

UNIWIN VERSION 10.4.0

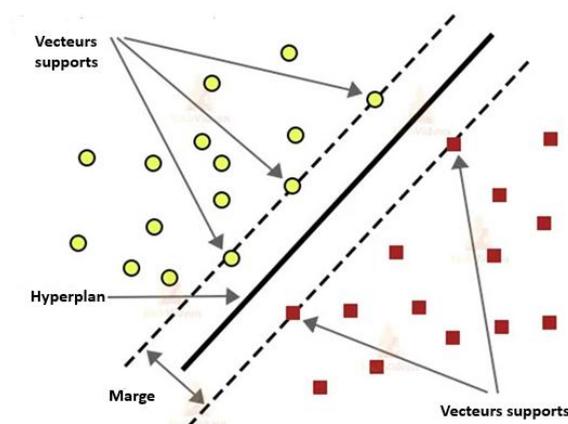
SEPARATEURS A VASTES MARGES

Révision : 15/09/2025

Définition.....	1
Entrée des données	2
Données manquantes	3
Exemple 1 : Classement binaire - Fichier Ann.....	4
Exemple 2 : Classement binaire - Fichier Moons	6
Exemple 3 : Classement binaire - Fichier Infarct2	7
L'option Rapports	10
L'option Graphiques	14
Exemple 2 : Classement multi-classes - Fichier Iris3	17
Exemple 3 : Régression - Fichier Boston	25
Enregistrement des résultats	30
Formules des calculs	31
Références	33

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 découpe 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.

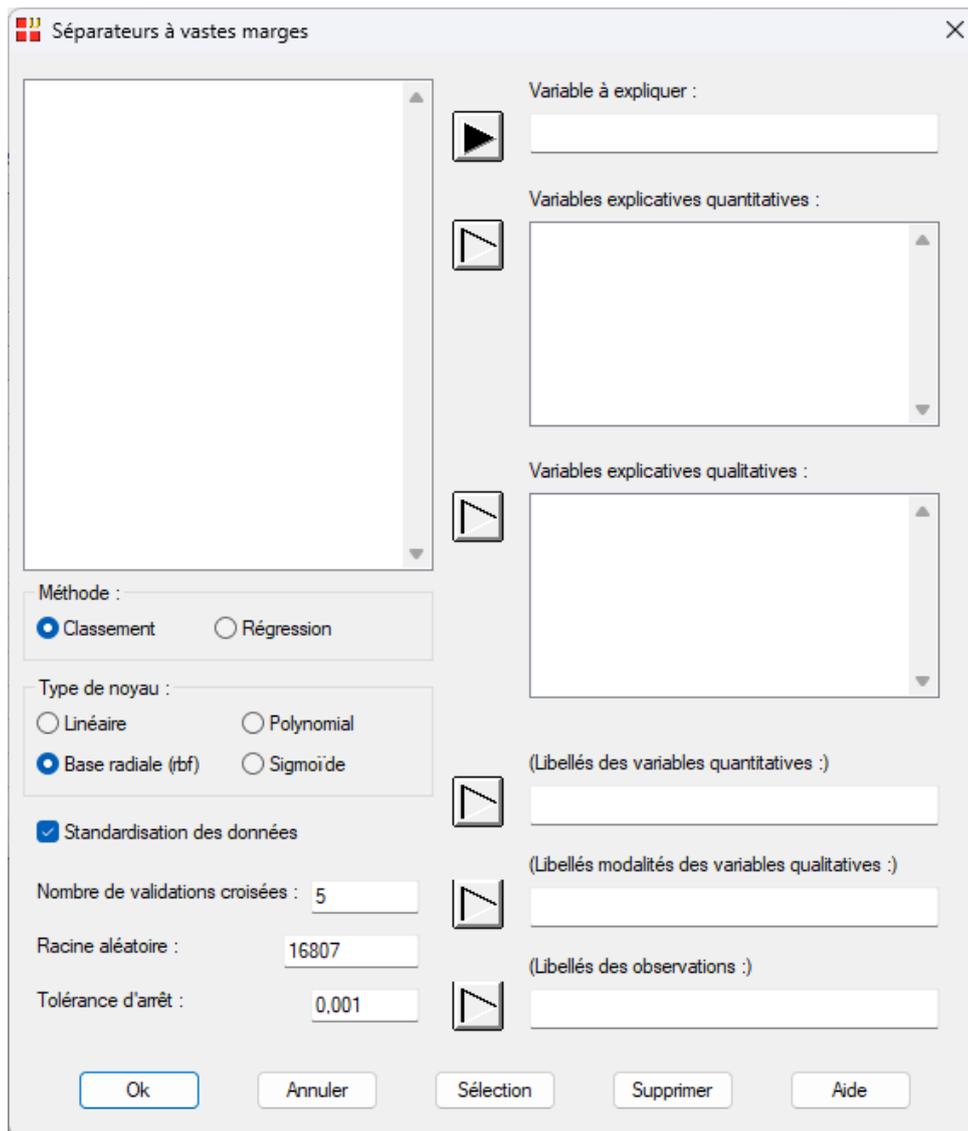
Elle permet également, en option, d'indiquer le nom de la variable contenant les libellés des variables quantitatives, le nom de la variable contenant les libellés des modalités des variables qualitatives ainsi que le nom de la variable contenant les libellés des observations.

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.

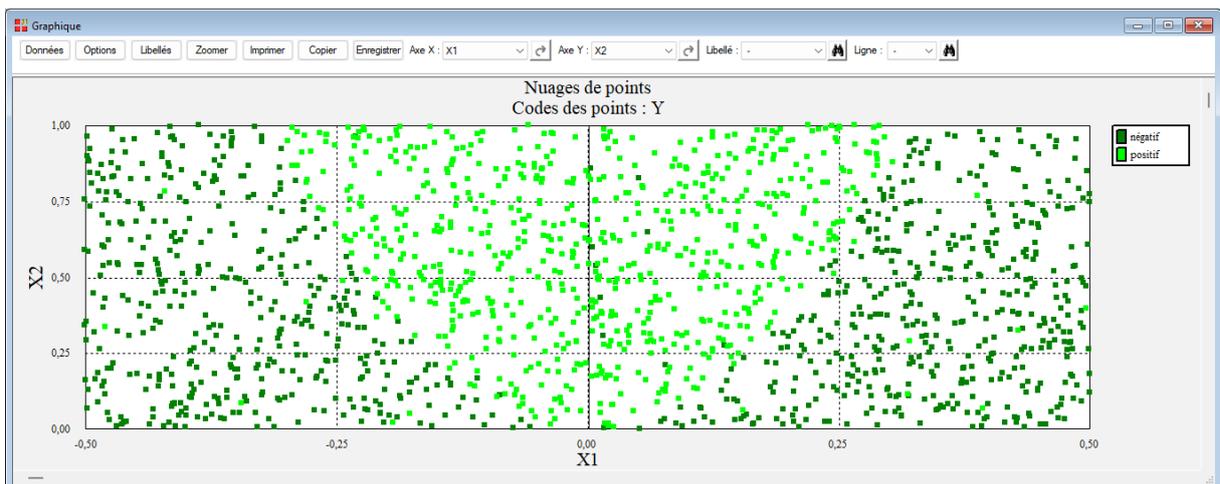
X1 et X2 sont deux variables aléatoires uniformes définies respectivement sur les intervalles $[-0,5 ; 0,5]$ et $[0 ; 1]$.

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

si $(0,1 * X2 > X1^2)$ 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.

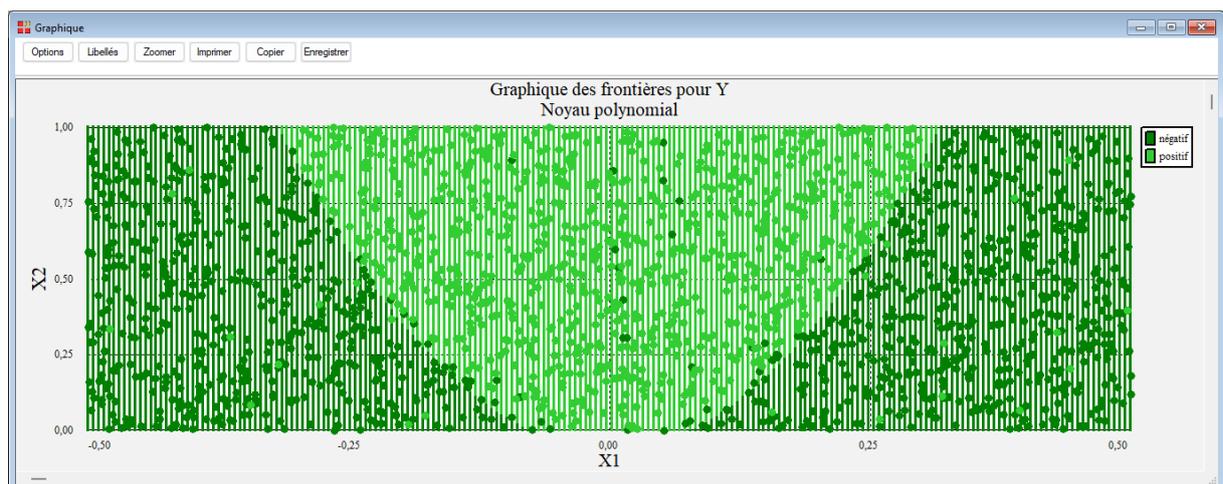
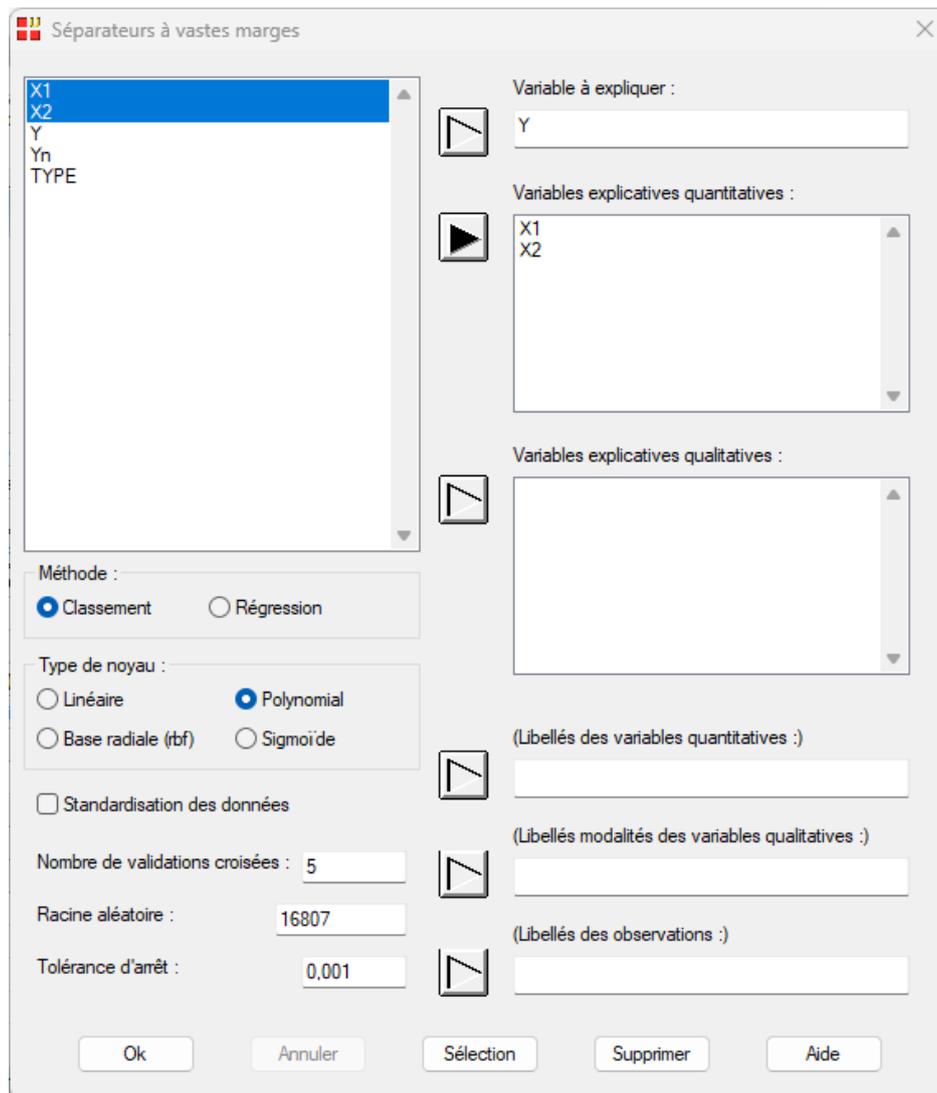


Cliquons sur l'icône SVM dans le ruban Expliquer et renseignons la boîte de dialogue comme montré ci-après.

