Présentation du didacticiel

Dans ce didacticiel, vous allez apprendre à mettre en œuvre une méthode d'apprentissage supervisé.

Le fichier d'exemple utilisé est « breast.txt ».

Ce fichier, bien connu, est issu du domaine médical, et contient les caractéristiques physiologiques de cellules ponctionnées sur des patientes atteintes (ou non) du cancer du sein.

Nous mettons en œuvre dans ce didacticiel la méthode d'apprentissage ID3 (arbre de décision).

Vous apprendrez à utiliser les opérateurs suivants :

Onglet	**Opérateur**	**Commentaire**
Feature selection	Define status	Précision des variables à utiliser
Meta-spv learning	Supervised learning	Encapsule l'opérateur d'apprentissage
Spv learning	ID3	Opérateur d'apprentissage

Charger les données dans Tanagra

- Ouvrir un diagramme existant

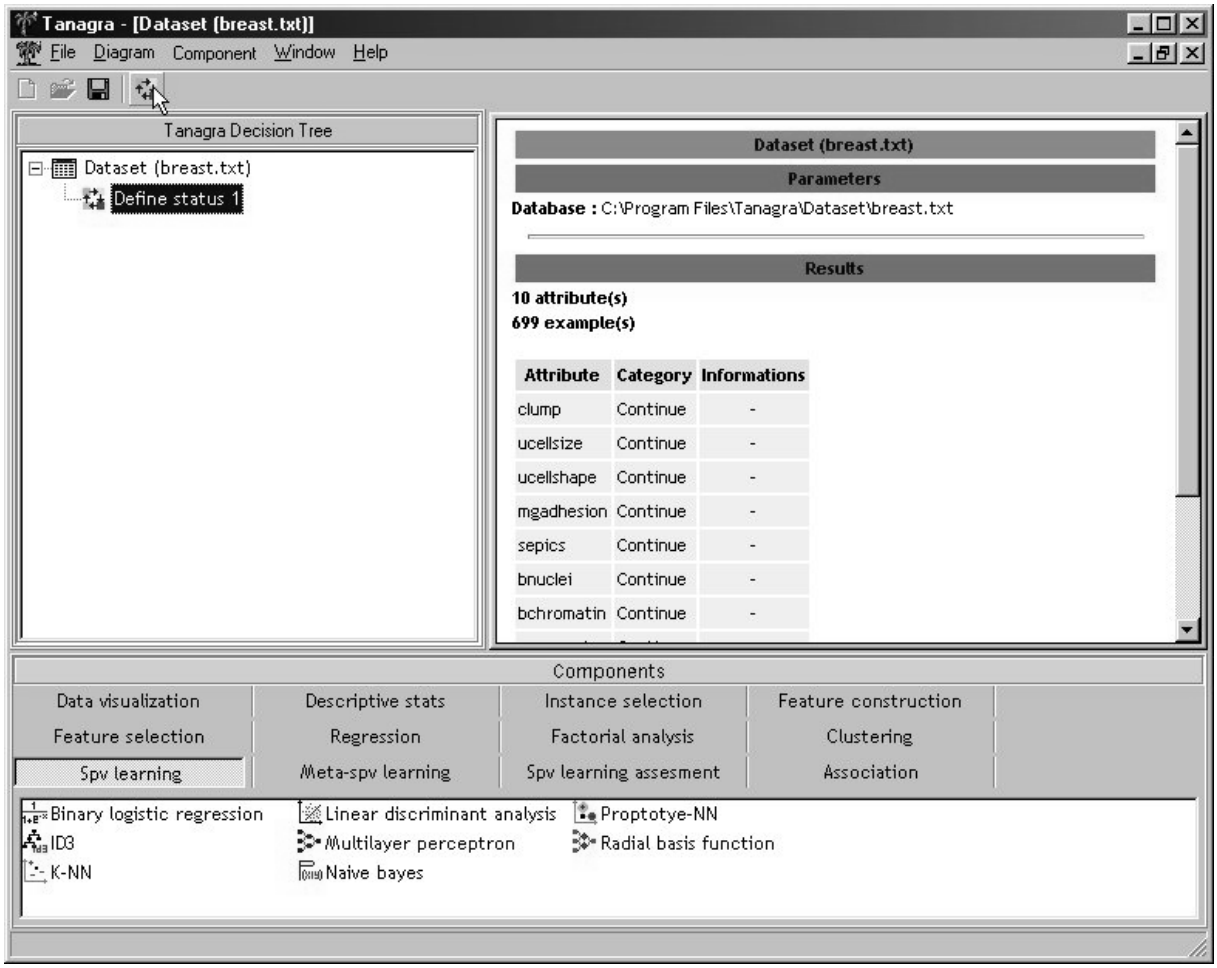
1 – Choisissez *File/Open…* dans le menu de Tanagra.

