Objectif

Avec la version **1.4.8,** il est possible de sauver une partie du diagramme de traitements. L'idée est de pouvoir reproduire des séquences entières d'analyses sur des bases de données différentes sans avoir à redéfinir manuellement le diagramme.

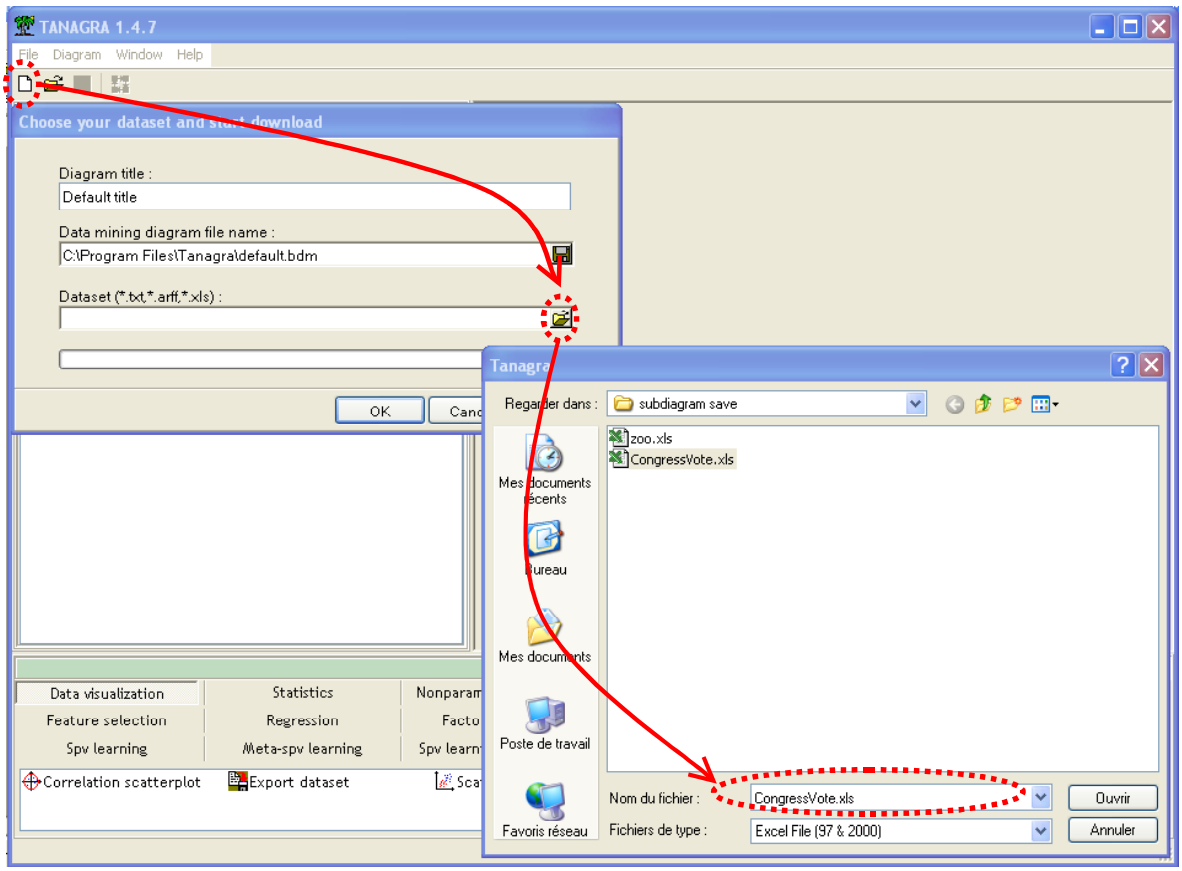
Fichier

Nous utilisons deux fichiers : CONGRESSVOTE.XLS et ZOO.XLS. Nous voulons prédire les valeurs de l'attribut classe à partir de descripteurs discrets, avec ou sans sélection de variables. Les performances sont évaluées à l'aide de la validation croisée. Les traitements définis sur le premier fichier de données seront reproduits sur le second fichier avec un minimum de manipulations.

