Objectif

Comment appliquer attribuer une classe à de nouveaux individus en appliquant un modèle de prédiction sur un fichier de données externe ?

Cette procédure dépasse largement le cadre dans lequel a été défini TANAGRA qui est, rappelons-le, l'évaluation et la comparaison des méthodes de fouille de données. Il rend opérationnel le logiciel dans le déploiement effectif des modèles construits, et par-là, ouvre la porte à une utilisation professionnelle. Mais face à des demandes répétées d'utilisateurs d'horizons très différents, nous proposons dans ce didacticiel une démarche, assez tarabiscotée, reconnaissons-le, qui permet de mettre en œuvre le logiciel dans ce contexte.

Fichier

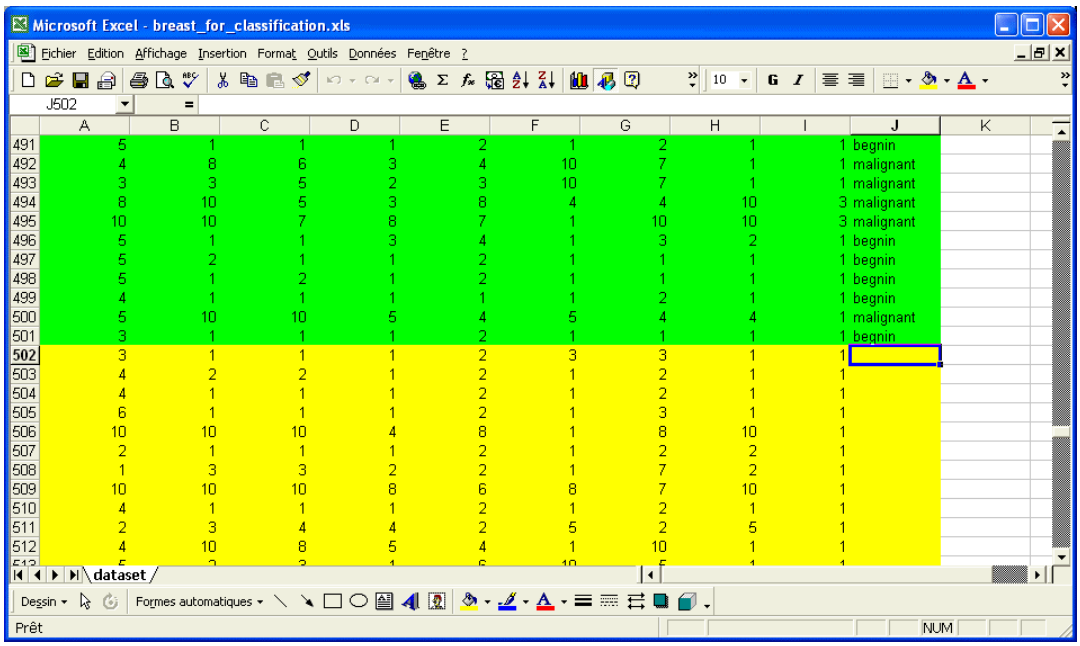
La préparation du fichier est une étape primordiale dans cette procédure. En effet, par construction, il est impossible de manipuler simultanément deux fichiers différents dans un diagramme de traitements TANAGRA. De fait, appliquer un modèle de prédiction sur un fichier externe n'est donc théoriquement pas possible. L'astuce réside dans une préparation adéquate du fichier de données.

Dans ce didacticiel, nous utiliserons le fichier BREAST CANCER WISCONSIN que nous avons subdivisé en 2 parties : 500 individus pour l'apprentissage, 199 individus pour le classement. Nous ne sommes donc pas censés connaître la classe d'appartenance de ces 199 d'individus, il revient au modèle de prédiction construit avec TANAGRA de les classer.

