

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.

	A	B	C	D	E	F	G	H	I	J	K
491	5	1	1	1	2	1	2	1	1	1 benign	
492	4	8	6	3	4	10	7	1	1	1 malignant	
493	3	3	5	2	3	10	7	1	1	1 malignant	
494	8	10	5	3	8	4	4	10	3	3 malignant	
495	10	10	7	8	7	1	10	10	3	3 malignant	
496	5	1	1	3	4	1	3	2	1	1 benign	
497	5	2	1	1	2	1	1	1	1	1 benign	
498	5	1	2	1	2	1	1	1	1	1 benign	
499	4	1	1	1	1	1	2	1	1	1 benign	
500	5	10	10	5	4	5	4	4	1	1 malignant	
501	3	1	1	1	2	1	1	1	1	1 benign	
502	3	1	1	1	2	3	3	1	1	1 malignant	
503	4	2	2	1	2	1	2	1	1	1 malignant	
504	4	1	1	1	2	1	2	1	1	1 malignant	
505	6	1	1	1	2	1	3	1	1	1 malignant	
506	10	10	10	4	8	1	8	10	1	1 malignant	
507	2	1	1	1	2	1	2	2	1	1 malignant	
508	1	3	3	2	2	1	7	2	1	1 malignant	
509	10	10	10	8	6	8	7	10	1	1 malignant	
510	4	1	1	1	2	1	2	1	1	1 malignant	
511	2	3	4	4	2	5	2	5	1	1 malignant	
512	4	10	8	5	4	1	10	1	1	1 malignant	
513	5	2	3	1	6	10	5	1	1	1 malignant	
514	2	1	1	1	2	1	2	1	1	1 malignant	

(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 ».

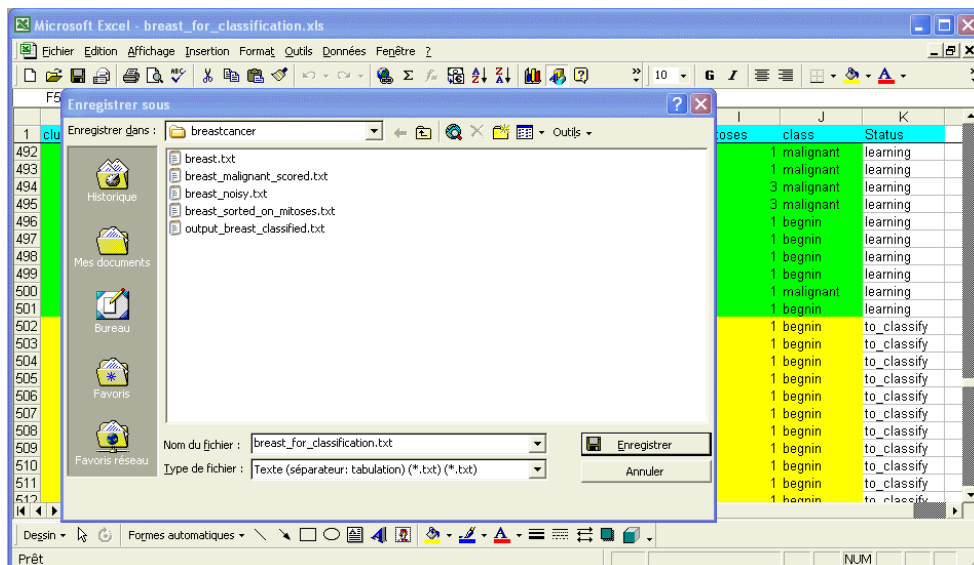
	A	B	C	D	E	F	G	H	I	J	K
1	clump	ucellsize	ucellshape	mgadhesion	sepics	bnclei	bchromatin	normnucl	mitoses	class	Status
485	8	7	8	5	10	10	7	2	1	1 malignant	learning
486	3	1	1	1	3	1	2	1	1	1 benign	learning
487	3	1	1	3	2	1	2	1	1	1 benign	learning
488	10	5	6	10	6	10	7	7	10	10 malignant	learning
489	6	9	7	5	5	8	4	2	1	1 benign	learning
490	4	1	1	1	2	1	3	1	1	1 benign	learning
491	5	1	1	1	2	1	2	1	1	1 benign	learning
492	4	8	6	3	4	10	7	1	1	1 malignant	learning
493	3	3	5	2	3	10	7	1	1	1 malignant	learning
494	8	10	5	3	8	4	4	10	3	3 malignant	learning
495	10	10	7	8	7	1	10	10	3	3 malignant	learning
496	5	1	1	3	4	1	3	2	1	1 benign	learning
497	5	2	1	1	2	1	1	1	1	1 benign	learning
498	5	1	2	1	2	1	1	1	1	1 benign	learning
499	4	1	1	1	1	1	2	1	1	1 benign	learning
500	5	10	10	5	4	5	4	4	1	1 malignant	learning
501	3	1	1	1	2	1	1	1	1	1 benign	learning
502	3	1	1	1	2	3	3	1	1	1 malignant	to_classify
503	4	2	2	1	2	1	2	1	1	1 malignant	to_classify
504	4	1	1	1	2	1	2	1	1	1 malignant	to_classify
505	6	1	1	1	2	1	3	1	1	1 malignant	to_classify
506	10	10	10	4	8	1	8	10	1	1 malignant	to_classify
507	2	1	1	1	2	1	2	2	1	1 malignant	to_classify
508	1	3	3	2	2	1	7	2	1	1 malignant	to_classify
509	10	10	10	8	6	8	7	10	1	1 malignant	to_classify
510	4	1	1	1	2	1	2	1	1	1 malignant	to_classify
511	2	3	4	4	2	5	2	5	1	1 malignant	to_classify
512	4	10	8	5	4	1	10	1	1	1 malignant	to_classify
513	5	2	3	1	6	10	5	1	1	1 malignant	to_classify
514	2	1	1	1	2	1	2	1	1	1 malignant	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 « beginn » aux 199 individus à classer.

