1 Objectif

SQL Server Data Mining Add-Ins (incluant Data Mining Client pour Excel).

Excel – le tableur de manière général – est très populaire auprès des « data scientist » [1]. Mais, s'il est effectivement performant pour la manipulation et la préparation des données, il est moyennement apprécié concernant les calculs statistiques, parce que moins précis et affublé d'une bibliothèque de fonctions limitée[2]. Pour palier cette faiblesse, de nombreux add-ins (macro complémentaires) existent pour compléter ses capacités d'analyse.

Nous avons maintes fois abordé le sujet, décrivant des solutions basées sur des technologies différentes[3]. Dans ce tutoriel, nous décrivons la macro complémentaire SQL Server Data Mining Add-ins. Il rajoute un onglet « Data Mining » dans le ruban Excel, avec plusieurs menus permettant d'accéder aux principales méthodes de la fouille de données, pour les plus marquantes d'entres elles : l'analyse prédictive, la classification automatique – clustering, la construction de règles d'association. L'add-in a de particulier qu'il s'appuie sur le service « Analysis Services Data Mining » (SSAS) de SQL Server. Ce dernier doit être installé, configuré et démarré pour disposer des fonctionnalités d'analyse de données sous Excel. Chaque traitement data mining implique le lancement, de manière transparente pour l'utilisateur, d'une session de calcul sous SQL Server. Les modèles et les données intermédiaires sont stockés dans une base dédiée d'Analysis Services. Les résultats peuvent être visualisés dans des fenêtres spécifiques ou insérés dans une feuille de calcul. Un post-traitement est donc possible en utilisant les fonctionnalités courantes du tableur. Les caractéristiques générales de l'outil sont assez similaires à celles de l'add-in SAS pour Excel[4].

Nous décrivons plusieurs cas d'utilisation des techniques de data mining fournies par l'add-in « SQL Server Data Mining Add-ins ». Plusieurs traits distinctifs apparaissent clairement a posteriori. L'outil couvre parfaitement la pratique usuelle du data mining (ex. le module d'apprentissage supervisé permet de construire différents modèles sur un échantillon d'apprentissage, d'en mesurer et comparer les performances sur un échantillon test, d'effectuer le déploiement sur des données non-étiquetées). Il mise sur une très grande

[1] KDnuggets Polls, « What Analytics, Data Mining, Data Science software/tools you used in the past 12 months for a real project », June 2014.

[2] K. Keeling, R. Pavur, « Statistical Accuracy of Spreadsheet Software », The Amercial Statistician, 65:4, 265-273, 2011.

[3] « L'add-in Tanagra pour Excel 2007 et 2010 », Août 2010 ; « L'add-in Real Statistics pour Excel », Juin 2014 ; « Connexion entre R et Excel via RExcel », Décembre 2011.

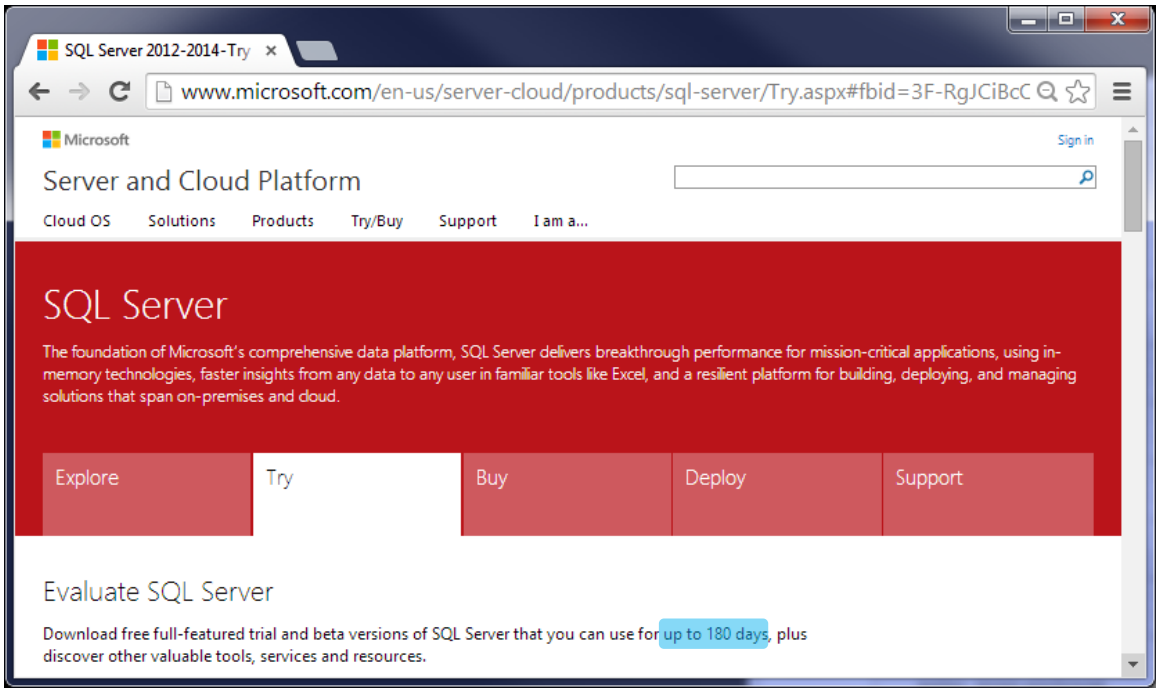
[4] « SAS Add-in 4.3 pour Excel », Avril 2012.

simplicité d'utilisation. Les manipulations sont intuitives. Il n'est pas nécessaire de scruter des heures durant l'aide en ligne[5] pour réaliser une étude. Cette apparente simplicité ne doit pas masquer la complexité de certaines opérations. L'outil effectue automatiquement des choix (ex. discrétisation automatique des variables, sélection des variables pertinentes, etc.) dont on doit avoir conscience pour apprécier pleinement la pertinence des résultats présentés.

2 Chargement et installation

2.1 Chargement et installation de SQL Server

Nous pouvons charger une version d'essai de SQL Server 2014 sur le site de Microsoft. Elle est valable 180 jours (6 mois), il y a de quoi voir venir.

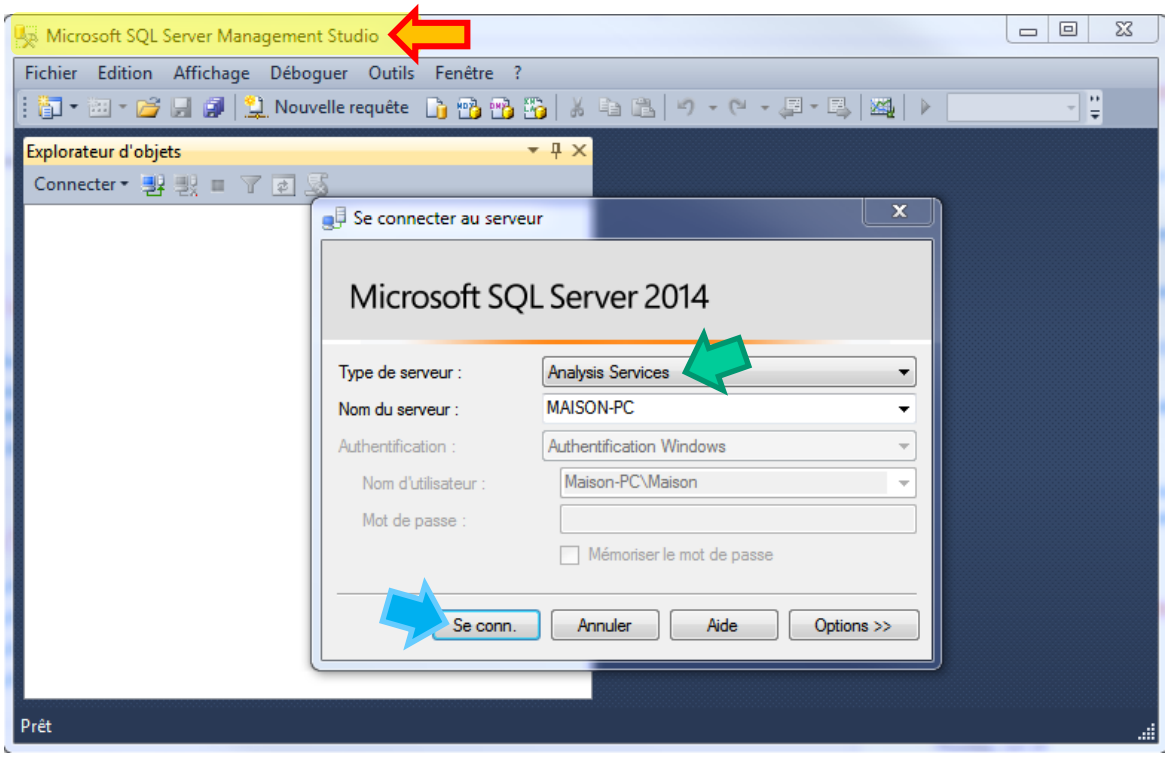


L'installation est très facile. Attention ! Il faut veiller à inclure dans votre configuration : Analysis Services Data Mining (SSAS) et SQL Server Management Studio. SQL Server est automatiquement démarré à l'issue de la procédure. Je n'en dis pas plus, d'une part parce que je ne suis pas un spécialiste des SGBD (Système de Gestion de Bases de Données), encore moins de SQL Server, et d'autre part, parce qu'il y a de très nombreux didacticiels à ce sujet sur le web.

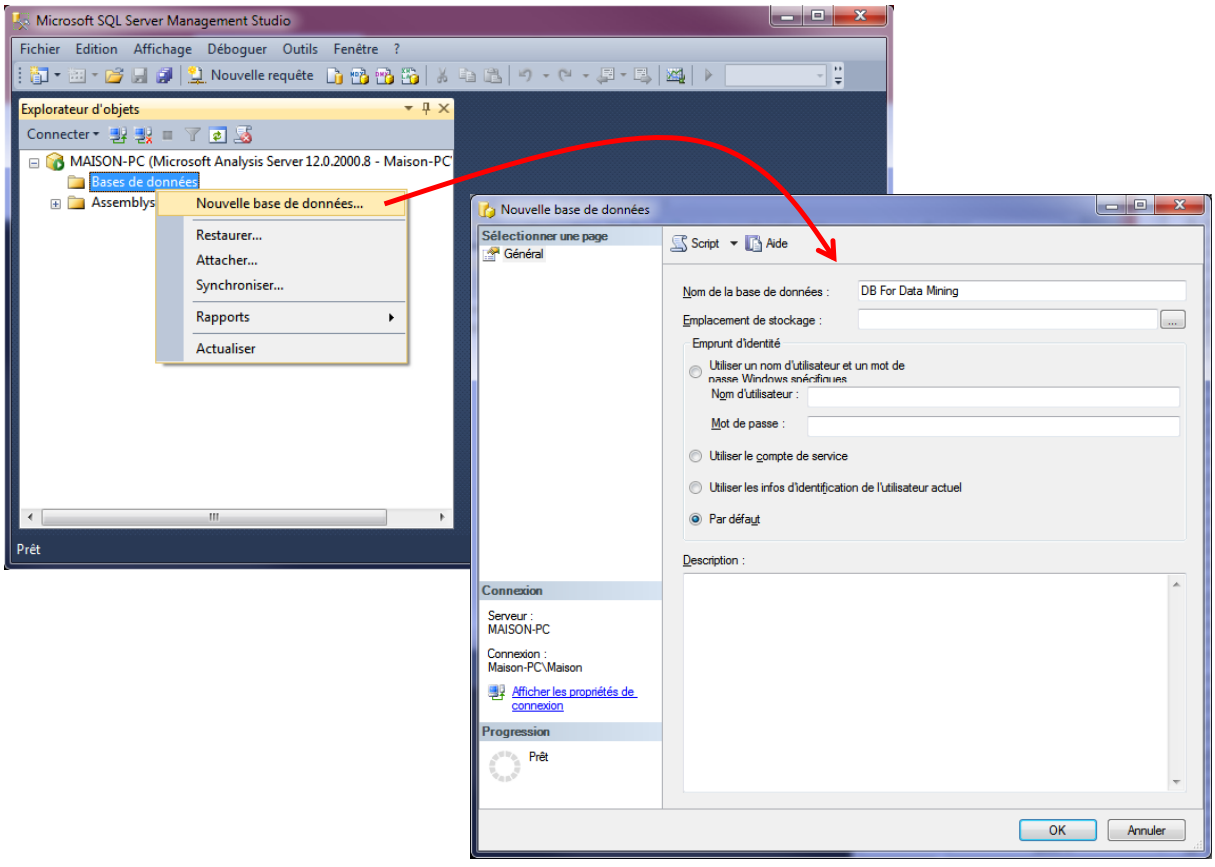
2.2 Configuration de Analysis Services

SQL Server a besoin d'un espace pour stocker les modèles et les données intermédiaires. Nous devons le créer explicitement avec l'outil SQL Server Management Studio.

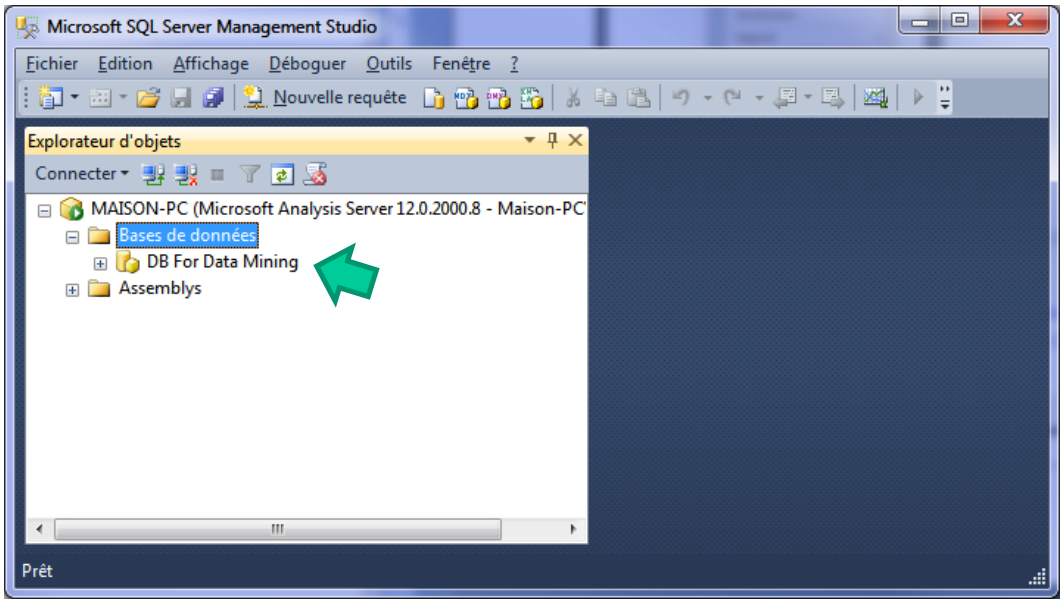
[5] msdn.microsoft.com – Data Mining (SSAS).



Le choix du type de serveur apparaît au démarrage de l'outil. Nous sélectionnons « **Analysis Services** » et nous cliquons sur le bouton « **Se Conn.** » (pour « Se connecter »). Une nouvelle base de données « DB For Data Mining » est créée : qu'importe son nom, la seule base existante deviendra la base par défaut.

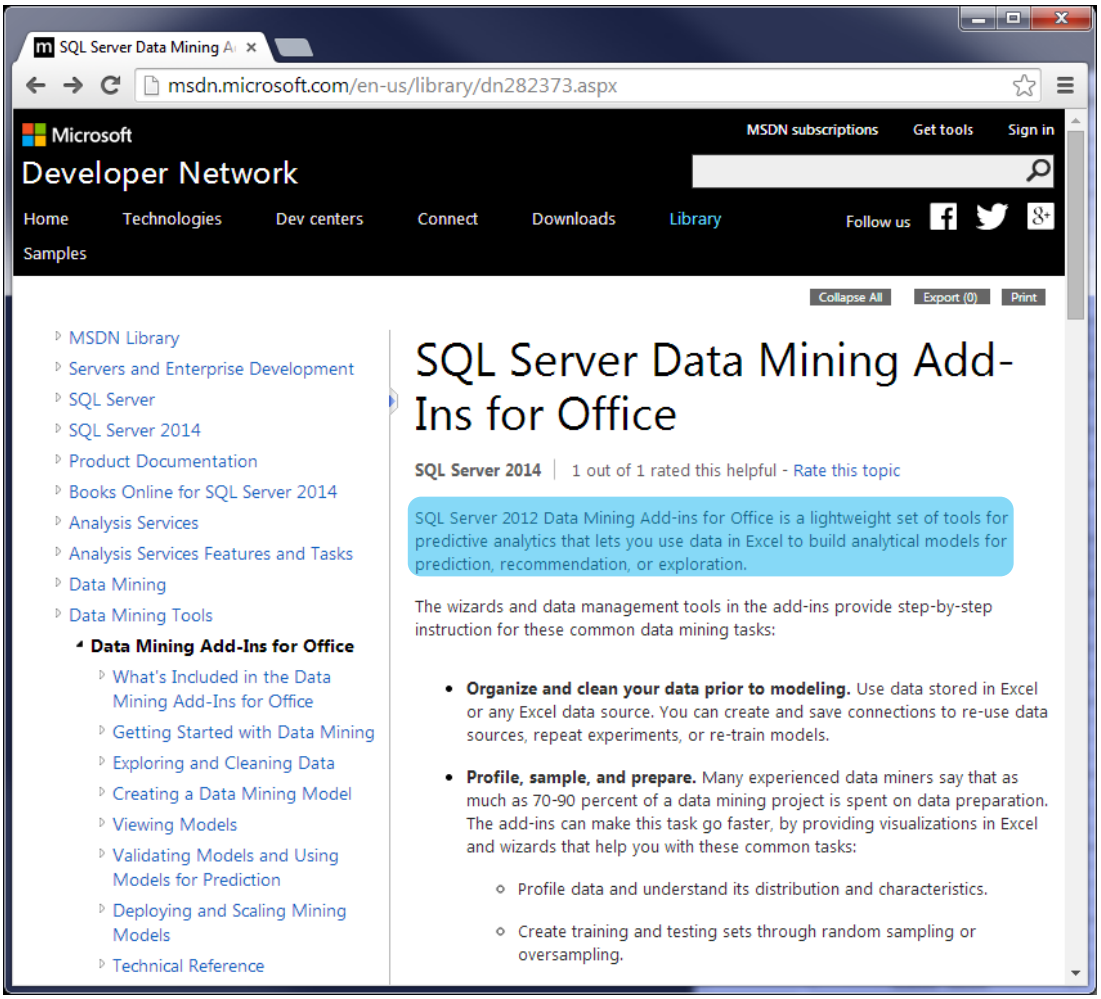


Elle est maintenant disponible pour les traitements.

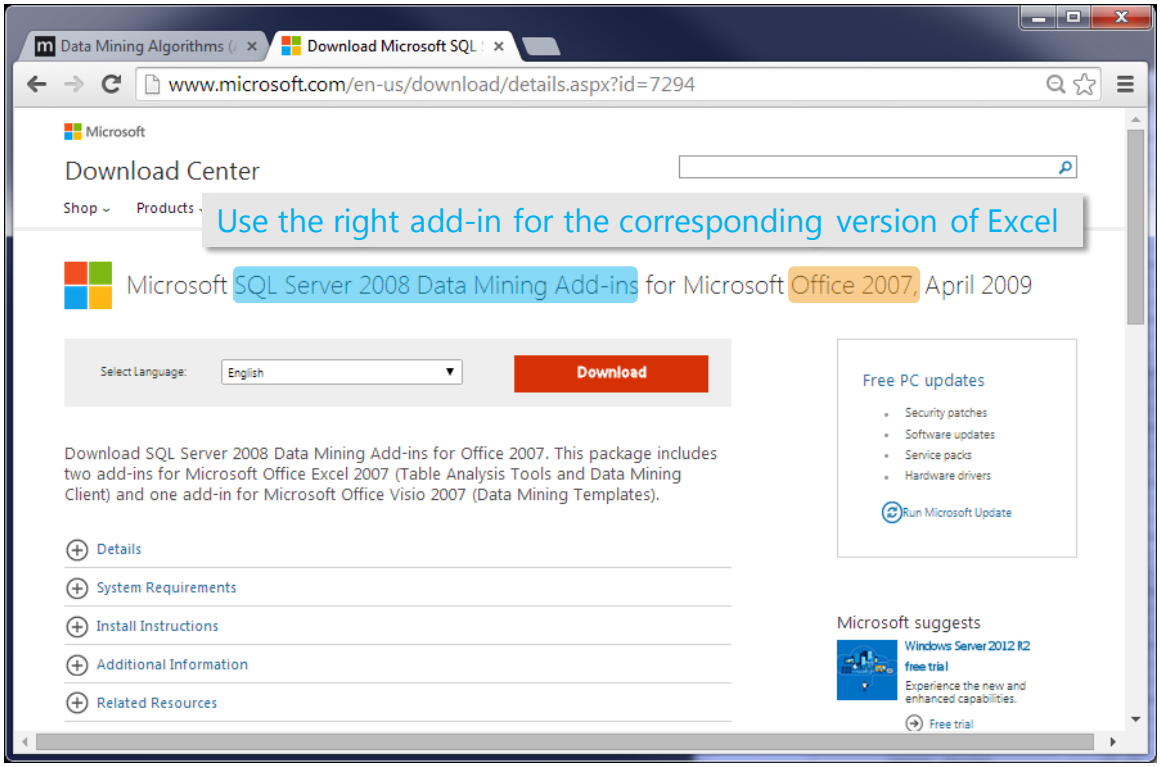


2.3 Chargement et installation de la librairie pour Office

L'étape suivante consiste à charger la librairie pour Office (et par extension, pour Excel) à partir du site de Microsoft.



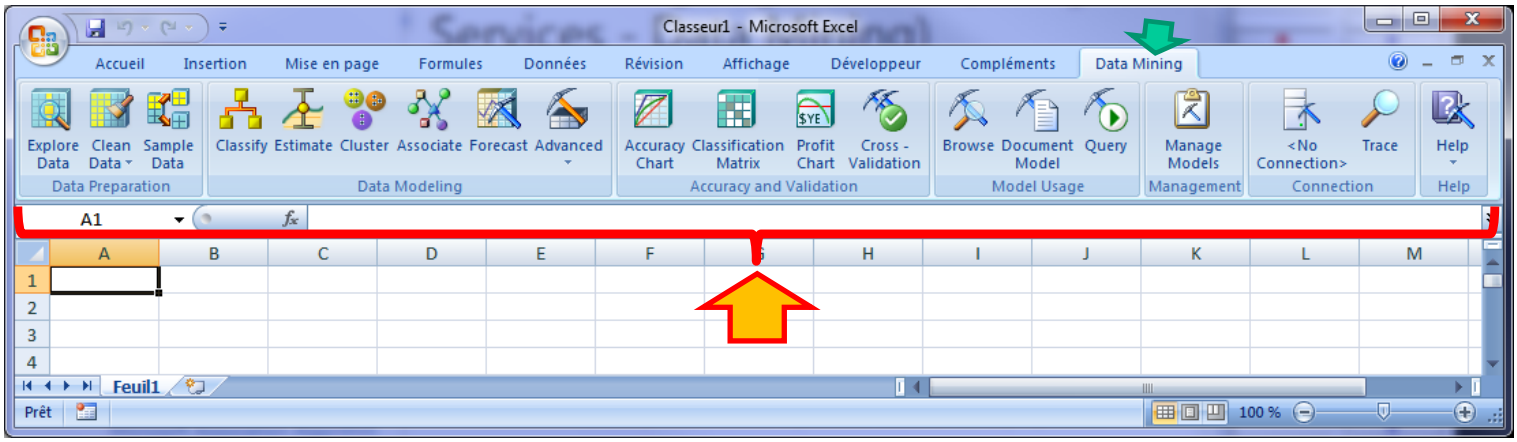
Attention, il faut utiliser une version compatible avec votre Excel. Pour ma part, j'ai récupéré **SQL Server 2008 Data Mining Add-Ins** parce que je dispose d'**Excel 2007**. Il inclut, entres autres, **Data Mining Client for Excel** qui rend accessible sous Excel les fonctionnalités de Data Mining du service.



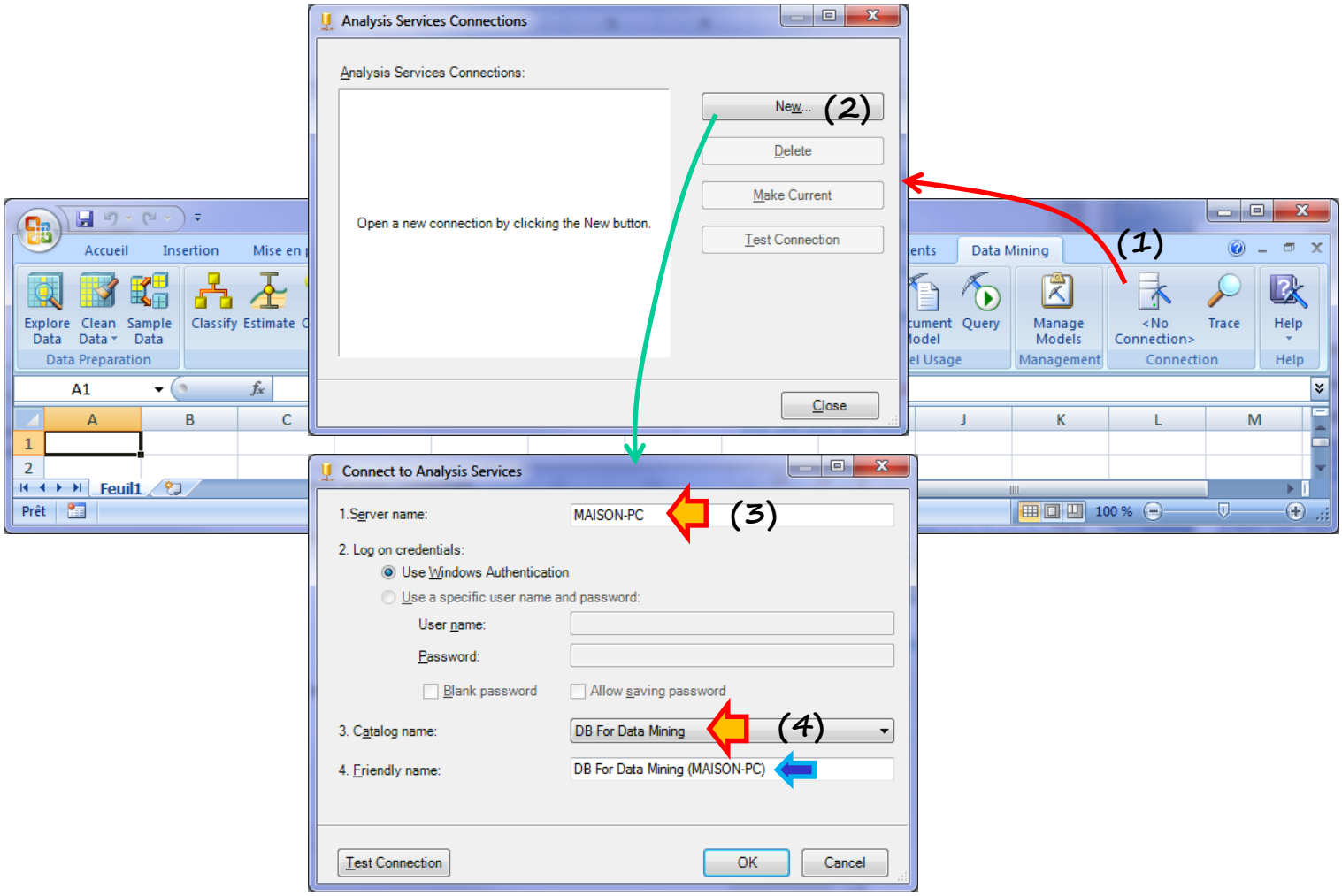
Une fois le fichier chargé, nous lançons l'installation. La procédure n'appelle pas de commentaires particuliers.

2.4 Configuration de l'add-in « Data Mining Client » dans Excel

Au démarrage d'Excel, un nouveau bandeau « Data Mining » est disponible. Nous y retrouvons entres autres : les méthodes de modélisation (data modeling) ; les outils d'évaluation et de mesure de performances (accuracy and validation) ; les outils d'exploitation (model usage), avec notamment une procédure de déploiement (Query).



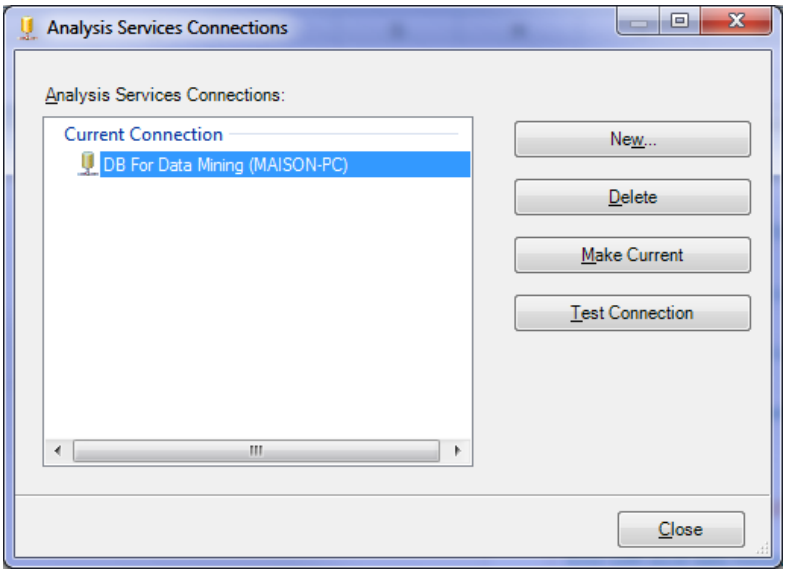
Lors du premier lancement, il faut établir la connexion avec SQL Server pour bénéficier du service Data Mining, et disposer d'un espace de stockage pour les modèles et les données intermédiaires. Nous cliquons sur l'icône <No Connection> de la section « Connection » (**1**).



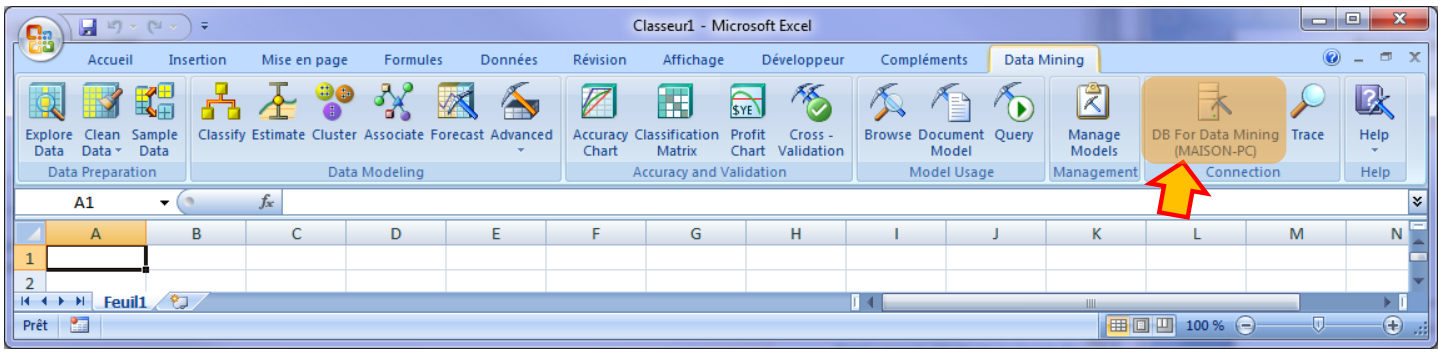
Dans la boîte de dialogue qui vient : nous établissons une nouvelle connexion en cliquant sur le bouton « New » (**2**) ; le nom du serveur « Server Name » est le nom de la machine, puisque SQL a été installé localement en ce qui me concerne (**3**)[6] ; enfin, nous sélectionnons le nom de la base de données précédemment créée dans Management Studio (section 2.2).

La connexion est maintenant établie, elle est visible dans la boîte de paramétrage « Analysis Services Connection ». On remarquera qu'il est possible d'en concevoir plusieurs.