2 – Allez chercher le fichier « breast.bdm » situé dans le sous-répertoire « Dataset » de Tanagra.

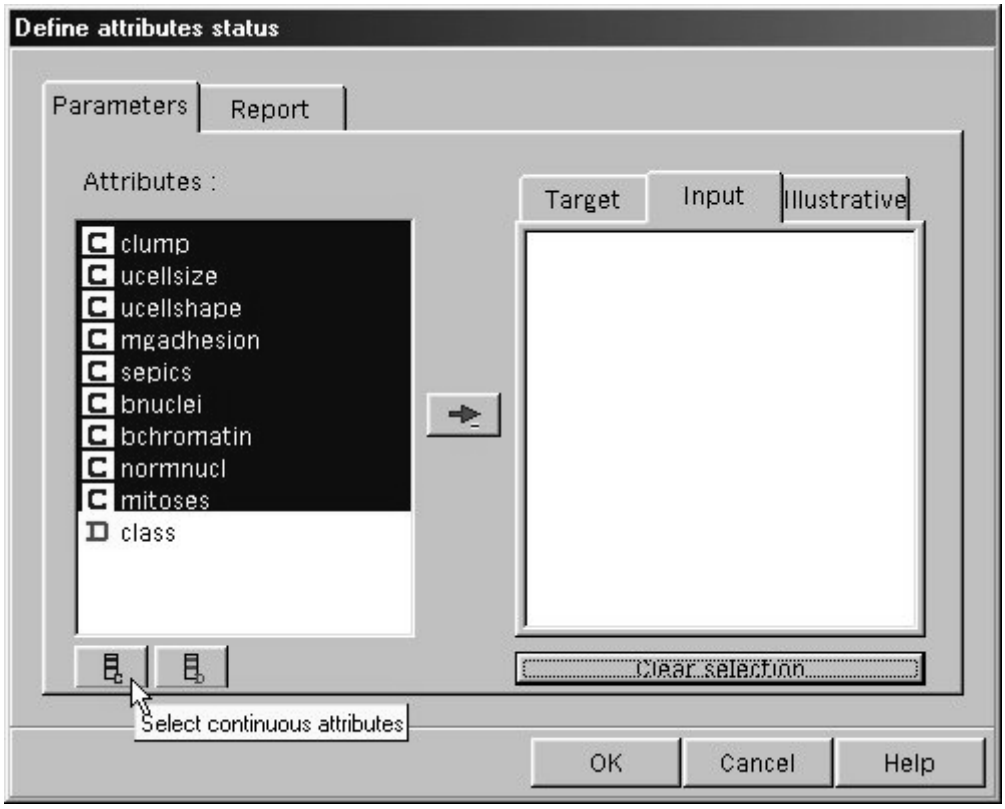
Utiliser un opérateur d'apprentissage supervisé

