1 Objectif

Présentation des nouveaux arbres de décision interactifs de SPAD version 8.

Les arbres de décision interactifs font partie de la panoplie d'outils privilégiés du data miner[1]. D'une part parce que l'induction par arbres en elle-même est une méthode intéressante : elle se positionne honorablement par rapport aux autres techniques prédictives en termes de performance ; elle fournit une connaissance intelligible, facilement interprétable ; ses conditions d'utilisation sont particulièrement larges, aucune hypothèse sur les distribution n'est nécessaire, nous pouvons directement mixer les variables prédictives quantitatives et qualitatives, elle sait effectuer les codages les plus appropriés en fonction de la variable cible. D'autre part, du fait qu'elle soit interactive, elle donne la possibilité aux experts du domaine de guider l'exploration des solutions en accord avec des connaissances qui ne sont pas directement disponibles dans les données traitées. De fait, tous les grands éditeurs de logiciels de statistique et de data mining se doivent de proposer les outils – c'est un vrai critère de différenciation entre les logiciels – qui permettent aux utilisateurs d'interagir avec l'arbre de décision élaboré au préalable par les approches bien connues telles que CHAID, CART, C4.5[2] ou leurs variantes.

J'avais déjà présenté les arbres de décision de la version 7 de SPAD précédemment[3]. Aujourd'hui, je décris le module proposé par SPAD 8. En effet, il a connu une évolution importante, tant en qualité graphique, qu'en matière d'utilisabilité (grosso modo, un mix d'efficacité et d'ergonomie)[4]. Il me semblait intéressant d'étudier cette nouvelle mouture pour cerner ce que nous pouvons faire avec les arbres de décision interactifs. Je me concentre sur les fonctionnalités d'exploration dans ce tutoriel. Pour ce qui est du stockage du modèle et de son déploiement, le mieux est de lire/relire le précédent document (Janvier 2010).

2 Données

Nous cherchons à déterminer les facteurs prédisposant à la naissance de bébés à faibles poids. La variable cible « LowBirthWeigth » prend 2 valeurs possibles {yes, no}. Les variables prédictives décrivent les caractéristiques et le comportement de la mère (âge, poids, fumer

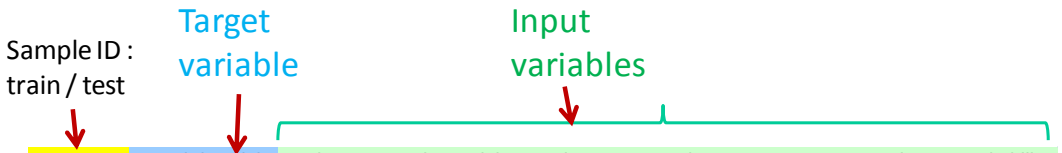
[1] S. Tufféry, « Quelle méthode de data mining utilisez vous le plus ? », sur le site « http://data.mining.free.fr/ ; Kdnuggets Polls, « Algorithms for data analysis / data mining », Novembre 2011, sur le site www.kdnuggets.com

[2] R. Rakotomalala, « Les méthodes d'induction d'arbres de décision – CHAID, CART, C4.5 et les autres… ».

[3] Tutoriel Tanagra, « Arbres de décision interactifs avec SPAD », Janvier 2010.

[4] Dixit Wikipédia, le terme recouvre 5 notions clés : l'efficience, la satisfaction, la facilité d'apprentissage, la facilité d'appropriation, et la fiabilité. http://fr.wikipedia.org/wiki/Utilisabilité

durant la grossesse, etc.). Il s'agit d'une version modifiée du jeu de données référencé, entres autres, dans l'ouvrage de Hosmer & al. (2013). La colonne SAMPLE permet d'identifier les individus appartenant aux échantillons d'apprentissage (TRAIN, 348 individus) et de test (TEST, 100). Voici les 10 premières observations du fichier « **faible poids bebes.xlsx** ».

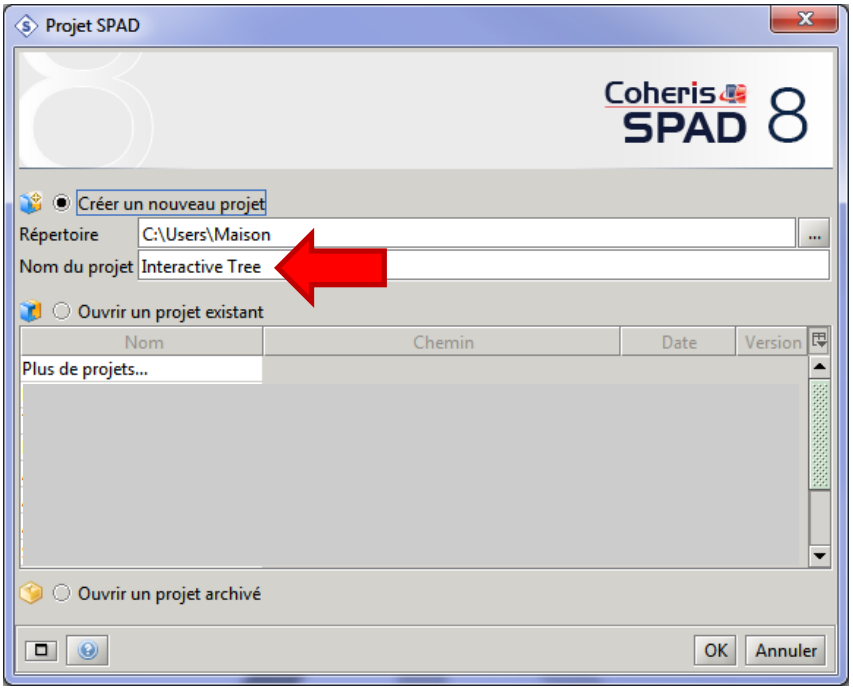


SAMPLE	LowBirthWeight	MotherAge	MotherWeight	SmokePregnant	HistPremature	Hypertension	UterIrritability
train	yes	17	142	no	no	yes	no
train	no	23	115	yes	no	no	no
test	no	26	160	no	no	no	no
train	no	25	155	no	no	no	no
train	no	31	215	yes	no	no	no
train	no	29	135	no	no	no	no
train	yes	17	120	no	no	no	no
train	no	20	169	no	yes	no	yes
train	no	27	124	yes	no	no	no
train	yes	25	115	no	no	no	no

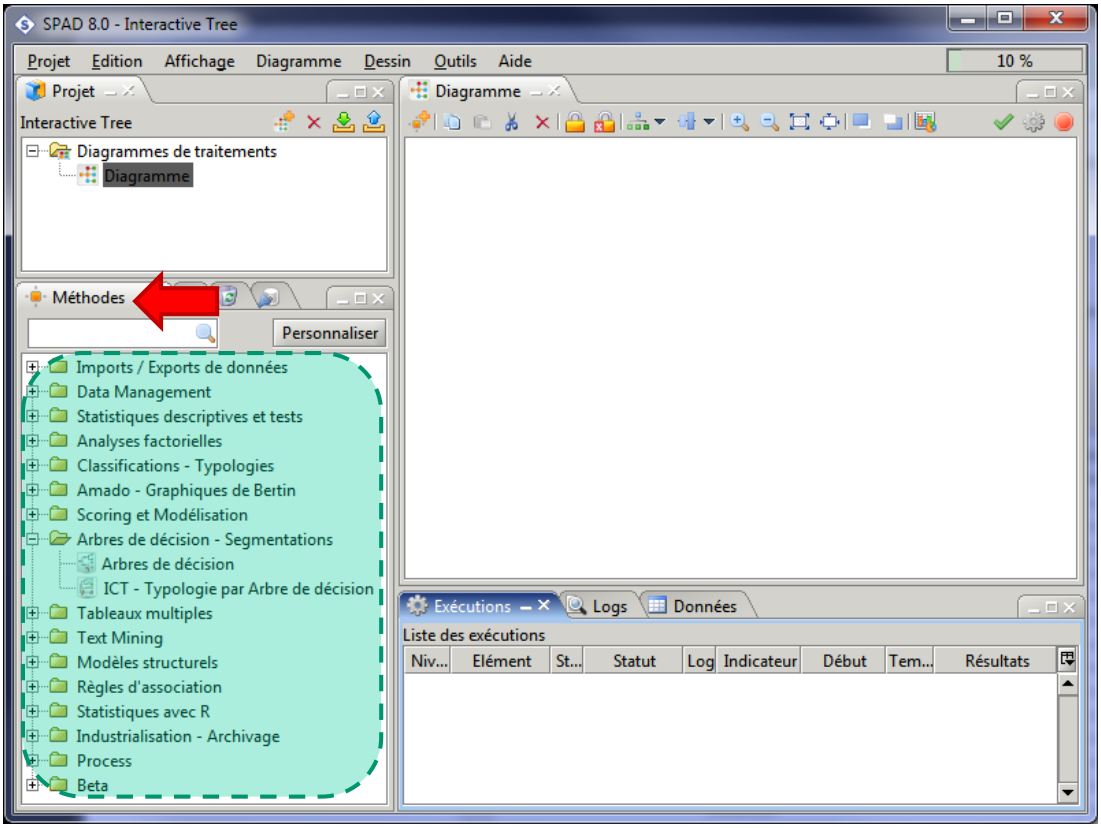
3 Analyse interactive avec SPAD 8

3.1 Démarrage du logiciel et création d'un projet

Au démarrage de SPAD, le logiciel nous demande de choisir entre ouvrir un projet existant ou en créer un nouveau. Nous sélectionnons la seconde option et nous donnons le nom de « Interactive Tree » à notre projet.

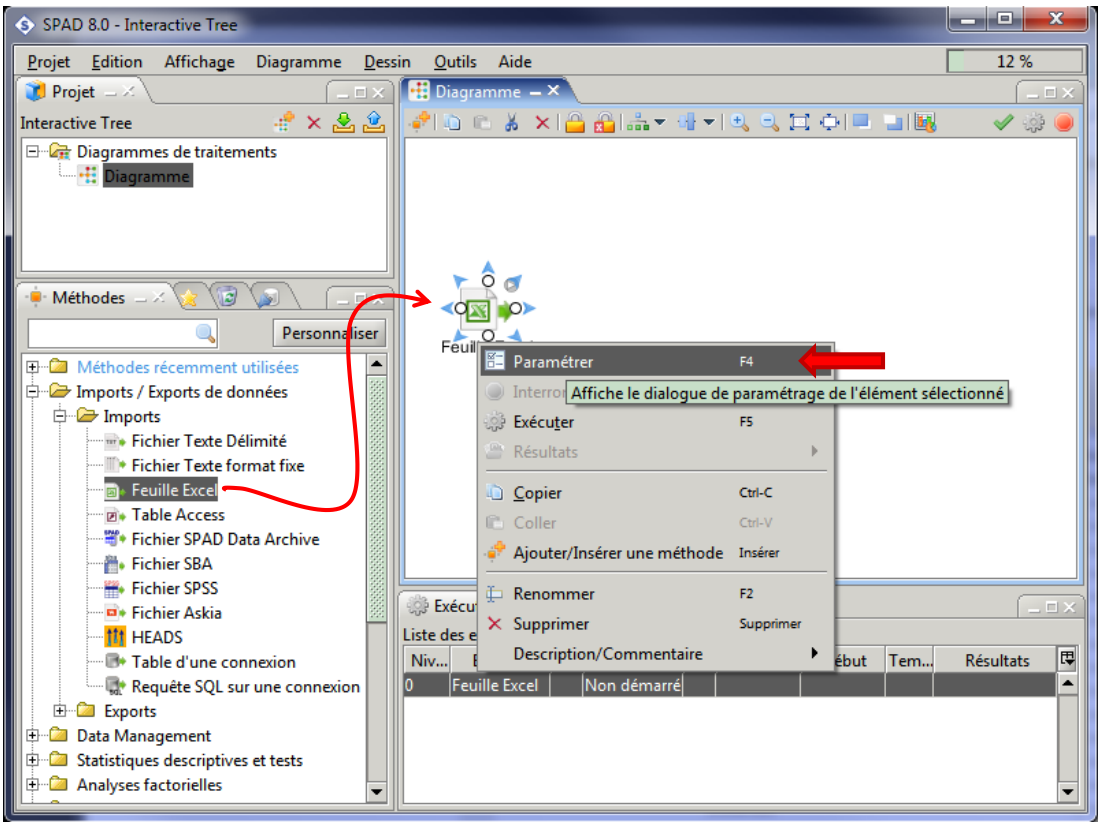


Un diagramme vide est créé. Nous pouvons définir les traitements à effectuer en utilisant les outils disponibles dans le panneau « **Méthodes** » en bas à gauche de la fenêtre principale.

