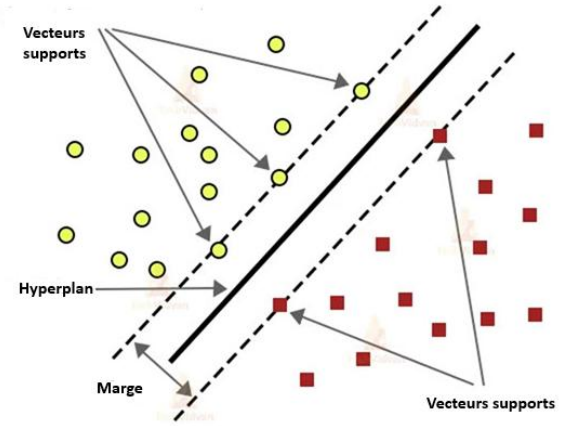
UNIWIN VERSION 10.4.0

SEPARATEURS A VASTES MARGES

Révision : 15/09/2025

Définition

La procédure *Séparateurs à vastes marges* implémente une procédure d'apprentissage machine pour prévoir des observations à partir de données. Elle crée des modèles de deux formes : *modèles de classement* qui découpent des observations en groupes en se basant sur les caractéristiques observées, *modèles de régression* qui prévoient la valeur d'une variable cible.



Dans le cas d'un modèle de classement, l'algorithme découpent les observations en groupes en générant des marges autour des groupes aussi vastes que possible.

Dans le cas d'un modèle de régression, l'algorithme minimise les coefficients d'un modèle dans lequel la distance des observations à une région autour du modèle ajusté définie par un montant d'erreur acceptable est aussi petite que possible.

Les observations sont classiquement divisées en trois jeux :

- Un jeu d'*apprentissage* utilisé pour construire le modèle.
- Un jeu de *validation*, pour lequel le groupe ou la valeur est connu, utilisé pour valider le modèle.
- Un jeu de *prévision*, pour lequel le groupe ou la valeur n'est pas connu, utilisé pour faire les prévisions désirées.

La variable à expliquer est qualitative (classement) ou quantitative (régression).

Cette procédure est basée sur le package R ‘e1071’.

Entrée des données

Cliquons sur l’icône SVM dans le ruban Expliquer pour afficher la boîte de dialogue montrée ci-après.

Cette boîte de dialogue permet de préciser la variable à expliquer ainsi que les variables explicatives quantitatives (s’il y en a) et qualitatives (s’il y en a) à utiliser.

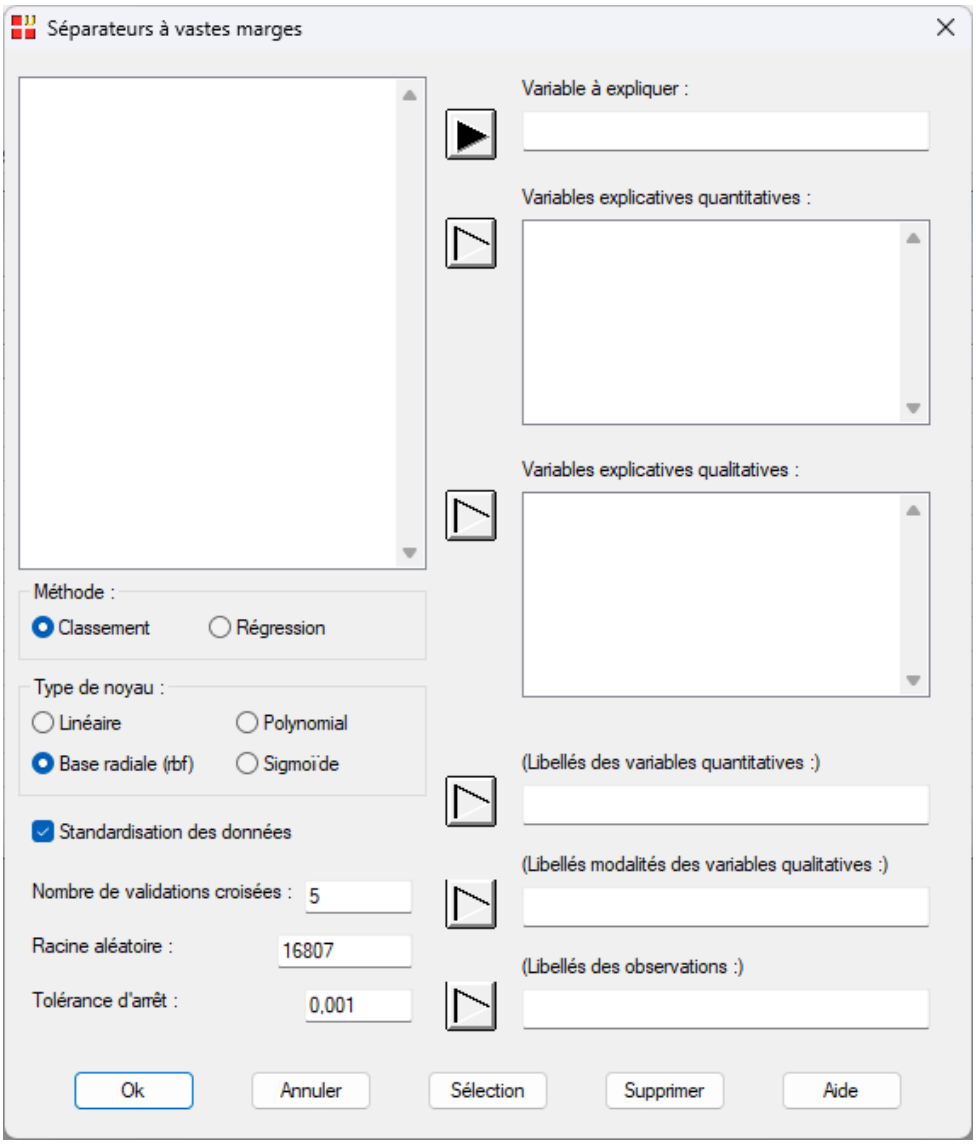
Elle permet également, en option, d’indiquer le nom de la variable contenant les libellés des variables quantitatives, le nom de la variable contenant les libellés des modalités des variables qualitatives ainsi que le nom de la variable contenant les libellés des observations.

Par défaut, les données sont standardisées (centrées et réduites).

Le choix de la méthode est proposé : Classement (si la variable à expliquer est qualitative) ou Régression (si la variable à expliquer est quantitative).

Quatre types de noyaux sont possibles : linéaire, polynomial, base radiale et sigmoïde.

Le nombre de validations croisées, la racine aléatoire et la tolérance d’arrêt peuvent être précisés.



Données manquantes

Les observations ayant des données manquantes pour les variables explicatives sont automatiquement éliminées des calculs.

Les observations ayant des données manquantes pour la variable à expliquer constituent le jeu de prévision.

Exemple 1 : Classement binaire - Fichier Ann

Pour ce premier exemple « académique » que nous présenterons rapidement, nous utiliserons le fichier Ann.

Ce fichier contient 2000 observations décrites par trois variables *X1*, *X2* et Y.

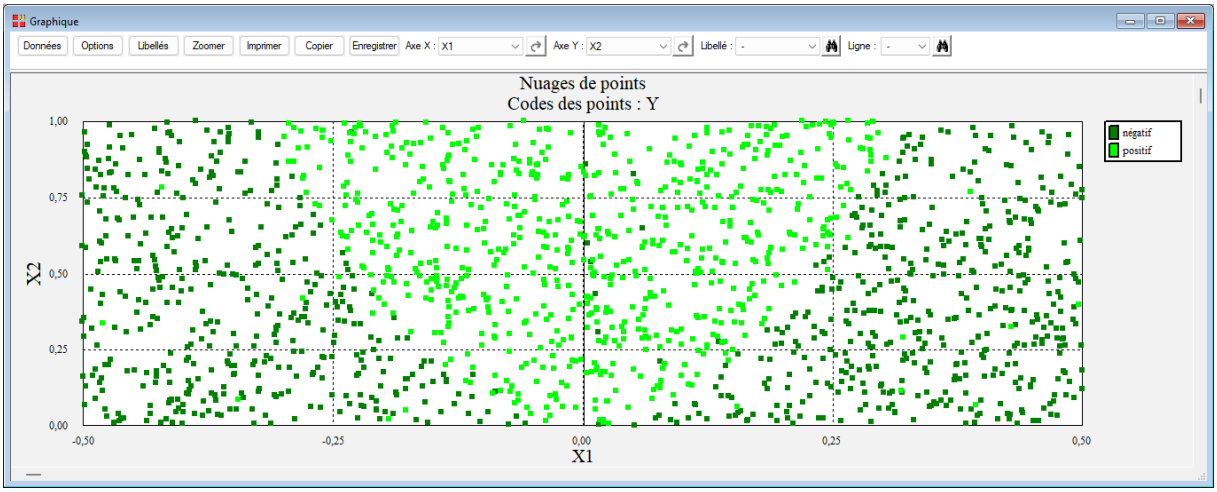
X1 et X2 sont deux variables aléatoires uniformes définies respectivement sur les intervalles [-0,5 ; 0,5] et [0 ; 1].

Y est une variable à deux modalités prenant les valeurs 'négatif' et 'positif' selon la formule bien évidemment inconnue :

si (0,1 * X2 > X1²) alors Y = 'positif' sinon Y = 'négatif'.

A noter qu'il y a 6 valeurs manquantes pour Y.

Visualisons les données dans un nuage de points codifiés.

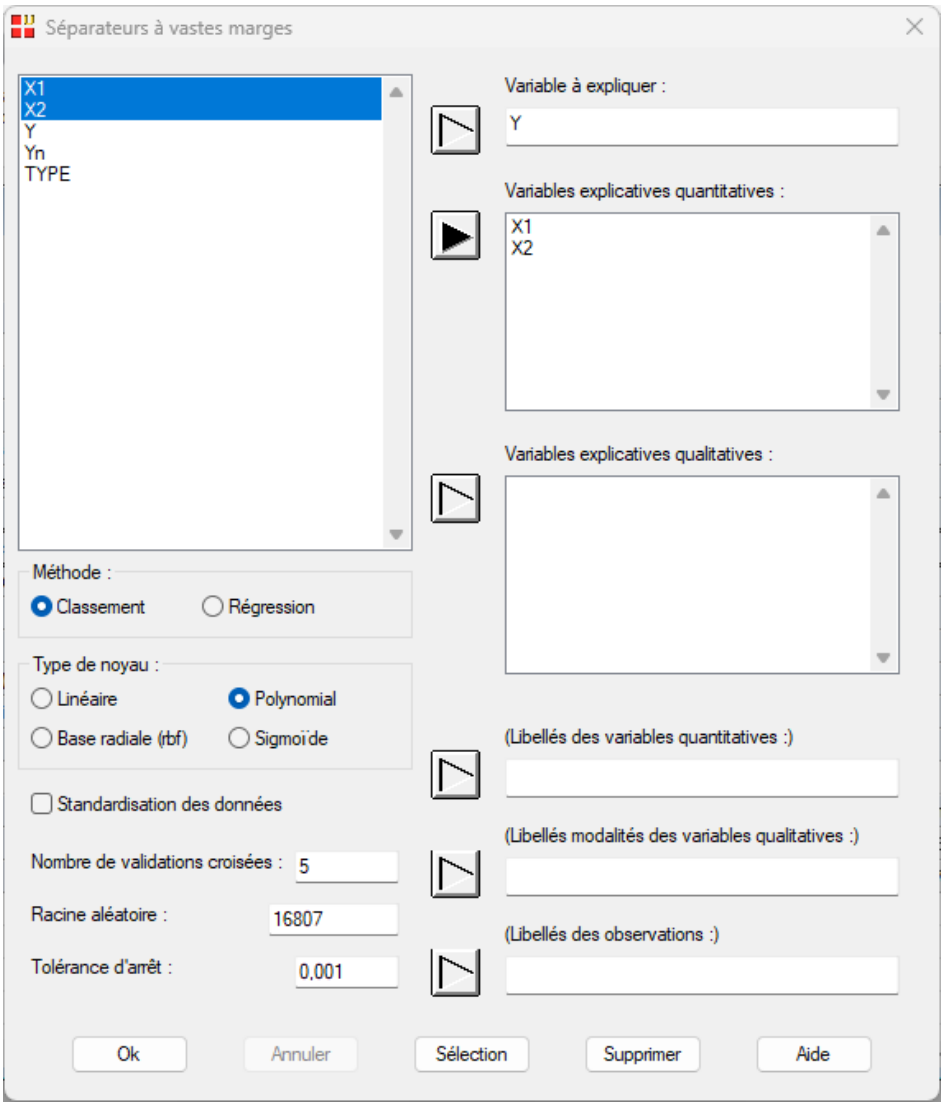


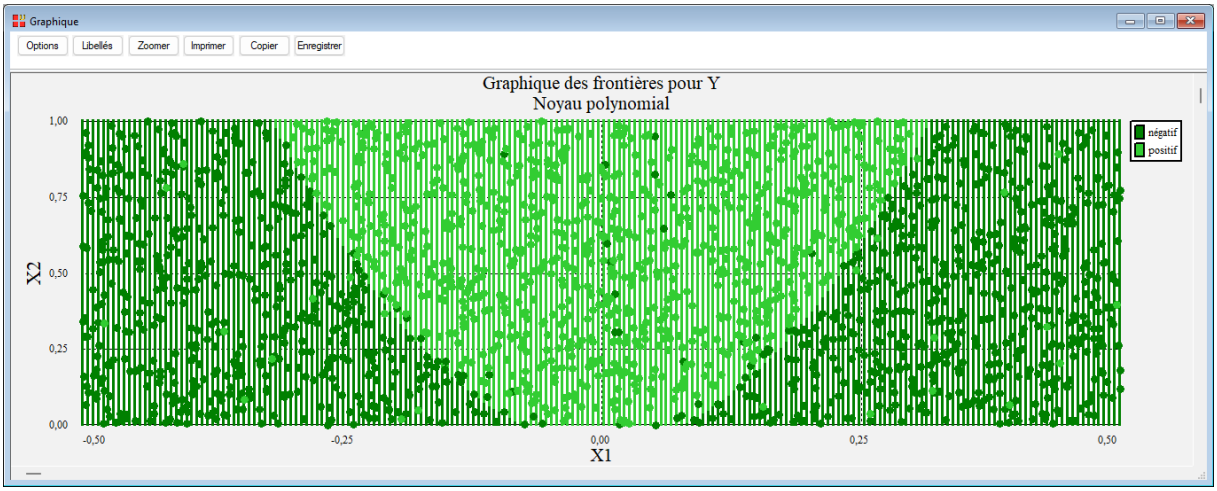
Cliquons sur l'icône SVM dans le ruban Expliquer et renseignons la boîte de dialogue comme montré ci-après.

Demandons un ajustement *Polynomial de degré 2*.

Après exécution des calculs, les paramètres optimaux obtenus sont :

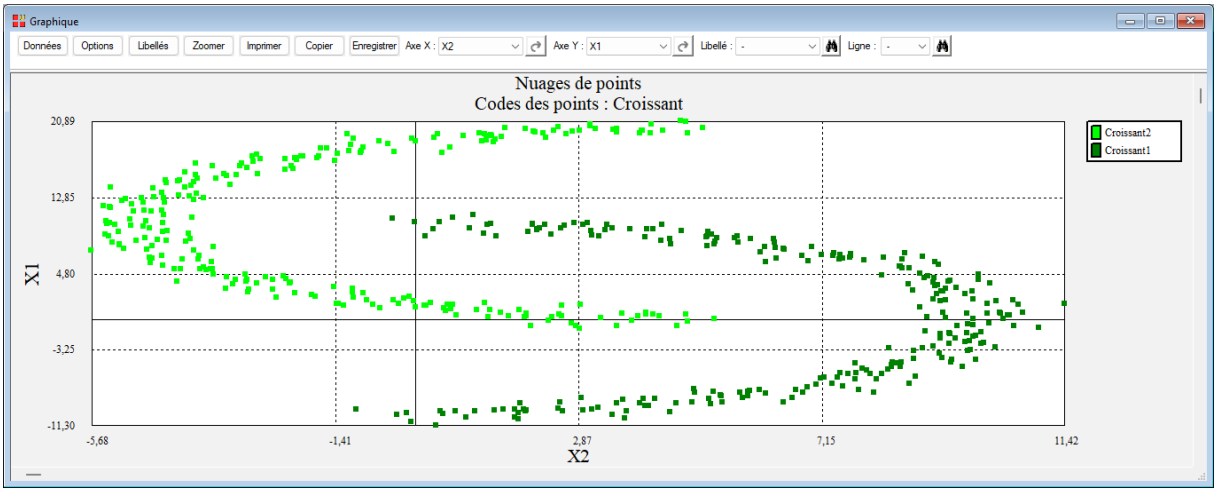
Noyau de type polynomial
- degré du polynôme : 2
- terme constant coef0 : 0,1
- coefficent C de coût : 1000
- coefficient gamma : 1


Séparateurs à vastes marges
X1
X2
Y
Yn
TYPE
Variable à expliquer :
Y
Variables explicatives quantitatives :
X1
X2
Variables explicatives qualitatives :
Méthode :
Classement
Régression
Type de noyau :
Linéaire
Polynomial
Base radiale (rbf)
Sigmoïde
Standardisation des données
Nombre de validations croisées : 5
Racine aléatoire : 16807
Tolérance d'arrêt : 0,001
(Libellés des variables quantitatives :)
(Libellés modalités des variables qualitatives :)
(Libellés des observations :)
Ok
Annuler
Sélection
Supprimer
Aide


Graphique
Options
Libellés
Zoomer
Imprimer
Copier
Enregistrer
Graphique des frontières pour Y
Noyau polynomial
négatif
positif
X2
X1
1,00
0,75
0,50
0,25
0,00
-0,50
-0,25
0,00
0,25
0,50

Exemple 2 : Classement binaire - Fichier Moons

Pour ce deuxième exemple « académique » que nous présenterons rapidement, nous utiliserons le fichier Moons. Ce fichier contient 500 observations décrites par trois variables *X1*, *X2* et *Croissant*. Visualisons les données dans un nuage de points codifiés.