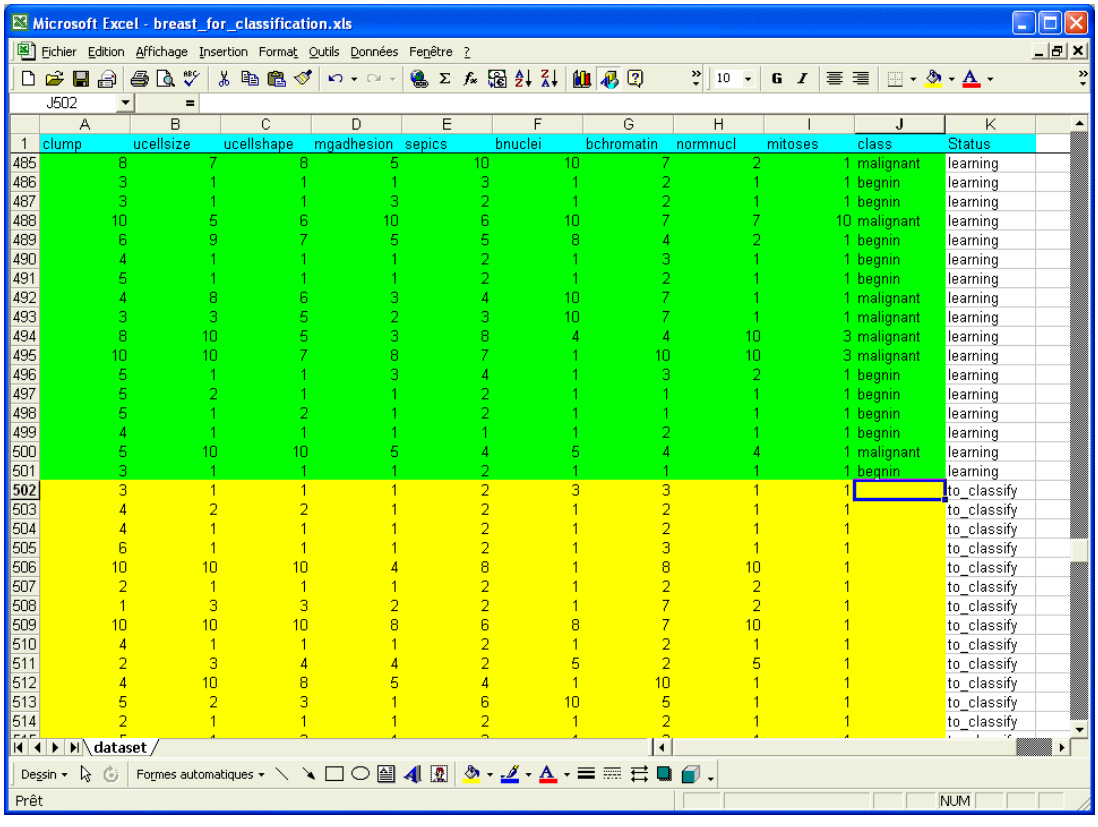
Pour mémoire, le fichier BREAST retranscrit des données issues d'analyses cytologiques : l'objectif est de reconnaître le caractère malin ou bénin d'une tumeur à partir de l'analyse des cellules ponctionnées. Ce fichier est très utilisé par la communauté de l'apprentissage automatique.

La construction du fichier est réalisée en plusieurs étapes :

(1) réunir les données dans un même et seul fichier, en utilisant EXCEL par exemple, placer les individus étiquetés en premier, les individus à classer à la suite.