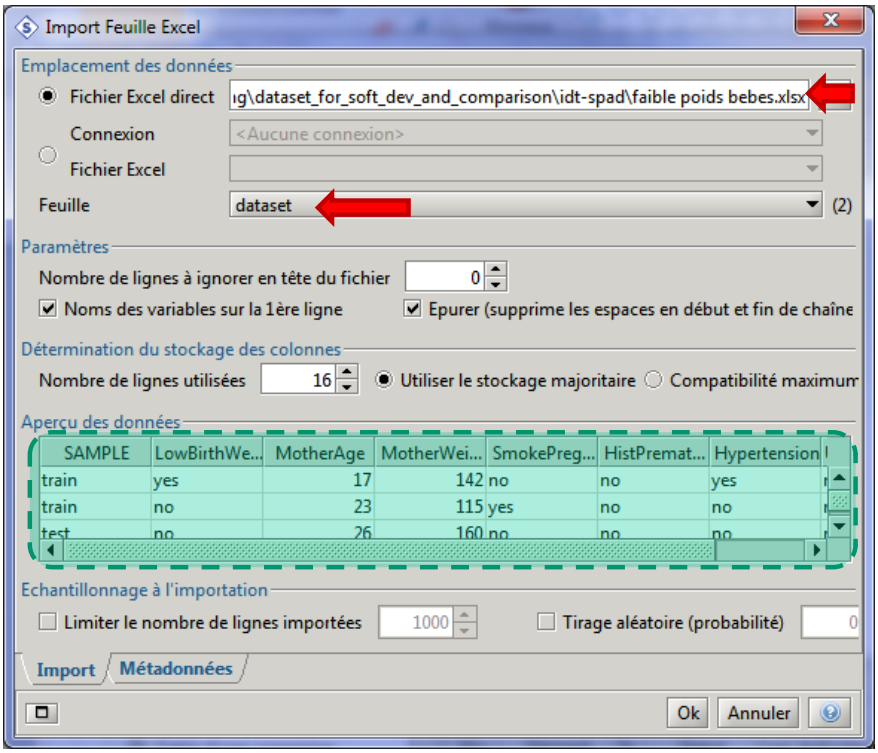


3.2 Importation des données

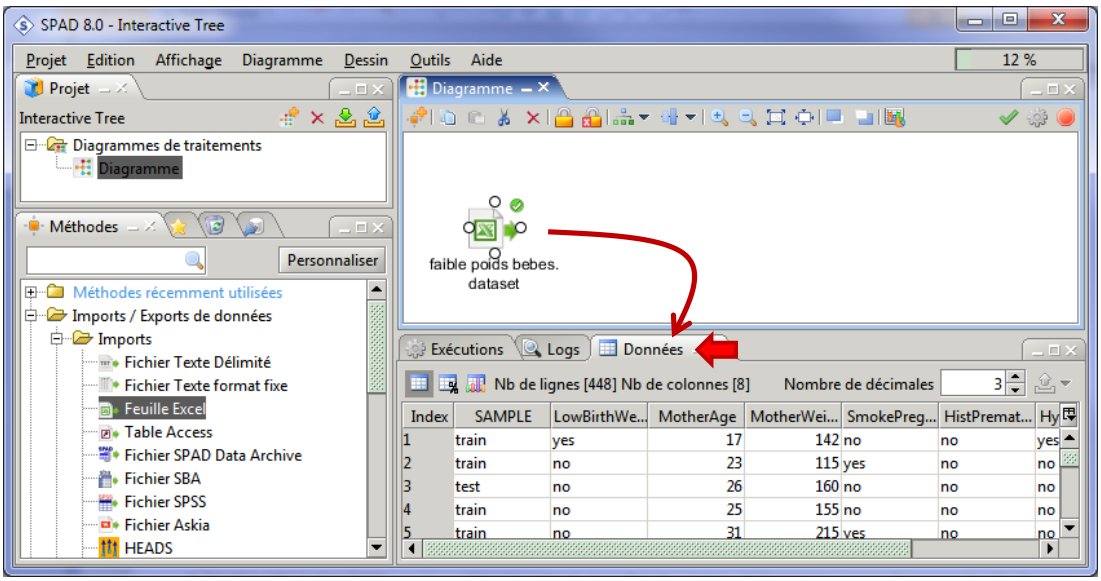
Réaliser un traitement consiste à piocher l'outil idoine dans la palette de méthodes, le placer dans l'espace de travail « **Diagramme** », le paramétrer, puis l'exécuter.



Nous insérons l'outil « **Feuille Excel** » (disponible dans la branche **Imports - Exports de données / Imports**) dans le diagramme. Sous cliquons sur l'item « **Paramétrer** » du menu contextuel. Nous sélectionnons le fichier « **faible poids bebes.xlsx** » dans la boîte de dialogue et la première feuille « **dataset** ». La prévisualisation indique que SPAD a bien identifié les noms de variables et a lu correctement les premières lignes du fichier.



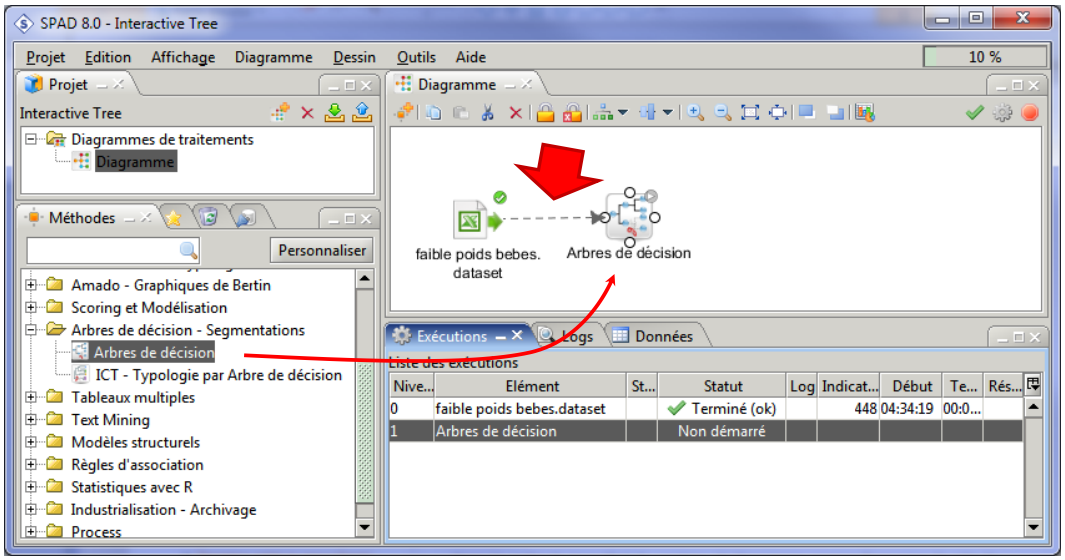
Après validation des paramètres, les données peuvent être consultées dans le panneau « **Données** » en bas à droite de la fenêtre principale (attention, il est nécessaire que l'outil soit sélectionné – symbolisé par les 4 ronds autour de l'icône – dans le diagramme).



L'importation s'est bien déroulée, avec 448 observations et 8 colonnes.

3.3 L'outil « Arbres de décision » - Paramétrage

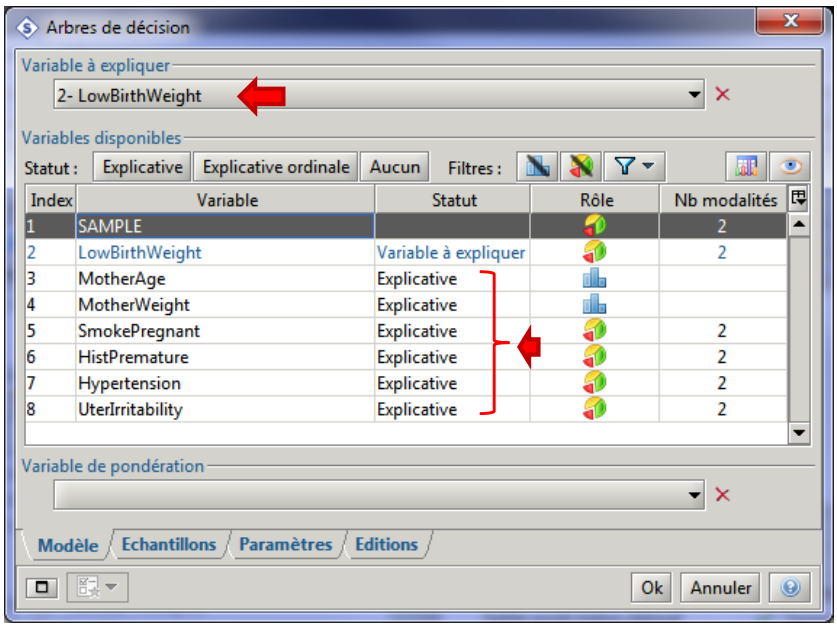
Nous trouvons l'outil « **Arbres de décision** » dans la branche « **Arbres de décision – Segmentations** » des méthodes. Nous le plaçons dans l'espace de travail et nous lui relions le composant « Fichier Excel » d'importation des données



Nous actionnons le menu contextuel « **Paramétrer** ». Nous aurons à travailler dans plusieurs onglets de la boîte de dialogue qui apparaît.

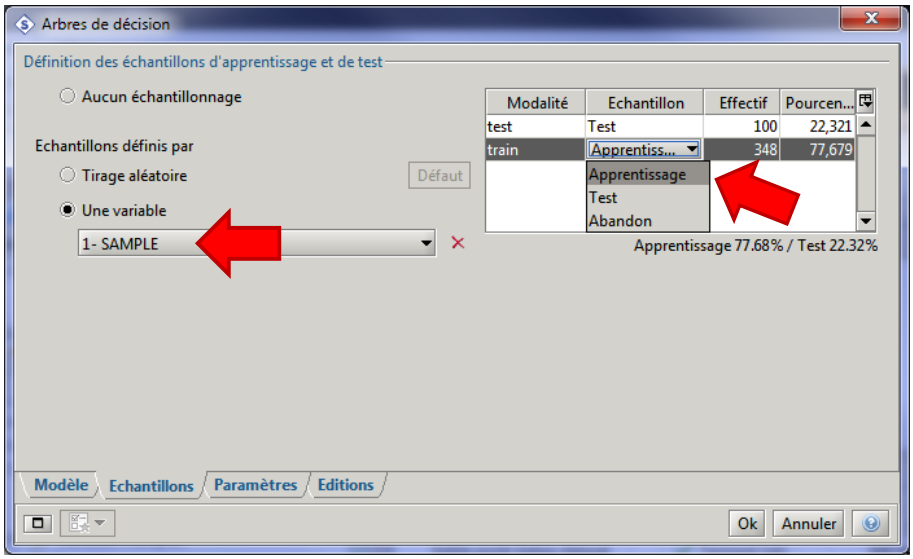
3.3.1 Paramétrage – Statut des variables

Dans « **Modèle** », nous spécifions le statut des variables. LowBirthWeight est à expliquer ; les autres constituent les explicatives. SAMPLE ne joue pas à ce stade. Notons que le type des variables est symbolisé par un histogramme pour les quantitatives, par un diagramme en secteurs (camemberts) pour les qualitatives.