Sauver des séquences de traitements

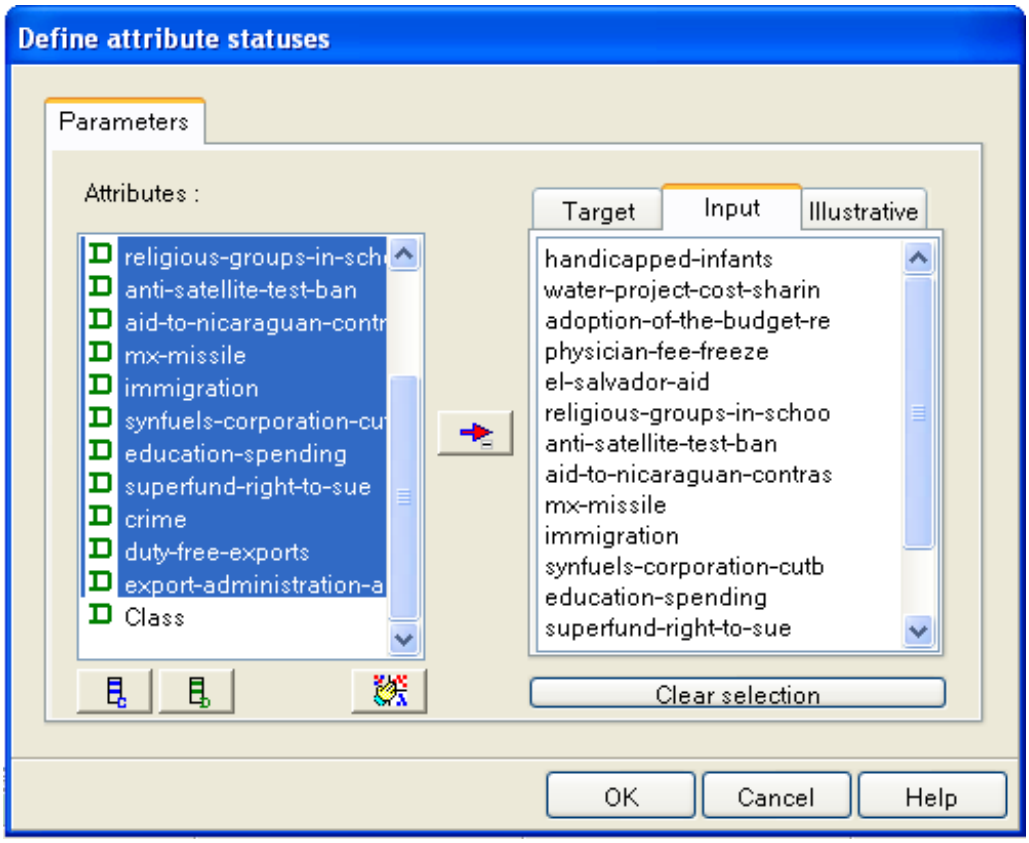
Importer les données

Première étape, créer un diagramme et importer les données : nous cliquons sur le menu FILE/NEW et sélectionnons le fichier CONGRESSVOTE.XLS.



Définir les variables de l'analyse

L'étape suivante consiste à définir l'attribut à prédire (TARGET : CLASS) et les descripteurs (INPUT : Tous les autres). Nous ajoutons le composant DEFINE STATUS en utilisant le raccourci dans la barre d'outils.



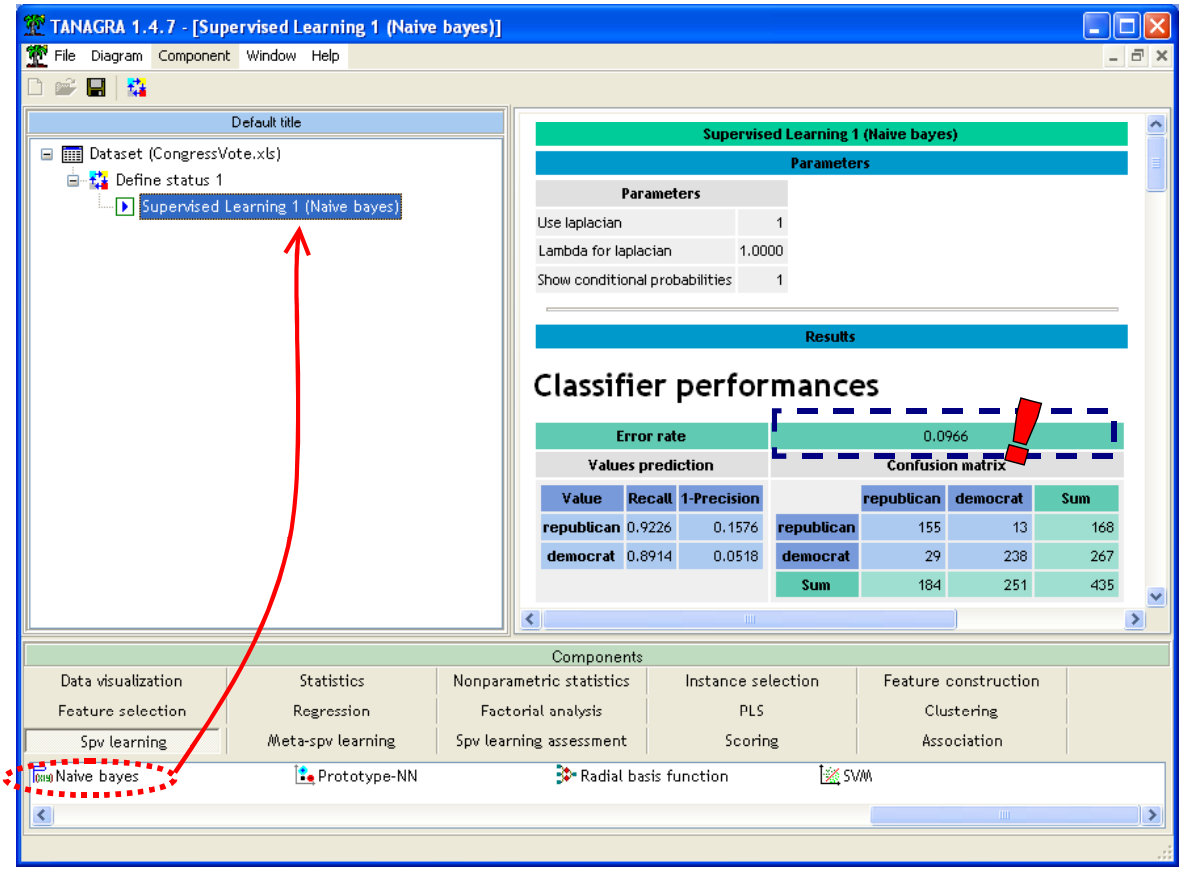
Placer la méthode d'apprentissage et son évaluation

Jusqu'à maintenant (version 1.0.0 → 1.4.6), l'adjonction d'une méthode supervisée dans le diagramme se faisait en deux étapes : d'abord placer le composant qui instancie la méthode (META SPV LEARNING), puis y insérer l'algorithme d'apprentissage (SPV LEARNING).

