquettes observées des individus à classer dans la troisième feuille « étiquette ». Nous nous efforçons donc de confronter nos prédictions (la colonne « conclusion » du fichier « predictions.xls ») avec les vraies valeurs de la variable à prédire.

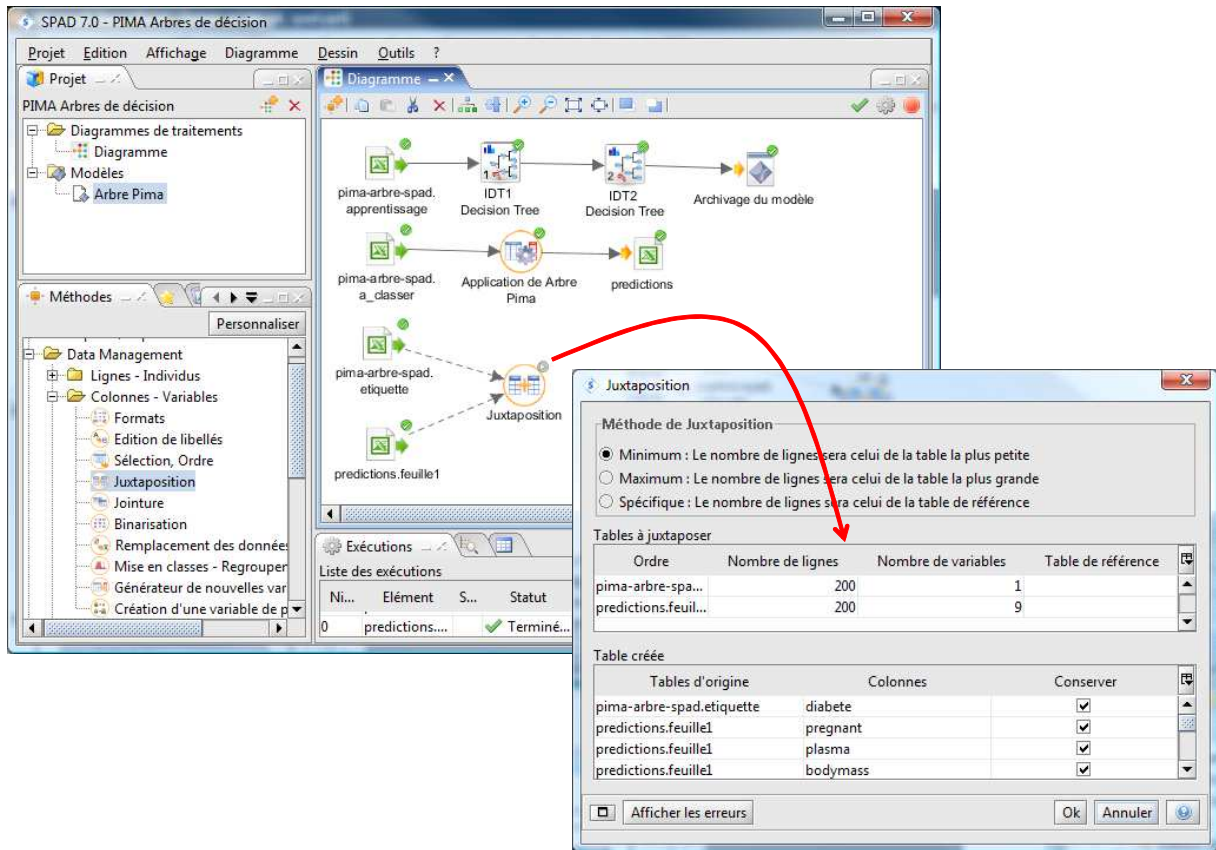
Pour cela, nous devons accéder aux deux fichiers, les fusionner pour mettre en correspondance les individus en provenance des deux sources, puis construire un tableau croisé pour confronter les prédictions avec les vraies valeurs observées.

