

## UNIWIN VERSION 9.7.0

# ARBRES DE DECISION ET DE REGRESSION

Révision : 02/09/2023

Définition.....	1
Entrée des données .....	2
Données manquantes ou non sélectionnées.....	3
Exemple 1 : Fichier IRIS3 (arbre de décision) .....	3
L'option Rapports .....	7
L'option Graphiques .....	12
Exemple 2 : Fichier DIABETES (arbre de décision) .....	14
Exemple 3 : Fichier GRAISSE (arbre de régression).....	18
Exemple 4 : Fichier WINES3 (arbre de régression).....	23
Exemple 5 : Fichier TITANIC .....	28
Les variables internes créées par la procédure .....	29
Références .....	30

### Définition

Les arbres de décision et de régression sont des méthodes permettant d'obtenir des modèles explicatifs et prédictifs. Ils sont faciles à comprendre du fait de l'affichage des résultats sous la forme d'arbres et de la génération d'un ensemble de règles en langage naturel. Les arbres de décision (classement) permettent d'expliquer et de prévoir l'appartenance d'observations à une classe d'une variable qualitative en se basant sur un ensemble de variables explicatives quantitatives et qualitatives. Les arbres de régression permettent d'expliquer et de prévoir la valeur prise par une variable quantitative à expliquer en fonction de variables explicatives quantitatives et qualitatives.

Les données brutes sont utilisées pour les calculs car la structure de l'arbre n'est pas impactée par les habituelles transformations monotones des données.

La procédure propose l'étude des jeux d'apprentissage, de validation et de prévision. Un rapport général de synthèse est construit ainsi que les graphiques des coefficients de complexité, de l'importance des variables, des arbres complet et élagué, de la courbe ROC (décision), des valeurs estimées par rapport aux valeurs observées (régression) et des résidus par rapport aux valeurs estimées (régression).

Cette procédure est basée sur les packages R 'rpart' et 'rpart.plot'.

## Entrée des données

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

Arbres de décision et de régression

Variable à expliquer :

Variables explicatives quantitatives :

Variables explicatives qualitatives :

(Poids des observations :)

(Libellés des variables quantitatives :)

(Libellés des variables qualitatives :)

(Libellés des observations :)

Type d'arbre :  
 Classement  Régression

Mesure de l'impureté (classement) :  
 Indice de Gini  Gain d'information

Taille minimale pour découpage : 5

Taille minimale d'un noeud terminal : 2

Profondeur maximale de l'arbre : 30

Coefficient de complexité : 0,01

Nombre de validations croisées : 10

Racine aléatoire : 1023129506

Ok Annuler Sélection Supprimer Aide

Cette boîte de dialogue permet de définir la variable à expliquer, les variables explicatives quantitatives et qualitatives et les poids optionnels des observations (par défaut tous égaux à 1).

Elle permet également, en option, d'indiquer les noms des variables contenant les libellés des variables quantitatives et qualitatives et les libellés des observations.

Le type d'arbre peut être précisé : 'Classement' (décision) pour une variable à expliquer qualitative alphanumérique, 'Régression' pour une variable à expliquer quantitative.

Dans le cas d'un arbre de décision, la mesure de l'impureté des nœuds peut être choisie : indice de Gini ou gain d'information.

Pour un ensemble d'observations appartenant à  $K$  classes où  $p_k$  est la fraction des observations dans la classe  $k$  :

Indice de Gini

$$G = \sum_{k=1}^K p_k(1 - p_k) = 1 - \sum_{k=1}^K p_k^2$$

Gain d'information

$$H = - \sum_{k=1}^K p_k \log p_k$$

Dans le cas d'un arbre de régression, c'est la somme des carrés des erreurs qui est minimisée.

Différents critères pour la construction de l'arbre peuvent être précisés : taille minimale pour un découpage, taille minimale pour un nœud terminal, profondeur maximale de l'arbre, valeur minimale du coefficient de complexité, nombre de validations croisées et racine aléatoire pour tester différents élagages de l'arbre.

## Données manquantes ou non sélectionnées

Les valeurs manquantes dans les variables à expliquer quantitatives et qualitatives ne sont pas autorisées. Les valeurs manquantes de la variable à expliquer définissent le jeu de prévision. Les observations non sélectionnées définissent le jeu de validation.

### Exemple 1 : Fichier IRIS3 (arbre de décision)

Pour ce premier exemple, nous utiliserons le fichier Iris3.

Ce fichier contient les données relatives à 150 iris de trois espèces : Iris Setosa, Iris Versicolor et Iris Virginica.

Les mesures effectuées sont : longueur du sépale (lonsepal), longueur du pétale (lonpetal), largeur du sépale (larsepal), largeur du pétale (larpetal).

Ce fichier contient 6 iris pour lesquels les classes d'appartenance sont inconnues. Ils définissent l'échantillon de prévision.