Dorénavant (version 1.4.7), dans le cadre de l'apprentissage simple, avec une seule instance de l'algorithme, il sera possible de placer directement le composant de la palette apprentissage supervisé dans le diagramme, TANAGRA effectue automatiquement le nécessaire pour que le composant META SPV soit correctement ajouté.

NOTE : Attention, si l'on veut mettre en oeuvre une procédure d'agrégation particulière, le « bagging » par exemple, l'insertion en deux temps est toujours requise : placer d'abord le meta-apprentissage avant d'y intégrer le composant représentant l'algorithme d'apprentissage.

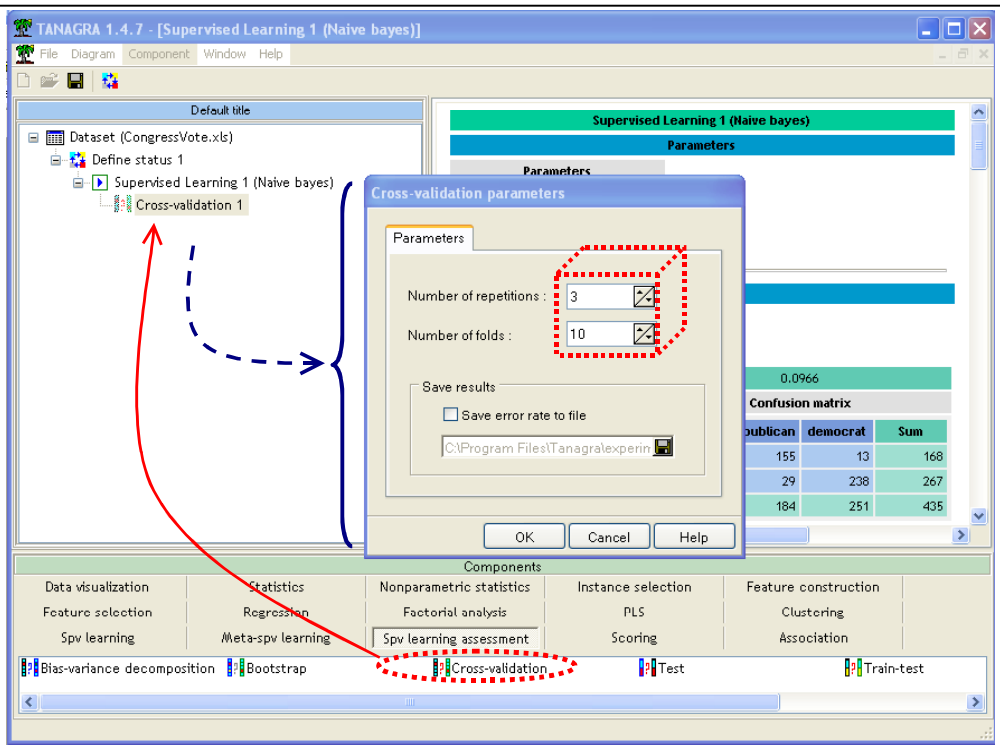
Nous glissons donc le composant NAIVE BAYES (onglet SPV LEARNING) dans le diagramme, après avoir relâché la souris, nous observons qu'il a été correctement complété.



L'apprentissage donne un taux de mauvais classement de 9.66%.

Nous savons que ce taux, estimé sur le fichier d'apprentissage, est biaisé. Pour obtenir une estimation plus crédible, nous voulons utiliser la validation croisée que nous retrouvons dans la palette SPV LEARNING ASSESSMENT. Nous modifions le paramétrage par défaut[1], nous le configurons de manière à ré-itérer 3 fois la « 10-fold cross validation ». Cette précision est importante, nous devrons utiliser le même protocole pour toutes les évaluations à venir dans notre didacticiel.

[1] Le paramétrage par défaut est de 5 fois une 2-fold validation croisée.



L'exécution indique un taux d'erreur de 10.08%. Notons que la procédure a bien respecté le paramétrage que nous avons défini (TRIALS = 3 ; FOLDS = 10).