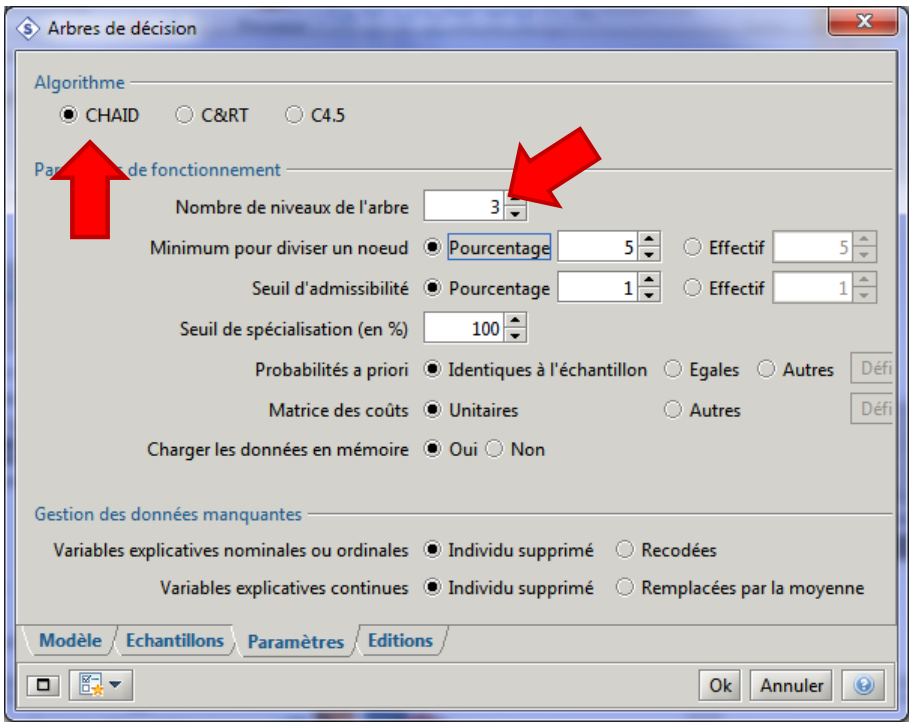
3.3.2 Paramétrage - Echantillons

Dans l'onglet « **Echantillons** » nous désignons la colonne SAMPLE comme identifiant des sous-ensembles d'apprentissage et de test. Il est possible de spécifier explicitement le rôle de chaque « modalité » de SAMPLE dans la grille de visualisation.



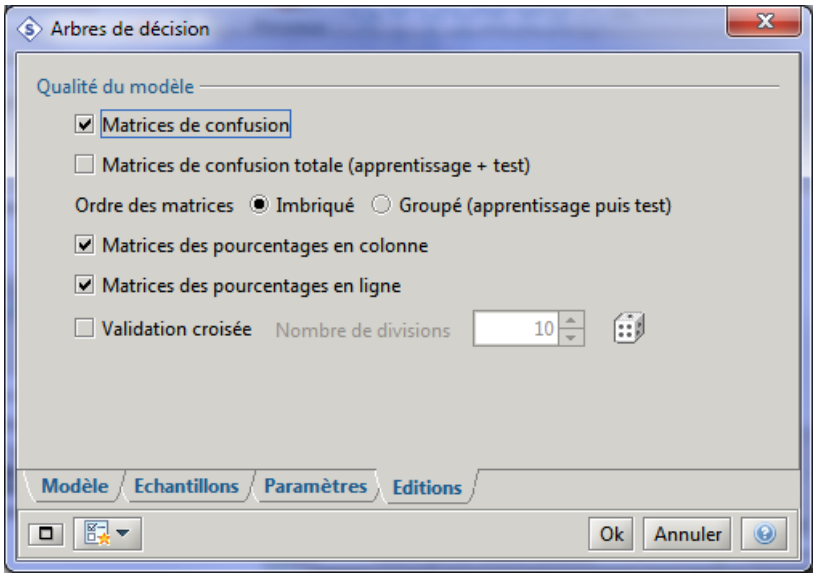
3.3.3 Paramétrage – Paramètres

L'onglet « **Paramètres** » permet de choisir les algorithmes de data mining à utiliser. Trois méthodes de référence sont disponibles : **CHAID**, C&RT (CART) et C4.5. Chacune peut être paramétrée spécifiquement. Nous choisissons de limiter – complètement arbitrairement – le nombre de niveaux de l'arbre à 3 (la racine est le niveau 0) afin d'en faciliter la lecture.



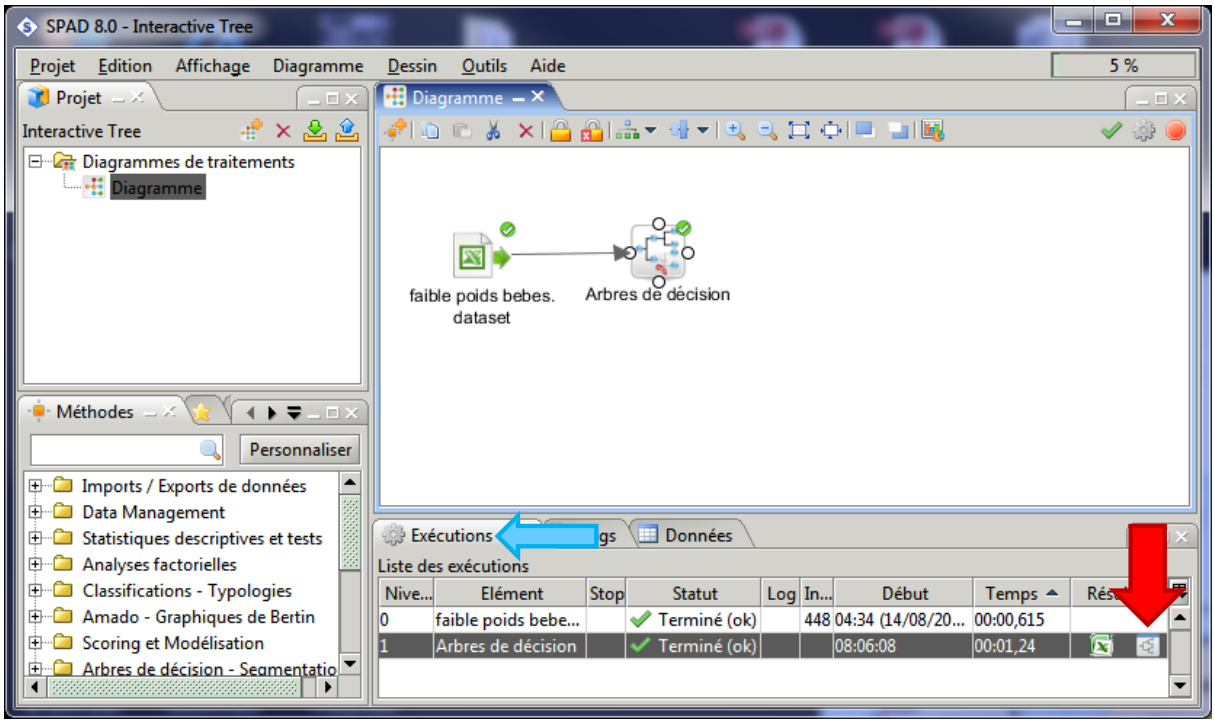
3.3.4 Paramétrage – Editions

« **Editions** » permet d'indiquer le contenu des sorties. Nous laissons les valeurs par défaut.



3.4 Affichage(s) de l'arbre

Dès validation des paramètres, les calculs sont exécutés. Les résultats sont disponibles. Nous pouvons visualiser l'arbre en cliquant sur l'icône dédié du panneau « **Exécutions** » (le menu contextuel « Résultats / Arbres interactifs » aurait fait l'affaire également).



La fenêtre d'affichage de l'arbre apparaît avec 2 onglets : « **Graphique** » et « **Rapport** ». Pour l'instant, intéressons-nous au premier et inspectons l'arbre. Les sommets sont numérotés pour faciliter la lecture.

3.4.1 L'onglet « Graphiques »

Le modèle propose 7 règles puisque que nous avons obtenu 7 feuilles. Elles correspondent aux sommets : 7, 8, 9, 10, 11, 12, 6. Concernant ce dernier par exemple, la règle s'écrit :

(Feuille 6) **SI** SmokePregnant = yes **ET** HistPremature = yes **ALORS** LowBirthWeight = yes

Sur la partie droite nous disposons d'informations détaillées sur le sommet sélectionné, en l'occurrence la racine dans la copie d'écran ci-dessous.