Demandons un ajustement *Polynomial de degré 2*.

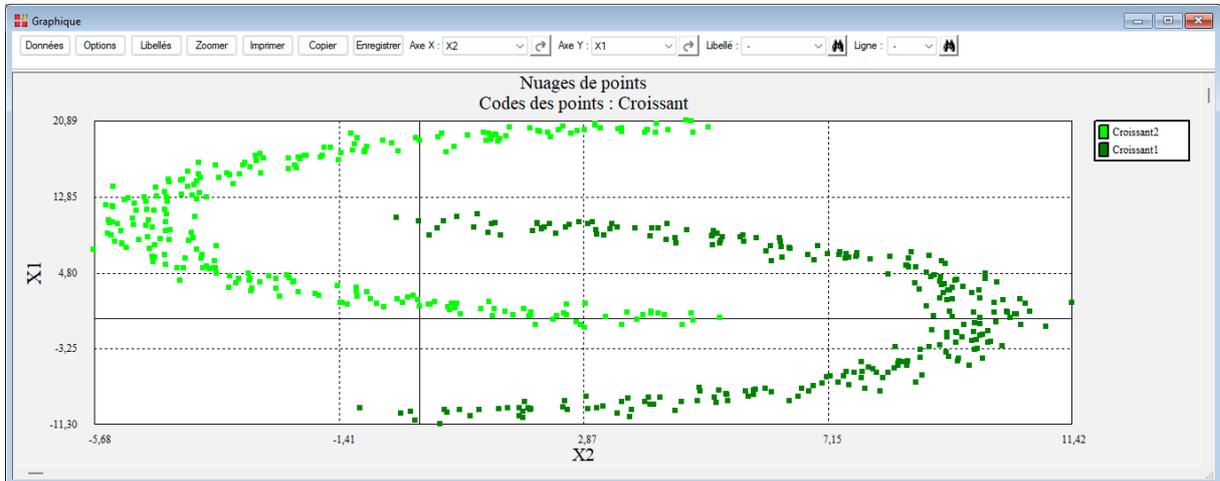
Après exécution des calculs, les paramètres optimaux obtenus sont :

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



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.

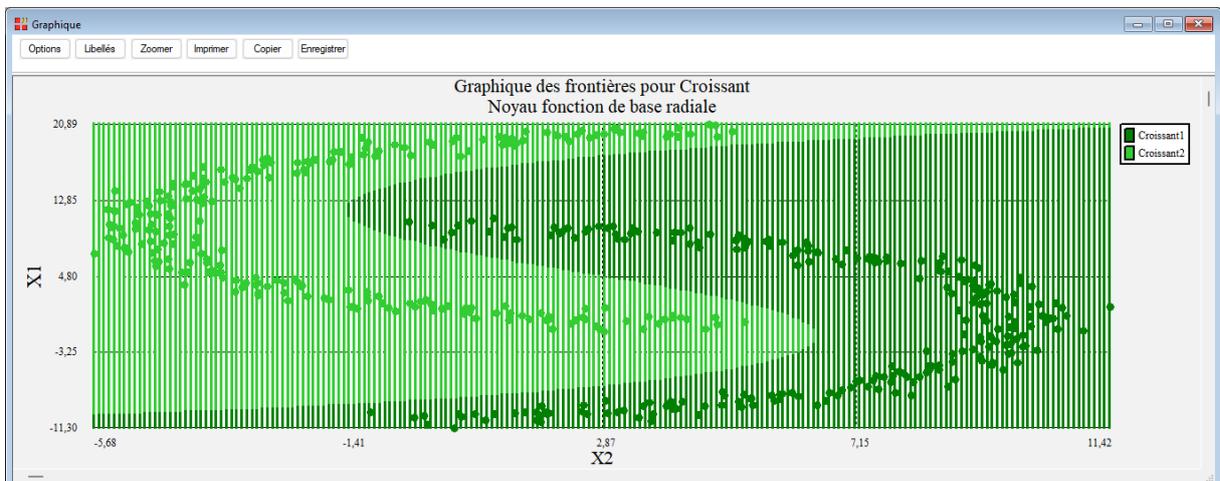


Cliquons sur l'icône SVM dans le ruban Expliquer et renseignons la boîte de dialogue en indiquant $Croissant$ comme variable à expliquer, $X1$ et $X2$ comme variables explicatives et ne standardisons pas les données.

Demandons un ajustement *Base radiale*.

Après exécution des calculs, les paramètres optimaux obtenus sont :

Noyau de type fonction de base radiale
- coefficient 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 :

<i>frcar</i>	fréquence cardiaque	<i>incar</i>	index cardiaque
<i>insys</i>	index systolique	<i>prdia</i>	pression diastolique
<i>papul</i>	pression artérielle pulmonaire	<i>pvent</i>	pression ventriculaire
<i>repul</i>	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.

Séparateurs à vastes marges

frcar
incar
insys
prdia
papul
pvent
repul
groupe
type
libobs

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

Variable à expliquer :
groupe

Variables explicatives quantitatives :
frcar
incar
insys
prdia
papul
pvent
repul

Variables explicatives qualitatives :

(Libellés des variables quantitatives :)

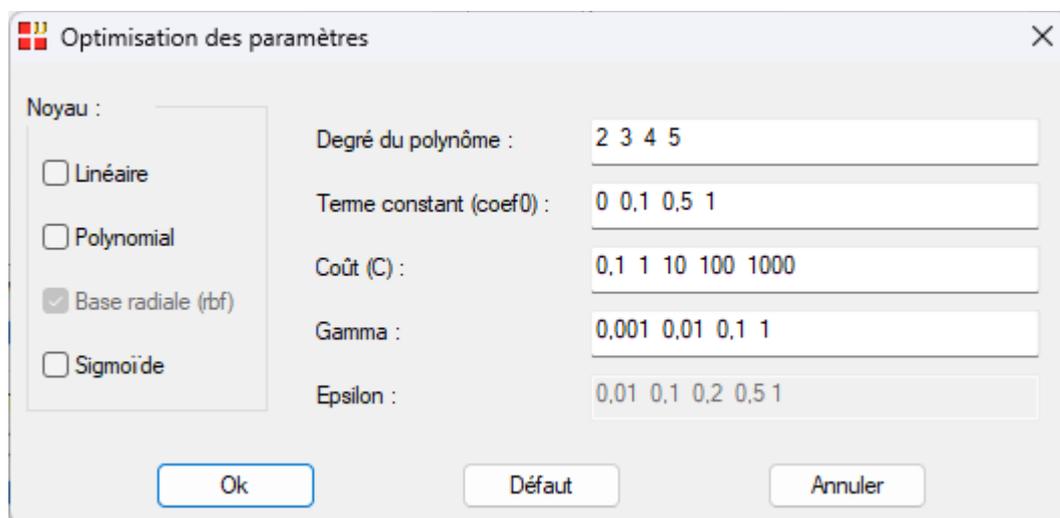
(Libellés modalités des variables qualitatives :)

(Libellés des observations :)
libobs

Ok Annuler Sélection Supprimer Aide

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.

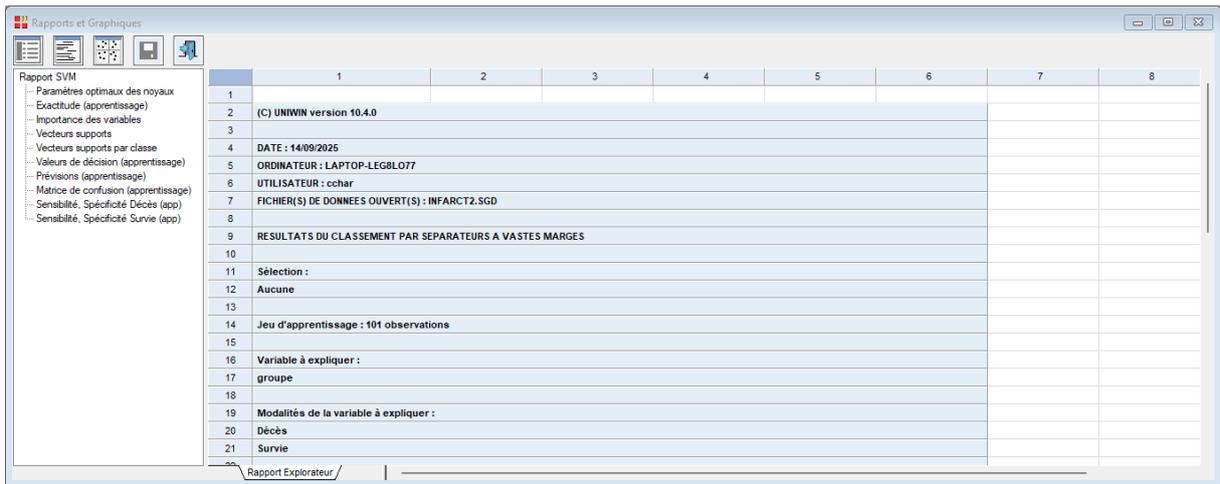
Voir le paragraphe '**Formules des calculs**' pour plus de détails sur les types de noyaux et les paramètres.

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'.

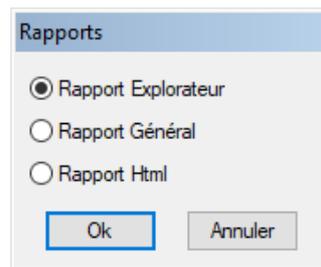
Cliquons sur *Défaut*.

