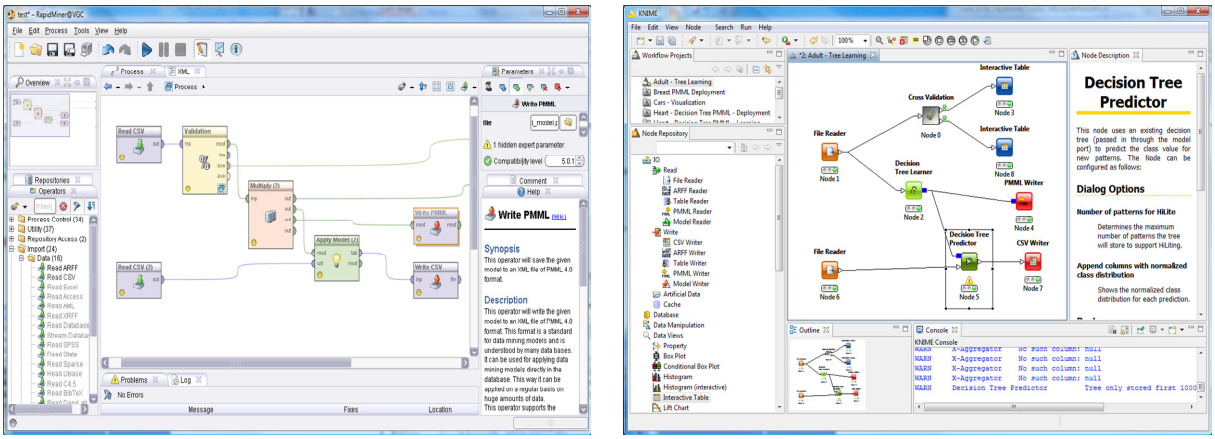
1 Objectif

Analyse de la nouvelle interface du logiciel RapidMiner 5.0.

La société Rapid-I, à travers leur logiciel phare RapidMiner, est un acteur très dynamique du l'informatique décisionnelle. Au-delà de l'outil, elle propose des solutions et des services dans le domaine de l'analyse prédictive, data mining et du text mining. Son site web regorge d'informations (blog, tutoriels, vidéos, forum, newsletter, wiki, etc.). Manifestement, elle sait créer le « buzz » (comment on le dit si bien de nos jours) pour promouvoir son activité. Ainsi, leur produit RapidMiner arrive systématiquement parmi les premiers du sondage annuel « kdnuggets » sur l'utilisation des logiciels de data mining (2010, 2009, 2008, 2007) ; s'attirant d'ailleurs les foudres d'acteurs installés, sûrs de leurs bon droit, tels que Weka. Nous ne sommes pas dupes. Derrière cette petite guéguerre se cache des enjeux commerciaux non négligeables, RapidMiner et Weka (via la suite Pentaho), proposent tous deux une version étendue commerciale[1], payante donc, dont le succès repose en partie sur la popularité de la version gratuite. Entre nous, nous savons qu'il faut toujours être prudent s'agissant de sondages avec des répondants volontaires en ligne sur le net. Néanmoins, la régularité du résultat sur plusieurs années mérite d'être soulignée.



RapidMiner Knime

Figure 1 - Fenêtres principales des logiciels RapidMiner et Knime

La version 5.0 de RapidMiner propose une interface profondément remaniée, s'inspirant visiblement de Knime. Les ressemblances entre les deux produits sont frappantes (Figure 1). L'organisation de l'interface est très proche. Les composants sont regroupés dans une structure arborescente située sur la partie gauche du logiciel. L'enchaînement des traitements est représenté par un graphe (et non plus par arbre dans RapidMiner). On parle de « Process » dans la documentation. En cela, la version 5.0 rejoint la présentation adoptée par la grande majorité des logiciels de data mining, commerciaux ou gratuits. Il est possible d'établir une connexion avec le logiciel R pour profiter de l'immense bibliothèque de calcul de ce dernier. Les métas-nœud permettent d'encapsuler une série d'opérations dans un composant du graphe (lors de la validation croisée par exemple). Enfin, toujours à l'instar de Knime, la description du composant sélectionné est constamment visible dans la partie droite de la fenêtre principale. Ouh là là, j'imagine que certains doivent en faire des

[1] Les éditions payantes s'accompagnent surtout d'une assistance professionnelle. Argument choc pour les entreprises. Sur le site de RapidMiner, nous pouvons observer les différences entre les versions (http://rapid-i.com/content/view/181/190/). La « Enterprise Edition – Developer » est vendue 10.000 euros (au 01/10/2010).

cauchemars. Mais bon, on sait depuis le litige qui a opposé Borland à Lotus, le tableur Quattro Pro s'inspirant très largement de Lotus 1-2-3, qu'il ne peut pas y avoir de copyright sur une interface et des fonctionnalités. Et puis, après tout, il n'y a pas 36.000 manières d'organiser un logiciel dont les tâches sont représentées par un flux de traitements.

RapidMiner ayant évolué de manière substantielle, je me suis dit qu'il était opportun d'étudier cela en détail, en évaluant son comportement dans le cadre d'une analyse type. Nous souhaitons mettre en place le processus suivant : (1) construire et afficher un arbre de décision à partir d'un ensemble d'observations étiquetées ; (2) sauvegarder l'arbre dans un fichier au format PMML en vue d'un déploiement ultérieur ; (3) évaluer les performances en généralisation du classifieur à travers la validation croisée ; (4) utiliser le modèle pour classer un ensemble d'observations non étiquetées contenues dans un second fichier, les résultats (descripteurs et étiquette attribuée) doivent être consignés dans un troisième fichier au format CSV. Ce sont là des tâches très classiques du data mining. Nous les avons maintes fois décrites dans nos didacticiels. Raison de plus pour vérifier s'il est aisé de les mener à bien avec cette nouvelle version de RapidMiner. En effet, avec la précédente mouture, certains enchaînements étaient compliqués. Mettre en place une validation croisée par exemple demandait une organisation, certes très rigoureuse dans son esprit, mais peu intuitive (http://tutoriels-data-mining.blogspot.com/2008/11/validation-croise-comparaison-de.html).

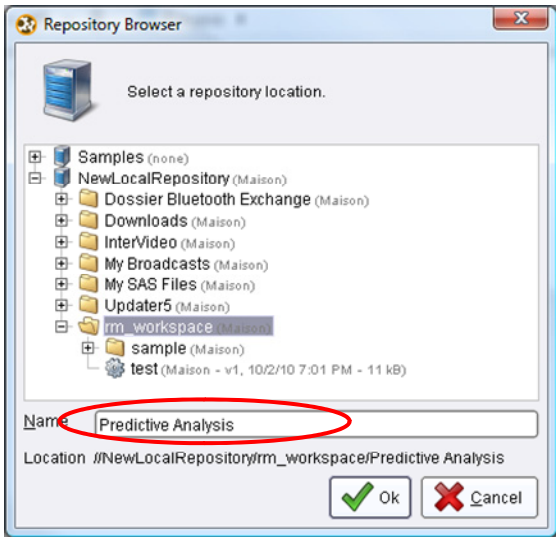
2 Données

Nous utilisons deux versions de la base « adult » (http://archive.ics.uci.edu/ml/datasets/Adult) du serveur UCI. La première, « adult_labeled.csv », contient les observations étiquetées. La colonne « classe » indique la classe d'appartenance des individus. La seconde, « adult_unlabeled.csv », contient les observations non étiquetées que nous devons classer à l'aide du modèle prédictif.