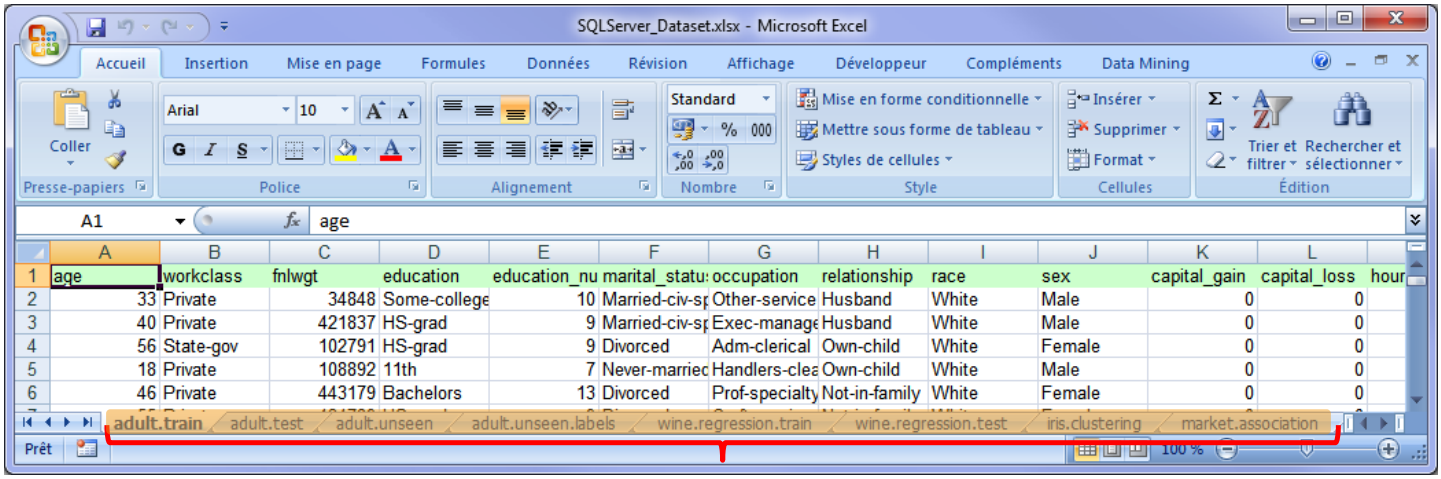
[6] Oui, oui, ça va, MAISON-PC est un nom très très original…



La base courante est visible dans le ruban « Data mining » d'Excel. La connexion est persistante c.-à-d. au prochain démarrage d'Excel, elle est instaurée automatiquement. Nous pouvons cependant la reconfigurer si nous souhaitons modifier la base à utiliser.



A ce stade, tout est prêt pour pouvoir mener des analyses. Nous chargeons le fichier « **SQLServer_Dataset.xlsx** ». Il comporte plusieurs feuilles relatives aux différentes procédures de Data Mining Client pour Excel que nous souhaitons évaluer dans ce qui suit.



3 Apprentissage supervisé - Classement

Nous souhaitons mener une démarche typique de l'analyse prédictive : construire plusieurs modèles sur un échantillon d'apprentissage (**train**) ; comparer leurs performances sur un échantillon test (**test**) ; déployer le meilleur modèle sur des données non étiquetés (**unseen**).

Nous disposons des étiquettes de ce troisième ensemble de données dans notre cas, puisque que nous sommes dans une démarche expérimentale. Nous pourrons vérifier si les performances estimées sur l'échantillon test sont effectivement crédibles. Dans les études réelles, nous ne disposons pas de cette possibilité. Les indications de l'échantillon test font foi. Nous avons intérêt à multiplier les précautions pour obtenir des résultats fiables.

3.1 Données à traiter

Nous utilisons les données ADULT du serveur UCI[7]. Nous cherchons à prédire le niveau de revenu de personnes (>50K$ annuels ou pas, variable TARGET) à partir de leurs caractéristiques (âge, niveau d'éducation, type d'emploi, etc.). Les observations ont été scindées en 3 parties : « **adult.train** » avec 30030 observations fait office d'échantillon d'apprentissage ; « **adult.test** » correspond à l'échantillon test (10000 observations) ; « **adult.unseen** » représente la population sur laquelle sera déployé le modèle (8812 observations, la colonne supplémentaire ID sert d'identifiant).

3.2 Construction d'un arbre de décision

Lancement de la procédure. SSAS propose plusieurs méthodes d'apprentissage supervisé. Les arbres de décision, méthodes phares du data mining[8], sont bien évidemment présents. Fait remarquable, une documentation particulièrement abondante est accessible en ligne[9]. A l'usage des scientifiques, l'algorithme est décrit[10], sans que l'on dispose cependant du détail des formules. Microsoft s'inscrit dans la tendance actuelle des éditeurs de logiciels statistiques qui rendent publiques les manuels de leurs produits (ex. SAS Enterprise Miner, IBM-SPSS, STATISTICA, etc.). Ce sont de vraies mines d'or pour tout curieux désireux de mieux comprendre les tenants et aboutissants des méthodes statistiques. Je les consulte très souvent pour situer mes propres implémentations et les sorties d'autres logiciels tels que R.

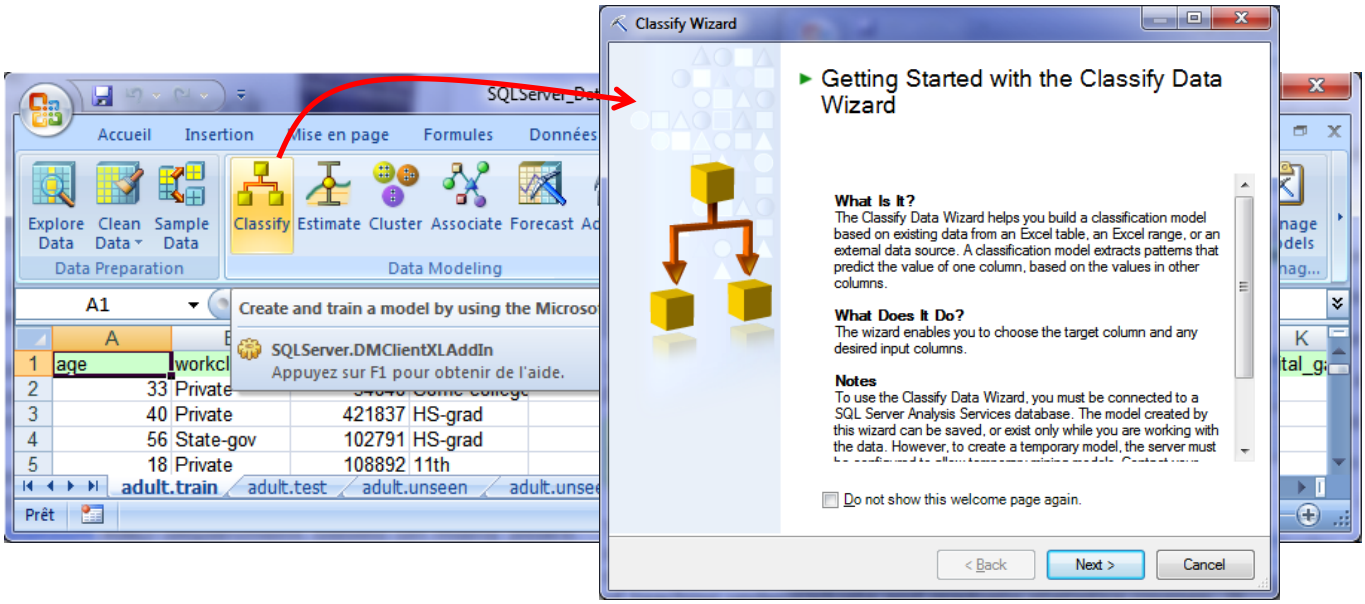
[7] https://archive.ics.uci.edu/ml/datasets/Adult

[8] KDnuggets Polls, « Which methods/algorithms did you use for data analysis in 2011 », Nov. 2011.

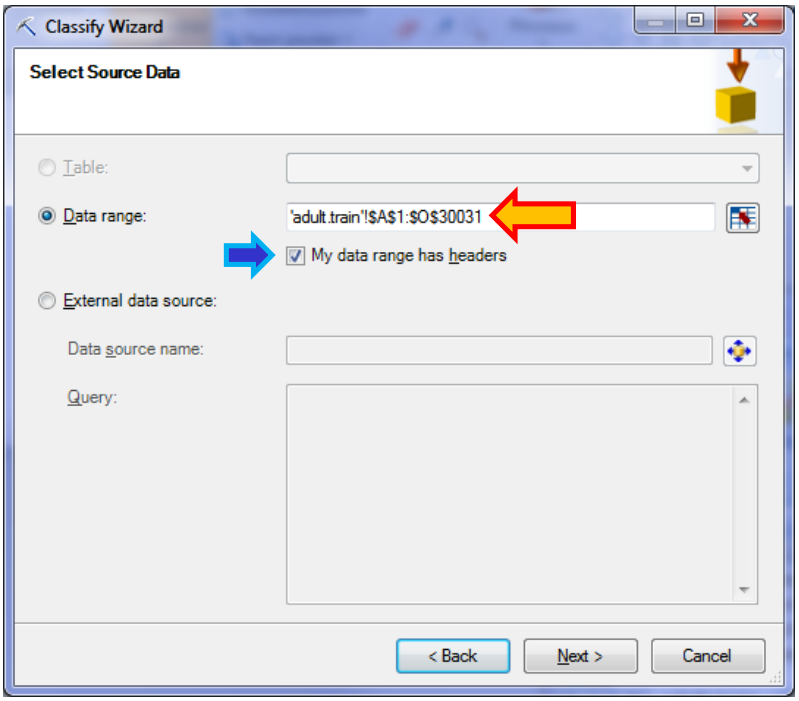
[9] « Microsoft Decision Trees Algorithm », http://technet.microsoft.com/en-us/library/ms175312.aspx

[10] « Microsoft Decision Trees Algorithm – Technical Reference ».

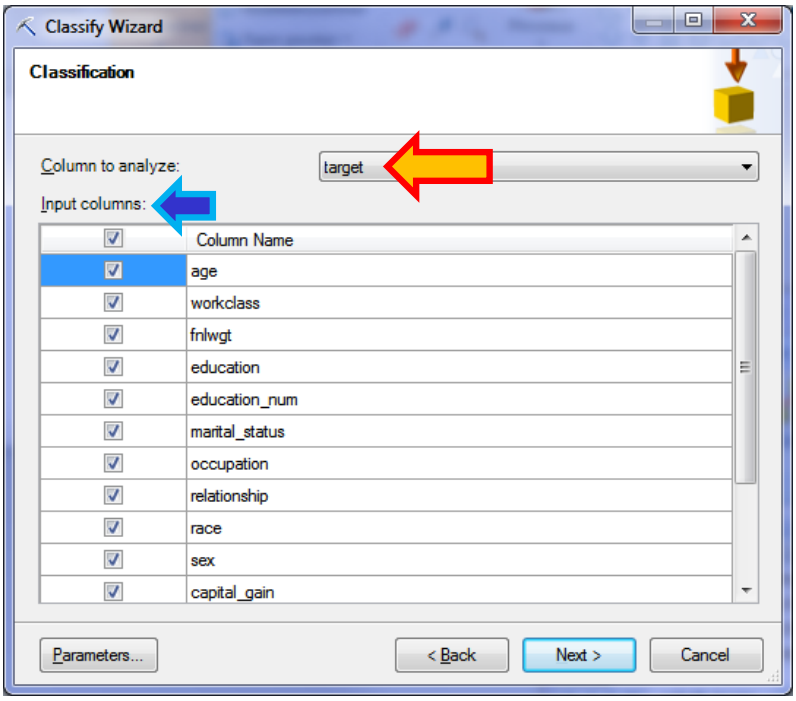
Nous cliquons sur l'icône CLASSIFY dans le ruban « Data Mining » d'Excel. Un « wizard » (guide) apparaît, nous précisant la teneur de l'analyse à venir.



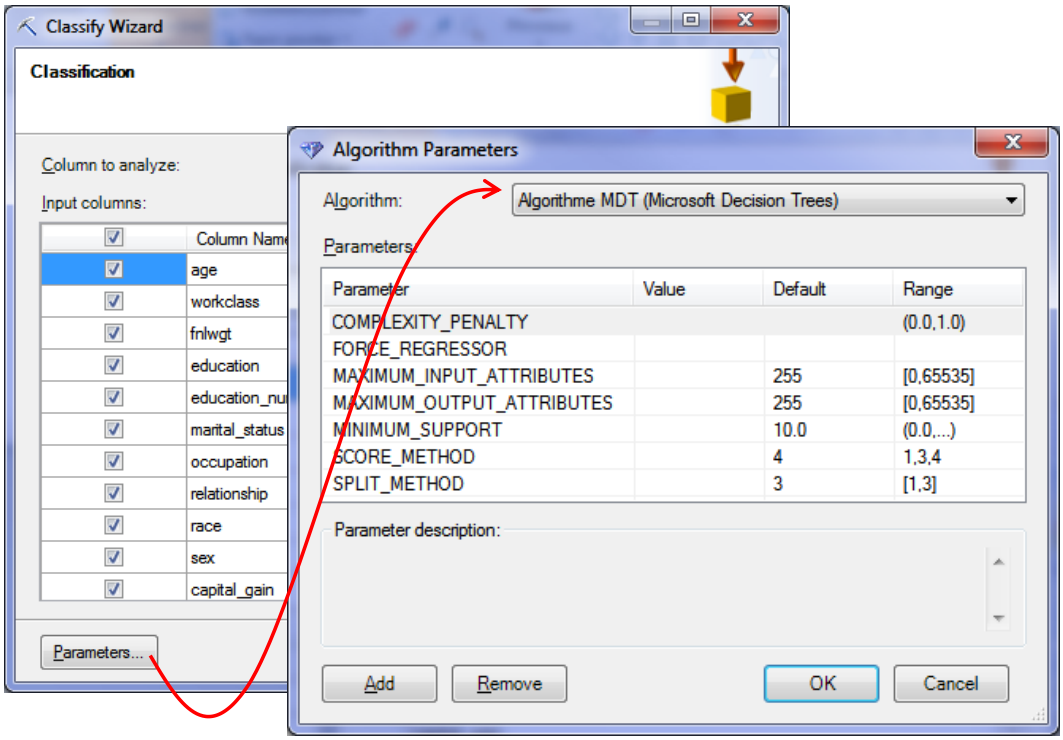
Sélection des données. Nous cliquons sur le bouton NEXT pour accéder à l'étape suivante. Nous sélectionnons la plage de données d'apprentissage située dans la feuille « adult.train ». La première ligne correspond aux noms des variables. Notons au passage qu'il est possible de brancher l'outil sur des sources de données externes.



Rôle des variables dans l'analyse. Dans la fenêtre suivante, nous précisons le rôle des variables. TARGET est la variable cible. Les colonnes à utiliser ou à exclure de l'analyse sont précisées à l'aide de l'option INPUT COLUMNS.



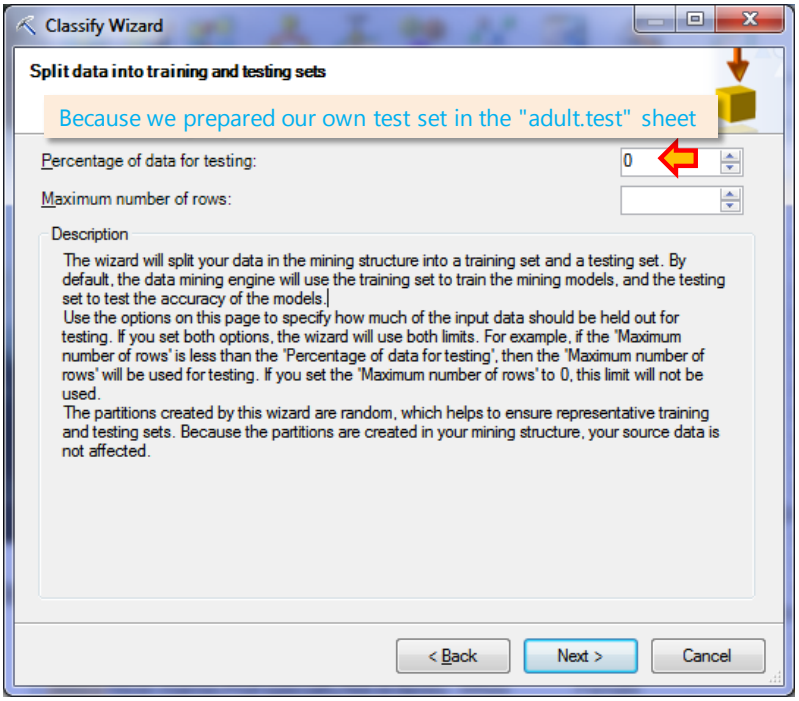
Choix de la méthode de data mining. Toujours dans la même fenêtre, nous cliquons sur le bouton PARAMETERS pour spécifier la méthode d'apprentissage.



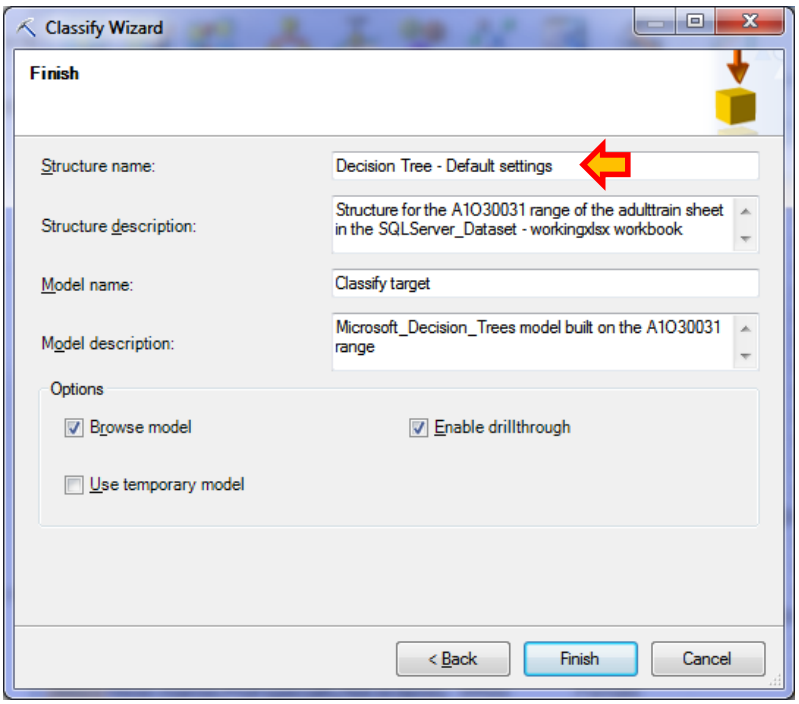
Nous choisissons les arbres de décision avec les options par défaut. Nous reviendrons sur le paramétrage des méthodes de data mining plus loin (section 3.3).

Subdivision apprentissage-test. L'outil propose automatiquement une partition des données en échantillons d'apprentissage (70%) et de test (30%). Dans notre cas, ce n'est pas nécessaire parce nous avons déjà préparé à l'avance notre classeur en copiant dans une feuille à part

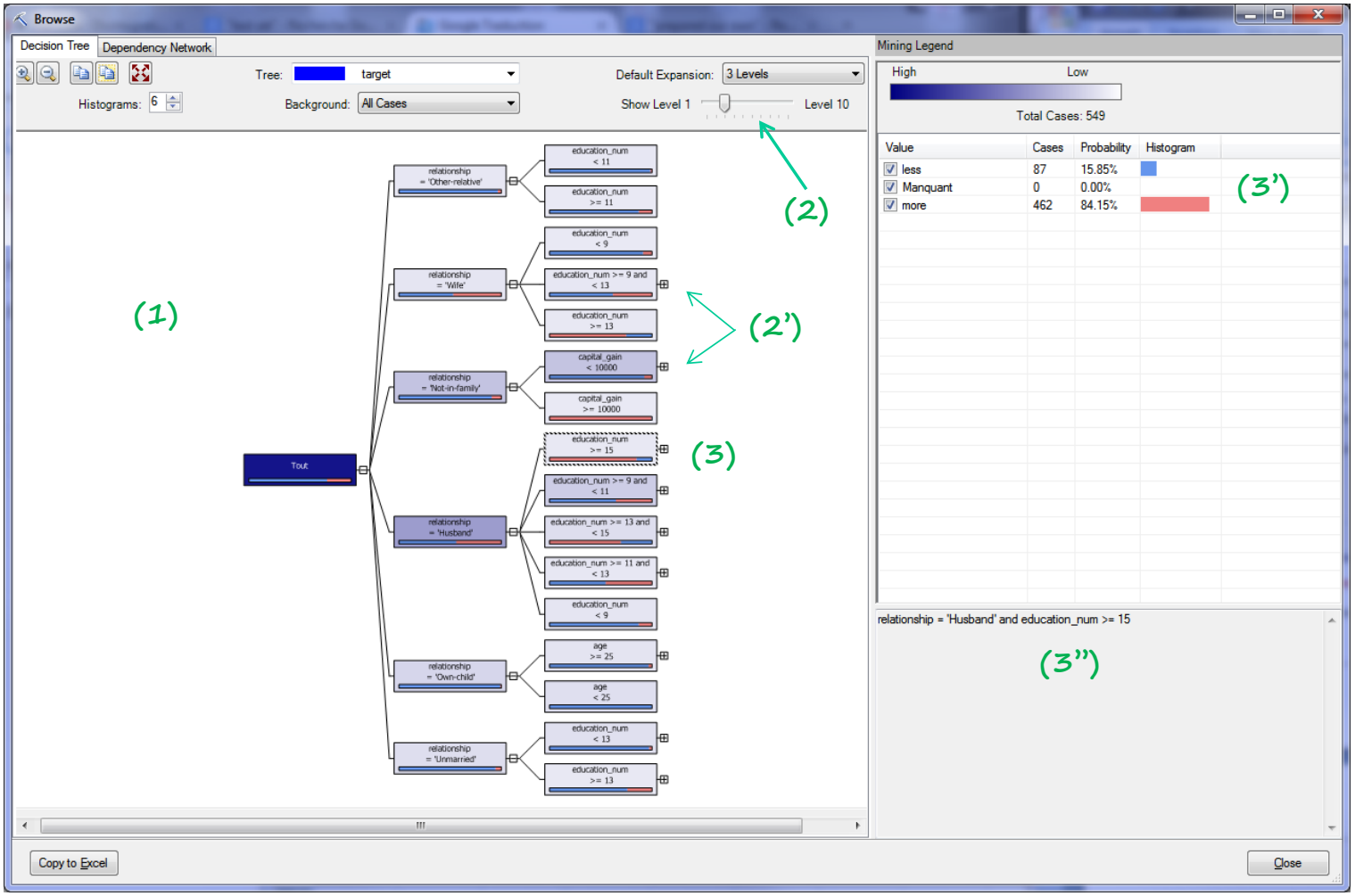
l'ensemble de test. Nous avons ainsi la garantie d'utiliser exactement les mêmes données pour l'évaluation et la comparaison des modèles. Ainsi, nous mettons la valeur 0 pour la taille du second échantillon dans notre configuration.



Stockage du modèle. Dernière étape, nous attribuons le nom « Decision Tree – Default Settings » au modèle pour son stockage dans la base de données.

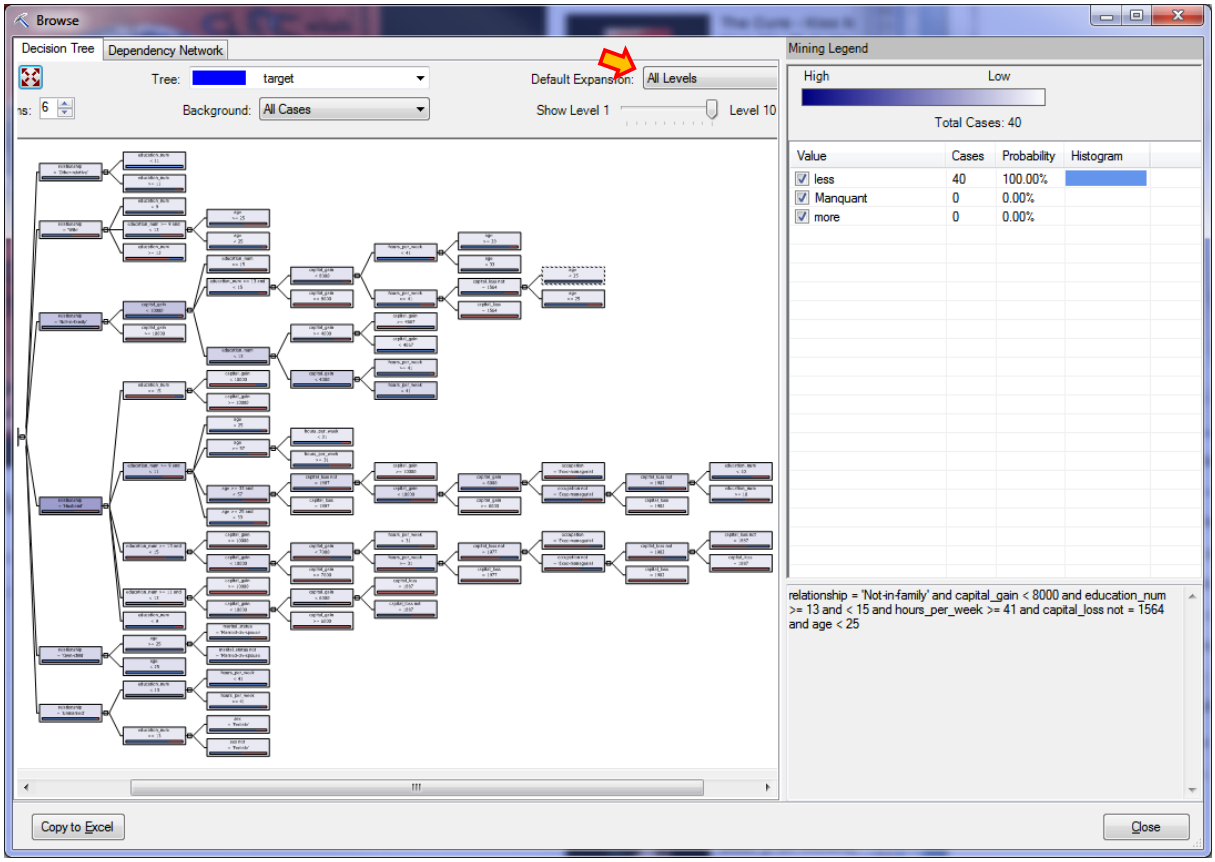


Visualisation de l'arbre. Nous lançons la construction effective du modèle en cliquant sur le bouton « Finish ». L'arbre est affiché dans une fenêtre spécifique (1).



Il n'est pas affiché dans sa totalité. Nous pouvons contrôler le niveau de profondeur avec l'outil « Show Level » (2). De même, en cliquant sur l'icône « + » sur les feuilles, nous pouvons consulter les sommets suivants (2'). Attention, l'outil sert uniquement à la visualisation. Il n'est pas question de manipulations interactives visant à élaguer l'arbre, ou encore introduire de nouvelles segmentations qui n'ont pas été calculées lors de la modélisation initiale c.-à-d. il n'est pas possible de partitionner une feuille où le symbole « + » n'est pas présent. De même il n'est pas possible de modifier la variable de segmentation sur un sommet. En sélectionnant un nœud (3), nous avons la distribution de fréquences des modalités de la variable cible (3') et la retranscription de la règle (3'').

Pour obtenir une vue globale de l'arbre, nous sélectionnons l'item « All Levels » de l'option « Default Expansion ». Nous effectuons un zoom arrière pour qu'il tienne dans la fenêtre de visualisation.



Graphe des dépendances. Un graphe des dépendances entre la cible et les variables prédictives est affiché dans l'onglet « Dependency network ». Les liens mis en évidence dépendent du niveau de profondeur sélectionné. Si l'on s'en tient aux 3 plus fortes liaisons, nous obtiendrons le graphe suivant.

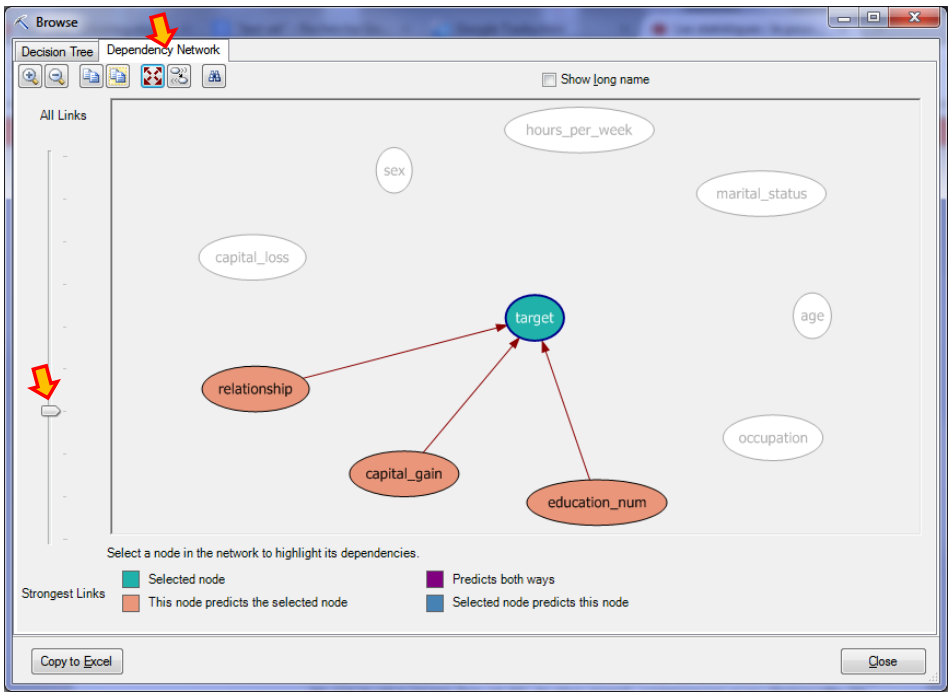
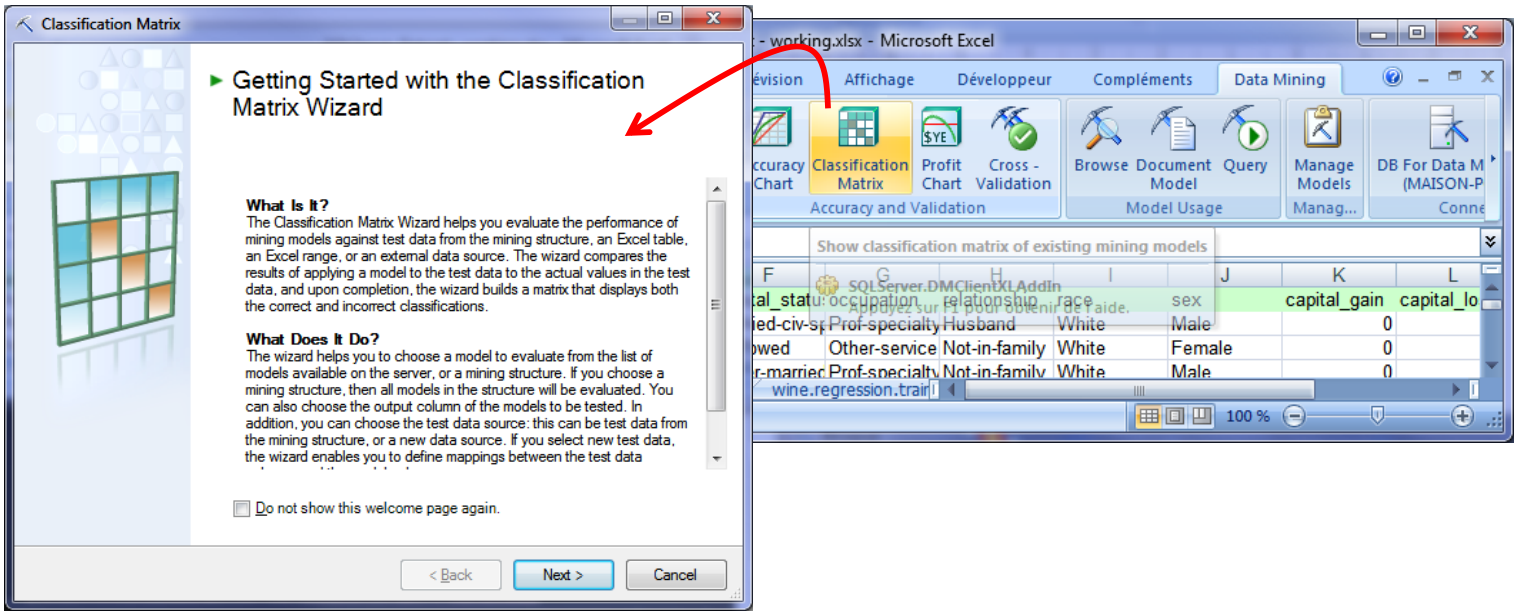


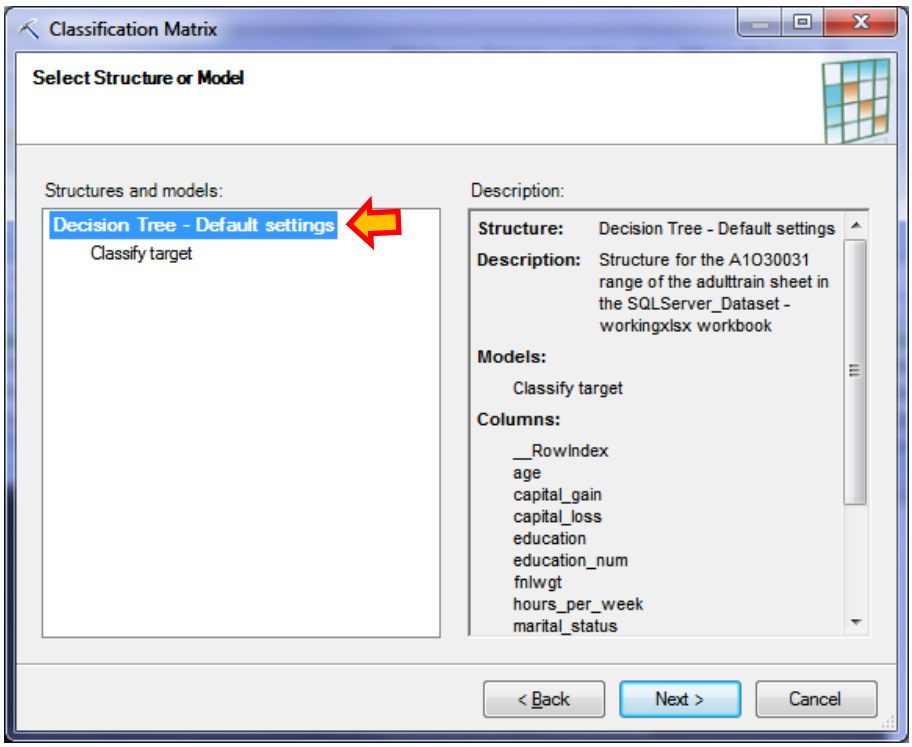
Figure 1 - Graphe des dépendances - Arbre de décision

L'enjeu consiste alors à comprendre ces indications en les mettant en relation avec l'arbre de décision. Nous notons par exemple que « relationship » est la première variable de segmentation à la racine de l'arbre, etc. Nous cliquons sur CLOSE pour terminer l'analyse.

