

UNIWIN VERSION 9.7.0

METHODE KNN

Révision : 02/09/2023

Définition.....	1
Entrée des données	2
Données manquantes	2
Exemple 1 : Fichier IRIS3.....	3
L'option Rapports	5
L'option Graphiques	6
Exemple 2 : Fichier INFARCT2	9
Exemple 3 : Fichier TITANIC	12
Les variables créées par la procédure.....	14
Références	15

Définition

La méthode des K plus proches voisins (KNN) a pour objectif de classer des observations dont les classes sont inconnues (échantillon de prévision) en fonction de leurs distances euclidiennes (calculées en utilisant les variables explicatives quantitatives précisées) à des observations dont les classes sont connues (échantillon d'apprentissage).

Une plage de valeurs de K est précisée.

La première étape de cette méthode consiste à classer les observations de l'échantillon d'apprentissage par validation croisée (méthode retirer 1 à la fois). Chaque observation retirée est affectée à la classe la plus fréquente de ses K plus proches voisins par un vote majoritaire. Le taux d'erreur de classement est alors calculé pour chaque valeur de K et le K optimal est déterminé.

La seconde étape consiste à classer les données de l'échantillon de prévision en utilisant ce K optimal. Chaque nouvelle observation est affectée à la classe la plus fréquente de ses K plus proches voisins.

Cette procédure est basée sur le package R 'class'.

Entrée des données

Cliquons sur l'icône KNN dans le ruban Expliquer. La boîte de dialogue montrée ci-dessous s'affiche :

Méthode KNN

Facteur de classement qualitatif :

Variables explicatives quantitatives :

(Libellés des individus :)

Prétraitement des données

Aucun

Centrage et réduction

Min-Max

Plage pour les nombres de voisins : 1 à 10

Racine aléatoire : 1023129506

Ok Annuler Sélection Supprimer Aide

Cette boîte de dialogue permet de définir le facteur de classement qualitatif, la liste des variables explicatives quantitatives et les libellés optionnels des individus. Elle permet également de préciser si les données explicatives doivent être standardisées (centrées et réduites), la plage des valeurs de K à étudier et la racine aléatoire pour la validation croisée en cas de votes identiques pour différentes classes.

Les résultats obtenus peuvent être différents de ceux présentés dans cette documentation si la racine aléatoire n'est pas celle utilisée dans les exemples.

Données manquantes

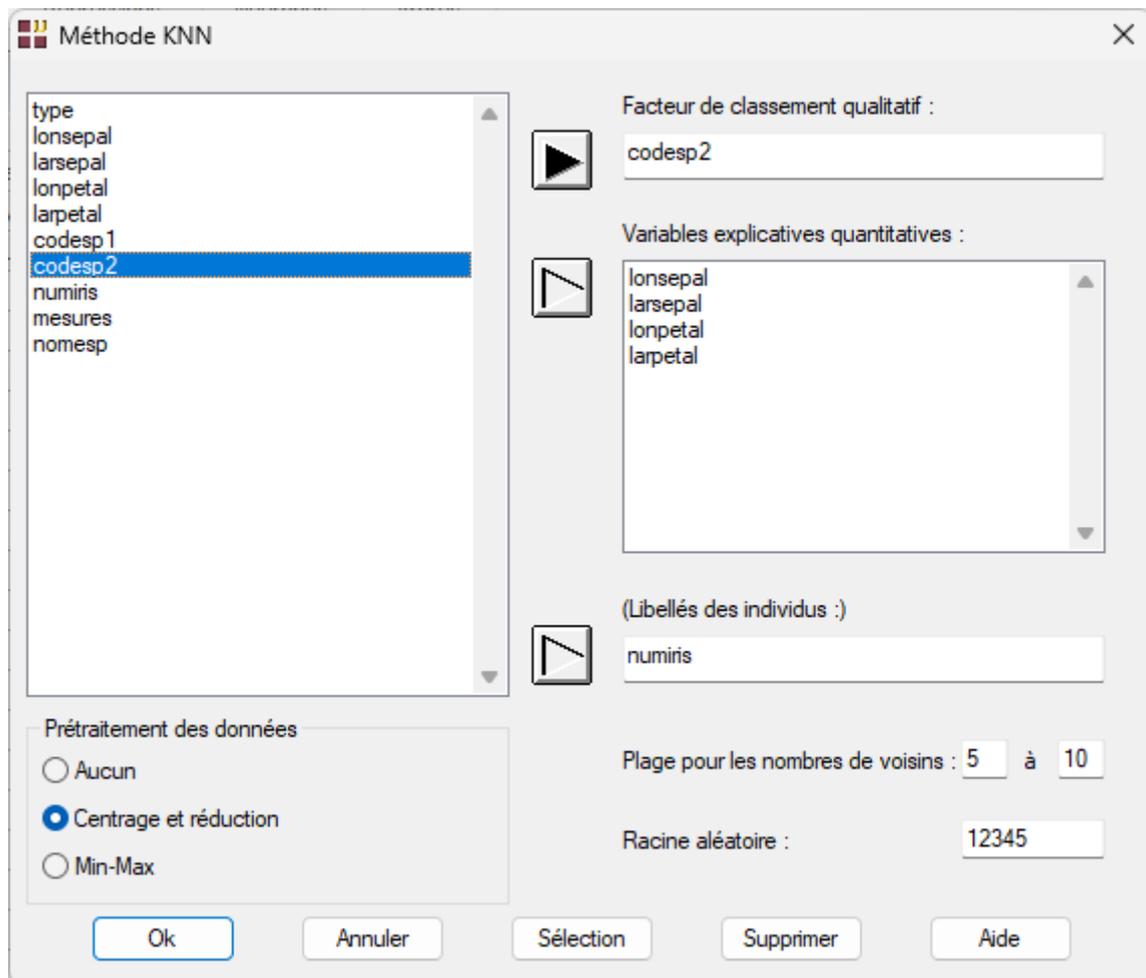
Dans cette procédure, les valeurs manquantes du facteur de classement permettent de définir l'échantillon de prévision. Les lignes ayant des valeurs manquantes pour les variables explicatives ne sont pas prises en compte par l'analyse.

Exemple 1 : 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 l'échantillon de prévision.