	A	B	C	D	E	F	G	H	I	J	K
1	clump	ucellsize	ucellshape	mgadhesion	sepics	bnuclei	bchromatin	normnucl	mitoses	class	Status
486	3	1	1	1	3	1	2	1	1	1 beginn	learning
487	3	1	1	3	2	1	2	1	1	1 beginn	learning
488	10	5	6	10	6	10	7	7	10	1 malignant	learning
489	6	9	7	5	5	8	4	2	1	1 beginn	learning
490	4	1	1	1	2	1	3	1	1	1 beginn	learning
491	5	1	1	1	2	1	2	1	1	1 beginn	learning
492	4	8	6	3	4	10	7	1	1	1 malignant	learning
493	3	3	5	2	3	10	7	1	1	1 malignant	learning
494	8	10	5	3	8	4	4	10	3	1 malignant	learning
495	10	10	7	8	7	1	10	10	3	1 malignant	learning
496	5	1	1	3	4	1	3	2	1	1 beginn	learning
497	5	2	1	1	2	1	1	1	1	1 beginn	learning
498	5	1	2	1	2	1	1	1	1	1 beginn	learning
499	4	1	1	1	1	1	2	1	1	1 beginn	learning
500	5	10	10	5	4	5	4	4	1	1 malignant	learning
501	3	1	1	1	2	1	1	1	1	1 beginn	learning
502	3	1	1	1	2	3	3	1	1	1 beginn	to_classify
503	4	2	2	1	2	1	2	1	1	1 beginn	to_classify
504	4	1	1	1	2	1	2	1	1	1 beginn	to_classify
505	6	1	1	1	2	1	3	1	1	1 beginn	to_classify
506	10	10	10	4	8	1	8	10	1	1 beginn	to_classify
507	2	1	1	1	2	1	2	2	1	1 beginn	to_classify
508	1	3	3	2	2	1	7	2	1	1 beginn	to_classify
509	10	10	10	8	6	8	7	10	1	1 beginn	to_classify
510	4	1	1	1	2	1	2	1	1	1 beginn	to_classify
511	2	3	4	4	2	5	2	5	1	1 beginn	to_classify
512	4	10	8	5	4	1	10	1	1	1 beginn	to_classify
513	5	2	3	1	6	10	5	1	1	1 beginn	to_classify
514	2	1	1	1	2	1	2	1	1	1 beginn	to_classify
515	5	1	3	1	2	1	2	1	1	1 beginn	to_classify