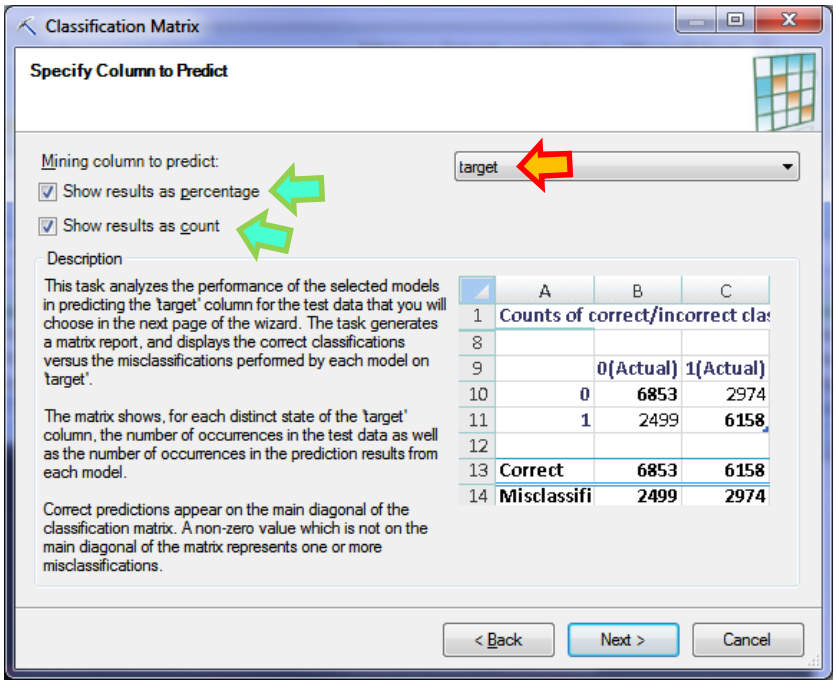
Evaluation du modèle. Pour construire la matrice de confusion et calculer le taux d'erreur sur l'échantillon test (feuille **adult.test**), nous actionnons le bouton CLASSIFICATION MATRIX de la section « Accuracy and validation » du ruban « Data Mining ».



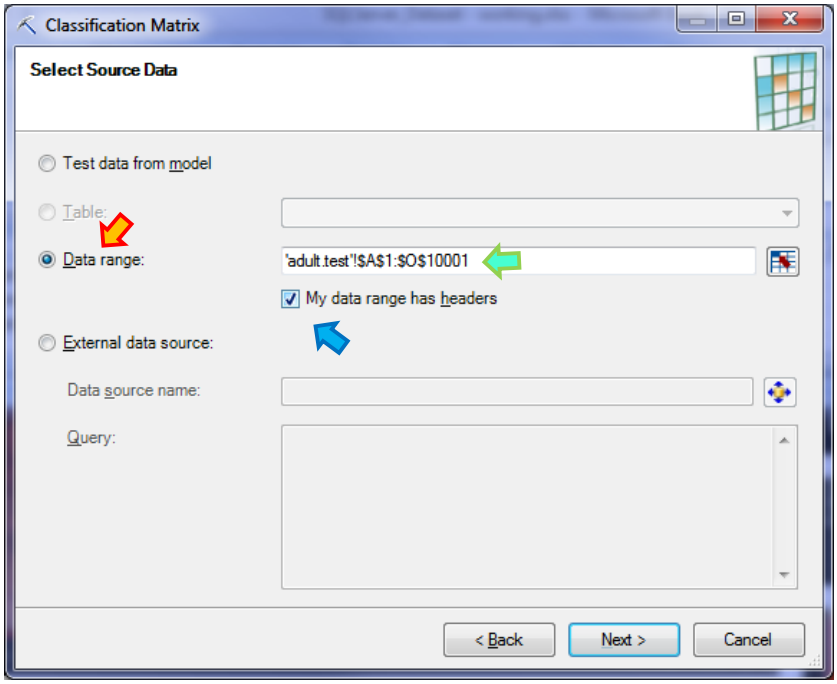
Dans la fenêtre suivante, nous sélectionnons le modèle à évaluer. Il n'y a que « Decision Tree – Default Settings », le premier arbre que nous avons construit, pour l'instant.



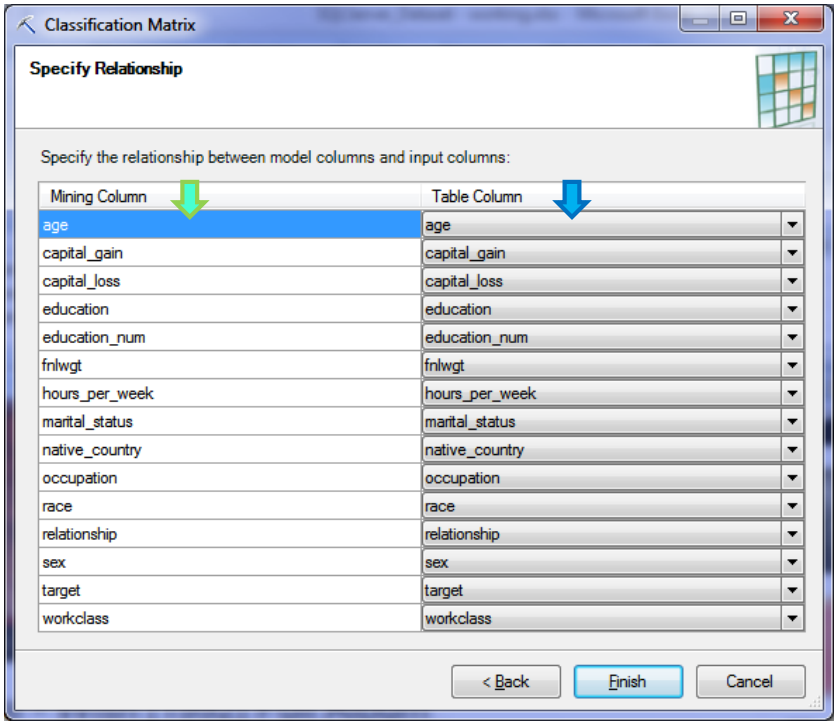
L'outil nous demande de préciser la variable cible, « target » en l'occurrence, et le mode d'affichage des effectifs (« as count » et/ou « as percentage »).



Nous devons ensuite spécifier l'échantillon à utiliser. Dans notre cas, nous activons l'option « Data Range » et nous sélectionnons la plage de données de la feuille « **adult.test** », en précisant que la première ligne correspond aux noms des variables.



Excel réalise automatiquement les correspondances entre les noms de variables incluses dans le modèles, et ceux disponibles dans l'échantillon test. Nous pouvons au besoin les préciser si des ambiguïtés persistent.



Nous cliquons sur FINISH pour lancer l'opération. La matrice de confusion est automatiquement insérée dans une nouvelle feuille Excel « **Classification Matrix** ».

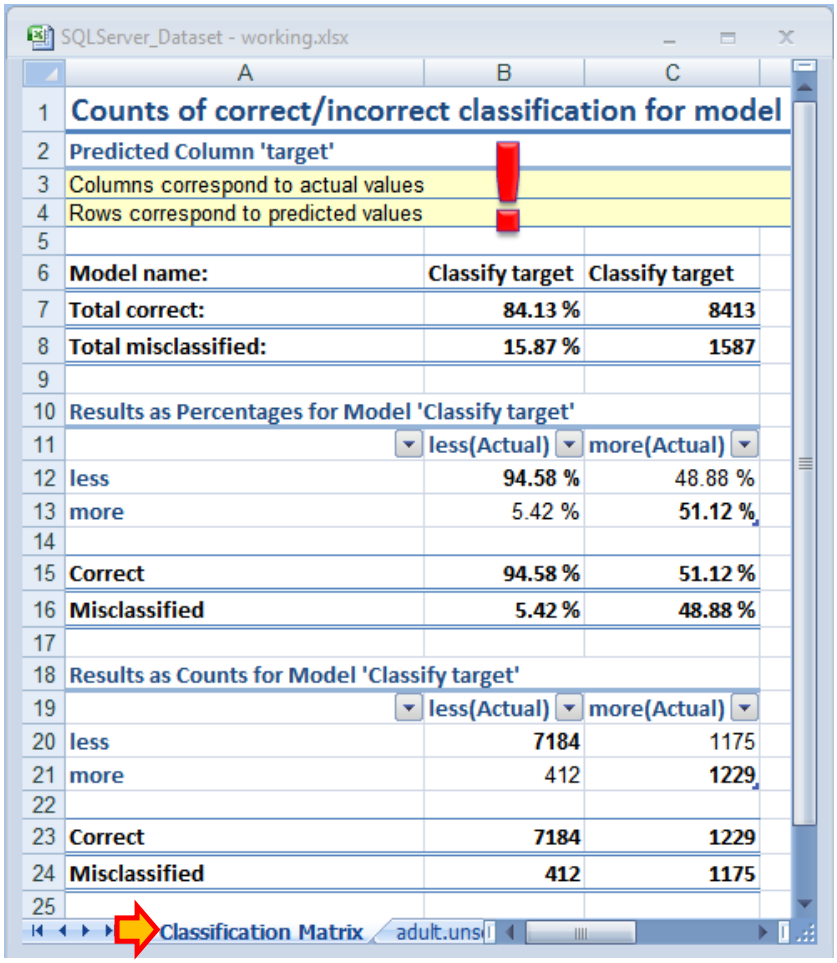


Figure 2 - Matrice de confusion sur l'échantillon test

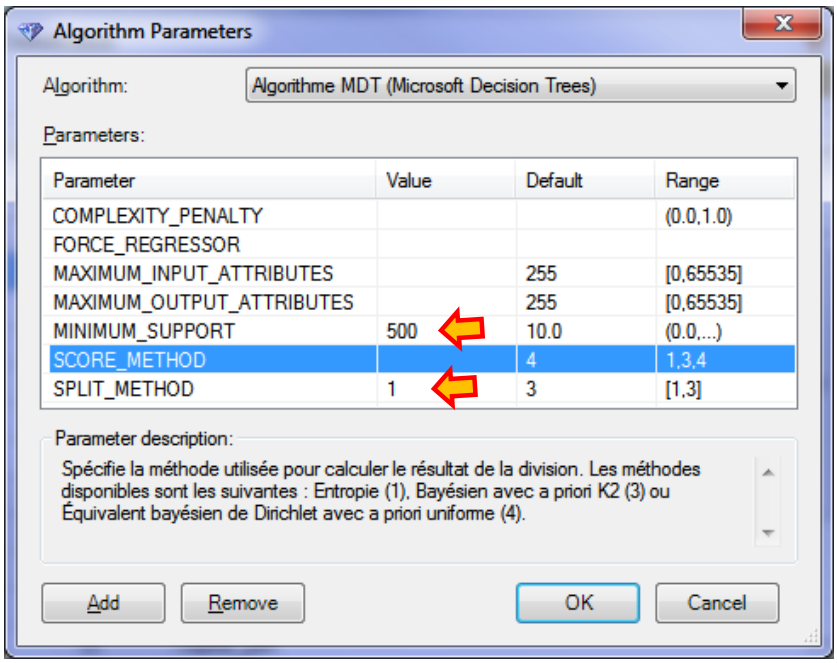
Contrairement à la présentation habituelle (dans le monde francophone en tous les cas), les modalités observées sont en colonnes[11], les prédictions sont en ligne. Nous remarquerons essentiellement les informations suivantes :

- Le taux d'erreur du modèle est de **15.87%** (1587 individus mal classés sur 10000)
- Sur 7596 individus « less » observés, 7184 (94,58%) ont été attribués à raison à la modalité « less », 412 (5.42%) ont été prédits « more » de manière erronée.
- La même lecture en colonne est possible pour la modalité « more ».

Ainsi s'achève cette première analyse. Elle a été plus longue à décrire qu'à réaliser. Nous allons reproduire la même démarche avec d'autres techniques de modélisation dans les sections suivantes.

3.3 Arbre de décision – Paramètres modifiés

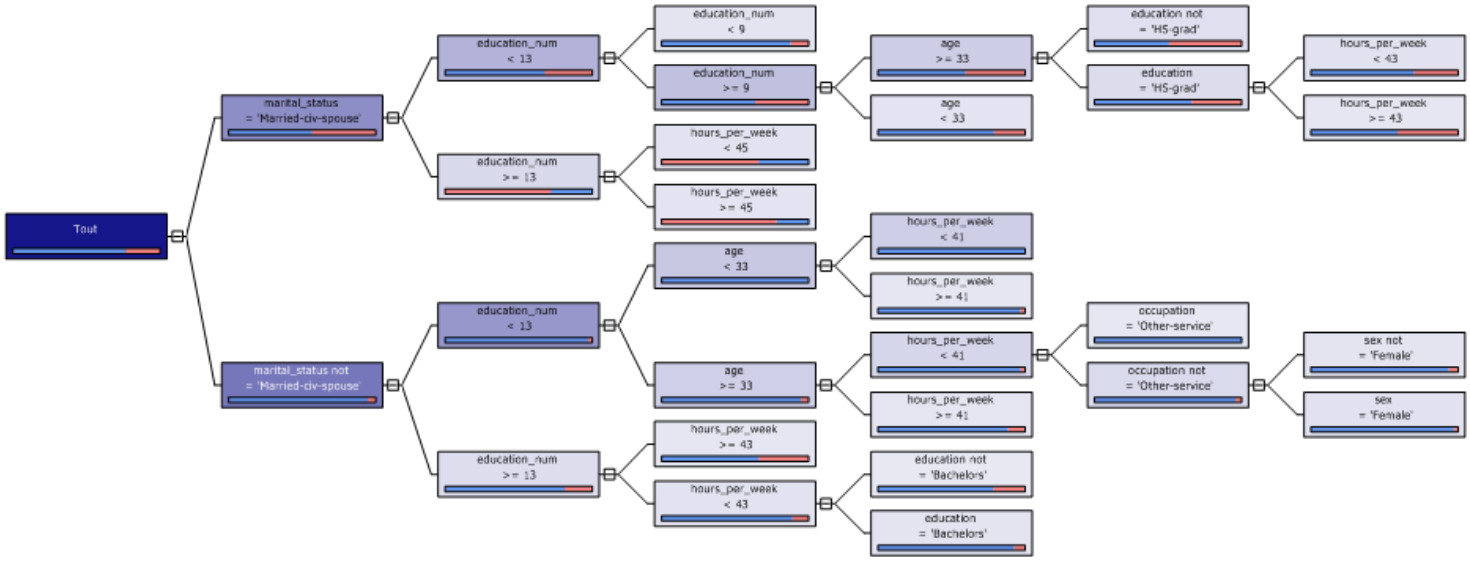
Notre arbre précédent était (peut-être) surdimensionné. Nous décidons de réitérer l'analyse en modifiant les paramètres de l'algorithme. De nouveau, nous actionnons le bouton CLASSIFY et nous enchainons les mêmes spécifications. A la différence que, lors de la phase du choix de l'algorithme, nous effectuons les modifications suivantes :



MINIMUM_SUPPORT = 500, une feuille (une règle) doit couvrir au moins 500 observations pour être validée ; SPLIT_METHOD = 1, un arbre binaire est construit. Ces modifications devraient aboutir à l'élaboration d'un arbre plus compact. Rien ne nous dit cependant qu'il sera plus performant, nous le saurons lorsque nous l'appliquerons sur l'échantillon test.

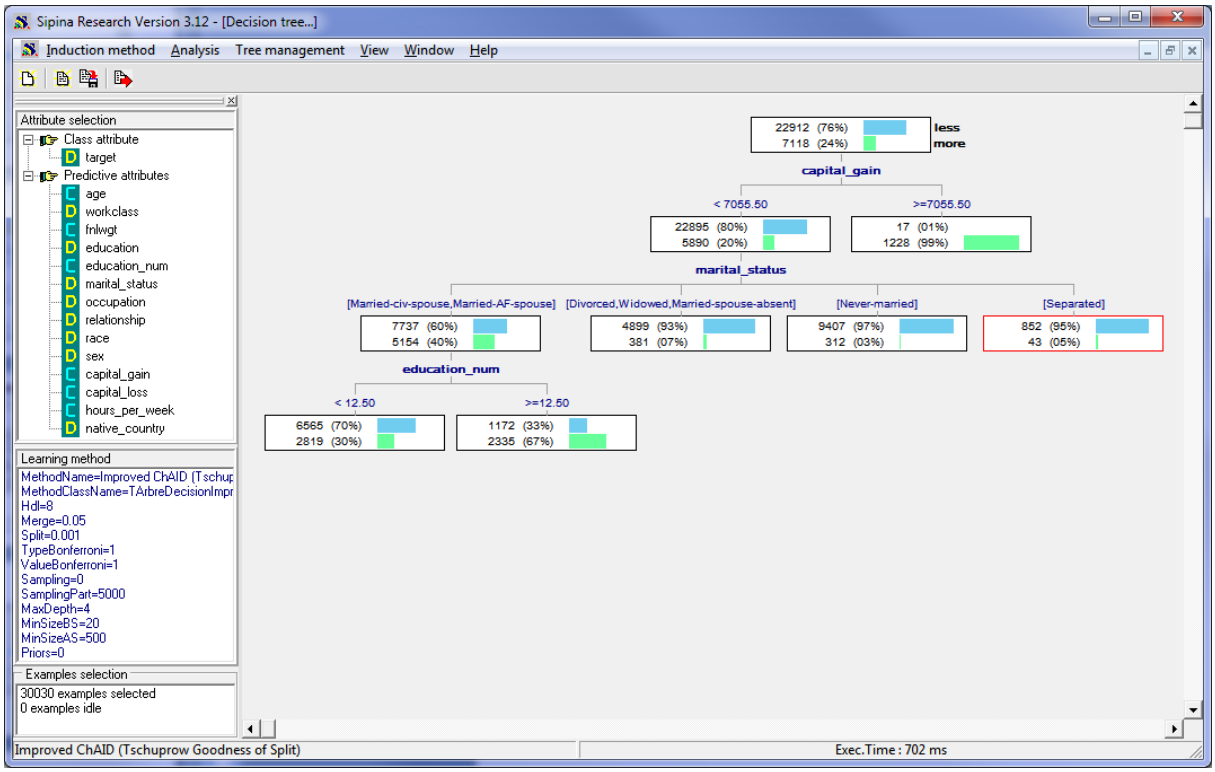
[11] http://technet.microsoft.com/en-us/library/ms174811.aspx

Nous appelons le modèle « Decision Tree – Modified Settings » et nous obtenons.



L'arbre est moins développé. La binarisation s'appuie sur la stratégie une modalité contre les autres (ex. la première segmentation correspond au découpage « marital_status = married-civ-spouse » vs. « marital_status **not** = married-civ –spouse »), elle n'effectue pas de regroupements (approche utilisée par la méthode CART, Breiman et al., 1984). Ses performances en test seront mesurées lors de la comparaison de modèles (section 3.7).

Par comparaison, voici l'arbre fourni par le logiciel Sipina 3.12.



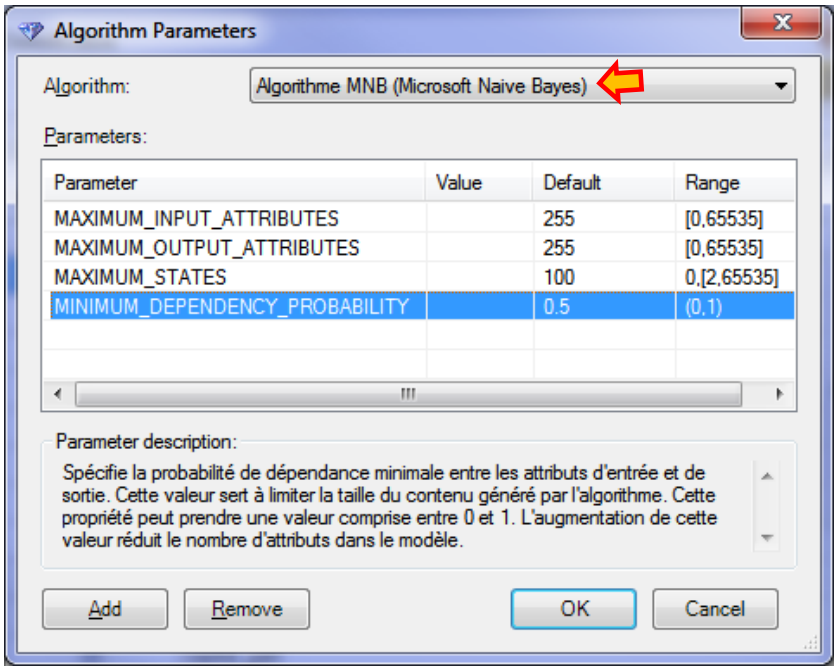
Nous avons utilisé une méthode apparentée à CHAID (Kass, 1980). Le critère de segmentation et la stratégie d'agrégation notamment sont différents. La taille minimale des feuilles est de 500 individus. Nous avons effectué un post-élagage basé sur les conclusions des feuilles c.-à-d. les feuilles sœurs porteuses de conclusions identiques sont supprimées, l'arbre est réduit de proche en proche en partant des sommets terminaux vers la racine.

« Capital_gain » joue un rôle important dans Sipina alors qu'elle n'apparaît nulle par dans le modèle de SSAS. C'est un des principaux reproches que l'on pourrait adresser aux arbres de décision. Le reste de l'arbre est fortement dépendant du choix de segmentation sur les premiers sommets. Des variables pertinentes peuvent être masquées par d'autres. Nous devons absolument explorer finement les différents sommets (segmentations concurrentes, statistiques descriptives comparatives, etc.) pour apprécier pleinement les résultats[12].

3.4 Bayésien Naïf (Modèle d'indépendance conditionnelle)

Le bayésien naïf (modèle d'indépendance conditionnelle) est un classifieur linéaire[13]. La difficulté ici réside dans la présence d'un mélange de variables prédictives qualitatives et quantitatives. Une harmonisation est nécessaire. Nous verrons comment procédera le logiciel.

Nous cliquons sur CLASSIFY pour initier une nouvelle analyse. Les étapes sont identiques à celle des arbres de décision, excepté au moment du choix de l'algorithme d'apprentissage.



[12] Voir Tutoriel Tanagra, « Analyse interactive avec Sipina », mars 2008.

[13] Wikipédia, « Classification naïve bayésienne ». Tanagra, « Le classifieur bayésien naïf revisité », mars 2010 ; « Bayésien naïf pour prédicteurs continus », octobre 2010.

Nous sélectionnons « Algorithme MNB (Microsoft NaiveBayes) ». Notons deux informations importantes concernant l'approche implémentée[14] : les prédicteurs continus sont automatiquement discrétisés ; la méthode intègre une technique de sélection de variables basée sur une mesure d'association avec la cible. Nous pouvons paramétrer un seuil d'acceptation pour incorporer plus ou moins de variables dans notre modèle.

« Naive Bayes » sera le nom de notre modèle. Plusieurs informations regroupées dans différents onglets sont fournies à l'issue du processus d'apprentissage.

Graphe des dépendances (Dependency Newtwork). Il hiérarchise les variables selon leur pertinence dans la prédiction.

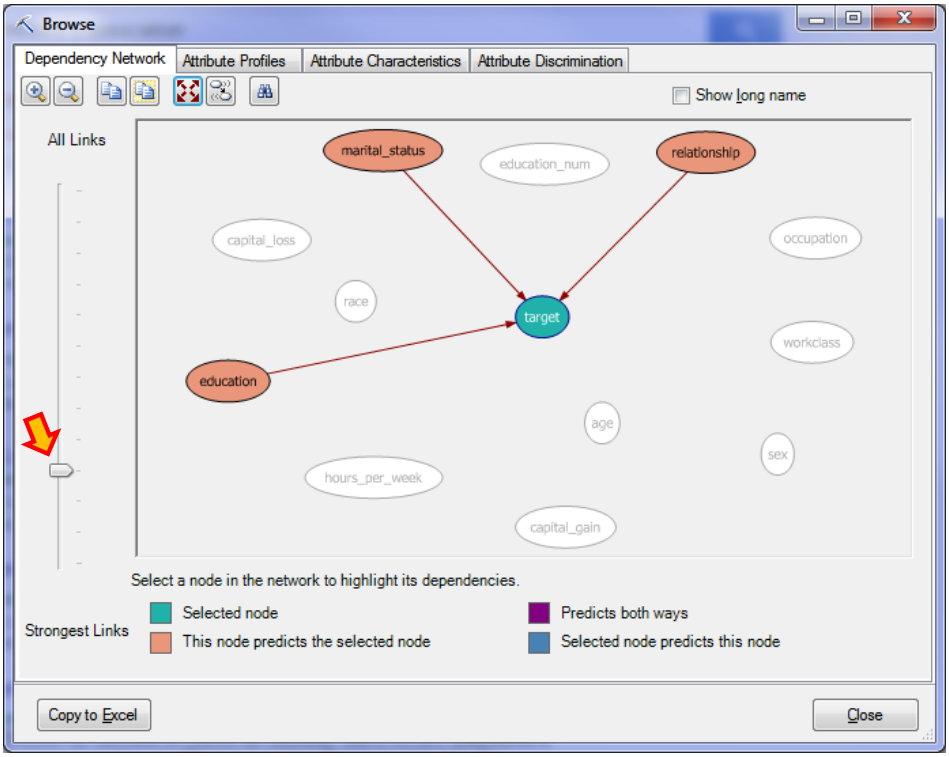
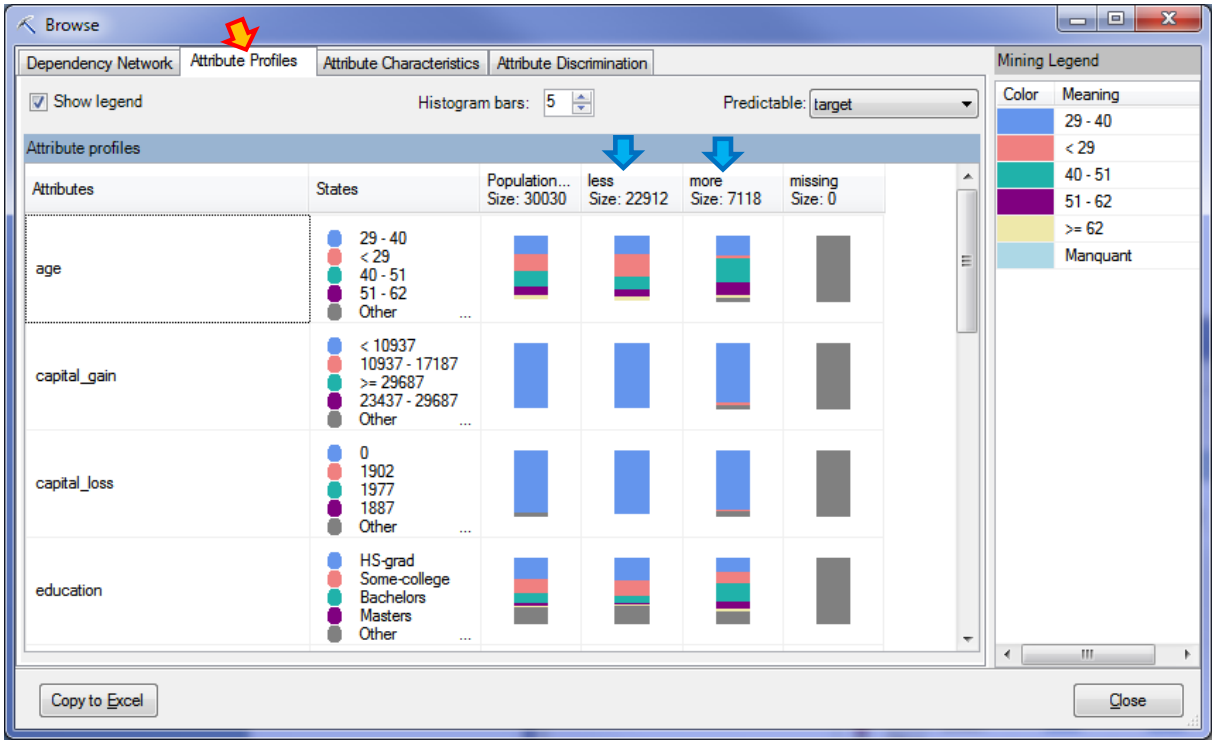


Figure 3 - Graphe des dépendances - Naive bayes

Si l'on se tient aux 3 premières variables, il semble que les plus pertinentes sont : « relationship », « marital_status », et « education ». Ce résultat est plus ou moins cohérent avec le premier arbre de décision où « relationship », « capital_gain », « education_num » étaient les plus décisifs (Figure 1). Que la hiérarchie des variables prédictive ne soit pas strictement identique n'est pas étonnant, les modèles reposent des structures de représentation de connaissances différentes. Il faut s'inquiéter cependant si des contradictions fortes apparaissent. Ce n'est pas le cas ici.

[14] « Microsoft Naive Bayes Algorithm », http://msdn.microsoft.com/en-us/library/ms174806.aspx - Qu'il faut absolument lire attentivement. Les férus de statistique trouveront leur bonheur dans la documentation technique.

Profils des attributs (Attributes Profiles). Cette fenêtre montre les distributions des modalités des variables conditionnellement aux valeurs de la cible.



Voyons le détail pour « age ». C'est une variable quantitative, elle a été automatiquement découpée en 5 intervalles (le nombre de classes est paramétrable[15]) : (1)]- ∞, 29] ; (2) [29, 40[; (3) [40 ; 51[; (4) [51, 62[; (5) [62, +∞[. Les histogrammes empilés indiquent la proportion des individus dans chaque intervalle, pour les modalités « less » et « more » de TARGET.

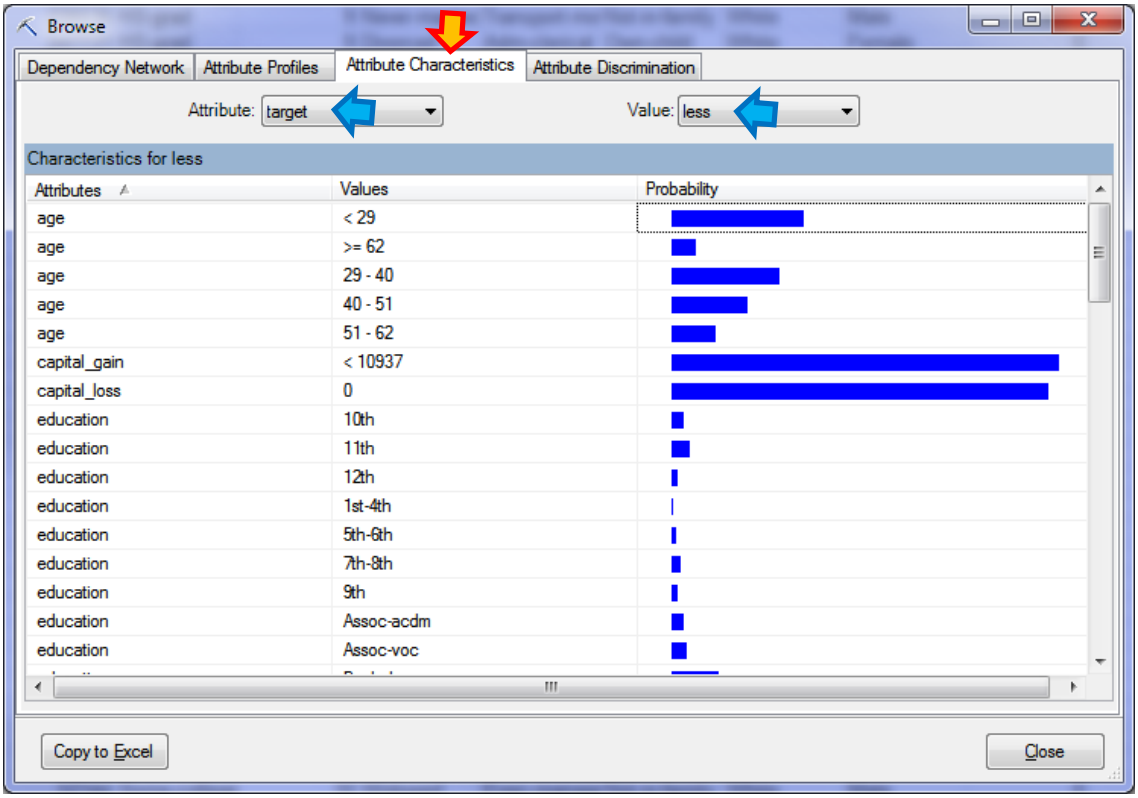
Nous disposons d'une autre perspective en formant à l'aide des tableaux croisées dynamiques (après recodage de « age ») d'Excel le tableau de contingence croisant « age » et « target ».