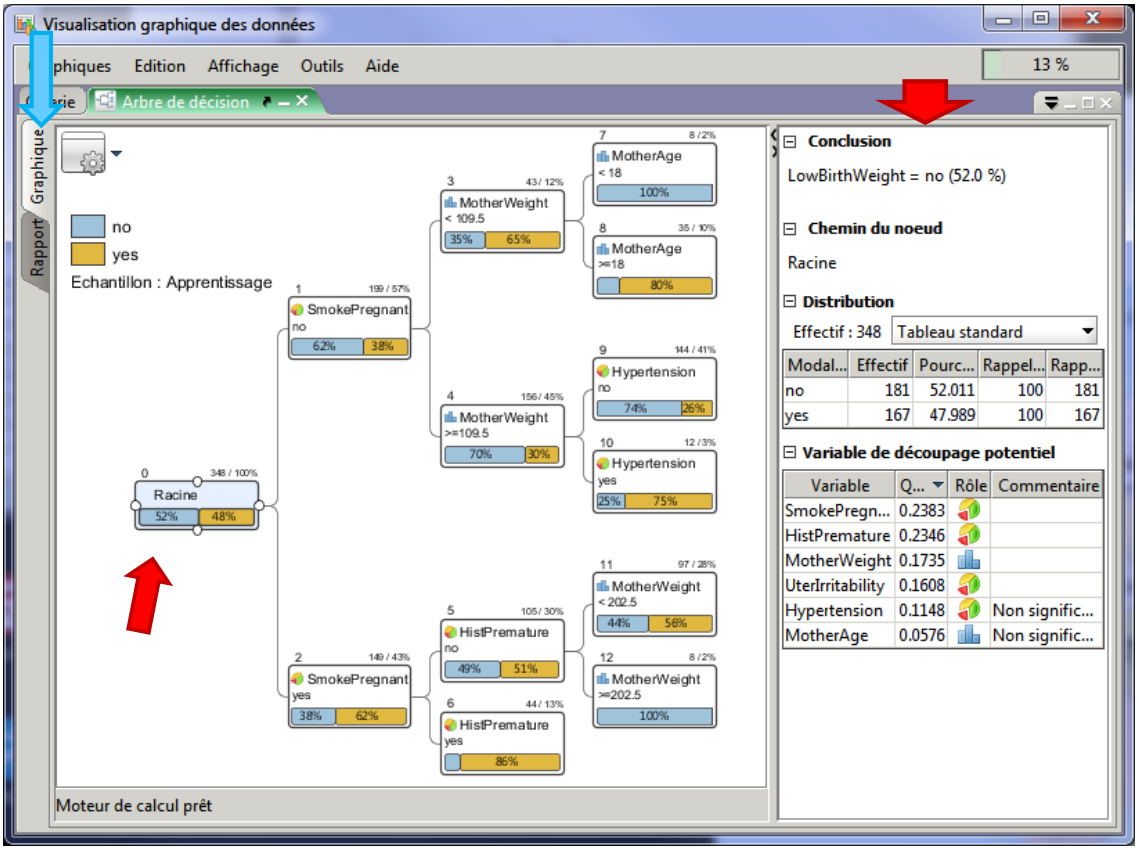
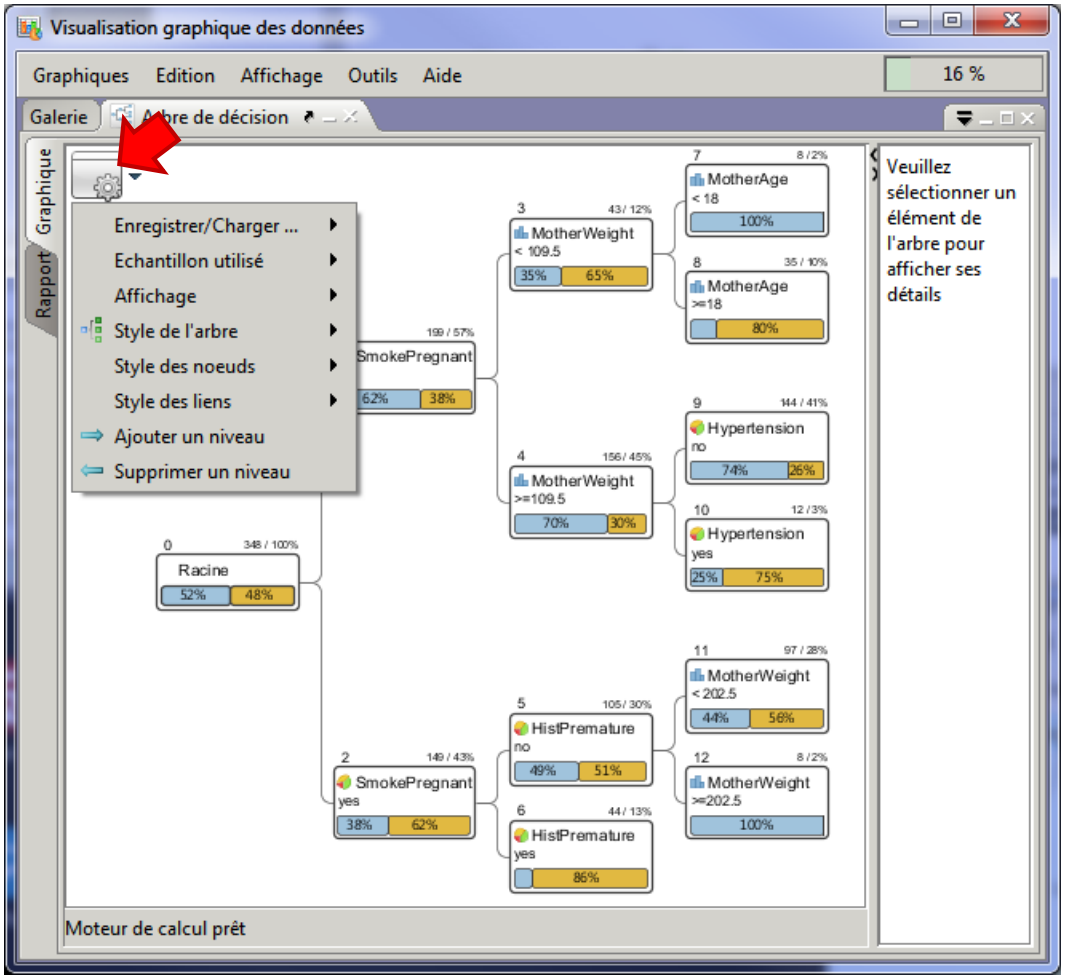


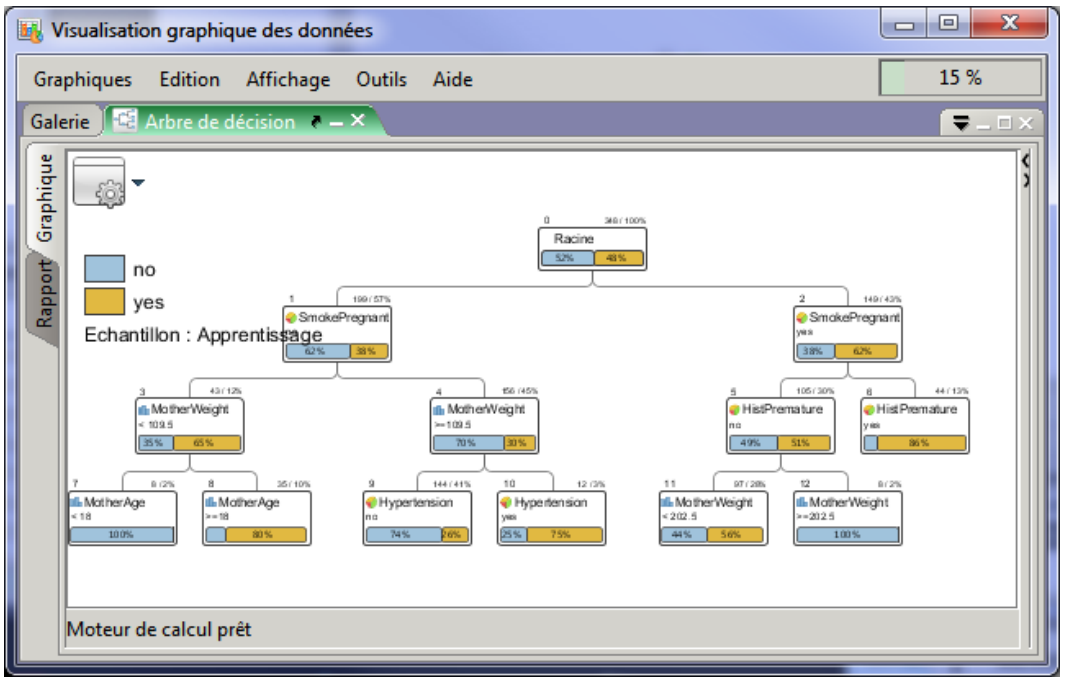
Figure 1 - Vue globale de l'arbre - Informations sur la racine - Echantillon d'apprentissage

Nous observons dans la section « **Distribution** » qu'il y a bien 348 observations dans l'échantillon d'apprentissage avec 181 « no » (52%) et 167 « yes » (48%). Dans « **Variable de découpage potentiel** », la plus performante est « SmokePregnant, que l'arbre a utilisé d'ailleurs pour segmenter la racine, avec une qualité (t de Tschuprow) de 0.2383. Nous notons que « HistPremature » est quasiment équivalente (t = 0.2346), et que nous pourrions le substituer à « SmokePregnant » pour le traitement de la racine. Nous reviendrons sur le thème de la construction interactive de l'arbre plus bas (section 3.6).

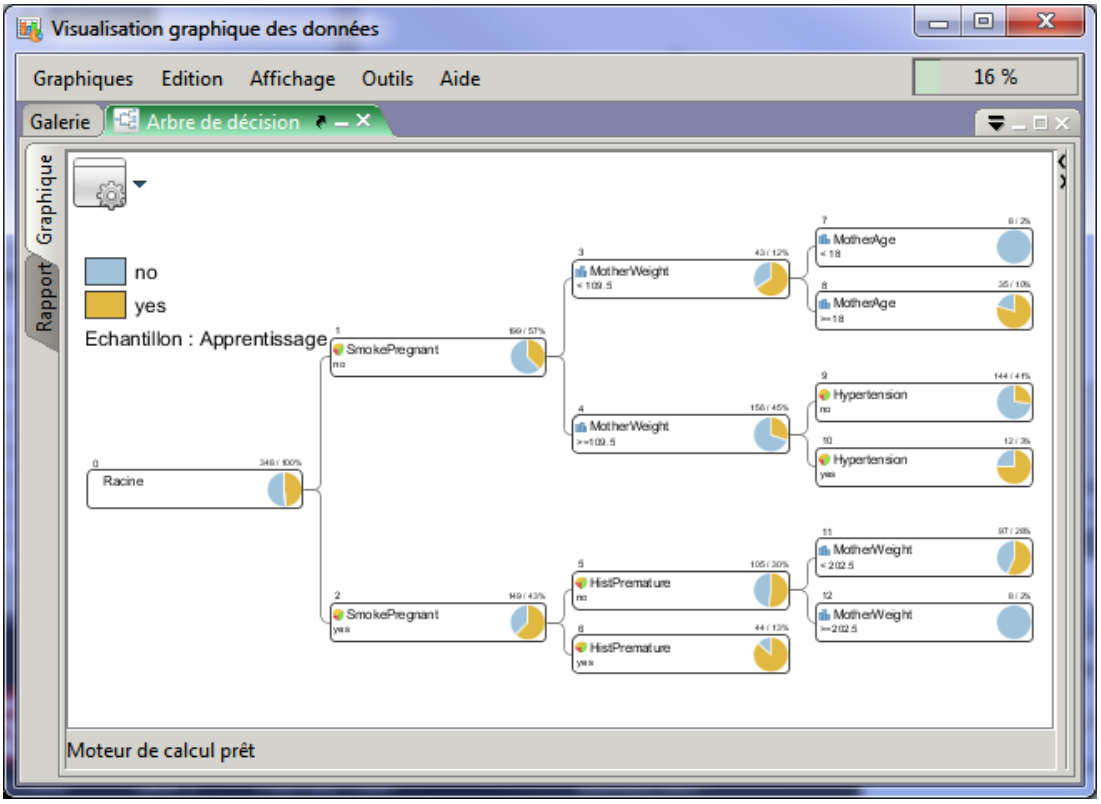
Plusieurs options d'affichage sont proposées via l'outil situé en haut et à gauche de la fenêtre. Certaines sont essentiellement cosmétiques. D'autres sont pertinentes pour une meilleure appréhension des résultats.