3 Analyse avec RapidMiner 5.0

3.1 Création d'un projet

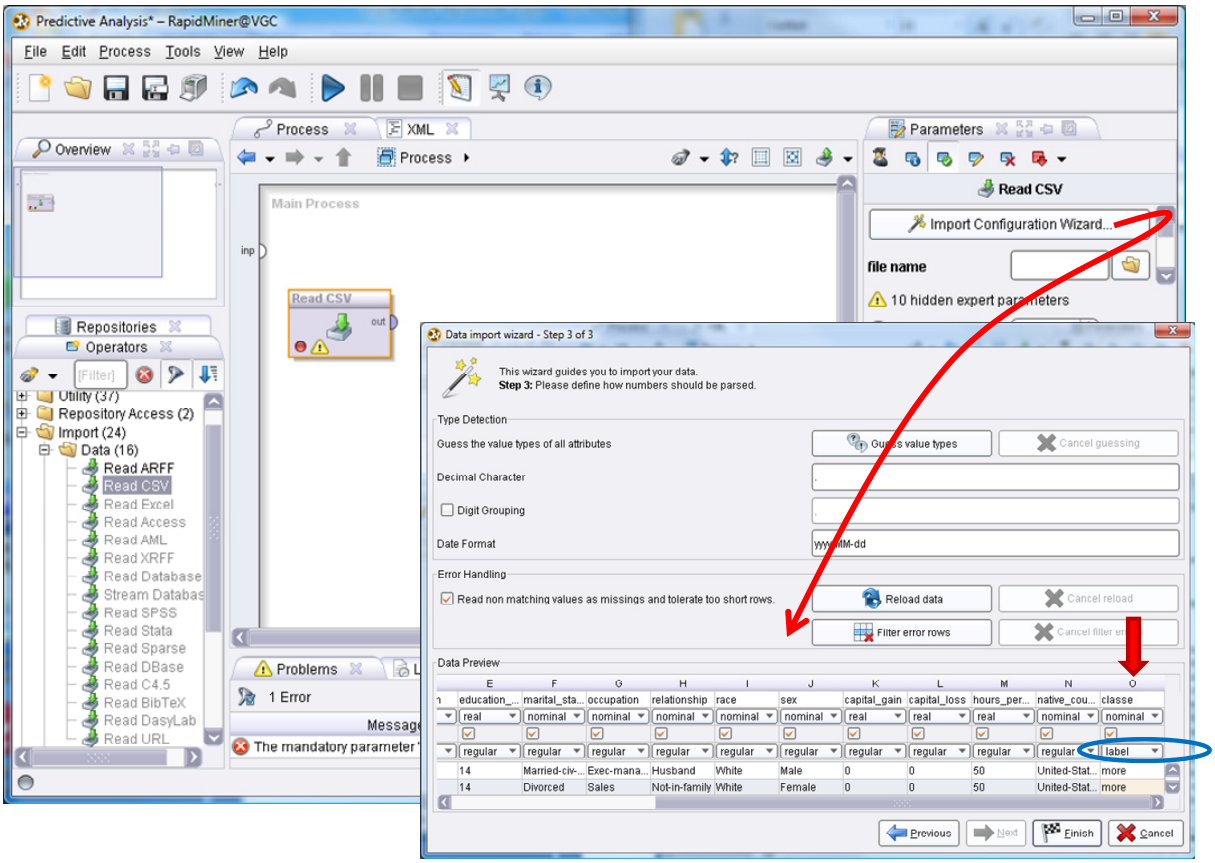
Au démarrage du logiciel, nous initions un nouveau projet en actionnant le menu FILE / NEW. Dans la boîte de paramétrage, nous indiquons le nom du processus qui sera sauvegardé dans l'espace de travail dédié à RapidMiner sur notre disque local. Nous l'appelons « Predictive Analysis ».



Nous disposons maintenant d'un espace de travail pour définir notre analyse.

3.2 Importation des données étiquetées

Nous utilisons le composant « READ CSV » pour lire le fichier des données étiquetées. Nous pouvons spécifier directement le nom du fichier dans la section de paramétrage située dans la section en haut et à droite de la fenêtre principale. Néanmoins il est préférable d'utiliser le « wizard » d'importation. Il nous permet entre autres de typer correctement les variables (quantitatives ou qualitatives) et de désigner directement la variable cible.

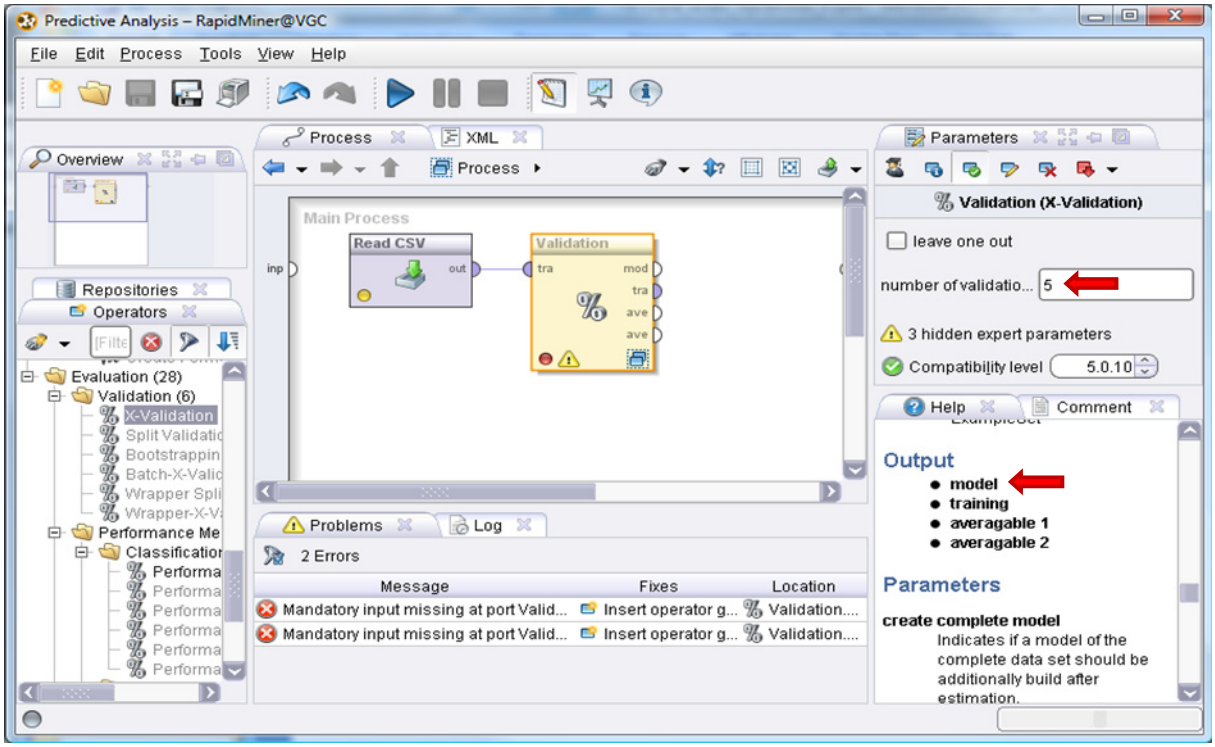


Pour simplifier, nous désignons comme variable continue (réelle) toutes les colonnes numériques. La colonne « classe » représente la variable cible « label ». Nous validons en cliquant sur le bouton « FINISH ».