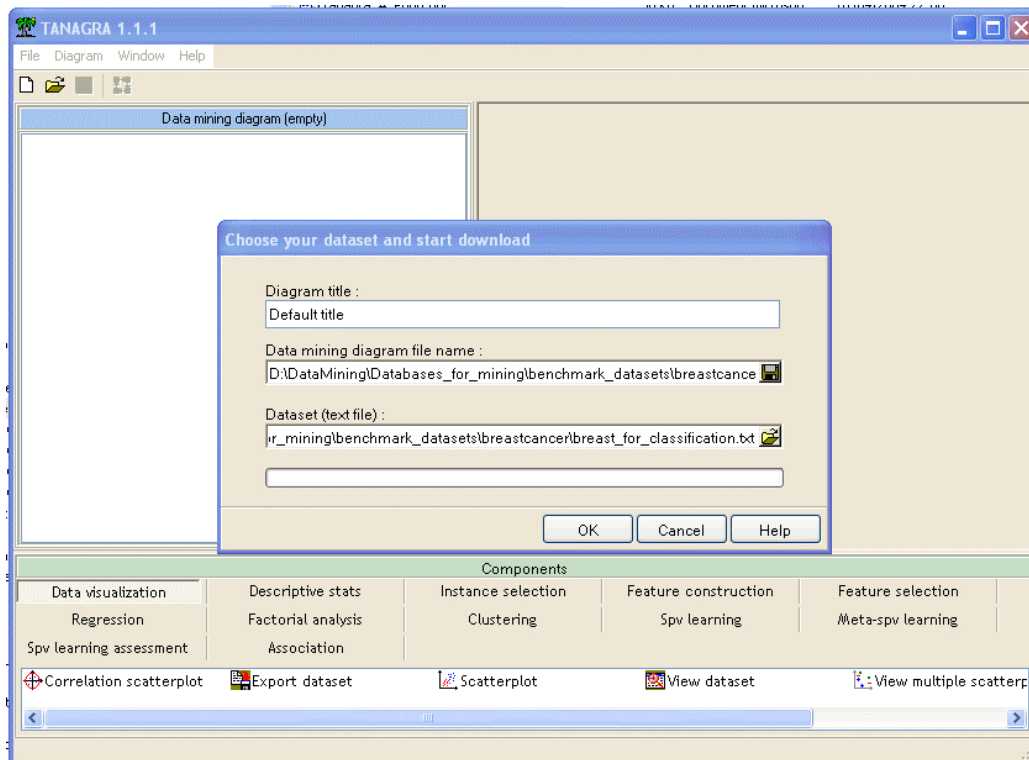
(4) reste alors à exporter le fichier au format texte avec séparateur tabulation reconnu par TANAGRA.



