Objectif

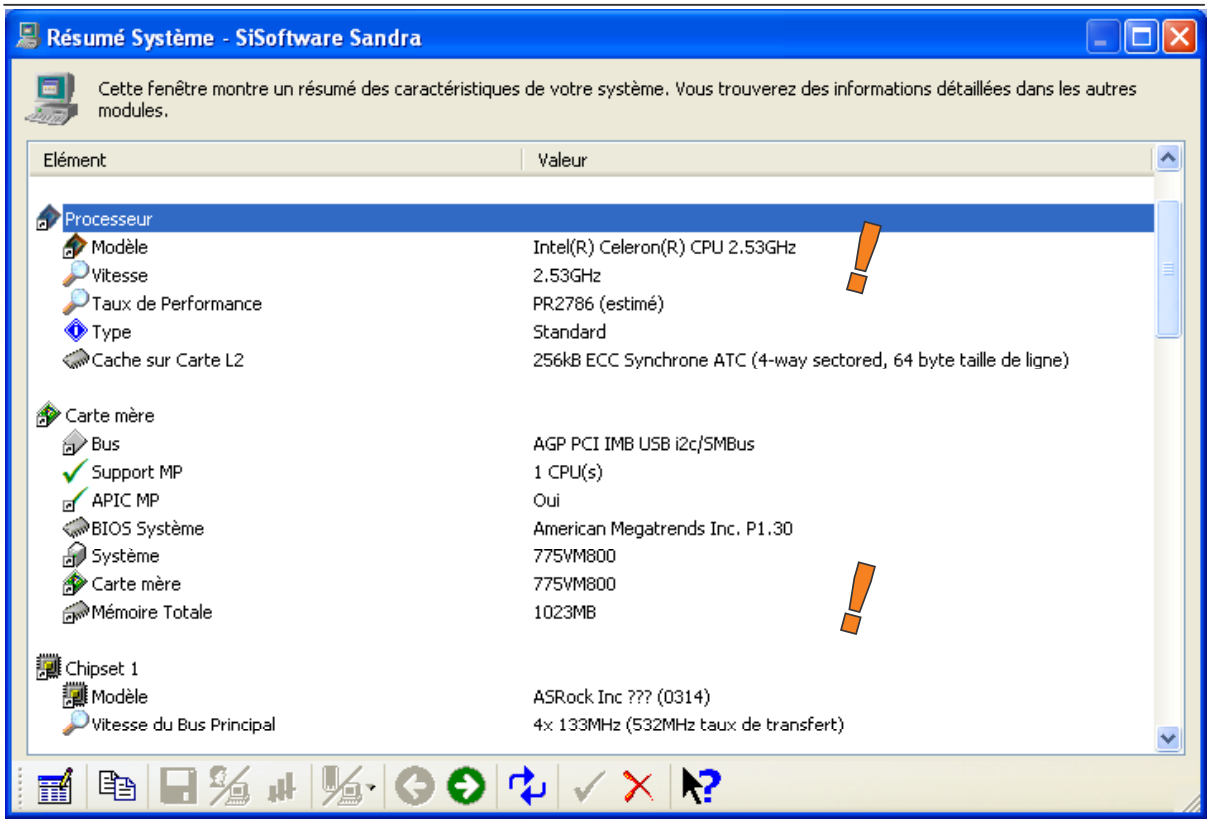
Traitement sur un gros fichier de données : importation et construction d'un arbre de décision avec la méthode ID3.

Une des principales nouveautés de ces dernières années est l'évolution quasi-exponentielle du volume des fichiers que nous sommes emmenés à traiter. Il y a une dizaine d'années encore, un tableau de 5000 observations avec 22 variables, les fameuses « ondes de Breiman », faisait figure de « gros fichier » au sein de la communauté de l'apprentissage automatique. Aujourd'hui, les fichiers connaissent une inflation plus que galopante avec, selon les domaines, une augmentation importante du nombre d'observations (les bases marketing par exemple) et/ou du nombre de descripteurs (en bio-informatique par exemple, en réalité tous les domaines où les descripteurs sont générés automatiquement).