Nombre de targ	Étiquettes		
Étiquettes de	**less**	**more**	**Total général**
1	34.25%	4.83%	27.28%
2	28.07%	31.30%	28.84%
3	19.69%	37.47%	23.91%
4	11.37%	20.53%	13.54%
5	6.61%	5.87%	6.44%
Total général	**100.00%**	**100.00%**	**100.00%**

Effectivement, pour « less », les individus sont surtout présents dans le 1er intervalle ; pour « more », le 3ème est prépondérant.

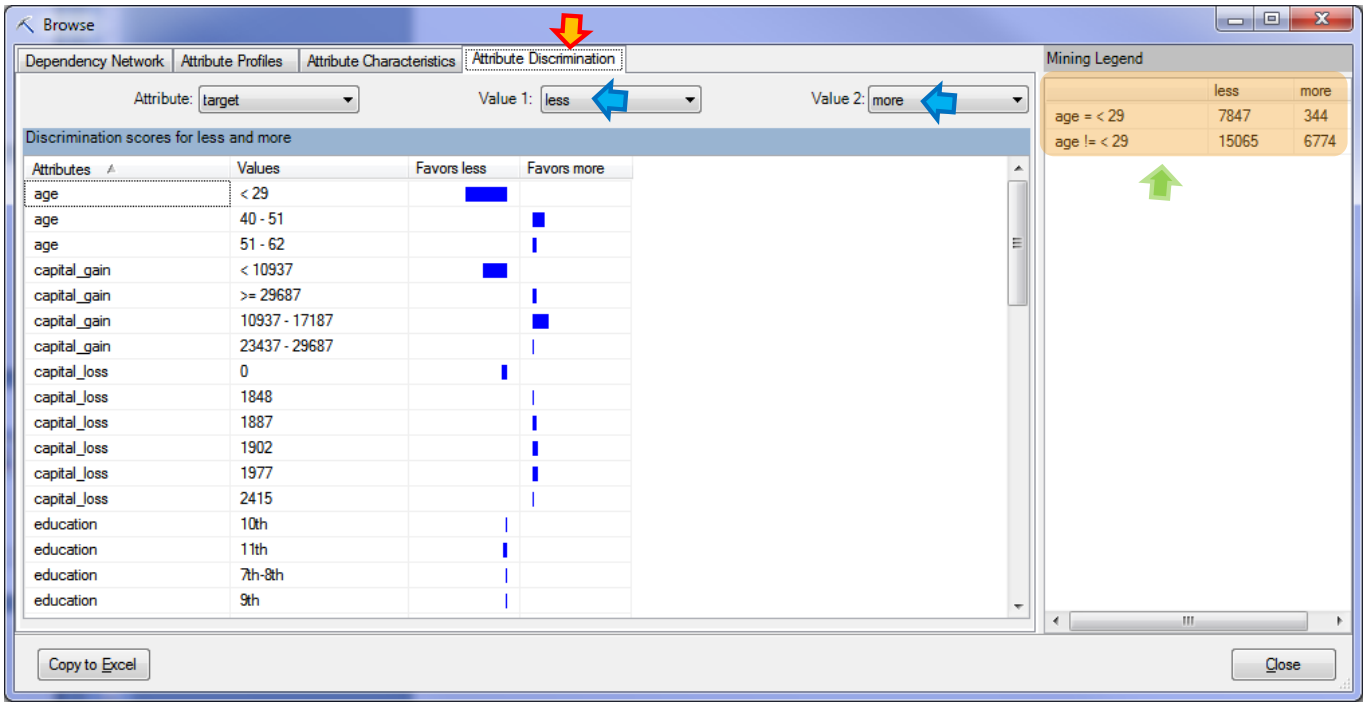
Caractéristiques des attributs (Attribute Characteristics). Les mêmes distributions conditionnelles sont disponibles dans l'onglet « Attribute Characteristics ».

[15] « Discretization Methods (Data Mining) », http://msdn.microsoft.com/en-us/library/ms174512.aspx



On note par exemple que l'intervalle]- ∞, 29] (< 29) est la plus fréquente pour la modalité « less » de la variable « target ».

Attribute Discrimination. Dans cet onglet est retracé l'impact de chaque modalité de variable prédictive (ou intervalle pour les quantitatives) dans la prédiction des valeurs de la cible. Voyons le détail pour l'intervalle]-∞, 29] de la variable « age ».



Le tableau de contingence est visible dans la partie droite de la fenêtre.

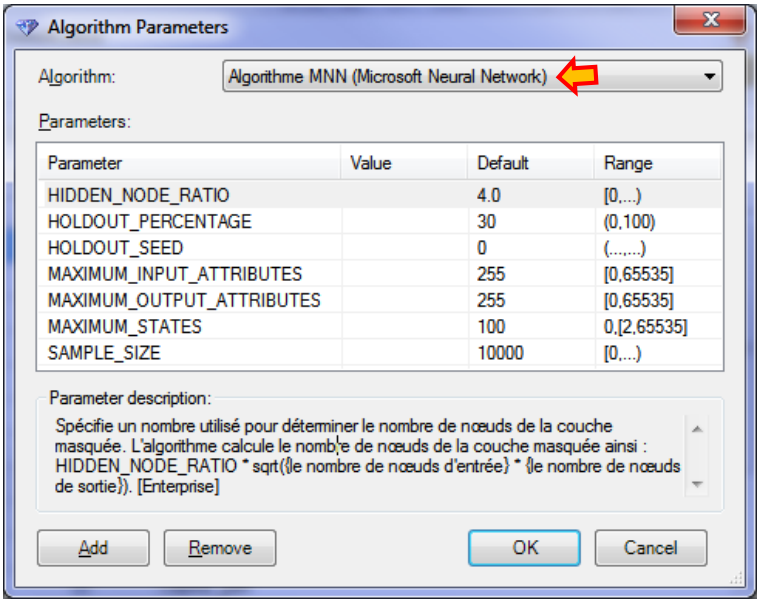
Lorsque nous la transformons en profils colonnes, la surreprésentation de « age < 29 » lorsque « target = less » (par rapport à « target = more ») apparaît clairement.

Nombre de target	Étiquettes		
Étiquettes de lignes	less	more	Total général
⊞age < 29	34.25%	4.83%	27.28%
⊞age >= 29	65.75%	95.17%	72.72%
Total général	100.00%	100.00%	100.00%

La documentation[16] n'est pas très diserte concernant la formule utilisée pour calculer l'importance sous-jacente à la barre affichée dans « Favors less ». Nous n'en saurons pas plus.

3.5 Réseaux de Neurones (Perceptron)

SSAS s'appuie sur un perceptron multicouche[17]. Voici le paramétrage utilisé.



Le nombre effectif de neurones dans la couche cachée dépend d'une formule savante intégrant le nombre de neurones en entrée et en sortie (HIDDEN_NODE_RATIO). Notons qu'une partie des observations peut être utilisée pour apprécier l'évolution de l'erreur en généralisation (HOLDOUT_PERCENTAGE), un échantillonnage est possible (SAMPLE_SIZE) pour l'apprentissage des poids synaptiques lorsque la taille des données est trop importante.

Nous nommons le modèle « Neural Network ». Le temps d'exécution est plus élevé par rapport aux précédents modèles.

Visualisation du modèle. Les sorties sont relativement succinctes pour le réseau de neurones. De nouveau, nous constatons qu'une discrétisation des variables quantitatives a été opérée. Nous observons dans la fenêtre l'influence des modalités (ou intervalles) des variables

[16] http://msdn.microsoft.com/en-us/library/dn458537.aspx

[17] http://msdn.microsoft.com/en-us/library/cc645901.aspx

prédictives sur l'une ou l'autre valeur de la cible. Ici non plus, nous ne savons pas vraiment comment sont calculées les valeurs sous-jacentes aux barres.

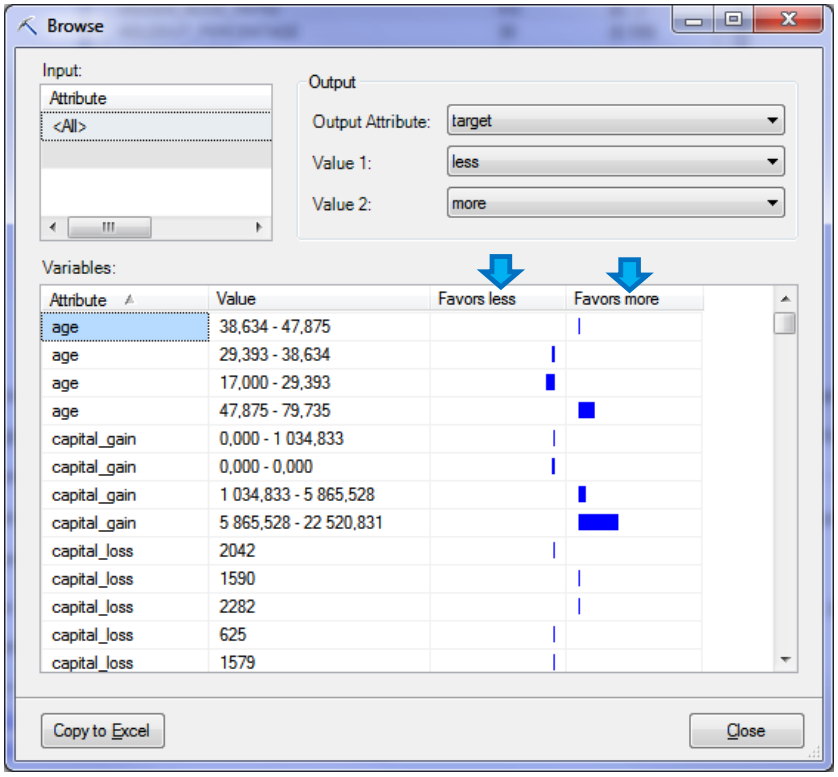
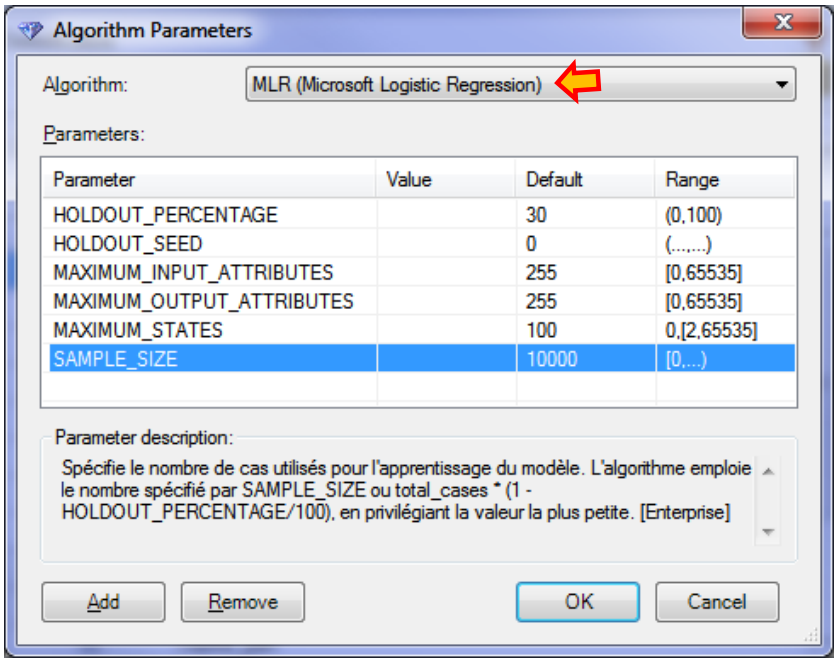


Figure 4 - Importance des variables - Réseau de neurones

3.6 Régression logistique

La régression logistique est une variante du réseau de neurones dans SSAS. Nous retrouvons à peu près les mêmes paramètres, à l'exception de la couche cachée.



De fait, nous disposons de la même fenêtre de visualisation.

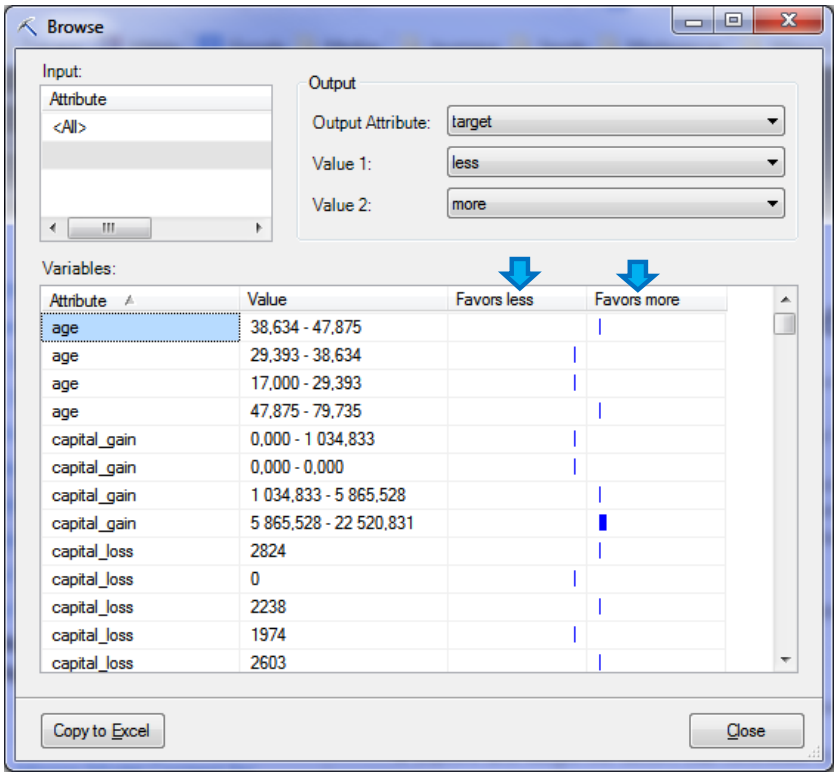


Figure 5 - Importance des variables - Régression logistique

Nous avons un classifieur linéaire avec la régression logistique. Le modèle étant de nature différente par rapport au perceptron multicouche, l'importance des variables n'est pas la même (comparer Figure 4 et Figure 5).

3.7 Comparaison des modèles

Suivant la procédure décrite à partir de la page 14 (Evaluation du modèle), nous avons appliqué les modèles sur l'échantillon test « adult.test ». Les taux d'erreur sont retranscrits dans le tableau suivant.

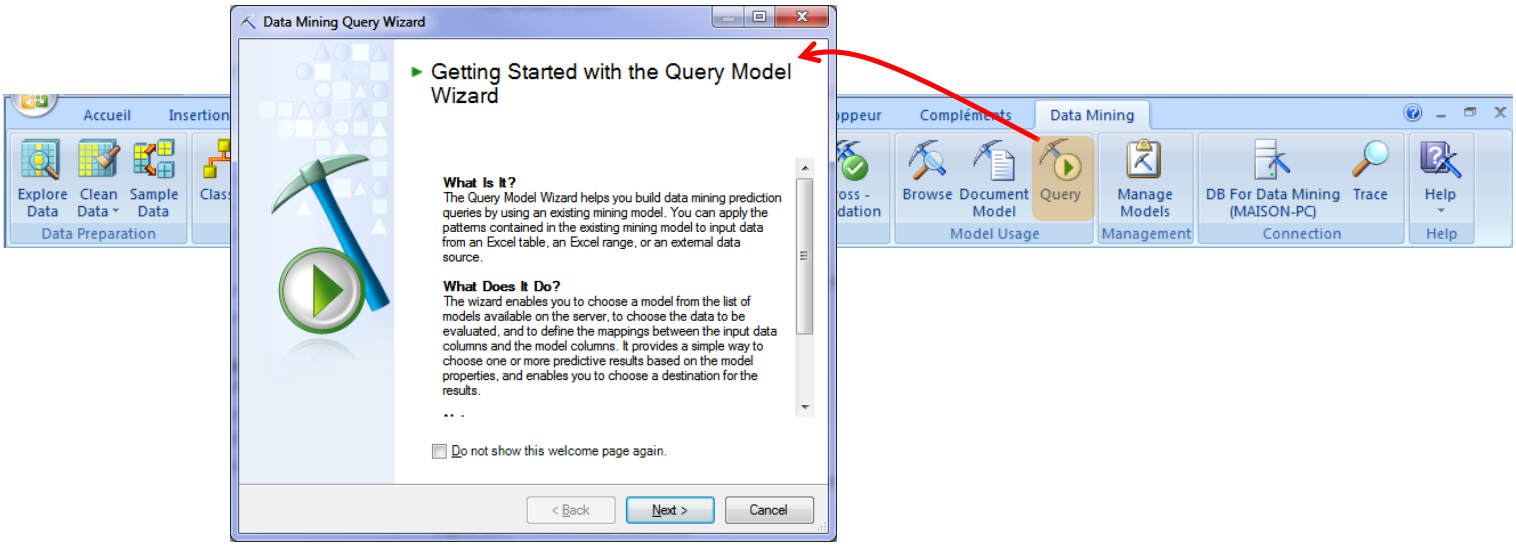
Modèle	Taux d'erreur en test	Nombre de mal classés
M1 - Decision Tree – Default Settings	15.87 %	1587
M2 - Decision Tree – Modified Settings	18.28 %	1828
M3 - Naive Bayes	17.36 %	1736
M4 - Neural Network	**14.69 %**	**1469**
M5 - Logistic Regression	**14.71 %**	**1471**

Le réseau de neurones (Neural Network) est le plus performant avec un taux d'erreur de 14.69%. Mais à bien y regarder, il ne présente que 2 individus mal classés en moins – sur les 10000 qui composent l'échantillon test – par rapport à la régression logistique (Logistic

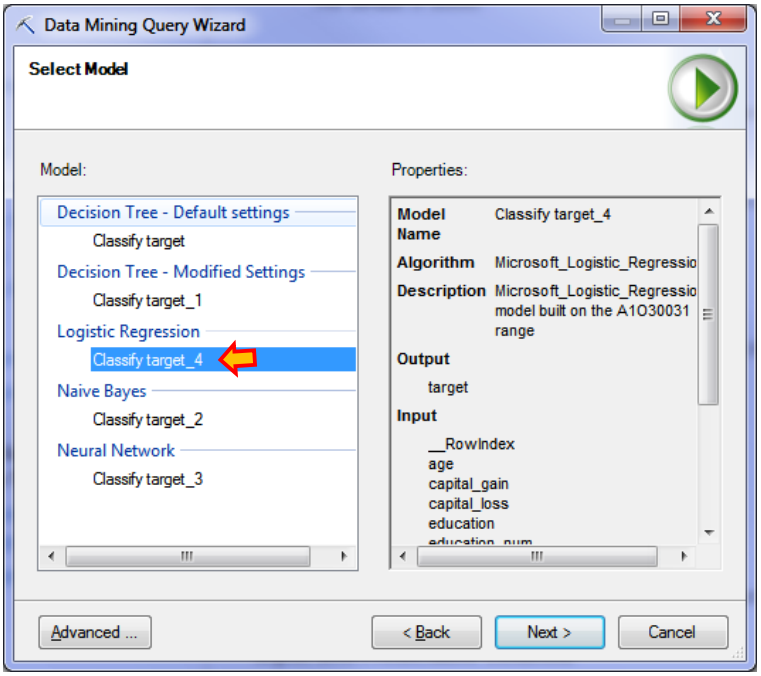
Regression). Dans une étude réelle, surtout dans les domaines où l'interprétation des résultats est importante, nous préférerons le dernier modèle parce qu'il est plus simple. Nous utiliserons donc la régression logistique lors du déploiement sur les individus non étiquetés.

3.8 Déploiement sur les individus non étiquetés

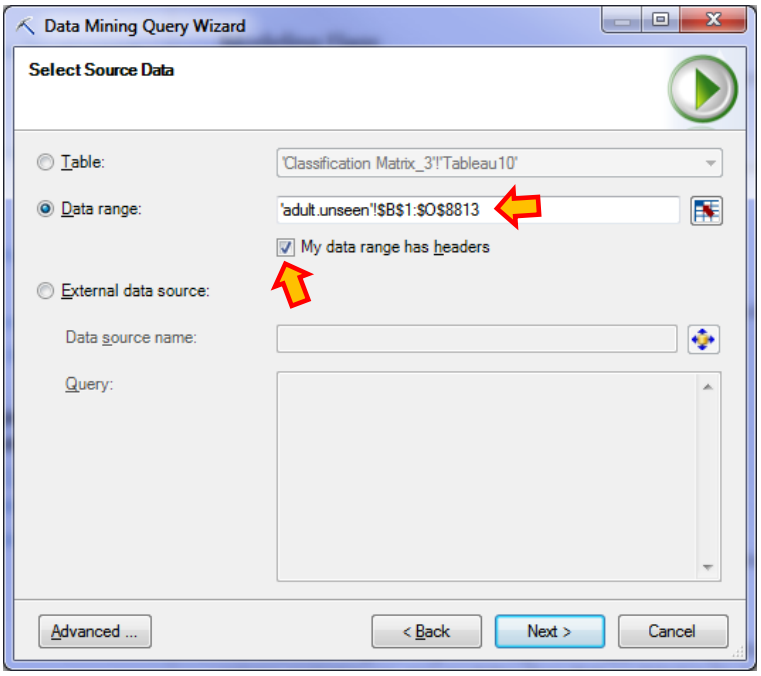
Nous souhaitons appliquer le modèle **M5** sur les individus non étiquetés de la feuille « adult.unseen ». Cette configuration correspond au déploiement du modèle dans la population. Sauf qu'au lieu de traiter individuellement chaque individu supplémentaire, nous effectuons plusieurs prédictions d'un coup (8812 observations). Nous actionnons le bouton QUERY (section MODEL USAGE) dans l'onglet Data Mining. Un wizard apparaît.



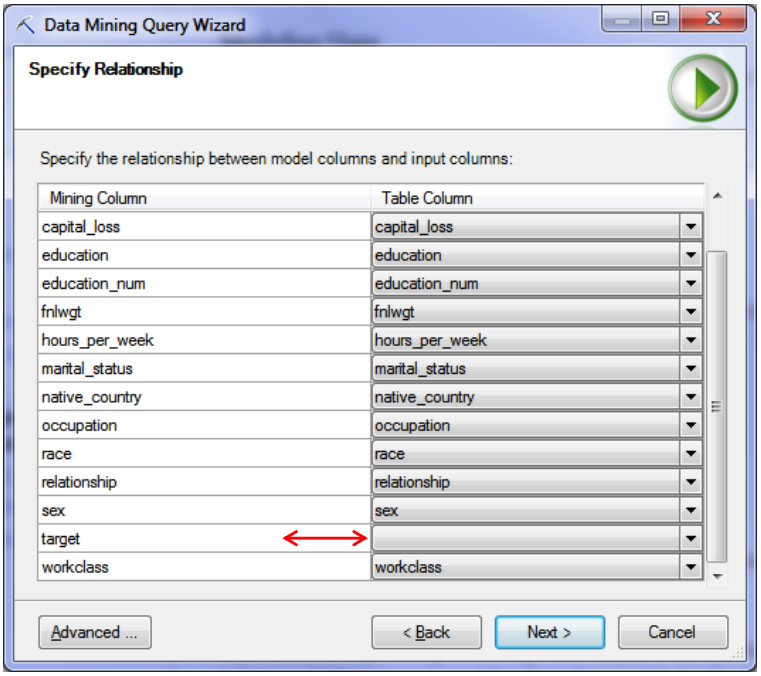
Nous sélectionnons le modèle « Logistic Regression ».



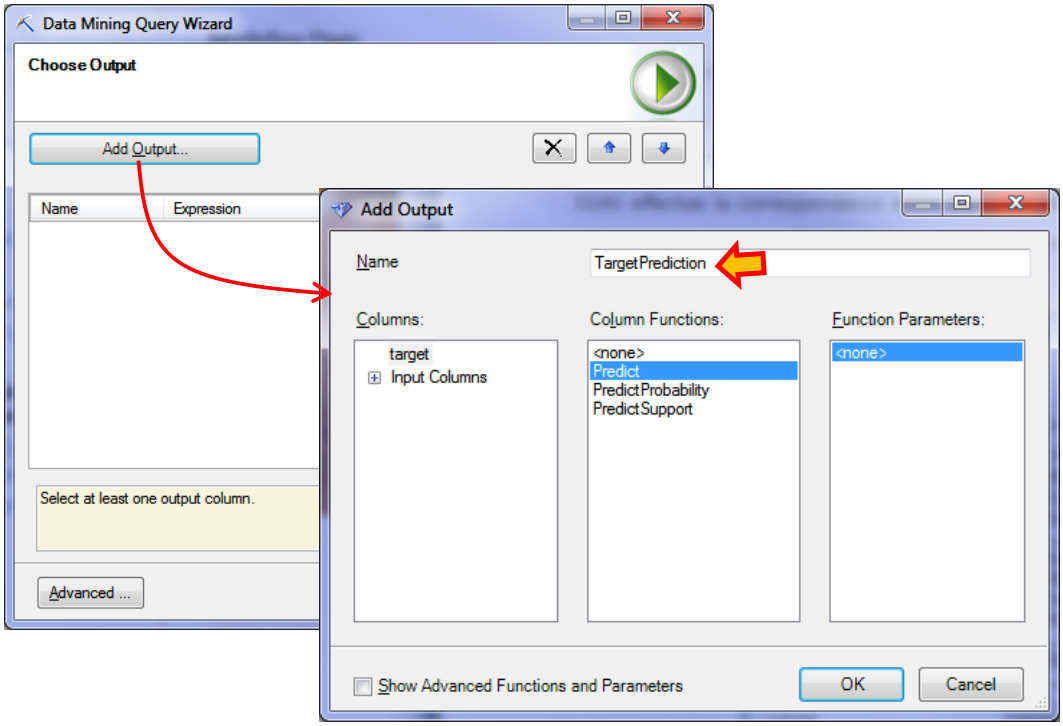
Nous spécifions la description des individus à prédire. Attention, il ne faut pas inclure ID dans la sélection.



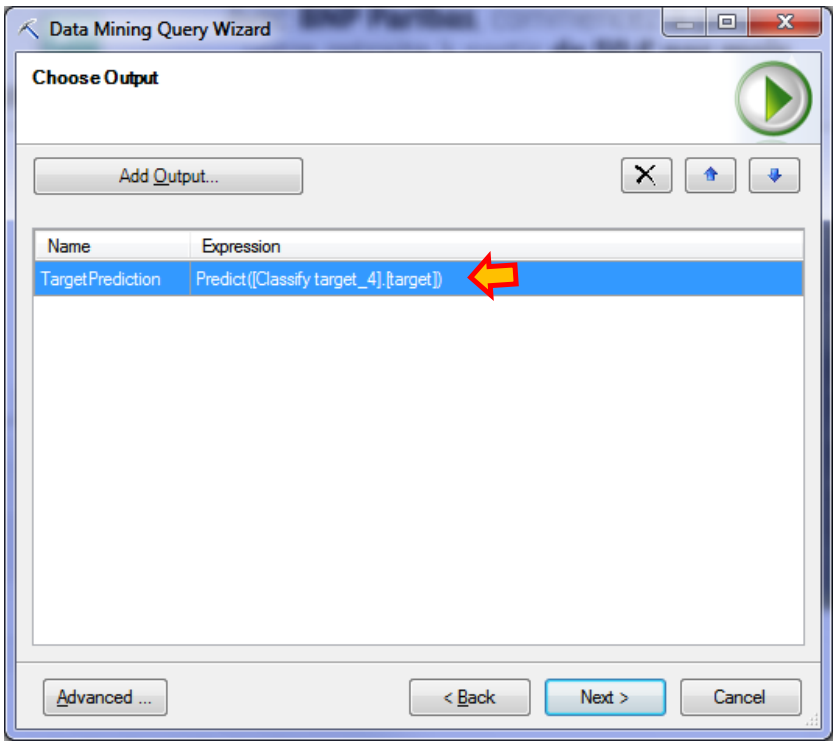
SSAS effectue la correspondance entre les noms des variables présents dans le modèle et les en-têtes de colonnes des données à prédire. Bien sûr, « target » n'a pas d'équivalent dans « adult.unseen ».



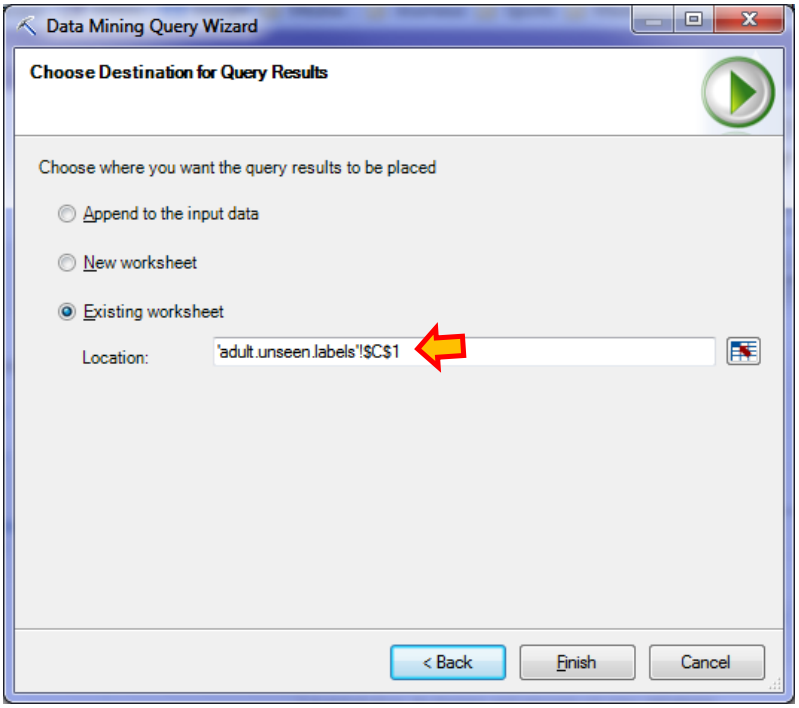
Le type de sortie à produire est spécifié dans la fenêtre suivante, en cliquant sur le bouton ADD OUTPUT.



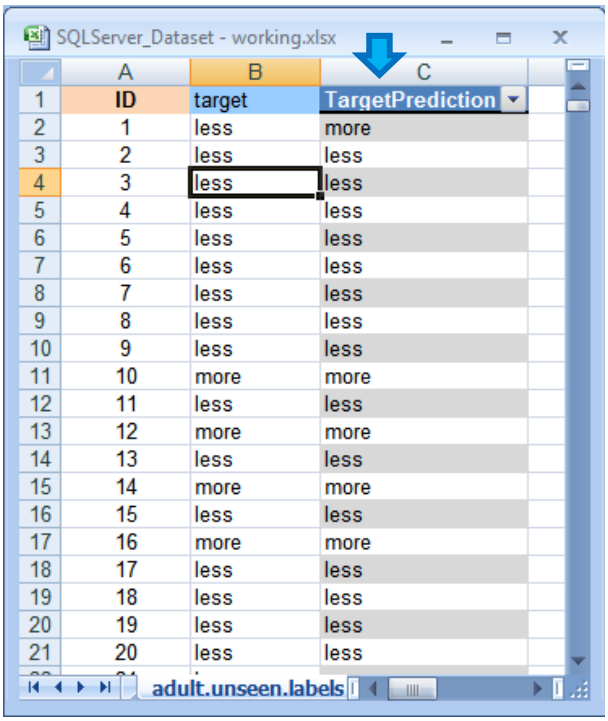
Nous souhaitons obtenir les classes prédites (Predict). La colonne s'appellera « TargetPrediction » dans Excel. Plusieurs sorties peuvent être produites (ex. probabilités d'affectation, etc.) comme nous le verrons plus loin.



Reste à définir l'endroit où devra être placée cette nouvelle variable. Nous sélectionnons la 3ème colonne de la feuille « adult.unseen.labels ». L'idée bien évidemment est de pouvoir comparer classes observées et prédites par la suite.



Nous cliquons sur FINISH. La prédiction est disponible. Voici les 20 premières observations.



	A	B	C
1	ID	target	TargetPrediction
2	1	less	more
3	2	less	less
4	3	less	less
5	4	less	less
6	5	less	less
7	6	less	less
8	7	less	less
9	8	less	less
10	9	less	less
11	10	more	more
12	11	less	less
13	12	more	more
14	13	less	less
15	14	more	more
16	15	less	less
17	16	more	more
18	17	less	less
19	18	less	less
20	19	less	less
21	20	less	less

En situation réelle, notre analyse s'arrête ici. Ce n'est que rétrospectivement que nous saurons si la prédiction a été correcte ou non. Pour l'instant, sur la base de la matrice de confusion calculée sur l'échantillon test, nous savons que nous avons 14.71% de chances (si on peut dire) de nous tromper pour chaque affectation.

Dans notre cadre expérimental, nous disposons des étiquettes dans « adult.unseen.labels », nous sommes en mesure de calculer la fréquence des erreurs.

3.9 Vérification des classements

Nous construisons un tableau croisé dynamique entre « target » et « TargetPrediction ». Nous suivons le formalisme de SSAS (Figure 2) c.-à-d. nous plaçons en colonne les modalités observées et en ligne les modalités prédites. Ce n'est pas l'usage dans les publications francophones, rappelons-le encore une fois.