Cross-validation 1

Parameters

Cross-validation parameters	
Folds	10
Trials	3

Results

CV error rate

Range	
MIN	0.0977
MAX	0.1047
Trial	**Err rate**
1	0.1047
2	0.1000
3	0.0977

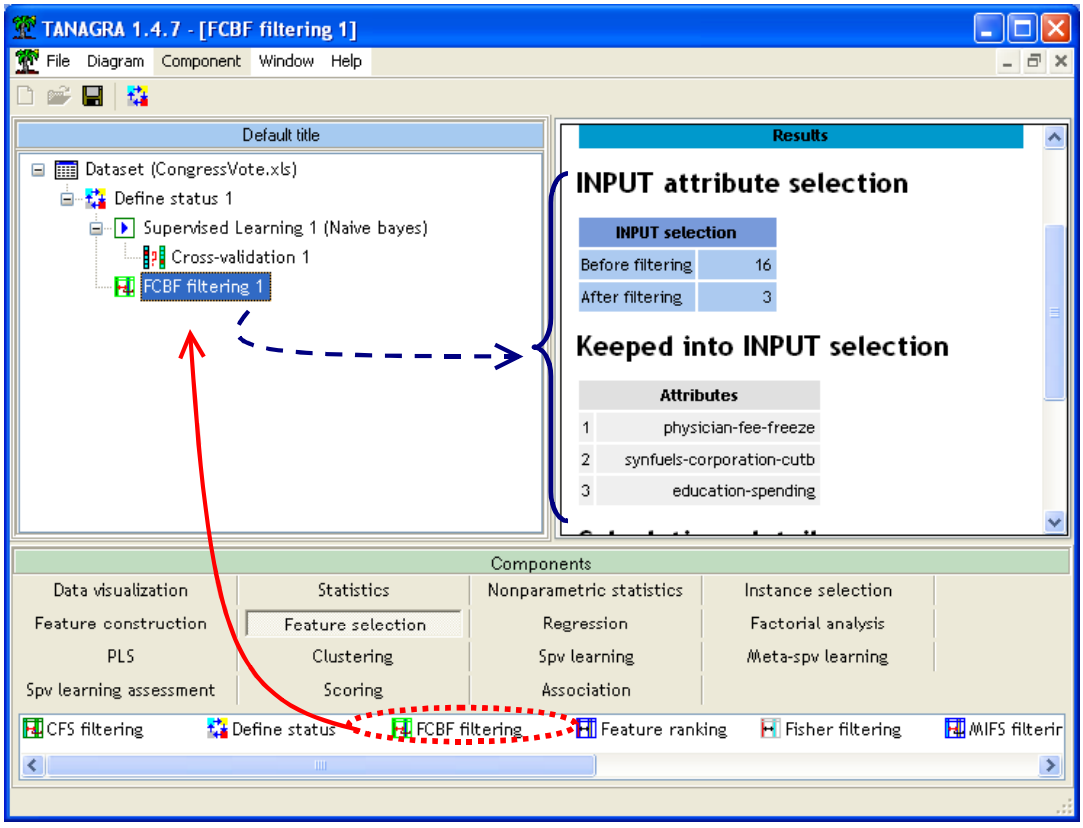
Overall cross-validation error rate

Error rate	0.1008
Values prediction	**Confusion matrix**

Sélection de variables dans le processus

Nous voulons ré-appliquer le même schéma, mais en insérant une procédure de sélection automatique de variables avant l'apprentissage du modèle de prédiction. En ne retenant que les variables pertinentes, nous espérons améliorer la qualité de l'apprentissage.

Nous le glissons dans le diagramme, en dessous du composant DEFINE STATUS, le composant FCBF[2] de la palette FEATURE SELECTION.

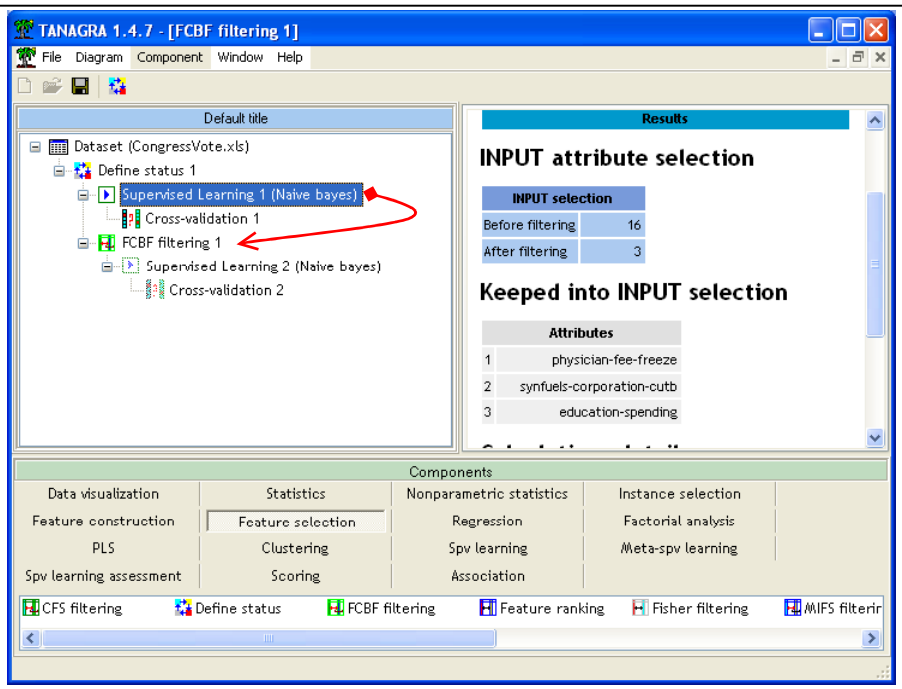


Nous constatons que la procédure FCBF a sélectionné 3 descripteurs parmi les 16.

Il nous reste alors à évaluer le processus global en insérant à la suite de FCBF l'apprentissage et son évaluation, toujours avec les mêmes paramètres.

Plutôt que de reproduire l'insertion et le paramétrage des composants, il nous est possible maintenant (version 1.4.7) de copier/coller des parties du diagramme. Pour ce faire, nous sélectionnons à l'aide de la souris l'item SUPERVISED LEARNING 1 (NAIVE BAYES) dans le diagramme, puis nous le glissons en dessous du composant FCBF. La séquence de traitements est dupliquée.

[2] FCBF est une méthode de filtrage qui s'appuie sur le calcul des corrélations. Le sous-ensemble de descripteurs proposé est indépendant de l'algorithme d'apprentissage mis à contribution par la suite.



En lançant l'exécution (menu VIEW du dernier composant du diagramme), nous obtenons un nouveau taux d'erreur de 5.5% en validation croisée.

Cross-validation 2

Parameters

Cross-validation parameters	
Folds	10
Trials	3

Results

CV error rate

Range	
MIN	0.0465
MAX	0.0628

Trial	Err rate
1	0.0465
2	0.0628
3	0.0558

Overall cross-validation error rate

Error rate	0.0550

Dans cet exemple précis, la réduction de la dimensionalité a permis d'améliorer significativement la qualité de l'apprentissage[3].