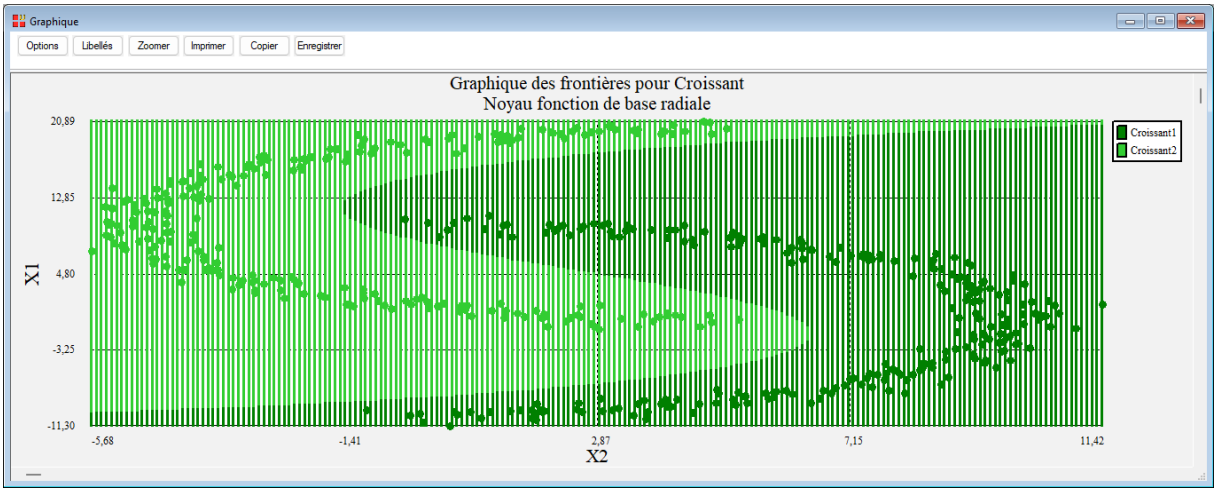


Cliquons sur l'icône SVM dans le ruban Expliquer et renseignons la boîte de dialogue en indiquant *Croissant* comme variable à expliquer, *X1* et *X2* comme variables explicatives et ne standardisons pas les données.

Demandons un ajustement *Base radiale.*

Après exécution des calculs, les paramètres optimaux obtenus sont :

Noyau de type fonction de base radiale
- coefficent C de coût : 100
- coefficient gamma : 0,001



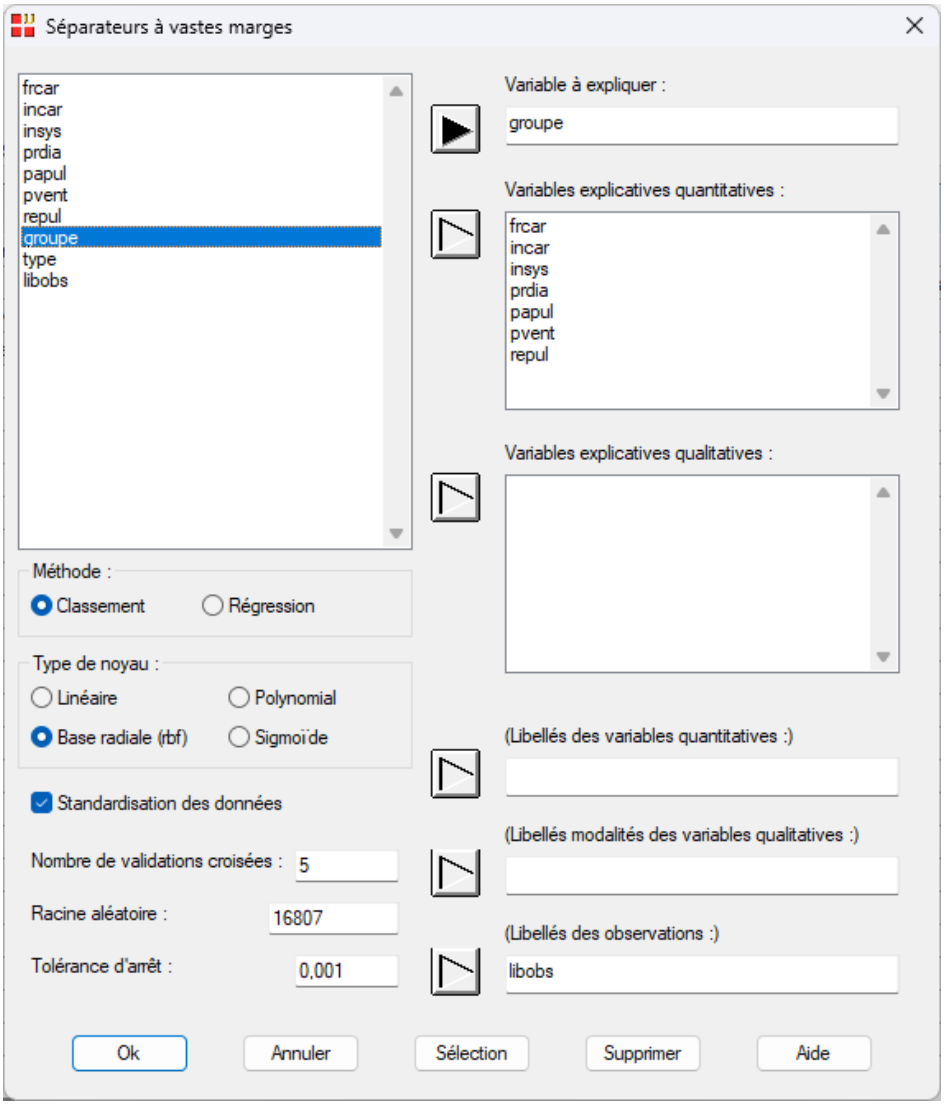
Exemple 3 : Classement binaire - Fichier Infarct2

Ce fichier contient des informations concernant 101 victimes d'un infarctus du myocarde. Les variables mesurées sont :

frcar fréquence cardiaque
incar index cardiaque
insys index systolique
prdia pression diastolique
papul pression artérielle pulmonaire
pvent pression ventriculaire
repul résistance pulmonaire

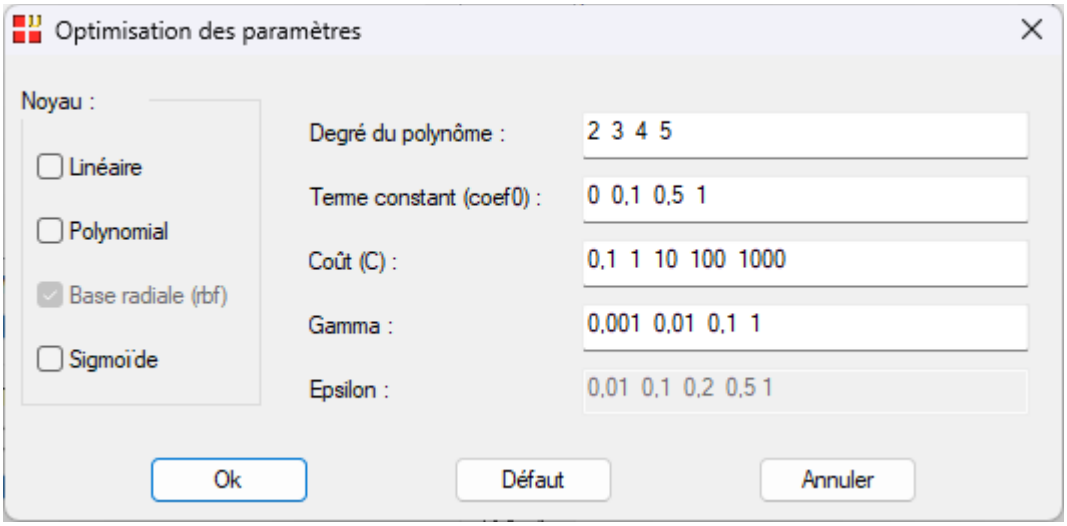
La variable *libobs* contient les libellés des observations et la variable qualitative *groupe* indique par ses deux codes les personnes décédées ou vivantes.

Sélectionnons la variable *groupe* comme variable à expliquer, les variables *frcar* à *repul* comme variables explicatives quantitatives, la variable *libobs* pour les libellés des observations, choisissons la méthode *Classement*, un noyau *Base radiale (rbf),* standardisons les données et laissons les autres paramètres aux valeurs par défaut.



Cliquons sur OK et confirmons que la méthode de classement utilise une variable à 2 modalités.

Une fenêtre nous demande ensuite de définir les options pour l'optimisation des paramètres de la méthode de classement :



Le noyau *Base radiale (rbf)* a été choisi par défaut. C'est celui le plus fréquemment utilisé. Il est toutefois possible de tester d'autres types de noyaux : *Linéaire*, *Polynomial* et *Sigmoïde*.

	Base radiale (rbf)	Linéaire	Polynomial	Sigmoïde
Degré du polynôme			X	
Terme constant (coef0)			X	X
Coût (C)	X	X	X	X
Gamma	X		X	X

Le paramètre Epsilon n'est utilisé que pour un modèle de régression.

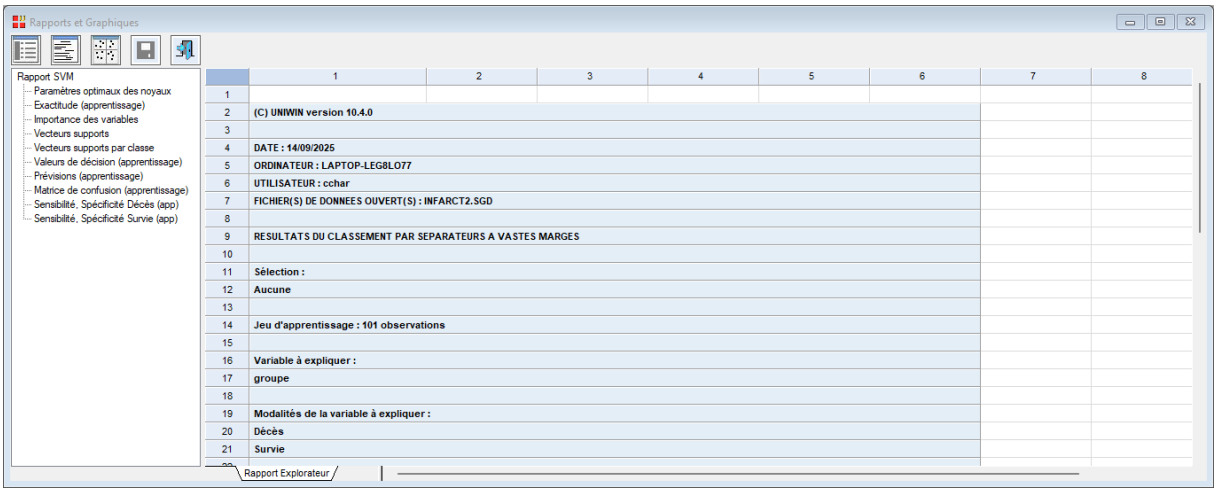
Voir le paragraphe '**Formules des calculs**' pour plus de détails sur les types de noyaux et les paramètres.

Des valeurs usuelles de ces paramètres sont proposées par défaut. Il est possible de saisir d'autres valeurs.

Note : Si de nombreuses valeurs sont entrées pour chacun des paramètres, le temps de calcul peut rapidement devenir important. En cas de besoin, il est possible de modifier le temps maximum d'exécution via le menu 'Fichier – Installer R et ses packages'.

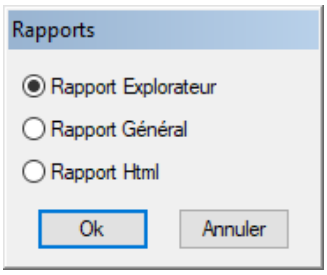
Cliquons sur *Défaut*.

Après l'affichage de messages indiquant la progression des calculs, la fenêtre suivante s'affiche :

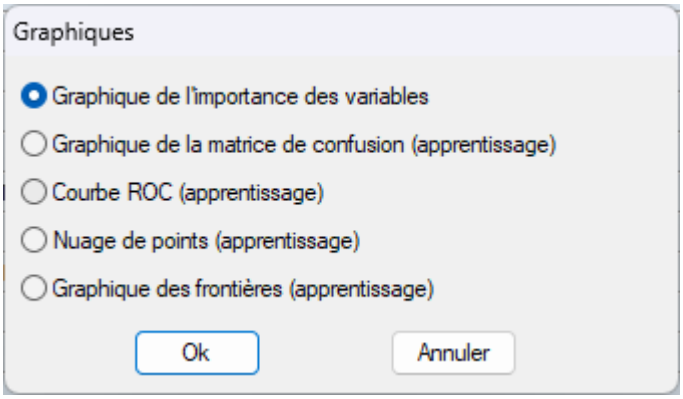


La barre d'outils 'Rapports et Graphiques' permet par l'icône 'Données'  de rappeler la boîte de dialogue d'entrée des données.

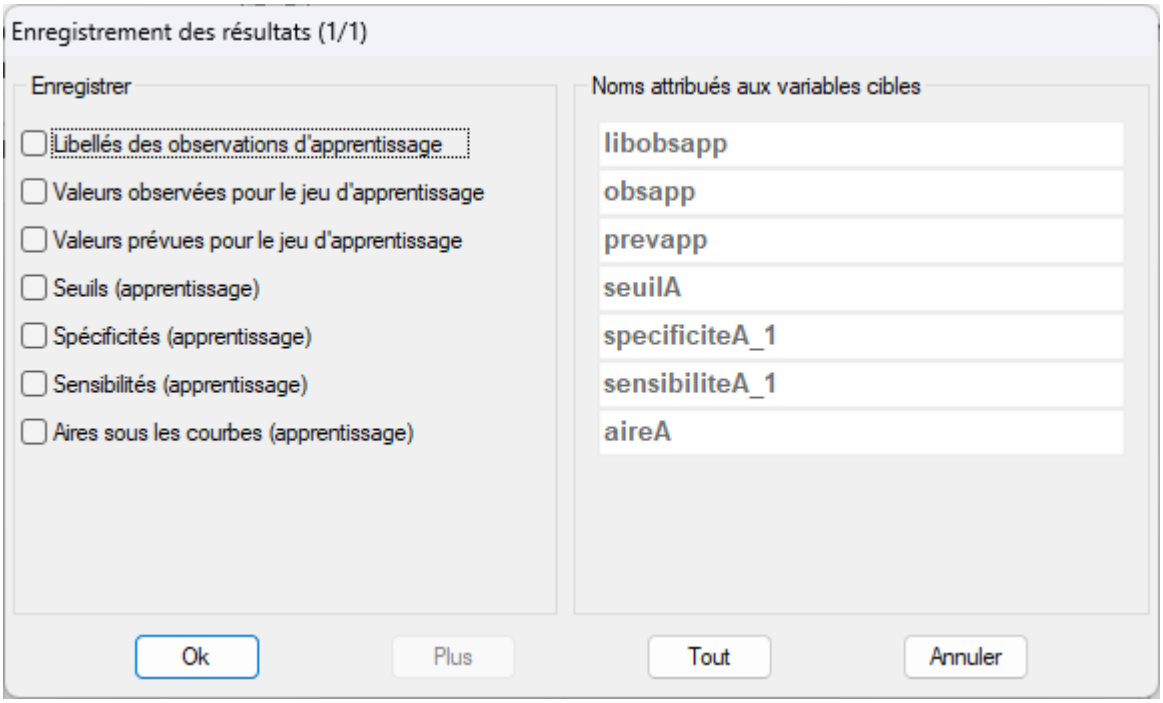
L'icône 'Rapports'  affiche la boîte de dialogue des options pour les rapports :



et l'icône 'Graphiques'  affiche la boîte de dialogue des options pour les graphiques.