Nombre de target	Étiquettes (		
Étiquettes de ligr	less	more	Total général
less	6267	891	7158
more	380	1274	1654
Total général	**6647**	**2165**	**8812**
Error rate	**14.42%**		

Le modèle s'est trompé dans **14.42%** des cas. Le taux mesuré sur l'échantillon test (**14.71%**) était assez fidèle. C'est tout l'intérêt d'utiliser un second échantillon n'ayant pas participé à la construction du modèle pour évaluer les performances en prédiction.

4 Apprentissage supervisé – Scoring (ciblage)

Le scoring est une activité clé du data mining. Nous sommes toujours dans le cadre de l'analyse prédictive, mais la nature du problème est un peu différente. Toujours sur les mêmes données, nous souhaitons cibler les individus à salaire élevé (la modalité positive est « target = more »). L'objectif est de répondre à la question suivante : si nous attribuons un score qui indique leur propension à être positif aux individus non étiquetés « adult.unseen », combien y aura-t-il de personnes effectivement « target = more » parmi les 440 premiers (c.-à-d. les 440 personnes qui présentent les scores les plus élevés) ?

Voici les principales étapes du processus :

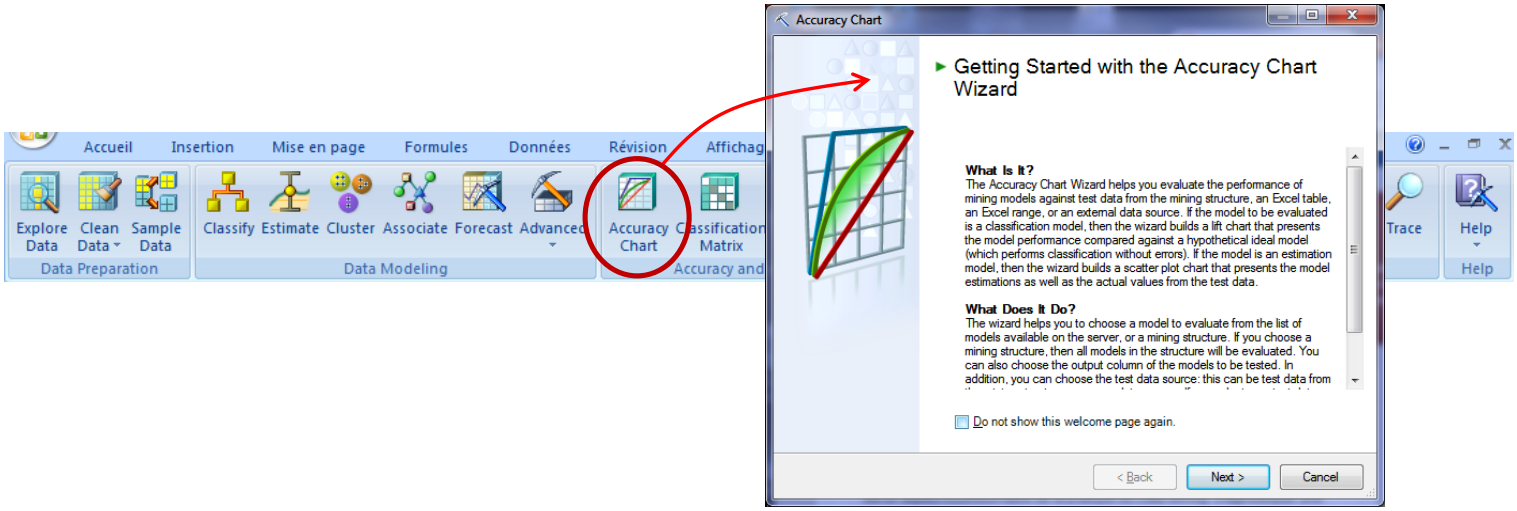
a. Nous avons besoin d'un modèle prédictif. Celui issu de la régression logistique construit précédemment (**M5**) conviendra très bien[18]. On sait que les modèles linéaires se comportent généralement très bien dans le scoring.
b. Nous utilisons les observations de l'échantillon test (adult.test) pour construire la courbe de gain (« Accuracy Chart » dans la terminologie SSAS) permettant d'évaluer les performances en ciblage. Nous pouvons répondre à la question ci-dessus à ce stade.
c. Nous appliquons le modèle sur l'ensemble non étiqueté « adult.unseen » pour attribuer un score (le degré de positivité) à chaque individu.

[18] Si le modèle n'était pas disponible, il suffisait de le construire à partir de l'échantillon d'apprentissage « adult.train » en suivant la procédure décrite dans la section 3.6.

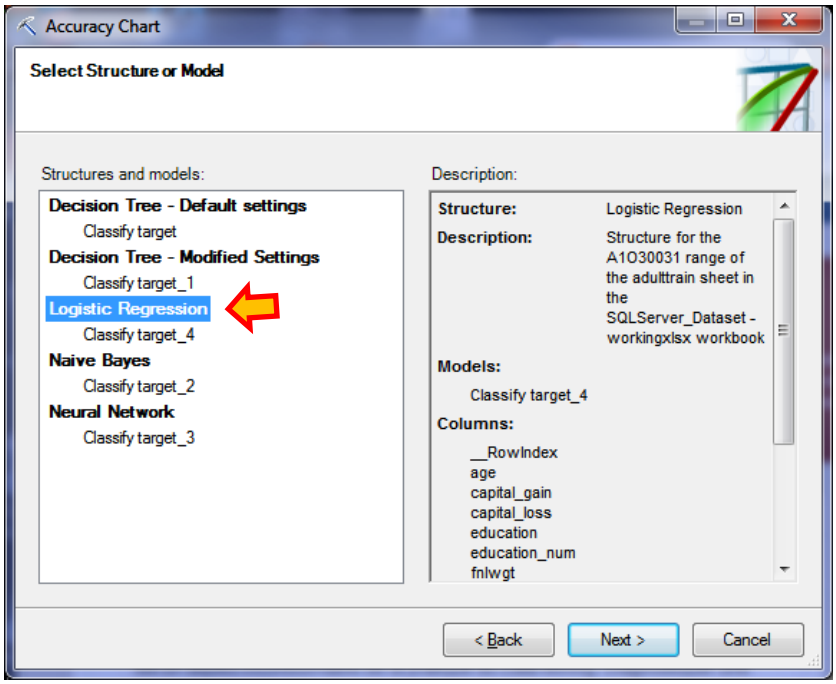
d. Et, luxe dont nous ne disposons pas dans les études réelles, nous pourrons vérifier à partir des étiquettes observées des individus « adult.unseen.labels.scoring » si la réponse calculée en (b) est suffisamment précise.

4.1 Construction de la courbe de gain

Nous cliquons sur l'icône ACCURACY CHART[19] (section Accuracy and Validation) de l'onglet Data Mining. Un wizard apparaît pour nous guider.

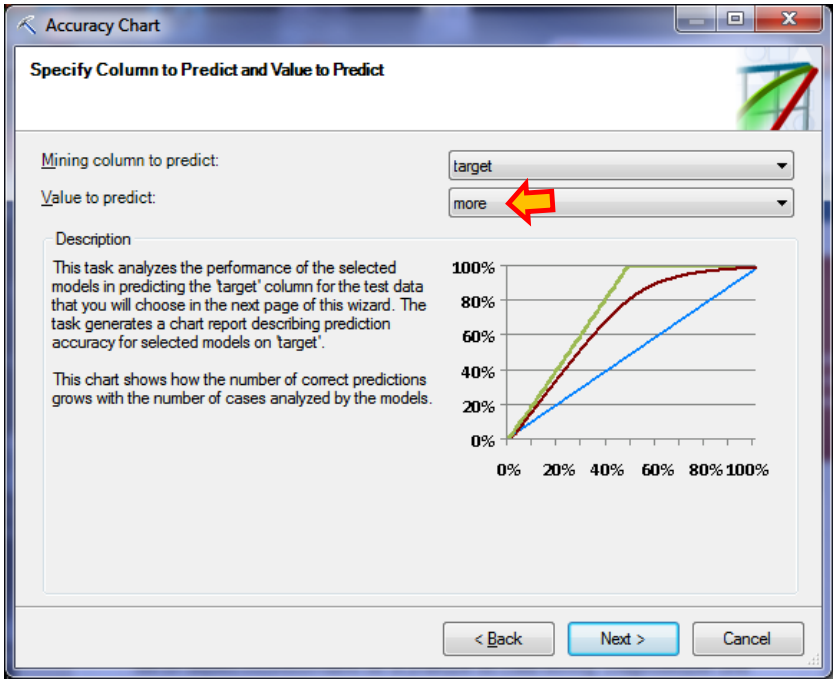


Nous sélectionnons le modèle à utiliser : « Logistic Regression ».

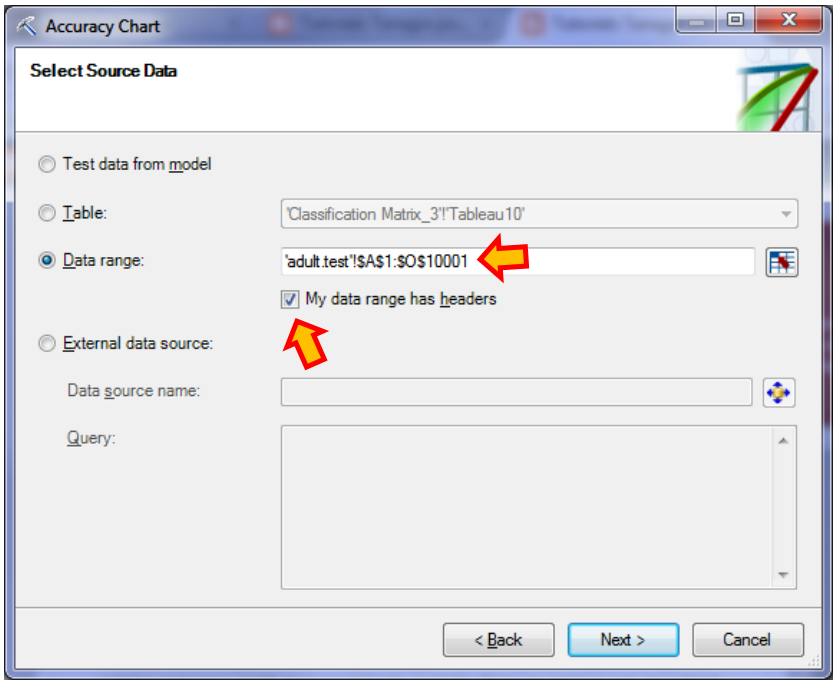


Nous précisons la modalité de « target » à cibler : « more » est la modalité positive.

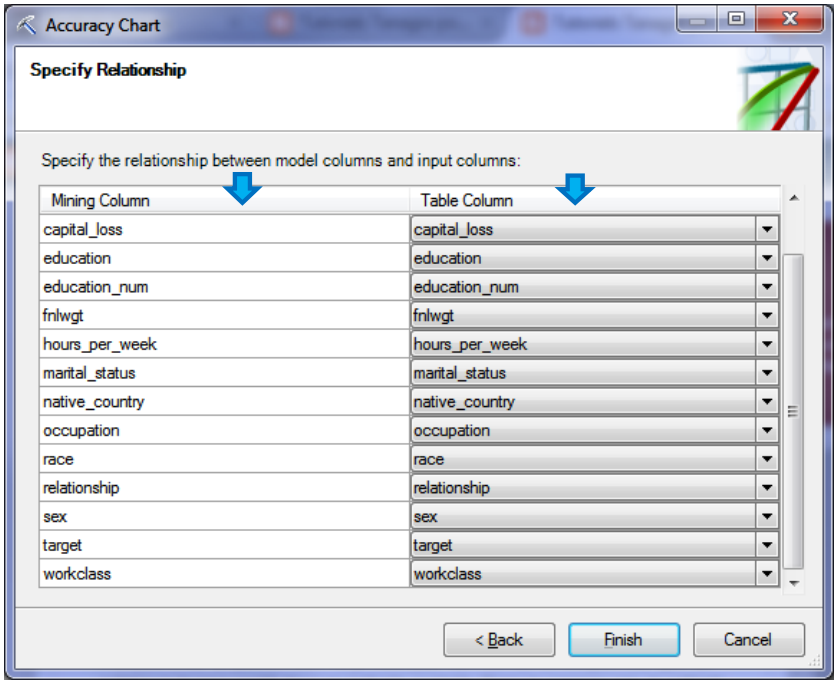
[19] L'aide en ligne fait mention également de LIFT CHART (http://technet.microsoft.com/en-us/library/ms175428.aspx).



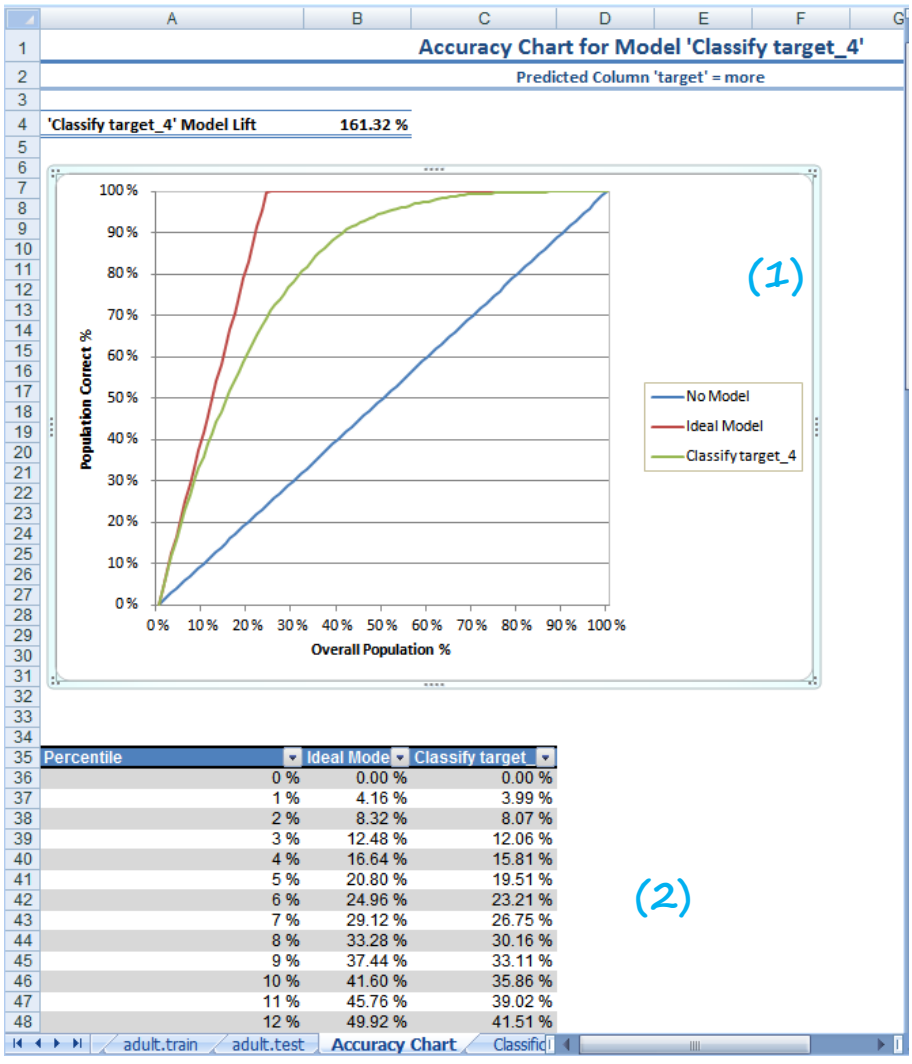
Nous sélectionnons l'échantillon test situé dans la feuille « adult.test ».



L'outil effectue les correspondances entre les variables du modèle et celles de l'ensemble test à partir de leurs noms.



La courbe de gain (1) et le détail des valeurs (2) sont insérés dans une nouvelle feuille « Accuracy Chart » lorsque nous cliquons sur FINISH.



Comment répondre à notre question initiale (page 30) à partir de ces éléments ?

- Nous disposons de n_t = 10000 individus dans l'échantillon test (adult.test) ;
- Dont $n_t(-)$ = 7596 « less » et $n_t(+)$ = 2404 « more »

Étiquettes de lign	Nombre de
less	7596
more	2404
Total général	**10000**

- Il y a n_u = 8812 observations dans l'échantillon non étiqueté (adult.unseen) qui représente la population.
- On *estime*, par une simple règle de trois, à $n_u(+)$ =2404 /10000 x 8812 = 2118.4048 le nombre de positifs (target = more) présents dans « adult.unseen ».
- 440 individus parmi 8812 correspondent (à peu près) 5% de la population[20].
- En nous référant à la courbe de gain, ou au tableau des valeurs pour une meilleure précision, nous constatons que cela représente 19.51% des individus positifs dans le ciblage élaboré à partir de la régression logistique.

Percentile	Ideal Mode	Classify targ
0 %	0.00 %	0.00 %
1 %	4.16 %	3.99 %
2 %	8.32 %	8.07 %
3 %	12.48 %	12.06 %
4 %	16.64 %	15.81 %
5 %	20.80 %	**19.51 %**
6 %	24.96 %	23.21 %
7 %	29.12 %	26.75 %
8 %	33.28 %	30.16 %
9 %	37.44 %	33.11 %
10 %	41.60 %	35.86 %

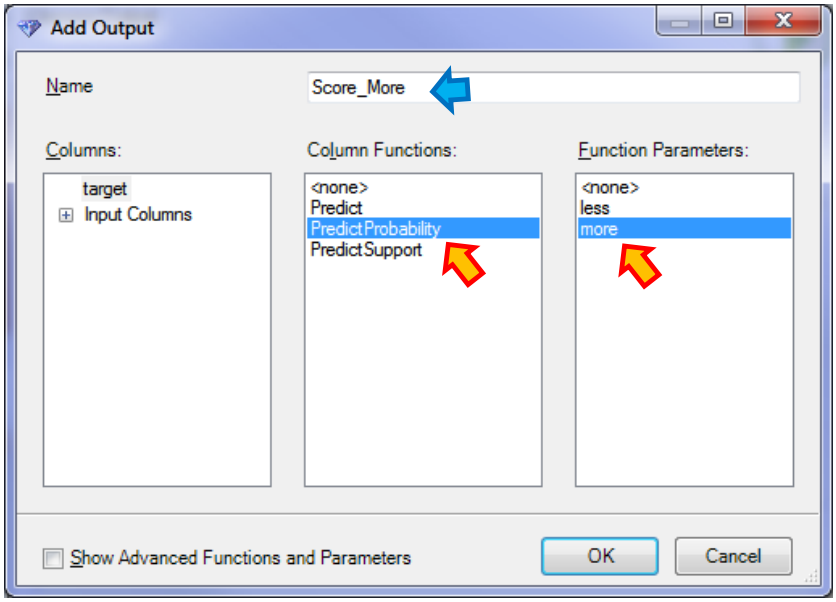
- Soit, 19.51% x 2118.4048 ≈ 413 individus positifs.

Conclusion : si l'on applique le modèle pour calculer les scores des individus dans une population de n_u = 8812 individus, on pense trouver (en espérance) 413 individus positifs (target = more) parmi les 440 individus présentant les scores les plus élevés. Vérifions tout de suite si notre dispositif tient la route.

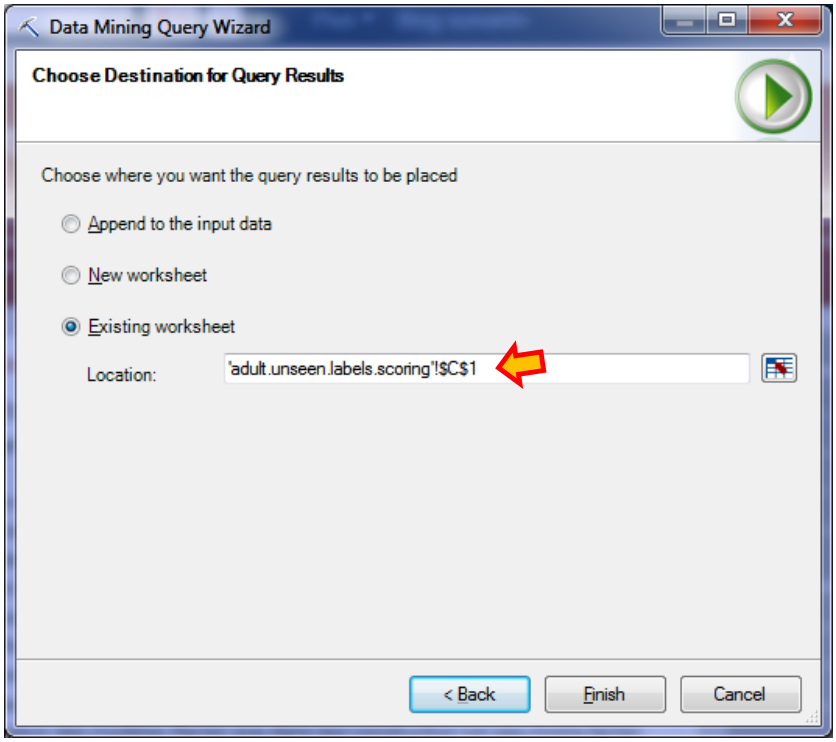
4.2 Déploiement sur les données non étiquetés

Nous déployons le même modèle « Logistic Regression » sur les individus non étiquetés. Comme plus haut (section 3.8), nous actionnons le bouton QUERY de la section « model Usage ». Sauf que lors de la définition de l'OUTPUT (Bouton ADD OUTPUT), nous choisissons de calculer la probabilité d'être positif que nous nommons « Score_More ».

[20] Il faut utiliser une interpolation linéaire si on tombe entre 2 percentiles.



Que nous choisissons de placer dans la colonne C de la feuille « adult.unseen.labels.scoring ».

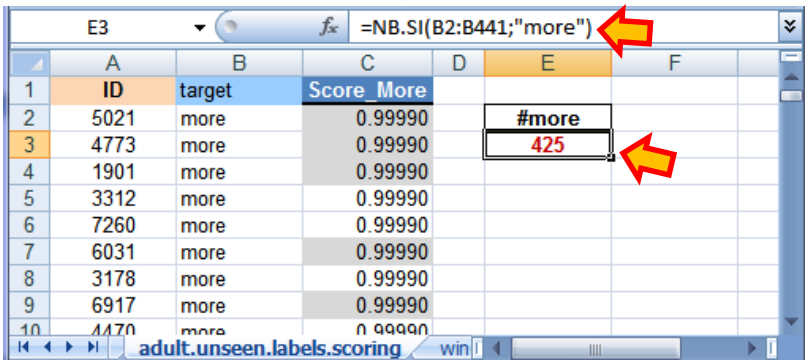


La colonne des scores est insérée dans la feuille. Voici les 10 premières lignes.

ID	target	Score_More
1	less	0.538066053
2	less	0.026390785
3	less	0.335151212
4	less	0.004378306
5	less	0.267170992
6	less	0.124751653
7	less	0.044880848
8	less	0.002986321
9	less	0.023183426
10	more	0.983735107

4.3 Vérification à partir des étiquettes disponibles

Nous disposons des scores et des étiquettes observées. Nous pouvons vérifier notre assertion ci-dessus. Nous trions la base de manière décroissante selon le score, et nous calculons le nombre de positifs (target = more) parmi les 440 premiers à l'aide de la fonction NB.SI(.).



	A	B	C	D	E	F
1	ID	target	Score_More			
2	5021	more	0.99990		#more	
3	4773	more	0.99990		425	
4	1901	more	0.99990			
5	3312	more	0.99990			
6	7260	more	0.99990			
7	6031	more	0.99990			
8	3178	more	0.99990			
9	6917	more	0.99990			
10	4470	more	0.99990			

Nous obtenons 425 individus « more » au final. On pensait en trouver 413. On peut penser que l'écart est important. Pas tant que ça en réalité. Il est nettement moins flagrant lorsqu'on passe en pourcentages : il y a 2165 individus « more » en tout dans l'échantillon, 425 / 2165 = 19.36%. La courbe de gain nous indiquait 19.51%.

5 Régression

La régression consiste à prédire et expliquer une variable cible quantitative (variable dépendante, variable endogène) à partir d'un ensemble de variables prédictives (variables indépendantes, variables exogènes). Nous évaluerons la régression linéaire multiple dans ce tutoriel. SASS propose une implémentation très singulière car, pour lui, la régression linéaire est un cas particulier des arbres de régression[21].

5.1 Données à traiter

On cherche à expliquer la qualité (quality) du vin, une note comprise entre 0 et 10, à partir de ses propriétés physico-chimiques (acidité, teneur en sucre, teneur en alcool, etc.)[22]. L'échantillon d'apprentissage contient 1500 observations (wine.regression.train), l'échantillon test 3398 (wine.regression.test).

Nous allons au plus simple dans notre étude. Nous considérons que la cible « quality » est quantitative. D'autres approches auraient été possibles. Par exemple, comme les valeurs sont entières, nous aurions pu tout aussi bien l'exploiter comme une variable qualitative ordinale. La documentation y fait mention. La régression linéaire de SSAS sait apparemment manipuler

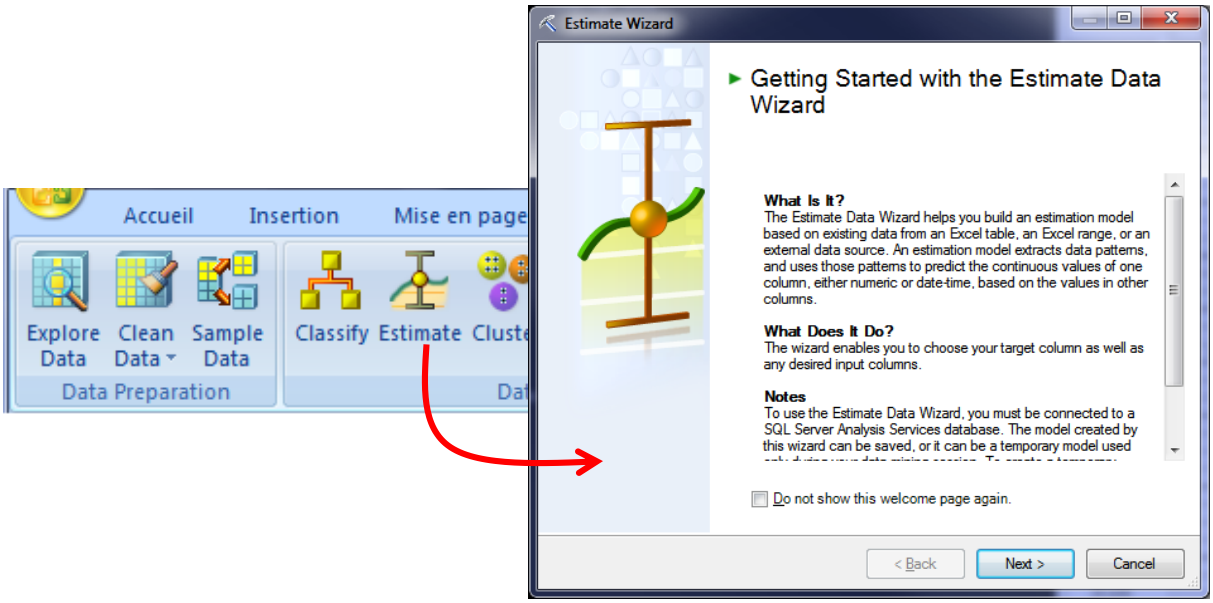
[21] http://technet.microsoft.com/en-us/library/cc645871.aspx

[22] https://archive.ics.uci.edu/ml/datasets/Wine+Quality ; seuls les vins blancs sont utilisés dans notre étude.

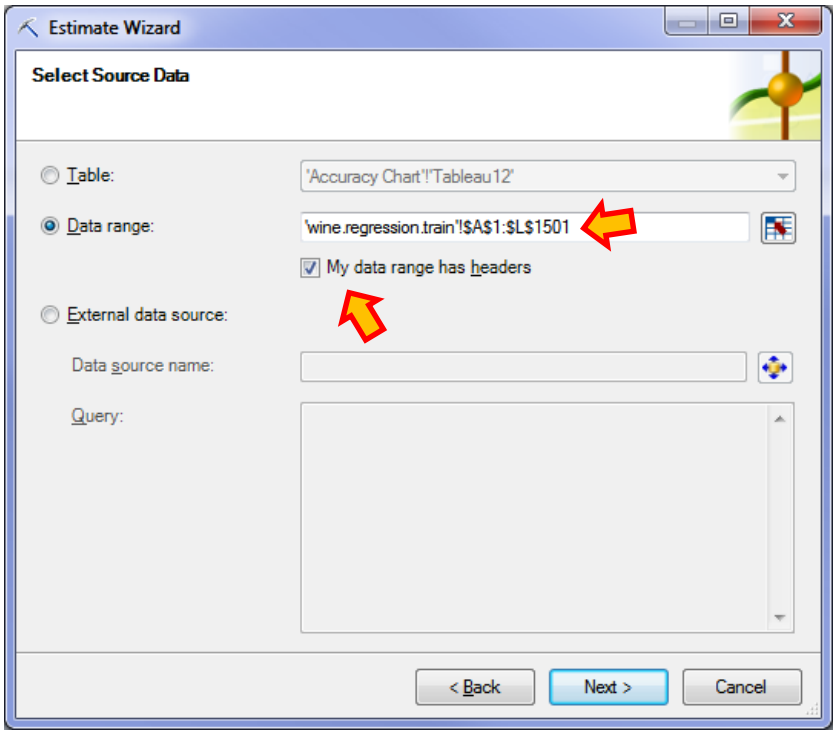
ce type de variable dépendante sans donner plus d'explications. Nous verrons ce qu'il en sera en déploiement. A priori, nous devrions prédire des valeurs non entières si nous appliquons directement l'équation de régression.

5.2 Modélisation

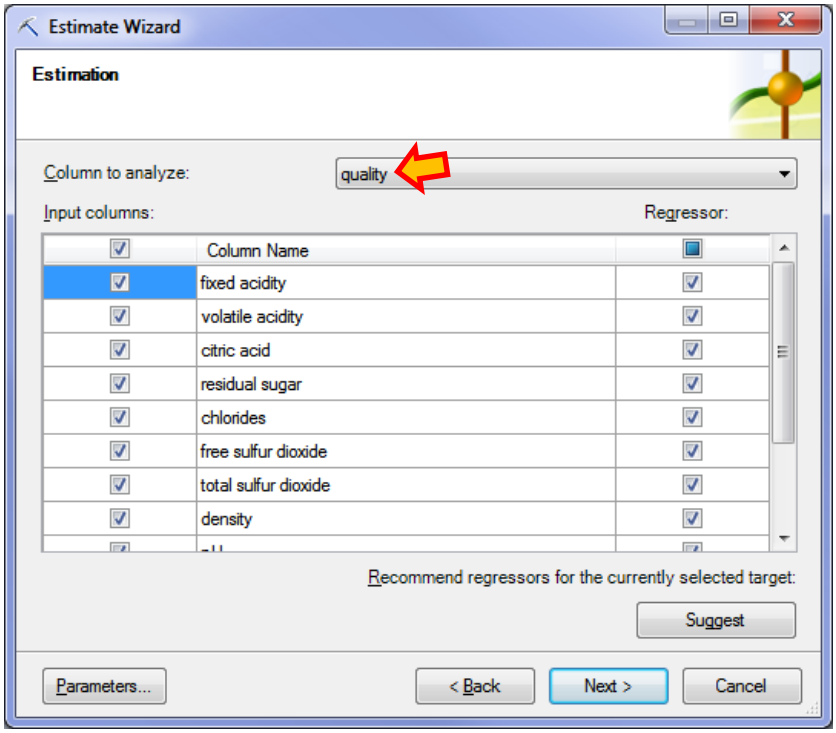
Nous cliquons sur l'icône ESTIMATE (section DATA MODELING) du ruban « Data Mining ». Un wizard prend en main le processus.



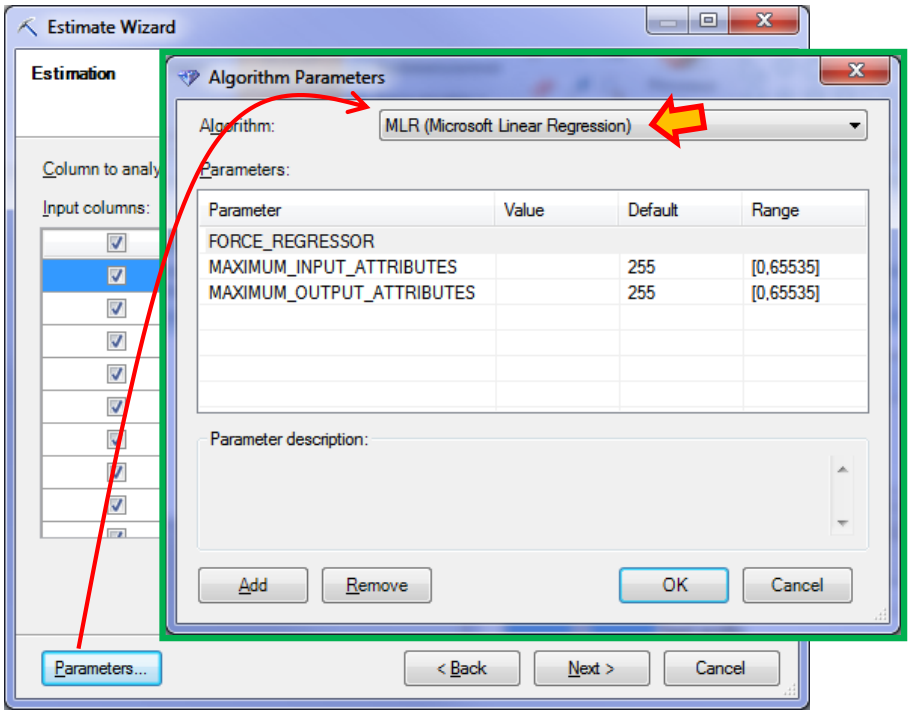
La plage de données d'apprentissage est sélectionnée dans la fenêtre suivante.