- Définir le statut des variables

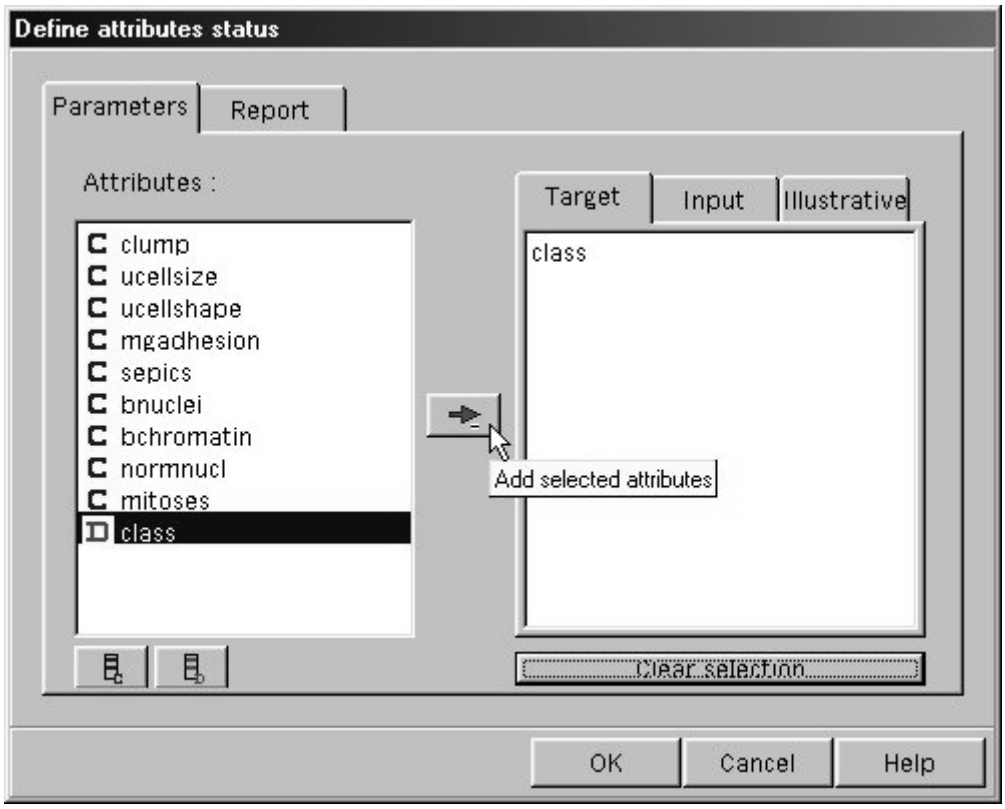
1 – Placez-vous sur le nœud « Dataset » et ajoutez un opérateur **Define Status** en cliquant sur son icône dans la barre des raccourcis. La fenêtre de dialogue permettant de définir le statut des variables apparaît automatiquement.



2 – Assurez-vous que c'est l'onglet « Input » qui est actif. Sélectionnez les variables continues de la liste en cliquant sur le bouton correspondant (cf ci-dessous), et cliquez enfin sur le bouton flèche pour les passer dans la liste des Input.



3 – Toujours en restant dans la fenêtre de dialogue, activez l'onglet Target. Cliquez sur la variable « class » pour la sélectionner, puis sur le bouton flèche.



4 – Vous venez de définir la variable à prédire (« class » = Target) et les variables explicatives (les autres = Input). Appuyez sur OK pour valider et fermer cette fenêtre.