[3] Attention, à chaque étape de la validation croisée, FBCF est ré-exécuté. Rien ne dit que ce sont ces mêmes trois variables qui sont sélectionnés à chaque passage, il se peut également qu'il y ait plus ou moins de 3 variables dans la sélection. C'est bien la démarche « FBCF + apprentissage avec le modèle bayesien naïf » que nous évaluons, et non pas les trois variables mis en avant lorsque nous appliquons FBCF sur la totalité des données.

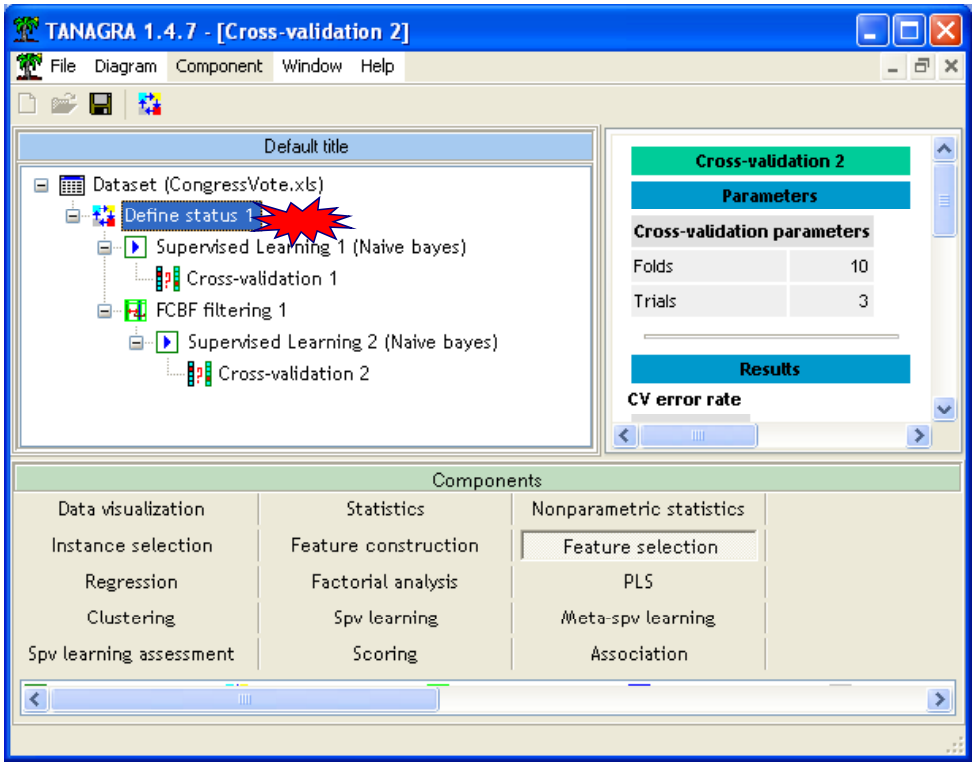
Reproduire la démarche sur un autre fichier

Ce premier résultat, très encourageant, nous pousse à vérifier l'efficacité de l'approche sur une autre base de données. Il faudrait dans ce cas reproduire toute la démarche ci-dessus en comparant les performances en prédiction avec et sans sélection de variables.

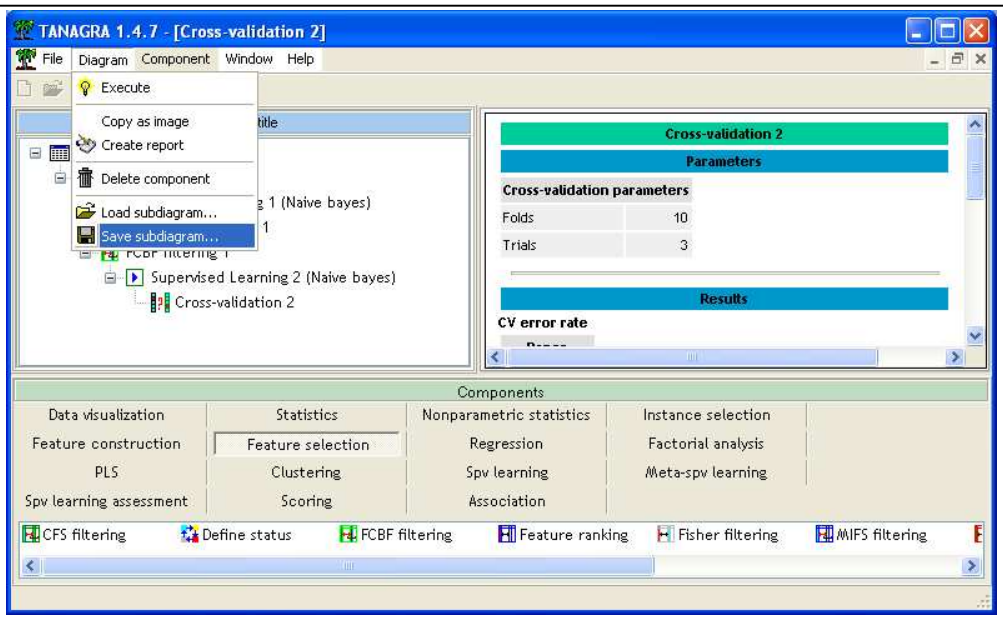
La nouvelle fonctionnalité que nous présentons dans ce didacticiel (version **1.4.8** de TANAGRA) permet de transposer des traitements définis sur une base de données sur une autre base de données. L'opération se fait en 3 étapes : sauver la séquence de traitements à reproduire ; importer les nouvelles données dans un nouveau diagramme ; insérer le sous-diagramme préalablement sauvegardé.