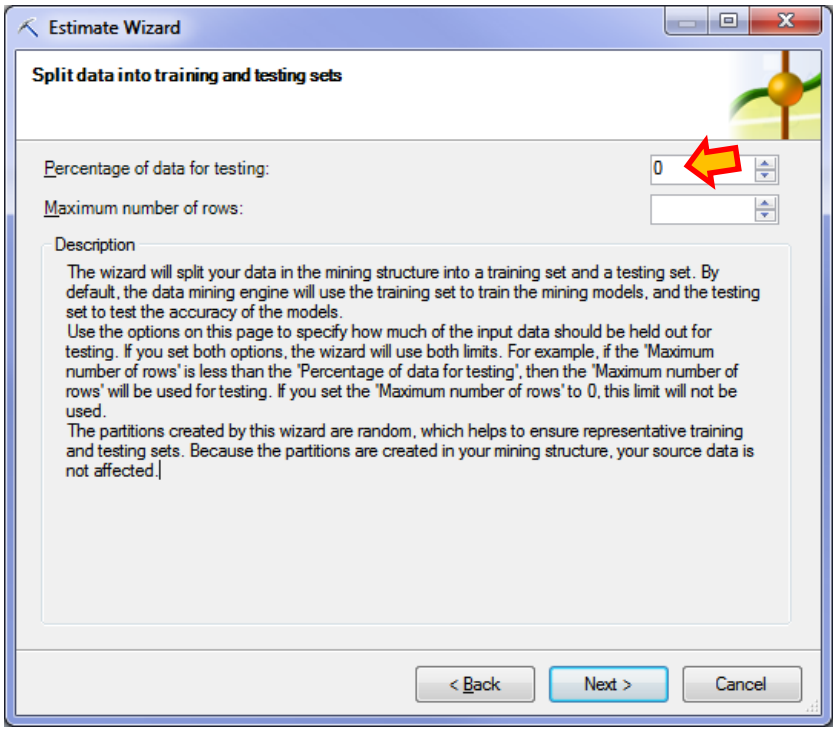
Puis nous désignons la variable cible QUALITY.



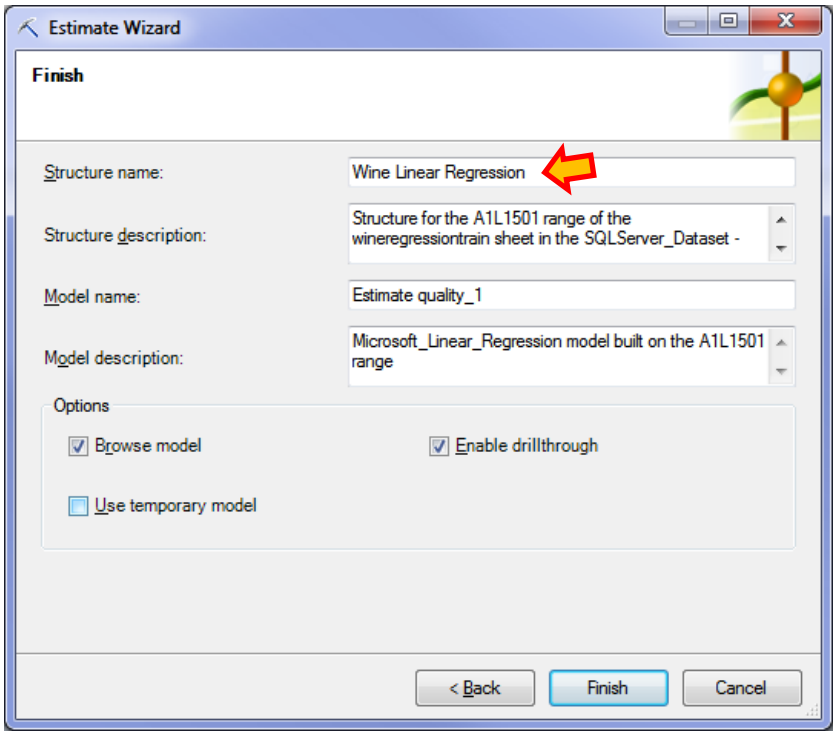
En cliquant sur le bouton PARAMETERS, nous pouvons choisir la technique de modélisation MLR (Microsoft Linear Regression).



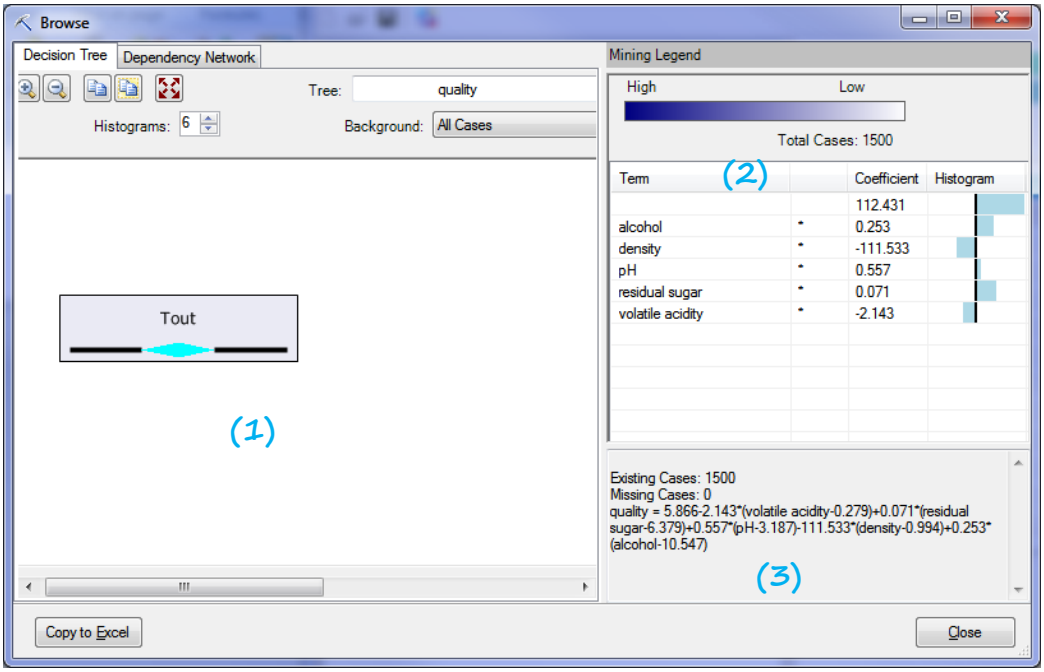
Tout comme lors de l'apprentissage supervisé, nous pouvons réserver une partie des observations pour l'évaluation du modèle. Ce n'est pas nécessaire en ce qui nous concerne puisque nous avions explicitement scindé les données en échantillons d'apprentissage et de test sur deux feuilles Excel distinctes. Nous mettons 0 pour « Percentage data for testing ».



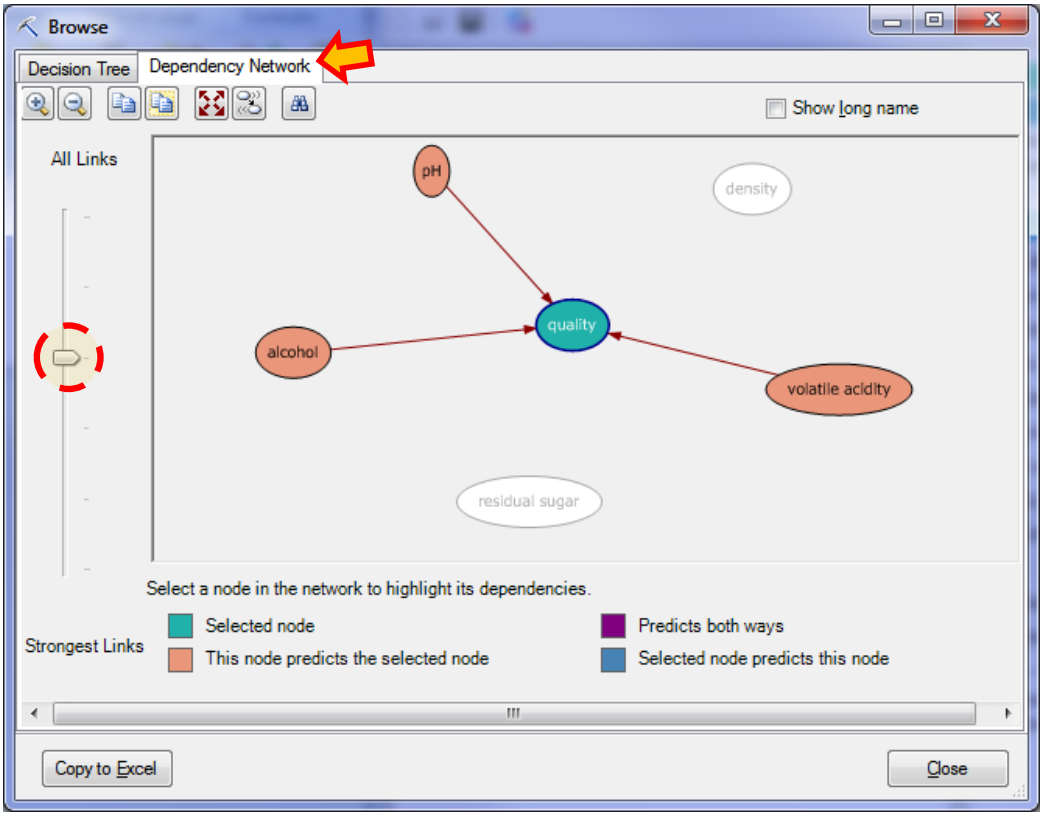
Nous sauvegardons le modèle sous le nom « Wine Linear Regression ».



La fenêtre de visualisation dérouterait n'importe quel féru d'économétrie. SSAS affiche la racine d'un arbre (1), avec en haut à droite les coefficients non standardisés de l'équation de régression (2), et en bas à droite, la même équation sur variables centrées (3). SASS a automatiquement effectué une sélection, seules 5 variables ont été retenues.



Pour se faire une idée de la hiérarchie des variables dans le modèle, nous allons dans l'onglet « Dependency Network ». Voici les 3 variables les plus influentes.



Le résultat n'est pas vraiment cohérent avec l'histogramme proposé dans (2) de la fenêtre « Decision Tree », on a du mal à en saisir la signification en définitive. Du reste, il faut avouer que le mode de présentation des résultats de la régression adopté par SSAS n'est quand même pas très habituel. Il est même déroutant je dirais.

5.3 Déploiement – Evaluation sur l'échantillon test

Remarque introductive : Mesurer les performances d'un modèle sur un échantillon test n'est pas très répandu en économétrie. Cette pratique nous vient de la culture « machine learning ». Pourtant, elle est justifiée car elle permet de mettre sur un pied d'égalité des modèles avec des complexités différentes, ou même s'appuyant sur des représentations différentes. Je vois cependant que les mentalités évoluent. Des critères comme le PRESS (Predicted Residual Sum Of Squares[23]) apparaissent de plus en plus souvent dans les ouvrages. Cette statistique applique le principe de leave-one-out, une forme particulière de la validation croisée. Son énorme intérêt en régression linéaire multiple est que l'on peut l'obtenir sans avoir à calculer explicitement autant de modèles (en retirant une observation à chaque fois) qu'il y a d'observations dans la base.

Pour l'évaluation de la régression, SSAS propose via l'icône « Accuracy Chart », qui produit donc un calcul différent selon le contexte (cf. Courbe de gain – section 4.1), un graphique nuage de points[24] croisant les valeurs observées et prédites sur l'échantillon test (Figure 6). Aucun indicateur numérique n'est mis en avant. C'est quand même très étonnant.

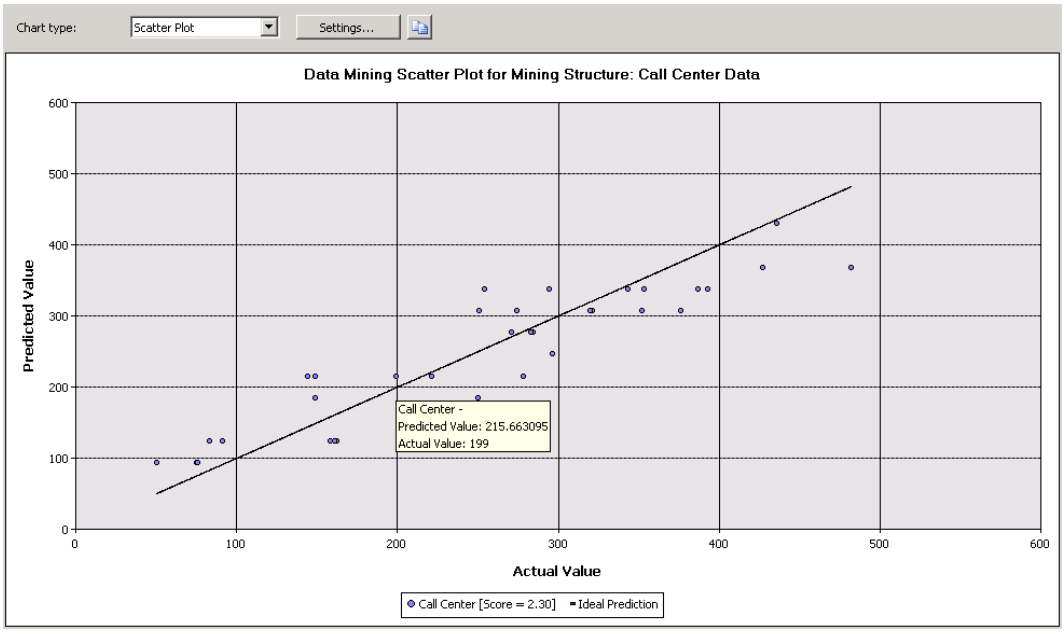


Figure 6 - Exemple de nuage de points pour évaluer une régression (Exemple MSDN)

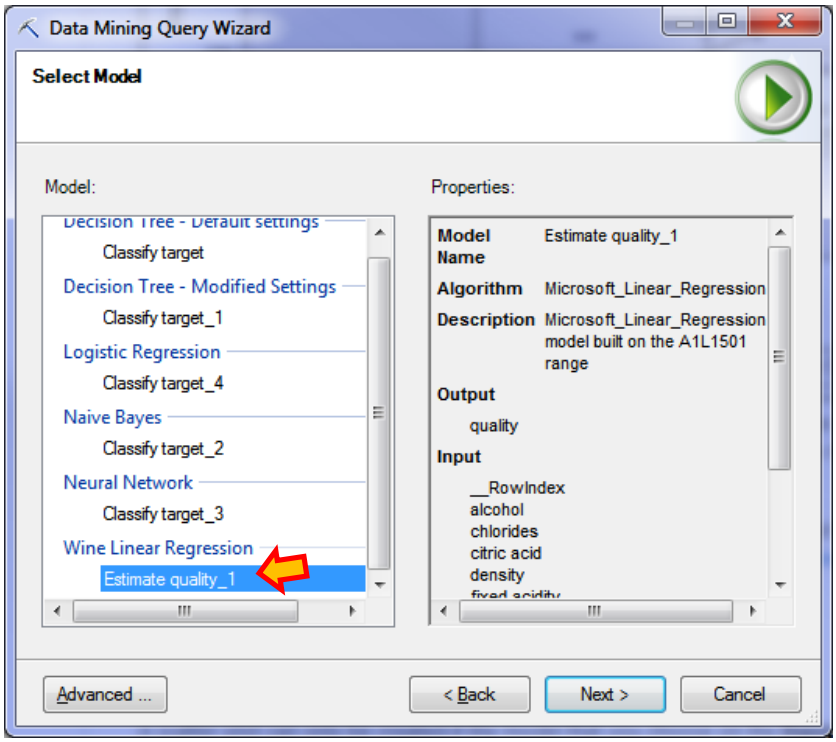
Pour dépasser ces insuffisances, nous allons produire manuellement les prédictions sur l'échantillon test, puis calculer nos propres indicateurs. Le tableur se prête parfaitement à ce type de manipulations additionnelles. Obtenir le détail de la démarche d'évaluation sera

[23] http://en.wikipedia.org/wiki/PRESS_statistic

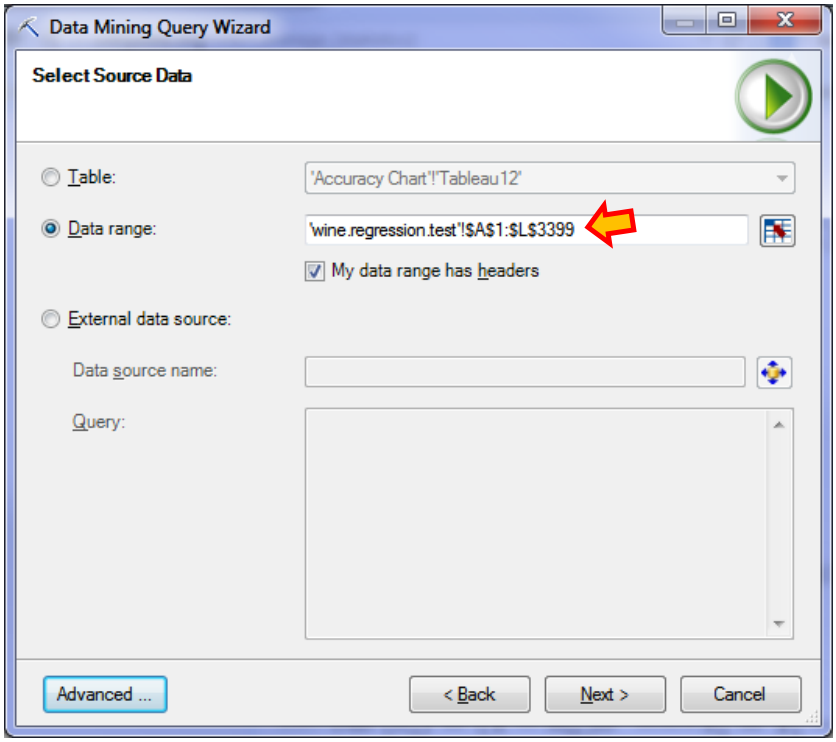
[24] « Scatter Plot » - http://technet.microsoft.com/en-us/library/bb895169.aspx

d'autant plus intéressant que la variable cible prenant des valeurs entières, j'étais vraiment curieux de scruter le comportement de SSAS en déploiement dans cette configuration (il ignore cette information ? il arrondit à l'entier le plus proche ? …).

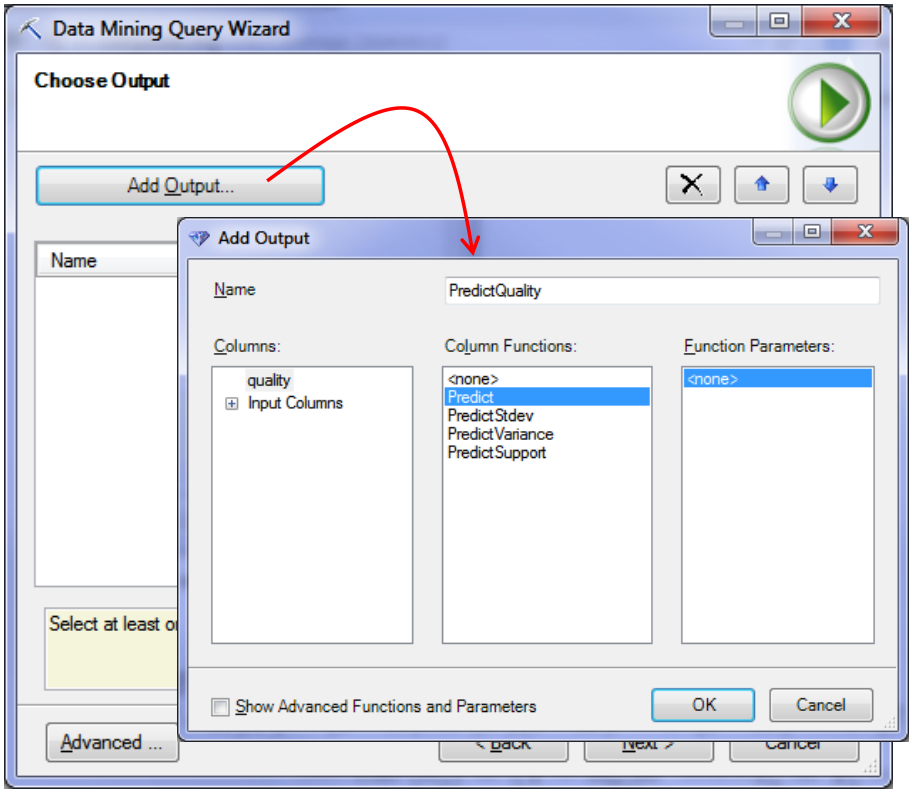
Déploiement sur l'échantillon test. Nos actionnons le bouton QUERY. Dans la fenêtre de sélection des modèles, nous choisissons « Wine Linear Regression ».



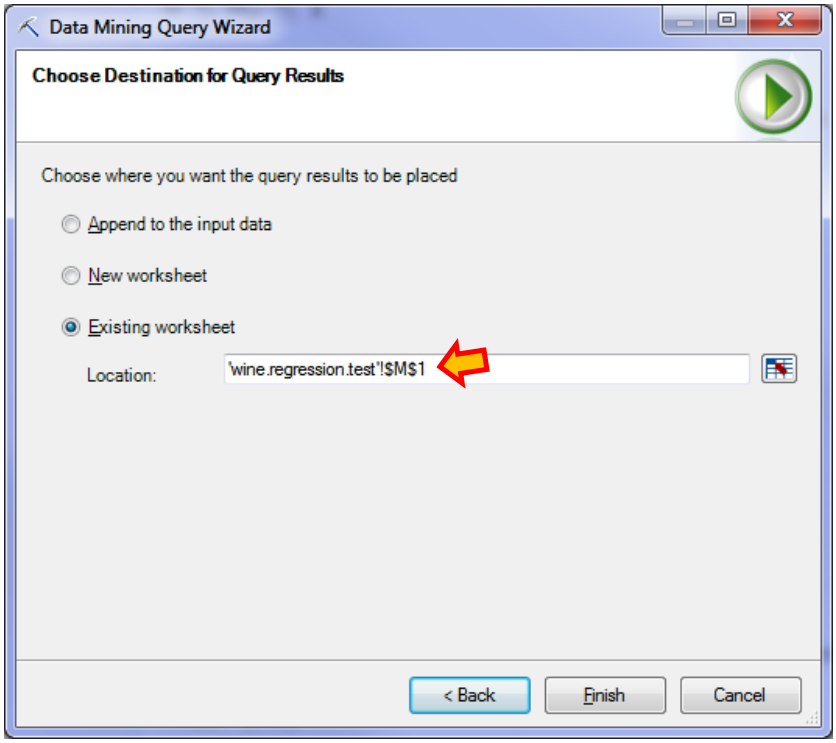
Nous sélectionnons ensuite l'échantillon test.



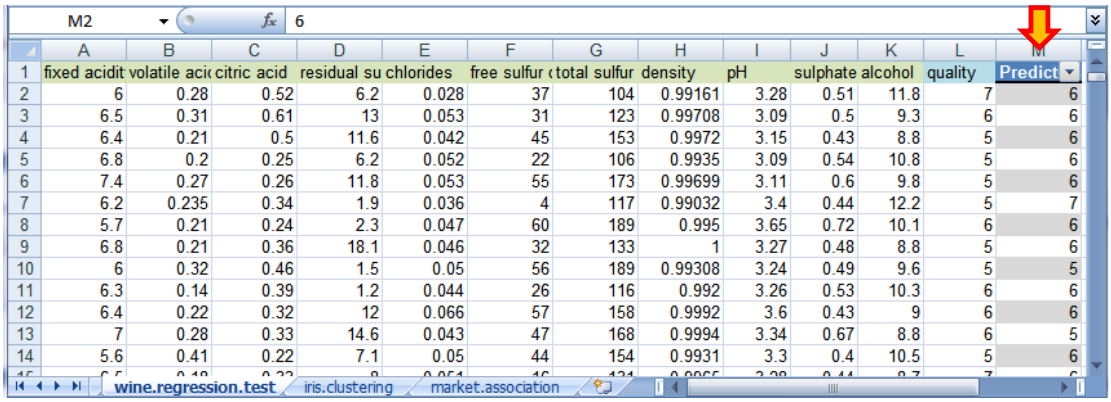
SSAS effectue les correspondances entre les variables et nous demande de préciser le type de valeurs à produire. Nous choisissons la prédiction ponctuelle. Nous remarquerons que nous pouvons également calculer les écarts-type de prédiction et, manuellement sous Excel, de construire après coup les fourchettes (intervalles) de prédiction.



Enfin, nous précisons l'emplacement de la sortie.



Valeurs prédites. Dans Excel, nous constatons que les valeurs prédites sont entières.



M2 : 6

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	fixed acidit	volatile aci	citric acid	residual su	chlorides	free sulfur	total sulfur	density	pH	sulphate	alcohol	quality	Predict
2	6	0.28	0.52	6.2	0.028	37	104	0.99161	3.28	0.51	11.8	7	6
3	6.5	0.31	0.61	13	0.053	31	123	0.99708	3.09	0.5	9.3	6	6
4	6.4	0.21	0.5	11.6	0.042	45	153	0.9972	3.15	0.43	8.8	5	6
5	6.8	0.2	0.25	6.2	0.052	22	106	0.9935	3.09	0.54	10.8	5	6
6	7.4	0.27	0.26	11.8	0.053	55	173	0.99699	3.11	0.6	9.8	5	6
7	6.2	0.235	0.34	1.9	0.036	4	117	0.99032	3.4	0.44	12.2	5	7
8	5.7	0.21	0.24	2.3	0.047	60	189	0.995	3.65	0.72	10.1	6	6
9	6.8	0.21	0.36	18.1	0.046	32	133	1	3.27	0.48	8.8	5	6
10	6	0.32	0.46	1.5	0.05	56	189	0.99308	3.24	0.49	9.6	5	5
11	6.3	0.14	0.39	1.2	0.044	26	116	0.992	3.26	0.53	10.3	6	6
12	6.4	0.22	0.32	12	0.066	57	158	0.9992	3.6	0.43	9	6	6
13	7	0.28	0.33	14.6	0.043	47	168	0.9994	3.34	0.67	8.8	6	5
14	5.6	0.41	0.22	7.1	0.05	44	154	0.9931	3.3	0.4	10.5	5	6

wine.regression.test | iris.clustering | market.association

Pourtant, si nous appliquons les coefficients de la régression sur le 1er individu, en utilisant les seules les variables sélectionnées, nous avons :

112.431 + 0.253 x 11.8 – 111.533 x 0.99161 + 0.557 * 3.28 + 0.071 x 6.2 – 2.143 x 0.28 = 6.486

La prédiction ponctuelle n'est pas entière. Il semble que SSAS, ayant détecté la nature de la variable cible, arrondisse automatiquement à l'entier le plus proche. Pourquoi pas après tout ? Mais il faudrait que la transformation soit annoncée clairement à l'utilisateur pour éviter les incompréhensions.

Mesurer la performance à l'aide du pseudo-R². Nous utilisons un pseudo-R² pour mesurer la qualité de la prédiction sur l'échantillon test. Il consiste à confronter les prédictions du modèle (PREDICTQUALITY) avec celles du modèle trivial, constitué uniquement d'une constante (sans intervention des variables prédictives).

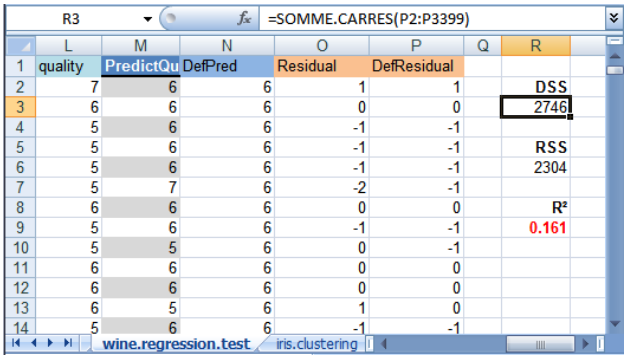
- La prédiction du modèle trivial est égale à la moyenne de l'endogène calculée sur l'échantillon d'apprentissage (la moyenne est la meilleure prédiction possible au sens des moindres carrés). Elle est égale à 5.872 sur « wine.regression.train ».
- En utilisant le même artifice que SSASS, la prédiction par défaut est donc égale à DEFPRED = 6 (arrondi à l'entier le plus proche)[25].
- Deux colonnes résidus sont construites à partir des prédictions : RESIDUAL et DEFRESIDUAL correspondent respectivement aux résidus de prédiction de notre modèle et du modèle trivial.
- Nous calculons le pseudo R², il confronte les deux prédictions à l'aide de la formule

$$R^2 = 1 - \frac{RSS}{DSS}$$

[25] « 6 » est aussi la valeur de « quality » qui apparaît le plus souvent (le mode) dans l'échantillon d'apprentissage. L'utiliser comme prédiction triviale peut aussi se justifier de cette manière.

- Où RSS (resp. DSS) correspond à la somme des carrés des résidus du modèle à évaluer (resp. du modèle trivial).
- Lorsque R^2 est proche de 0, notre modèle ne fait pas mieux que le modèle trivial ; lorsque R^2 est proche de 1, nous avons un modèle parfait puisque qu'il prédit sans erreurs.

Dans la copie d'écran ci-dessous sont retracées les opérations réalisées sous Excel.



	L	M	N	O	P	Q	R
1	quality	PredictQu	DefPred	Residual	DefResidual		
2	7	6	6	1	1		DSS
3	6	6	6	0	0		2746
4	5	6	6	-1	-1		
5	5	6	6	-1	-1		RSS
6	5	6	6	-1	-1		2304
7	5	7	6	-2	-1		
8	6	6	6	0	0		R²
9	5	6	6	-1	-1		0.161
10	5	5	6	0	-1		
11	6	6	6	0	0		
12	6	6	6	0	0		
13	6	5	6	1	0		
14	5	6	6	-1	-1		

Dans notre cas,

$$R^2 = 1 - \frac{2304}{2746} = 0.161$$

Le modèle n'est pas très performant. Ce n'est pas très étonnant. Arrondir à l'entier le plus proche occasionne une perte d'information.

6 Classification automatique (clustering)

La classification automatique consiste à regrouper les observations (ça peut être les variables aussi) de manière à ce que les observations qui appartiennent à un même groupe (cluster) présentent des caractéristiques proches. Il existe pléthores d'approches. SSAS s'appuie sur deux algorithmes itératifs : la méthode K-Means, où chaque individu appartient à un groupe et un seul à l'issue du traitement ; la méthode EM (Expectation Maximization) où un individu peut être rattaché à plusieurs groupes à différents degrés de probabilité[26].

6.1 Données à traiter

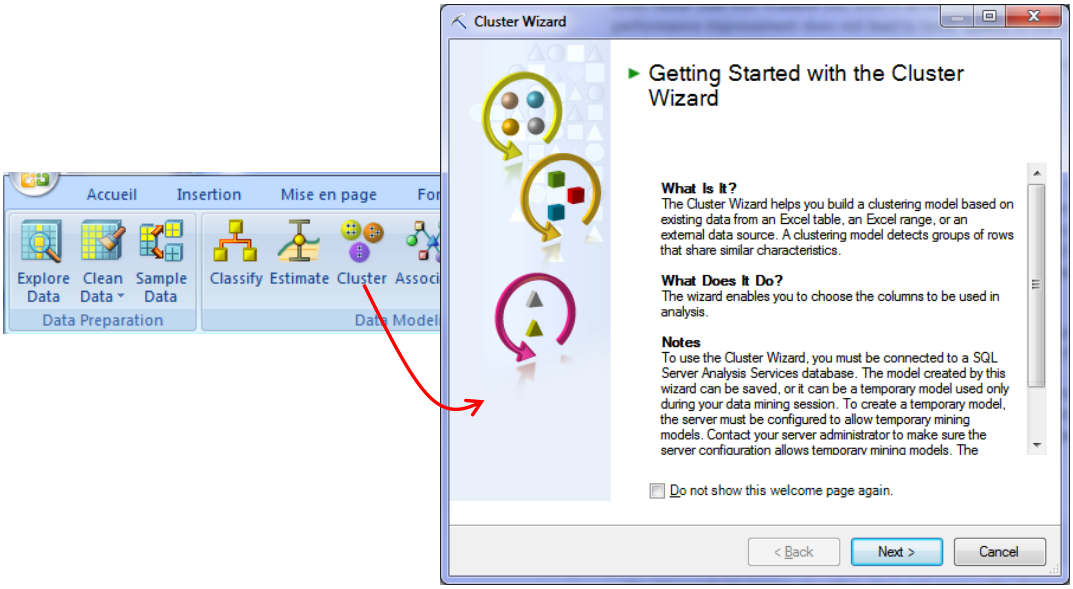
Nous traitons le très populaire fichier IRIS[27], incontournable en reconnaissance de formes. Nous l'utilisons dans un contexte non supervisé dans cette section. Nous essayons de les regrouper en 3 clusters à partir de leurs caractéristiques physiques (longueur et largeur des pétales et des sépales). Nous vérifierons a posteriori, on parle de validation externe, si les groupes correspondent aux types prédéfinis, décrits par la variable « class » = {setosa, versicolour, virginica}, auxquels sont rattachés les objets.