Déployer un modèle de prédiction

Importer les données dans TANAGRA

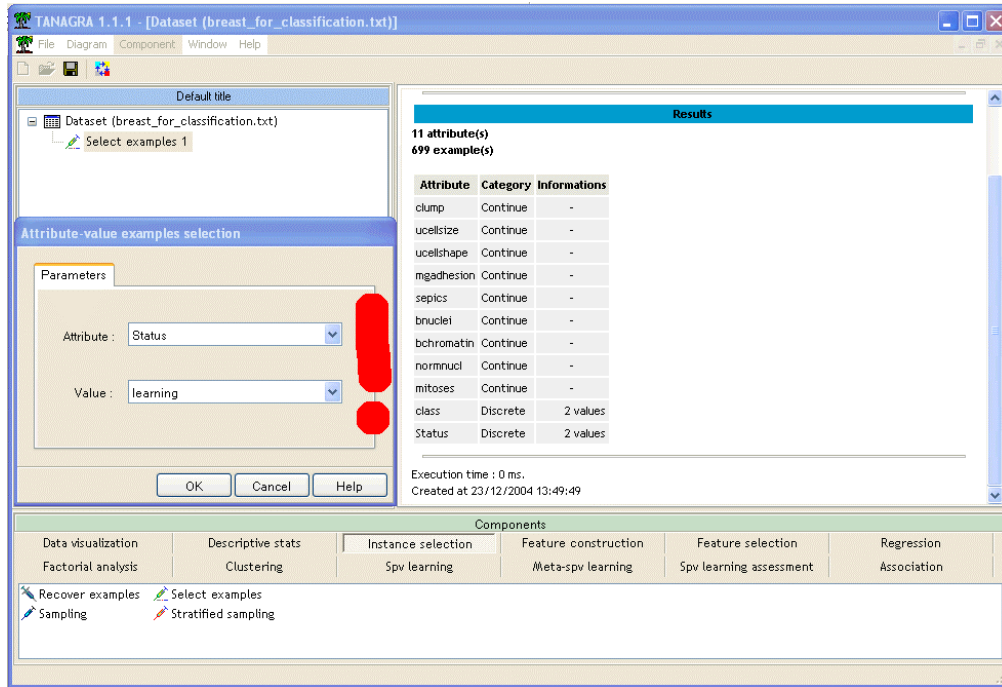
La première étape est tout à fait classique, il s'agit d'importer les données et de définir un diagramme de traitements.



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.

The screenshot shows the TANAGRA 1.1.1 interface with the 'Supervised Learning 1 (C-RT)' component selected. The 'Classifier characteristics' window displays the following data:

Classifier characteristics

Data partition

Growing set	335
Pruning set	165

Trees sequence (# 5)

N°	# Leaves	Err (growing set)	Err (pruning set)
5	1	0.3075	0.3697
4	2	0.0657	0.0909
3	4	0.0418	0.0667
2	7	0.0299	0.0606
1	10	0.0209	0.0485

Tree description

Number of nodes	13
Number of leaves	7

Decision tree

- ucellsize < 2,5000
 - normnucl < 3,5000 then class = **benign** (99.08 % of 218 examples)
 - normnucl >= 3,5000 then class = **malignant** (75.00 % of 8 examples)
- ucellsize >= 2,5000
 - ucellshape < 1,5000 then class = **benign** (100.00 % of 4 examples)
 - ucellshape >= 1,5000
 - clump < 6,5000
 - bnuclei < 5,5000
 - ucellsize < 9,0000 then class = **benign** (64.29 % of 14 examples)
 - ucellsize >= 9,0000 then class = **malignant** (100.00 % of 4 examples)
 - bnuclei >= 5,5000 then class = **malignant** (96.15 % of 26 examples)
 - clump >= 6,5000 then class = **malignant** (100.00 % of 61 examples)

Execution time : 0 ms.
Created at 23/12/2004 14:00:27

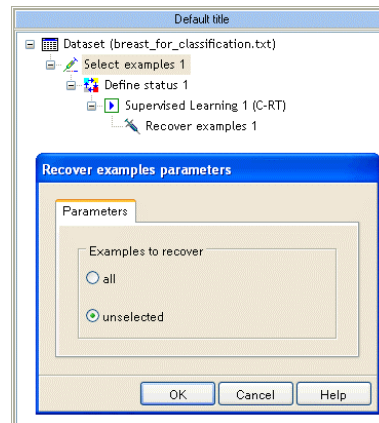
The bottom of the interface shows a 'Components' bar with various machine learning methods, and a taskbar at the very bottom with icons for 'Binary logistic regression', 'C-RT', 'ID3', 'K-NN', and 'Linear discriminant analysis'.

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 c-à-d rendre actifs les individus précédemment illustratifs, et inversement, en utilisant le composant RECOVER EXAMPLES.