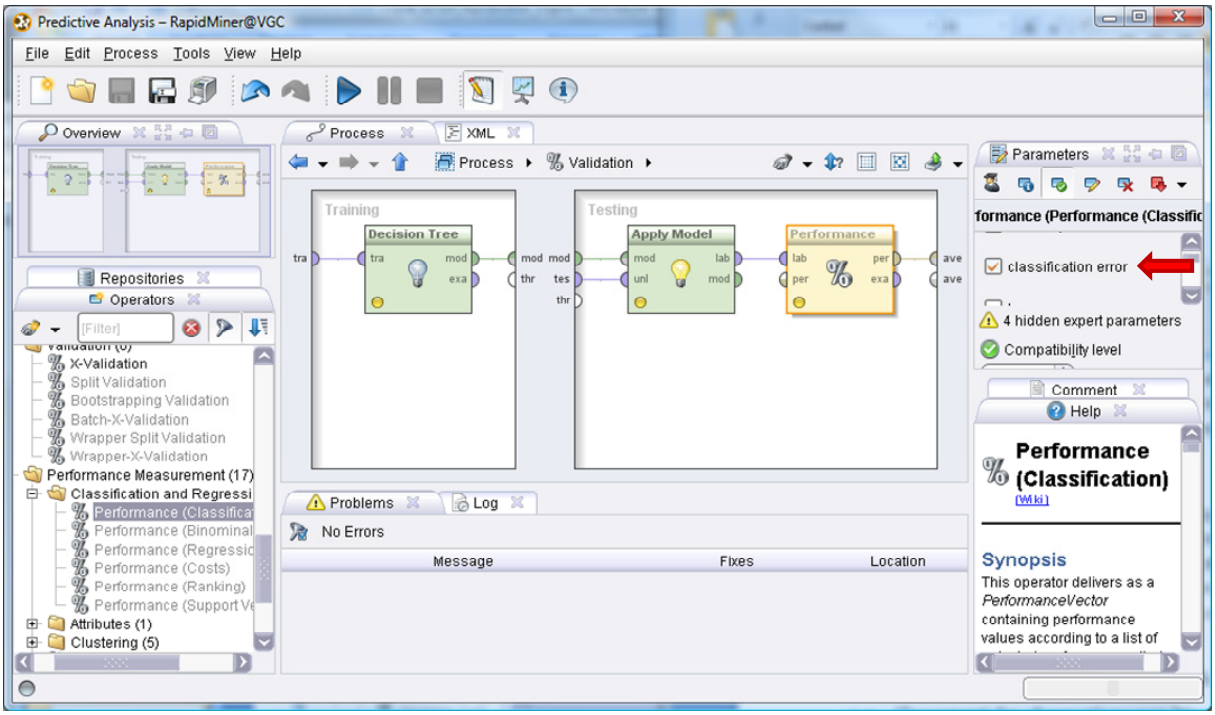
3.3 Construction de l'arbre et validation croisée

La construction de l'arbre sur la totalité des données et son évaluation en validation croisée sont curieusement intimement liées dans RapidMiner. Plutôt que de placer le composant « Arbre de décision » dans le diagramme, nous insérons le méta-composant « X-Validation ». Nous lui connectons l'outil d'accès aux données.

Nous demandons une validation croisée en 5 portions. On note surtout, en consultant l'aide contextuelle, que nous pouvons récupérer le modèle élaboré sur la totalité des observations en sortie du composant. Nous souhaitons le visualiser, le sauvegarder dans un fichier au format PMML, et l'utiliser pour le classement des individus non étiquetés.

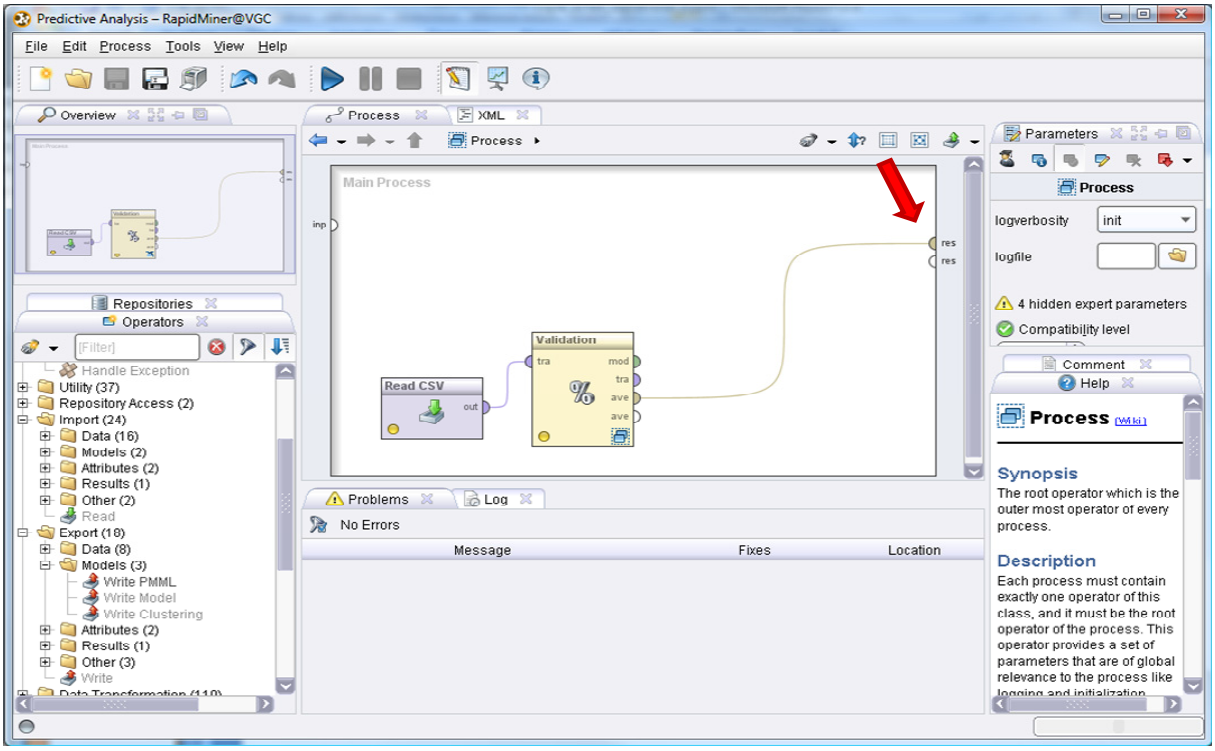


Il nous faut choisir la méthode d'apprentissage et le critère de performance maintenant. Pour ce faire, nous double-cliquons sur le composant. On se rend compte que la validation croisée est un méta-nœud, nous tombons sur une nouvelle interface permettant de spécifier l'enchaînement d'opérations qui la compose.