[26] http://technet.microsoft.com/en-us/library/cc280445.aspx

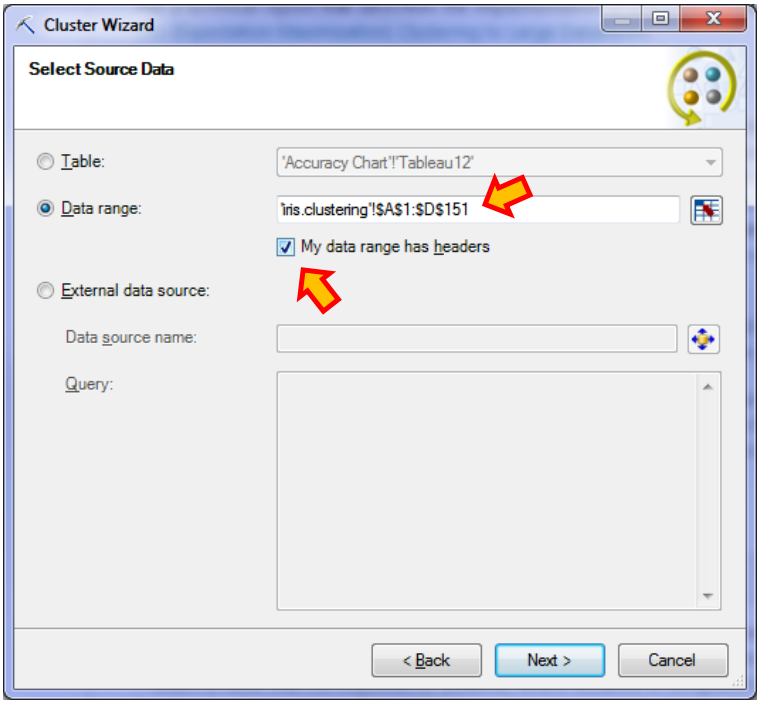
[27] https://archive.ics.uci.edu/ml/datasets/Iris

6.2 Construction des groupes

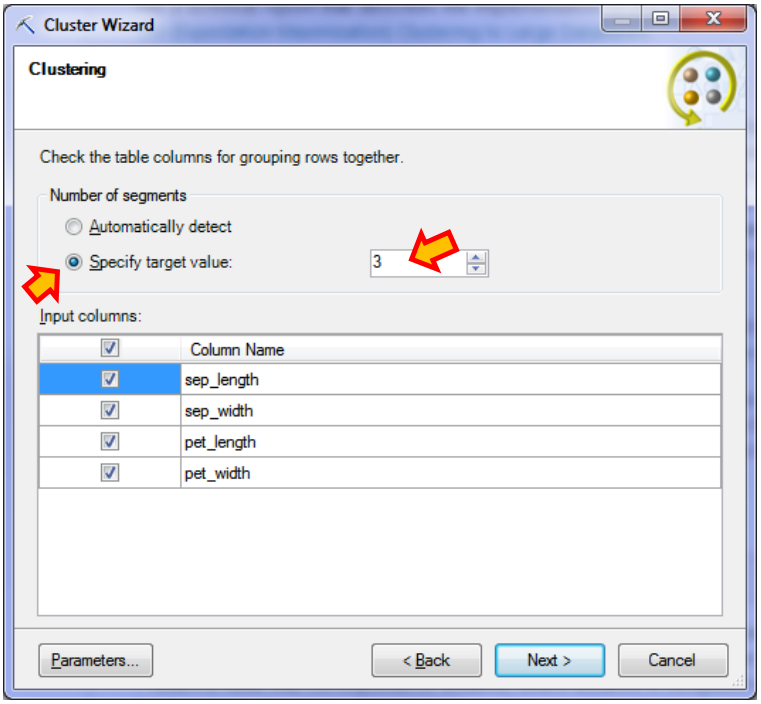
Lancement de l'analyse. Nous cliquons sur l'icône CLUSTER (section Data Modeling) du ruban « Data Mining » pour lancer le processus. Un nouveau wizard apparaît.



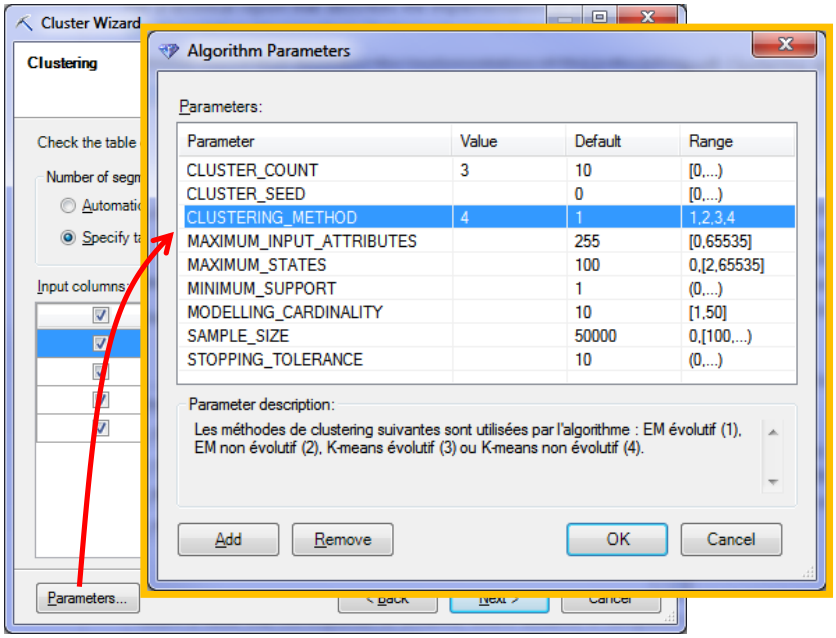
Plage de données. Nous spécifions la plage de données, les 4 premières colonnes de la feuille « **iris.clustering** »



Choix et paramétrage de la méthode. Nous fixons à 3 le nombre de groupes à construire. On notera que SSAS dispose d'une heuristique de détection automatique du nombre de groupe. Cela peut être très précieux dans une phase de première appréhension des données.



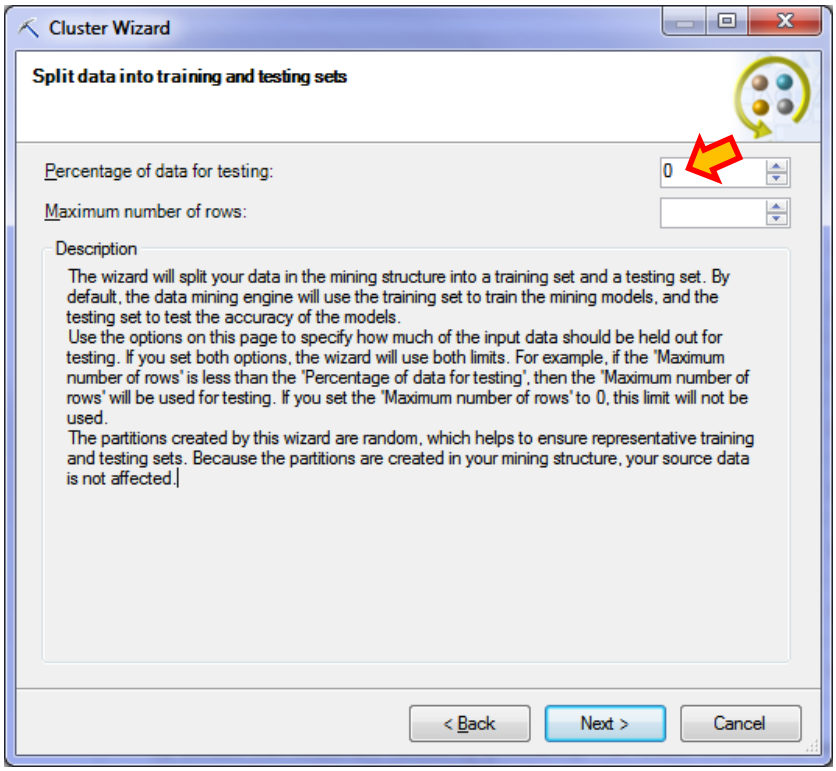
En cliquant sur PARAMETERS, nous pouvons paramétrer l'algorithme.



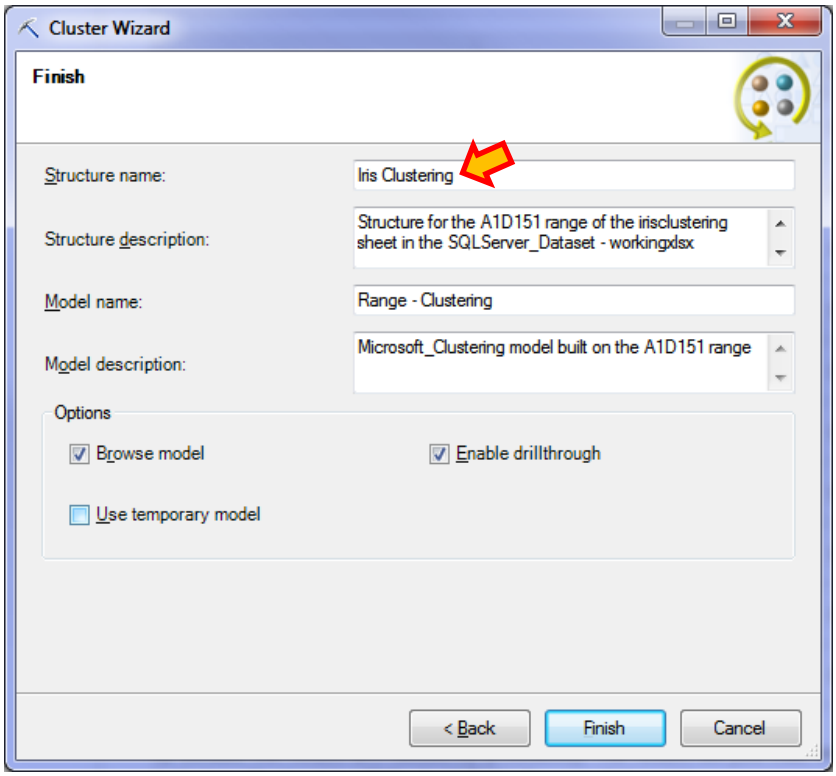
Nous choisissons un K-MEANS non évolutif (« Non scalable K-Means », la traduction de « scalable » en « évolutif » peut prêter à confusion…) qui correspond à l'approche usuelle c.-à-d. l'algorithme s'appuie directement sur l'ensemble des observations disponibles. L'outil sait gérer à la fois les variables quantitatives et qualitatives[28].

[28] « Microsoft Clustering Algorithm Technical Reference », http://technet.microsoft.com/en-us/library/cc280445.aspx

Echantillon d'apprentissage et de test. Très curieusement, SSAS nous propose de réserver un échantillon test pour évaluer le modèle. Pourquoi pas, mais ce n'est pas vraiment la pratique usuelle en classification automatique. Nous fixons la proportion à 0.

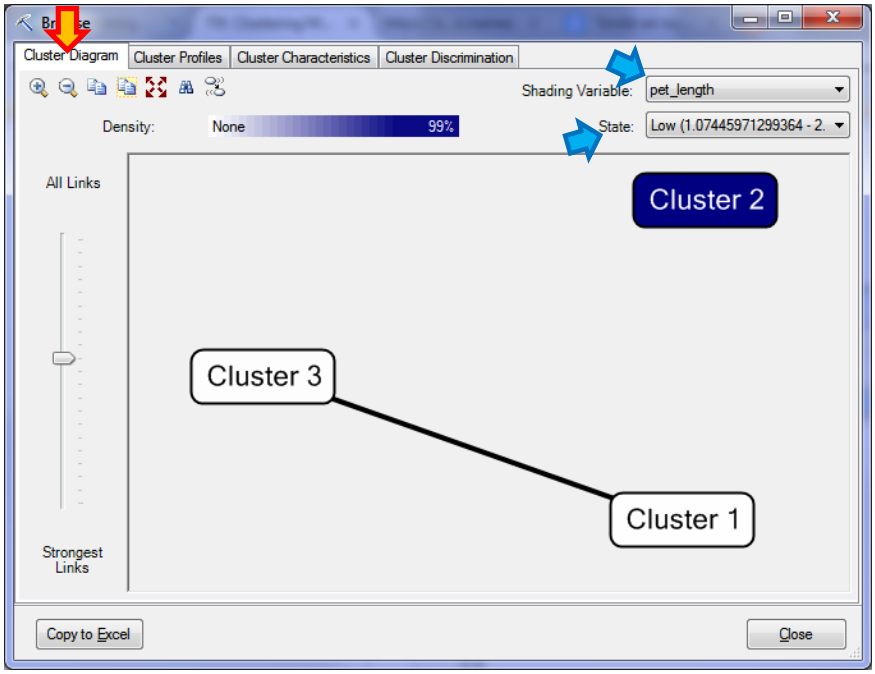


Stockage du modèle. Il reste enfin à stocker le modèle dans la base SSAS. Nous lui attribuons le nom « Iris Clustering ».

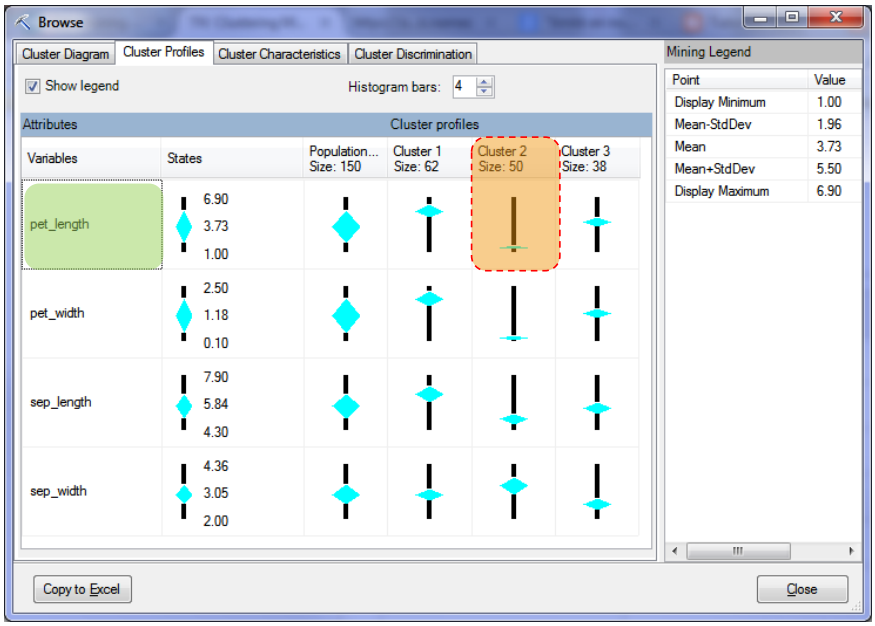


Visualisation des résultats. La fenêtre de visualisation apparaît lorsque nous cliquons sur le bouton FINISH. Elle comporte plusieurs onglets.

Cluster diagram indique la proximité entre les groupes d'une part, l'association entre les variables et les groupes d'autre part. Par exemple, dans la copie d'écran ci-dessous, nous constatons que « Cluster 1 » et « Cluster 3 » sont assez proches, et que tous deux sont éloignés de « Cluster 2 ». Une valeur faible (« low », comprise entre 1.07 et 2.84) de PET_LENGTH désigne le « Cluster 2 » en fond bleu foncé c.-à-d. on observe une forte densité d'individus appartenant au « Cluster 2 » dans cet intervalle.

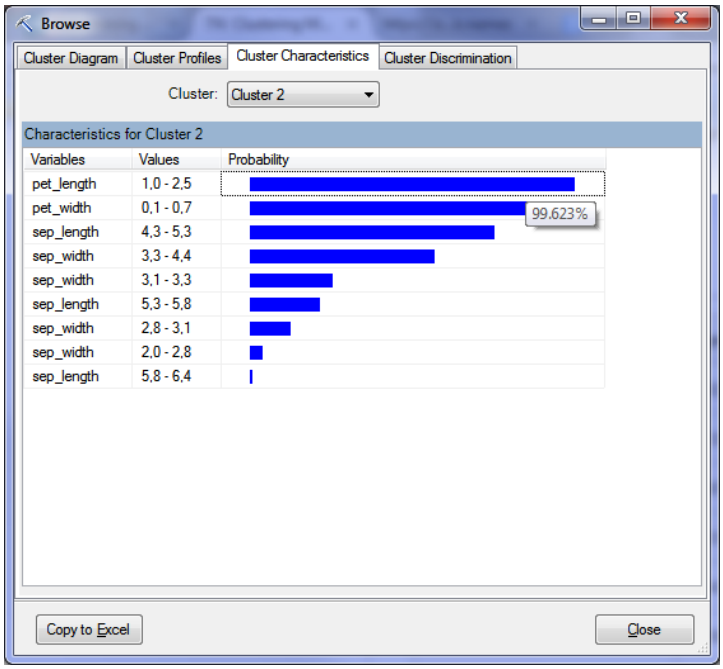


Cluster profile calcule les statistiques descriptives conditionnelles des variables.

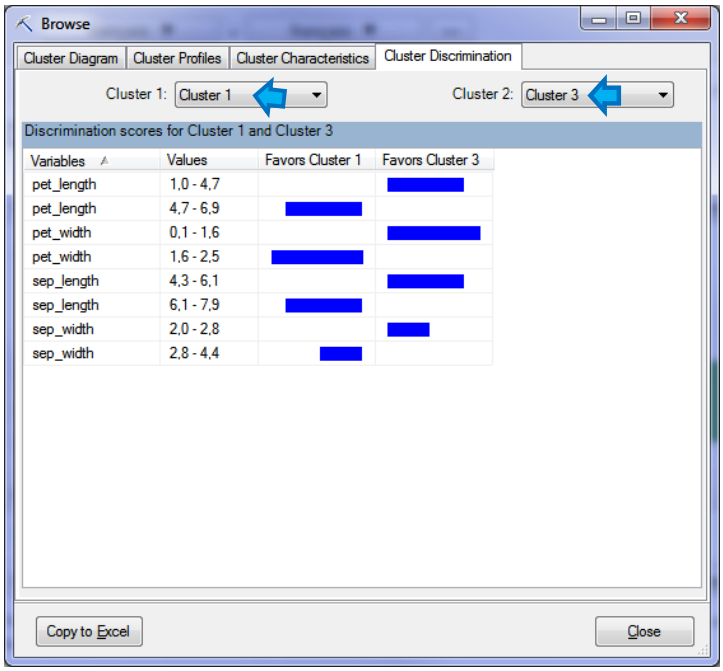


Tout comme dans l'onglet « Cluster Diagram », nous constatons qu'une valeur faible de PET_LENGTH est associée au 2nd cluster.

Cluster Characteristics propose les informations sous la forme de probabilités. Dans la copie d'écran ci-dessous, nous observons qu'une valeur de PET_LENGTH comprise entre 1.0 et 2.5 a 99.623% de chances d'apparaître dans le « Cluster 2 ».



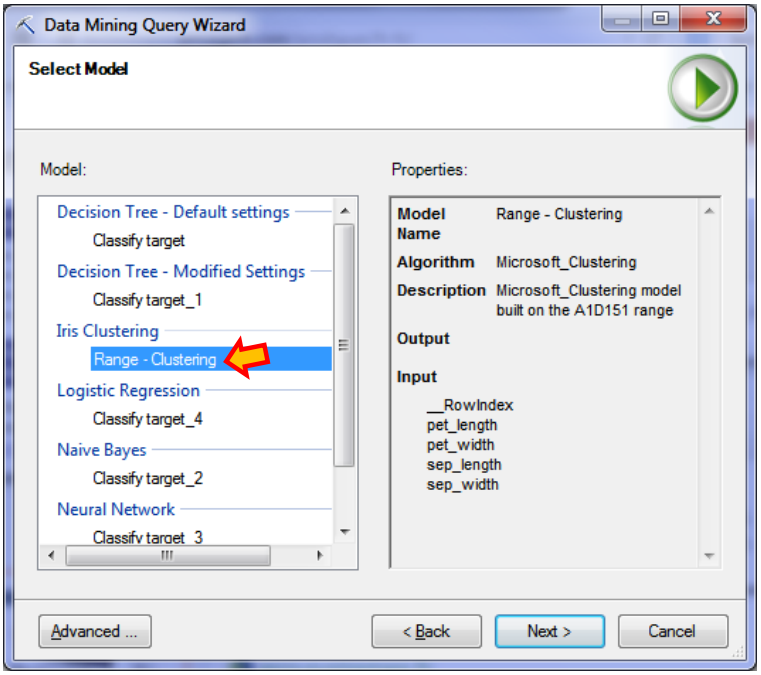
Cluster discrimination identifie les caractéristiques des données qui différencient les groupes.



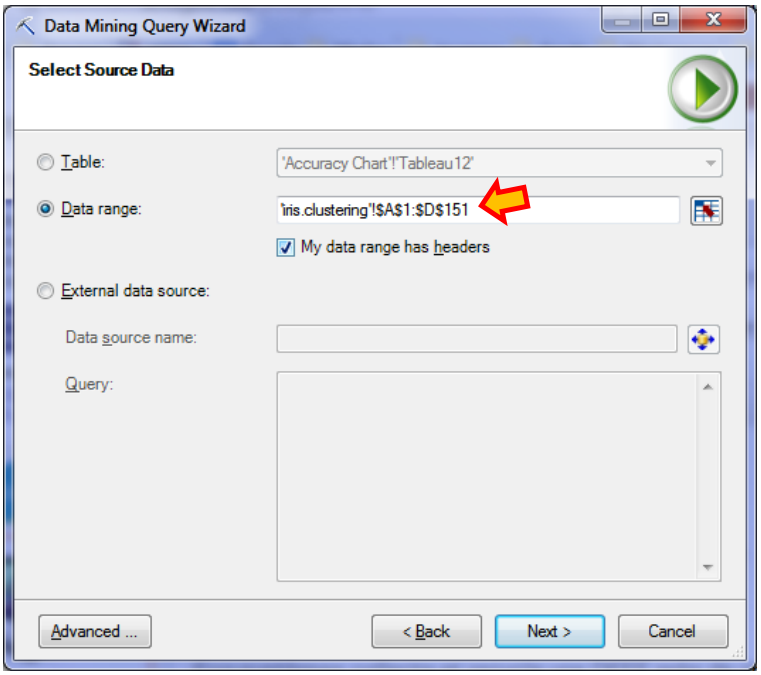
Dans notre exemple, PET_LENGTH comprise entre 1.0 et 4.7 désigne le « Cluster 3 », en opposition au « Cluster 1 » qui est plutôt associé à la plage de valeurs 4.7 et 6.9.

6.3 Déploiement

L'outil de déploiement permet d'associer chaque individu à son groupe d'appartenance. Nous cliquons sur l'icône QUERY et nous sélectionnons le modèle « Iris Clustering ».

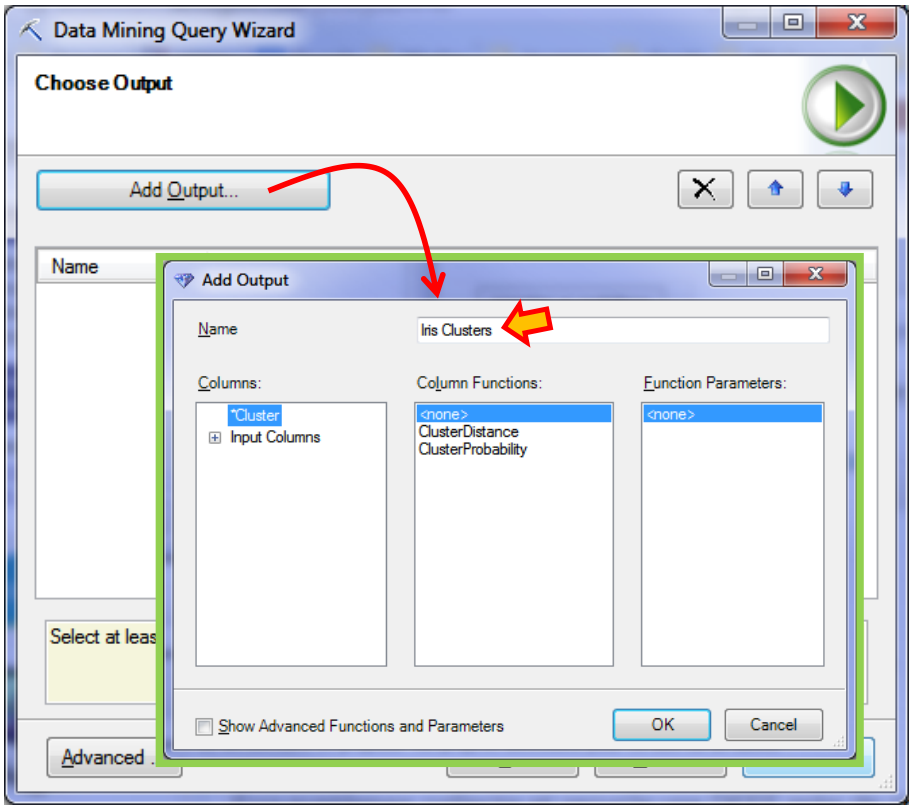


Nous précisons ensuite la plage de données de déploiement. Cela peut être n'importe quelle base de données pourvu que les variables de l'analyse soient présentes.

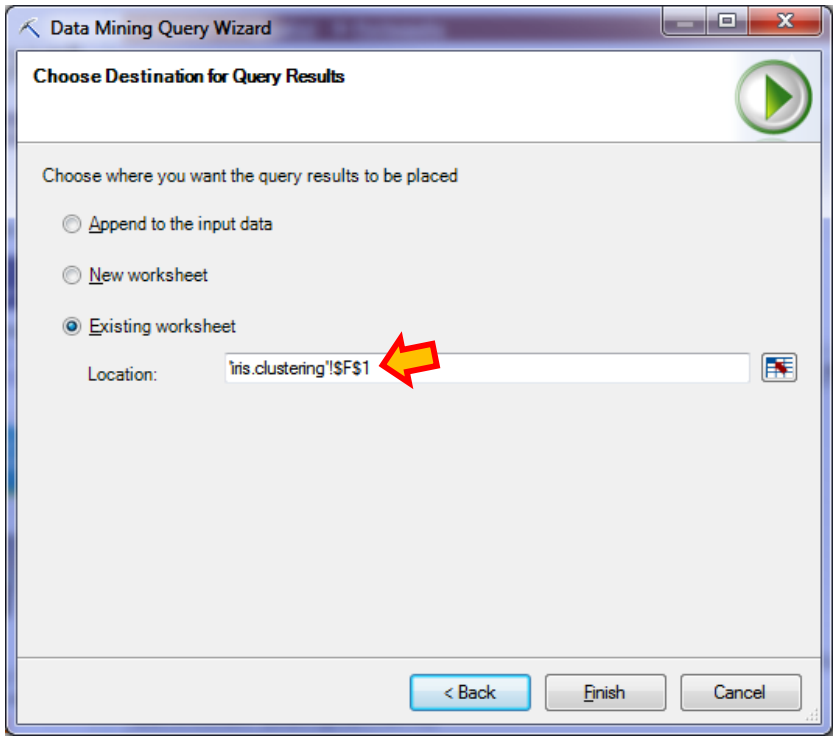


SSAS vérifie les correspondances puis nous demande de préciser le type d'information à produire. Le paramétrage suivant permet d'obtenir les groupes d'appartenance. D'autres

sorties sont possibles (distance aux groupes, probabilité d'appartenance). Nous appelons la nouvelle colonne « Iris Clusters ».



Enfin, nous précisons la plage de sortie.



La nouvelle colonne de données est disponible dans la feuille Excel. Voici les résultats pour les 10 premières observations.

	A	B	C	D	E	
1	sep_length	sep_width	pet_length	pet_width	class	Iris Clusters
2	5.1	3.5	1.4	0.2	Iris-setosa	Cluster 2
3	4.9	3.0	1.4	0.2	Iris-setosa	Cluster 2
4	4.7	3.2	1.3	0.2	Iris-setosa	Cluster 2
5	4.6	3.1	1.5	0.2	Iris-setosa	Cluster 2
6	5.0	3.6	1.4	0.2	Iris-setosa	Cluster 2
7	5.4	3.9	1.7	0.4	Iris-setosa	Cluster 2
8	4.6	3.4	1.4	0.3	Iris-setosa	Cluster 2
9	5.0	3.4	1.5	0.2	Iris-setosa	Cluster 2
10	4.4	2.9	1.4	0.2	Iris-setosa	Cluster 2
11	4.9	3.1	1.5	0.1	Iris-setosa	Cluster 2

6.4 Validation externe

Nous avons la chance de disposer avec la variable « class » des « vrais » groupes d'appartenance des observations. Bien sûr, ce type de configuration n'existe pas dans les études réelles. La classification sert à identifier les groupes dont on n'a pas idée a priori. La technique ne sert à rien s'il existe déjà par ailleurs des groupes prédéfinis. Cette démarche de confrontation entre classes existantes et prédites, dite de validation externe, ne se justifie que dans le cadre des publications scientifiques. Face à la difficulté de définir un critère objectif de qualité de partitionnement, s'appuyer sur des informations externes est une piste comme une autre pour valider et comparer les performances des algorithmes de clustering.

A l'aide de l'outil « tableau croisé dynamique », nous créons le tableau de contingence entre « class » (groupe prédéfini) et « Iris Clustering » (groupe attribué par l'algorithme).

Nombre de class	Étiquettes			
Étiquettes de lignes	Cluster 1	Cluster 2	Cluster 3	Total général
Iris-setosa		50		50
Iris-versicolor	14		36	50
Iris-virginica	48		2	50
Total général	**62**	**50**	**38**	**150**

Il y a équivalence entre « Iris-Setosa » et « Cluster 2 ». « Iris-Virginica » est rassemblé dans « Cluster 1 », qui comporte également 14 exemplaires de « Iris-Versicolor ». Le « Cluster 3 » est composé essentiellement de « Iris-Versicolor ». Les groupes calculés sont relativement cohérents par rapport aux groupes prédéfinis.

7 Règles d'association

La recherche des règles d'association vise à mettre en lumière les relations entre deux ou plusieurs variables d'une base de données[29]. Elle est utilisée notamment dans l'analyse des tickets de caisse des supermarchés où l'on cherche à mettre en évidence les cooccurrences des produits dans les caddies. A la sortie, elle produit des règles de type :

Si achats produit_1 **et** produit_2 **Alors** achat produit_3

[29] http://fr.wikipedia.org/wiki/Règle_d'association

La partie SI est appelée antécédent, la partie ALORS, conséquent. La qualité d'une règle est communément qualifiée par le support, qui indique la proportion de transactions (caddies) couvertes ; et la confiance, qui indique sa précision. D'autres mesures existent, les scientifiques sont particulièrement prolixes dans le domaine.

7.1 Données à traiter

Les données de la feuille « **market.association** » décrivent le contenu de 1361 tickets de caisse. Elles sont au format transactionnel[30] c.-à-d. nous disposons d'une liste de produits par caddie : ID est le numéro du caddie, PRODUCT correspond aux produits présents. Voici les premières lignes du fichier.

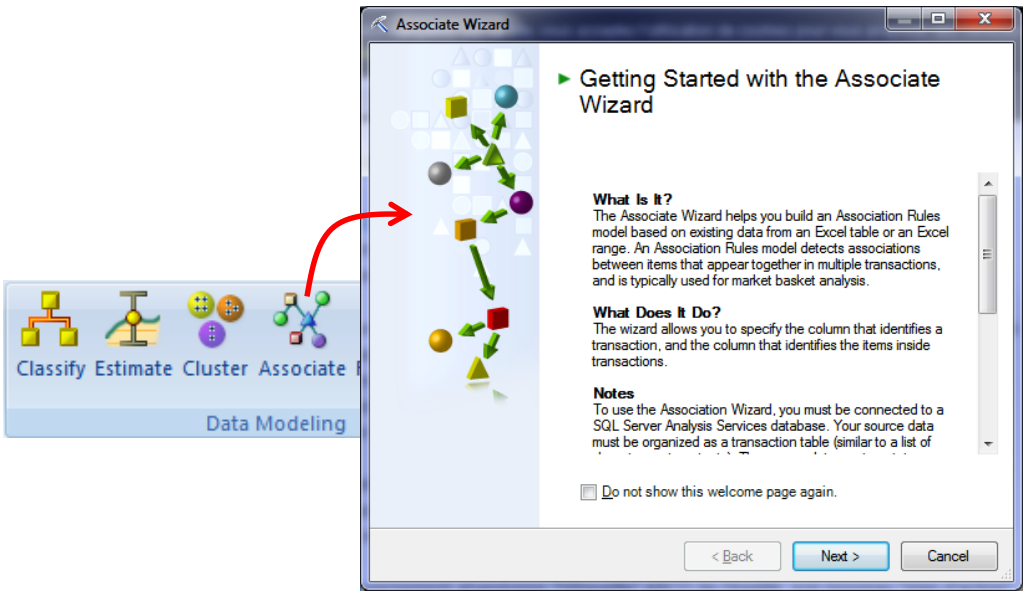
	A	B	C
1	ID	Product	
2	1	Peaches	
3	2	Vegetable_Oil	
4	2	Frozen_Corn	
5	3	Plums	
6	4	Pancake_Mix	
7	5	Cheese	
8	6	Cauliflower	
9	7	2pct_Milk	
10	8	98pct_Fat_Free_Hamburger	
11	8	Potato_Chips	
12	8	Sesame_Oil	
13	8	Ice_Cream_Sandwich	

market.association

Par exemple, nous observons 2 produits dans le caddie n°2 : Vegetable_Oil, Frozen_Com.