Sauver le sous-diagramme de traitements

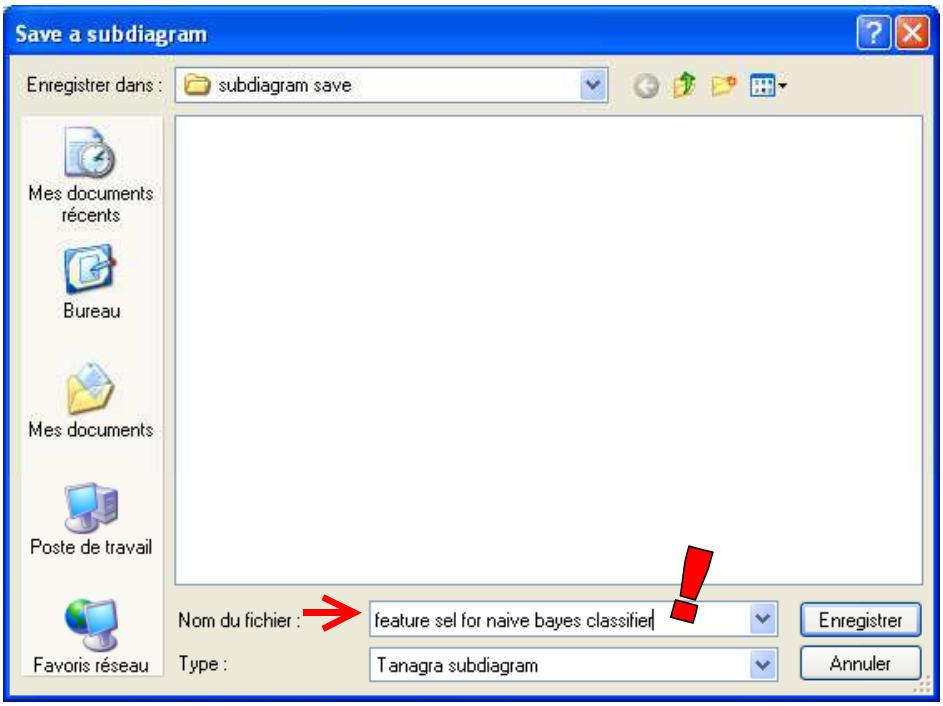
Nous voulons sauvegarder le processus d'apprentissage et d'évaluation, avec et sans la sélection de variables. Nous sélectionnons le composant DEFINE STATUS 1 dans le diagramme.



Puis nous activons le menu principal DIAGRAM / SAVE SUBDIAGRAM.



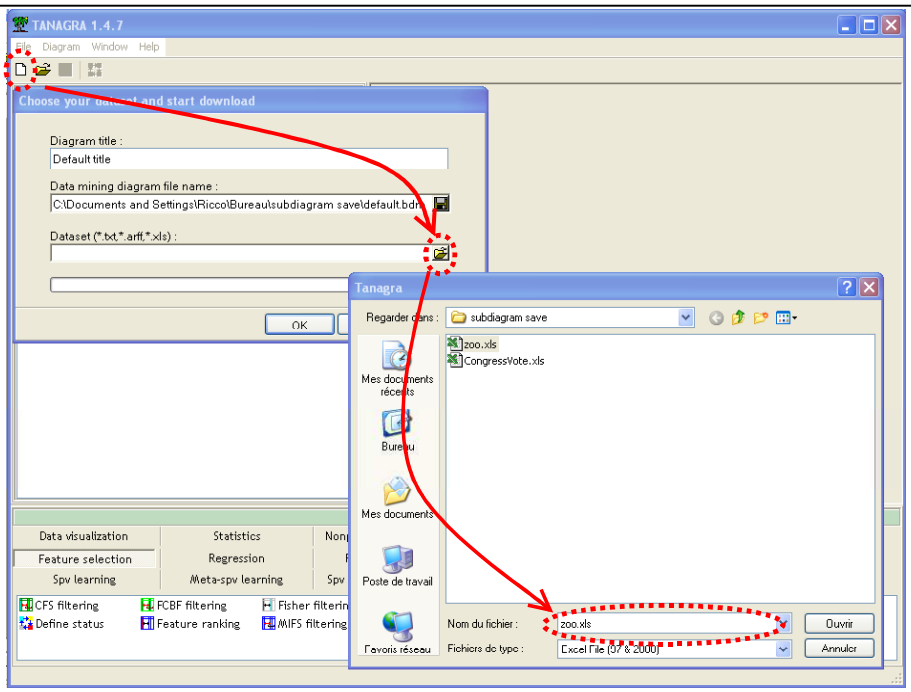
Une boîte de dialogue apparaît, elle nous permet de nommer le fichier de sauvegarde.



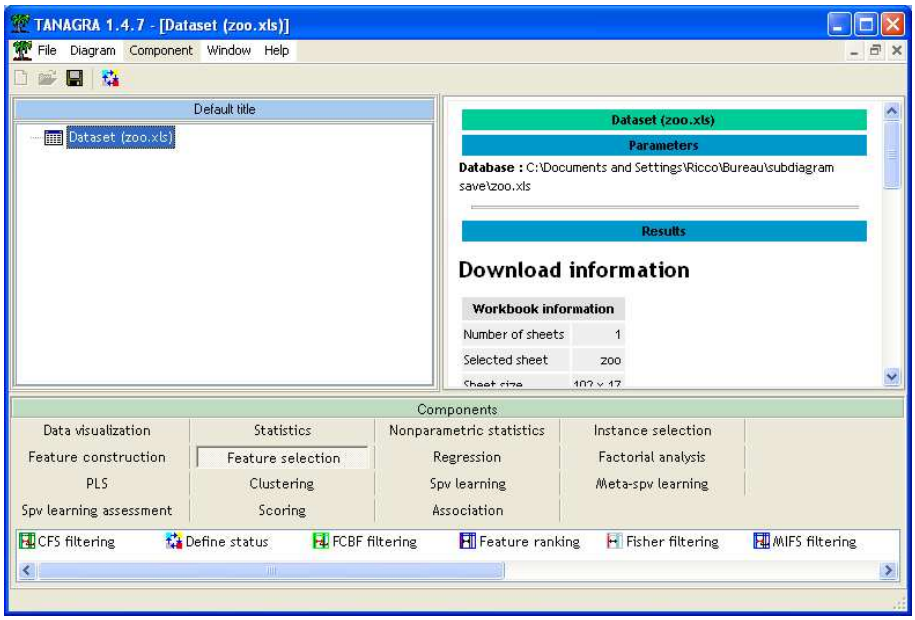
Une fois le sous-diagramme sauvegardé, nous pouvons fermer le diagramme courant à l'aide du menu FILE / CLOSE.