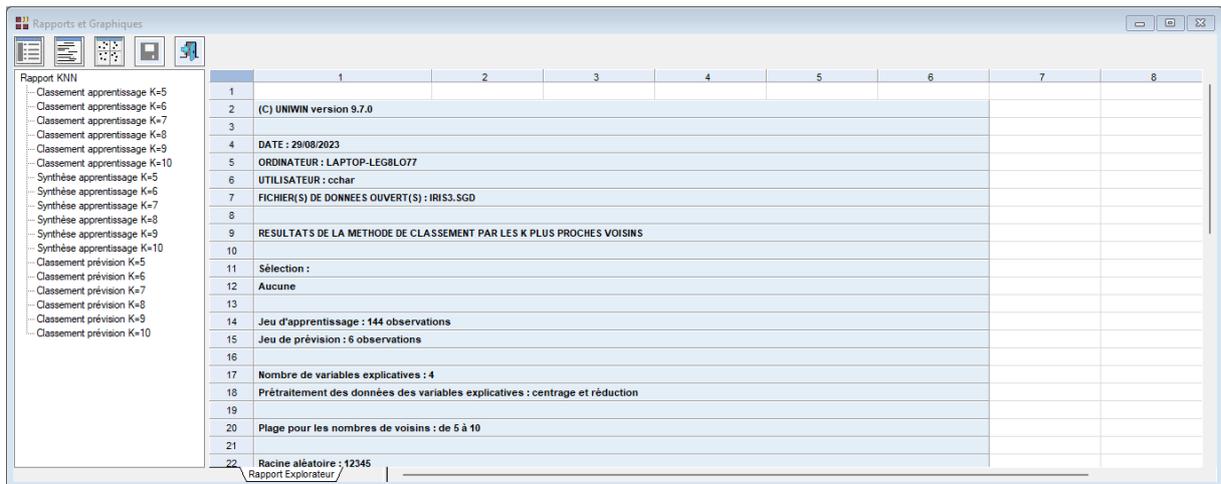
Renseignons la boîte de dialogue comme montré ci-dessous.



Sélectionnons la variable *codesp2* comme facteur de classement, les variables *lonsepal* à *larpetal* comme variables explicatives et la variable *numiris* pour les libellés des individus. Standardisons les données et définissons la plage des valeurs de K de 5 à 10.

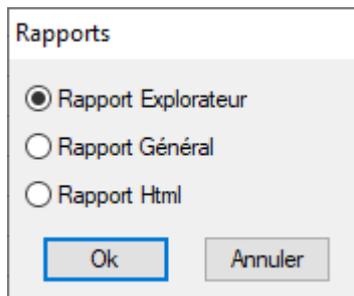
Cliquons sur le bouton Ok pour exécuter le traitement de l'analyse.

Après quelques instants, la fenêtre « Rapports et Graphiques » s'affiche :

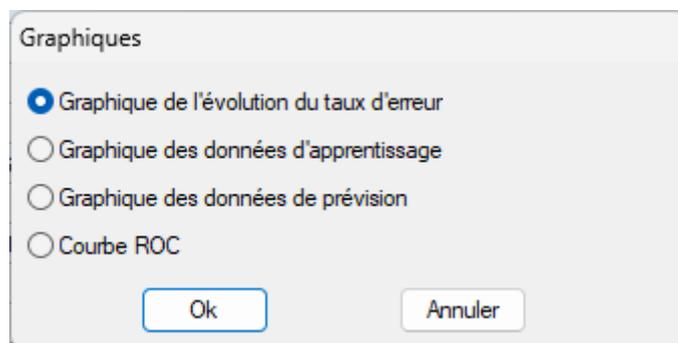


La barre d'outils 'Rapports et Graphiques' permet par l'icône 'Données'  de rappeler la boîte de dialogue d'entrée des données.

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 (1/4)

Enregistrer

- Libellés des individus (apprentissage)
- Classes des individus (apprentissage)
- Affectations K = 5 (apprentissage)
- Proportions des votes K = 5 (apprentissage)
- Affectations K = 6 (apprentissage)
- Proportions des votes K = 6 (apprentissage)
- Affectations K = 7 (apprentissage)
- Proportions des votes K = 7 (apprentissage)
- Affectations K = 8 (apprentissage)
- Proportions des votes K = 8 (apprentissage)

Noms attribués aux variables cibles

applibind

applcind

appaffect5

appprop5

appaffect6

appprop6

appaffect7

appprop7

appaffect8

appprop8

Ok Plus Tout Annuler

L'icône 'Quitter'  permet de quitter l'analyse.

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 affiche pour chacune des valeurs de K les classements de l'échantillon d'apprentissage effectués par validation croisée (méthode retirer 1 à la fois) et les proportions des votes (classes les plus fréquentes des K plus proches voisins).

Rapports et Graphiques

Rapport KNN

- Classement apprentissage K=5
- Classement apprentissage K=6
- Classement apprentissage K=7
- Classement apprentissage K=8
- Classement apprentissage K=9
- Classement apprentissage K=10
- Synthèse apprentissage K=5
- Synthèse apprentissage K=6
- Synthèse apprentissage K=7
- Synthèse apprentissage K=8
- Synthèse apprentissage K=9
- Synthèse apprentissage K=10
- Classement prévision K=5
- Classement prévision K=6
- Classement prévision K=7
- Classement prévision K=8
- Classement prévision K=9
- Classement prévision K=10

	1	2	3	4	5	6	7
1							
2	RESULTATS DU CLASSEMENT DE LA POPULATION D'APPRENTISSAGE PAR VALIDATION CROISEE						
3							
4	Nombre de plus proches voisins : 5						
5							
6	Individu - Groupe observé - Groupe prévu - Proportion des votes						
7							
8	(*) = mal classé						
9							
10							
11		Proportion des votes					
12	Individu : 1 - Observé : Setosa -> Prévu : Setosa						1,0
13	Individu : 2 - Observé : Setosa -> Prévu : Setosa						1,0
14	Individu : 4 - Observé : Setosa -> Prévu : Setosa						1,0
15	Individu : 5 - Observé : Setosa -> Prévu : Setosa						1,0
16	Individu : 6 - Observé : Setosa -> Prévu : Setosa						1,0
17	Individu : 7 - Observé : Setosa -> Prévu : Setosa						1,0
18	Individu : 8 - Observé : Setosa -> Prévu : Setosa						1,0
19	Individu : 9 - Observé : Setosa -> Prévu : Setosa						1,0
20	Individu : 10 - Observé : Setosa -> Prévu : Setosa						1,0
21	Individu : 11 - Observé : Setosa -> Prévu : Setosa						1,0
22	Individu : 12 - Observé : Setosa -> Prévu : Setosa						1,0

Rapport Explorateur

Les individus mal classés sont affichés avec une « * »

Le deuxième tableau affiche pour chacune des valeurs de K la matrice de confusion et le taux d'erreur de classement.