3.3.1 Importation et fusion des fichiers

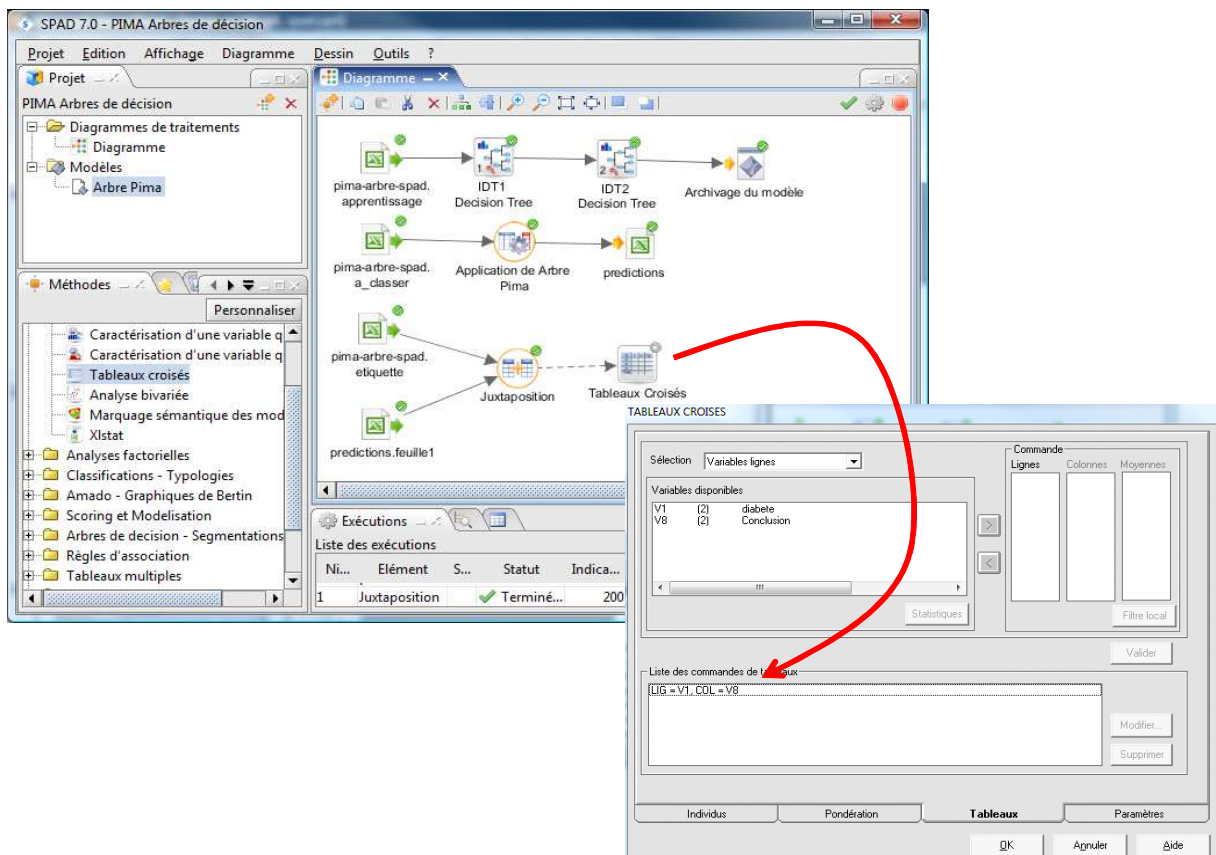
Nous importons les deux sources de données via deux composants **Feuille Excel**. Le premier est branché sur la feuille « étiquette » de « pima-arbre-spad.xls », le second sur la « feuille1 » de « predictions.xls ».



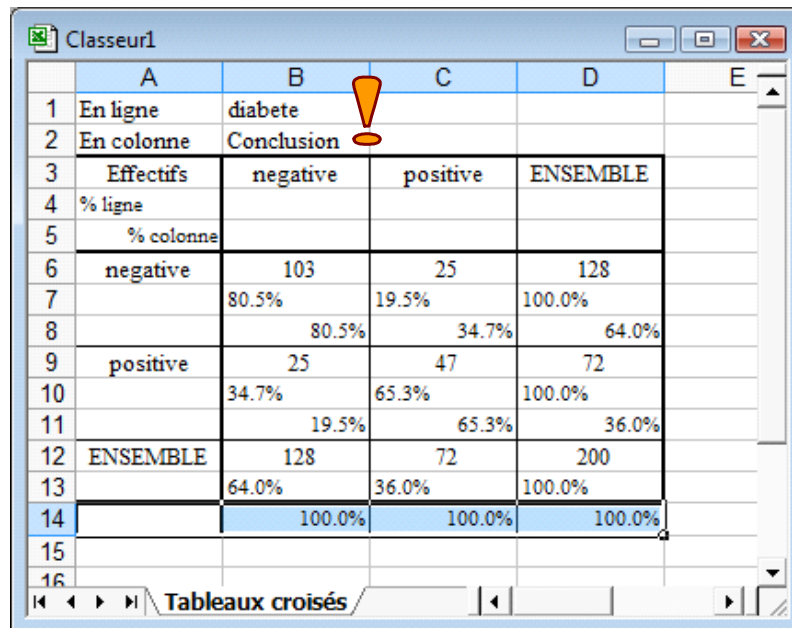
Nous réalisons la liaison entre les deux sources avec le composant **Juxtaposition**. La boîte de paramétrage nous fournit des indications sur le mode de création du nouvel ensemble de données. Cet outil est particulièrement pratique lorsque nous devons réunir régulièrement plusieurs fichiers épars.