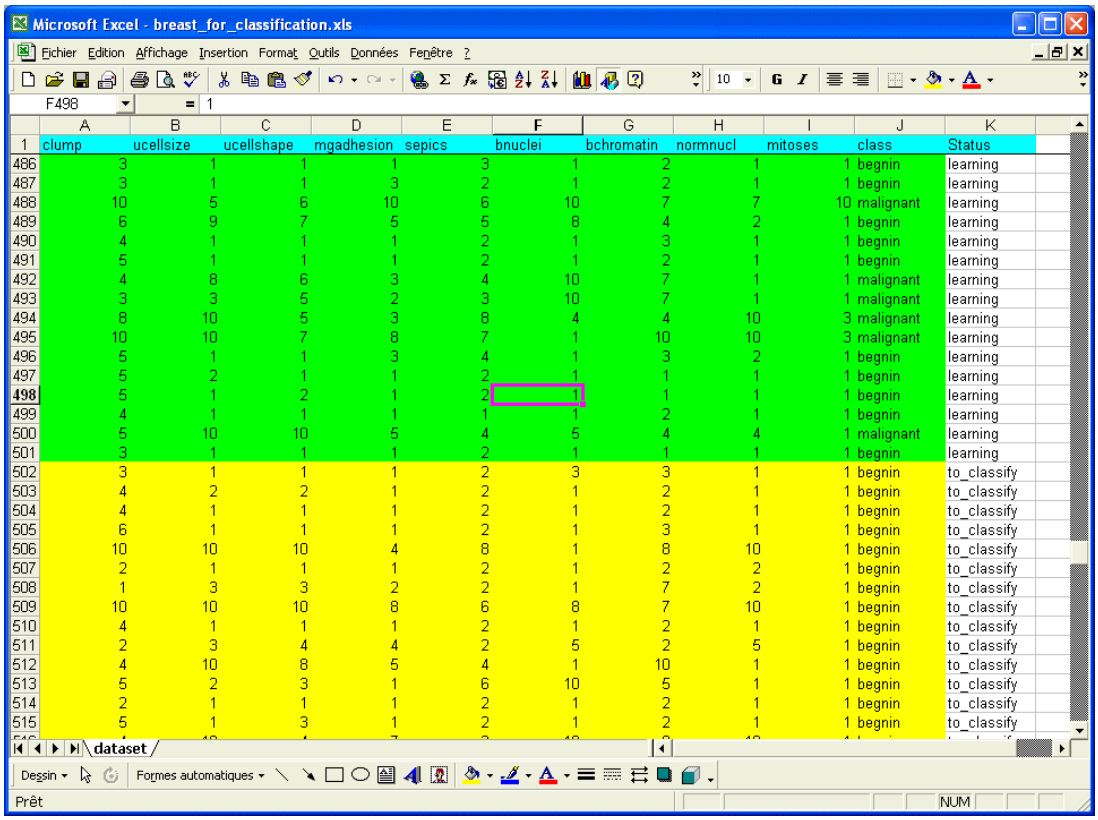


(2) définir alors une nouvelle variable « STATUS » qui permet de spécifier le rôle que joueront les individus dans le processus de modélisation, cette variable prend deux valeurs possibles « Learning » et « To_Classify ».

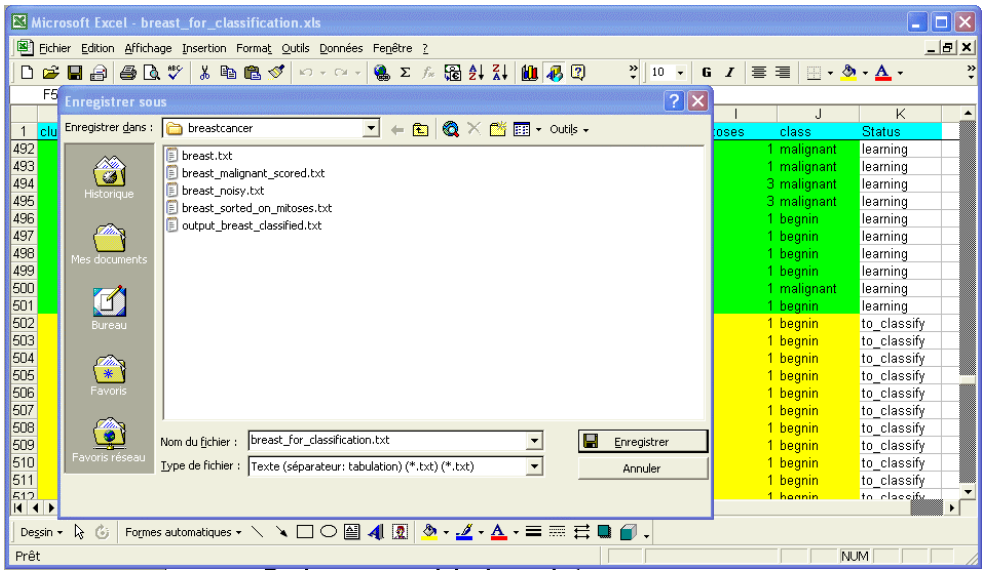


(3) enfin, et bien que cela soit tout à fait contre-intuitif, il est nécessaire d'attribuer au préalable une classe aux individus à classer. N'importe quelle valeur fera l'affaire, dans l'ensemble des valeurs

possibles bien entendu. Cela est nécessaire car TANAGRA ne gère pas les données manquantes, l'importation directe du fichier sous cette forme provoquerait une erreur. Par convention, on peut choisir d'attribuer arbitrairement la première valeur de la variable à prédire pour les individus à classer, de toute manière, cette information ne sera jamais utilisée par la suite. Dans notre cas, nous attribuons la valeur « begnin » aux 199 individus à classer.