La capacité à traiter les gros ensembles de données est un critère important de différenciation entre les logiciels de recherche et les logiciels commerciaux. Très souvent les outils commerciaux disposent de systèmes de gestion de données très performants, limitant la quantité de données chargée en mémoire à chaque étape du traitement. Les outils de recherche en revanche conservent toutes les données en mémoire, en les codant au mieux de manière à ce que l'occupation mémoire ne soit pas prohibitive. Les limites sont donc clairement les capacités de la machine utilisée.

Cette limitation constitue très certainement une barrière pour le traitement de gros fichiers. On se rend compte néanmoins qu'avec l'évolution actuelle des machines où, à moindre frais, on peut disposer d'ordinateurs très performants, elle est sans cesse reculée. Avec une stratégie d'encodage appropriée, nous pouvons faire tenir en mémoire et traiter facilement des fichiers de taille importante.

Dans ce didacticiel, nous montrons comment, avec TANAGRA, importer un fichier de 581012 observations et 55 variables, puis construire un arbre de décision avec la méthode ID3. Notre machine est un PC tout à fait banal dont les caractéristiques ont été mesurées avec une version shareware de SISOFTWARE SANDRA. Il s'agit d'un CELERON 2.53 GHz avec 1 GB de mémoire RAM fonctionnant sous Windows XP SP2. Ces informations sont importantes car elles vous permettront de comparer les performances rapportées dans ce didacticiel avec ceux que vous obtiendrez sur votre propre machine.



Fichier

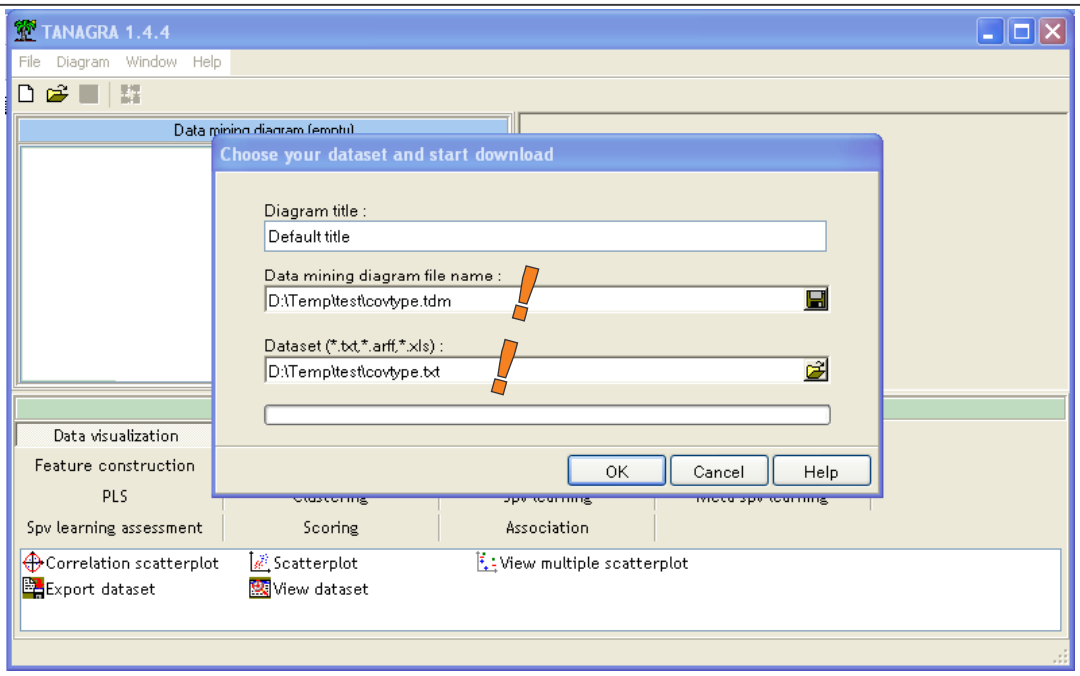
Le fichier COVTYPE contient 581102 observations, 54 descripteurs, tous discrets (ou discrétisés), la variable à prédire comporte 7 modalités. Le fichier texte associé occupe 62 Mo sur le disque dur.