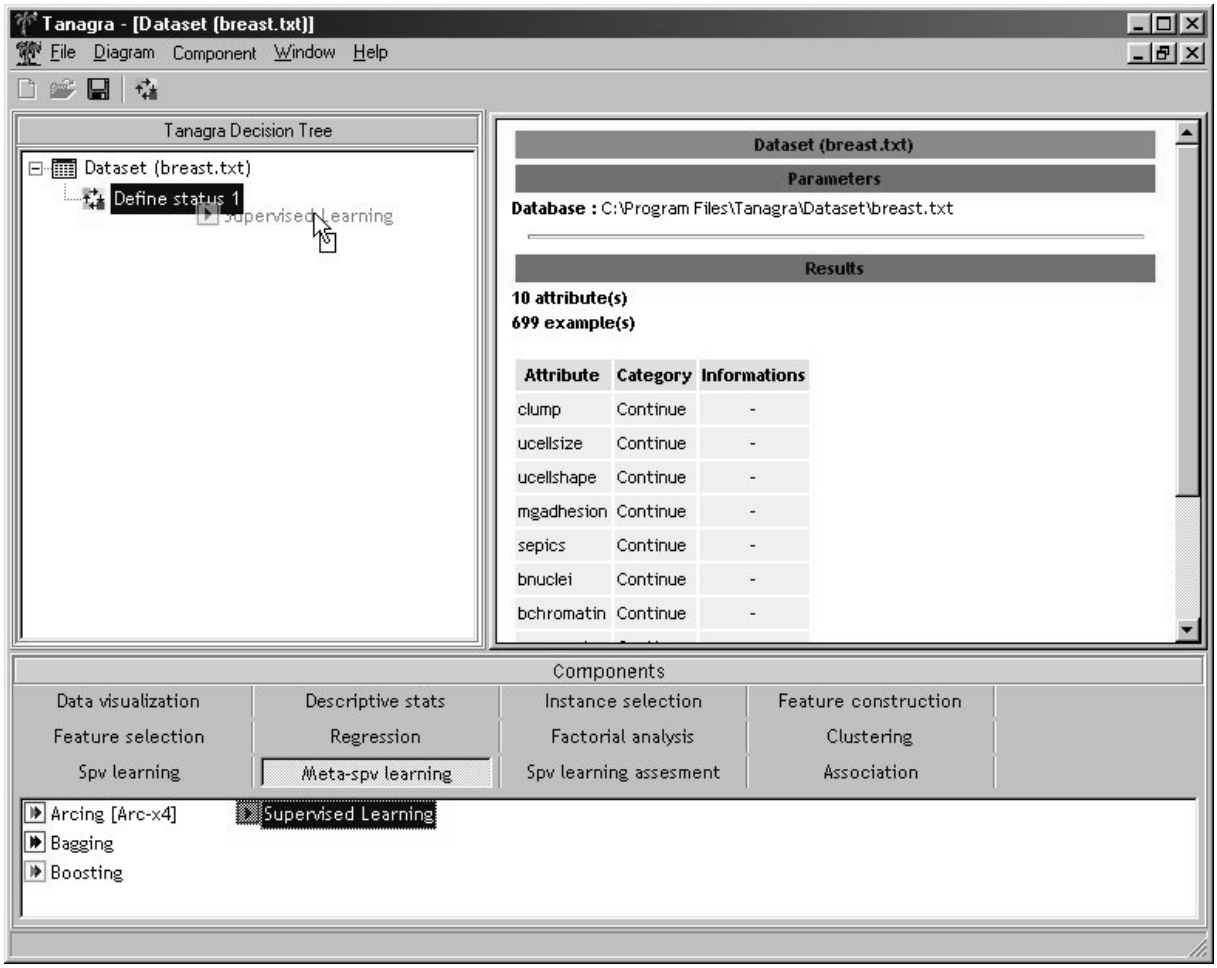
➢ Choisir un méta-opérateur d'apprentissage supervisé

Dans certaines expérimentations, on lance plusieurs fois l'apprentissage sur le même fichier, en pondérant les individus différemment à chaque fois. Le but est d'obtenir un meilleur modèle de prédiction.

Tanagra implémente ce chaînage d'un opérateur d'apprentissage via les méta-opérateurs.

Dans ce didacticiel nous ne rentrerons pas dans ce type d'expérimentations, nous ne lancerons qu'une seule fois l'opérateur ID3. Mais Tanagra oblige à utiliser un méta-opérateur. Il en propose toutefois un pour les lancements uniques de méthode. Il s'agit de l'opérateur **Supervised Learning**.