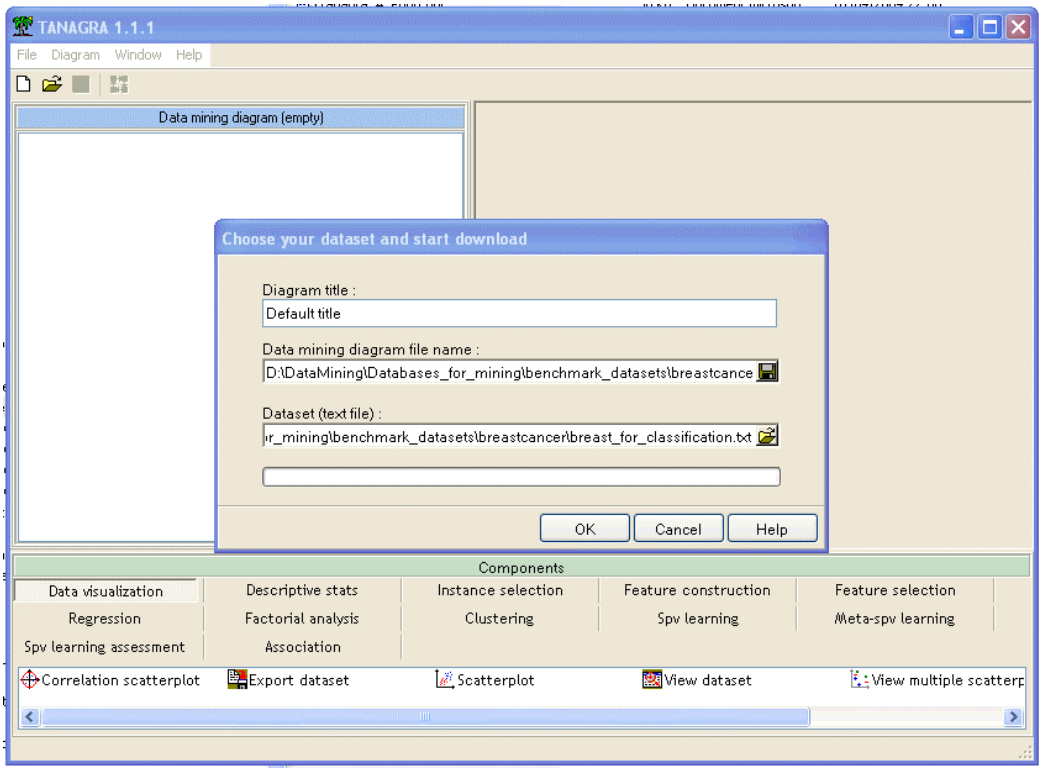
(4) reste alors à exporter le fichier au format texte avec séparateur tabulation reconnu par TANAGRA.



Déployer un modèle de prédiction

Importer les données dans TANAGRA

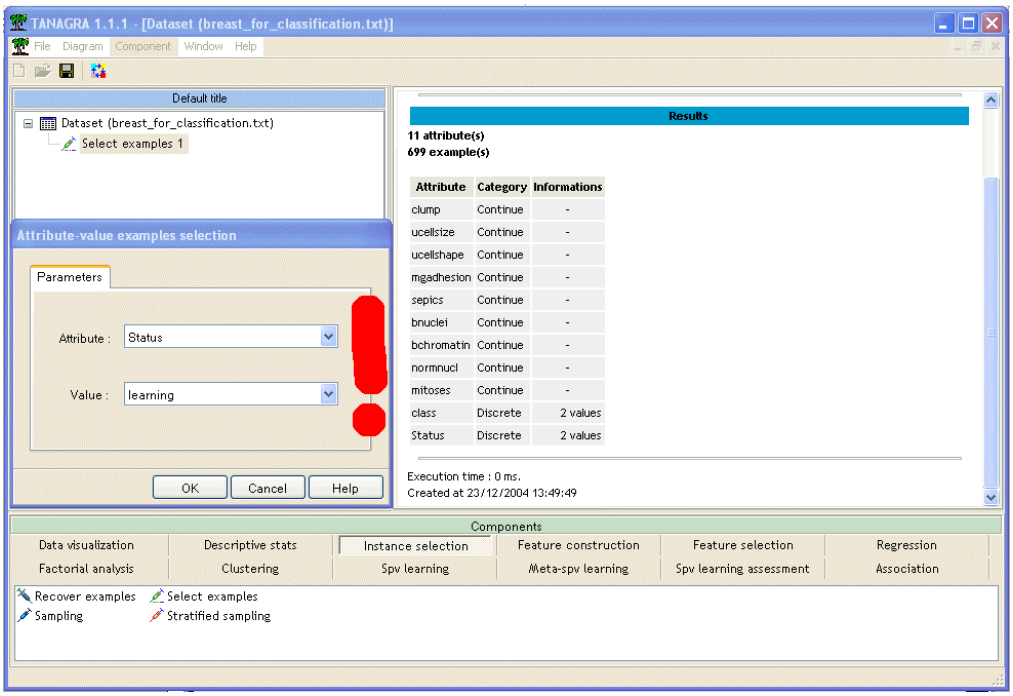
La première étape est tout à fait classique, il s'agit d'importer les données et de définir un diagramme de traitements.