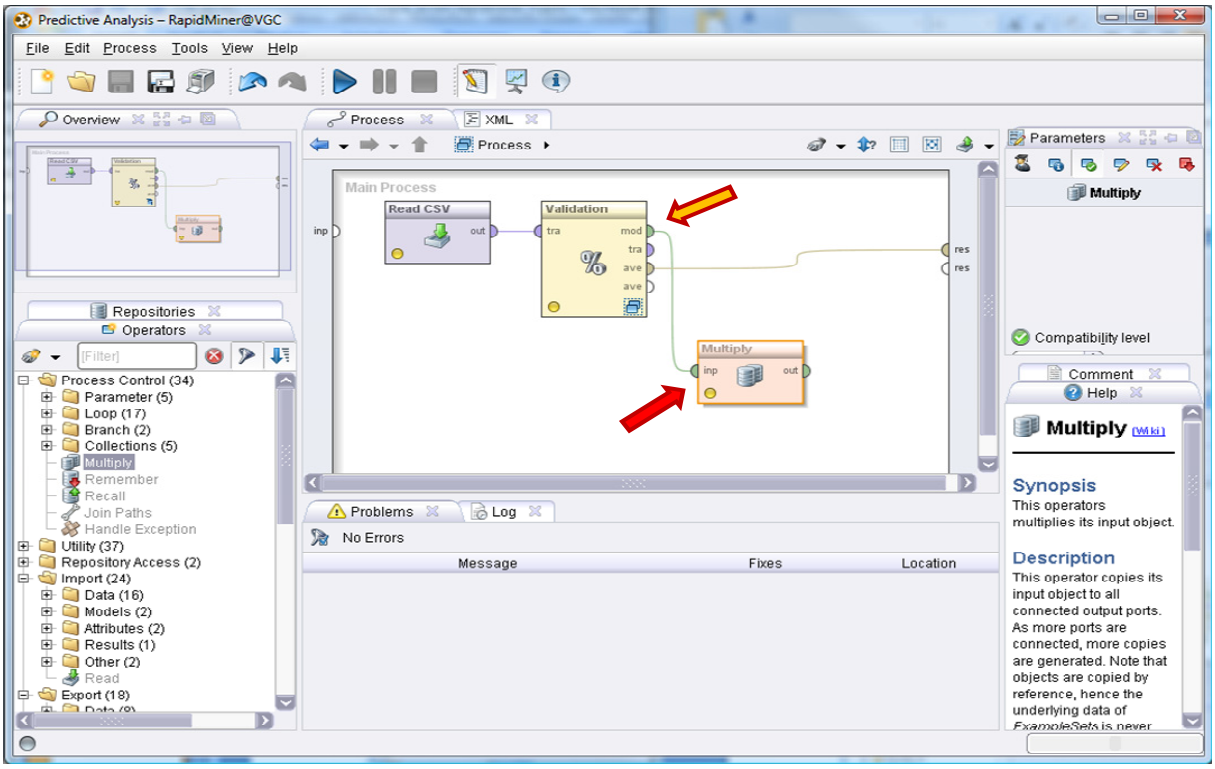


Dans la partie apprentissage « training », nous plaçons un arbre de décision « Decision Tree ». Dans la partie test « Testing », nous plaçons tour à tour l'applicateur du modèle sur la partie test des données « Apply Model » et l'outil de mesure des performances en classement « Performance (Classification) ». Nous choisissons le critère « taux d'erreur » (classification error). Remarquons les différentes connexions entre les composants, puis en entrée et sortie du méta-composant.

Pour que les résultats de la validation croisée (matrice de confusion et taux d'erreur) soient accessibles dans les sorties du logiciel, nous devons l'ajouter dans les résultats. Nous connectons VALIDATION (3ème slot) au slot RES du processus pour cela.



3.4 Démultiplication de l'objet modèle

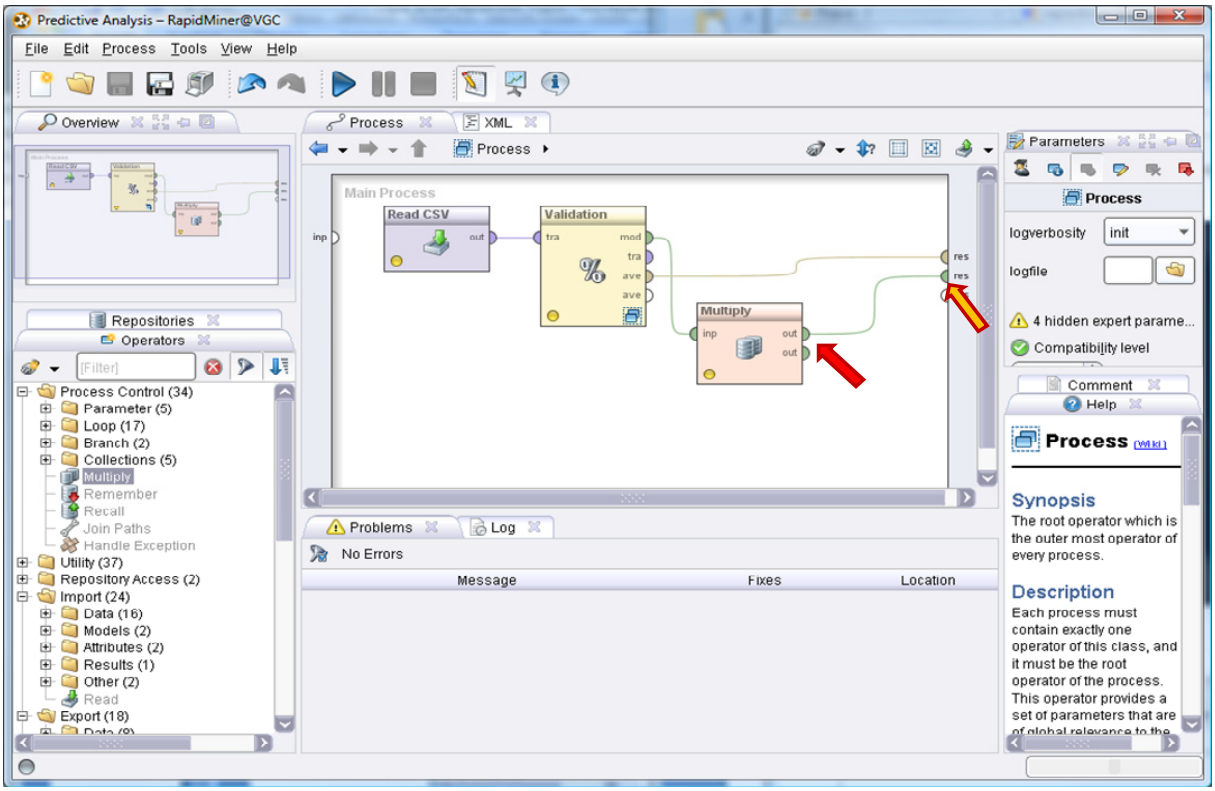


L'arbre de décision élaboré sur la totalité des données doit être affiché à l'issue des calculs. Nous l'appliquerons sur les données non étiquetées. Enfin, nous souhaitons le sauver dans un fichier au format PMML. Lorsqu'un élément (données, modèles, etc.) est exploité de différentes manières en