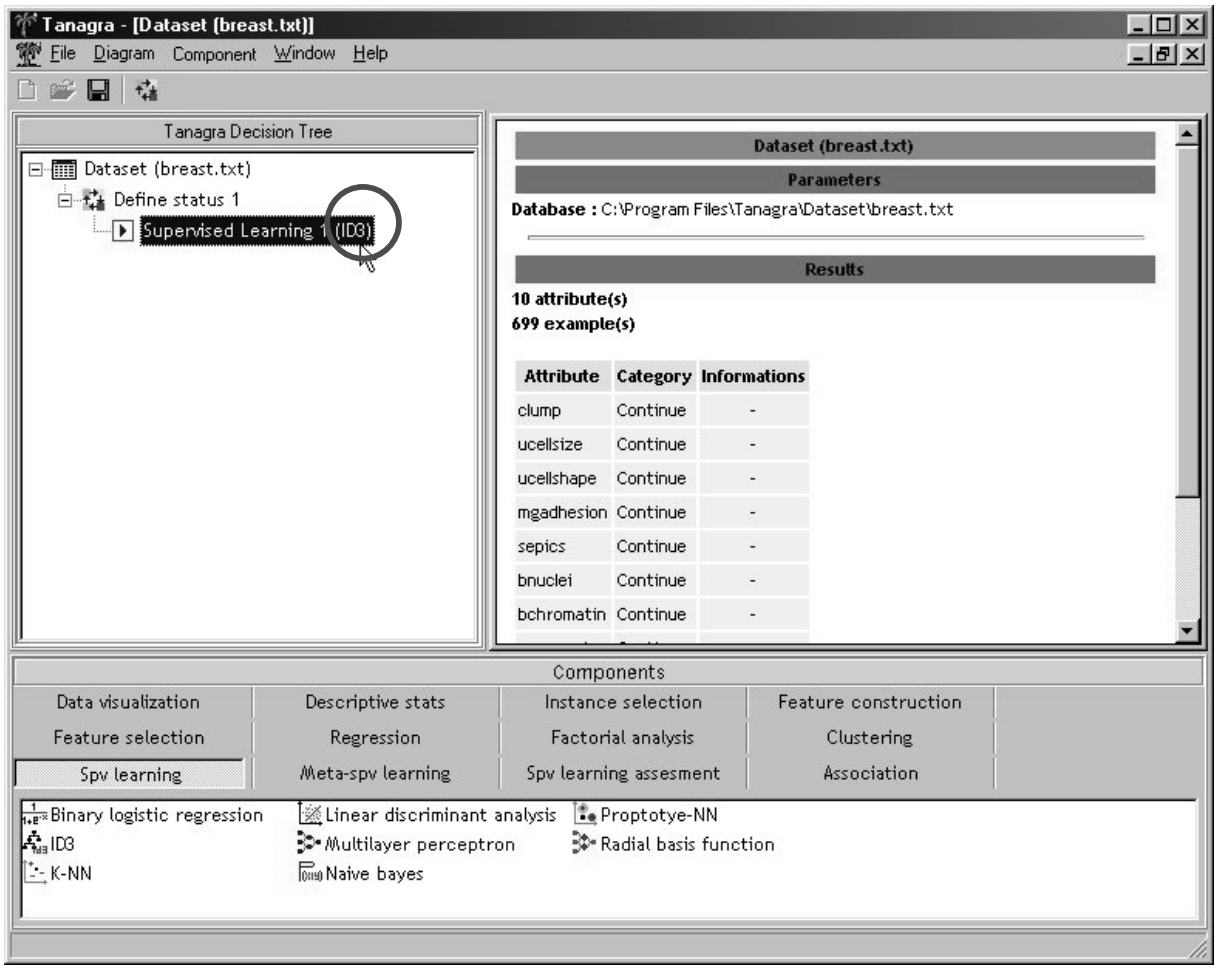
1 – Ajoutez un opérateur **Supervised learning** (onglet META-SPV LEARNING) au diagramme, sous le nœud «Define status 1».



➢ Choisir un opérateur d'apprentissage supervisé

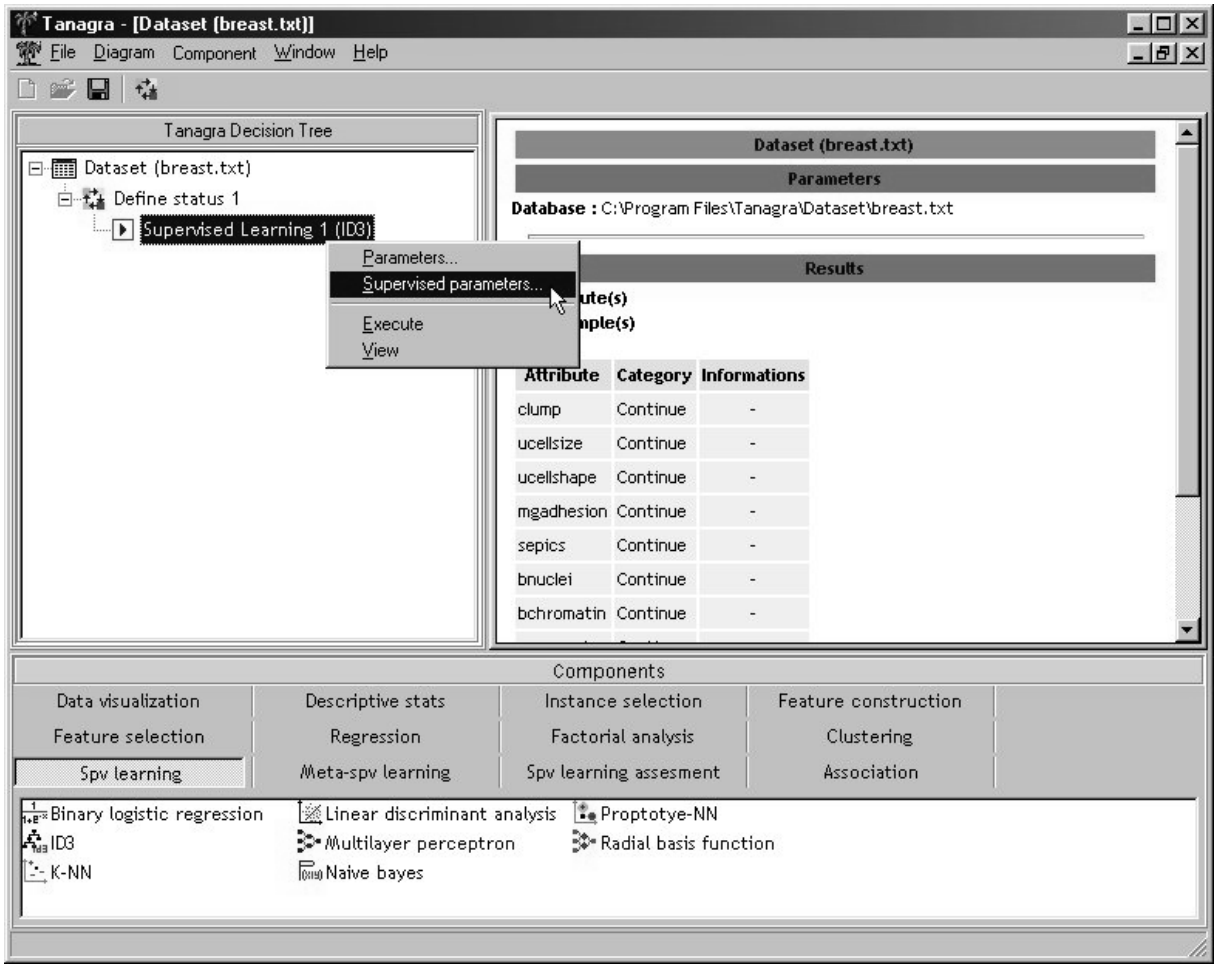
1 – Dans la palette des opérateurs, cliquez sur l'onglet SPV LEARNING, et faites glisser un opérateur **ID3** sur le nœud « Supervised Learning » que vous venez d'ajouter.