Spécifier les observations à utiliser pour l'apprentissage

Sélectionner par la suite les individus que nous utiliserons pour la construction du modèle de prédiction. Le composant SELECT EXAMPLES est tout à fait indiqué pour cela, nous utiliserons alors la variable STATUS définie précédemment.

La base est maintenant subdivisée en interne : 500 individus « actifs » à partir desquels nous construirons le modèle de prédiction, 199 individus « supplémentaires » sur lesquels nous effectuerons des projections.

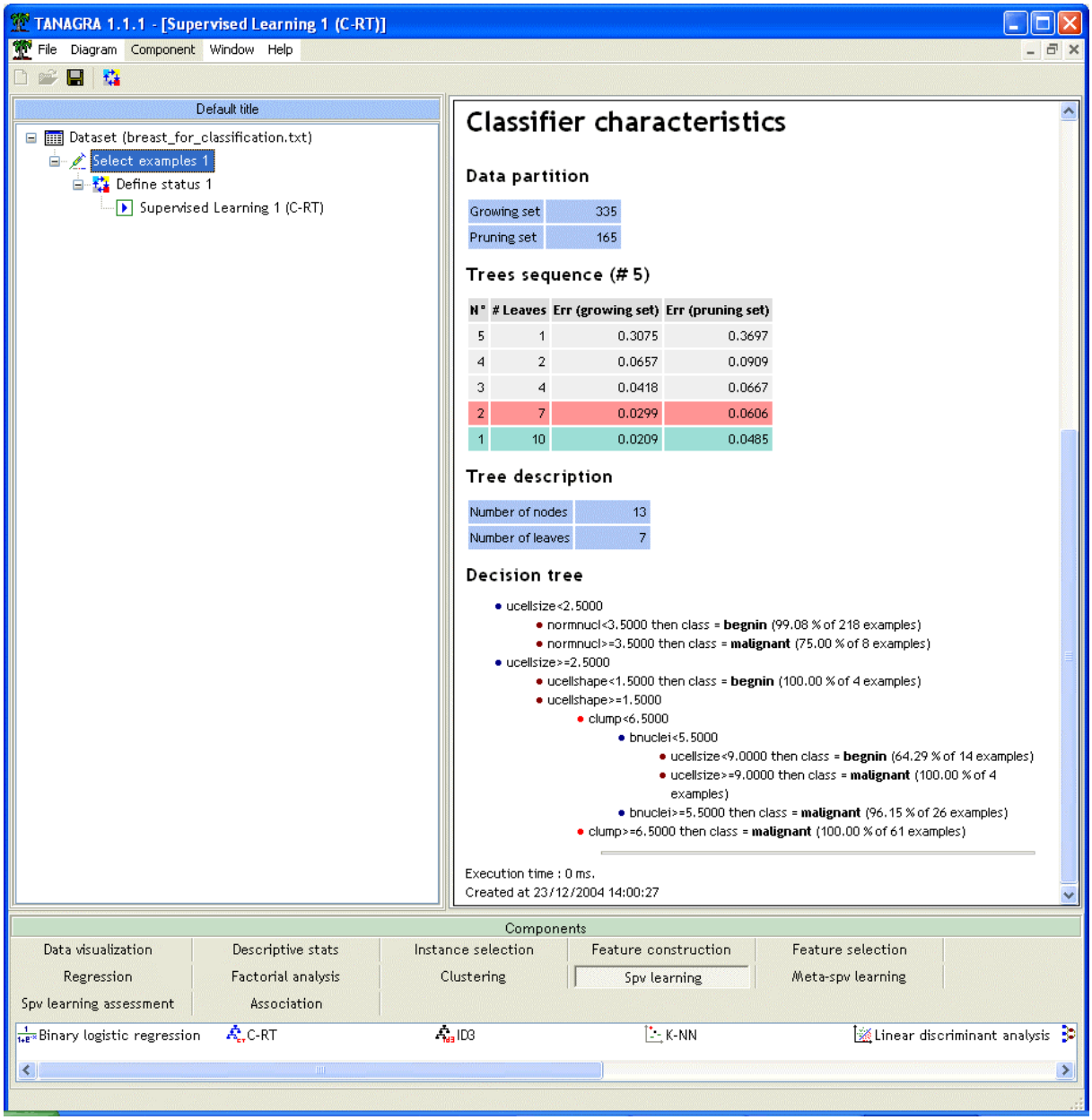


Choisir les variables de l'étude – Algorithme d'apprentissage

La suite reproduit les schémas standards de l'apprentissage supervisé : choix des variables et de l'algorithme d'apprentissage. Dans notre exemple, nous utiliserons la méthode d'induction d'arbre de décision C-RT (Breiman et al., 1984).