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

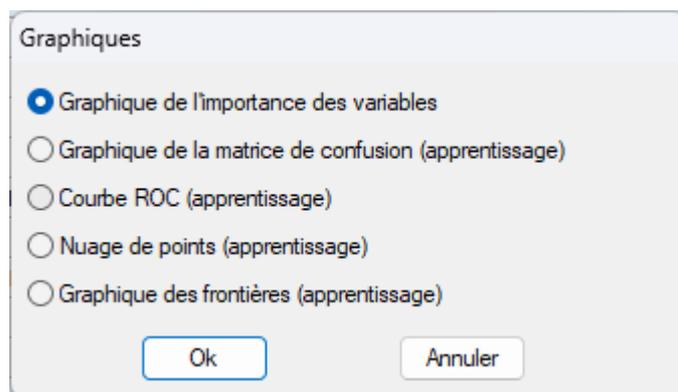


La barre d'outils 'Rapports et Graphiques' permet par l'icône 'Données'  de rappeler la boîte de dialogue d'entrée des données.

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.

Enregistrement des résultats (1/1)

Enregistrer

- Libellés des observations d'apprentissage
- Valeurs observées pour le jeu d'apprentissage
- Valeurs prévues pour le jeu d'apprentissage
- Seuils (apprentissage)
- Spécificités (apprentissage)
- Sensibilités (apprentissage)
- Aires sous les courbes (apprentissage)

Noms attribués aux variables cibles

libobsapp

obsapp

prevapp

seuilA

specificiteA_1

sensibiliteA_1

aireA

Ok Plus Tout Annuler

L'option Rapports

Cette option permet d'obtenir le rapport à l'écran sous la forme d'un explorateur, d'un tableau 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.

Rapports et Graphiques

Rapport SVM

- Paramètres optimaux des noyaux
- Exclure (apprentissage)
- Importance des variables
- Vecteurs supports
- Vecteurs supports par classe
- Valeurs de décision (apprentissage)
- Prévisions (apprentissage)
- Matrice de confusion (apprentissage)
- Sensibilité, Spécificité Décès (app)
- Sensibilité, Spécificité Survie (app)

	1	2	3	4	5	6	7	8
1								
2	(C) UNWIN version 10.4.0							
3								
4	DATE : 14/09/2025							
5	ORDINATEUR : LAPTOP-LEGL077							
6	UTILISATEUR : cchar							
7	FICHIER(S) DE DONNEES OUVERT(S) : INFARCT2.SGD							
8								
9	RESULTATS DU CLASSEMENT PAR SEPARATEURS A VASTES MARGES							
10								
11	Sélection :							
12	Aucune							
13								
14	Jeu d'apprentissage : 101 observations							
15								
16	Variable à expliquer :							
17	groupe							
18								
19	Modalités de la variable à expliquer :							
20	Décès							
21	Survie							

Rapport Explorateur /

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

The screenshot shows the 'Rapports et Graphiques' window with the 'Paramètres optimaux des noyaux testés' section selected. The table displays the following data:

	1	2	3	4	5	6	7	8
1								
2	Paramètres optimaux des noyaux testés							
3								
4	Type du noyau optimal : fonction de base radiale							
5								
6	Noyau de type fonction de base radiale							
7	- coefficient C de coût : 1							
8	- coefficient gamma : 0,142857142857143							
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								

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.

The screenshot shows the 'Rapports et Graphiques' window with the 'Exactitude calculée par validation croisée' section selected. The table displays the following data:

	1	2	3	4	5	6	7	8
1								
2	Exactitude calculée par validation croisée							
3	(pourcentage de bien classés du jeu d'apprentissage)							
4								
5	Moyenne = 86,04762							
6	Ecart-type = 9,00460							
7	Minimum = 75,00000							
8	Maximum = 95,23810							
9								
10								
11	Exactitude							
12	Validation croisée n° 1	95,00000						
13	Validation croisée n° 2	80,00000						
14	Validation croisée n° 3	85,00000						
15	Validation croisée n° 4	75,00000						
16	Validation croisée n° 5	95,23810						
17								
18								
19								
20								
21								

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

The screenshot shows the 'Rapports et Graphiques' window with the 'Importance des variables' section selected. The table displays the following data:

	1	2	3	4	5	6	7	8
1								
2	Importance des variables							
3								
4	(basée sur la diminution de l'exactitude du modèle suite à la permutation aléatoire							
5	des données du jeu d'apprentissage dans chacune des variables explicatives)							
6								
7								
8	Importance							
9	incar	0,10495						
10	insys	0,05347						
11	repu	0,03366						
12	pvent	0,00990						
13	pspul	0,00792						
14	prdia	0,00396						
15	frcar	0,00198						
16								
17								
18								
19								
20								
21								

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.

Observation	Classe	frcar	incar	insys	prdia	papul	pvei
15 Obs011	1	0,78167	-0,00871	-0,37625	-0,21679	-0,54626	0,1151
16 Obs012	1	-0,31399	-0,97986	-0,79608	2,53740	2,04846	2,4187
17 Obs013	1	-1,65313	-0,99503	-0,28547	-0,73320	-1,09251	0,8062
18 Obs018	1	-1,34878	-0,61588	-0,02449	-0,04465	0,06828	0,3455
19 Obs021	1	-1,04443	-0,96468	-0,53510	-0,04465	-0,27313	-1,2689
20 Obs023	1	1,45124	0,00646	-0,55780	2,36526	2,18503	0,8062
21 Obs024	1	-0,00964	0,18855	0,06628	-0,21679	0,13656	-1,4973

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

Classe	Nombre de vecteurs
Décès	28
Survie	28

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

Observation	Décès/Survie
12 Obs011	-0,15340
13 Obs012	1,00057
14 Obs013	0,64981
15 Obs014	1,33747
16 Obs015	1,59708
17 Obs016	1,73324
18 Obs017	1,06324
19 Obs018	0,65388
20 Obs019	1,58859
21 Obs020	1,48650

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.

The screenshot shows a software window titled 'Rapports et Graphiques' with a sidebar menu on the left. The main area displays a table with 8 columns and 21 rows. The table content is as follows:

	1	2	3	4	5	6	7	8
1								
2	Prévisions pour le jeu d'apprentissage							
3								
4	Classes :							
5								
6	1 = Décès							
7	2 = Survie							
8								
9	(*) = mauvais classement							
10								
11								
12	Observation	Classe observée	Classe prévue	Proba(Décès)	Proba(Survie)			
13	Obs011 (*)	1	2	0,42469	0,57531			
14	Obs012	1	1	0,89893	0,10107			
15	Obs013	1	1	0,80672	0,19328			
16	Obs014	1	1	0,84844	0,05156			
17	Obs015	1	1	0,96968	0,03012			
18	Obs016	1	1	0,97737	0,02263			
19	Obs017	1	1	0,91056	0,08944			
20	Obs018	1	1	0,80808	0,19192			
21	Obs019	1	1	0,96934	0,03066			

Le tableau suivant affiche la matrice de confusion.

The screenshot shows the same software window with a different table selected in the sidebar. The table content is as follows:

	1	2	3	4	5	6	7	8
1								
2	Matrice de confusion pour le jeu d'apprentissage							
3								
4	En lignes, les classes observées							
5	En colonnes, les classes prévues							
6								
7	Pourcentage de mal classés : 9,901 %							
8	Pourcentage de bien classés (exactitude) : 90,099 %							
9								
10	Précision = VP / (VP + FP)							
11	Rappel = VP / (VP + FN)							
12	Score F1 = 2 x (Précision x Rappel) / (Précision + Rappel)							
13								
14								
15	Observé \ Prévu	Taille	Décès	Survie	Précision	Rappel	Score F1	
16	Décès	51	46	5	0,90196	0,90196	0,90196	
17	Survie	50	5	45	0,90000	0,90000	0,90000	
18								
19								
20								
21								

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).

Rapports et Graphiques

Rapport SVM

- Paramètres optimaux des noyaux
- Exactitude (apprentissage)
- Importance des variables
- Vecteurs supports
- Vecteurs supports par classe
- Valeurs de décision (apprentissage)
- Prévisions (apprentissage)
- Matrice de confusion (apprentissage)
- Sensibilité, Spécificité Décès (app)**
- Sensibilité, Spécificité Survie (app)

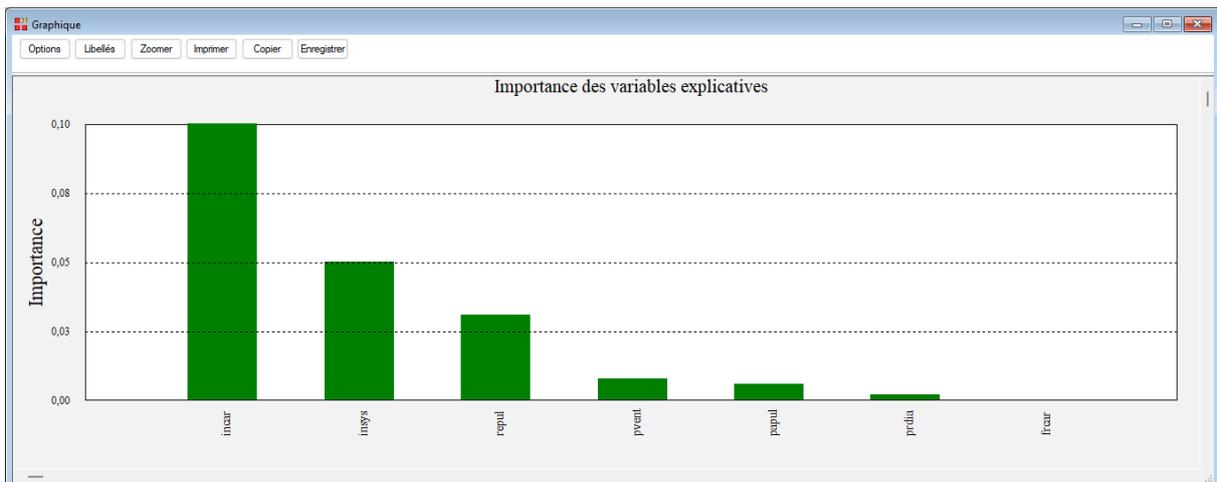
	1	2	3	4	5	6	7	8
1								
2	SENSIBILITE, SPECIFICITE POUR LE JEU D'APPRENTISSAGE							
3								
4	Classe : Décès							
5								
6	Sensibilité en %							
7	Spécificité en %							
8								
9	Aire sous la courbe (AUC) = 1,000							
10								
11								
12			Seuil	Sensibilité	Spécificité			
13	1		-Infini	100	100,00000			
14	2		0,01257	100	100,00000			
15	3		0,01316	100	99,00990			
16	4		0,01545	100	98,01980			
17	5		0,01707	100	97,02970			
18	6		0,01779	100	96,03960			
19	7		0,01953	100	95,04950			
20	8		0,02756	100	94,05941			
21	9		0,02987	100	93,06931			