Var01	Var02	Var03	Var04	Var05	Var06	Var07	Var08	Var09	Var10	Var11	Var12	Var13
A	A	A	A	A	A	A	A	A	A	A	A	A
A	A	A	A	A	A	A	A	A	A	A	A	A
B	B	A	A	A	B	A	A	A	A	A	A	A
B	C	B	A	A	B	A	A	A	A	A	A	A
A	A	A	A	A	A	A	A	A	A	A	A	A
A	B	A	B	A	A	A	A	A	A	A	A	A
A	A	A	A	A	A	A	A	A	A	A	A	A
A	A	A	A	A	A	A	A	A	A	A	A	A
A	A	A	A	A	A	A	A	A	A	A	A	A
A	A	A	A	A	A	A	A	A	A	A	A	A
A	C	A	A	A	A	A	A	B	A	A	A	A
B	C	A	B	A	C	A	A	A	B	A	A	A
B	B	B	A	A	B	A	A	C	A	A	A	A
A	C	A	A	A	A	A	A	B	A	A	A	A
A	C	A	A	A	A	A	A	A	C	A	A	A
A	A	A	A	A	A	A	A	A	C	A	A	A
A	D	A	A	A	A	A	A	B	A	A	A	A

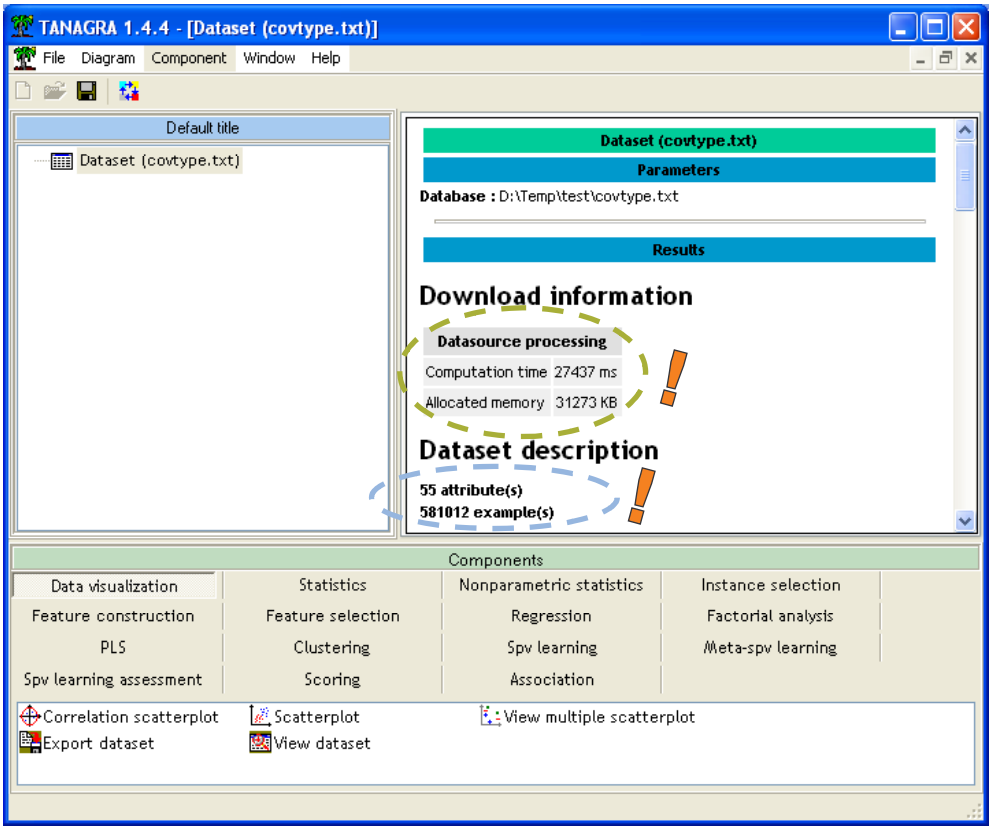
Traitement sur un gros fichier de données

Importation des données

Première étape toujours dans TANAGRA, créer un diagramme et importer les données (FILE / NEW).