L'opérateur est inclus dans le méta-opérateur, aussi voit-on son libellé dans celui du nœud du méta-opérateur, et non pas en-dessous de celui-ci.



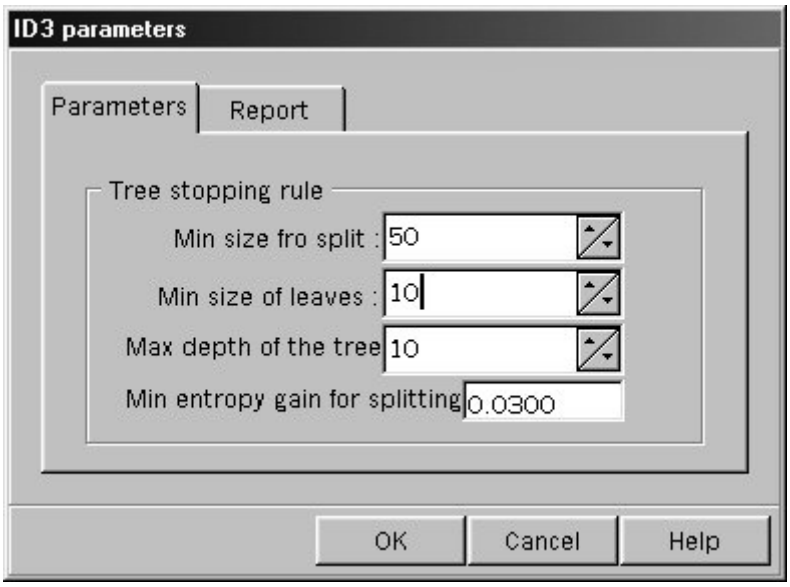
➢ Définir les paramètres de l'apprentissage (opérateur ID3)

1 – Faites apparaître le menu contextuel du nœud « Supervised learning (ID3) par clic droit sur ce dernier. En plus de la commande Parameters… habituelle, on trouve une commande Supervised parameters…



La première s'applique au méta-opérateur, la seconde à l'opérateur d'apprentissage supervisé. Choisissez celle-ci.