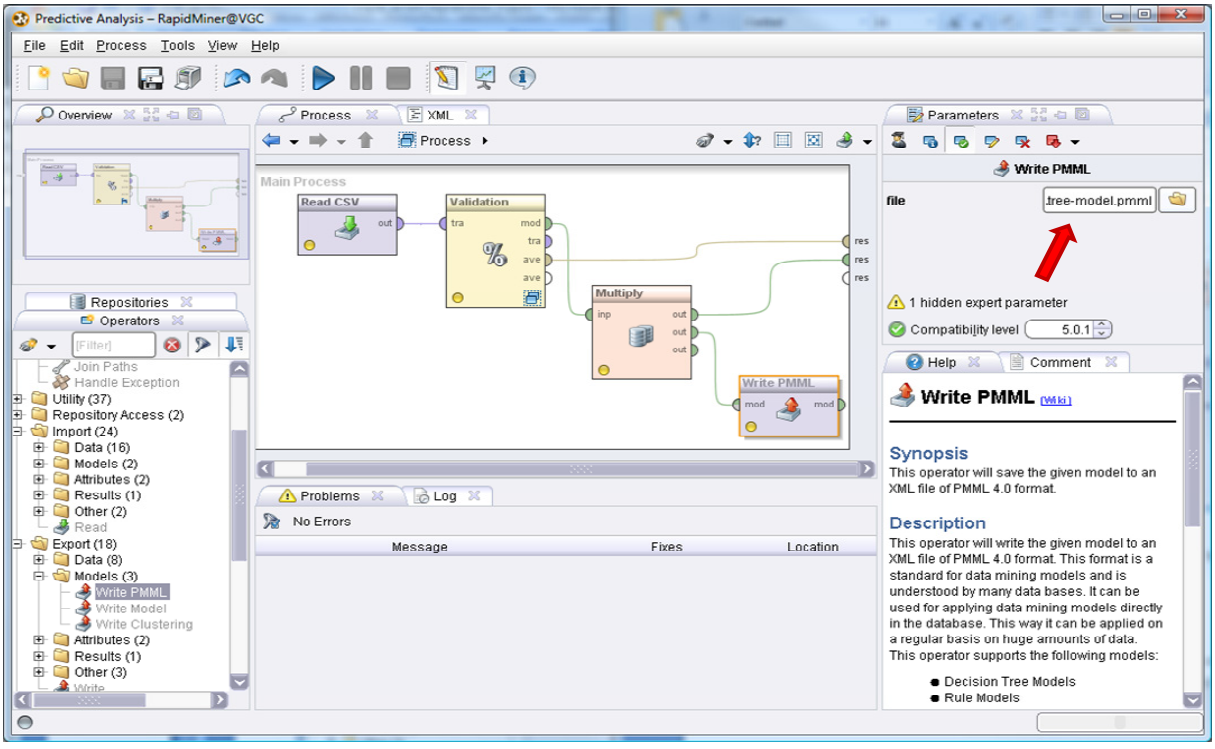
aval d'un composant, nous devons le démultiplier à l'aide du composant MULTIPLY. Remarquons le slot utilisé en sortie de VALIDATION (1er slot : « mod » pour modèle).

3.5 Affichage de l'arbre

Pour afficher l'arbre, nous l'ajoutons dans les résultats du processus. Nous remarquons qu'un slot supplémentaire est alors disponible en sortie de MULTIPLY.



3.6 Sauvegarde de l'arbre au format PMML



Nous ajoutons le composant WRITE PMML pour exporter le modèle au format PMML. Nous spécifions le nom du fichier.

3.7 Classement des observations non étiquetées

Pour appliquer le modèle sur les observations non étiquetées et sauvegarder les résultats (descripteurs + classe attribuée) dans un fichier de sortie au format CSV, nous ajoutons la séquence suivante dans le processus.

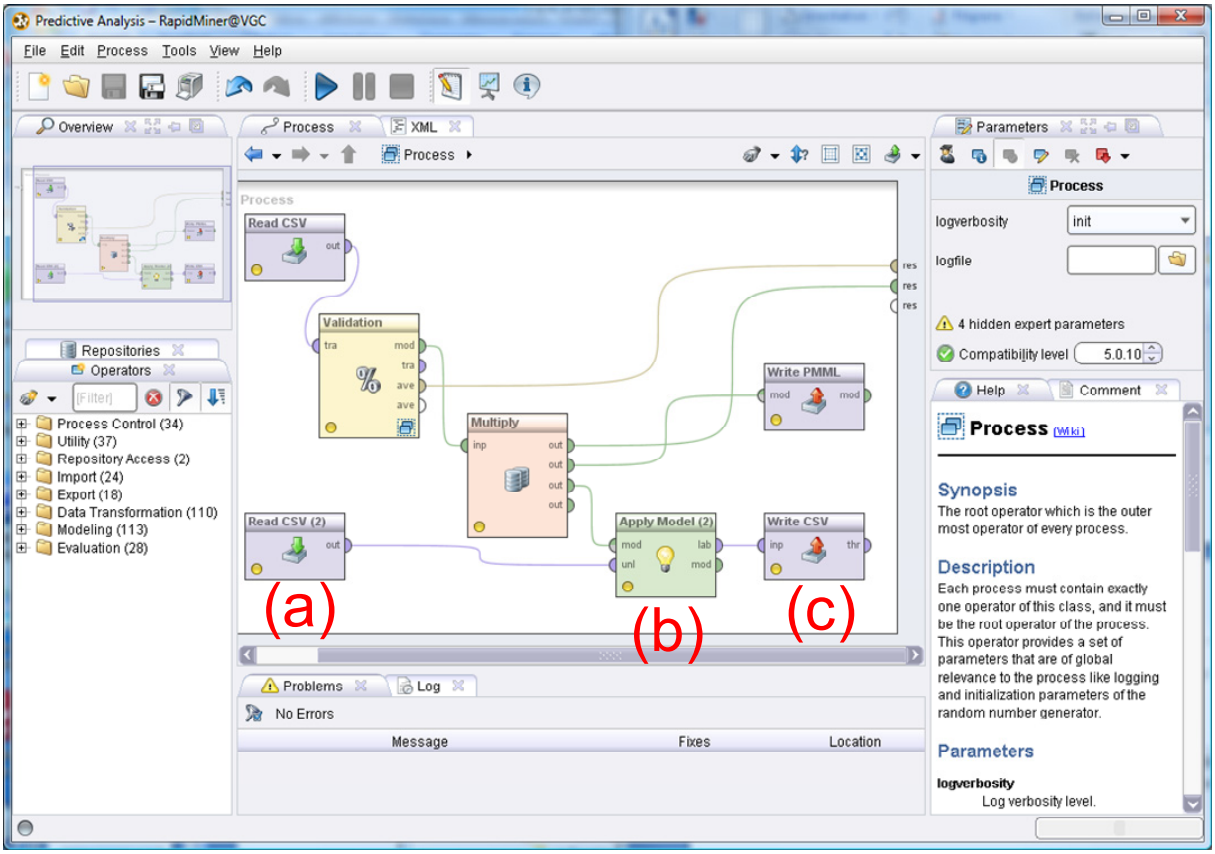


Figure 2 - L'analyse complète sous RapidMiner

(a) Nous importons le fichier « adult_unlabeled.csv » à l'aide de READ CSV. Il contient les descripteurs mais pas la classe. Nous devons veiller à typer correctement les variables (REAL ou NOMINAL) en calquant notre définition avec celle des données d'apprentissage.