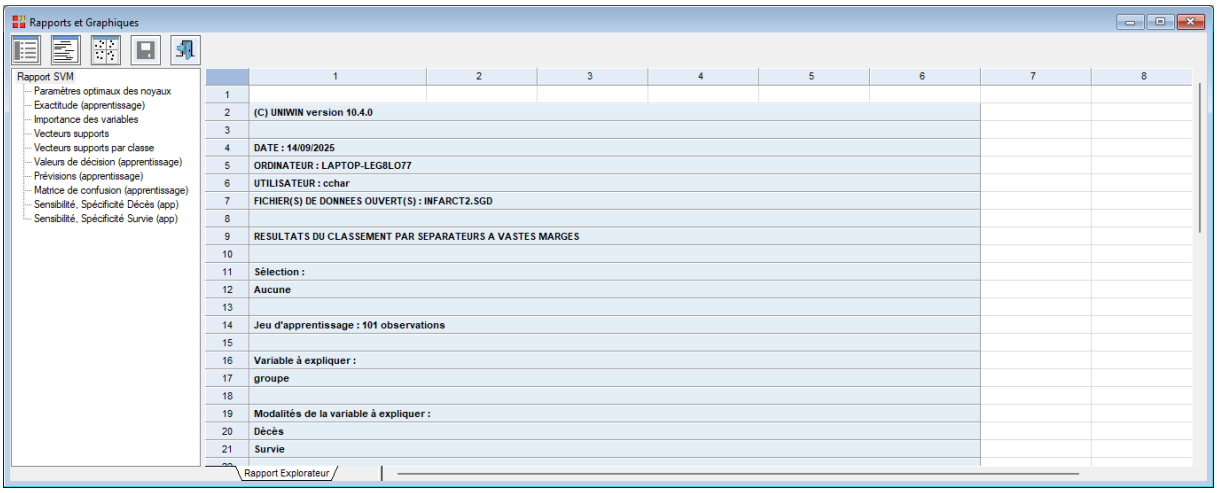
L'icône 'Enregistrer'  permet de sélectionner les résultats de l'analyse à enregistrer dans un fichier.



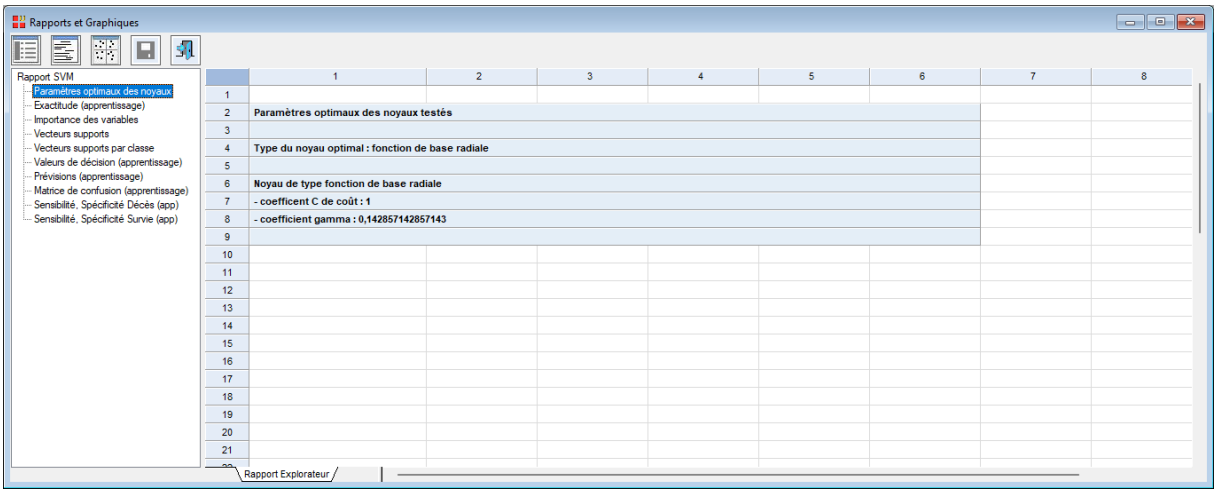
L'option Rapports

Cette option permet d'obtenir le rapport à l'écran sous la forme d'un explorateur, d'un tableur ou au format HTML.

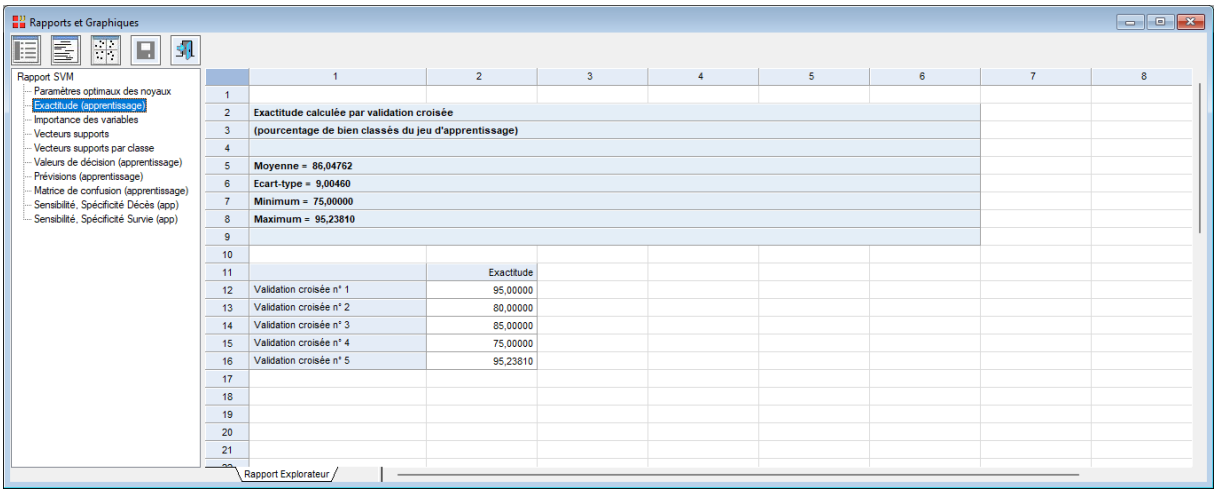
Le premier tableau indique que le jeu d'apprentissage est composé de 101 observations et rappelle les modalités de la variable à expliquer, les variables explicatives utilisées et quelques paramètres de l'analyse.



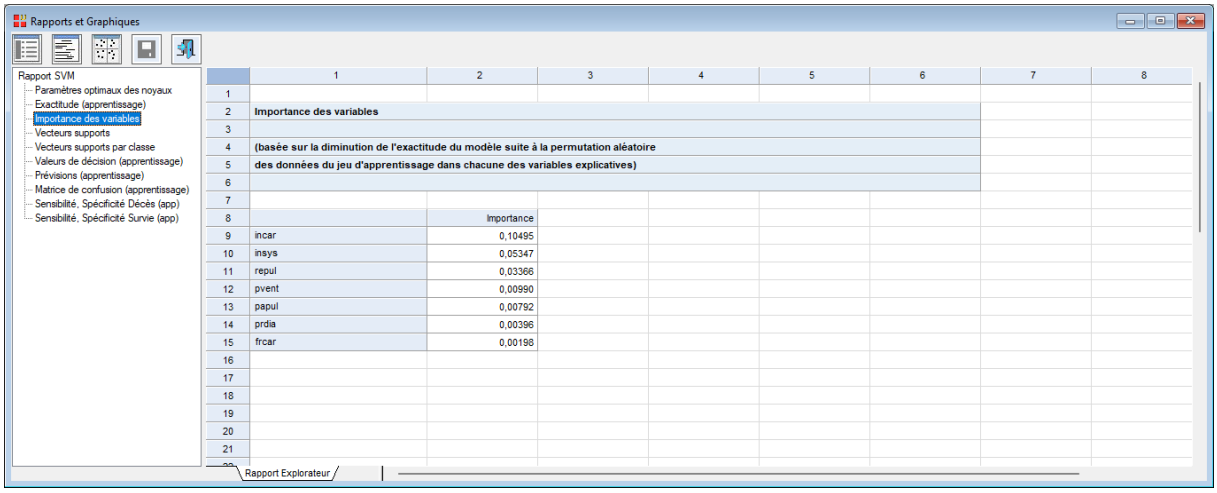
Le deuxième tableau indique les paramètres optimaux calculés pour le noyau *Base radiale (rbf)* : C = 1 et Gamma = 0,1428.



Le troisième tableau affiche l'exactitude calculée pour chacune des 5 validations croisées. Un faible écart-type de ces exactitudes est souhaité car il indique un modèle stable.

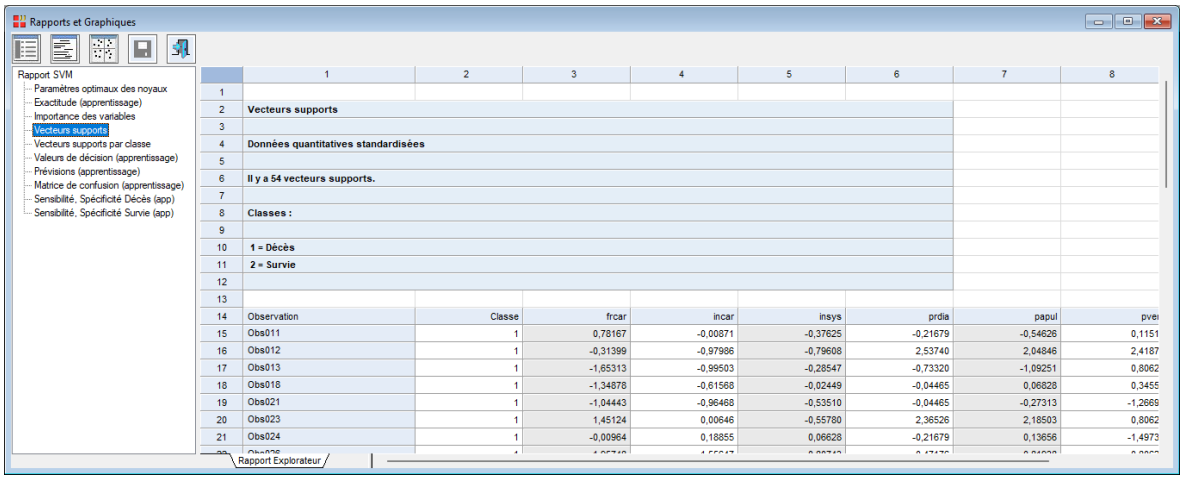


Le quatrième tableau affiche les importances des variables dans le modèle.

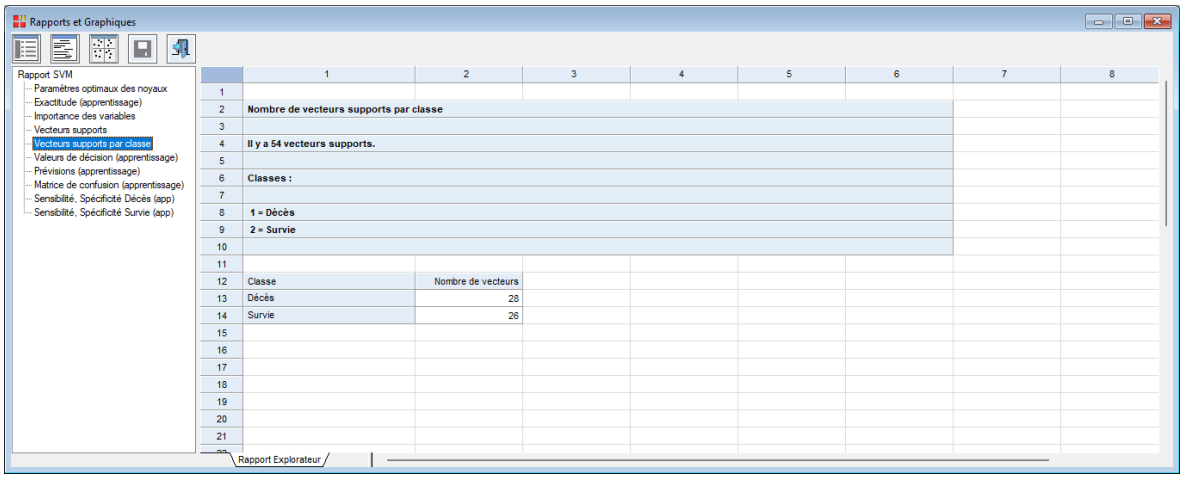


L’importance est basée sur la diminution de l'exactitude du modèle suite à la permutation aléatoire des données du jeu d'apprentissage dans chacune des variables explicatives.

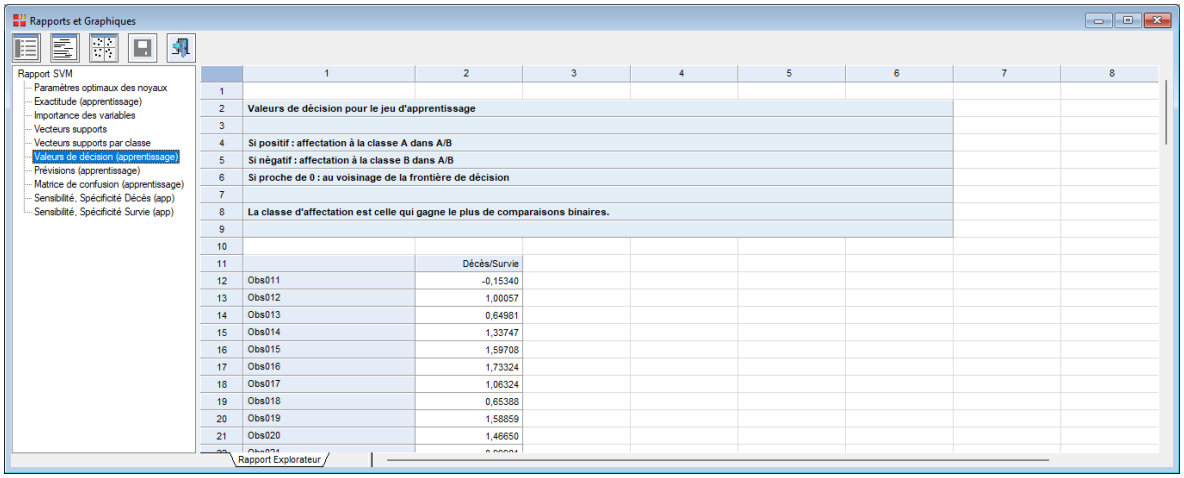
Le cinquième tableau affiche la liste des vecteurs supports. Un nombre pas trop important de vecteurs est souhaité.



Le sixième tableau indique le nombre de vecteurs supports par classe.



Le septième tableau affiche les valeurs de décision pour les données du jeu d’apprentissage.

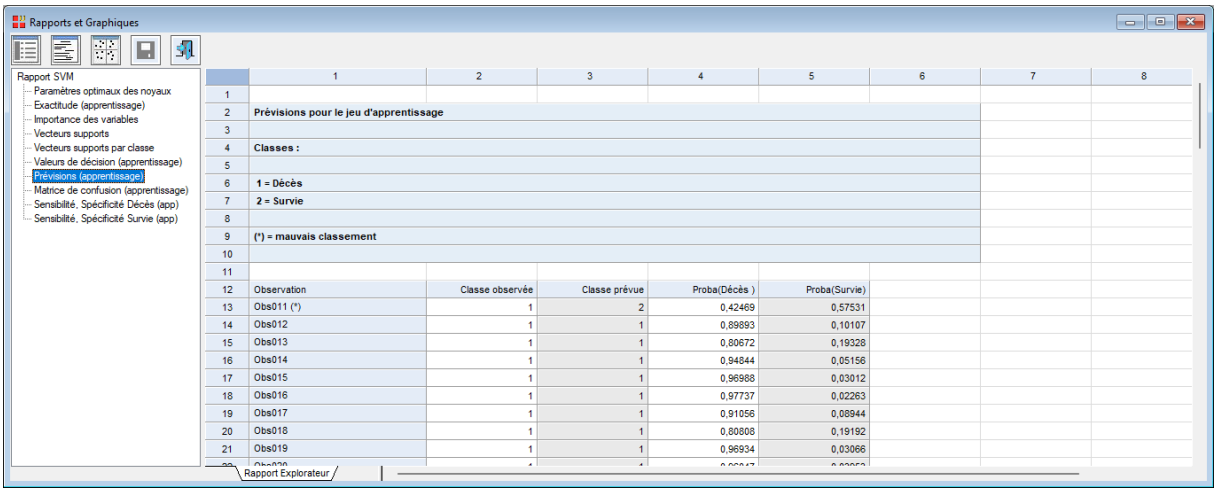


Si la valeur de décision est positive, l'observation est affectée à la classe A dans A/B.

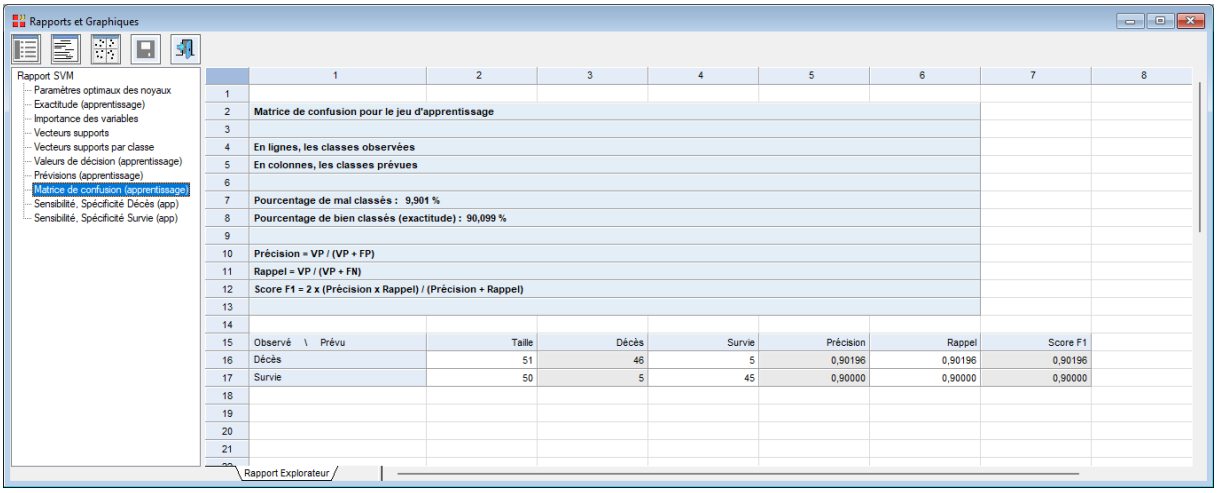
Si la valeur de décision est négative, l'observation est affectée à la classe B dans A/B.