Rapport Explorateur /

L'option Graphiques

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

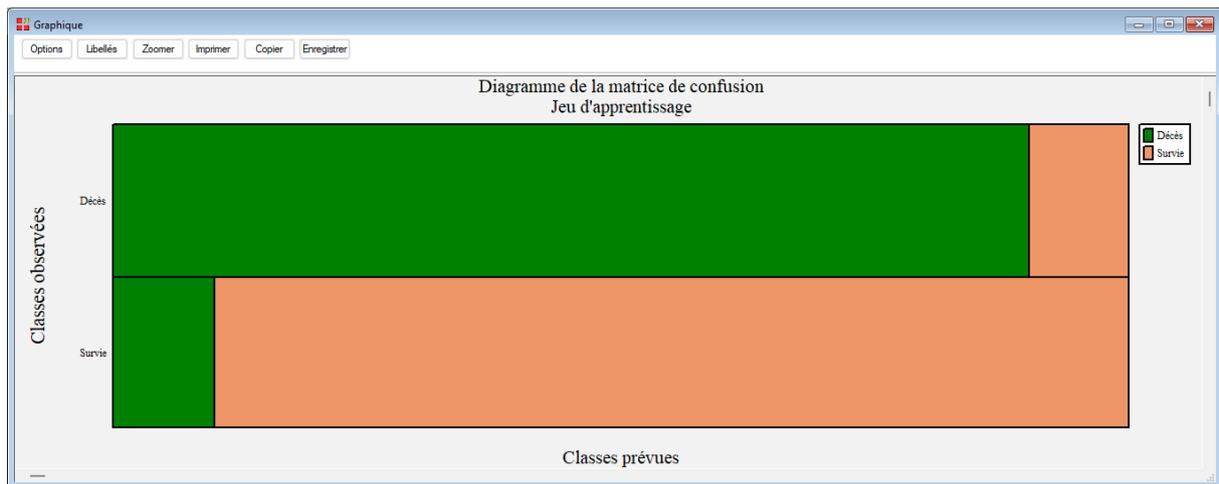
- Graphique de l'importance des variables



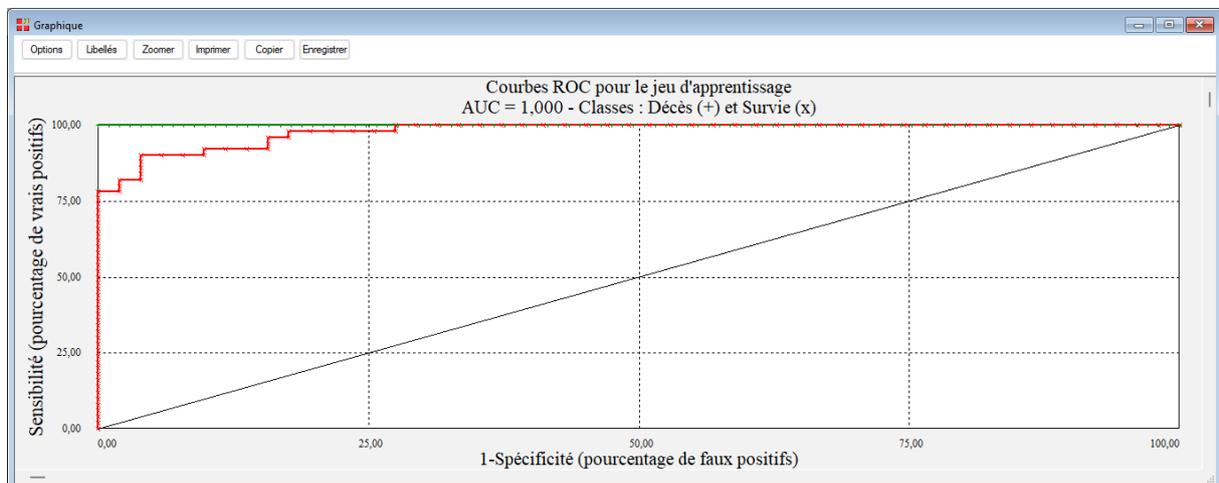
Les deux variables les plus importantes sont *incar* (index cardiaque) et *insys* (index systolique).

- 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)



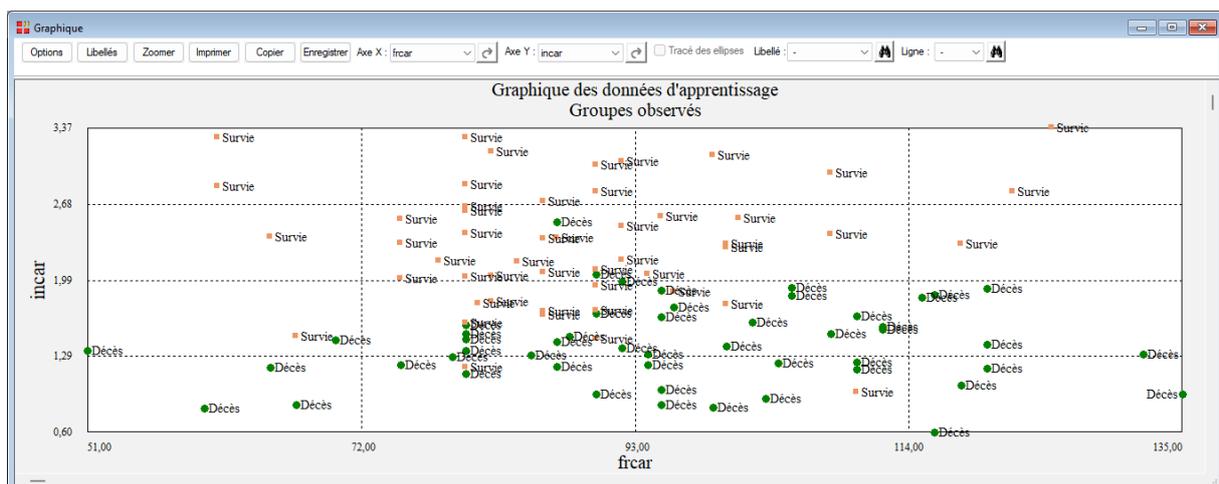
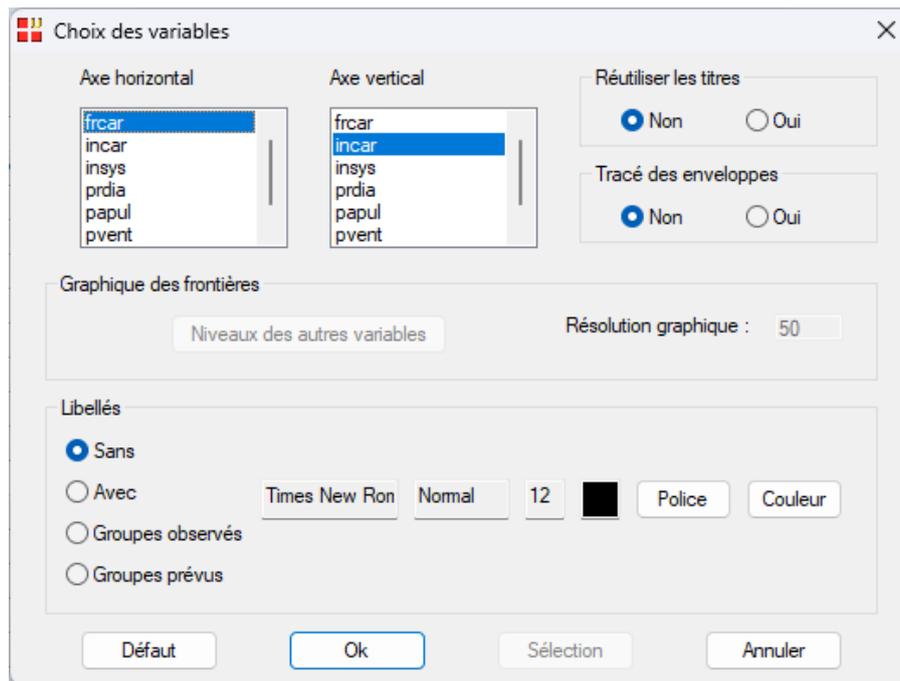
L'aire sous la courbe (AUC) nous indique l'efficacité du modèle construit. Plus la valeur de cette aire est élevée, meilleures sont les performances du modèle pour faire la distinction entre les classes *Survie* et *Décès*.

- 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'.

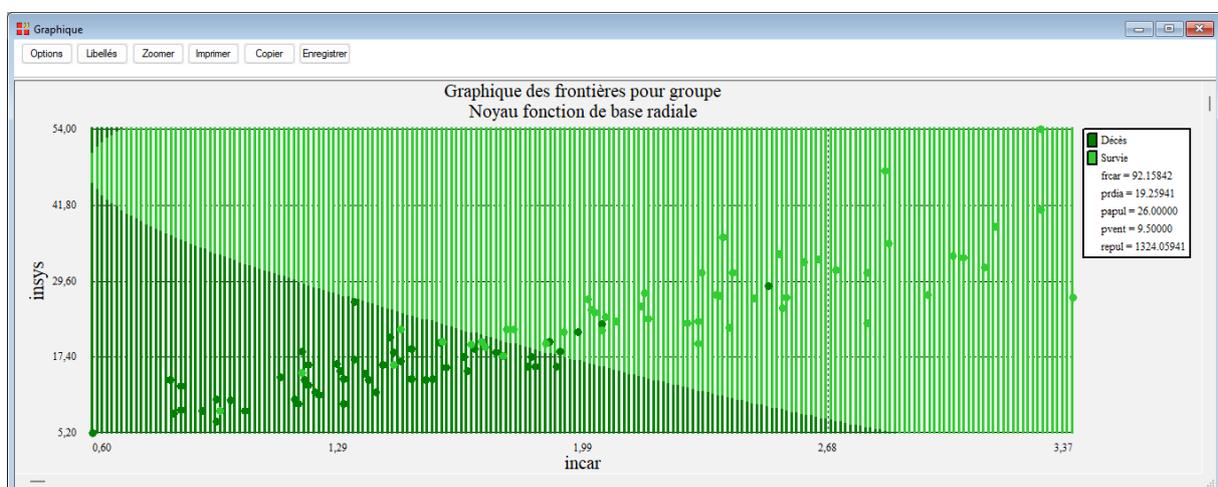
Par défaut les autres variables quantitatives sont maintenues aux valeurs moyennes.

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

	Variable	Niveau
1	frcar	92,15842
2	prdia	19,25941
3	papul	26,00000
4	pvent	9,50000
5	repul	1324,05941