Nous souhaitons confronter la colonne « diabète » observée, et celle prédite par l'arbre. Le composant **Tableau Croisé** fera l'affaire. Nous le paramétrons de la manière suivante.



Il ne reste plus qu'à lire le tableau, nous actionnons le menu contextuel et nous choisissons l'option RESULTATS / SORTIES EXCEL (F9).



	A	B	C	D	E
1	En ligne	diabete			
2	En colonne	Conclusion			
3	Effectifs	negative	positive	ENSEMBLE	
4	% ligne				
5	% colonne				
6	negative	103	25	128	
7		80.5%	19.5%	100.0%	
8		80.5%	34.7%	64.0%	
9	positive	25	47	72	
10		34.7%	65.3%	100.0%	
11		19.5%	65.3%	36.0%	
12	ENSEMBLE	128	72	200	
13		64.0%	36.0%	100.0%	
14		100.0%	100.0%	100.0%	
15					
16					

Le taux d'erreur sur le fichier de déploiement est

$$\text{Taux d'erreur} = (25 + 25) / 200 = 1.0 - (103 + 47) / 200 = 25\%$$

Ce qui est tout à fait conforme avec le taux d'erreur sur l'échantillon test, annoncé lors de la construction interactive (qui était de 25.4%, section 3.1.7).

Le tableau fournit également les profils lignes et colonnes. Il est par conséquent très facile de lire les autres indicateurs : en ligne, la sensibilité = $47 / 72 = 65.3\%$; en colonne, la précision = $47 / 72 = 65.3\%$ (c'est un hasard si nous avons une valeur identique à la sensibilité) ; en ligne, la spécificité = $103 / 128 = 80.5\%$; en ligne, le taux de faux positifs = $25 / 128 = 19.5\%$.

4 Traitements sous R

Je le dis, et je le dirai toujours : « sans compétences, le logiciel n'est rien³ ». Les différentes opérations ci-dessus peuvent être reproduites à l'aide de la grande majorité des logiciels de Data Mining, mais peut être pas avec la même aisance.

Si je prends l'exemple du logiciel R, quelques lignes de codes suffisent. Avec un sérieux bémol cependant : il faut connaître les bonnes instructions et savoir les enchaîner ; les arbres ne sont pas interactifs dans R.

Voici les commandes permettant de réaliser l'ensemble des étapes ci-dessus.

³ Ok, ok. Ceci est un pastiche de la fameuse pub avec Marie-Jo Perec, belle comme une sylphide, qui, au terme d'une série de foulées aériennes dans un décor dantesque, nous affirmait que « Sans contrôle, la puissance n'est rien ». Ceux de ma génération auront compris.

```
#vider la mémoire
rm (list=ls())

#####
#chargement des données
#####
library(xlsReadWrite)
#charger les différentes feuilles du classeur XLS
pima.train <- read.xls(file="pima-arbre-spad.xls",colNames=T,sheet="apprentissage")
pima.unlabeled <- read.xls(file="pima-arbre-spad.xls",colNames=T,sheet="a_classer")
pima.label <- read.xls(file="pima-arbre-spad.xls",colNames=T,sheet="etiquette")

#####
#induction d'un arbre de décision avec la méthode rpart du package éponyme
#####
library(rpart)
#création de l'arbre maximal, non (post) élagué
tree.unpruned <- rpart(diabete ~ ., data = pima.train)
#détection de la valeur adéquate de cp (complexity parameter) pour l'élagage
plotcp(tree.unpruned)
#la valeur cp = 0.04 est utilisé pour élaguer l'arbre
tree.pruned <- prune(tree.unpruned,cp=0.04)
#affichage de l'arbre de décision
print(tree.pruned)

#####
#déploiement et vérification
#####
#appliquer l'arbre sur les données non étiquetées
prediction <- predict(tree.pruned,newdata = pima.unlabeled,type="class")

#comparer la prédiction et les étiquettes observées – matrice de confusion
conf.matrix <- table(pima.label$diabete,prediction)
print(conf.matrix)

#calculer le taux d'erreur
error.rate <- 1.0 - (conf.matrix[1,1]+conf.matrix[2,2])/sum(conf.matrix)
print(error.rate)
```