The screenshot shows a software window titled 'Rapports et Graphiques'. On the left is a tree view with 'Synthèse apprentissage K=5' selected. The main area displays a table with the following content:

1								
2	SYNTHÈSE DU CLASSEMENT DE LA POPULATION D'APPRENTISSAGE PAR VALIDATION CROISEE							
3								
4	Nombre de plus proches voisins : 5							
5								
6	En lignes, les groupes observés							
7	En colonnes, les groupes prévus							
8								
9	Pourcentage de mal classés : 4,167 %							
10	Pourcentage de bien classés : 95,833 %							
11								
12								
13		Setosa	Versicolor	Virginica	Total			
14	Setosa	48	0	0	48			
15	Versicolor	0	45	3	48			
16	Virginica	0	3	45	48			
17	Total	48	48	48	144			
18								
19								
20								
21								
22								

Le troisième tableau affiche pour chacune des valeurs de K les classes prévues pour les données du jeu de prévision.

Ce tableau n'est disponible que s'il y a un jeu de prévision, c'est-à-dire des valeurs manquantes pour le facteur de classement.

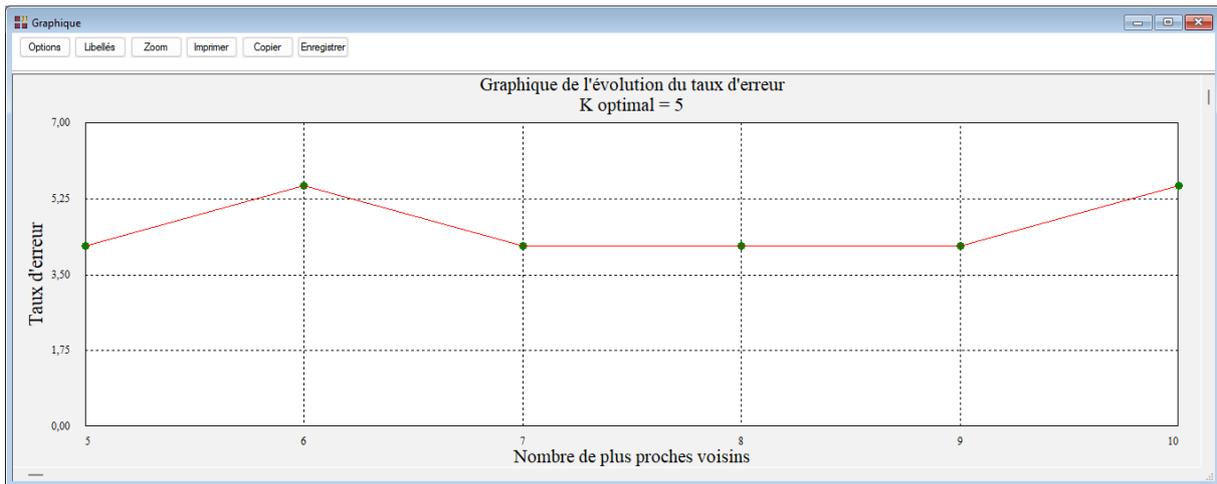
The screenshot shows the same software window with 'Classement prévision K=5' selected in the tree view. The main table displays the following content:

1									
2	RESULTATS DU CLASSEMENT DE LA POPULATION DE PREVISION								
3									
4	Nombre de plus proches voisins : 5								
5									
6	Individu - Groupe prévu - Proportion des votes								
7									
8									
9		Proportion des votes							
10	Individu : 3 -> Prévu : Setosa							1	
11	Individu : 36 -> Prévu : Setosa							1	
12	Individu : 62 -> Prévu : Versicolor							1	
13	Individu : 84 -> Prévu : Virginica							1	
14	Individu : 104 -> Prévu : Virginica							1	
15	Individu : 125 -> Prévu : Virginica							1	
16									
17									
18									
19									
20									
21									
22									

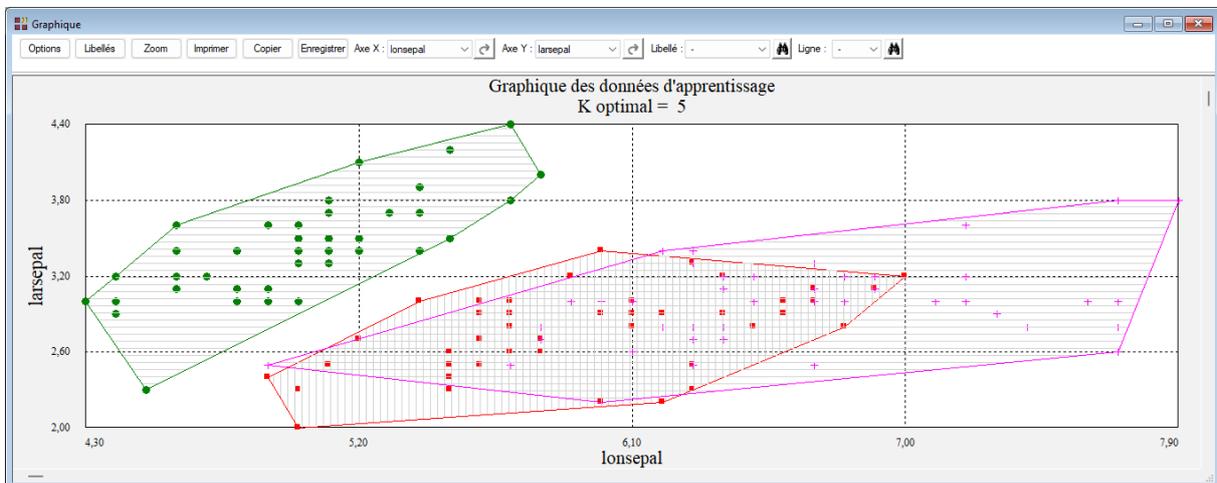
L'option Graphiques

- Graphique de l'évolution du taux d'erreur

Ce graphique affiche l'évolution du taux d'erreur en fonction des valeurs de K. Il permet de choisir la valeur optimale du nombre de voisins utilisés par la méthode KNN. Pour rappel, le taux d'erreur est obtenu suite au classement des observations de l'échantillon d'apprentissage par validation croisée (méthode retirer 1 à la fois).