La variable à prédire est la variable « CLASS », les 9 premières variables de la base représentent les descripteurs.

Le diagramme de traitements et les résultats associés sont reproduits dans la copie d 'écran ci-dessous.

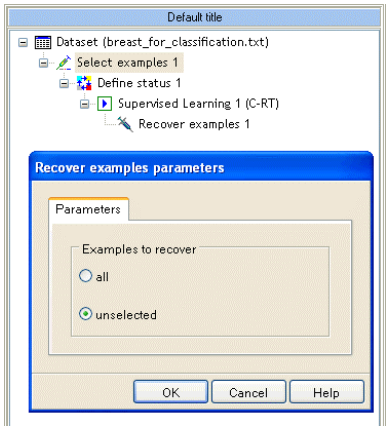


Comme pour toutes les méthodes d'apprentissage supervisé, un attribut supplémentaire a été rajouté à l'ensemble de données : la projection produite par l'algorithme d'apprentissage.

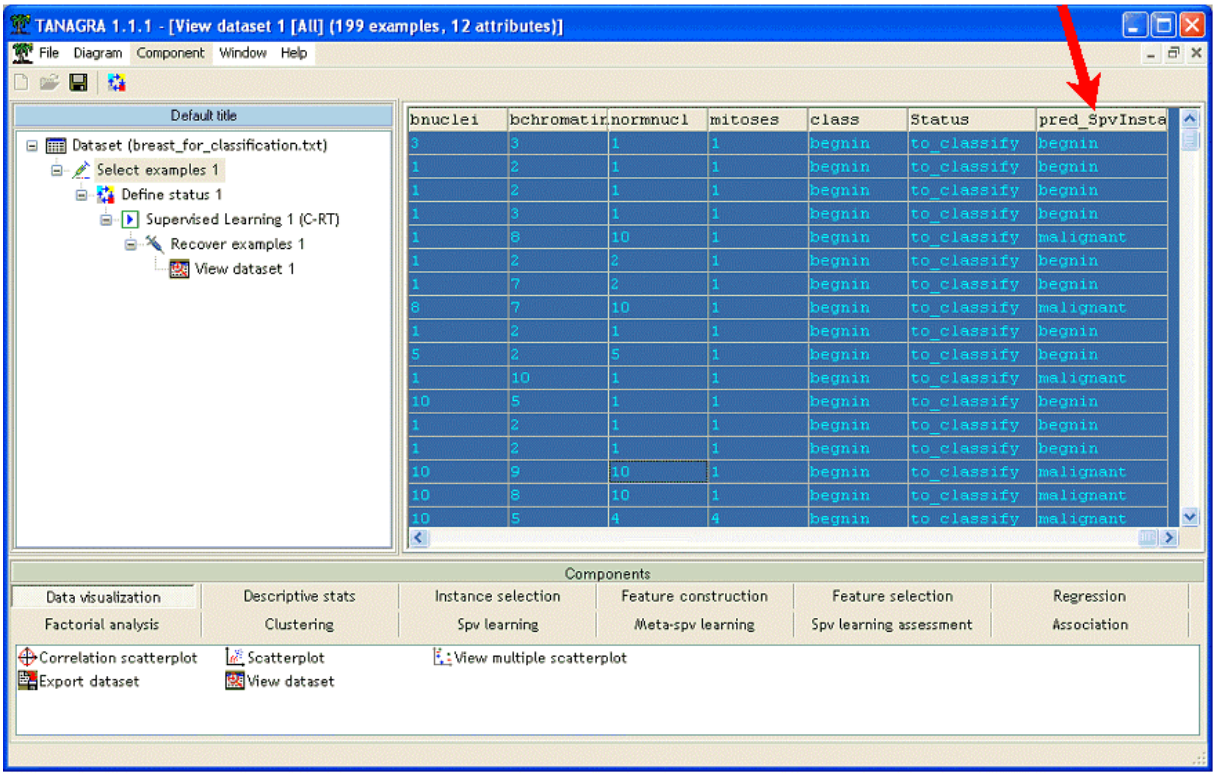
Il est très important de noter à ce stade que si l'apprentissage a bien été accompli sur les individus actifs sélectionnés par le composant « SELECT EXAMPLES », la projection est réalisée sur l'ensemble des observations, y compris donc les individus supplémentaires que nous avons mis de côté. C'est cette propriété que nous exploitons pour le déploiement des modèles sur de nouvelles bases.