Voici les résultats les plus importants : (1) L'arbre de décision obtenu après élagage.


```

R Console
> print(tree.pruned)
n= 568

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

1) root 568 196 negative (0.6549296 0.3450704)
2) plasma< 130.5 385 79 negative (0.7948052 0.2051948) *
3) plasma>=130.5 183 66 positive (0.3606557 0.6393443)
6) bodymass< 29.95 47 16 negative (0.6595745 0.3404255) *
7) bodymass>=29.95 136 35 positive (0.2573529 0.7426471) *
>

```

« Plasma » et « Bodymass » sont les deux variables qui interviennent dans la construction du modèle, tout comme dans IDT avant que nous modifiions manuellement l'arbre.

(2) La matrice de confusion et le taux d'erreur obtenu sur le fichier non étiqueté

```

R Console
> print(conf.matrix)
      prediction
      negative positive
negative   114      14
positive   36      36
>
> #computing the error rate
> error.rate <- 1.0 - (conf.matrix[1,1]+conf.matrix[2,2])/sum(conf.matrix)
> print(error.rate)
[1] 0.25

```

Le taux d'erreur est le même qu'avec IDT de SPAD. La structure de la matrice de confusion est différente en revanche.

5 Conclusion

Un logiciel n'a pas vocation à réfléchir à notre place. En revanche, il doit nous fournir les instruments pour nous dégager de tout un tas de tâches fastidieuses et répétitives : l'accès à des fichiers parfois dispersés, la préparation des données (ex. nettoyage, création de variables intermédiaires), la mise en forme des rapports, etc.

Nous pouvons ainsi nous consacrer à l'essentiel : réfléchir à la pertinence de ce que l'on est en train de faire (*est-ce que je réponds vraiment à la question initialement posée ?*); vérifier la validité des résultats (*est-ce que le modèle est transposable dans la population ?*); évaluer les performances; préparer un déploiement efficace ; etc.

Dans ce didacticiel, nous avons montré un enchaînement type de traitements à l'aide du logiciel SPAD. Il s'agissait de charger les données, créer un arbre de décision, l'appliquer sur de nouvelles observations, vérifier sa validité. Bien sûr, nous avons vu dans la foulée qu'il était possible de reproduire les mêmes étapes avec le logiciel R qui, lui, est libre. Sauf que : (1) il faut faire

l'apprentissage du langage de programmation, certains y sont définitivement réfractaires ; (2) les arbres ne sont interactifs dans R.

Pour le premier écueil, on peut se rabattre vers des outils libres qui fonctionnent par diagramme de traitements (ex. Tanagra, Knime, Orange, Weka, etc. – cf. <http://tutoriels-data-mining.blogspot.com/2008/03/dploiement-de-modles-avec-tanagra.html> ou <http://tutoriels-data-mining.blogspot.com/2008/04/schma-apprentissage-test-avec-orange.html>). Mais les fonctionnalités interactives lors de la construction de l'arbre manquent à l'appel.

Sur ce second écueil précisément, on peut se tourner vers SIPINA qui propose une série d'outils permettant à l'utilisateur de guider l'induction par arbre (cf. <http://tutoriels-data-mining.blogspot.com/2008/03/analyse-interactive-avec-sipina.html>). Mais, la possibilité d'automatiser des séquences de traitements, le déploiement⁴ et l'édition de rapports ne sont pas beaucoup développés.

Le principal attrait des logiciels commerciaux, entres autres, est d'intégrer dans le même environnement une série d'outils destinés à faciliter le travail quotidien du praticien du Data Mining. Ils sont importants, mais absolument pas valorisables en recherche. C'est une des raisons pour lesquelles ils sont généralement absents des logiciels issus du monde universitaire. Je me vois très mal, pour ma part, écrire un article scientifique sous prétexte que Tanagra sait lire les fichiers Excel (<http://tutoriels-data-mining.blogspot.com/2008/03/importation-fichier-xls-excel-macro.html>). Et pourtant, c'est un plus considérable quant à l'utilisabilité⁵ du logiciel.

⁴ L'opération est possible certes, mais au prix d'une préparation minutieuse des fichiers -- <http://tutoriels-data-mining.blogspot.com/2008/03/dploiement-de-modles-avec-sipina.html>

⁵ Si, si, le terme existe. Cf. <http://fr.wikipedia.org/wiki/Utilisabilité>