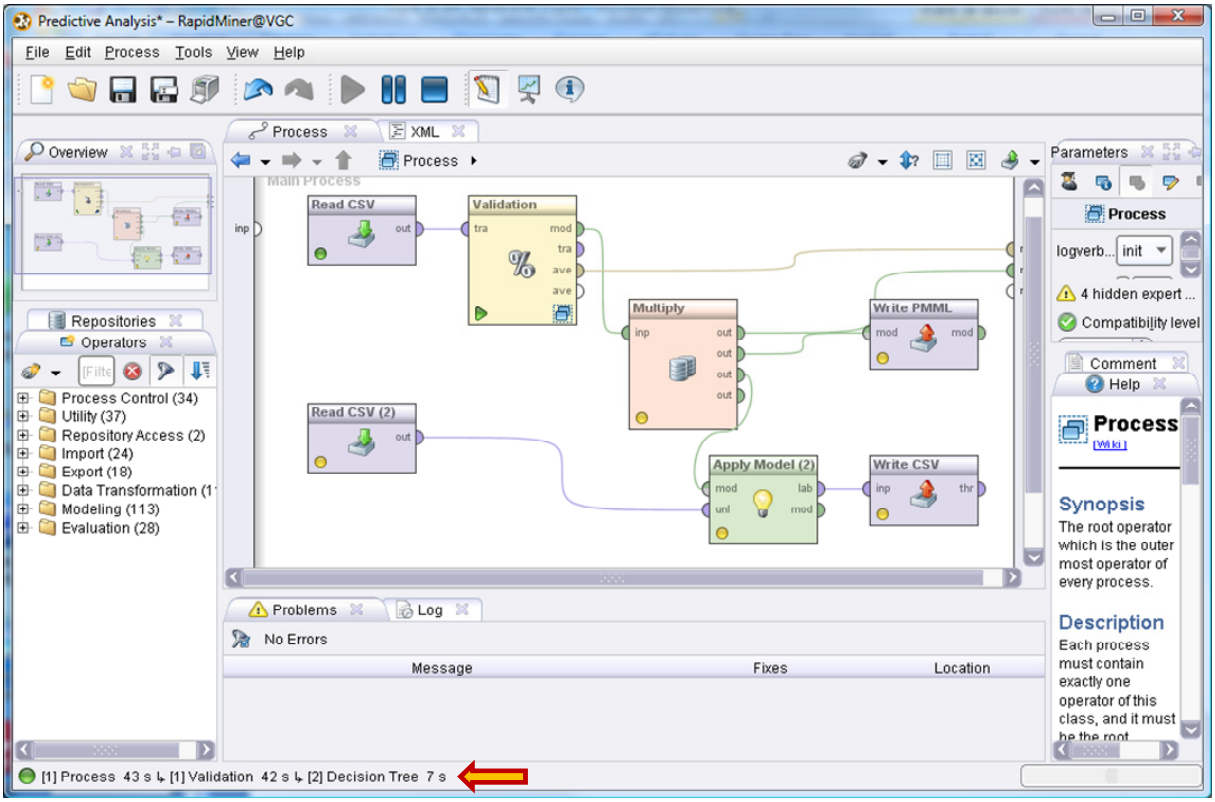
(b) APPLY MODEL sert à appliquer le classifieur sur les données non étiquetées.

(c) WRITE CSV se charge de sauver les données dans un nouveau fichier que nous avons nommé « adult_with_predictions.csv ».

3.8 Exécution des calculs

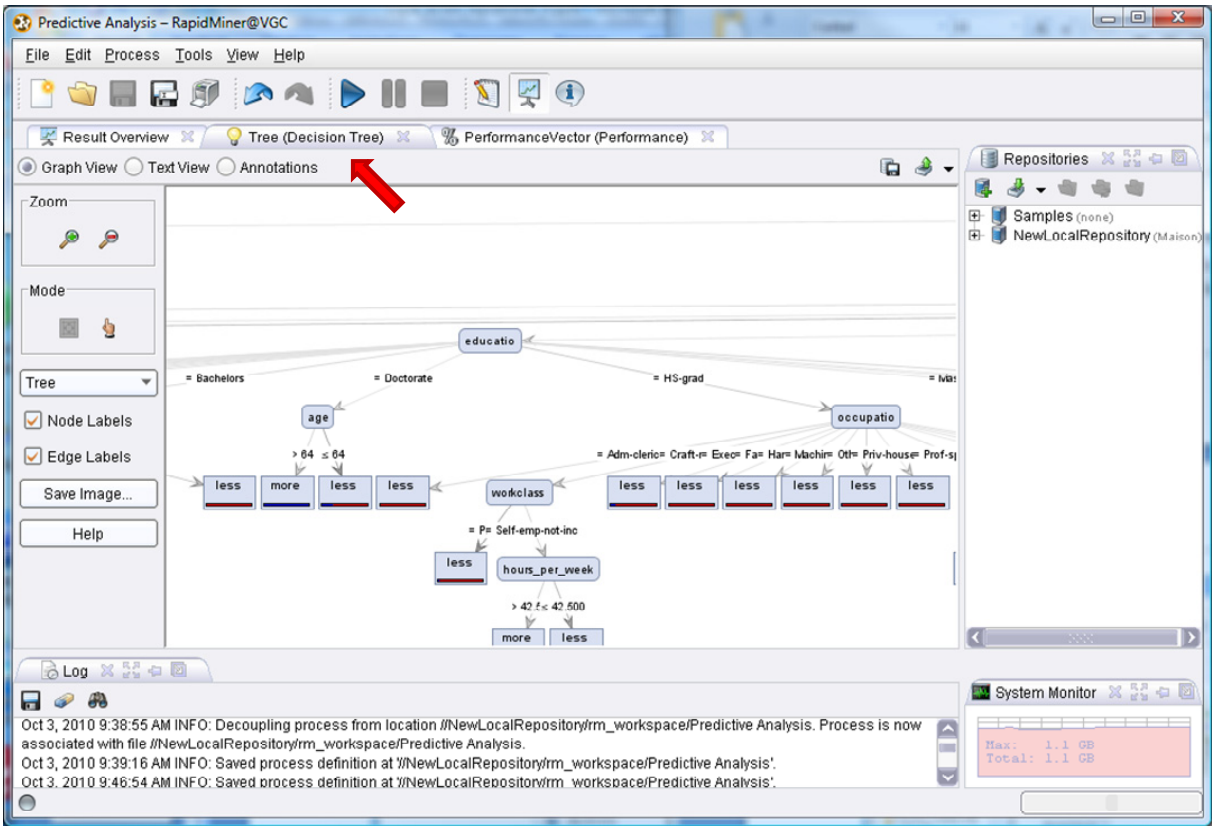
Après avoir sauvegardé le processus, nous lançons les calculs en actionnant le menu PROCESS / RUN. Il est également possible de cliquer directement sur le bouton idoine dans la barre d'outils.