Par exemple, l'observation 11 est affectée à la classe *Survie* et l'observation 12 à la classe *Décès*.

Le huitième tableau affiche les valeurs prévues de la variable à expliquer pour chacune des observations du jeu d'apprentissage.



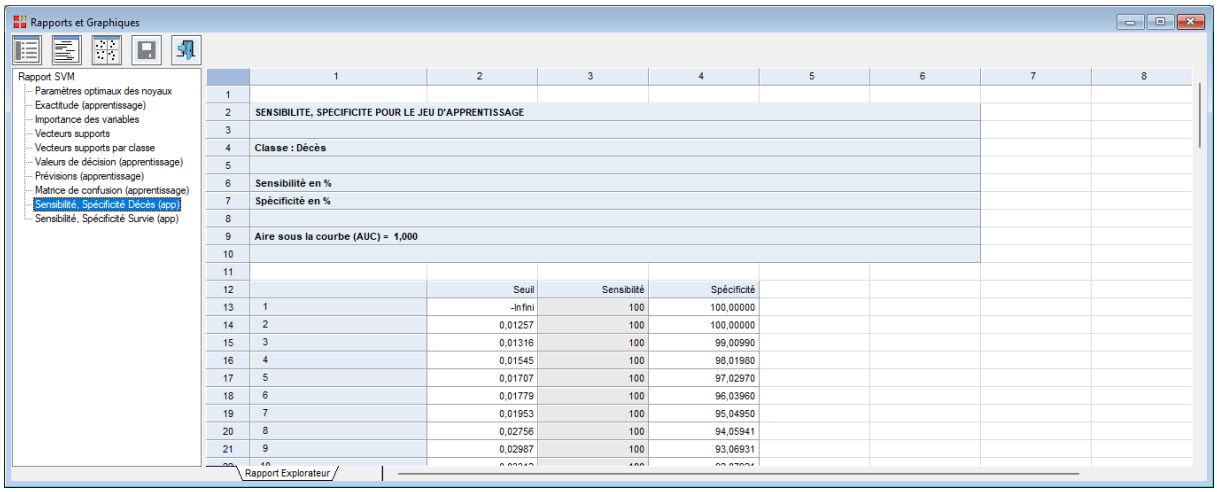
Le tableau suivant affiche la matrice de confusion.



Environ 90% des observations sont bien classées par le modèle construit.

Les deux tableaux suivants affichent les sensibilités et spécificités ainsi que l'aire sous la courbe ROC.

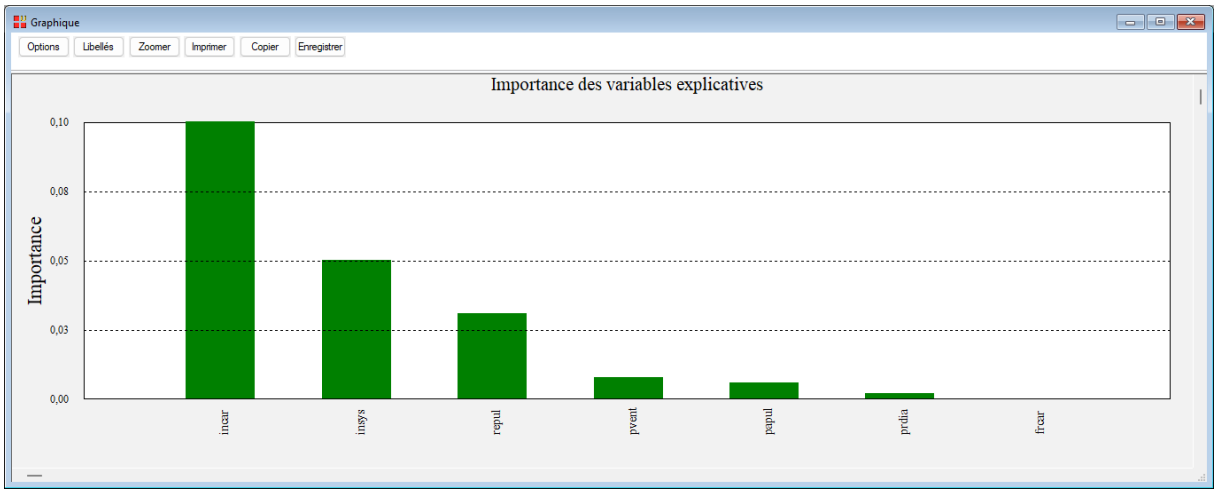
Ces tableaux ne sont disponibles que pour un classement binaire (deux modalités pour la variable à expliquer).



L'option Graphiques

Cette option permet d'obtenir divers graphiques pour l'analyse SVM.

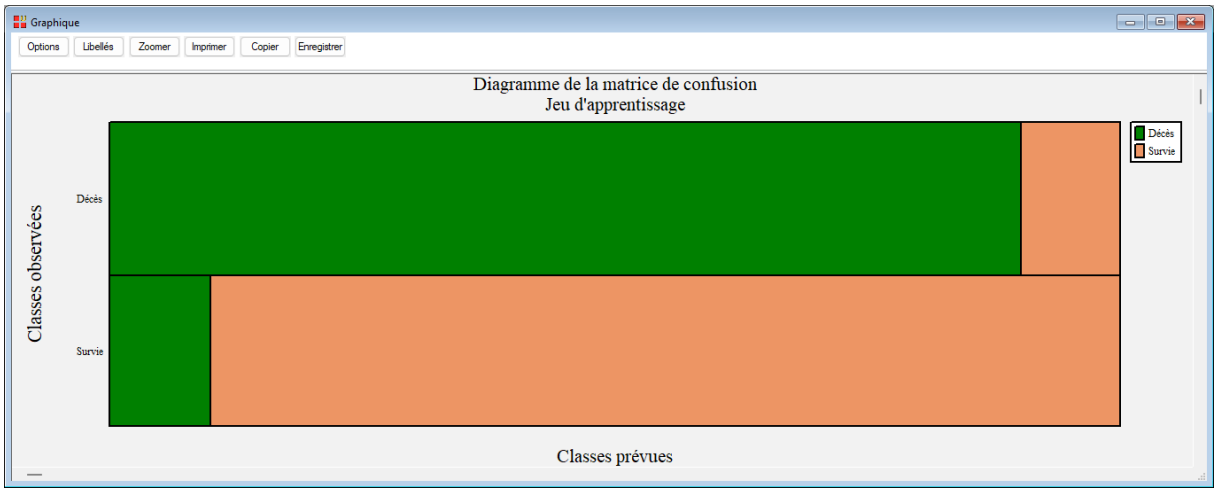
- Graphique de l'importance des variables



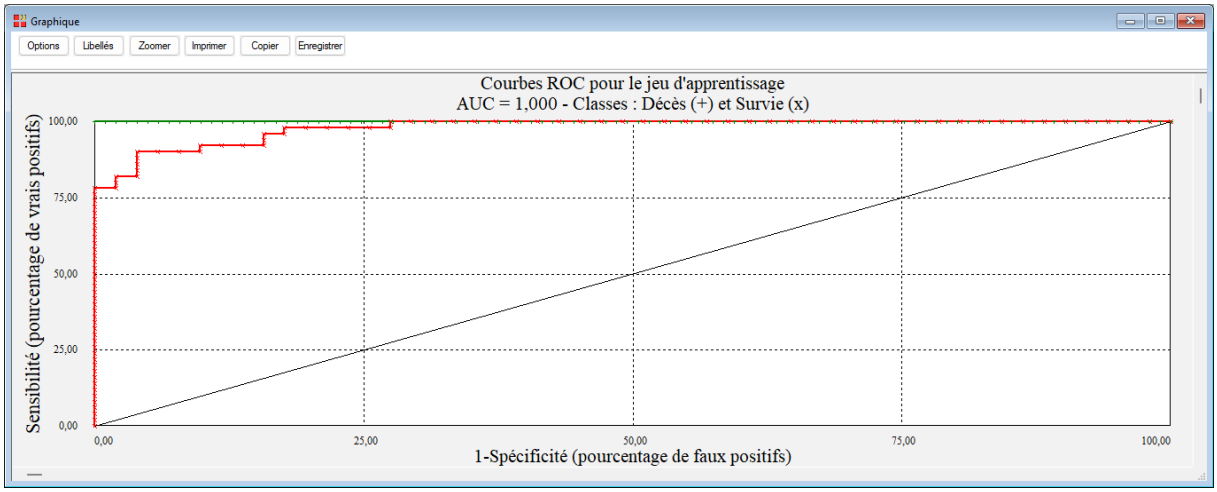
Les deux variables les plus importantes sont *incar* (index cardiaque) et *insys* (index systolique).

- Graphique de la matrice de confusion (apprentissage)

Ce graphique permet de visualiser les résultats obtenus dans le tableau 'Matrice de confusion'.



- Graphique de la courbe ROC (apprentissage)



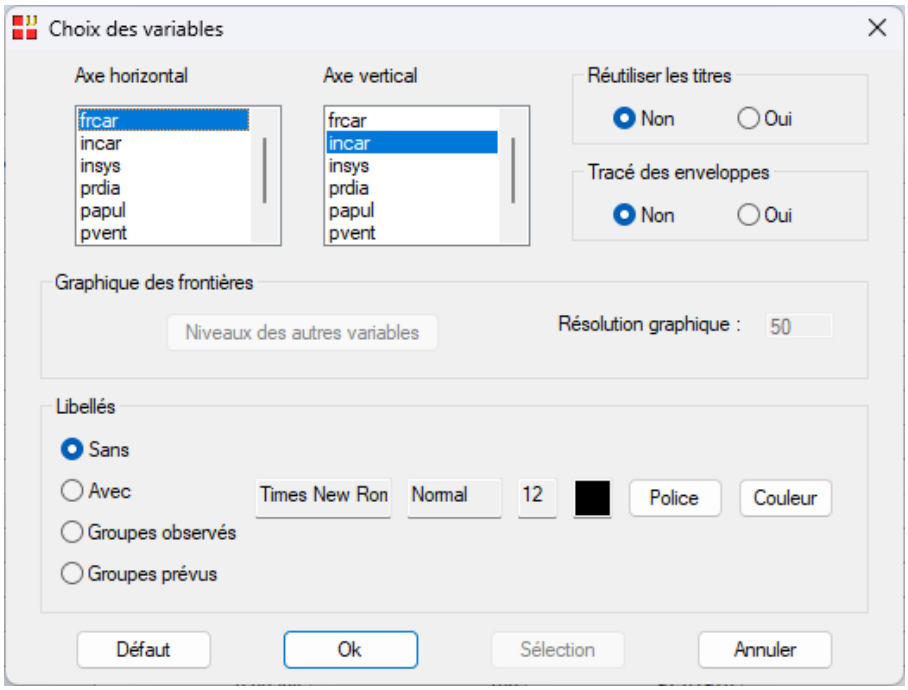
L'aire sous la courbe (AUC) nous indique l'efficacité du modèle construit. Plus la valeur de cette aire est élevée, meilleures sont les performances du modèle pour faire la distinction entre les classes *Survie* et *Décès*.

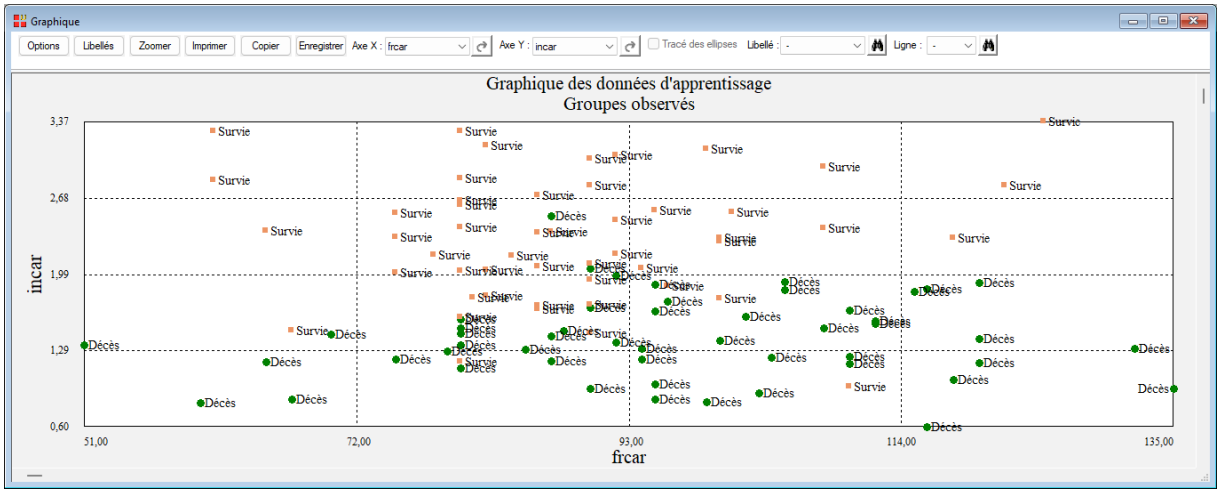
- Graphique Nuage de points (apprentissage)

Cette option permet d'afficher un nuage des observations pour deux variables sélectionnées.

Il est possible d'afficher les points avec ou sans libellés, avec les codes des groupes observés ou avec les codes des groupes prévus.

Dans le graphique, la barre d'outils permet de modifier les variables affichées.





- Graphique des frontières (apprentissage)

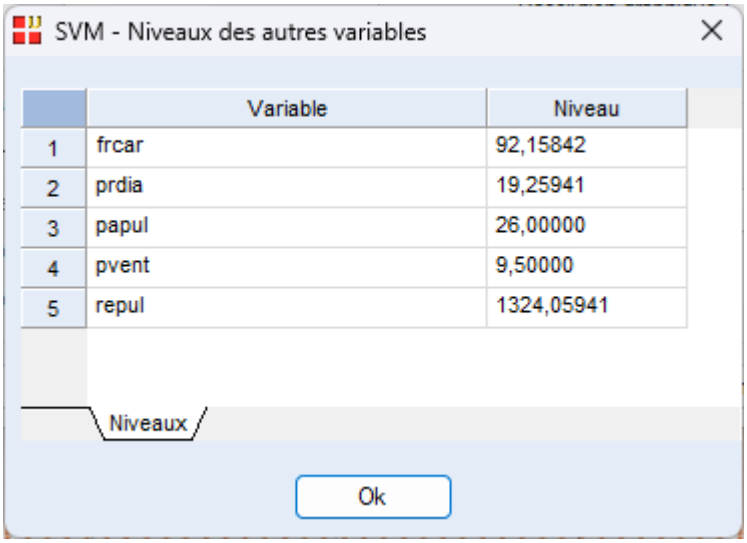
Cette option permet d'afficher un nuage des observations pour deux variables sélectionnées et de préciser les niveaux auxquels les autres variables sont maintenues.

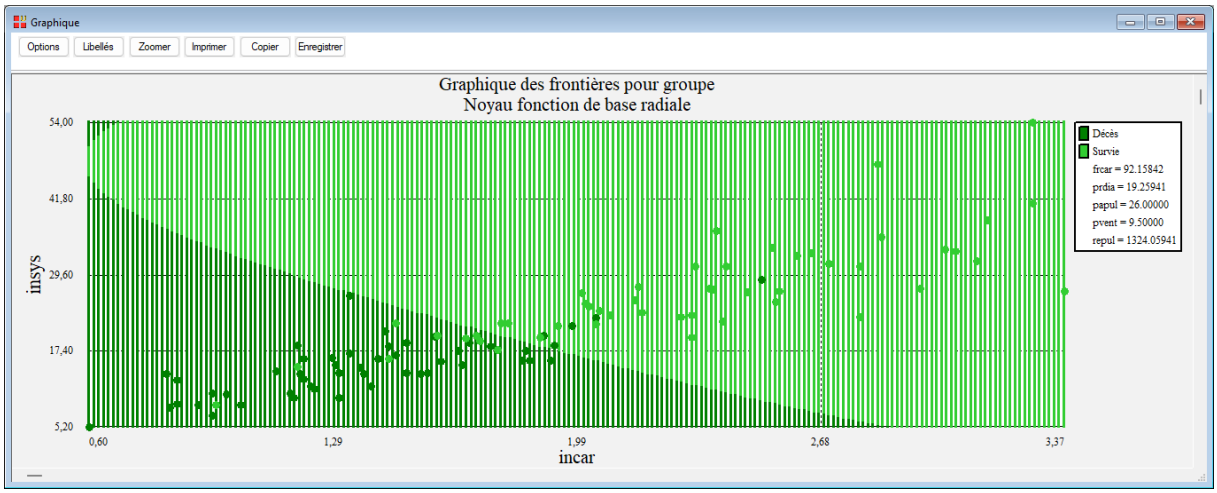
Par défaut les deux variables automatiquement sélectionnées sont les deux variables les plus importantes, à savoir *incar* et *insys*.

Pour cela, il faut cliquer sur le bouton 'Niveaux des autres variables' dans la boîte de dialogue 'Choix des variables'.

Par défaut les autres variables quantitatives sont maintenues aux valeurs moyennes.

L'option 'Résolution graphique' permet de préciser le nombre d'évaluations de chacune des variables.