Visualisation des projections

Pour s'en persuader, nous allons visualiser les projections sur ces individus supplémentaires. Pour ce faire nous allons inverser la sélection càd rendre actifs les individus précédemment illustratifs, et inversement, en utilisant le composant RECOVER EXAMPLES.

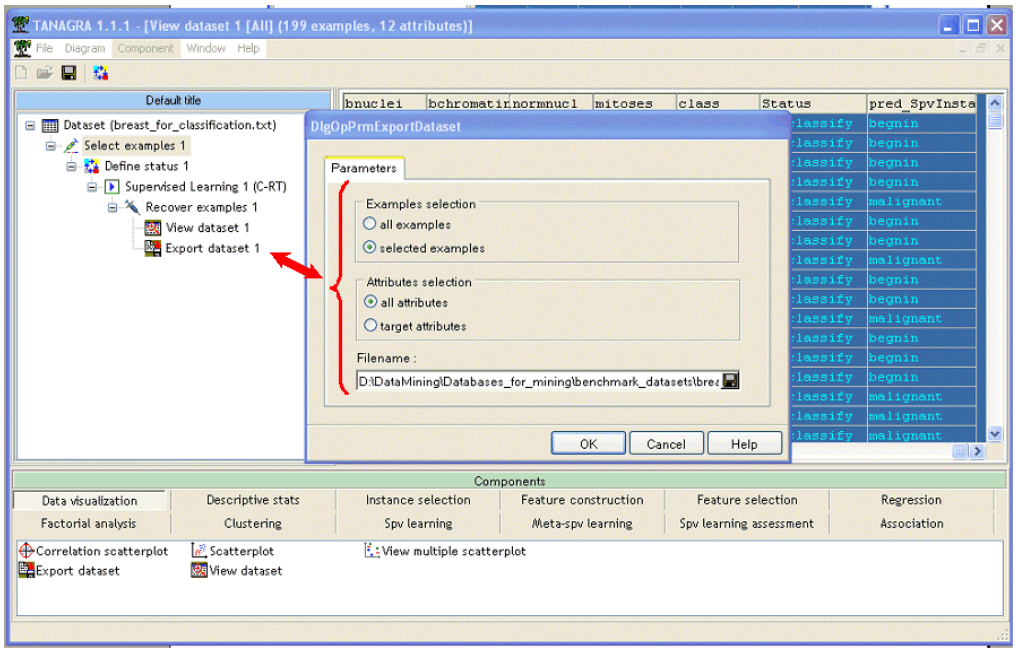


Reste alors à visualiser les données dans une grille. La nouvelle colonne, la projection, a été rajoutée en dernière position dans l'ensemble courant des données.