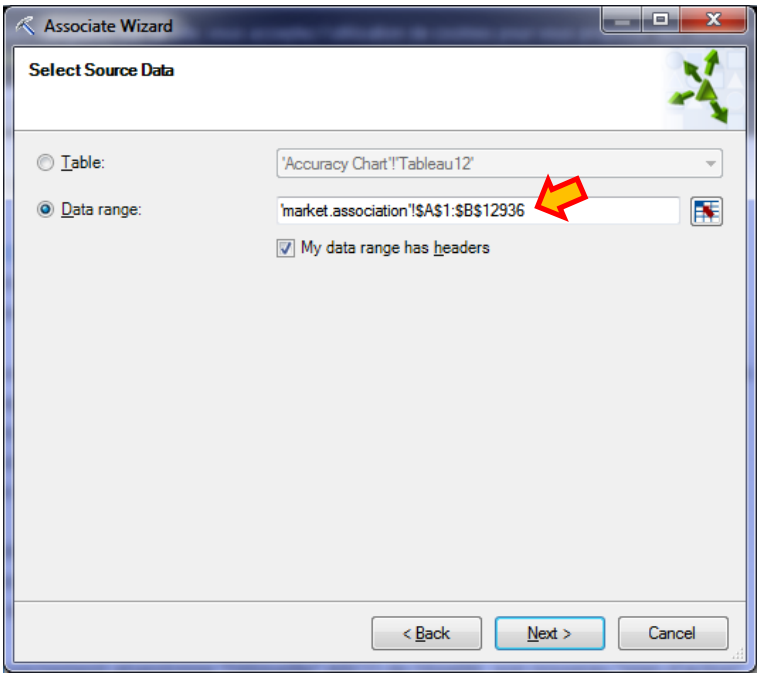
7.2 Construction des règles

Nous actionnons le bouton ASSOCIATE (section Data Modeling) du ruban Data Mining.

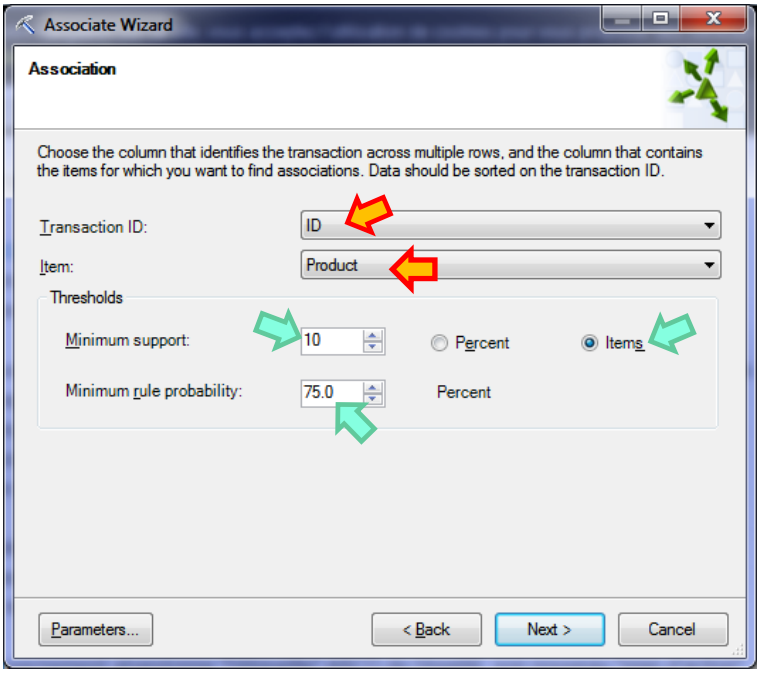


[30] http://tutoriels-data-mining.blogspot.fr/2010/12/regles-dassociation-donnees.html

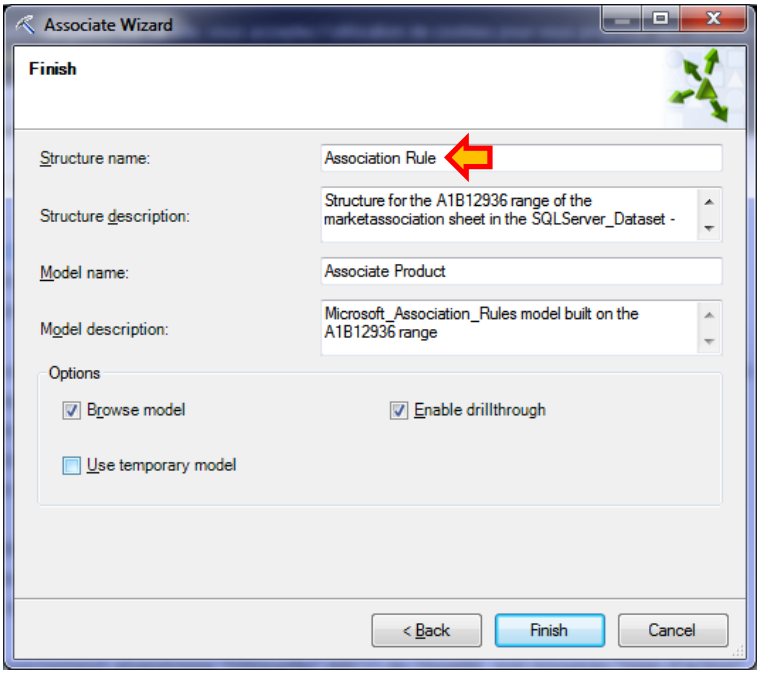
Nous enchaînons alors les étapes. Tout d'abord, la sélection des données.



Puis le paramétrage de la méthode : ID est bien la colonne des identifiants de transactions, PRODUCT correspond aux items, le support minimum est de 10 items (en valeur absolue), et la confiance minimum est fixée à 75%. En cliquant sur le bouton PARAMETERS nous accédons à d'autres paramètres pour les analyses plus sophistiquées.

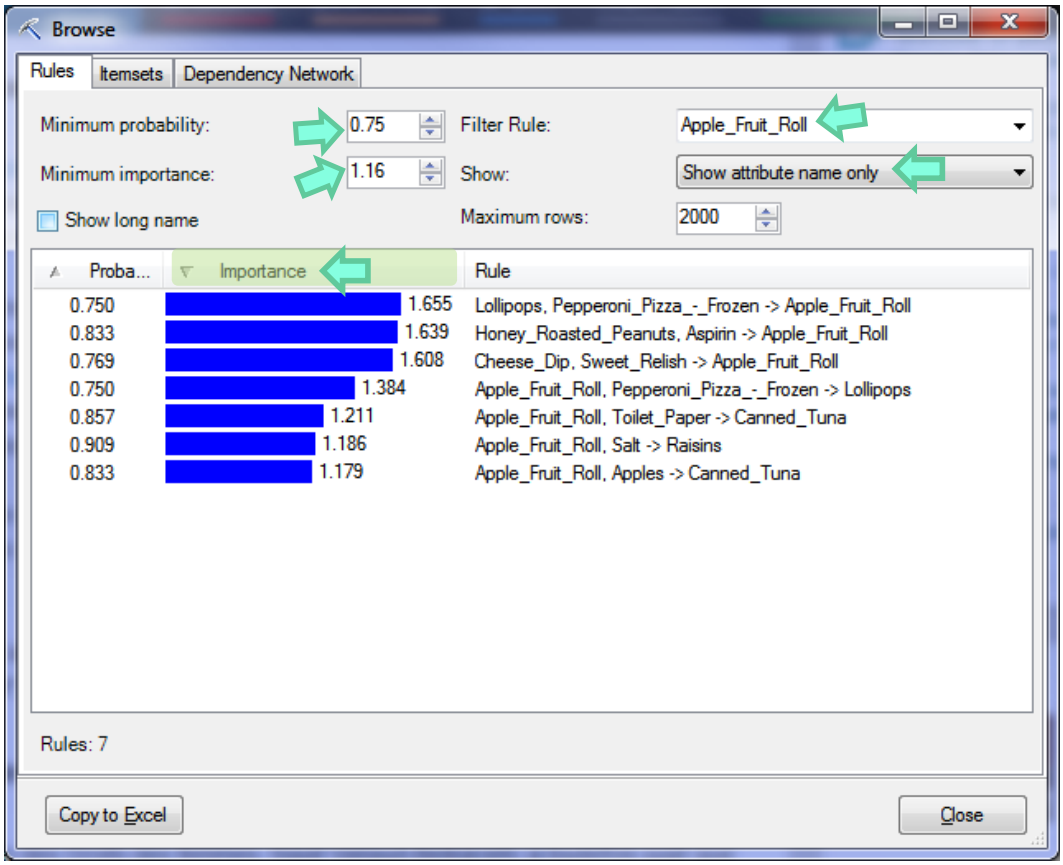


Nous nommons la structure « Association Rule » et nous cliquons sur FINISH.



7.3 Visualisation des règles

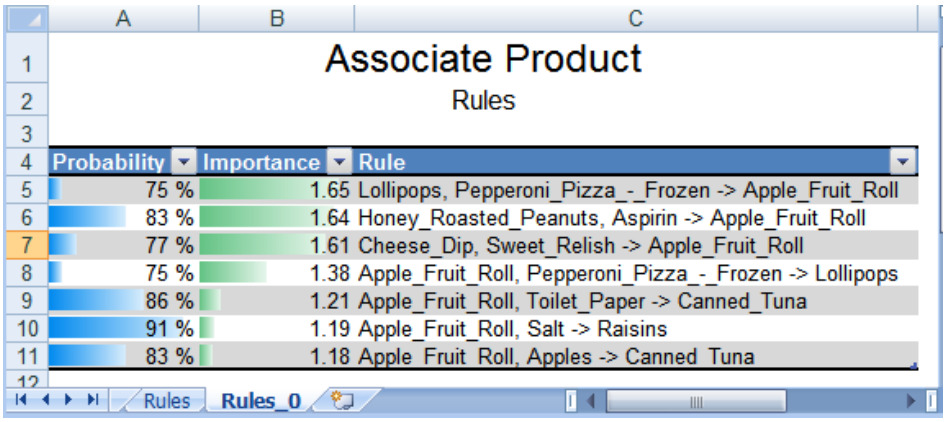
Nous disposons de la liste des règles dans l'onglet « **Rules** ».



Nous pouvons les filtrer a posteriori avec plusieurs critères. Nous avons fixé : Minimum Probability = 0.75, Minimum Importance = 1.16, et Filter Rule = Apple_Fruit_Roll c.-à-d. seules

les règles contenant le produit « Apple_Fruit_Roll » nous intéressent. 7 règles répondent à ces critères. Nous les avons triées selon leur « Importance »[31].

Les mêmes règles peuvent être visualisées dans le tableur Excel en cliquant sur le bouton COPY TO EXCEL. Nous avons un tableau que nous pouvons filtrer et manipuler à souhait[32].

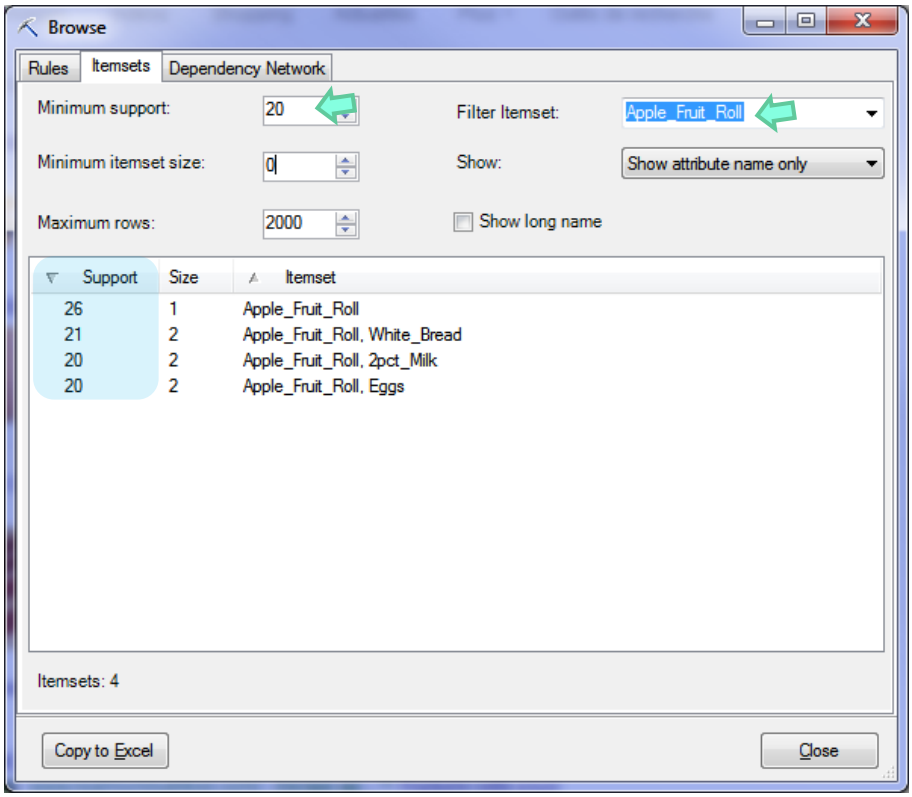


Associate Product
Rules

Probability	Importance	Rule
75 %	1.65	Lollipops, Pepperoni_Pizza_-_Frozen -> Apple_Fruit_Roll
83 %	1.64	Honey_Roasted_Peanuts, Aspirin -> Apple_Fruit_Roll
77 %	1.61	Cheese_Dip, Sweet_Relish -> Apple_Fruit_Roll
75 %	1.38	Apple_Fruit_Roll, Pepperoni_Pizza_-_Frozen -> Lollipops
86 %	1.21	Apple_Fruit_Roll, Toilet_Paper -> Canned_Tuna
91 %	1.19	Apple_Fruit_Roll, Salt -> Raisins
83 %	1.18	Apple_Fruit_Roll, Apples -> Canned_Tuna

Rules | Rules_0

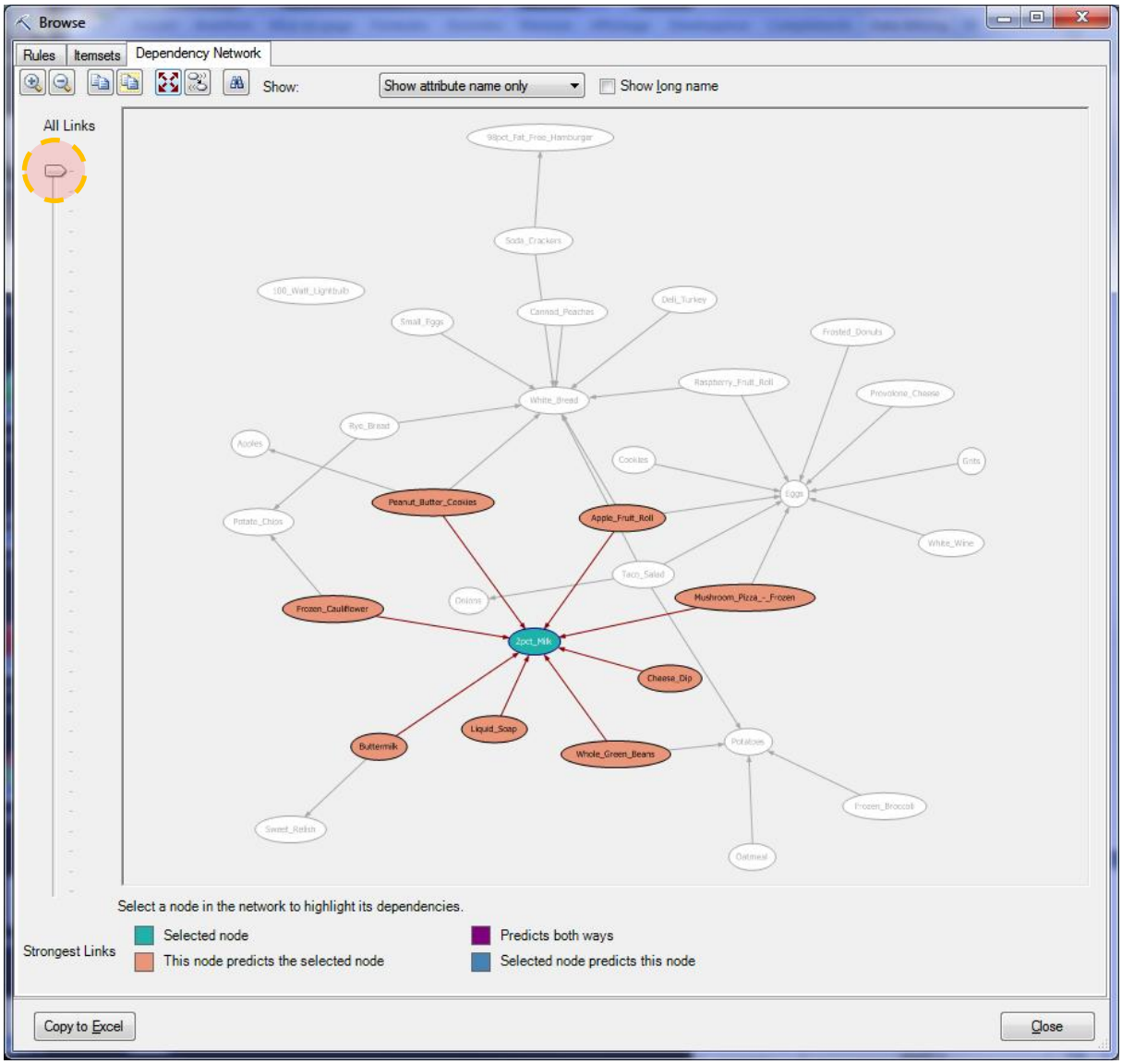
La liste des cooccurrences entre les produits et leurs supports associés sont accessibles dans l'onglet « **Itemsets** ». Diverses options de filtrage sont disponibles. Par exemple, nous visualisons ci-dessous les itemsets concernés par « Apple_Fruit_Roll ». Le produit apparaît dans 26 caddies, il apparaît en compagnie de « White_Bread » dans 21 caddies, etc.



[31] Voir « Microsft Association Algorithm Technical Reference » pour la définition des indicateurs utilisés par SSAS.

[32] Les atouts d'Excel en matière de post-manipulation des règles (filtrage, tri) sont décrits dans « Associations dans la distribution SIPINA », avril 2013.

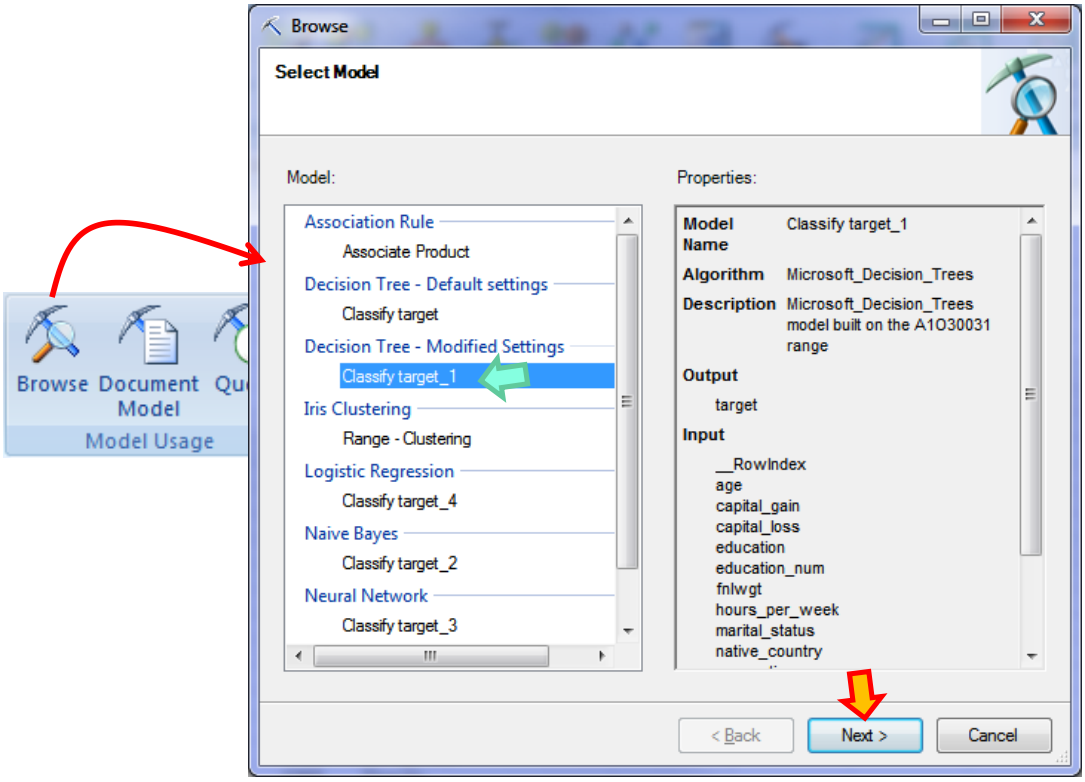
Enfin, le graphe des dépendances est visible dans l'onglet « **Dependency Network** ».



Dixit la documentation, les liaisons entre les items indiquent les associations. La direction des flèches retrace traduit le sens des relations en accord avec les règles extraites. Le curseur sur la gauche permet de filtrer les associations selon la confiance.

8 Visualisation des modèles déjà construits

A l'issue de chaque traitement, les modèles sont stockés dans la base que nous avions créée initialement (section 2.2). Nous pouvons y accéder à tout moment en cliquant sur l'icône BROWSE (section Model Usage) du ruban « Data Mining ». Une fenêtre listant les modèles disponibles apparaît. Nous sélectionnons celui qui nous intéresse et nous cliquons sur NEXT.



Le modèle apparaît dans la fenêtre de visualisation qui lui est propre.

9 Traitement des gros volumes sous Excel

9.1 Préambule

Tout d'abord, signalons que nous pouvons directement travailler sur un ensemble de données chargé dans une base SQL Server avec SSAS. Lors du lancement des techniques statistiques à partir du ruban « data mining », nous les branchons directement sur des bases externes. Dans ce cas, les limitations inhérentes au tableur n'entrent pas en ligne de compte.

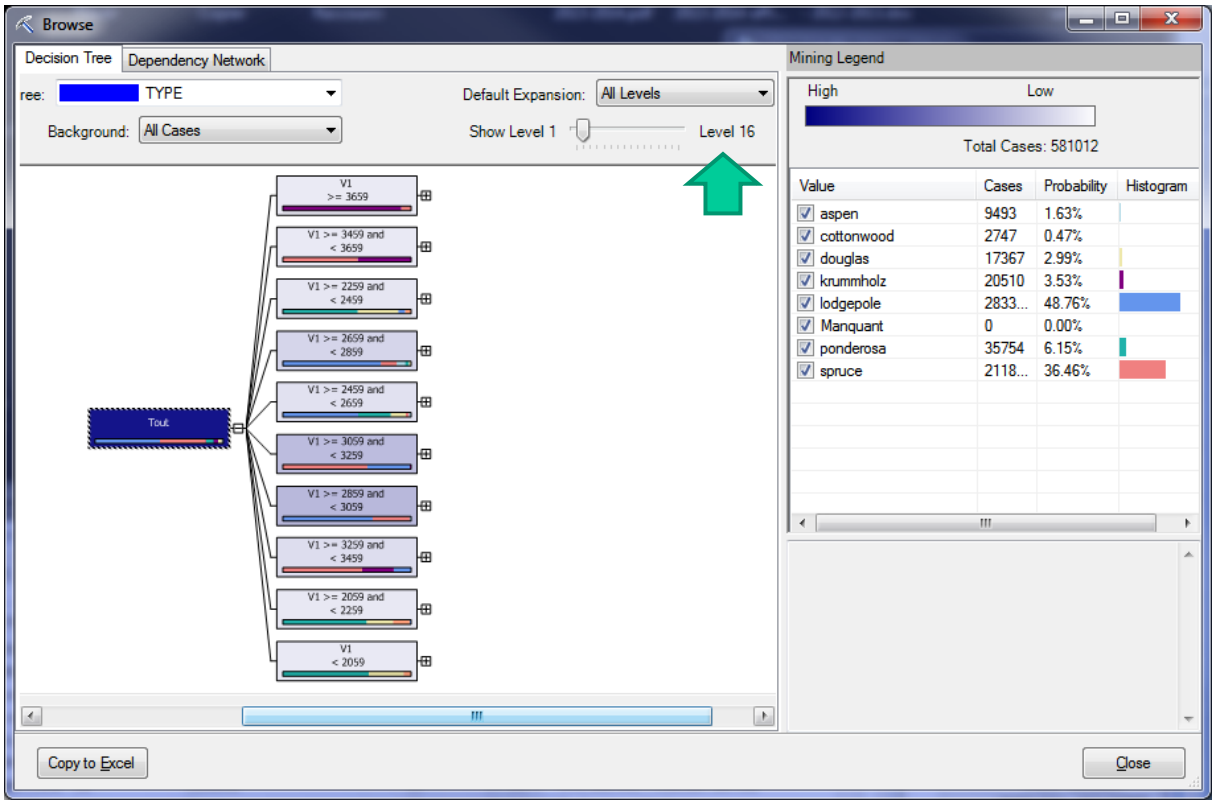
L'affaire est différente si nous souhaitons analyser des données situées dans un classeur au format XLSX. Parler de grandes bases peut paraître une hérésie en nous référant à Excel. Mais d'ailleurs, est-il vraiment indispensable de traiter de très gros volumes ? Je ne suis pas sûr qu'il soit utile et nécessaire d'extraire de la connaissance à partir de millions d'observations, sauf à vouloir déceler les phénomènes de niche. Dans la majorité des cas, travailler à partir d'échantillons permet de produire des modèles efficaces. On le sait peu, certains logiciels qui font référence intègrent l'éventualité de travailler automatiquement sur des échantillons. Dans ce tutoriel, nous avions noté par exemple que le réseau de neurones prévoyait de ne traiter qu'une sous-partie des observations (section 3.5). De fait, Excel a tout à fait sa place. Nous disposons d'une capacité de 1.048.575 observations (la première ligne correspondant aux noms des colonnes) et 16384 variables. Ca nous laisse une certaine marge de manœuvre.

9.2 Traitement du fichier « Covertype »

Le fichier COVERTYPE[33] comporte 581012 observations, une variable cible à 7 modalités (TYPE) et 54 variables prédictives, dont 10 sont quantitatives. Le fichier au format TXT a été importé dans Excel. La construction d'un arbre de décision est demandée avec les paramètres par défaut. Ma machine est Quad Core Q9664 avec 8 Go de RAM. Les différentes étapes avec l'occupation mémoire et le temps d'exécution sont consignés dans le tableau ci-dessous.

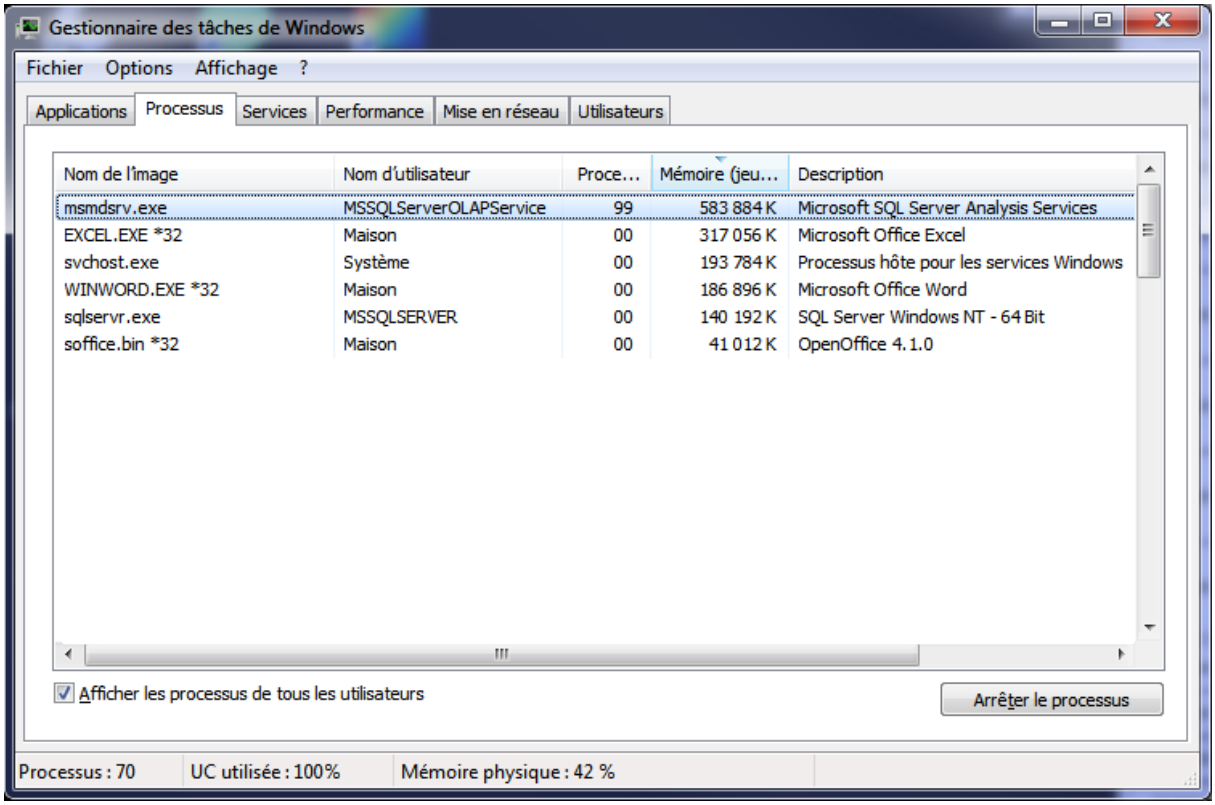
Opération	Occupation mémoire Excel (Mo)	Occupation mémoire SQL Server (Mo)	Durée des traitements (sec.)
0. Démarrage	19.4	136.9	-
1. Importation	288.2	136.9	26
2. Sélection des données	309.6	136.9	85
3. Construction du modèle	309.6	707.1 (SQL Server + SSAS)	338

Modèle. L'arbre est un surdimensionné (16 niveaux). Nous ne montrons que les 2 premiers niveaux ici. Manifestement, le paramétrage par défaut n'est pas adapté pour un tel volume.



[33] https://archive.ics.uci.edu/ml/datasets/Covertype

Phase de calcul. Le gestionnaire de tâches de Windows a permis de surveiller les traitements. Excel est mis à contribution dans la phase de préparation de données (boîte de dialogue « Select Source Data ») (2). Au démarrage de la modélisation, SSAS prend le relais comme on peut le voir dans la copie d'écran ci-dessous (3). Les opérations sont scindées en 2 parties : transfert des données, puis construction du modèle.



Autre élément très important toujours dans cette copie d'écran, une partie des calculs sont multithreads. Les 4 cœurs de ma machine (99%) sont utilisés pendant une fraction non négligeable du processus[34]. Le passage à l'échelle (scalabilité) sur les machines multiprocesseurs ou à processeurs multi-cœurs semble naturel. La capacité de traitement paraît d'autant plus intéressante que l'occupation mémoire reste assez raisonnable compte tenu de la taille de notre ensemble de données.

Bien évidemment, pour les très gros volumes, le mieux est de ne pas passer par Excel du tout, et de travailler directement à partir d'une base importée dans SQL Server.

[34] Chapeau ! Ca rigole plus du tout là. Pour avoir moi-même exploré de nombreuses pistes dans le domaine, je sais combien il est difficile d'implémenter des solutions multithreads efficaces pour les arbres de décision. « Multithreading pour les arbres de décision », novembre 2010.

10 Conclusion

Je suis assez bluffé par la facilité d'utilisation du dispositif. Une demi-journée de prise en main suffit largement pour peu que l'on sache précisément ce que l'on souhaite faire. L'add-in démontre que l'on peut réaliser des études de qualité à partir d'Excel. La possibilité d'exploiter les fonctionnalités usuelles du tableur pour le prétraitement des données ou le post-traitement des résultats nous ouvre des perspectives intéressantes. Excel a encore de beaux jours devant lui dans le domaine du « data analytics ».

Tout n'est pas rose cependant. Certaines méthodes sous-jacentes aux outils sont très originales. Malgré une importante documentation en ligne, nous avons du mal à identifier précisément leur contenu scientifique. Les sorties sont parfois déroutantes, celles de la régression linéaire notamment. Microsoft a certainement le poids nécessaire pour imposer ses propres approches. Mais il est difficile d'adhérer à un résultat dont on ne cerne pas totalement la teneur.