Niveaux

Ok



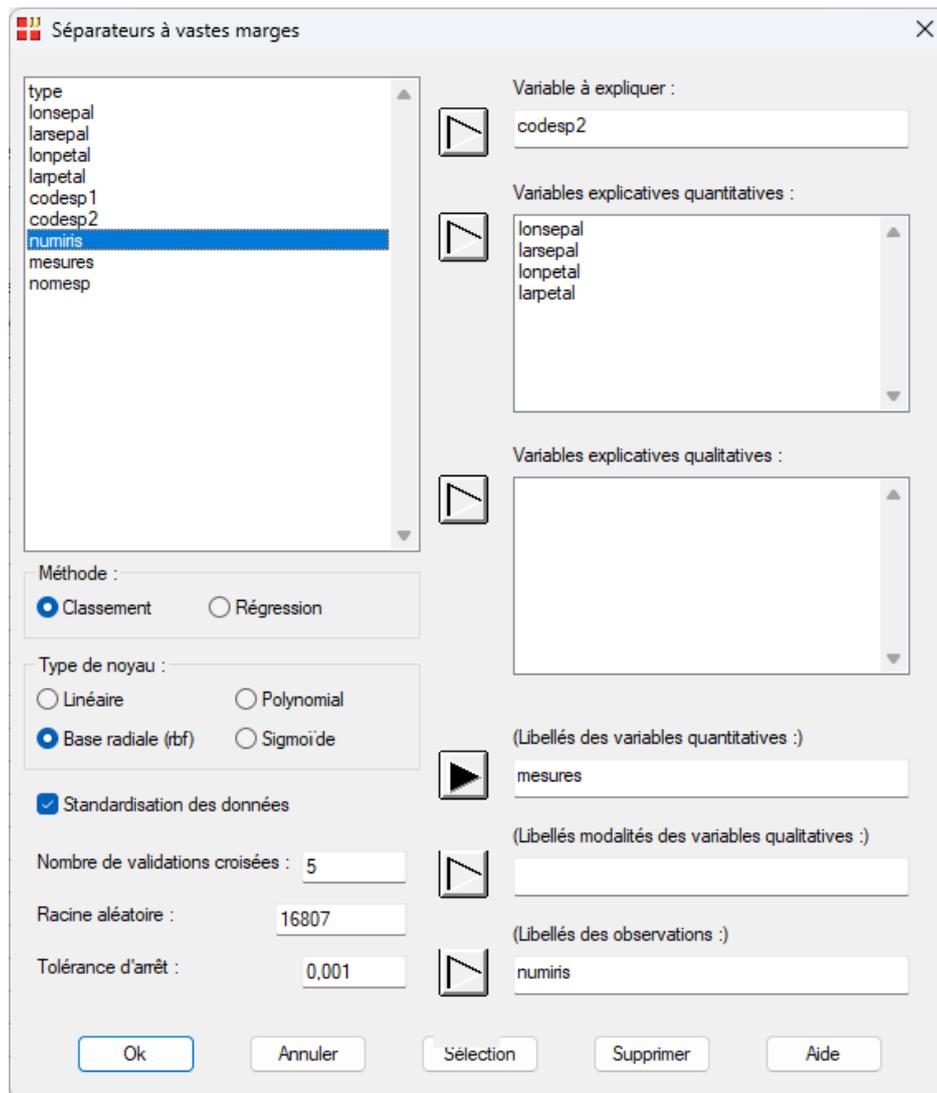
Exemple 2 : Classement multi-classes - Fichier Iris3

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.

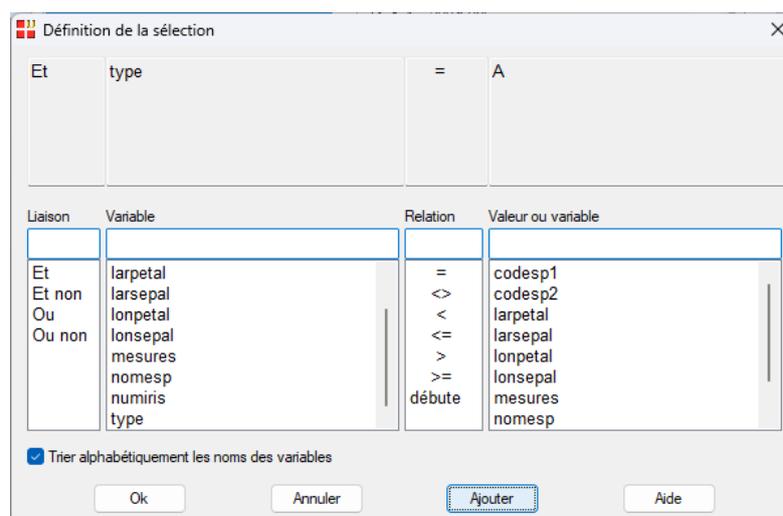
Ce fichier contient 6 iris pour lesquels les classes d'appartenance sont inconnues. Ils définiront le jeu de prévision.

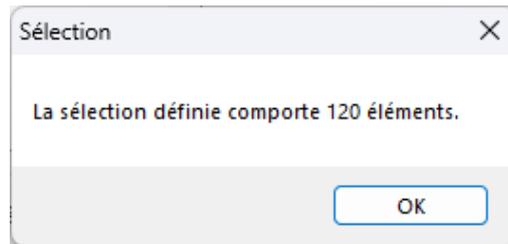
Sélectionnons la variable *codesp2* comme variable à expliquer, les variables *lonsepal* à *larpetal* comme variables explicatives, la variable *mesures* pour les libellés des variables explicatives et la variable *numiris* pour les libellés des observations.

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

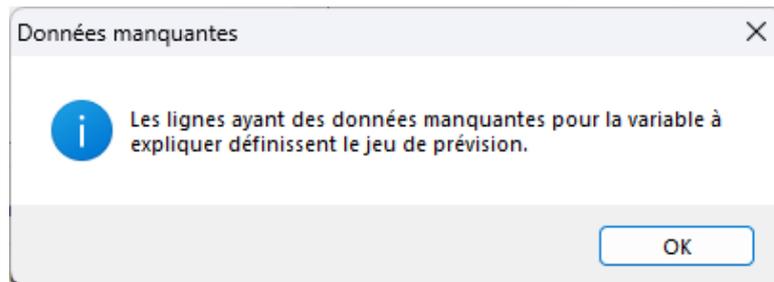


Cliquons sur le bouton *Sélection* pour définir le jeu d'apprentissage :

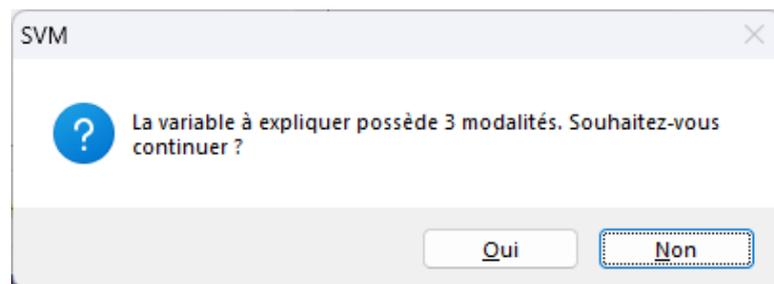




120 iris définissent le jeu d'apprentissage. Cliquons sur le bouton OK.

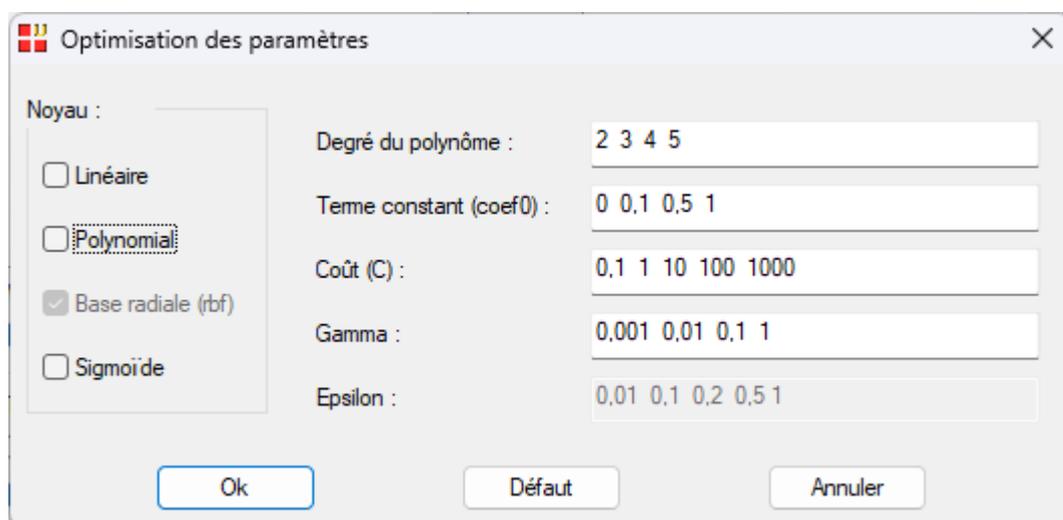


Un message nous informe que les lignes ayant des données manquantes pour la variable à expliquer définissent le jeu de prévision. Cliquons sur OK.



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

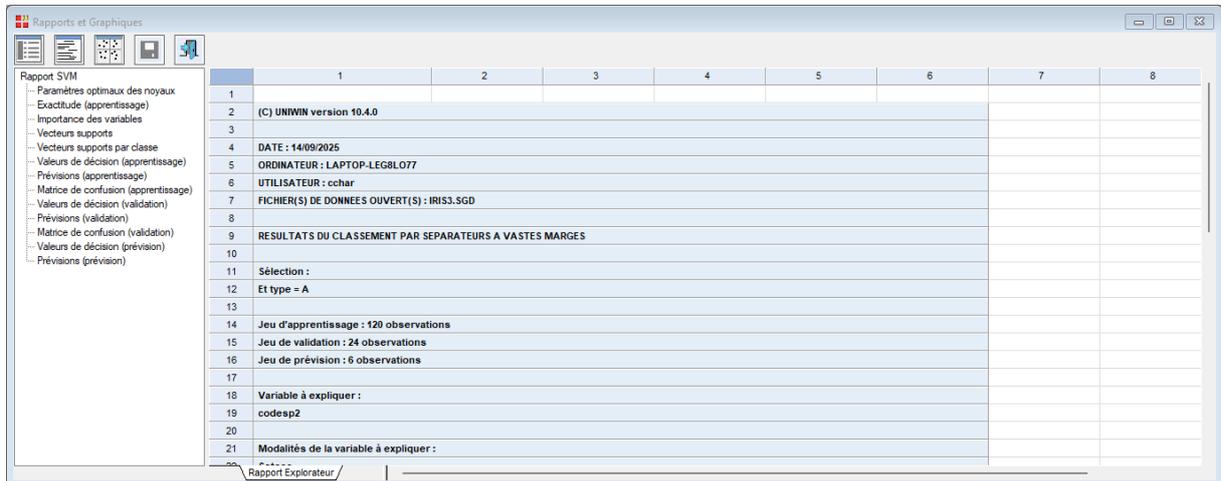
Une fenêtre nous demande ensuite de définir les options pour l'optimisation des paramètres de la méthode de classement :



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 :

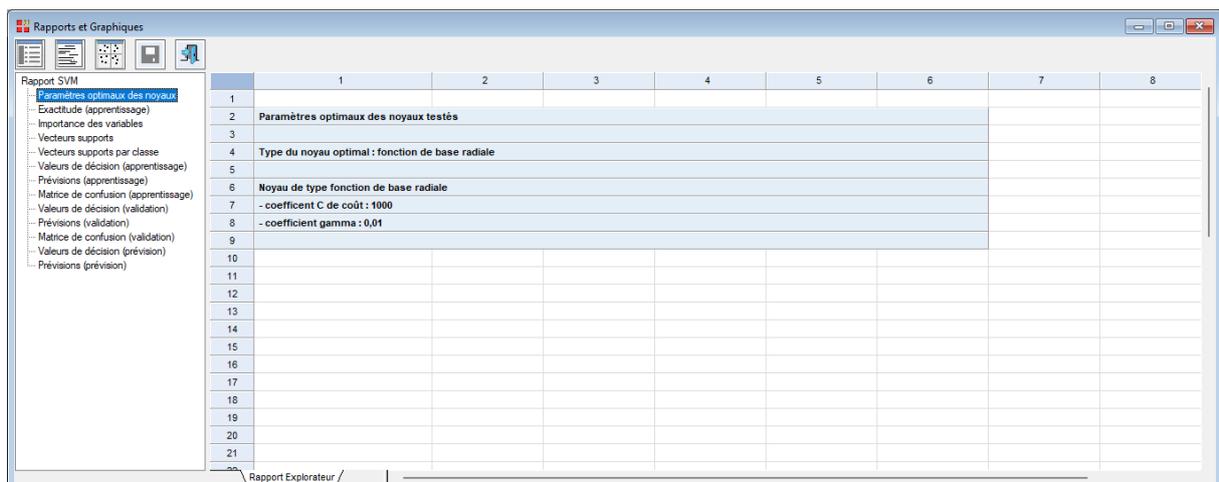


	1	2	3	4	5	6	7	8
1								
2	(C) UNIMIN version 10.4.0							
3								
4	DATE : 14/09/2025							
5	ORDINATEUR : LAPTOP-LEG8L077							
6	UTILISATEUR : cchar							
7	FICHIER(S) DE DONNEES OUVERT(S) : IRIS3.SGD							
8								
9	RESULTATS DU CLASSEMENT PAR SEPARATEURS A VASTES MARGES							
10								
11	Sélection :							
12	Et type = A							
13								
14	Jeu d'apprentissage : 120 observations							
15	Jeu de validation : 24 observations							
16	Jeu de prévision : 6 observations							
17								
18	Variable à expliquer :							
19	codesp2							
20								
21	Modalités de la variable à expliquer :							

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 $\text{Gamma} = 0,01$.