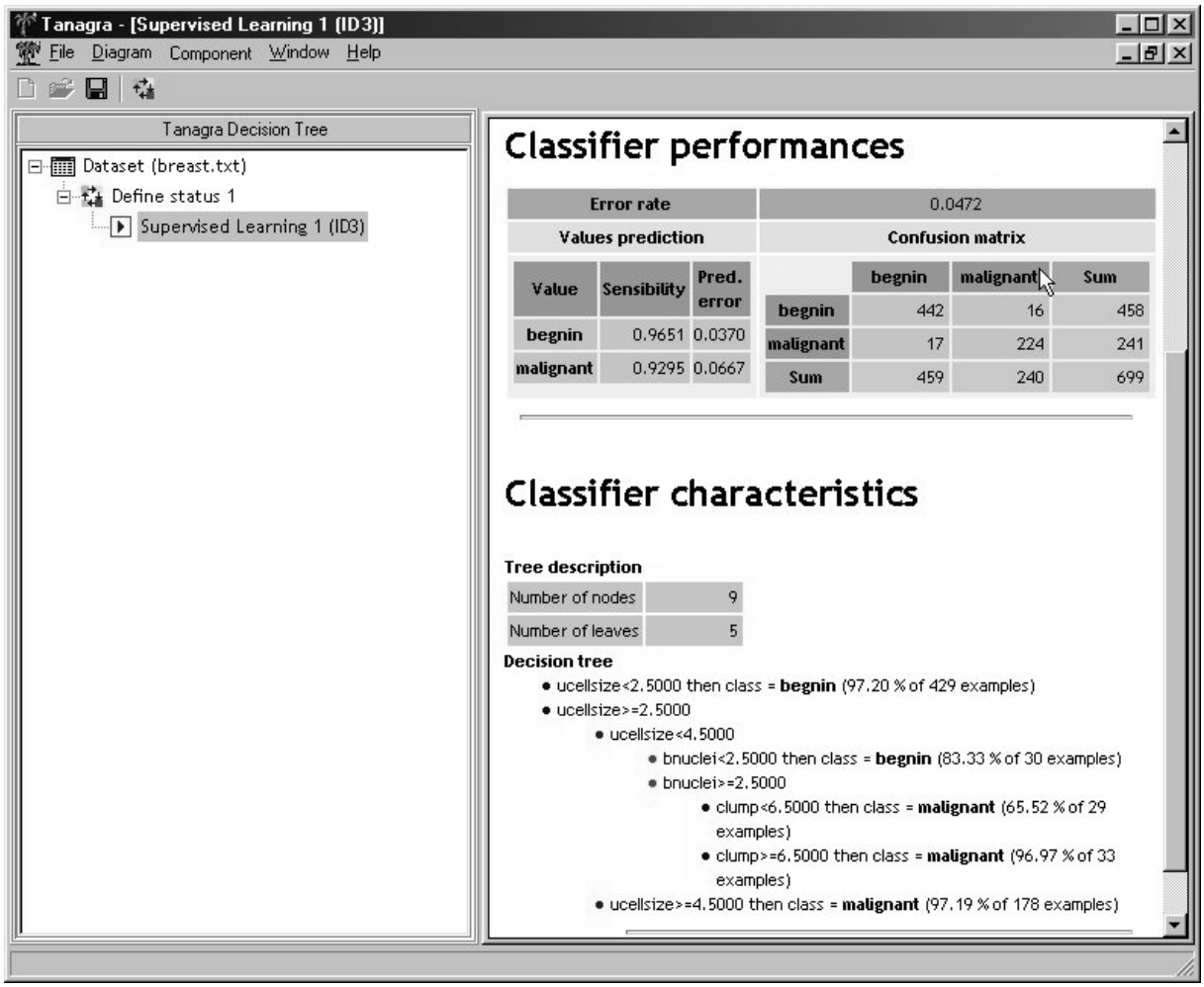
2 – Dans la fenêtre de dialogue qui s'affiche, compte tenu de la taille du fichier étudié (699 individus), modifiez les paramètres de ID3 comme suit :



3 – Validez en cliquant sur OK.

Effectuer l'apprentissage

1 – Dans le menu contextual du noeud, choisissez *View.* Les résultats s'affichent dans le cadre de droite.



Le taux d'erreur calculé sur l'apprentissage paraît bon (4,72 %). On voit dans la matrice de confusion que l'erreur se répartit également entre diagnostiquer un cancer à tort et ne pas diagnostiquer un cancer existant.

On constate dans l'arbre retranscrit l'importance de la variable ucellsize dans le diagnostic automatique.