La taille du fichier étant relativement importante et la validation croisée gourmande en ressources, les calculs peuvent prendre un certain temps (un temps certain même). Nous pouvons suivre leur avancée dans la barre d'état située dans la partie basse de la fenêtre principale.

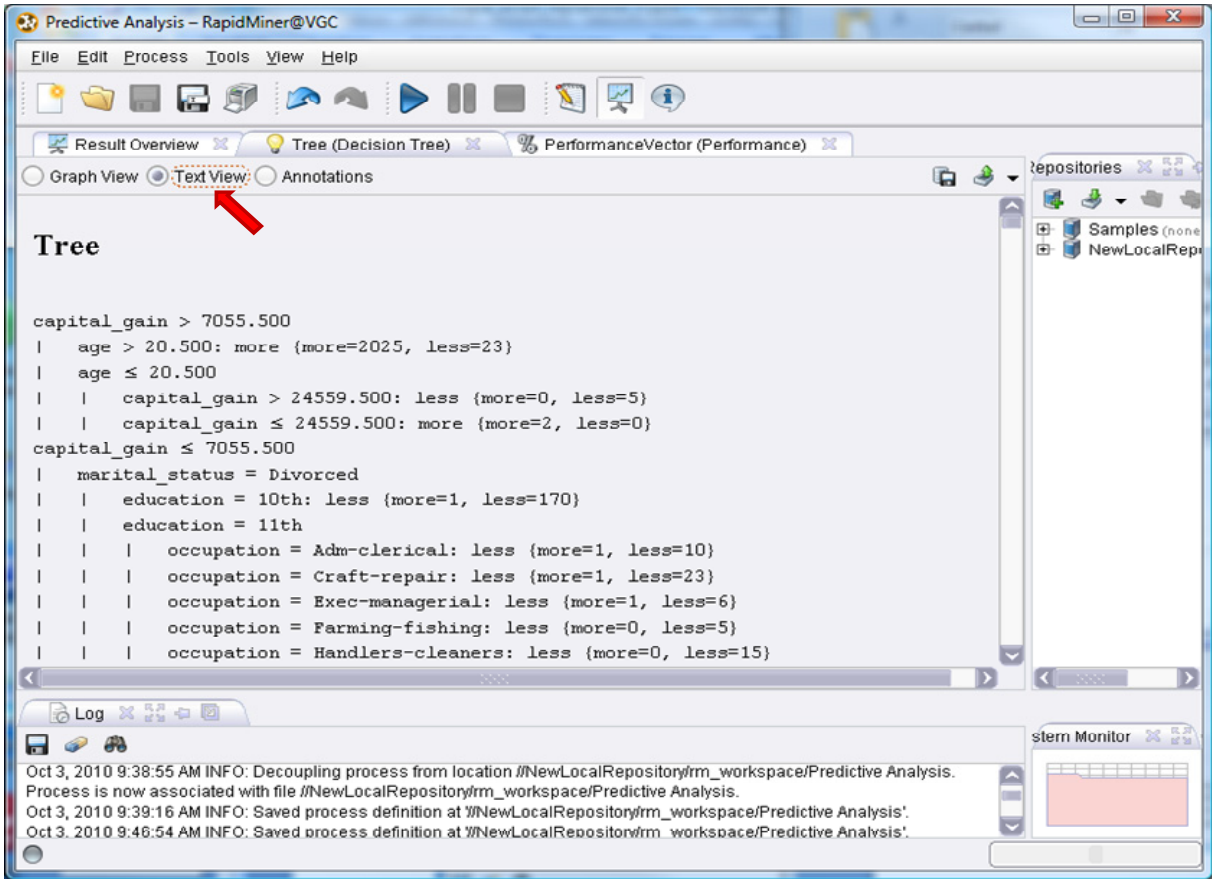


3.9 Résultats

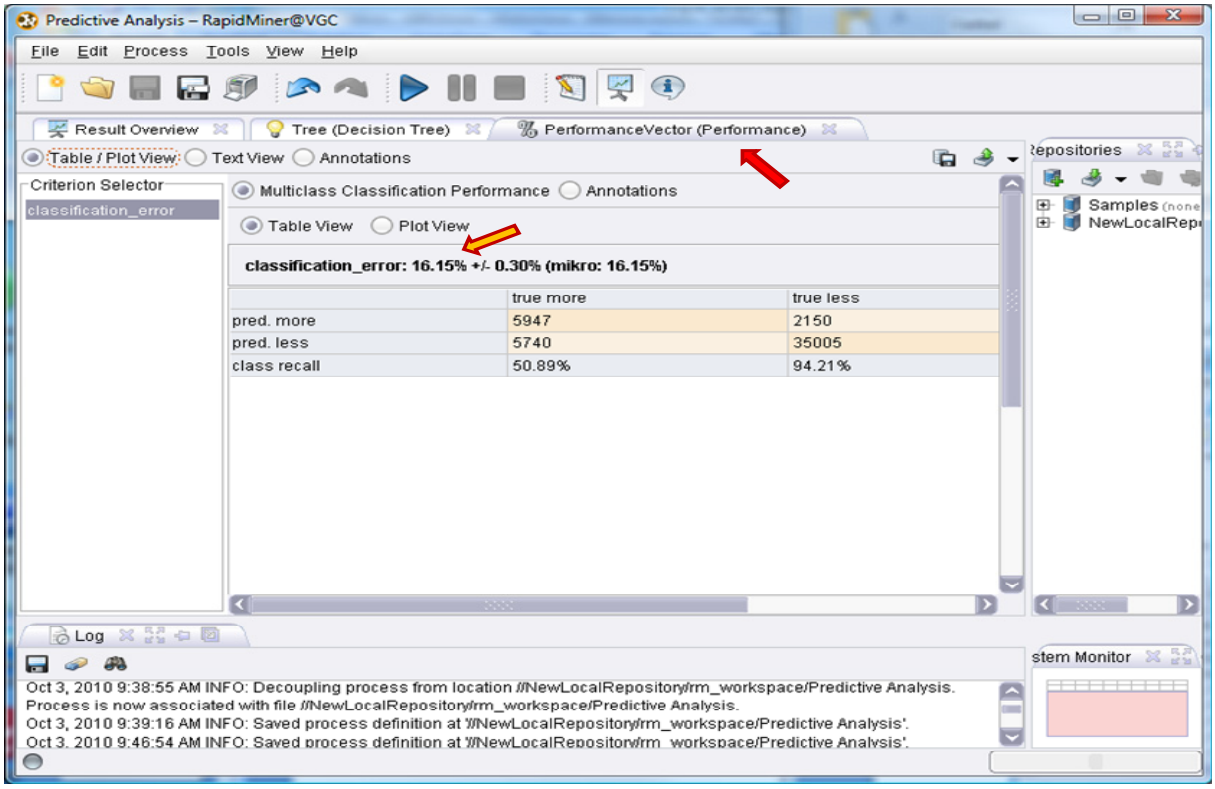
Plusieurs fenêtres sont générées à l'issue des calculs. Dans l'onglet TREE, nous avons accès à l'arbre de décision. La représentation graphique n'est pas très lisible compte tenu de la taille de l'arbre.