- Graphique des données d'apprentissage



Le graphique est affiché par défaut pour la valeur optimale de K, ici égale à 5.

Le bouton 'Libellés' permet de préciser les libellés à afficher :

Libellés des points

Libellés

Non

Oui

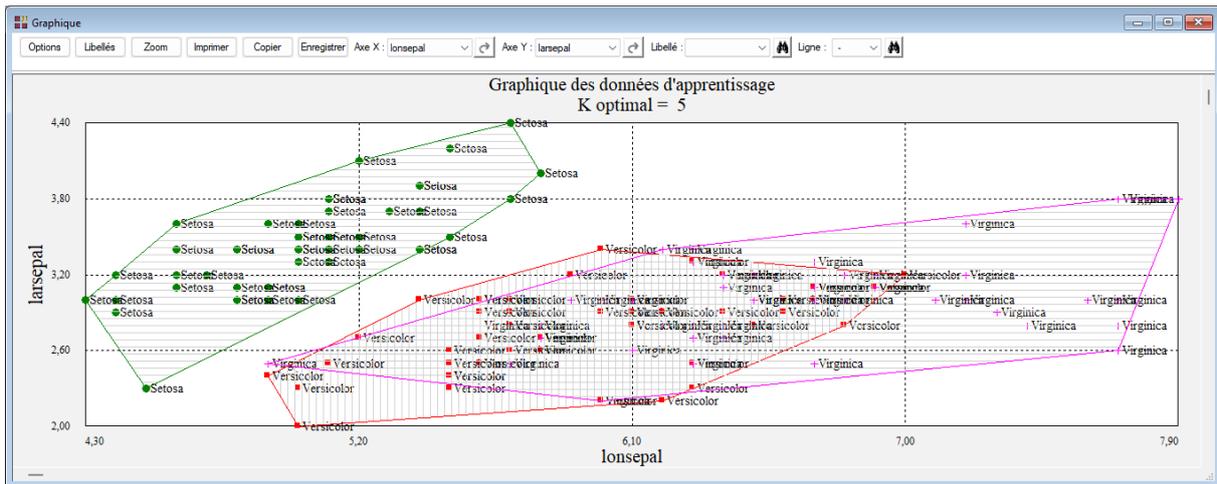
Groupes observés

Groupes prévus

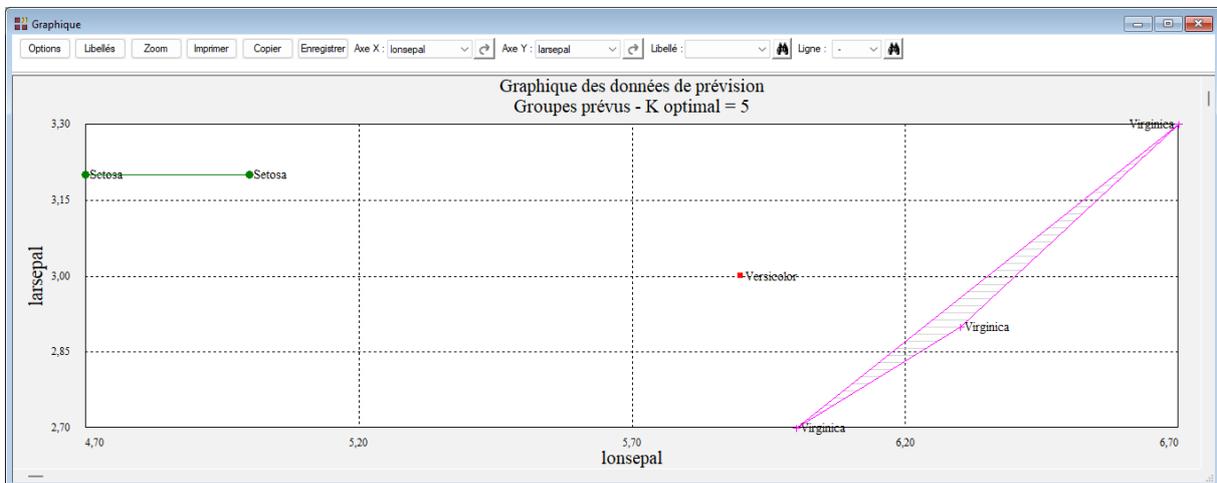
Times New Roman Normal 12 Police Couleur

Défaut Ok Annuler

Quatre options : pas de libellés, les libellés des individus, les groupes observés ou les groupes prévus.



- Graphique des données de prévision



- Courbe ROC

Dans cet exemple, la courbe ROC n'est pas disponible car la variable à expliquer possède plus de deux classes.

Exemple 2 : Fichier INFARCT2

Pour ce deuxième exemple, nous utiliserons le fichier INFARCT2. Ce fichier contient des informations concernant 101 victimes d'un infarctus du myocarde.