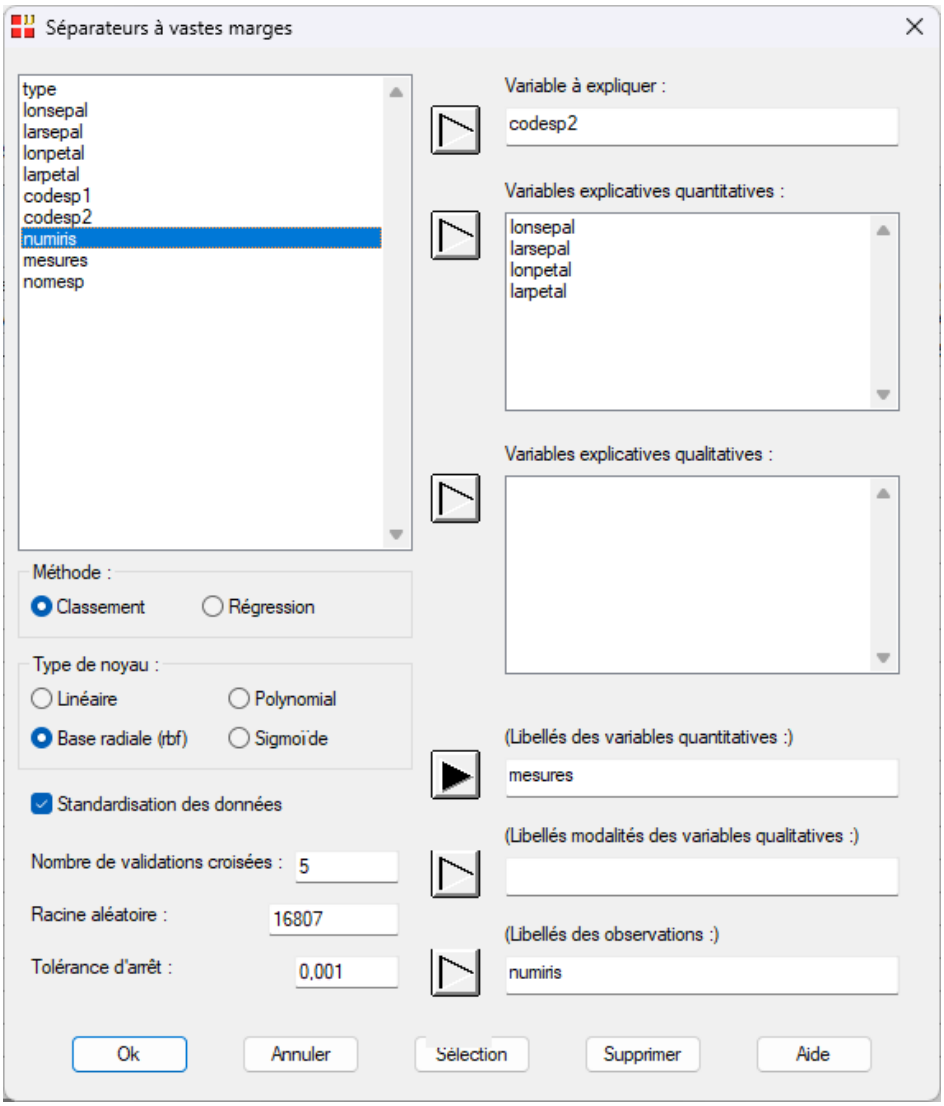
Exemple 2 : Classement multi-classes - Fichier Iris3

Nous utiliserons le fichier IRIS3 pour illustrer cette procédure. Ce fichier contient pour 150 iris de trois espèces différentes les mesures des quatre caractéristiques suivantes exprimées en millimètres : longueur du sépale, largeur du sépale, longueur du pétale, largeur du pétale. Les trois espèces sont : Setosa, Versicolor et Virginica.

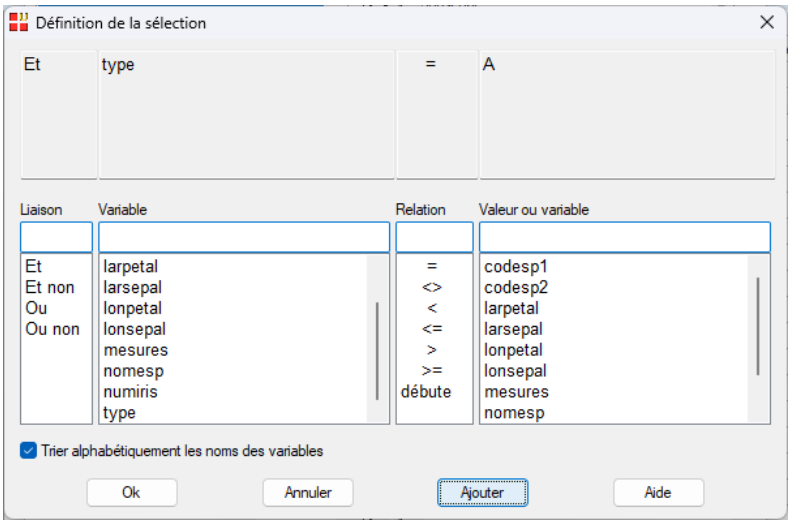
Ce fichier contient 6 iris pour lesquels les classes d'appartenance sont inconnues. Ils définiront le jeu de prévision.

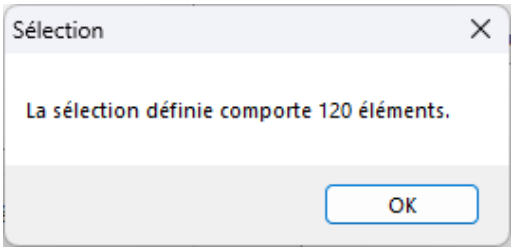
Sélectionnons la variable *codesp2* comme variable à expliquer, les variables *lonsepal* à *larpetal* comme variables explicatives, la variable *mesures* pour les libellées des variables explicatives et la variable *numiris* pour les libellés des observations.

Choisissons la méthode *Classement*, un noyau *Base radiale (rbf),* standardisons les données et laissons les autres paramètres aux valeurs par défaut.

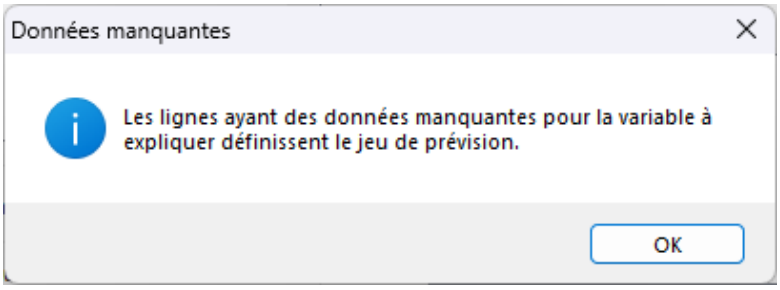


Cliquons sur le bouton *Sélection* pour définir le jeu d'apprentissage :

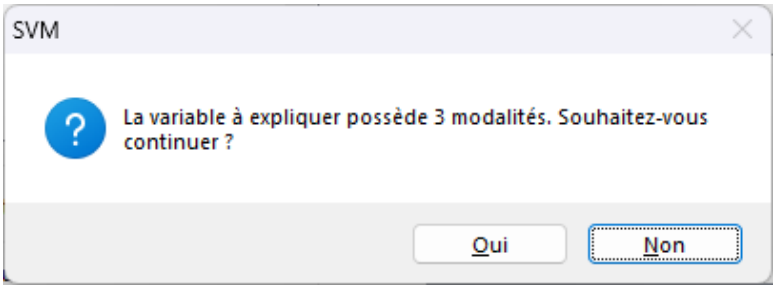




120 iris définissent le jeu d'apprentissage. Cliquons sur le bouton OK.

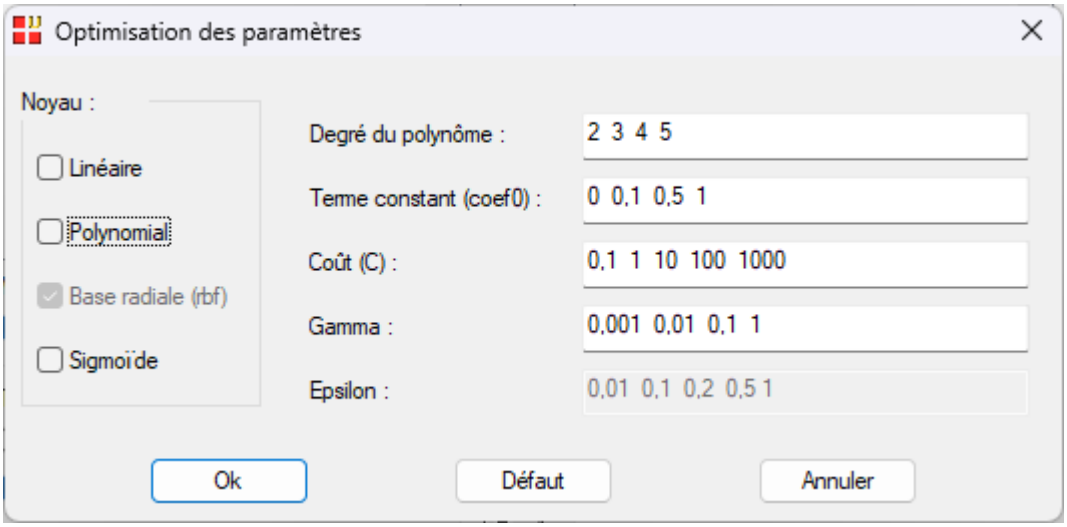


Un message nous informe que les lignes ayant des données manquantes pour la variable à expliquer définissent le jeu de prévision. Cliquons sur OK.



Confirmons que la méthode de classement utilise une variable à 3 modalités.

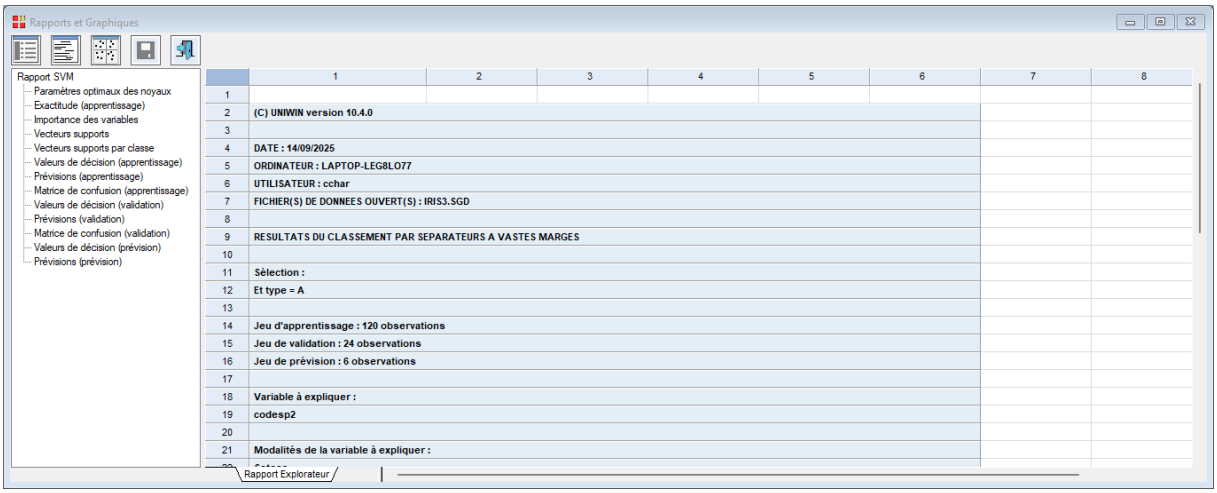
Une fenêtre nous demande ensuite de définir les options pour l'optimisation des paramètres de la méthode de classement :



Le noyau *Base radiale (rbf)* a été choisi par défaut. Il est toutefois possible de tester d'autres types de noyaux : *Linéaire*, *Polynomial* et *Sigmoïde*.

Cliquons sur le bouton *Ok*.

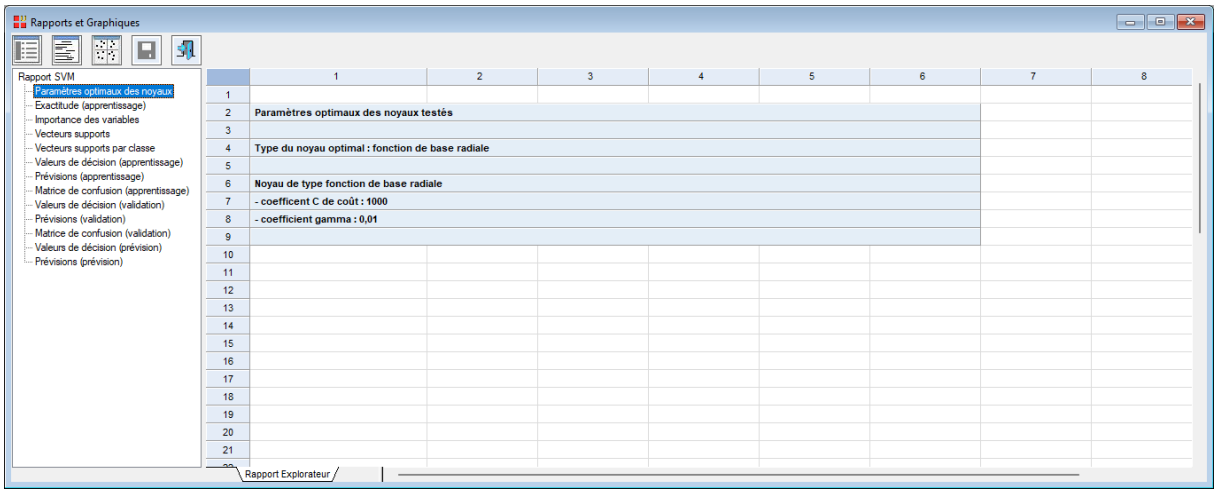
Après l'affichage de messages indiquant la progression des calculs, la fenêtre suivante s'affiche :



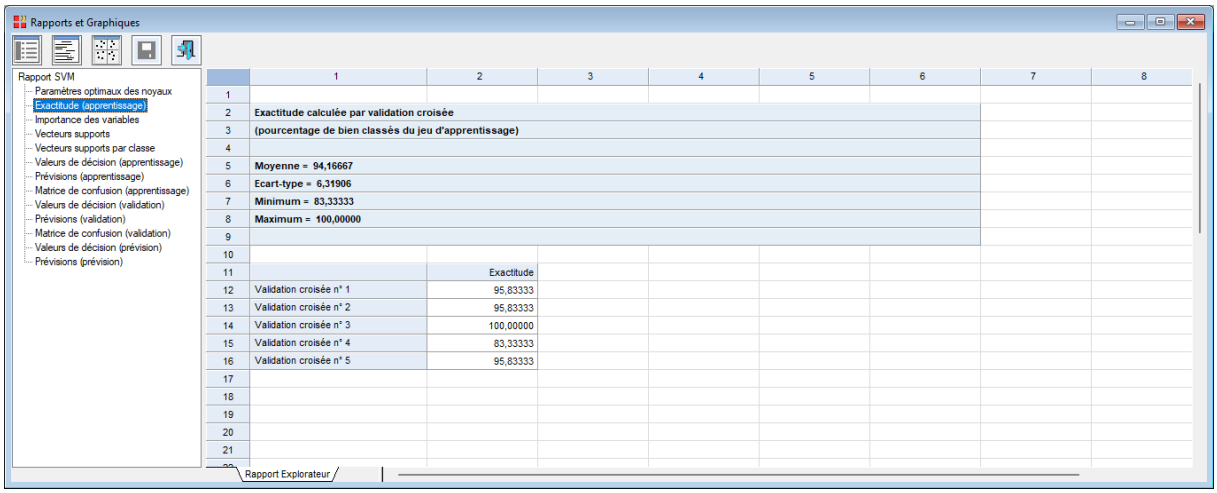
Le premier tableau, affiché ci-dessus, indique que le jeu d'apprentissage est composé de 120 observations, le jeu de validation de 24 observations et le jeu de prévision de 6 observations.

Il rappelle les modalités de la variable à expliquer, les variables explicatives utilisées et quelques paramètres de l'analyse.

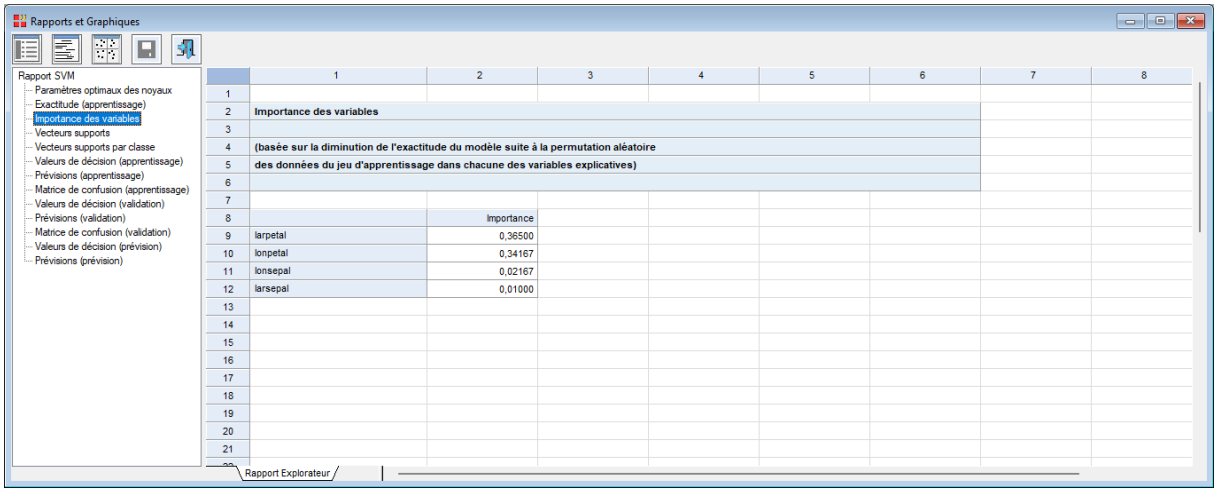
Le deuxième tableau indique les paramètres optimaux calculés pour le noyau *Base radiale (rbf)* : C = 1000 et Gamma = 0,01.