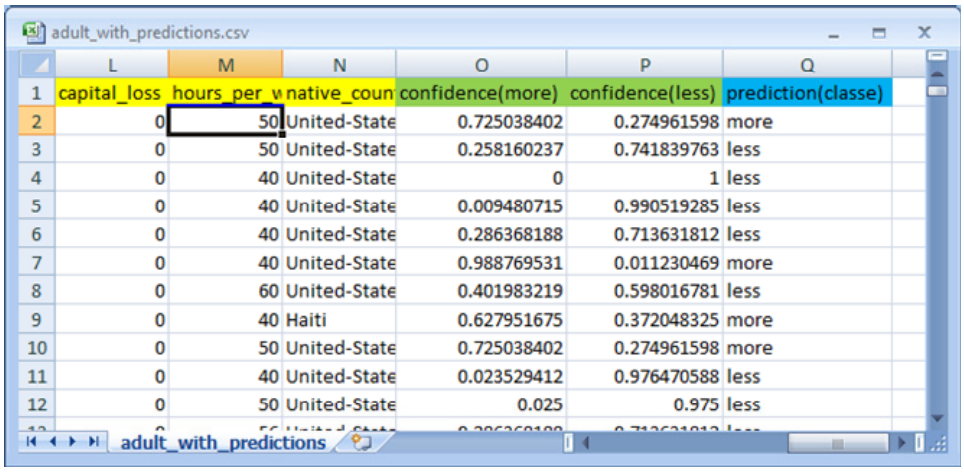
Nous préférerons l'affichage textuel sous forme de « text view ».



Dans l'onglet « Performance Vector », nous avons accès à la matrice de confusion et au taux d'erreur en validation croisée. Si nous appliquons le modèle sur un individu pris au hasard dans la population, nous avons 16.15% de chances de faire une mauvaise prédiction.



Enfin, concernant le déploiement sur le fichier des observations non étiquetées, nous avons chargé le fichier « adult_with_predictions.csv » dans un tableur.



adult_with_predictions.csv

	L	M	N	O	P	Q
1	capital_loss	hours_per_w	native_coun	confidence(more)	confidence(less)	prediction(classe)
2	0	50	United-State	0.725038402	0.274961598	more
3	0	50	United-State	0.258160237	0.741839763	less
4	0	40	United-State	0	1	less
5	0	40	United-State	0.009480715	0.990519285	less
6	0	40	United-State	0.286368188	0.713631812	less
7	0	40	United-State	0.988769531	0.011230469	more
8	0	60	United-State	0.401983219	0.598016781	less
9	0	40	Haiti	0.627951675	0.372048325	more
10	0	50	United-State	0.725038402	0.274961598	more
11	0	40	United-State	0.023529412	0.976470588	less
12	0	50	United-State	0.025	0.975	less

adult_with_predictions

En sus de la classe attribuée (en bleu), nous obtenons les probabilités d'affectation (en vert). Elles nous permettent de juger de la crédibilité de la prédiction.

4 Conclusion

Nous avons présenté succinctement la nouvelle version 5.0 du logiciel RapidMiner dans ce tutoriel. Elle constitue une évolution notable par rapport à la précédente. On ne manquera pas de souligner encore une fois les similitudes avec Knime. Pour le même processus, nous aurions défini un diagramme (Figure 3 - L'analyse complète sous Knime), très proche dans sa forme de celui que nous avions spécifié dans RapidMiner (Figure 2 - L'analyse complète sous RapidMiner). La concurrence est manifestement exacerbée.

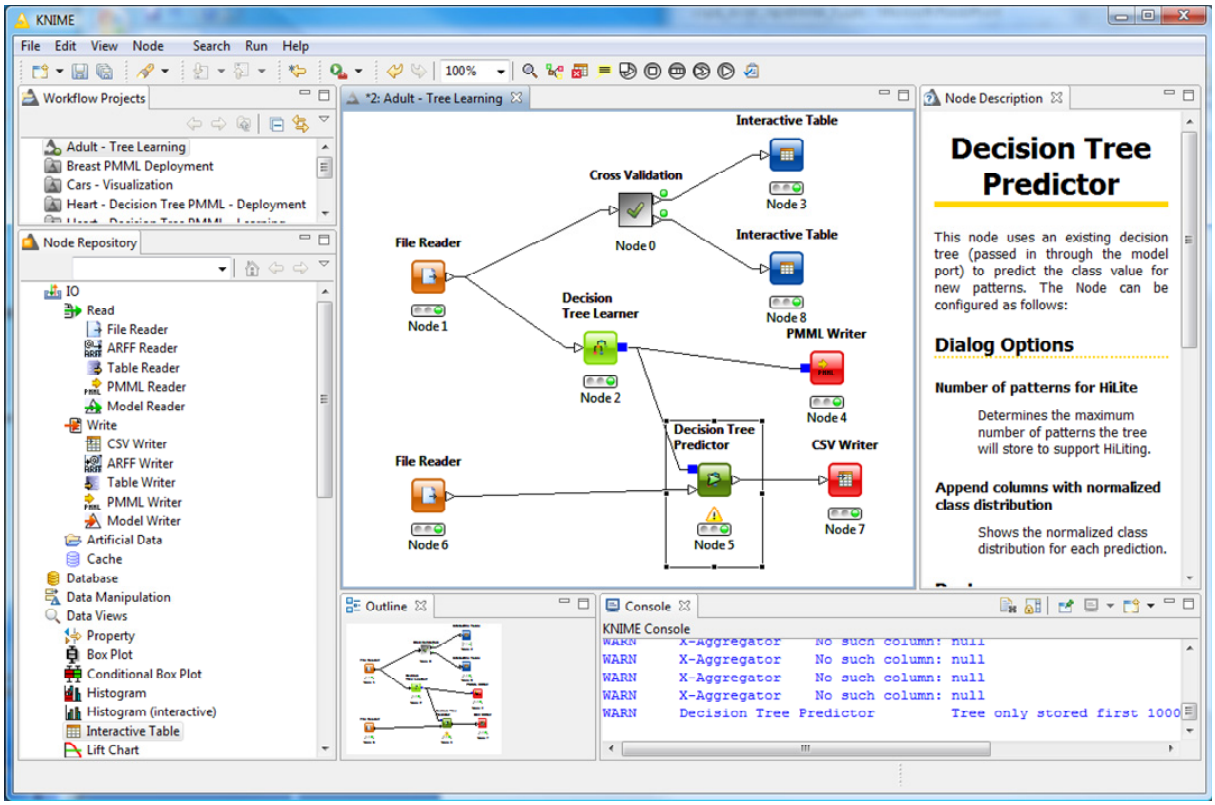


Figure 3 - L'analyse complète sous Knime

Lequel choisir ? Je resterai très prudent sur ce terrain là. RapidMiner propose une panoplie d'outils de modélisation impressionnante. Knime, de son côté, est plus simple d'utilisation et fait un effort méritoire sur les outils de management des données (préparation, recodage, etc.). Je note en tous les cas que nous, utilisateurs (chercheurs, étudiants, analystes, chargés d'études, etc.), sommes les principaux bénéficiaires de cette saine émulation qui pousse ces sociétés à mettre en ligne des versions gratuites de leurs logiciels. Tant mieux ! D'autant plus que tous deux sont d'excellente facture.