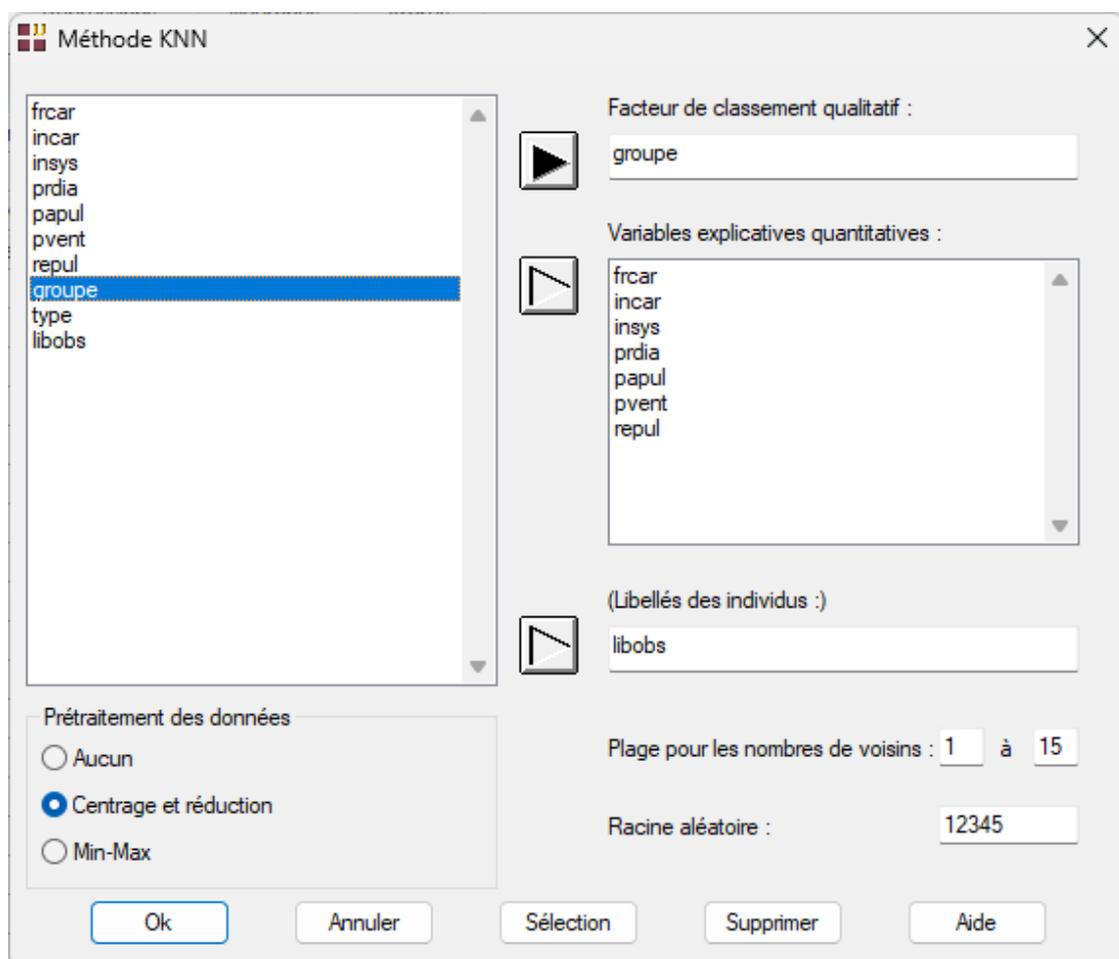
Les variables mesurées sont :

- *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 individus.

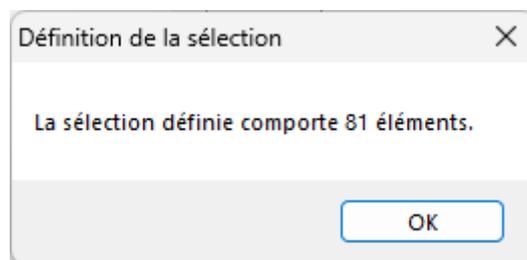
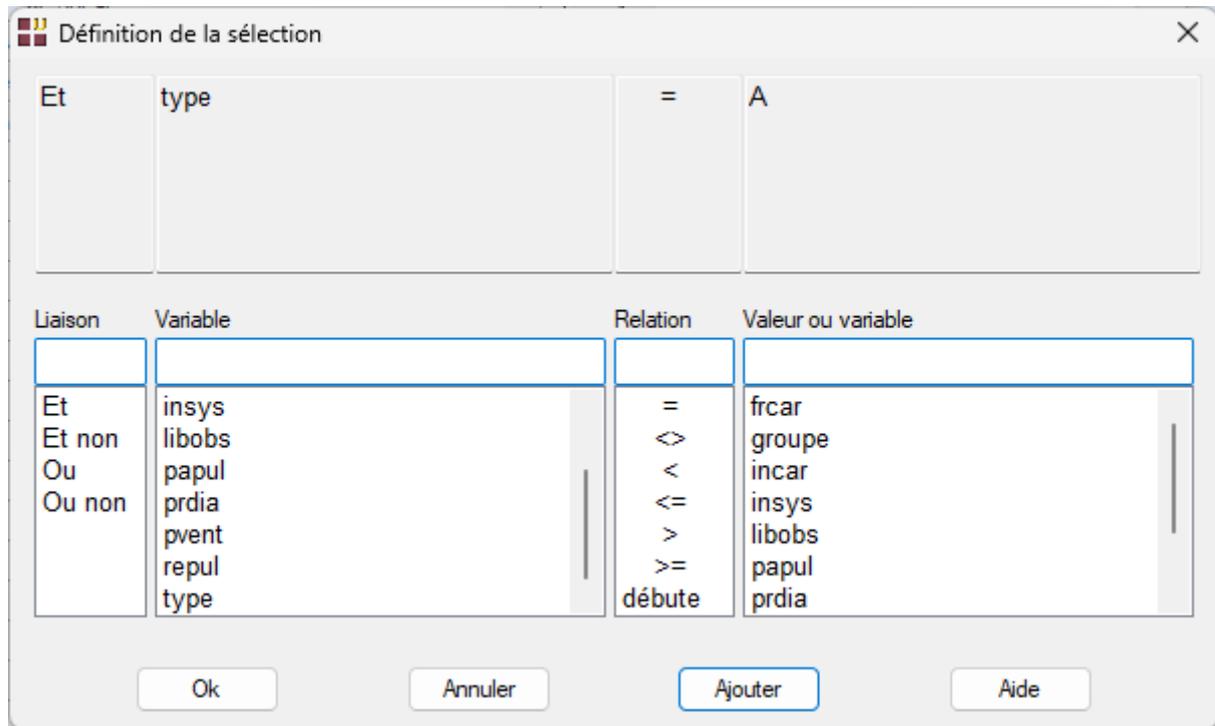
La variable qualitative *groupe* indique par ses deux codes les personnes décédées ou vivantes.