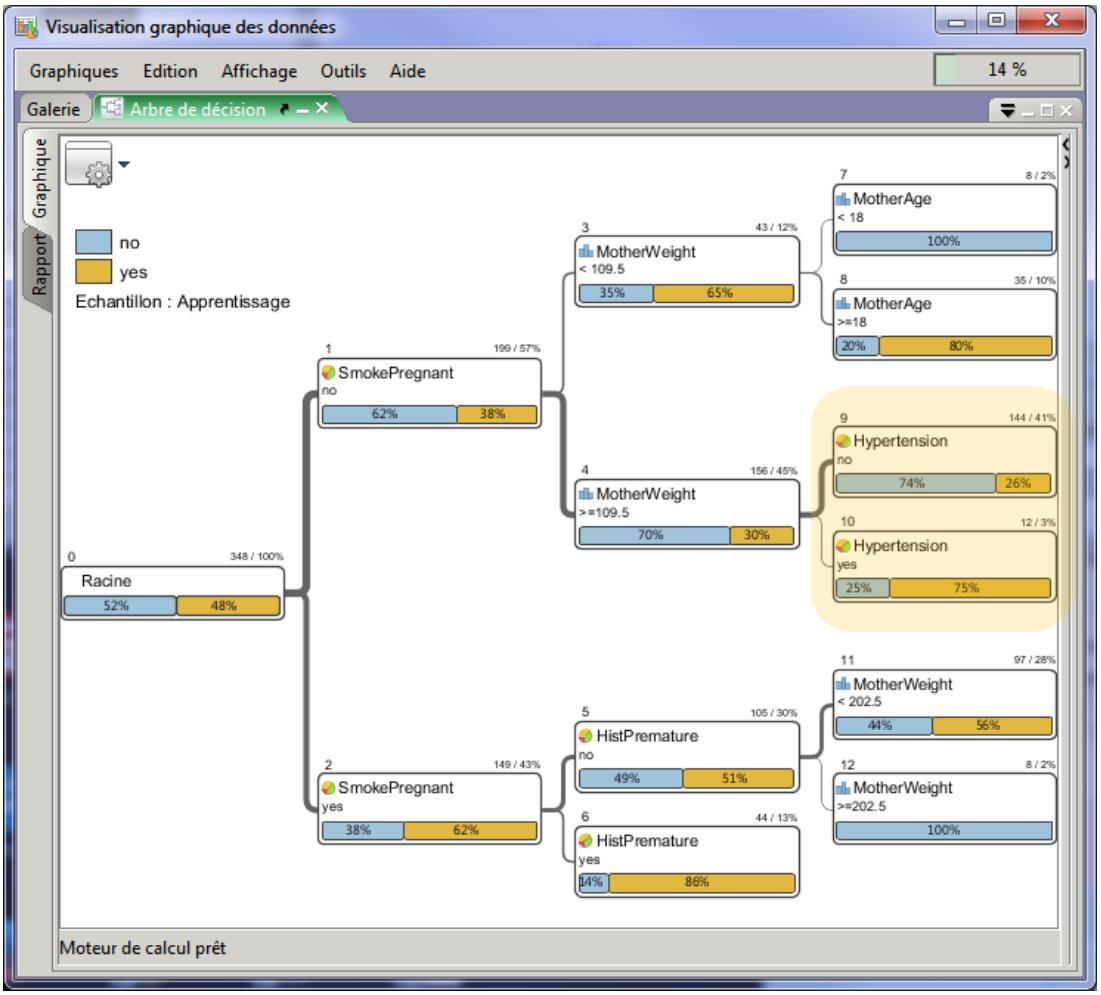
Style de l'arbre – Arbre bas. L'arbre s'affiche de haut en bas.



Style des nœuds – Secteur. Un camembert remplace l'histogramme empilé.



Style des liens – Epaisseur des liens en fonction des effectifs.



La fiabilité d'une règle dépend de son support. Avec cette option, nous pouvons évaluer en un coup d'œil les sommets qui couvrent une portion significative de la population. Par exemple, les règles issues sommets 9 et 10 semblent équivalentes en terme de confiance (~ 75% de précision). En revanche, la première (feuille 9) couvre 144 individus (41% de 448, taille de l'échantillon d'apprentissage), tandis que la seconde (feuille 10) ne concerne que 12 individus (3%). L'épaisseur du trait permet de jauger cette différence immédiatement.

Echantillon utilisé – Echantillon de test. Cette option est fondamentale. On sait qu'un des défauts des arbres est de produire des règles sur-spécialisées sur l'échantillon d'apprentissage, faussement précises, surtout lorsque l'on traite des petits effectifs. Avec cet outil, nous pouvons évaluer la crédibilité (la reproductibilité) des pourcentages à l'aide d'une fraction des données – l'échantillon test – qui n'a pas contribué directement à la construction du modèle. Les informations fournissent un contrepoint très utile à l'arbre initial (Figure 1).

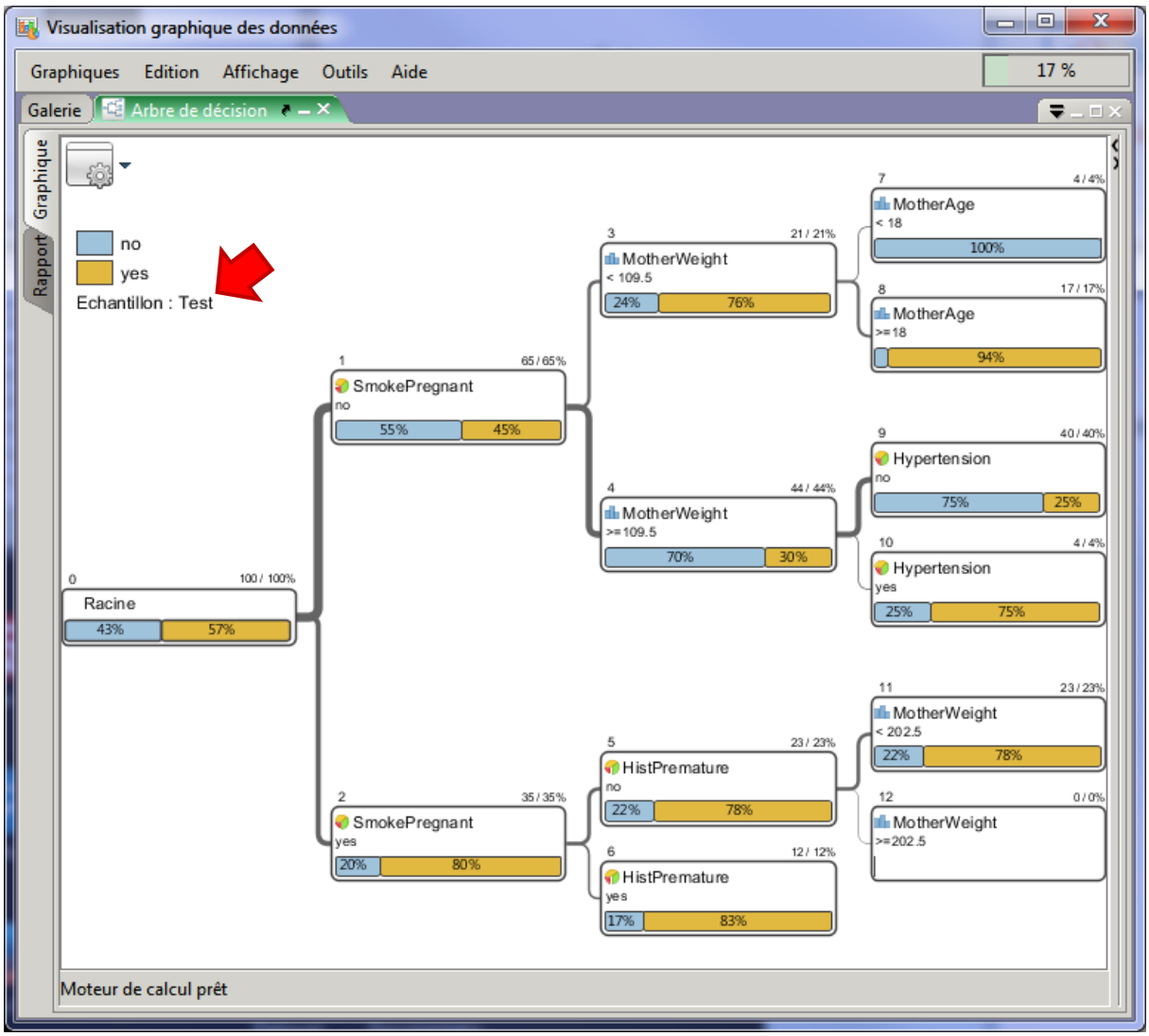
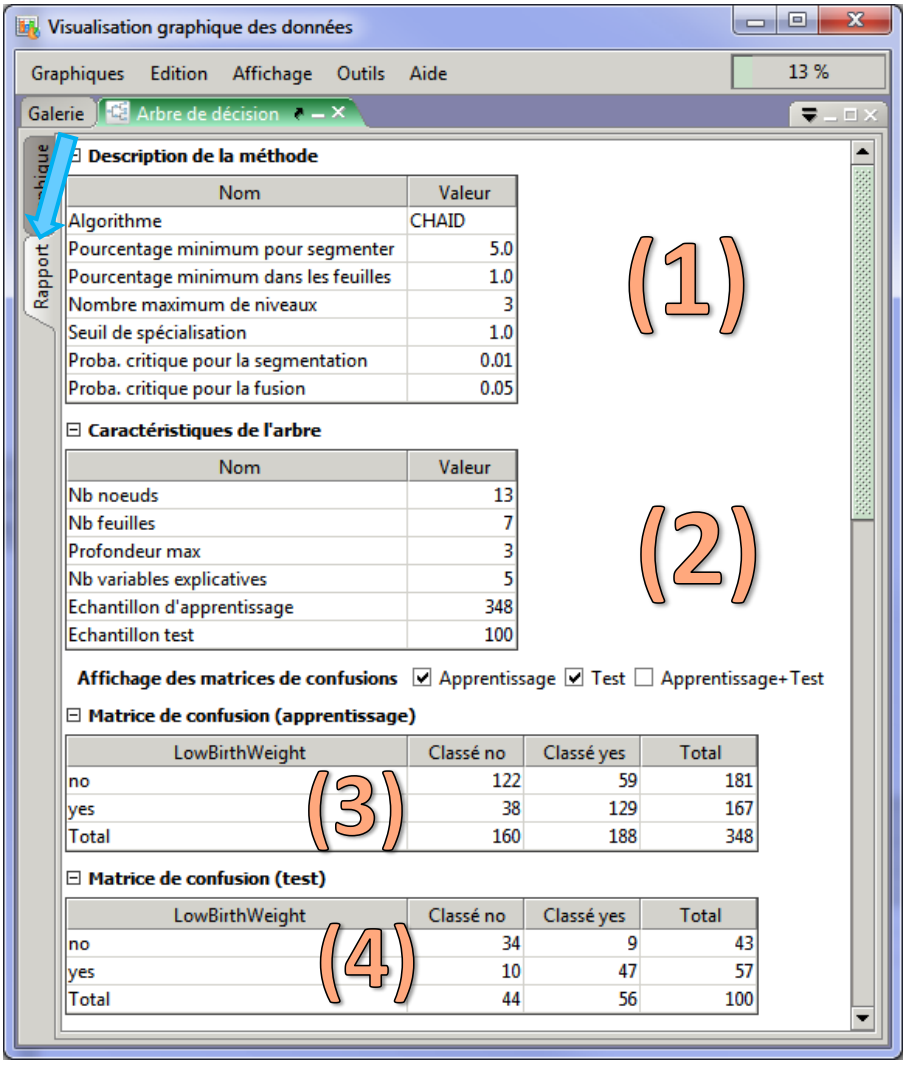


Figure 2 - Vue globale de l'arbre - Echantillon test (à comparer avec Figure 1)

Nous observons qu'il n'y a pas de contradictions notoires – en termes de proportions des classes dans les sommets – par rapport à l'arbre de l'échantillon d'apprentissage (mis à part le sommet 12 qui se révèle vide, cela n'est pas choquant dans la mesure où l'effectif de l'échantillon test est relativement faible, 100 observations). Le modèle semble refléter des phénomènes qui existent réellement dans la population. Si les deux arbres étaient très différents, là il faudrait s'inquiéter. Cela voudrait dire que l'arbre élaboré est trop spécifique aux données d'apprentissage. Soit parce que les échantillons ont été mal construits, dans ce cas il faut revenir au début de l'étude ; soit parce que l'arbre est trop spécialisé, dans ce cas les dissensions apparaissent surtout dans les parties basses (vers les feuilles), il faudrait réduire sa taille par post-élagage (par suppression des feuilles incriminées).