Nouveau diagramme et importation des données

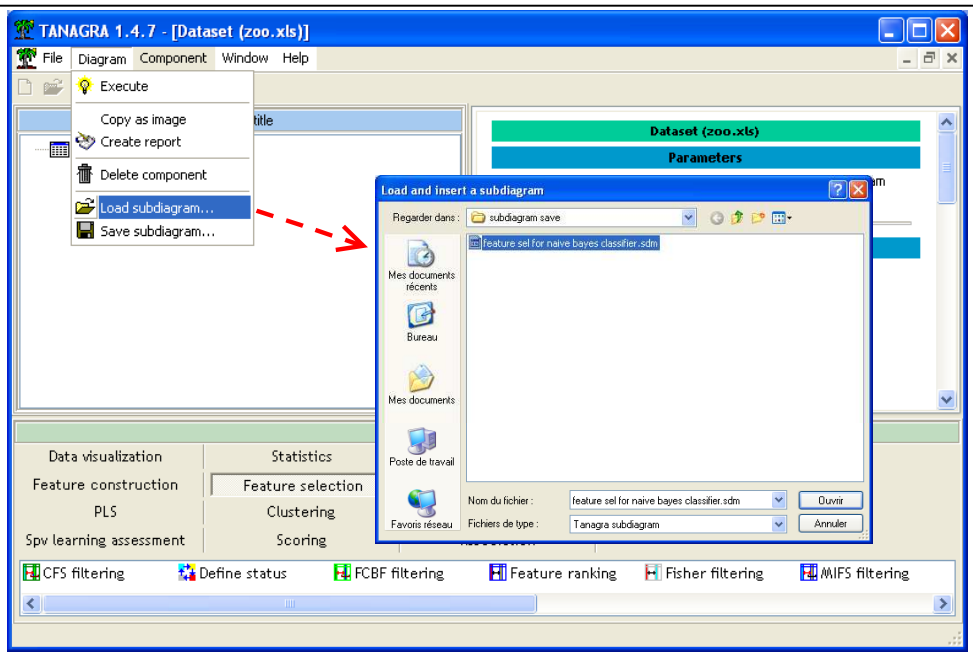
Nous voulons créer un nouveau diagramme et importer les données ZOO.XLS. De nouveau, nous activons le menu FILE / NEW, puis nous sélectionnons le fichier.



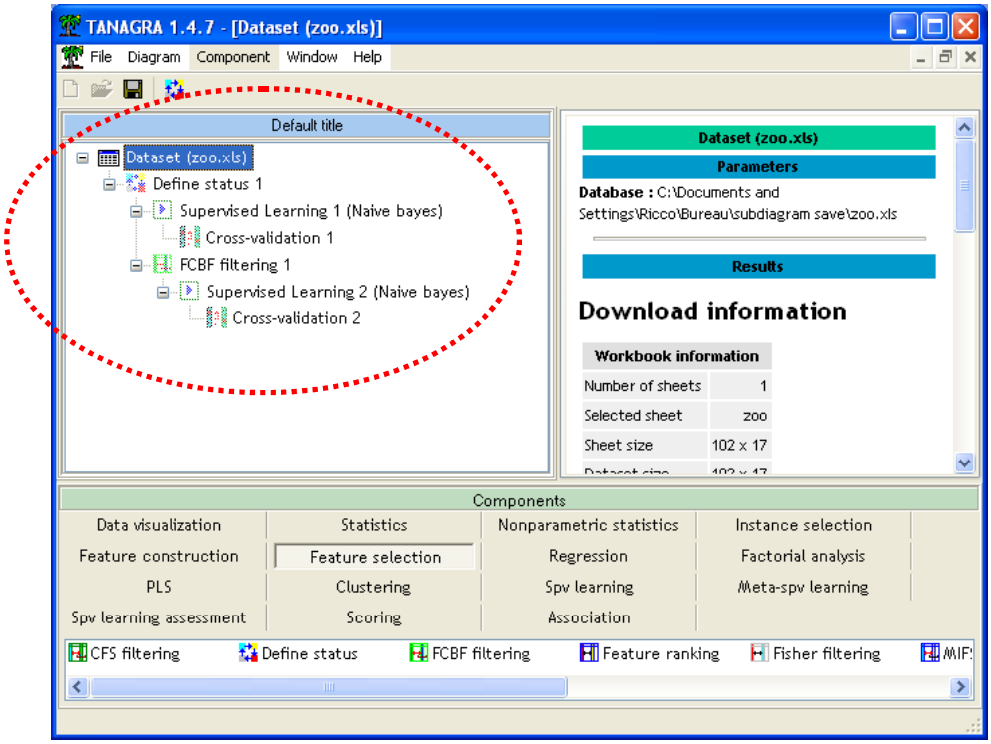
La fenêtre principale de l'application apparaît comme suit.