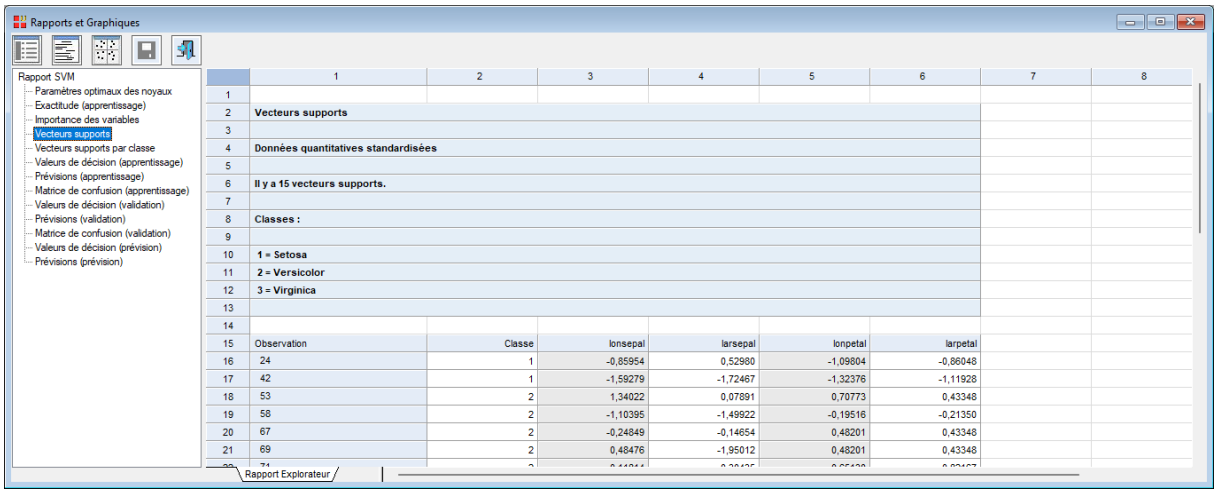
Le troisième tableau affiche l'exactitude calculée pour chacune des 5 validations croisées.



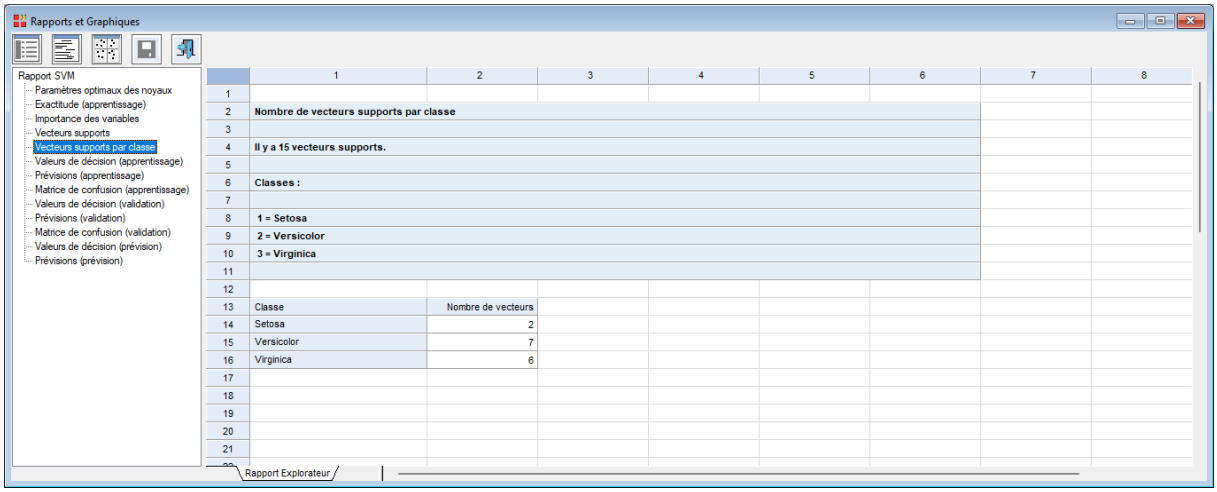
Le quatrième tableau affiche les importances des variables dans le modèle.



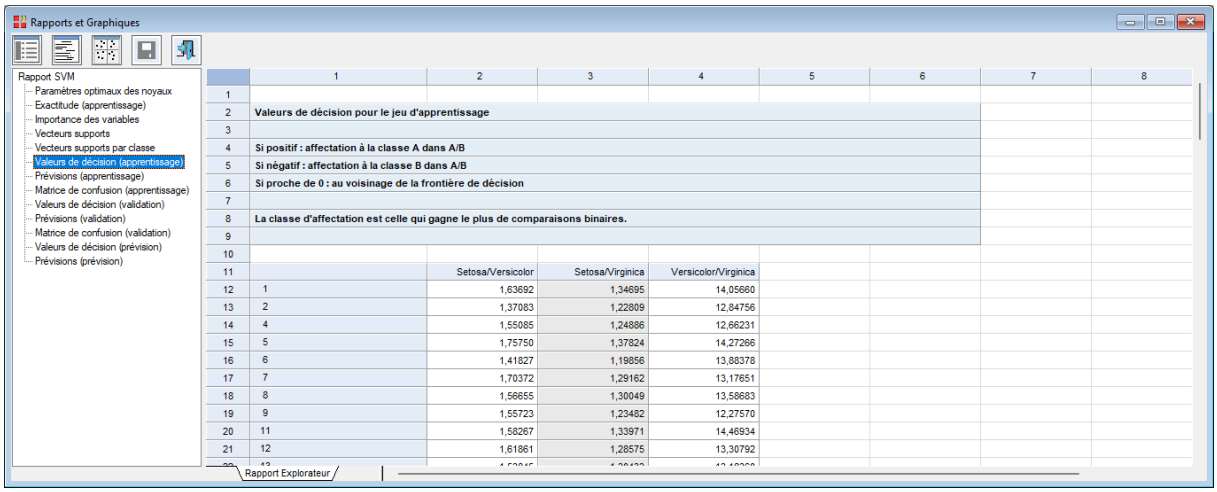
Le cinquième tableau affiche la liste des vecteurs supports. Un nombre pas trop important de vecteurs est souhaité.



Le sixième tableau indique le nombre de vecteurs supports par classe.



Le septième tableau affiche les valeurs de décision pour les données du jeu d'apprentissage.



Il y a 3 classes donc 3 * (3-1) / 2 = 3 comparaisons.

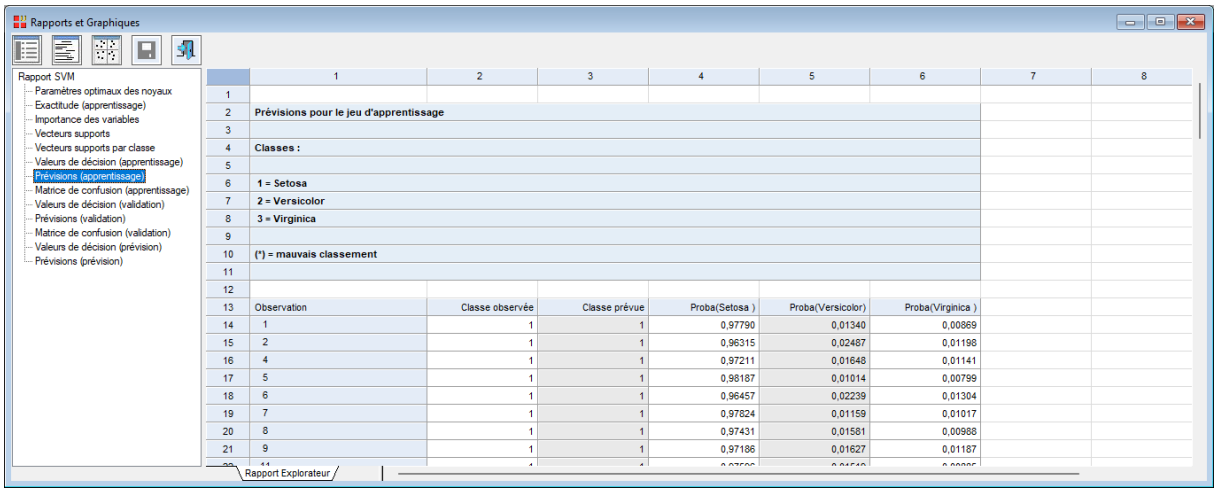
Si la valeur de décision est positive, l'observation est affectée à la classe A dans A/B.

Si la valeur de décision est négative, l'observation est affectée à la classe B dans A/B.

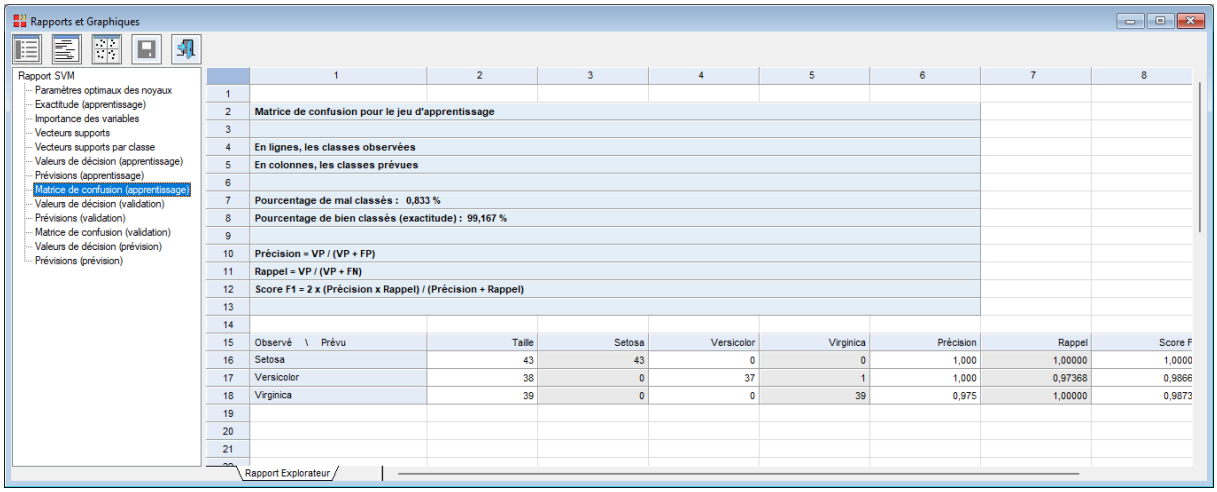
La classe finale est déterminée par un système de vote : la classe qui "gagne" le plus de comparaisons binaires est choisie.

Le huitième tableau affiche les prévisions pour le jeu d'apprentissage.

Par exemple, l'observation 1 (Setosa) est affectée à la classe 1 (Setosa) car cette classe a gagné 2 votes sur 3. Par contre l'observation 71 (Versicolor) est affectée à la classe 3 (Virginica) car cette classe a gagné 2 votes sur 3.

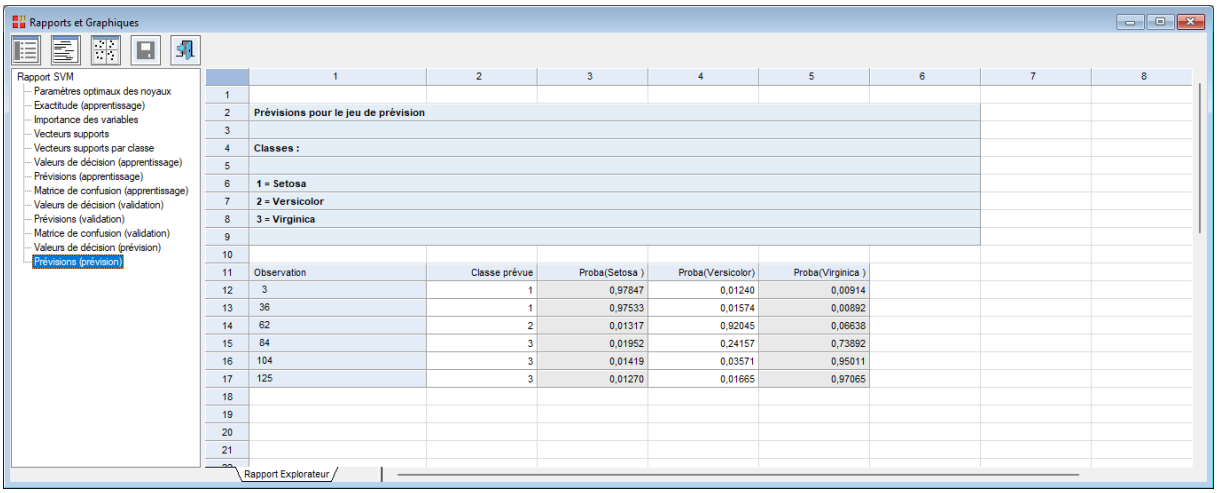


Le tableau suivant affiche la matrice de confusion pour le jeu d'apprentissage.

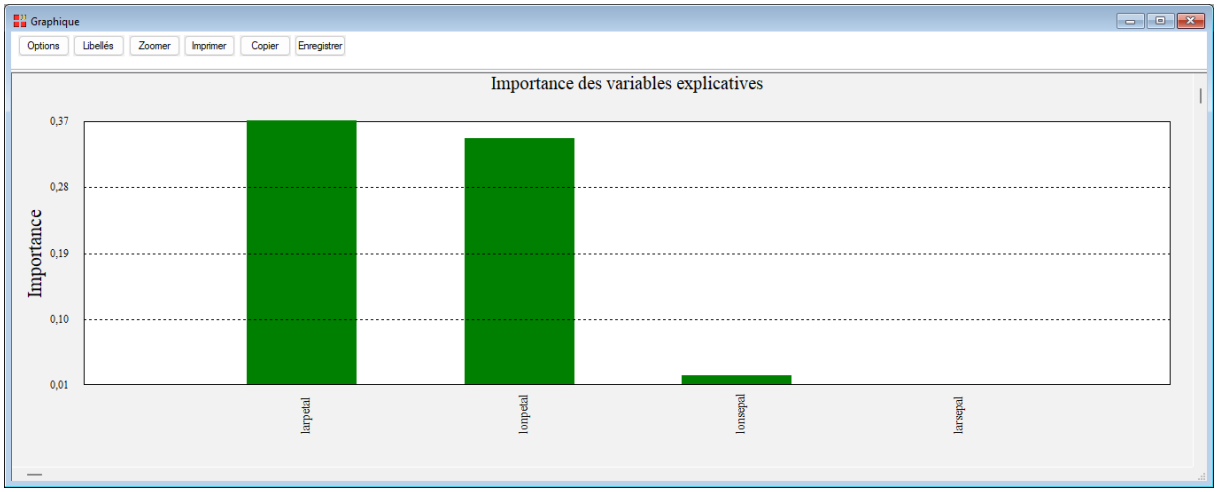


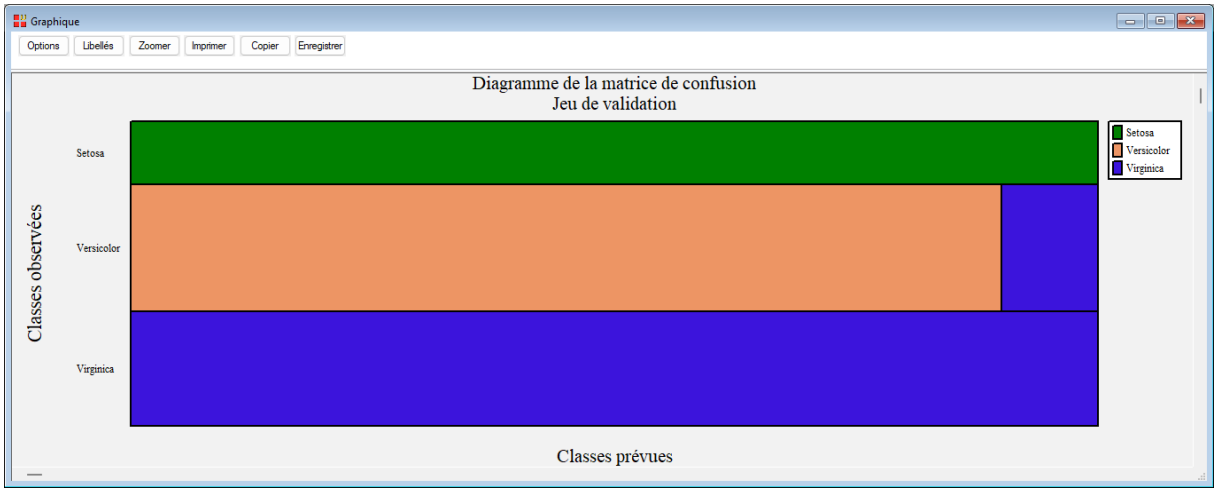
Les trois tableaux suivants affichent les valeurs de décision, les prévisions et la matrice de confusion pour le jeu de validation.