3.4.2 L'onglet « Rapport »

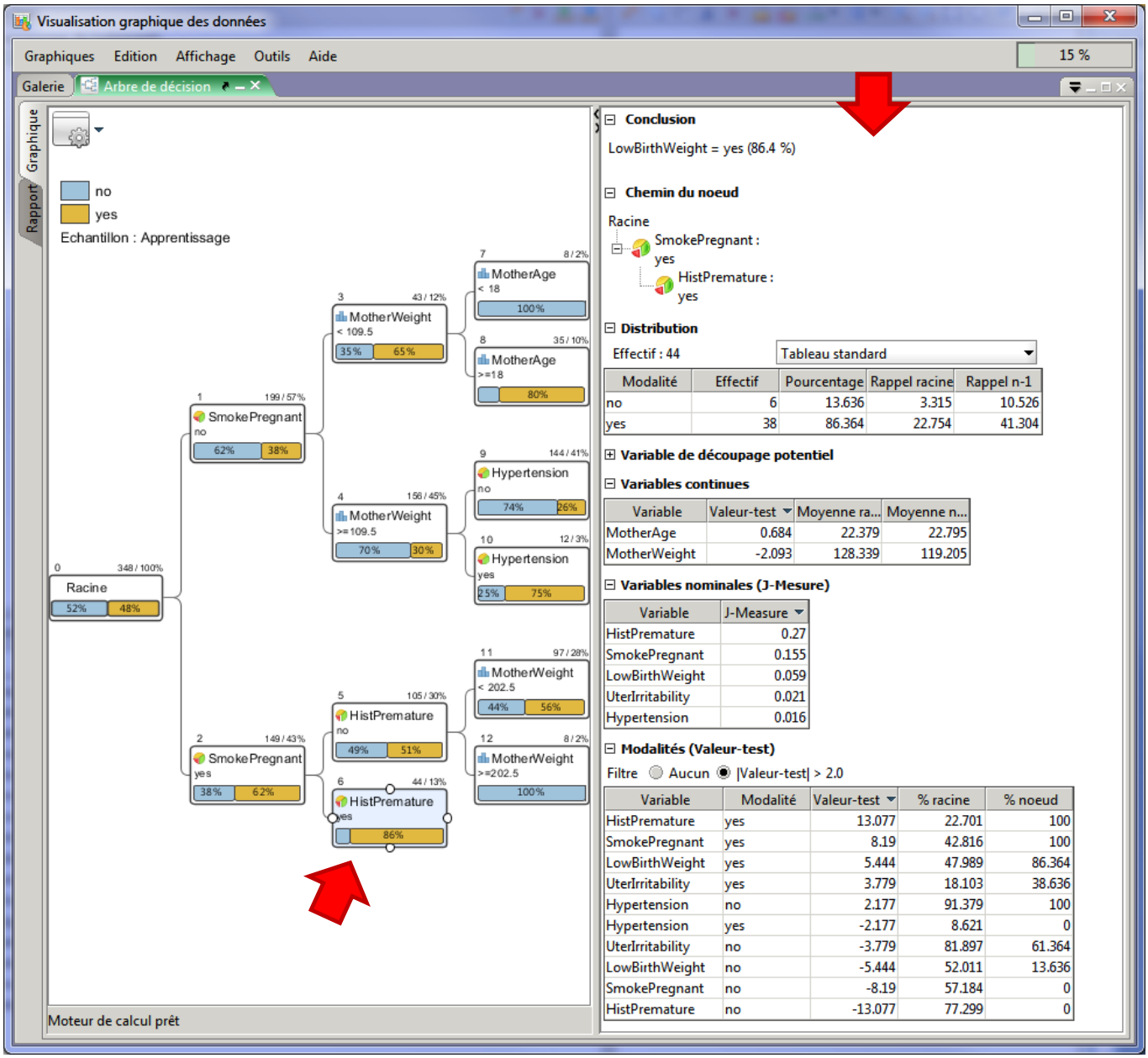
Dans l'onglet « **Rapport** » sont affichés – entres autres – les paramètres de la technique utilisée « Description de la méthode » (1), les « Caractéristiques de l'arbre » obtenu (2), les « Matrices de confusion » en apprentissage (3) et en test (4).



L'arbre comporte 7 feuilles, 5 variables ont été utilisées parmi les 6 disponibles, le taux d'erreur en test (affiché plus bas dans la fenêtre) est de (10 + 9) / 100 = 19%.

3.5 Exploration locale d'un sommet

Tout le sel de l'outil réside dans la possibilité d'explorer finement les sous-populations circonscrites par les sommets de l'arbre. Intéressons-nous à la règle établie par la feuille **n°6**. Elle s'écrit : **SI** SmokePregnant = yes **ET** HistPremature = yes **ALORS** LowBirthWeight = yes. Est-ce qu'on s'en tient uniquement à ces informations ? Est-ce qu'il est possible d'aller plus loin dans la caractérisation de la sous-population ainsi définie ? Qu'en est-il justement des autres variables ? Le panneau de droite fournit les informations permettant d'affiner l'analyse. Détaillons-en le contenu.



Conclusion indique la prédiction associée à la règle « LowBirthWeight = Yes ».

Chemin du nœud décrit la règle en reconstituant le chemin partant de la racine à la feuille : « SmokePregnant = yes » tout d'abord, « HistPremature = yes » ensuite.

Distribution indique les fréquences absolues et relatives des classes.

Les tableaux suivants permettent d'approfondir l'interprétation de la sous-population et de la règle associée, en précisant le rôle des variables qui n'apparaissent pas explicitement dans le chemin. Pour les **variables continues**, l'analyse s'appuie sur une comparaison de moyennes entre la population initiale représentée par la racine de l'arbre, et la sous-population définie par le sommet ciblé. Concernant « MotherWeight » par exemple, nous notons que le poids moyen dans la population globale est de **119.205** (livres, j'imagine), alors que dans la sous-population associée à la feuille n°6, il est de **128.339**. Ces personnes sont en surpoids visiblement selon l'indicateur « valeur-test » qui correspond, en simplifiant, à une statistique de test de comparaison de moyennes pour échantillons imbriqués[5]. La valeur-test est distribuée approximativement selon une loi normale. A un risque de 5%, nous considérerons l'écart significatif lorsqu'il est supérieur à 2 en valeur absolue. En pratique, on s'intéressera surtout aux forts écarts c.-à-d. les parties hautes et basses du tableau.

Le tableau suivant traite des **variables nominales**. Il indique les disparités des distributions de fréquences globales (racine de l'arbre) et locales (sommet sélectionné) à l'aide de l'indicateur J-Mesure (J-Measure en anglais). Plus la valeur est grande, plus forte sont les disparités. Une valeur nulle correspond à une équivalence des distributions. Il n'y a pas de seuil. Il faut avant tout se concentrer sur les parties hautes du tableau. Nous notons sans surprise ici que les variables « SmokePregnant » et « HistPremature » sont en bonne position ici puisqu'elles définissent le chemin. La cible « LowBirthWeight » est assez bien discriminée puisque nous passons d'une répartition de (52% : no, 48% : yes) sur la racine, à (14%, 86%) sur le nœud (J-Mesure = 0.059). Il faudrait par la suite se pencher sur le cas de « UterIrritability ».

Le tableau des **Modalités** sert justement à préciser les sur ou sous représentations. Les positions des 3 premières modalités sont sans surprises : 100% de « HistPremature = yes » correspond au sommet analysé, leur proportion était de 22.701% sur la racine de l'arbre ; le schéma est le même pour « SmokePregnant = yes » (100% vs. 42.816%) ; la situation de

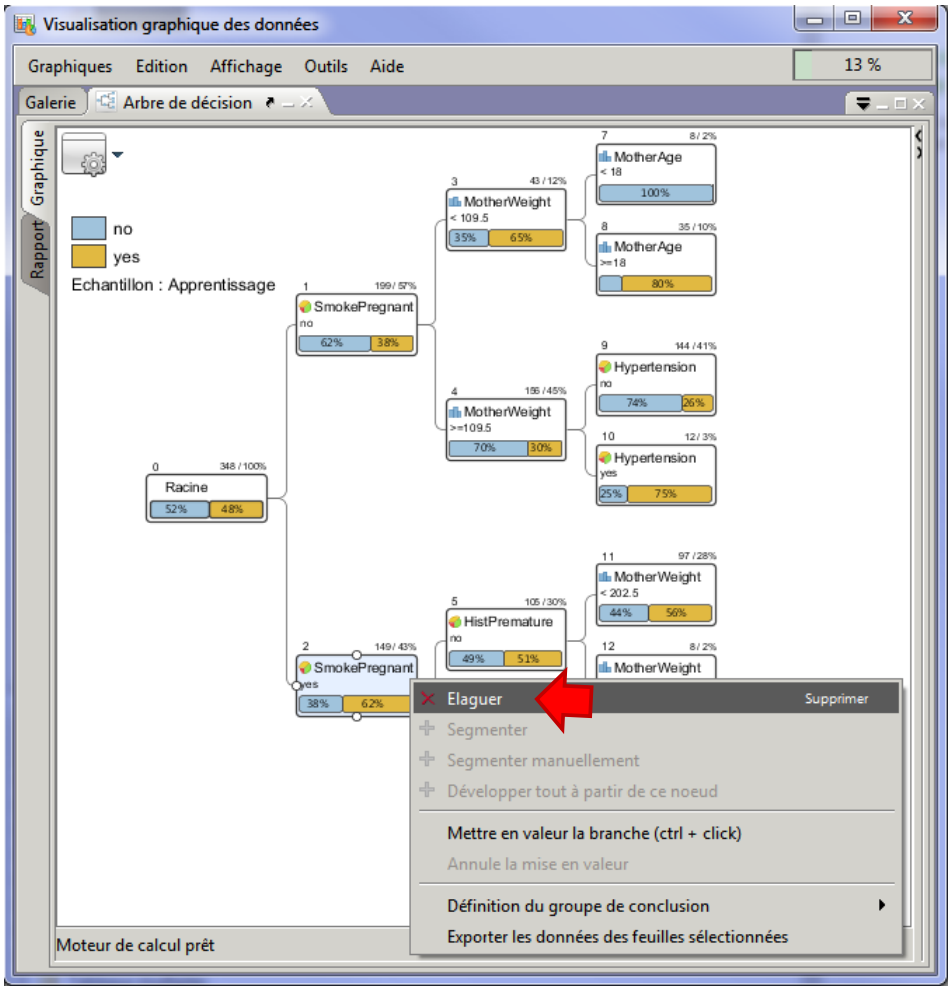
[5] Voir http://tutoriels-data-mining.blogspot.fr/2008/04/interprter-la-valeur-test.html ou http://tutoriels-data-mining.blogspot.fr/2008/03/statistiques-descriptives-comparatives.html ; voir aussi et surtout la copie de l'article original, http://www.deenov.com/analyse-de-donnees/article-valeur-test.aspx

« LowBirthWeigth = yes » (86% vs. 48%) était directement visible dans l'affichage graphique de l'arbre. Nous constatons de plus, et ça ce n'était pas visible directement, que cette sous population se distingue également par une **sur représentation** des modalités « UterIrritability = yes » (38.636% vs. 18.103% initialement) et « Hypertension = no » (100% vs. 91.379% sur la racine). Si nous ne disposons pas de cet outil, ces informations sont ignorées, et la caractérisation de la sous-population est incomplète. Des éléments de compréhension importants de la formation de la règle sont laissés de côté. La **valeur test** des modalités correspond à une statistique de comparaison de proportions.