Cliquons sur l'icône KNN dans le ruban Expliquer et renseignons la boîte de dialogue comme montrée ci-dessous.

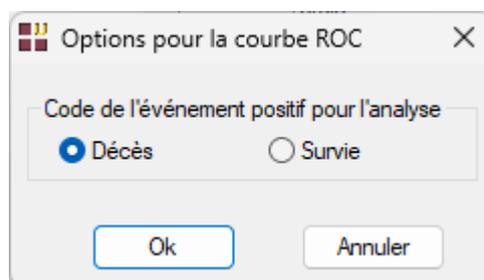


Standardisons les données et définissons la plage des valeurs de K de 1 à 15.

Par le bouton Sélection, sélectionnons la population d'apprentissage :



Puisque la variable à expliquer possède deux classes, la procédure demande de préciser le code de l'événement positif qui sera utilisé pour le tracé de la courbe ROC.

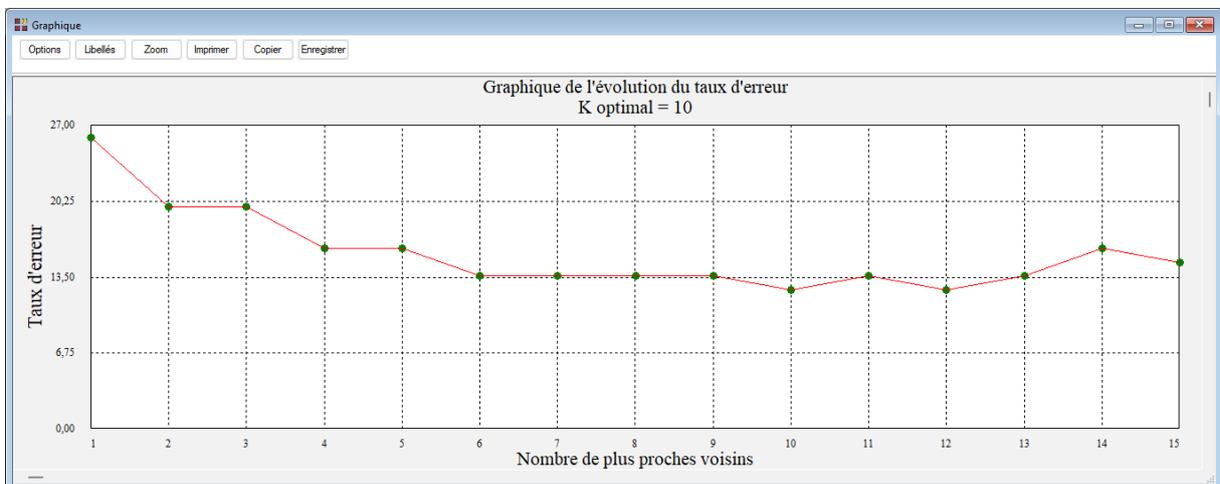


Cliquons sur le bouton Ok pour exécuter le traitement de l'analyse.

Après quelques instants, la fenêtre « Rapports et Graphiques » s'affiche.

	1	2	3	4	5	6	7	8
1								
2	(C) UNIWIN version 9.7.0							
3								
4	DATE : 29/08/2023							
5	ORDINATEUR : LAPTOP-LEGL077							
6	UTILISATEUR : cchar							
7	FICHER(S) DE DONNEES OUVERT(S) : INFARCT2.SGD							
8								
9	RESULTATS DE LA METHODE DE CLASSEMENT PAR LES K PLUS PROCHES VOISINS							
10								
11	Sélection :							
12	Et type = A							
13								
14	Jeu d'apprentissage : 81 observations							
15	Jeu de prévision : 0 observations							
16								
17	Nombre de variables explicatives : 7							
18	Prétraitement des données des variables explicatives : centrage et réduction							
19								
20	Plage pour les nombres de voisins : de 1 à 15							
21								
22	Racine aléatoire : 12345							

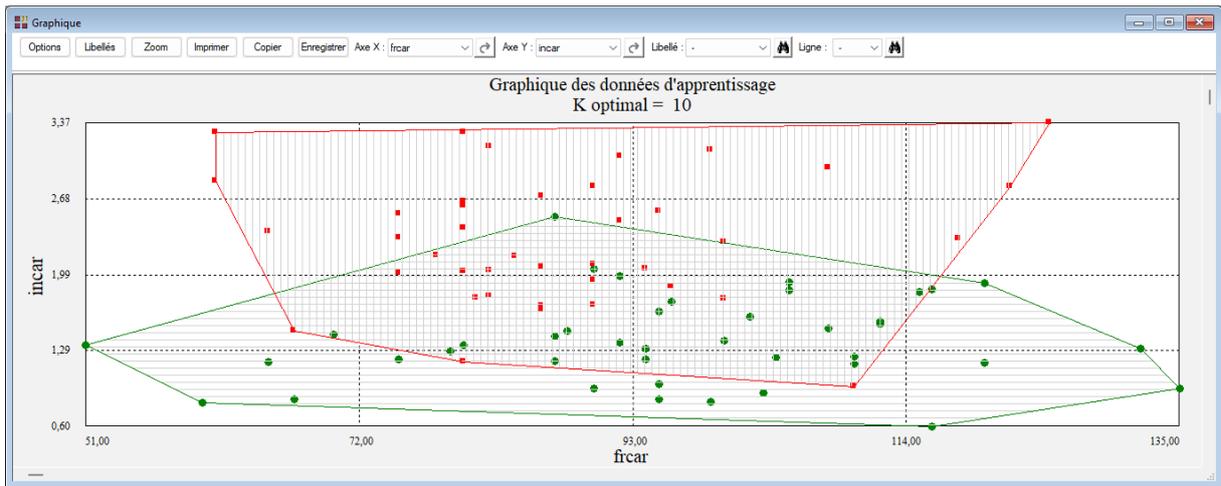
Visualisons le graphique de l'évolution du taux d'erreur.