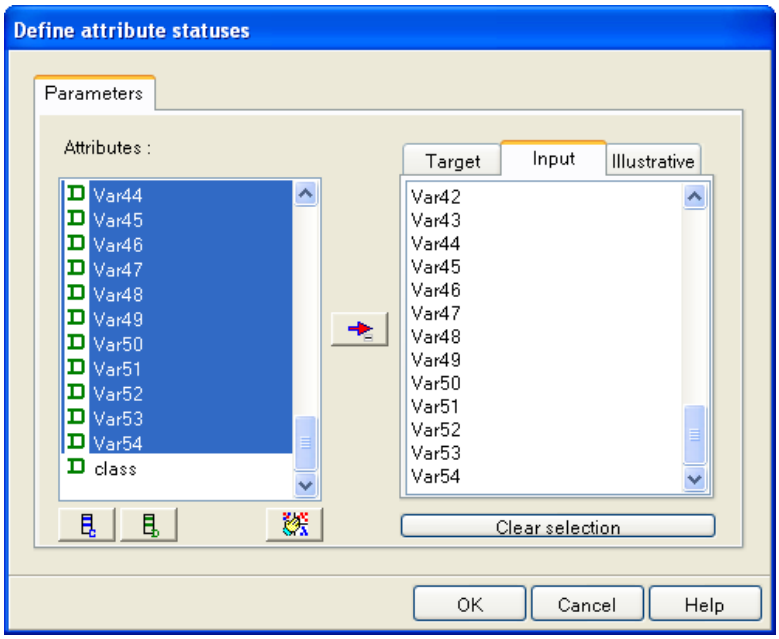


Nous validons en cliquant sur OK, nous observons alors le déroulement des opérations à l'aide de la barre de progression. Lorsque les données sont toutes chargées, un résumé apparaît avec le temps de traitement (# 27 s) et l'occupation mémoire des données après encodage (#31 MB).



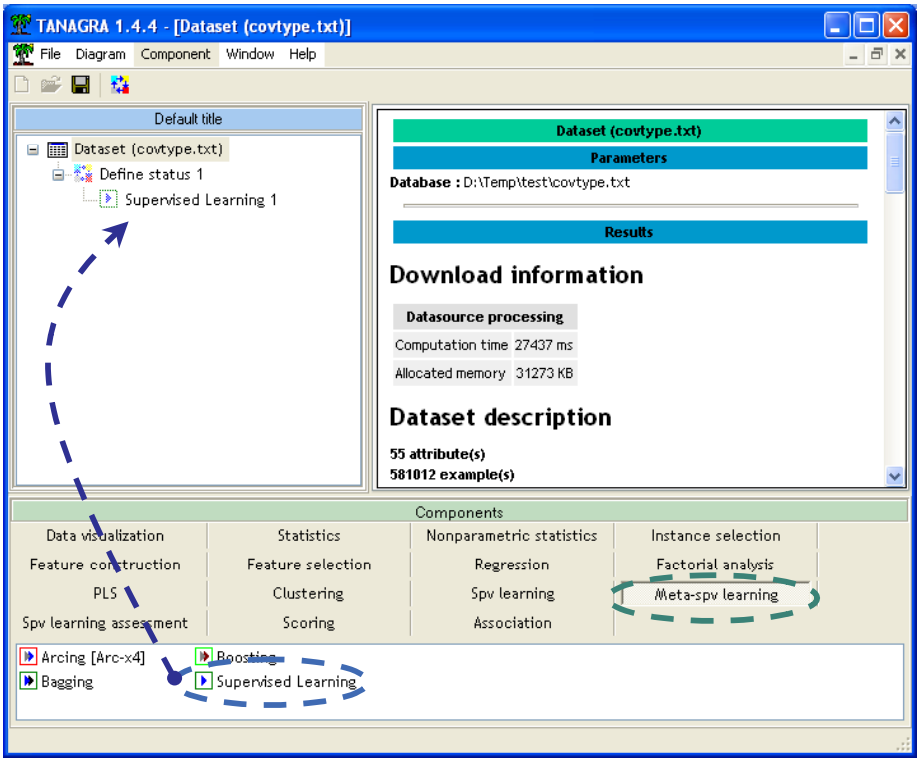
Sélection des variables

Nous plaçons le composant DEFINE STATUS (utiliser le raccourci dans la barre d'outils) dans le diagramme. L'attribut CLASS est la cible (TARGET), tous les autres sont les descripteurs (INPUT).