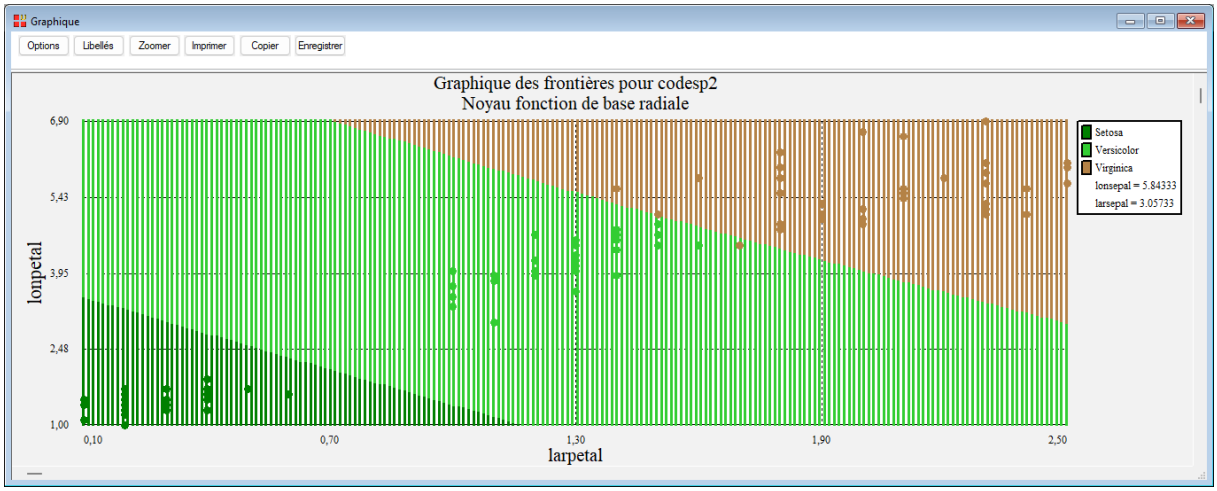
Enfin les deux derniers tableaux affichent les valeurs de décision et les prévisions pour le jeu de prévision.



Voici quelques graphiques obtenus par cette analyse.







Exemple 3 : Régression - Fichier Boston

Ce fichier contient des informations collectées par le « U.S Census Service » concernant la valeur des habitations dans l'agglomération de Boston (source : http://lib.stat.cmu.edu/datasets/boston).

Les caractéristiques contenues dans le fichier sont les suivantes :

crim	taux de criminalité par habitant par ville
zn	proportion de terrains résidentiels zonés pour des lots de plus de 25 000 pieds carrés.
indus	proportion d'acres non commerciales par ville
chas	indicatrice de proximité à la Charles River (1 = proche, 0 = sinon)
nox	concentration d'oxides nitriques (partie pour 10 millions)
rm	nombre moyen de pièces par habitation
age	proportion des habitations construites avant 1940
dis	distance pondérée à cinq lieux d'emplois de Boston
rad	indice d'accessibilité aux autoroutes
tax	taux d'imposition foncière par $ 10.000
ptratio	ratio élèves / enseignants par ville
black	1000(Bk-0,63)^2 où Bk = proportion de noirs par ville
lstat	% statut inférieur de la population
medv	valeur médiane de l'habitation (en milliers de $)
type	indique si l'observation est utilisée pour l'apprentissage (A) ou la validation (V)

Le but est de prévoir la variable *medv* en utilisant ces données.

Cliquons sur l'icône SVM dans le ruban Expliquer et renseignons la boîte de dialogue comme montré ci-après.

La variable à expliquer est *medv* et les variables explicatives sont *crim, zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, black, lstat*.

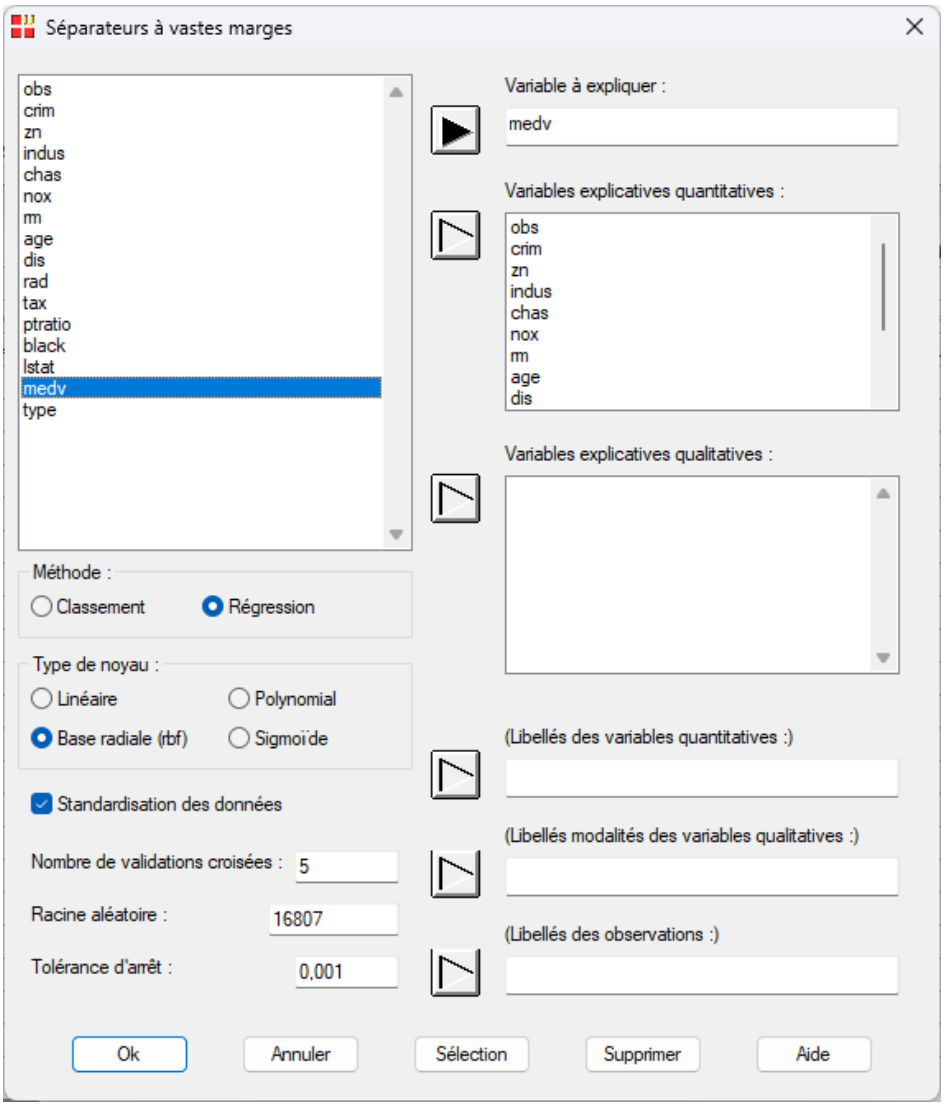
Cliquons sur le bouton 'Sélection' pour sélectionner les données du jeu d'apprentissage (type=A).

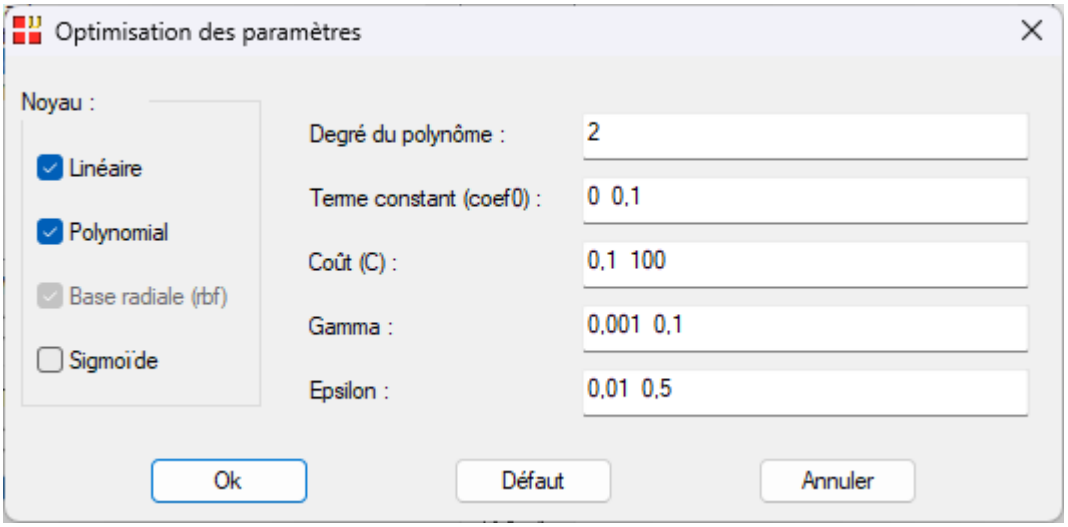
391 observations sont sélectionnées, 105 observations définissent le jeu de validation et 10 le jeu de prévision.

Cliquons sur Ok.

Une fenêtre nous demande ensuite de définir les options pour l'optimisation des paramètres de la méthode de régression.

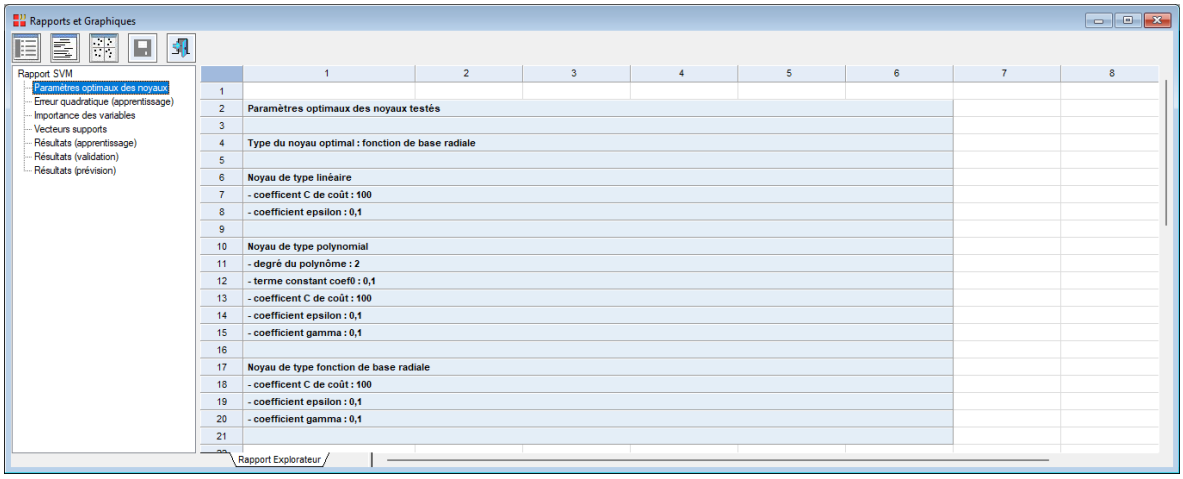
A noter que le paramètre *Epsilon* est maintenant actif.





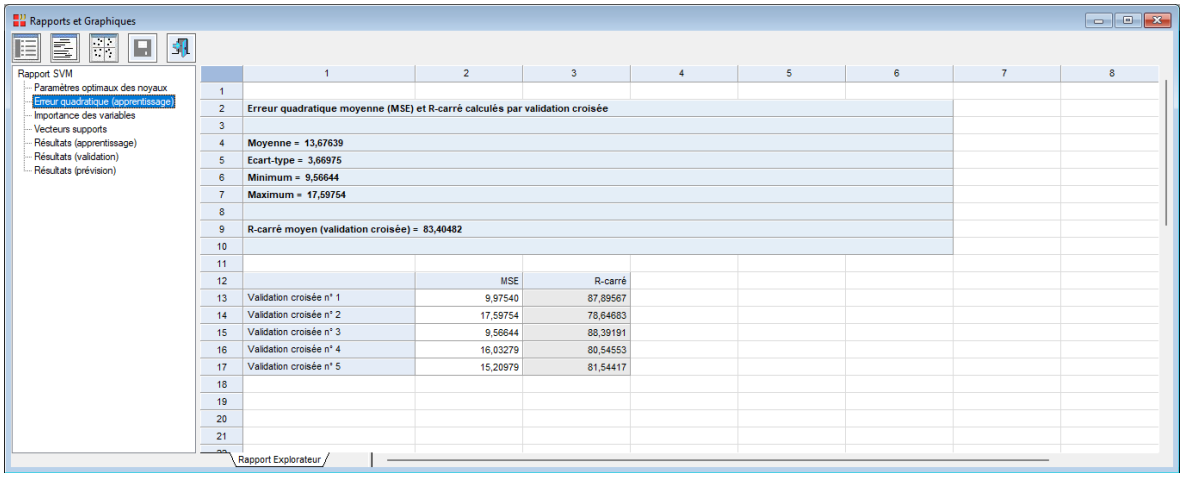
Trois types de noyaux sont testés : *Linéaire*, *Polynomial* et *Base radiale*.

Après attente pour l'exécution des divers modèles, visualisons la fenêtre affichant les paramètres optimaux des noyaux.

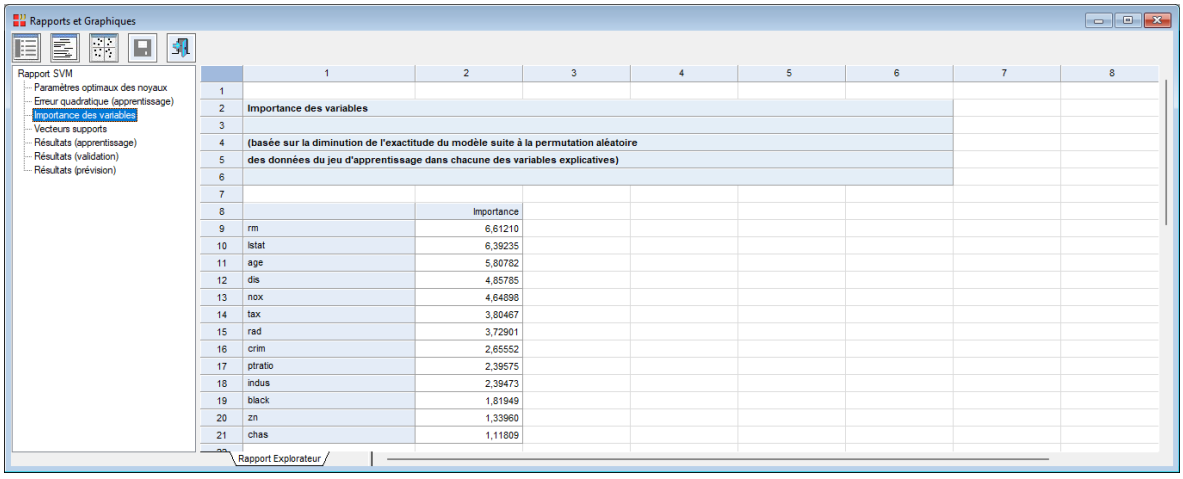


Un noyau *Base radiale* avec C = 100, Epsilon = 0,1 et Gamma = 0,1 est sélectionné.

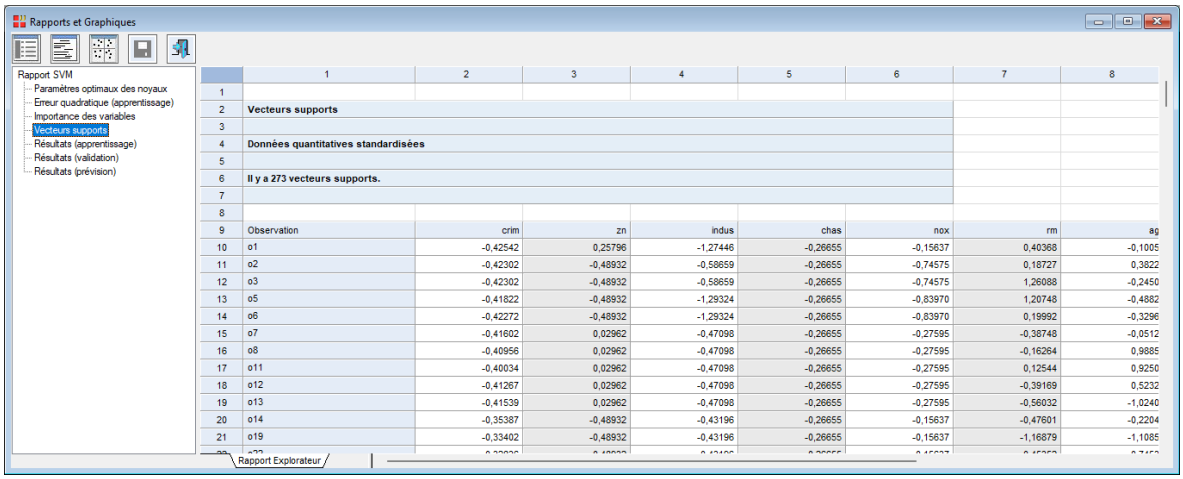
Un tableau affiche les erreurs quadratiques moyennes et les R-carrés calculés par validations croisées. Le R-carré moyen est de 83,4 %.