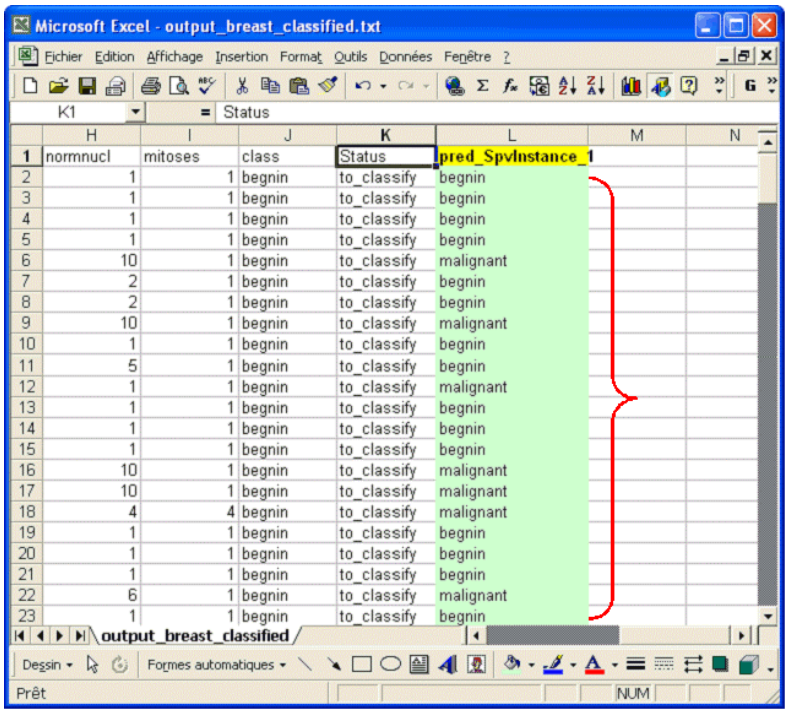


Exporter les résultats

La dernière étape consiste à exporter les données à l'aide du composant EXPORT DATASET. Vous avez la possibilité de restreindre l'exportation sur un sous-ensemble des observations, de la même manière, vous pouvez également restreindre l'ensemble des variables à exporter. Dans notre exemple, nous exportons toutes les variables sur les observations à classer.



Les données exportées peuvent être visualisées dans EXCEL.



	H	I	J	K	L	M	N
1	normnucl	mitoses	class	Status	pred_SpvInstance_1		
2	1	1	begnin	to_classify	begnin		
3	1	1	begnin	to_classify	begnin		
4	1	1	begnin	to_classify	begnin		
5	1	1	begnin	to_classify	begnin		
6	10	1	begnin	to_classify	malignant		
7	2	1	begnin	to_classify	begnin		
8	2	1	begnin	to_classify	begnin		
9	10	1	begnin	to_classify	malignant		
10	1	1	begnin	to_classify	begnin		
11	5	1	begnin	to_classify	begnin		
12	1	1	begnin	to_classify	malignant		
13	1	1	begnin	to_classify	begnin		
14	1	1	begnin	to_classify	begnin		
15	1	1	begnin	to_classify	begnin		
16	10	1	begnin	to_classify	malignant		
17	10	1	begnin	to_classify	malignant		
18	4	4	begnin	to_classify	malignant		
19	1	1	begnin	to_classify	begnin		
20	1	1	begnin	to_classify	begnin		
21	1	1	begnin	to_classify	begnin		
22	6	1	begnin	to_classify	malignant		
23	1	1	begnin	to_classify	begnin		

Conclusion

Nous pouvons suivre la même démarche lorsque nous désirons utiliser un fichier externe pour la validation d'un algorithme d'apprentissage. Pour obtenir la matrice de confusion en test, il faut construire, sur les individus supplémentaires, un tableau de contingence qui croise la variable à prédire observée et la projection.

Je reconnais que la procédure est compliquée, mais c'est la démarche la plus simple pour appliquer une procédure d'apprentissage sur un nouvel ensemble d'individus.

Dans un premier temps j'avais été tenté de construire simplement un nouveau composant « PROJECTION » qui aurait permis d'appliquer le modèle de prédiction sur un nouvel ensemble de données. Très séduisante sur le papier, je me suis rapidement rendu compte que cette solution impliquait une programmation très lourde. En effet, très souvent, l'algorithme d'apprentissage arrive en bout de chaîne, après une série de manipulations sur les données (sélection de variables, transformations de variables, projections, etc.). Reproduire ces séquences de calculs sur des données qui ne transitent pas dans le diagramme de traitements demande une refonte complète des structures de données actuelles. Ce sera pour une prochaine fois peut être…