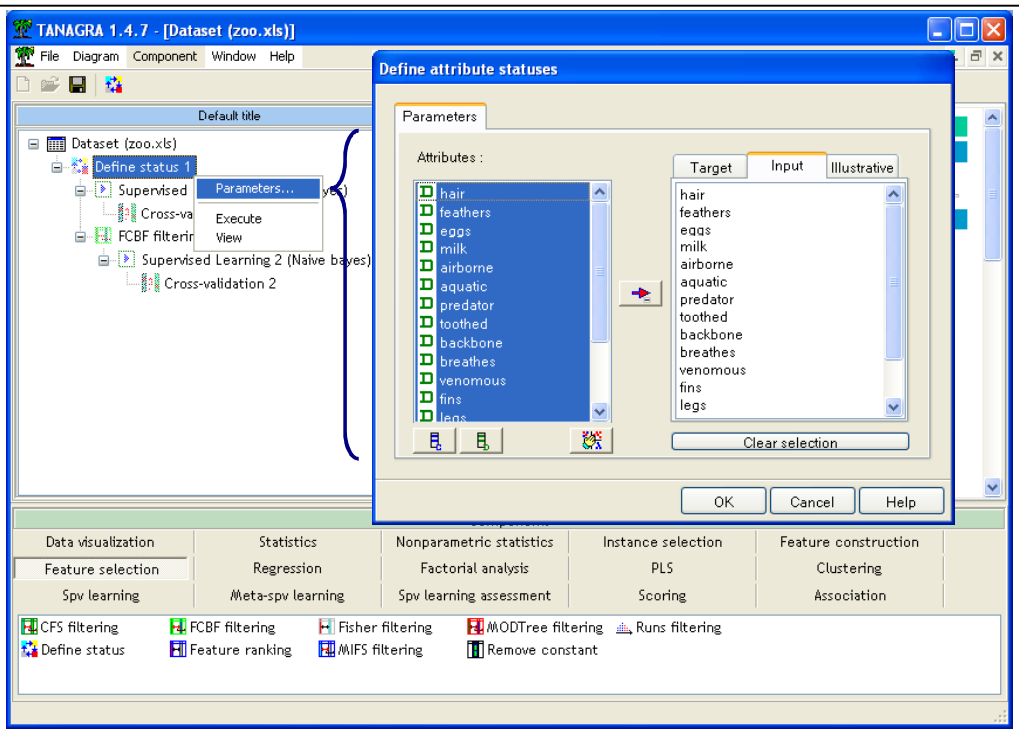
Notre idée maintenant est d'insérer la séquence de traitements définie dans l'analyse du fichier CONGRESSVOTE. Nous activons le menu principal DIAGRAM / LOAD SUBDIAGRAM. Nous sélectionnons le fichier précédemment créé.



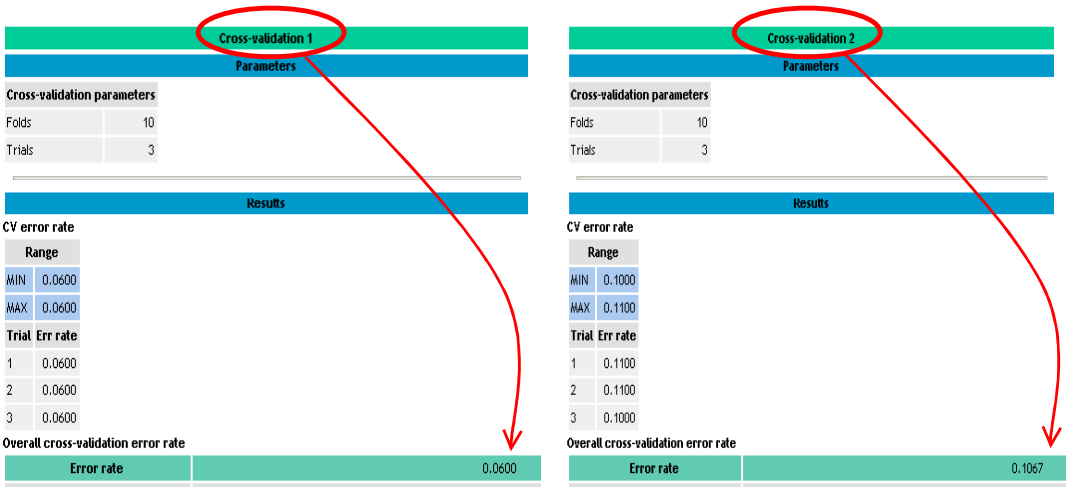
Le diagramme de traitements apparaît comme suit.



Il nous reste à configurer le composant DEFINE STATUS 1 de manière à définir correctement les descripteurs INPUT et l'attribut classe TARGET. Bien entendu, il faut supprimer les sélections réalisées dans le diagramme original (utiliser le bouton CLEAR SELECTION).



Nous lançons l'exécution de la validation croisée (Menu VIEW) de la validation croisée, sans (CROSS-VALIDATION 1) et avec sélection de variables (CROSS-VALIDATION 2).



Le taux d'erreur sans sélection de variables est de 6%, en introduisant FCBF dans la chaîne de traitements, il passe à 10.67%. Pour le fichier ZOO, à l'inverse de VOTE, la sélection FCBF est trop restrictive et dégrade les performances du modèle de prédiction.

Nous constatons surtout que cette nouvelle fonctionnalité permet de transposer très facilement des séquences de traitements d'un fichier de données à un autre en nous épargnant des manipulations fastidieuses.