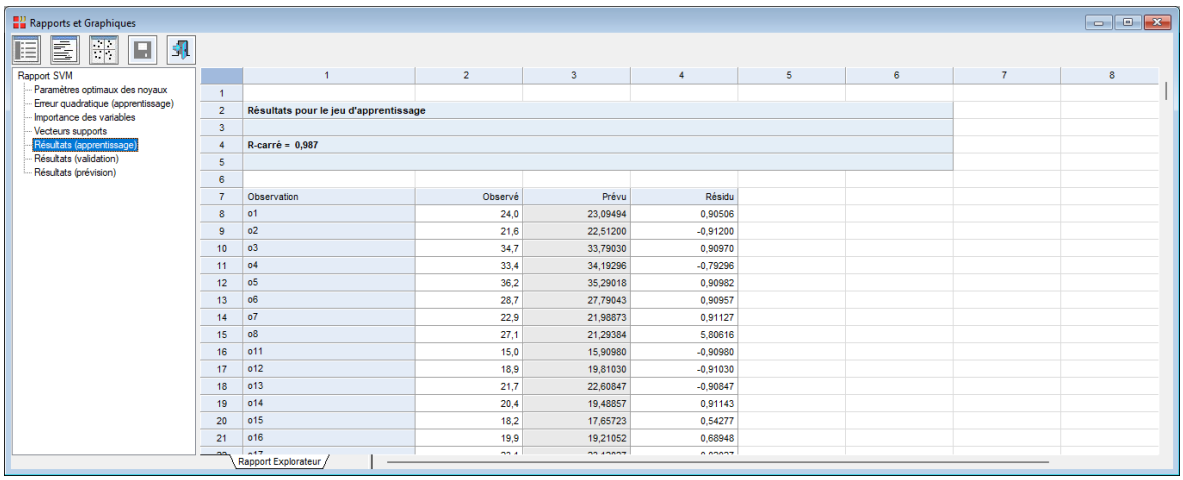
Un tableau affiche l'importance des variables explicatives.



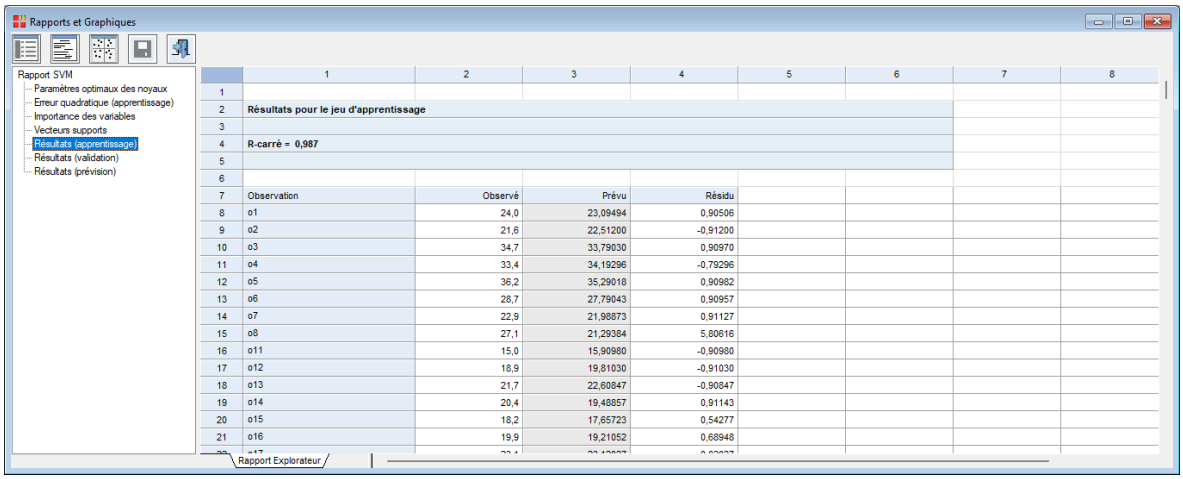
Le tableau suivant affiche les vecteurs supports.



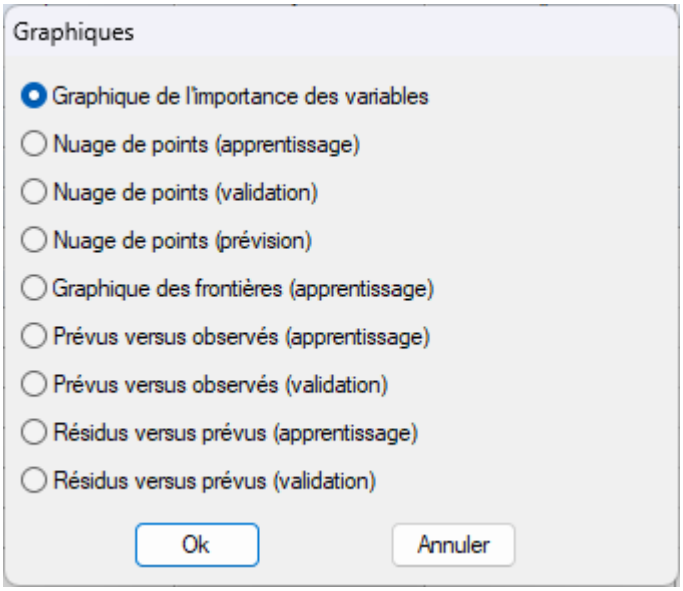
Les résultats pour les jeux d'apprentissage et de validation affichent les valeurs observées, prévues et les résidus ainsi que le R-carré.

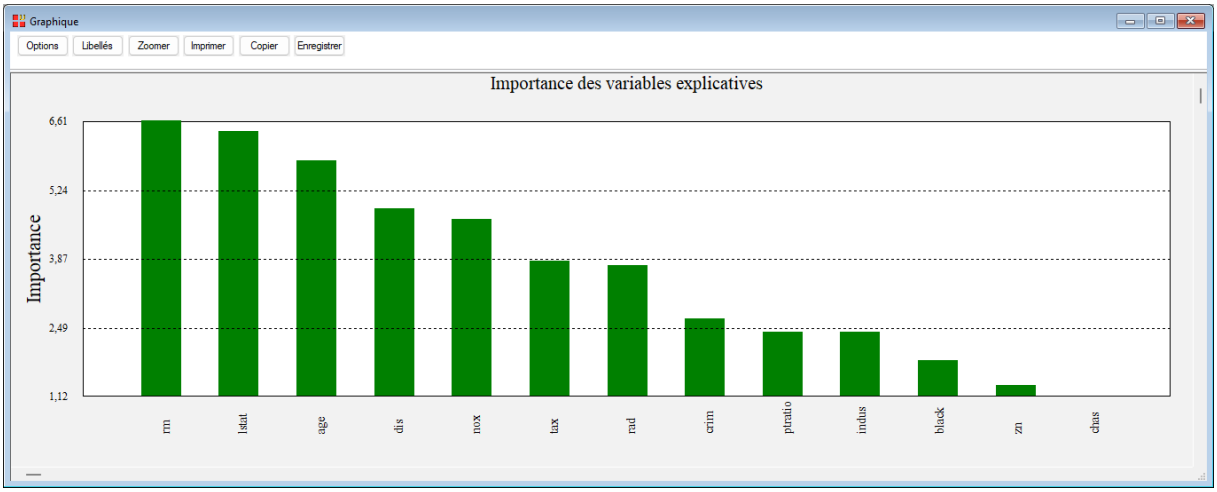


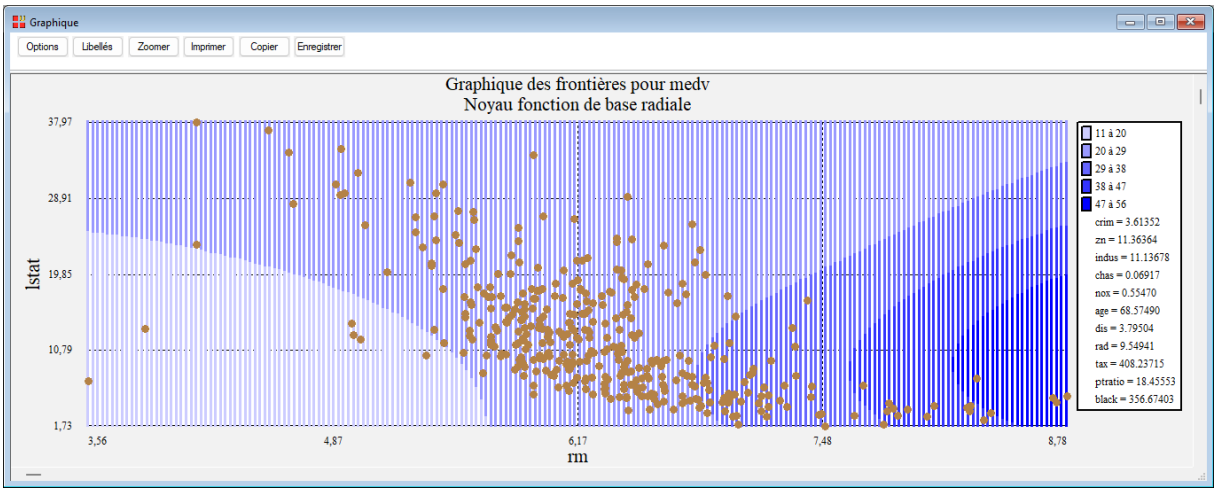
Enfin, un tableau affiche les valeurs prévues pour le jeu de prévision.

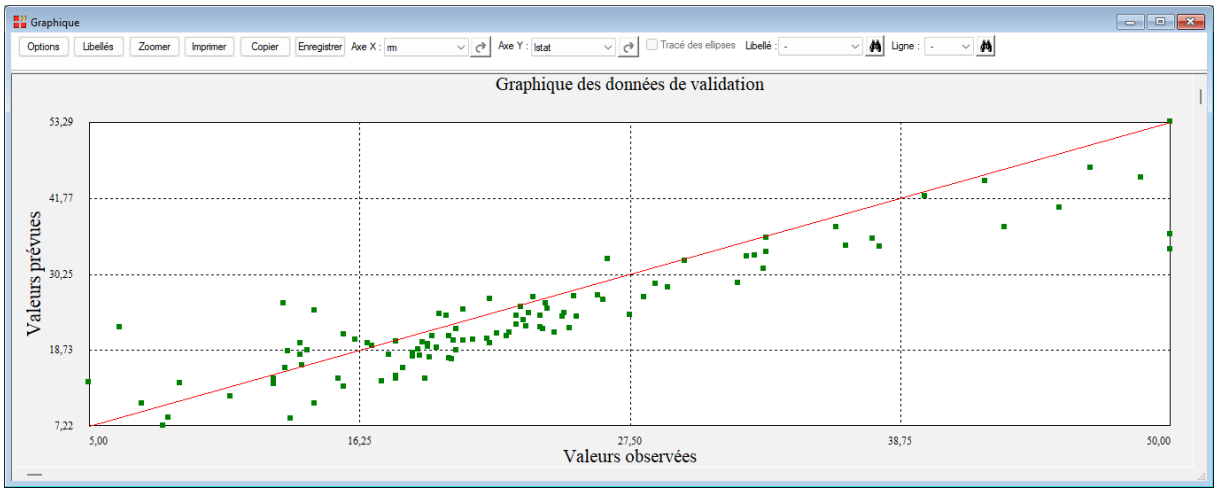


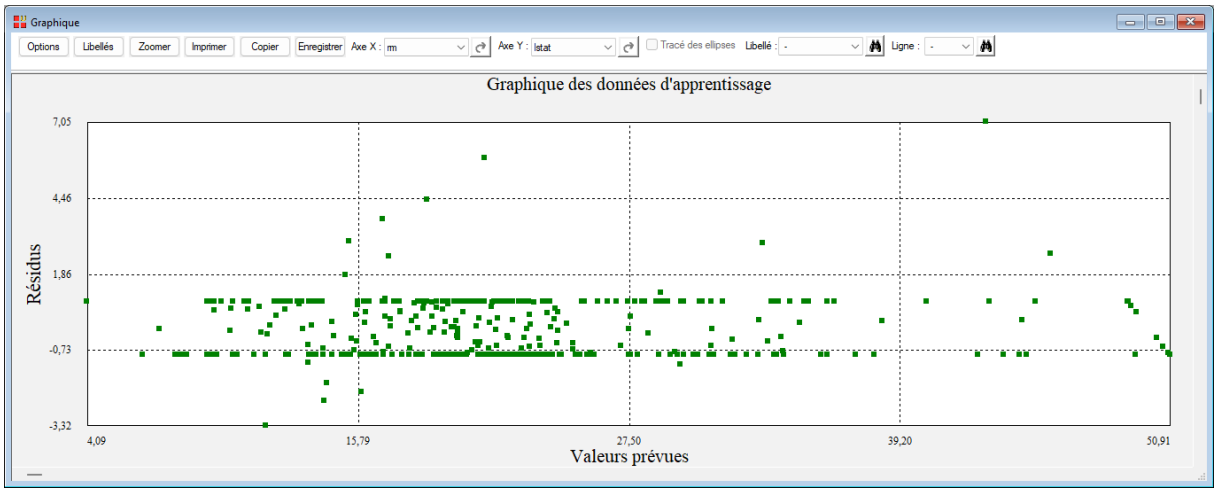
Visualisons les principaux graphiques :











Enregistrement des résultats

Voici la liste des variables créées par la procédure.

Variable	*Contenu*
libobsapp	Libellés des observations (apprentissage)
obsapp	Valeurs observées de la variable à expliquer (apprentissage)
prevapp	Valeurs prévues de la variable à expliquer (apprentissage)
residapp	Résidus pour le jeu d'apprentissage (régression)
seuilA	Seuils (apprentissage, classement binaire)
specificiteA	Spécificités (apprentissage, classement binaire)
sensibiliteA	Sensibilités (apprentissage, classement binaire)
aireA	Aires sous les courbes (apprentissage, classement binaire)
libobsvalid	Libellés des observations (validation)
obsvalid	Valeurs observées de la variable à expliquer (validation)
prevvalid	Valeurs prévues de la variable à expliquer (validation)

residapp	Résidus pour le jeu de validation (régression)
specificiteV	Spécificités (validation, classement binaire)
sensibiliteV	Sensibilités (validation, classement binaire)
aireV	Aires sous les courbes (validation, classement binaire)
libobsprev	Libellés des observations (prévision)
prevprev	Valeurs prévues de la variable à expliquer (prévision)

Formules des calculs

Noyaux

Les noyaux permettent de transformer l'espace des variables pour rendre les données linéairement séparables.

- Noyau linéaire

Le plus simple, utilisé quand les données sont déjà linéairement séparables :

K(x, y) = < x , y > où < x , y > est le produit scalaire de x et y

- Noyau polynomial

Permet de capturer des relations polynomiales entre les variables :

K(x, y) = [(gamma * < x , y >) + coef0]^d

où gamma = paramètre d'échelle, coef0 = terme constant, d = degré du polynôme

- Noyau base radiale (rbf)

Le plus populaire, efficace pour des frontières de décision complexes :

K(x , y) = exp(-gamma * ||x - y||²)

Plus gamma est grand, plus la frontière de décision sera complexe.

- Noyau sigmoïde

Inspiré des réseaux de neurones :

K(x, y) = tanh[(gamma *< x , y >) + coef0]

Considérations pratiques

Le choix du noyau dépend de vos données :

- Linéaire : pour des données simples et de grande dimension
- Base radiale (rbf) : choix par défaut pour la plupart des problèmes
- Polynomial : quand vous suspectez des relations polynomiales
- Sigmoïde : rarement utilisé en pratique

Paramètres des noyaux

	Base radiale (rbf)	Linéaire	Polynomial	Sigmoïde
Degré du polynôme			X	
Terme constant (coef0)			X	X
Coût (C)	X	X	X	X
Gamma	X		X	X
Epsilon	X	X	X	X

C : Le paramètre de régularisation C contrôle le compromis entre la complexité et la généralisation des modèles SVM. Une valeur C plus grande signifie une pénalité plus élevée pour les erreurs et une limite de décision plus complexe, tandis qu'une valeur C plus petite signifie une pénalité plus faible pour les erreurs et une limite de décision plus simple.

gamma : Le paramètre gamma contrôle la largeur et la forme de la fonction rbf et la limite de décision. Une valeur gamma plus grande signifie une fonction rbf plus étroite et plus pointue et une limite de décision plus complexe, tandis qu'une valeur gamma plus petite signifie une fonction rbf plus large et plus plate et une limite de décision plus simple.

Degré : Le degré contrôle le degré de la fonction polynomiale et la complexité de la limite de décision. Une valeur plus élevée signifie un polynôme de degré plus élevé et une limite de décision plus complexe, tandis qu'une valeur inférieure signifie un polynôme de degré inférieur et une limite de décision plus simple.

coef0 : Une valeur coef0 plus grande signifie une valeur à l'origine plus élevée et plus positive de la fonction polynomiale ou sigmoïde et de la limite de décision, tandis qu'une valeur plus petite signifie une valeur à l'origine inférieure et plus négative de la fonction polynomiale ou sigmoïde et de la limite de décision.

epsilon (uniquement pour la régression) : Le paramètre de marge epsilon pour la régression détermine la largeur et la flexibilité de la marge autour de la limite de décision. Une valeur epsilon plus grande signifie une marge plus large et plus flexible, tandis qu'une valeur epsilon plus petite signifie une marge plus étroite et plus rigide.

Références

Documentation du package R ‘e1071’
https://cran.r-project.org/web/packages/e1071/e1071.pdf

Vapnik V (1998). Statistical Learning Theory. Wiley, New York

Vapnik, V. (2000). The Nature of Statistical Learning Theory. Springer, 2ème édition.

Exemple ‘Boston’ : http://lib.stat.cmu.edu/datasets/boston