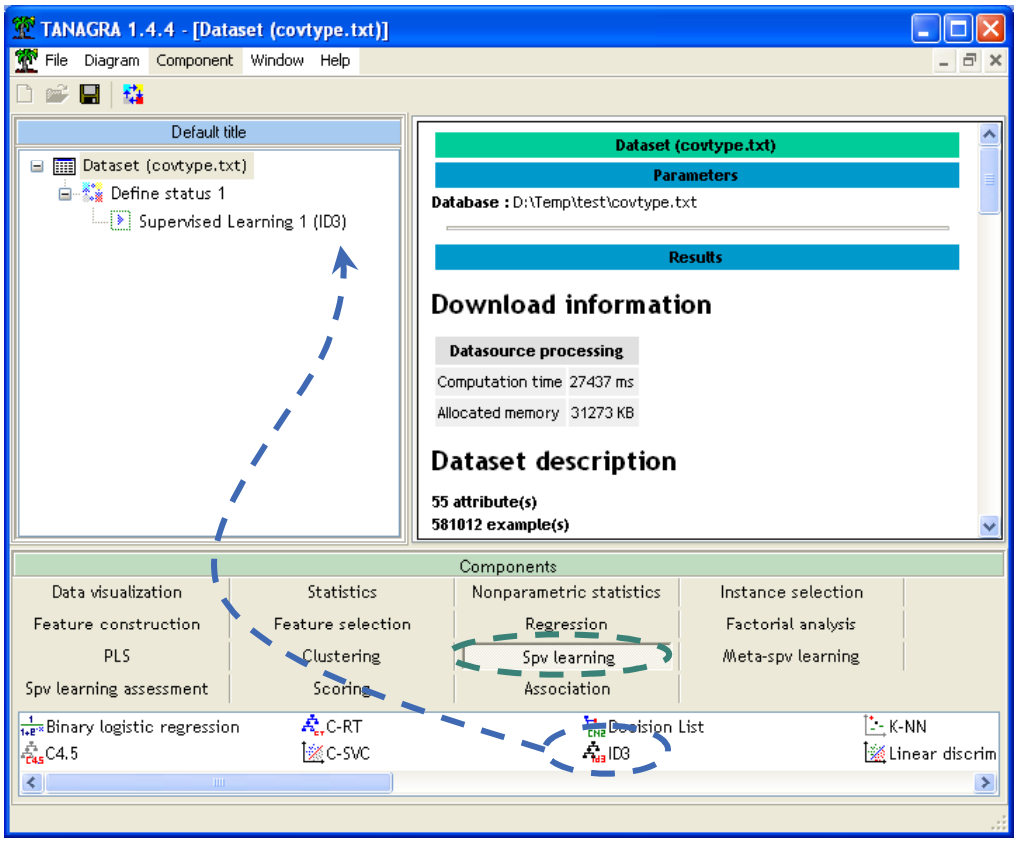


Construire d'un arbre de décision avec ID3

Il nous faut alors définir la méthode d'apprentissage. Nous voulons construire un arbre de décision avec la méthode ID3. Pour ce faire, nous procédons toujours en deux étapes : (1) placer le composant SPV LEARNING (onglet META SPV LEARNING)



(2) puis nous insérons la méthode d'apprentissage ID3 (onglet SPV LEARNING).



Il ne reste plus qu'à visualiser les résultats en cliquant sur VIEW du menu contextuel.