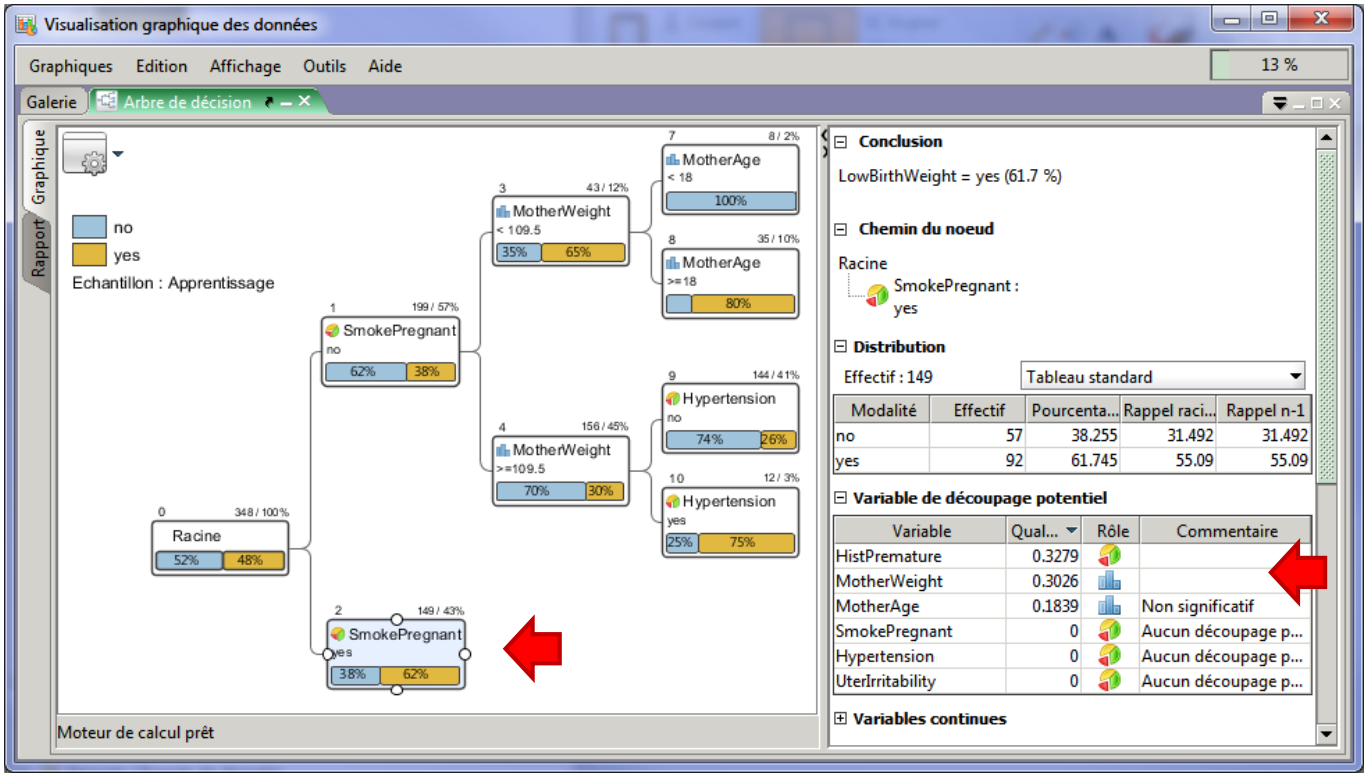
3.6 Construction interactive de l'arbre

Nous aspirons à guider nous-mêmes la construction de l'arbre, tout en conservant une certaine cohérence avec les indications numériques fournies par l'outil bien évidemment.

3.6.1 Elagage d'une partie de l'arbre

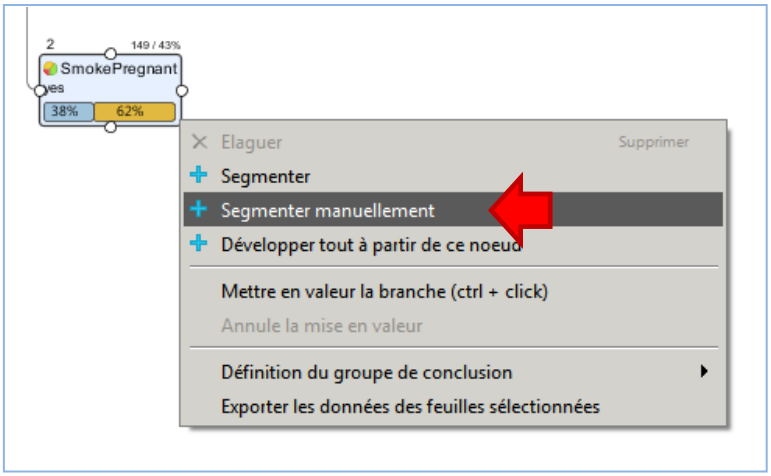


Mettons que nous souhaitons reconsidérer la partie située **en aval du sommet n°2**. Nous devons dans un premier temps supprimer les branches existantes. Pour ce faire, nous sélectionnons le nœud et nous actionnons le menu « Elaguer ». Nous obtenons.



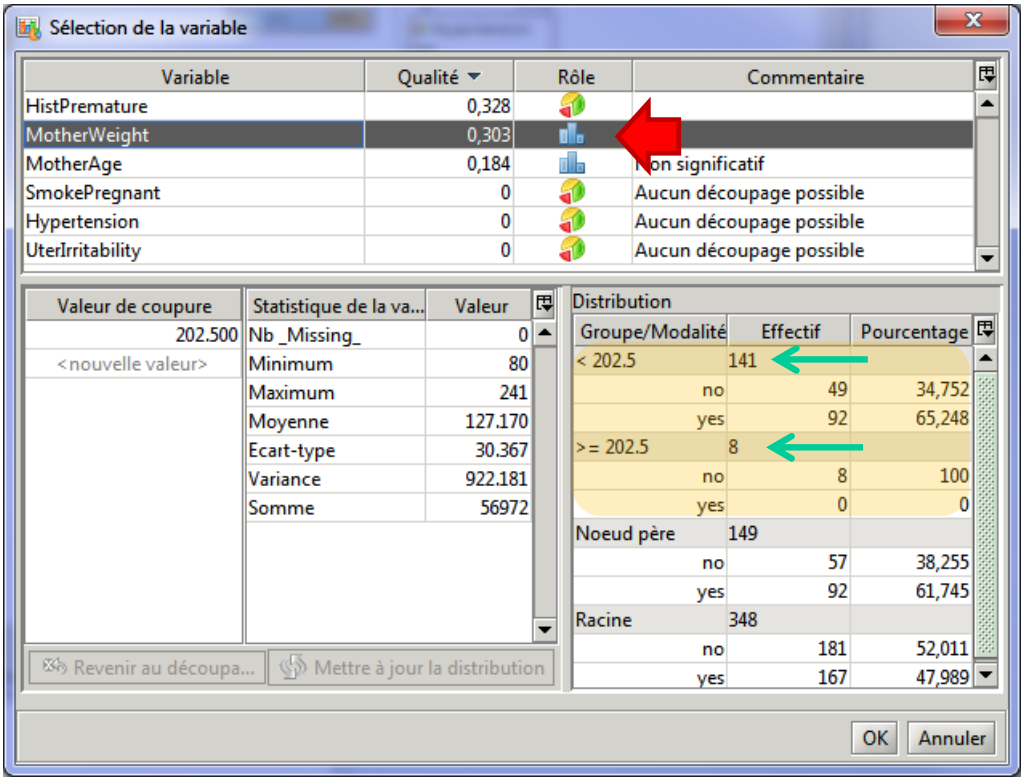
3.6.2 Choix de la variable de segmentation

Dans la copie d'écran ci-dessus, nous constatons (cf. « **Variable de découpage potentiel** ») que « HistPremature » (initialement sélectionnée) et « MotherWeight » sont les seules pertinentes pour la segmentation du sommet. Nous souhaitons exploiter cette dernière. Nous actionnons le menu contextuel et nous cliquons sur « Segmenter Manuellement »[6].



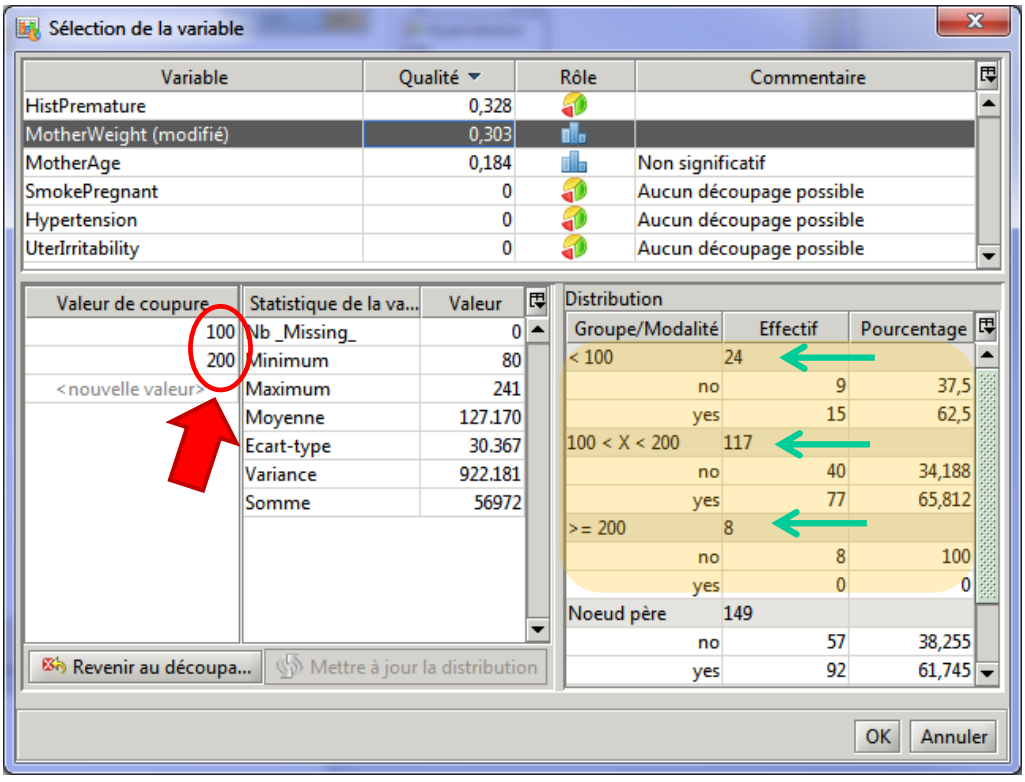
Une boîte de dialogue apparaît, énumérant les variables disponibles et les caractéristiques des découpages associés. Si nous choisissons « MotherWeight » par exemple, nous obtiendrions deux feuilles avec respectivement 141 et 8 observations.

[6] Cliquer sur « **Segmenter** » tout court reviendrait à sélectionner la variable présentant la qualité la plus élevée c.-à-d. « HistPremature ».