	1	2	3	4	5	6	7	8
1								
2	Paramètres optimaux des noyaux testés							
3								
4	Type du noyau optimal : fonction de base radiale							
5								
6	Noyau de type fonction de base radiale							
7	- coefficient C de coût : 1000							
8	- coefficient gamma : 0,01							
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								

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

	1	2	3	4	5	6	7	8
1								
2	Exactitude calculée par validation croisée							
3	(pourcentage de bien classés du jeu d'apprentissage)							
4								
5	Moyenne = 94,16667							
6	Ecart-type = 6,31906							
7	Minimum = 83,33333							
8	Maximum = 100,00000							
9								
10								
11								
12	Validation croisée n° 1		Exactitude					
13	Validation croisée n° 2		95,83333					
14	Validation croisée n° 3		100,00000					
15	Validation croisée n° 4		83,33333					
16	Validation croisée n° 5		95,83333					
17								
18								
19								
20								
21								

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

	1	2	3	4	5	6	7	8
1								
2	Importance des variables							
3	(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)							
4								
5								
6								
7								
8								
9	larpetal		Importance					
10	lonpetal		0,36500					
11	lonsepal		0,02167					
12	larsepal		0,01000					
13								
14								
15								
16								
17								
18								
19								
20								
21								

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

	1	2	3	4	5	6	7	8
1								
2	Vecteurs supports							
3	Données quantitatives standardisées							
4	Il y a 15 vecteurs supports.							
5	Classes :							
6	1 = Setosa							
7	2 = Versicolor							
8	3 = Virginica							
9								
10								
11								
12								
13								
14								
15	Observation	Classe	lonsepal	larsepal	lonpetal	larpetal		
16	24	1	-0,85954	0,52980	-1,09804	-0,86048		
17	42	1	-1,59279	-1,72467	-1,32376	-1,11928		
18	53	2	1,34022	0,07891	0,70773	0,43348		
19	58	2	-1,10395	-1,49922	-0,19516	-0,21350		
20	67	2	-0,24849	-0,14654	0,48201	0,43348		
21	69	2	0,48476	-1,95012	0,48201	0,43348		

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

Classe	Nombre de vecteurs
Setosa	2
Versicolor	7
Virginica	6

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

	Setosa/Versicolor	Setosa/Virginica	Versicolor/Virginica
1	1,63692	1,34695	14,05660
2	1,37063	1,22809	12,84756
4	1,55085	1,24886	12,66231
5	1,75750	1,37824	14,27266
6	1,41827	1,19956	13,88378
7	1,70372	1,29162	13,17651
8	1,56655	1,30049	13,58683
9	1,55723	1,23482	12,27570
11	1,58267	1,33971	14,46934
12	1,61861	1,28575	13,30792

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.

Rapport SVM

- Paramètres optimaux des noyaux
- Exactitude (apprentissage)
- Importance des variables
- Vecteurs supports
- Vecteurs supports par classe
- Valeurs de décision (apprentissage)
- Prévisions (apprentissage)
- Matrice de confusion (apprentissage)
- Valeurs de décision (validation)
- Prévisions (validation)
- Matrice de confusion (validation)
- Valeurs de décision (prévision)
- Prévisions (prévision)

1	2	3	4	5	6	7	8
1							
2	Prévisions pour le jeu d'apprentissage						
3							
4	Classes :						
5							
6	1 = Setosa						
7	2 = Versicolor						
8	3 = Virginica						
9							
10	(*) = mauvais classement						
11							
12							
13	Observation	Classe observée	Classe prévue	Proba(Setosa)	Proba(Versicolor)	Proba(Virginica)	
14	1	1	1	0,97790	0,01340	0,00869	
15	2	1	1	0,96315	0,02487	0,01198	
16	4	1	1	0,97211	0,01848	0,01141	
17	5	1	1	0,98107	0,01014	0,00799	
18	6	1	1	0,96457	0,02239	0,01304	
19	7	1	1	0,97824	0,01159	0,01017	
20	8	1	1	0,97431	0,01581	0,00988	
21	9	1	1	0,97186	0,01627	0,01187	

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

Rapport SVM

- Paramètres optimaux des noyaux
- Exactitude (apprentissage)
- Importance des variables
- Vecteurs supports
- Vecteurs supports par classe
- Valeurs de décision (apprentissage)
- Prévisions (apprentissage)
- Matrice de confusion (apprentissage)
- Valeurs de décision (validation)
- Prévisions (validation)
- Matrice de confusion (validation)
- Valeurs de décision (prévision)
- Prévisions (prévision)

1	2	3	4	5	6	7	8	
1								
2	Matrice de confusion pour le jeu d'apprentissage							
3								
4	En lignes, les classes observées							
5	En colonnes, les classes prévues							
6								
7	Pourcentage de mal classés : 0,833 %							
8	Pourcentage de bien classés (exactitude) : 99,167 %							
9								
10	Précision = VP / (VP + FP)							
11	Rappel = VP / (VP + FN)							
12	Score F1 = 2 x (Précision x Rappel) / (Précision + Rappel)							
13								
14	Observé \ Prévu	Taille	Setosa	Versicolor	Virginica	Précision	Rappel	Score F
15	Setosa	43	43	0	0	1,000	1,00000	1,0000
16	Versicolor	38	0	37	1	1,000	0,97368	0,9866
17	Virginica	39	0	0	39	0,975	1,00000	0,9873
18								
19								
20								
21								

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

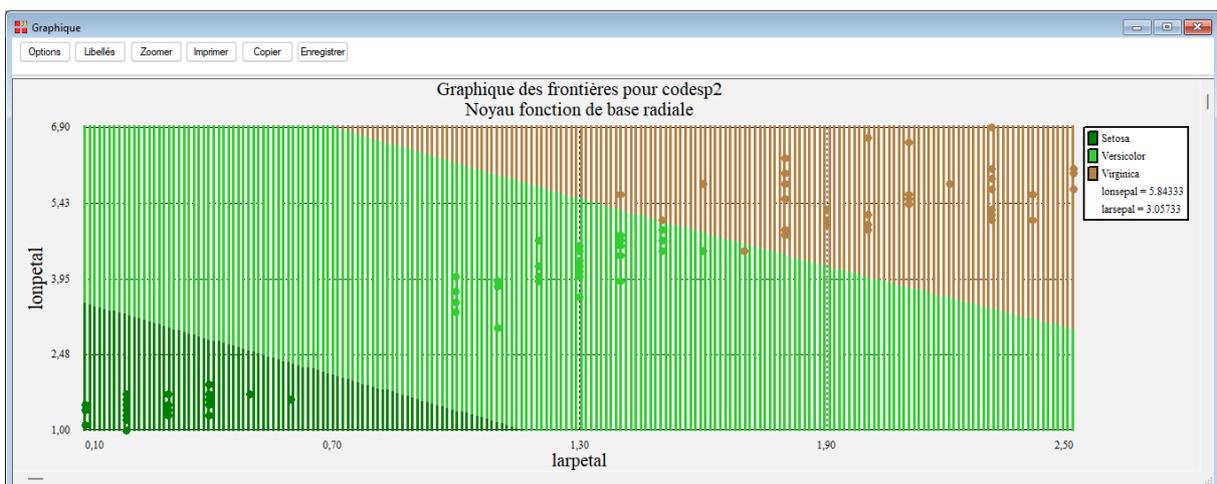
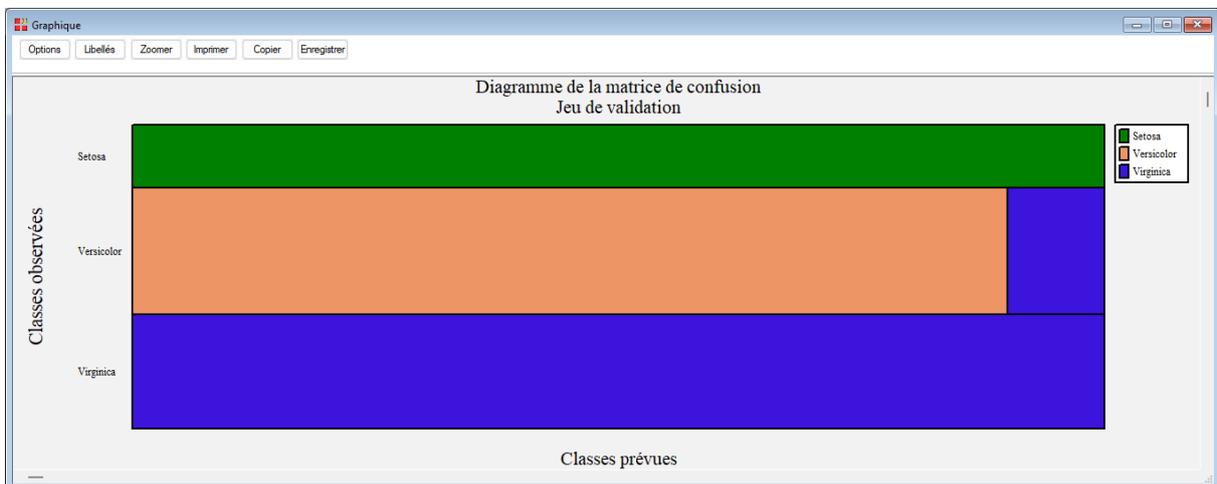
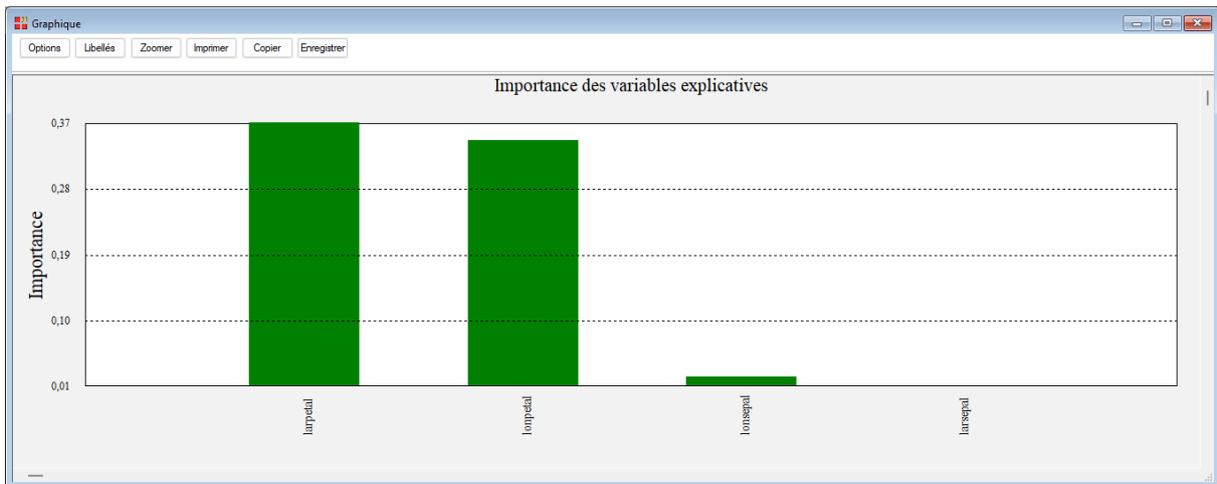
Enfin les deux derniers tableaux affichent les valeurs de décision et les prévisions pour le jeu de prévision.