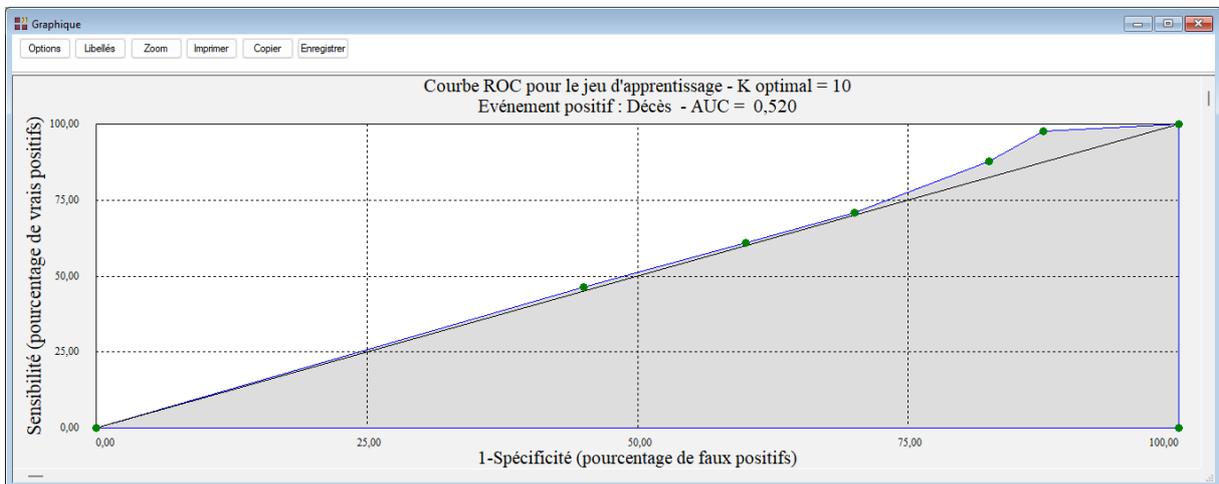
La valeur optimale de K est égale à 10 mais le taux d'erreur est toutefois élevé, plus de 12 %.

Visualisons la synthèse du classement pour K=10 puis affichons le graphique des données d'apprentissage sans libellés.

	1	2	3	4	5	6	7	8
1	SYNTHÈSE DU CLASSEMENT DE LA POPULATION D'APPRENTISSAGE PAR VALIDATION CROISEE							
2								
3								
4	Nombre de plus proches voisins : 10							
5								
6	En lignes, les groupes observés							
7	En colonnes, les groupes prévus							
8								
9	Pourcentage de mal classés : 12,346 %							
10	Pourcentage de bien classés : 87,654 %							
11								
12								
13		Décès	Survie	Total				
14	Décès	35	6	41				
15	Survie	4	36	40				
16	Total	39	42	81				
17								
18								
19								
20								
21								
22								



La visualisation de la courbe ROC pour l'événement 'Décès' confirme la mauvaise performance du modèle élaboré.



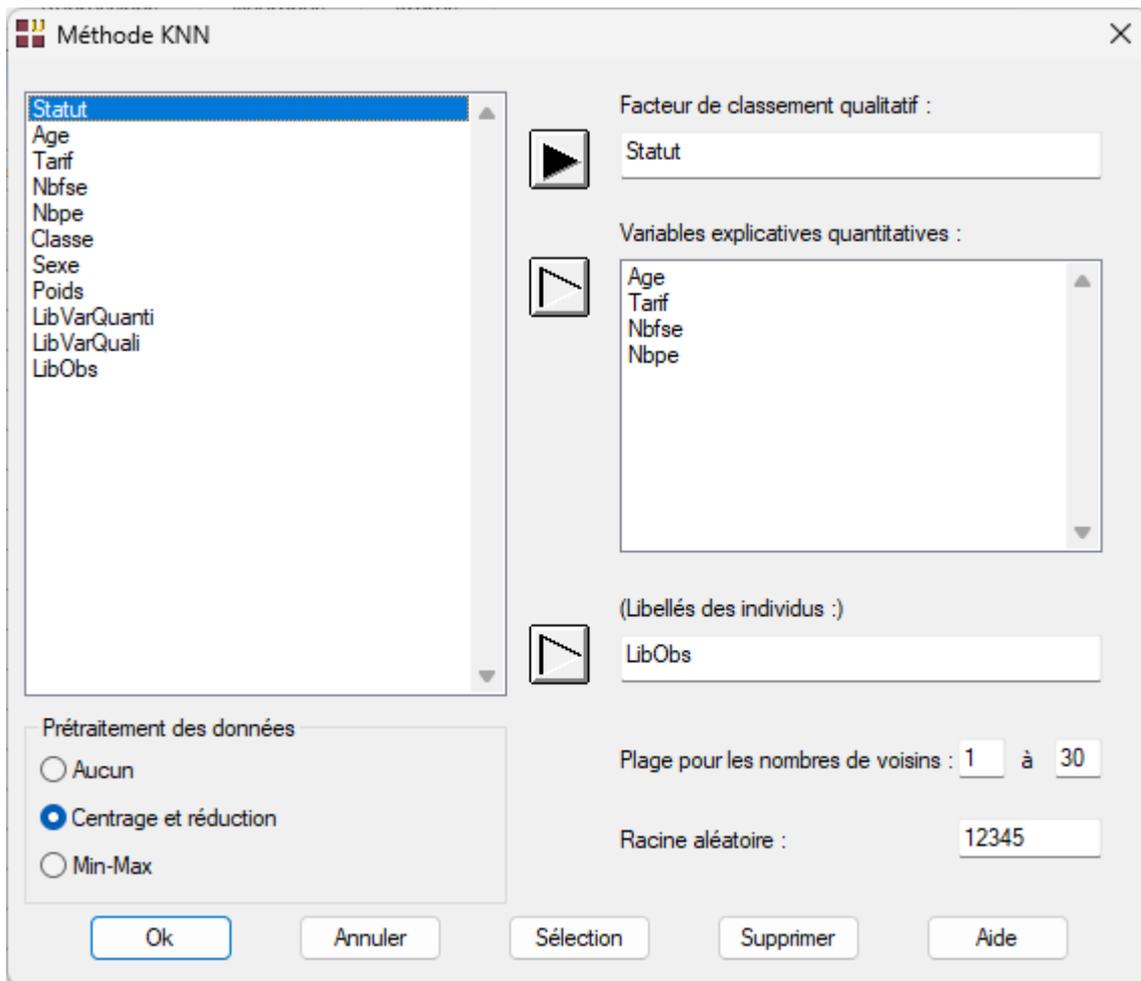
Exemple 3 : Fichier TITANIC

Pour ce troisième exemple, nous utiliserons le fichier TITANIC.