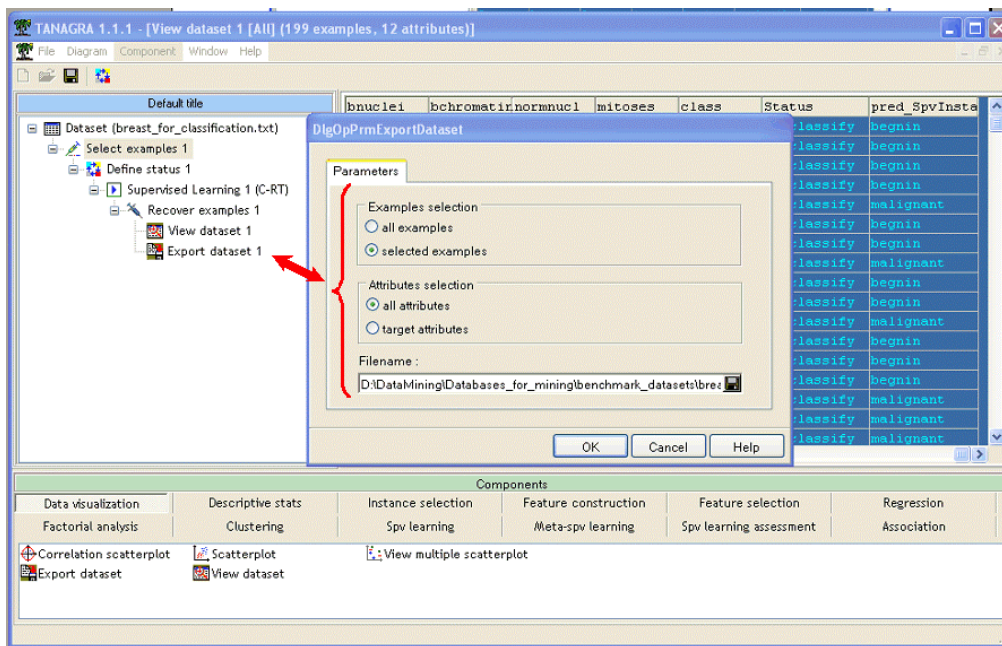
Reste alors à visualiser les données dans une grille. La nouvelle colonne, la projection, a été rajoutée en dernière position dans l'ensemble courant des données.

bnuclei	bchromatin	normnuclei	mitoses	class	Status	pred_SpvInsta
3	3	1	1	begin	to_classify	begin
1	2	1	1	begin	to_classify	begin
1	2	1	1	begin	to_classify	begin
1	3	1	1	begin	to_classify	begin
1	8	10	1	begin	to_classify	malignant
1	2	2	1	begin	to_classify	begin
1	7	2	1	begin	to_classify	begin
8	7	10	1	begin	to_classify	malignant
1	2	1	1	begin	to_classify	begin
5	2	5	1	begin	to_classify	begin
1	10	1	1	begin	to_classify	malignant
10	5	1	1	begin	to_classify	begin
1	2	1	1	begin	to_classify	begin
1	2	1	1	begin	to_classify	begin
10	9	10	1	begin	to_classify	malignant
10	8	10	1	begin	to_classify	malignant
10	5	4	4	begin	to_classify	malignant

The interface also shows a 'Components' panel at the bottom with various analysis options like 'Data visualization', 'Descriptive stats', 'Instance selection', etc.

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.

	normnucl	mitoses	class	Status	pred_SpvInstance 1
1	1	1	1	to_classify	begin
2	1	1	1	to_classify	begin
3	1	1	1	to_classify	begin
4	1	1	1	to_classify	begin
5	1	1	1	to_classify	begin
6	10	1	1	to_classify	malignant
7	1	2	1	to_classify	begin
8	2	1	1	to_classify	begin
9	10	1	1	to_classify	malignant
10	1	1	1	to_classify	begin
11	5	1	1	to_classify	begin
12	1	1	1	to_classify	malignant
13	1	1	1	to_classify	begin
14	1	1	1	to_classify	begin
15	1	1	1	to_classify	begin
16	10	1	1	to_classify	malignant
17	10	1	1	to_classify	malignant
18	4	4	1	to_classify	malignant
19	1	1	1	to_classify	begin
20	1	1	1	to_classify	begin
21	1	1	1	to_classify	begin
22	6	1	1	to_classify	malignant
23	1	1	1	to_classify	begin

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...