Rapport SVM

- Paramètres optimaux des noyaux
- Exactitude (apprentissage)
- Importance des variables
- Vecteurs supports
- Vecteurs supports par classe
- Valeurs de décision (apprentissage)
- Prévisions (apprentissage)
- Matrice de confusion (apprentissage)
- Valeurs de décision (validation)
- Prévisions (validation)
- Matrice de confusion (validation)
- Valeurs de décision (prévision)
- Prévisions (prévision)

1	2	3	4	5	6	7	8
1							
2	Prévisions pour le jeu de prévision						
3							
4	Classes :						
5							
6	1 = Setosa						
7	2 = Versicolor						
8	3 = Virginica						
9							
10	Observation	Classe prévue	Proba(Setosa)	Proba(Versicolor)	Proba(Virginica)		
11	3	1	0,97847	0,01240	0,00914		
12	36	1	0,97533	0,01574	0,00892		
13	62	2	0,01317	0,92045	0,06638		
14	84	3	0,01952	0,24157	0,73892		
15	104	3	0,01419	0,03571	0,95011		
16	125	3	0,01270	0,01665	0,97065		
17							
18							
19							
20							
21							

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'oxydes 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*.

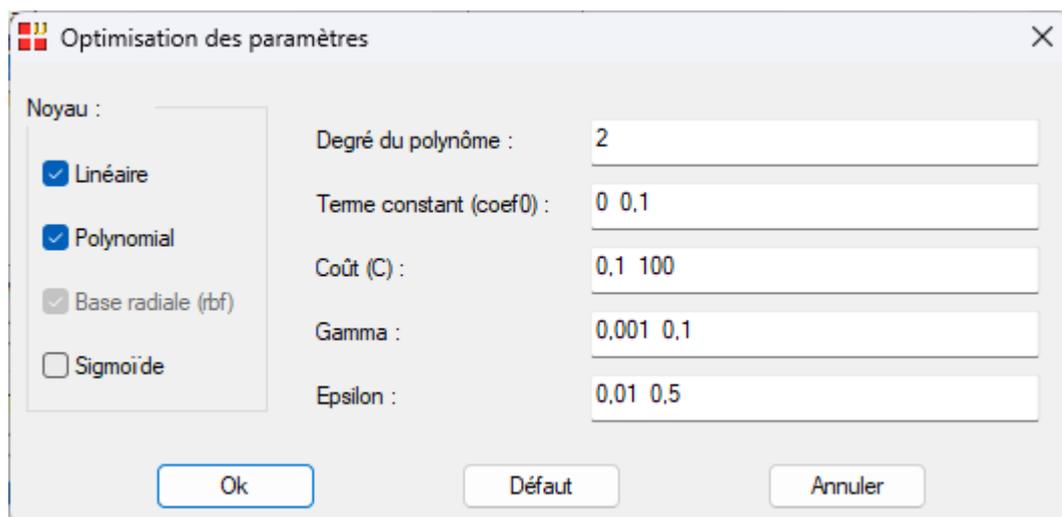
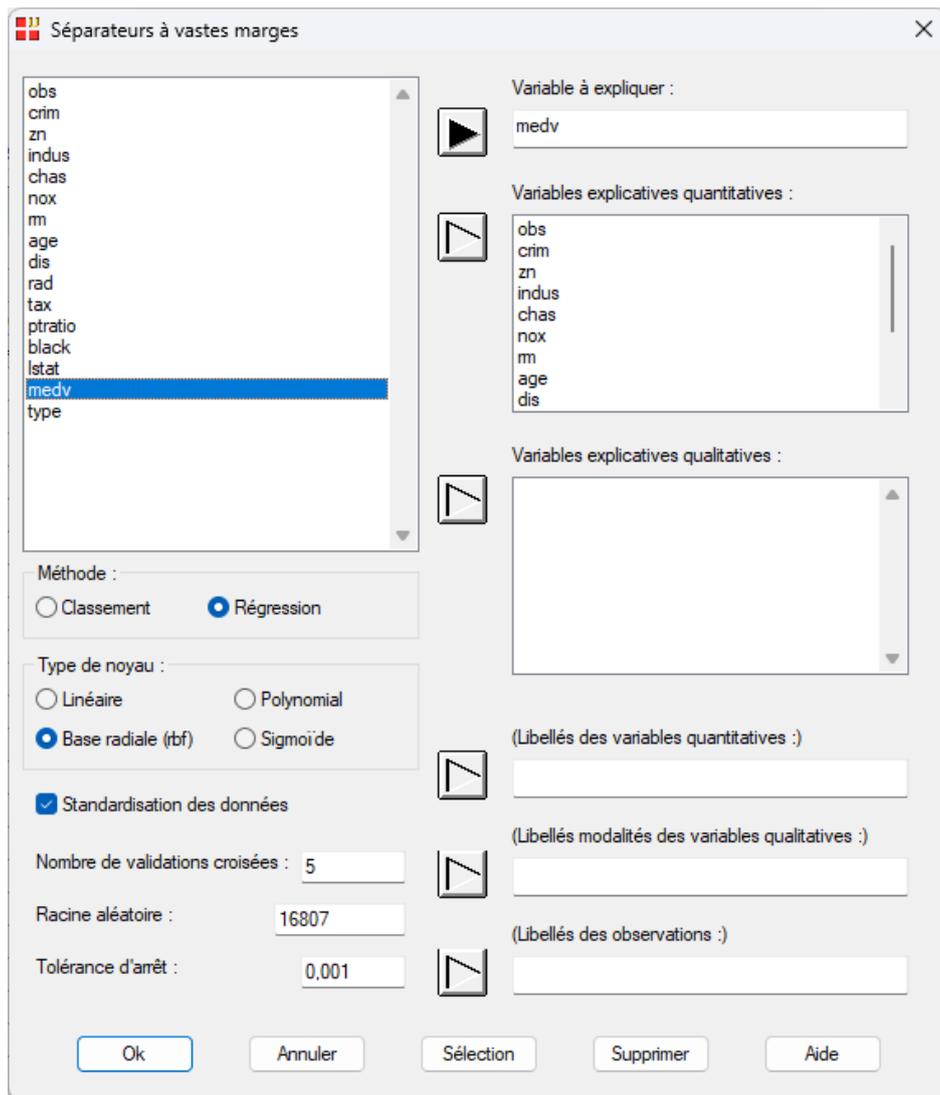
Cliquons sur le bouton 'Sélection' pour sélectionner les données du jeu d'apprentissage (type=A).

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.

	1	2	3	4	5	6	7	8
1								
2	Paramètres optimaux des noyaux testés							
3								
4	Type du noyau optimal : fonction de base radiale							
5								
6	Noyau de type linéaire							
7								
8								
9								
10	Noyau de type polynomial							
11								
12								
13								
14								
15								
16								
17	Noyau de type fonction de base radiale							
18								
19								
20								
21								

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 %.

	1	2	3	4	5	6	7	8
1								
2	Erreur quadratique moyenne (MSE) et R-carré calculés par validation croisée							
3								
4	Moyenne = 13,67639							
5	Ecart-type = 3,66975							
6	Minimum = 9,56644							
7	Maximum = 17,59754							
8								
9	R-carré moyen (validation croisée) = 83,40482							
10								
11								
12			MSE	R-carré				
13	Validation croisée n° 1		9,97540	87,89567				
14	Validation croisée n° 2		17,59754	76,64883				
15	Validation croisée n° 3		9,56644	88,39191				
16	Validation croisée n° 4		16,03279	80,54553				
17	Validation croisée n° 5		15,20979	81,54417				
18								
19								
20								
21								

Un tableau affiche l'importance des variables explicatives.

	1	2	3	4	5	6	7	8
1								
2	Importance des variables							
3								
4	(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)							
5								
6								
7								
8			Importance					
9	rm		6,61210					
10	lstat		6,39235					
11	age		5,80782					
12	dis		4,85785					
13	nox		4,64898					
14	tax		3,80467					
15	rad		3,72901					
16	crim		2,65552					
17	ptratio		2,39575					
18	indus		2,39473					
19	black		1,81949					
20	zn		1,33960					
21	chas		1,11809					

Le tableau suivant affiche les vecteurs supports.

	1	2	3	4	5	6	7	8
1								
2	Vecteurs supports							
3								
4	Données quantitatives standardisées							
5								
6	Il y a 273 vecteurs supports.							
7								
8								
9	Observation	crim	zn	indus	chas	nox	rm	ag
10	o1	-0,42542	0,25796	-1,27446	-0,26655	-0,15637	0,40368	-0,1005
11	o2	-0,42302	-0,48932	-0,58659	-0,26655	-0,74575	0,18727	0,3822
12	o3	-0,42302	-0,48932	-0,58659	-0,26655	-0,74575	1,26088	-0,2450
13	o5	-0,41822	-0,48932	-1,29324	-0,26655	-0,83970	1,20748	-0,4882
14	o6	-0,42272	-0,48932	-1,29324	-0,26655	-0,83970	0,19992	-0,3296
15	o7	-0,41602	0,02962	-0,47098	-0,26655	-0,27595	-0,38748	-0,0512
16	o8	-0,40956	0,02962	-0,47098	-0,26655	-0,27595	-0,16284	0,9885
17	o11	-0,40034	0,02962	-0,47098	-0,26655	-0,27595	0,12544	0,9250
18	o12	-0,41267	0,02962	-0,47098	-0,26655	-0,27595	-0,39169	0,5232
19	o13	-0,41539	0,02962	-0,47098	-0,26655	-0,27595	-0,58032	-1,0240
20	o14	-0,35387	-0,48932	-0,43196	-0,26655	-0,15637	-0,47601	-0,2204
21	o19	-0,33402	-0,48932	-0,43196	-0,26655	-0,15637	-1,16879	-1,1085