Ce fichier contient des informations concernant 714 passagers :

Statut	Survie ou Décès
Classe	Classe du passager (1 ^{ère} , 2 ^{ème} ou 3 ^{ème})
Sexe	Homme ou Femme
Age	Age du passager
Nbfse	Nombre de frères, sœurs, époux ou épouses à bord
Nbpe	Nombre de parents ou enfants à bord
Tarif	Tarif passager (en £)

Renseignons la boîte de dialogue comme montré ci-après.



Après exécution de la procédure, visualisons pour le K optimal égal à 11 le tableau de classement des données et la courbe ROC associée.

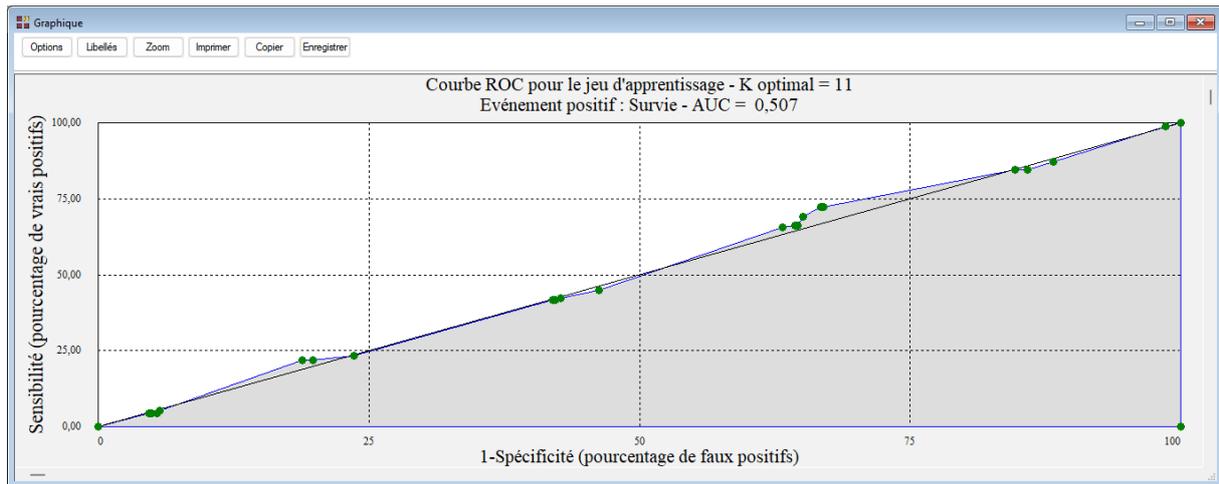
Rapports et Graphiques

Classement apprentissage K=27
Classement apprentissage K=28
Classement apprentissage K=29
Classement apprentissage K=30
Synthèse apprentissage K=1
Synthèse apprentissage K=2
Synthèse apprentissage K=3
Synthèse apprentissage K=4
Synthèse apprentissage K=5
Synthèse apprentissage K=6
Synthèse apprentissage K=7
Synthèse apprentissage K=8
Synthèse apprentissage K=9
Synthèse apprentissage K=10
Synthèse apprentissage K=11
Synthèse apprentissage K=12
Synthèse apprentissage K=13
Synthèse apprentissage K=14
Synthèse apprentissage K=15
Synthèse apprentissage K=16
Synthèse apprentissage K=17
Synthèse apprentissage K=18
Synthèse apprentissage K=19
Synthèse apprentissage K=20
Synthèse apprentissage K=21
Synthèse apprentissage K=22
Synthèse apprentissage K=23
Synthèse apprentissage K=24

	1	2	3	4	5	6	7	8
1								
2	SYNTHÈSE DU CLASSEMENT DE LA POPULATION D'APPRENTISSAGE PAR VALIDATION CROISEE							
3								
4	Nombre de plus proches voisins : 11							
5								
6	En lignes, les groupes observés							
7	En colonnes, les groupes prévus							
8								
9	Pourcentage de mal classés : 26,611 %							
10	Pourcentage de bien classés : 73,389 %							
11								
12								
13		Décès	Survie	Total				
14	Décès	364	60	424				
15	Survie	130	160	290				
16	Total	494	220	714				
17								
18								
19								
20								
21								
22								

Rapport Explorateur /

Le pourcentage d'erreur de classement est de près de 27 % et la courbe ROC (événement positif 'Survie') confirme la mauvaise performance du modèle élaboré.



Note : Pour comparer les performances de plusieurs méthodes d'analyse, cet exemple est traité dans les six analyses AFD, ADB, KNN, BAYES, ANN et ARBRE.

Les variables créées par la procédure

Voici la liste des variables créées par la procédure pour chaque valeur de K.

<i>Variable</i>	<i>Contenu</i>
applibind	libellés des individus du jeu d'apprentissage
applind	classes des individus du jeu d'apprentissage
appaflect	affectations pour le jeu d'apprentissage
appprop	proportions des votes pour le jeu d'apprentissage
prevlibind	libellés des individus du jeu de prévision
prevaflect	affectations pour le jeu de prévision
prevprop	proportions des votes pour le jeu de prévision
appaflectopt	affectations pour le jeu d'apprentissage (K optimal)
apppropopt	proportions des votes pour le jeu d'apprentissage (K optimal)
prevaflectopt	affectations pour le jeu de prévision (K optimal)
prevpropopt	proportions des votes pour le jeu de prévision (K optimal)
vp	vrais positifs
fn	faux négatifs
fp	faux positifs
vn	vrais négatifs
specificite	spécificité
sensibilité	sensibilité

Références

Documentation du package R 'class' (2021)

<https://cran.r-project.org/web/packages/class/class.pdf>

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge.

Venables, W. N. et Ripley, B. D. (2002) *Modern Applied Statistics with S*. 4ième édition. Springer.