Classifier performances

Error rate	0.2972
Values prediction	Confusion matrix

Value	Recall	1-Precision
A	0.0077	0.4931
B	0.7552	0.2716
C	0.6938	0.3297
D	0.5501	0.1422
E	0.8136	0.3019
F	0.3295	0.4907
G	0.4561	0.3062

	A	B	C	D	E	F	G	Sum
A	73	9119	0	0	142	159	0	9493
B	53	213962	63143	135	2983	3020	5	283301
C	0	63001	146967	1736	68	68	0	211840
D	0	124	9103	11283	0	0	0	20510
E	9	4022	39	0	29089	2117	478	35754
F	9	3475	4	0	8086	5723	70	17367
G	0	46	0	0	1299	149	1253	2747
Sum	144	293749	219256	13154	41667	11236	1806	581012

Classifier characteristics

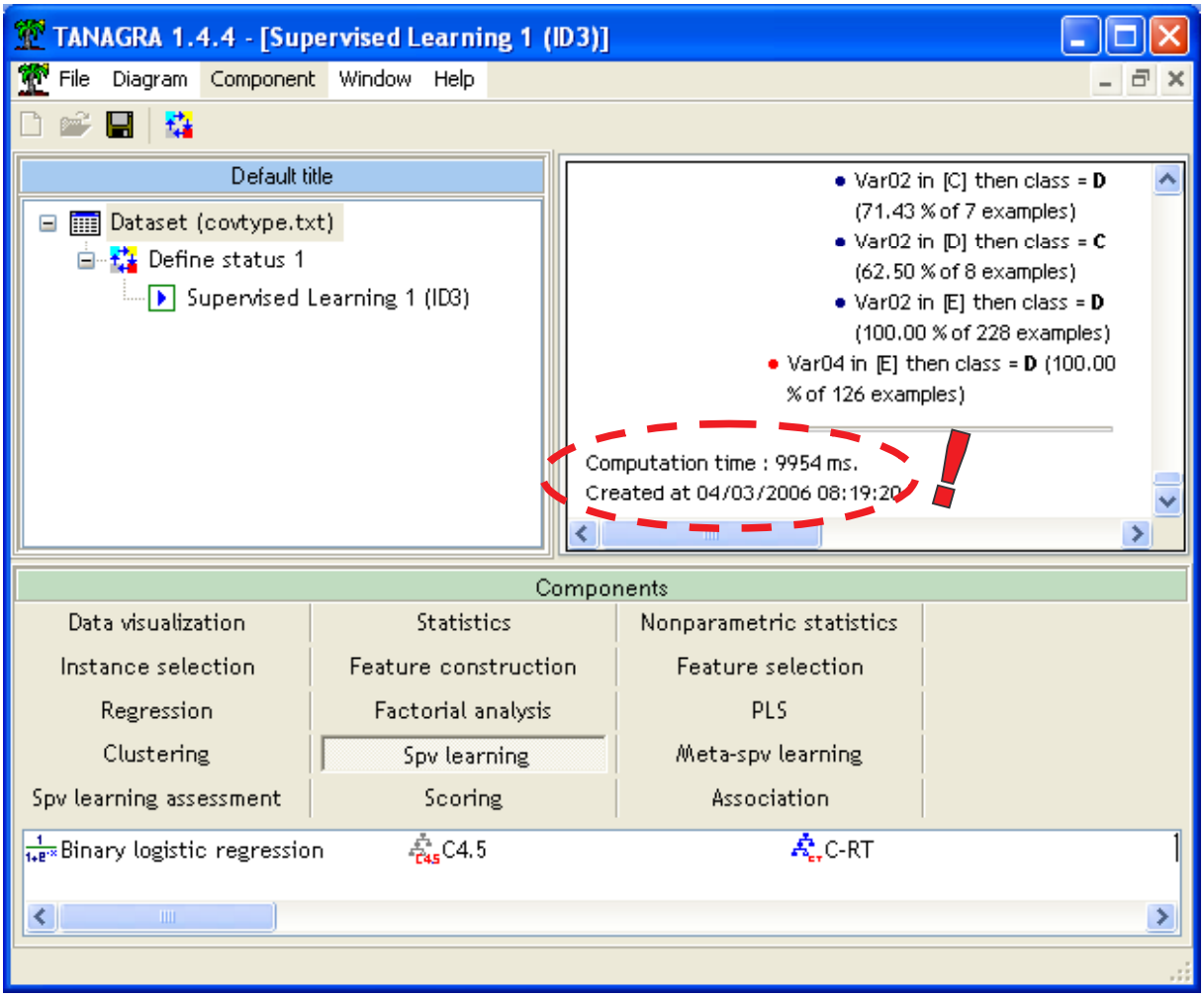
Data description

Target attribute	class (7 values)
# descriptors	54

Tree description

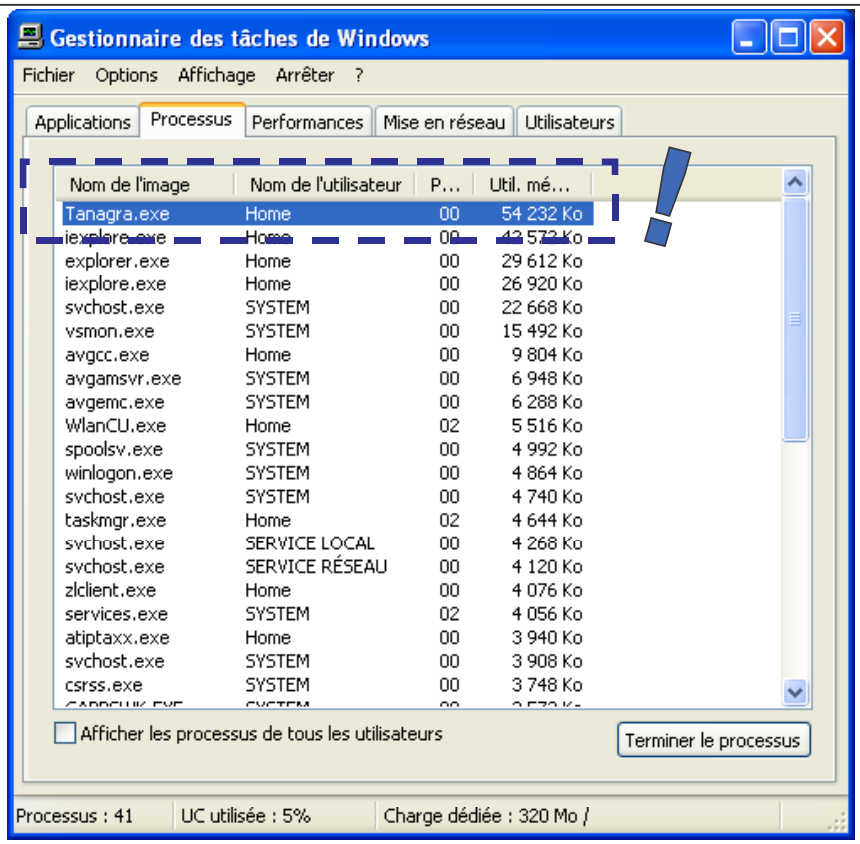
Number of nodes	1217
Number of leaves	927

L'arbre comporte 927 feuilles, ce qui est assez considérable. Lorsqu'un arbre atteint une telle taille, il est illusoire de vouloir interpréter les résultats. Plus intéressant dans le cadre de ce didacticiel est le temps de calcul. Nous déroulons l'affichage de l'arbre jusqu'à atteindre la partie basse de la fenêtre. Nous observons alors les informations suivantes[1].



Il a fallu 10 secondes pour construire un arbre de décision à partir d'un fichier de 581102 observations et 54 descripteurs. De plus nous sommes loin de saturer les 1 GB de mémoire vive, dans le gestionnaire de tâche de WINDOWS, nous observons que TANAGRA alloue réellement 54 MB. Ce qui reste très raisonnable compte tenu de la taille du fichier que nous avons traité.

[1] TANAGRA génère automatiquement un fichier DEBUGFILE.TXT qui trace toutes les opérations réalisées lors du traitement des données et les durées d'exécution associées. Il est possible de le consulter.



Conclusion

Charger les données en mémoire pour effectuer les traitements est une stratégie simple – simpliste -- qui permet d'atteindre de bonnes performances en temps de traitement : construire un arbre de décision en 10 secondes sur un fichier de 581102 observations et 54 descripteurs est un résultat intéressant.

Nous constatons que cette stratégie, a priori très pénalisante, s'avère souvent réaliste compte tenu des caractéristiques des machines actuelles. L'occupation mémoire reste contenue sur de gros ensembles de données pour peu que l'on choisisse les bonnes structures de programmation.