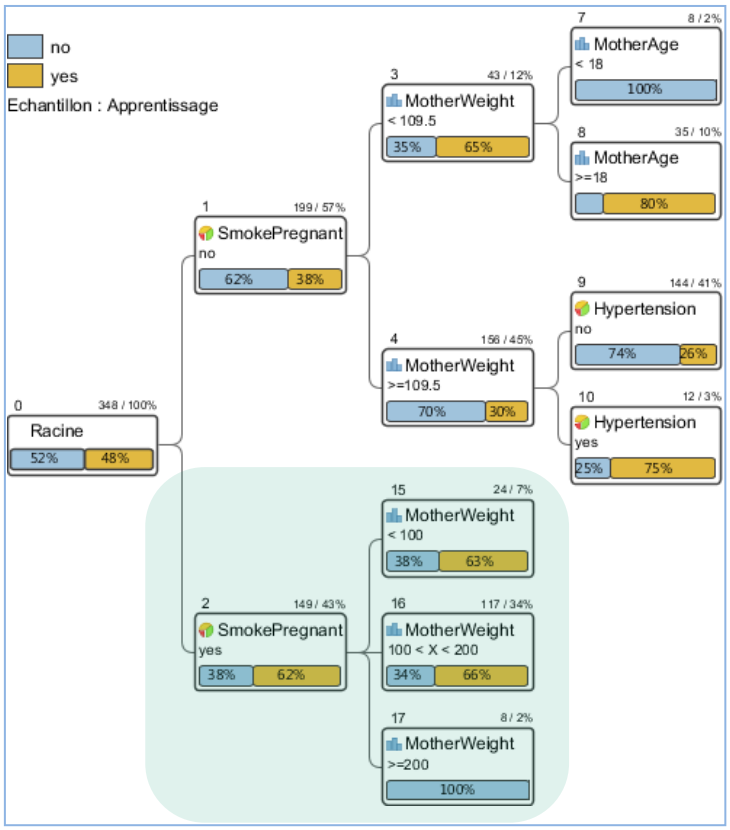


3.6.3 Modification des caractéristiques de la segmentation

Nous voulons utiliser cette variable mais avec nos propres bornes de découpage. Nous les introduisons dans la partie « **Valeur de coupure** ». Il est possible de produire plusieurs (K > 2) feuilles en saisissant plusieurs seuils (K – 1). Nous essayons les valeurs **100** et **200**.



Nous obtenons le modèle suivant après validation.

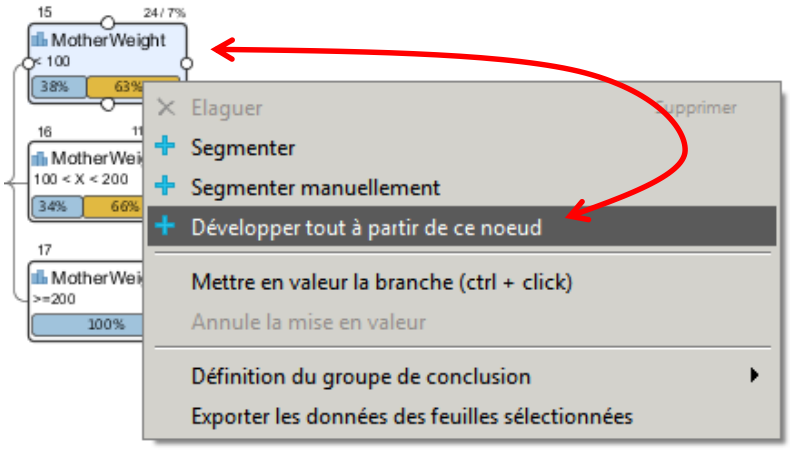


Notre arbre n'est pas très usuel certes. L'opération démontre surtout la souplesse de l'outil.

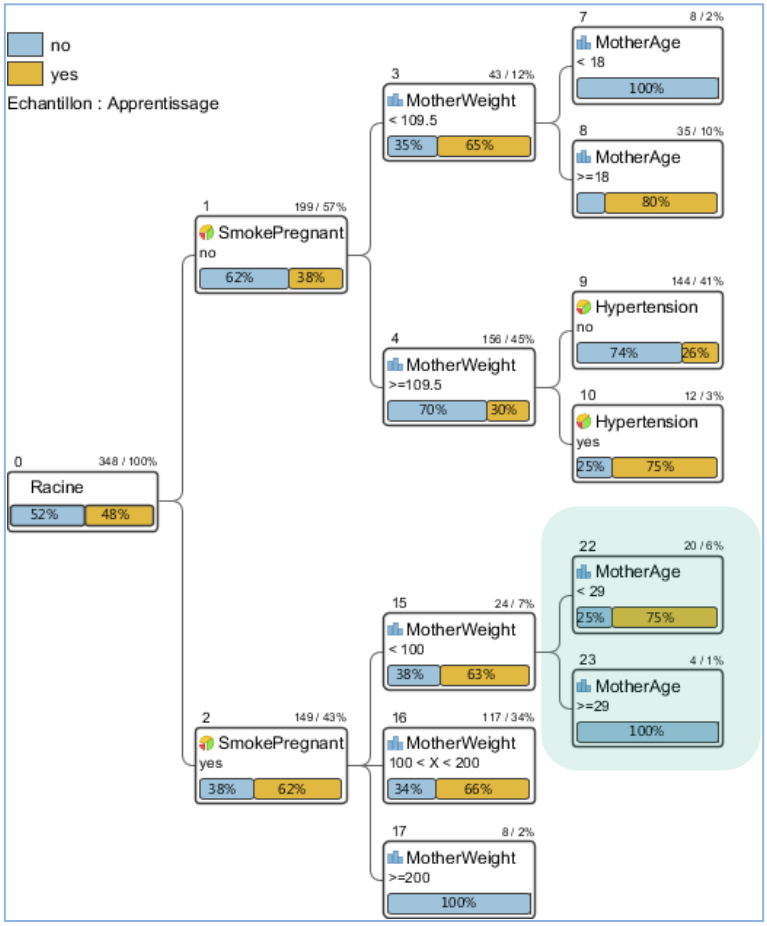
Remarque : Pour les **variables de segmentation qualitatives**, il est possible de modifier les regroupements, voire de procéder à un regroupement en (K > 2) groupes.

3.6.4 Développement automatique à partir d'un nœud

A partir de ce stade, une première possibilité serait de construire manuellement le reste de l'arbre en segmentant en proche en proche chaque feuille générée. Le procédé peut être très vite fastidieux. L'autre solution consiste à demander à SPAD de développer les branches à partir d'une feuille (temporaire pour le coup) quelconque. Par exemple, nous aimerions qu'un sous-arbre soit développé à la suite de la **feuille n°15**. Nous actionnons le menu contextuel.



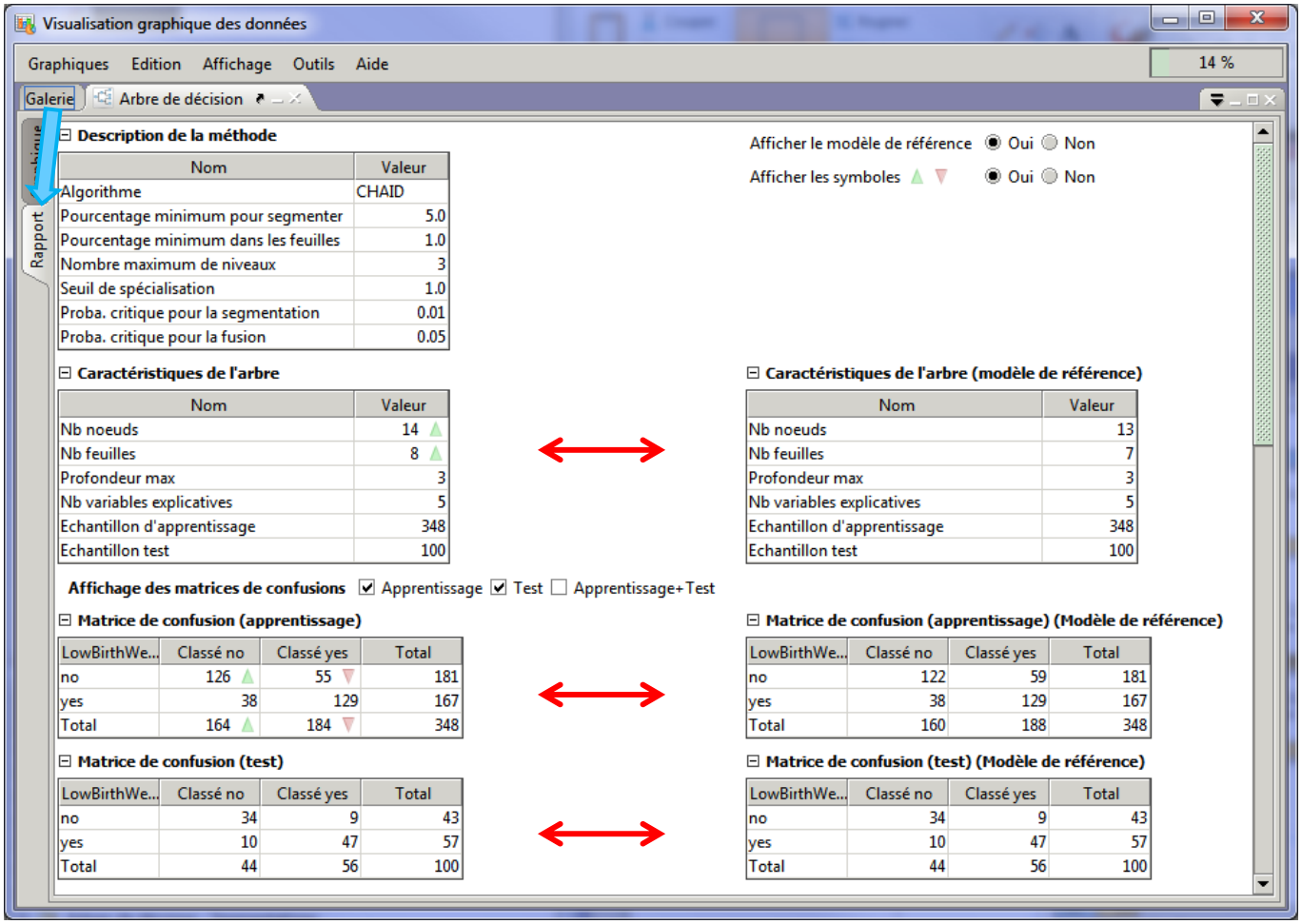
Deux feuilles sont produites finalement. Si d'autres segmentations avaient été possibles à la suite des sommets n°22 et n°23, elles auraient été concrétisées également.



3.6.5 Confrontation des modèles dans l'onglet « Rapport »

Manipuler un arbre pour obtenir des résultats qui nous conviennent est très facile comme nous pouvons le constater. Encore faut-il que le résultat tienne la route. La seule manière de le savoir est d'en évaluer les performances sur l'échantillon test dans l'onglet « **Rapport** ».

Nous constatons ci-dessous que, par rapport à l'arbre construit initialement par la méthode CHAID (Figure 1), le nouveau modèle est plus complexe avec une règle supplémentaire (8 feuilles vs. 7). Il semble plus performant **sur l'échantillon d'apprentissage** avec une sensibilité équivalente (« yes » est la modalité positive de la variable cible « LowBirthWeight », nous avons 129/181 vs. 129/181) et une meilleure précision (129/184 vs. 129/188). En réalité, les deux modèles sont totalement équivalents dixit la matrice de confusion construite **l'échantillon test**. Moralité, l'introduction de la variable « MotherWeight » au sommet n°2 et le découpage en 3 intervalles n'a pas été préjudiciable a la qualité de prédiction. A la seule différence que, mais ça seul l'expert de la spécialité peut nous le dire, l'arbre est peut être plus en adéquation avec les connaissances du domaine.



Remarque : Par expérience, j'ai remarqué que la construction manuelle des arbres – avec le choix des variables de segmentations « non-optimales » au sens du critère de qualité sur les sommets – peut conduire à des modèles apparemment bons en apprentissage, mais qui s'avèrent en réalité désastreux en déploiement. Il faut absolument être très prudent dans cette démarche, et **surveiller constamment le comportement sur l'échantillon test**.

4 Conclusion

Les arbres de décision interactifs sont précieux dans le processus de fouille des données. Ils permettent d'explorer finement les relations entre les variables dans les sous-populations. Nous pouvons allier, avec une grande facilité d'utilisation, une analyse guidée à la fois par des critères numériques (mesures d'association, comparaison de moyennes et de proportions, valeurs tests) et par l'expertise du domaine (priorisation des variables à qualité similaire, choix des seuils et des regroupements). Curieusement, peu de supports de cours académiques leurs sont consacrés.