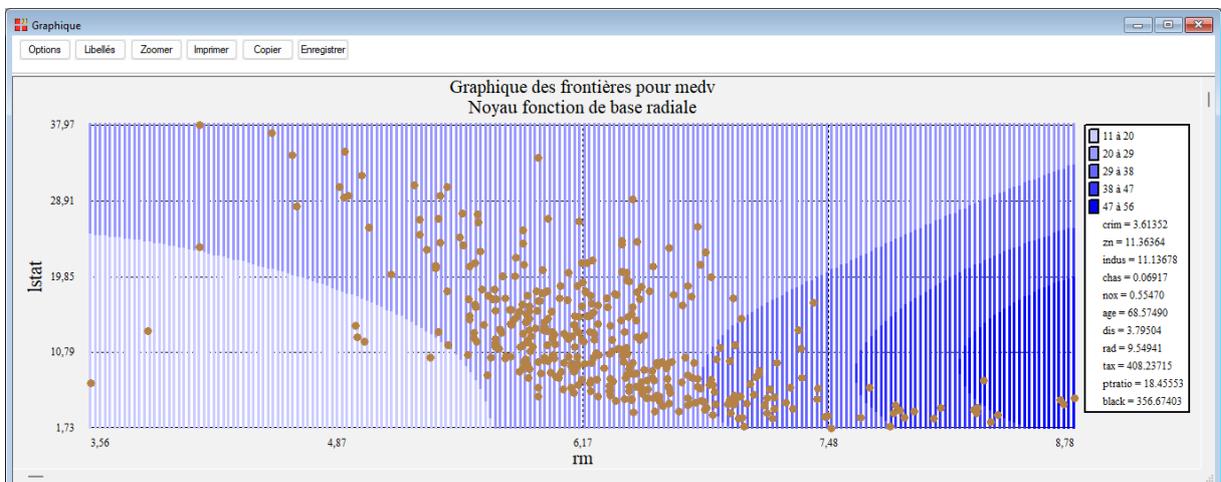
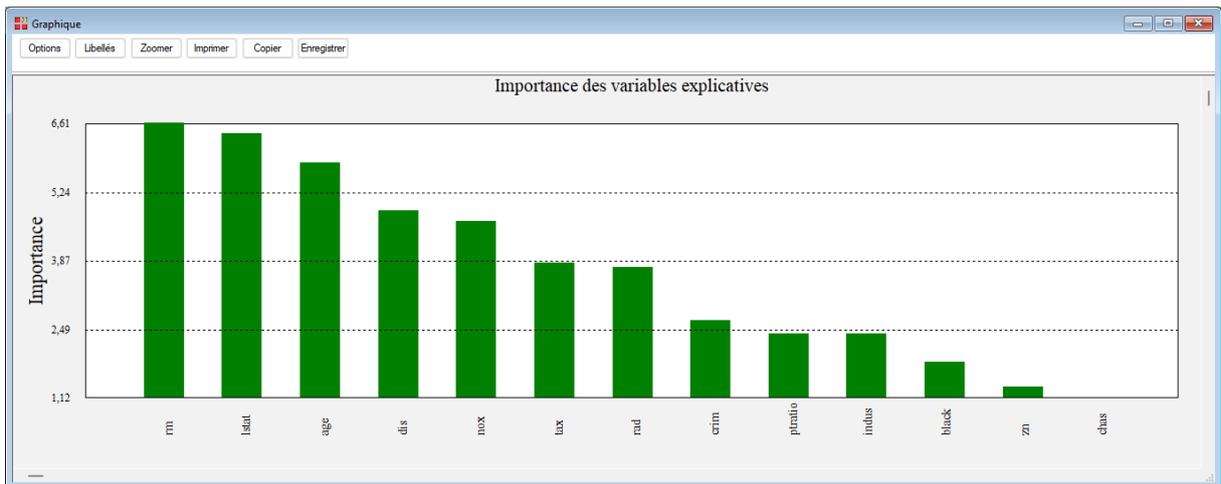
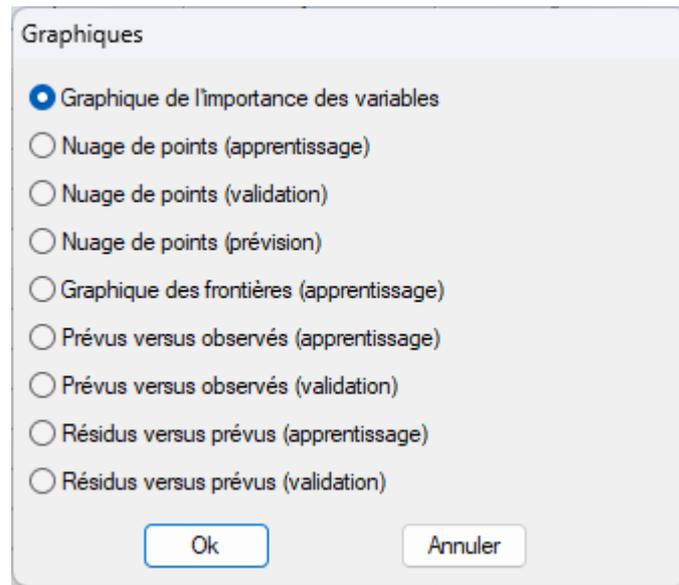
Les résultats pour les jeux d'apprentissage et de validation affichent les valeurs observées, prévues et les résidus ainsi que le R-carré.

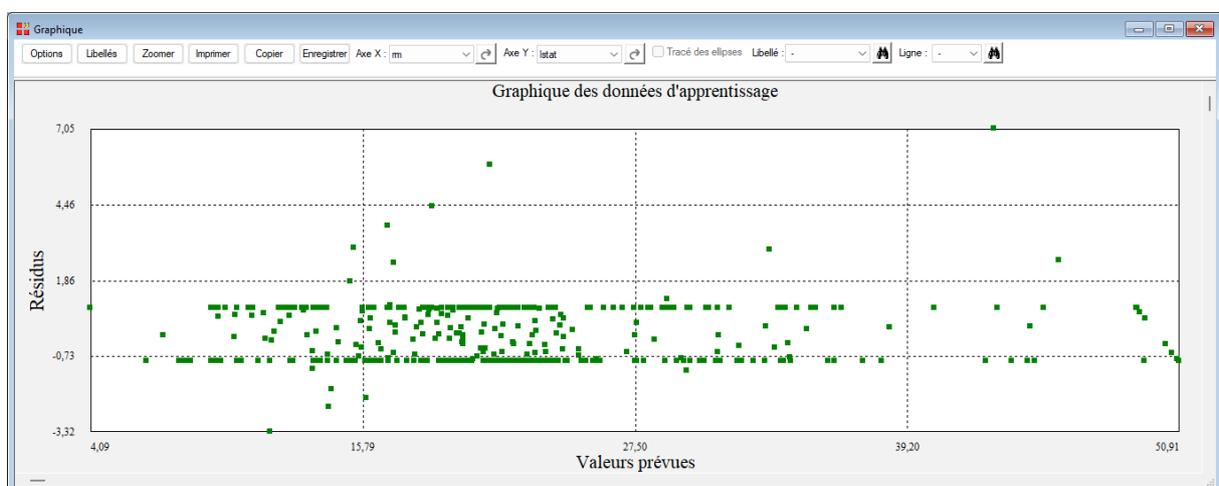
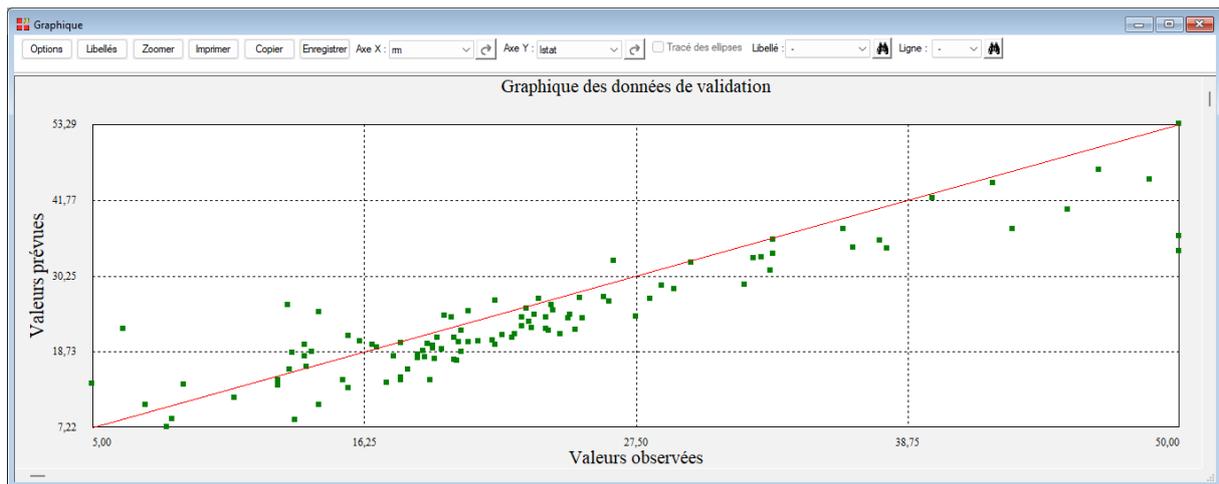
	1	2	3	4	5	6	7	8
1								
2	Résultats pour le jeu d'apprentissage							
3								
4	R-carré = 0,987							
5								
6								
7	Observation	Observé	Prévu	Résidu				
8	o1	24,0	23,09494	0,90506				
9	o2	21,6	22,51200	-0,91200				
10	o3	34,7	33,79030	0,90970				
11	o4	33,4	34,19296	-0,79296				
12	o5	36,2	35,29018	0,90982				
13	o6	28,7	27,79043	0,90957				
14	o7	22,9	21,98873	0,91127				
15	o8	27,1	21,29384	5,80616				
16	o11	15,0	15,90980	-0,90980				
17	o12	18,9	19,81030	-0,91030				
18	o13	21,7	22,60847	-0,90847				
19	o14	20,4	19,48857	0,91143				
20	o15	18,2	17,85723	0,54277				
21	o16	19,9	19,21052	0,68948				

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

	1	2	3	4	5	6	7	8
1								
2	Résultats pour le jeu de prévision							
3								
4	R-carré = 0,987							
5								
6								
7	Observation	Observé	Prévu	Résidu				
8	o1	24,0	23,09494	0,90506				
9	o2	21,6	22,51200	-0,91200				
10	o3	34,7	33,79030	0,90970				
11	o4	33,4	34,19296	-0,79296				
12	o5	36,2	35,29018	0,90982				
13	o6	28,7	27,79043	0,90957				
14	o7	22,9	21,98873	0,91127				
15	o8	27,1	21,29384	5,80616				
16	o11	15,0	15,90980	-0,90980				
17	o12	18,9	19,81030	-0,91030				
18	o13	21,7	22,60847	-0,90847				
19	o14	20,4	19,48857	0,91143				
20	o15	18,2	17,85723	0,54277				
21	o16	19,9	19,21052	0,68948				

Visualisons les principaux graphiques :





Enregistrement des résultats

Voici la liste des variables créées par la procédure.

<i>Variable</i>	<i>Contenu</i>
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 :

$K(x, y) = \langle x, y \rangle$ où $\langle x, y \rangle$ est le produit scalaire de x et y

- Noyau polynomial

Permet de capturer des relations polynomiales entre les variables :

$K(x, y) = [(\gamma * \langle x, y \rangle) + \text{coef0}]^d$

où γ = 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 γ est grand, plus la frontière de décision sera complexe.

- Noyau sigmoïde

Inspiré des réseaux de neurones :

$K(x, y) = \tanh[(\gamma * \langle x, y \rangle) + \text{coef0}]$

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
- Sigmoides : rarement utilisé en pratique

Paramètres des noyaux

	Base radiale (rbf)	Linéaire	Polynomial	Sigmoides
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, 2^{ème} édition.

Exemple 'Boston' : <http://lib.stat.cmu.edu/datasets/boston>