Iris Setosa (1)



Iris Versicolor (2)



Iris Virginica (3)



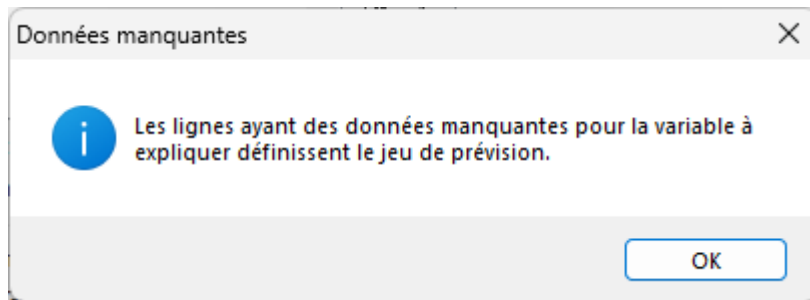
Cliquons sur l'icône ARBRE dans le ruban Expliquer.

La première boîte de dialogue montrée ci-après apparaît.

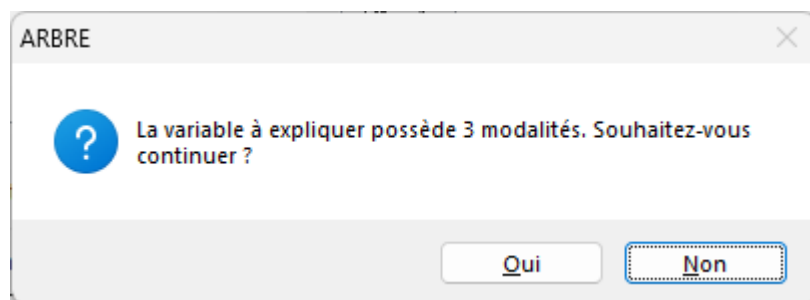
La variable codesp2 est la variable à expliquer. Elle contient pour chaque observation le libellé de son espèce d'appartenance. Nous choisissons les variables lonsepal à larpetal comme variables explicatives quantitatives et laissons les autres paramètres de l'analyse aux valeurs par défaut.

Cliquons sur le bouton Ok.

Un premier message nous indique que les lignes ayant des données manquantes pour la variable à expliquer seront utilisées comme jeu de prévision.

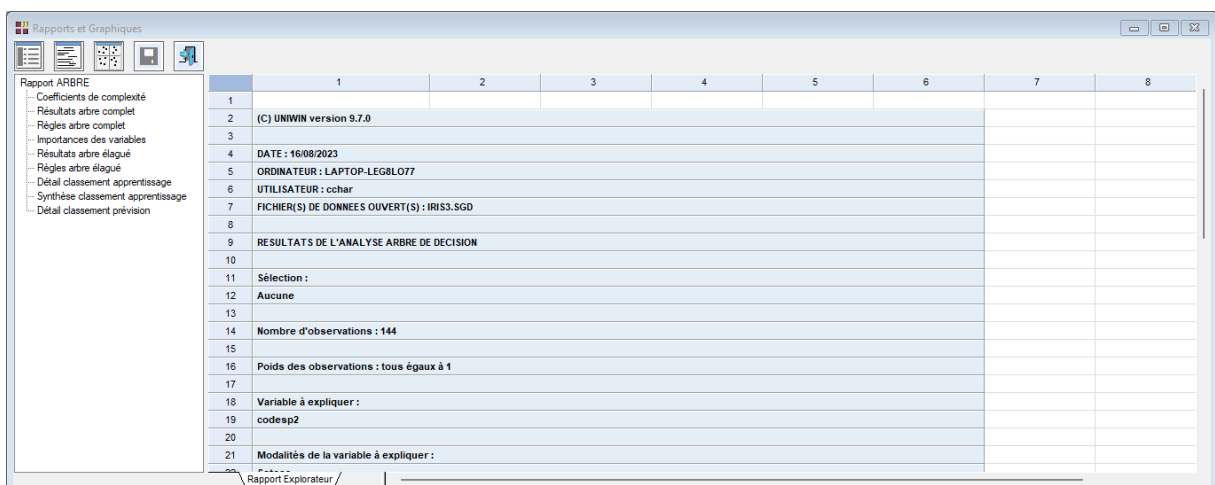



Un second message nous demande de confirmer notre choix d'un arbre de décision en fonction du nombre de modalités de la variable à expliquer :




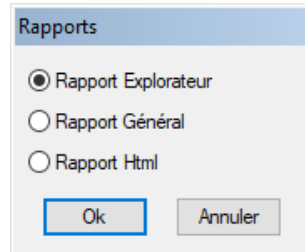
Cliquons sur Oui pour exécuter le traitement de l'analyse.


Après quelques instants, l'écran suivant s'affiche :

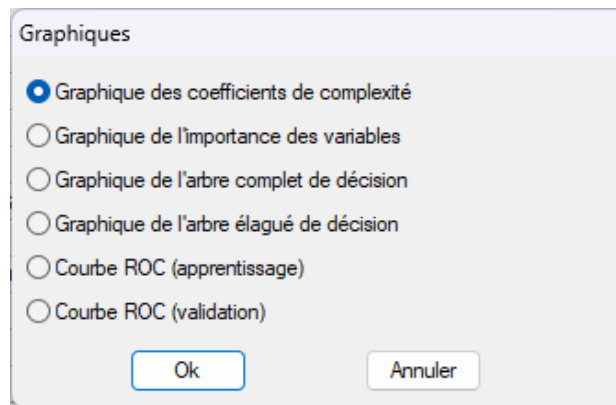



La barre d'outils 'Rapports et Graphiques' permet par l'icône 'Données'  de rappeler la boîte de dialogue d'entrée des données.

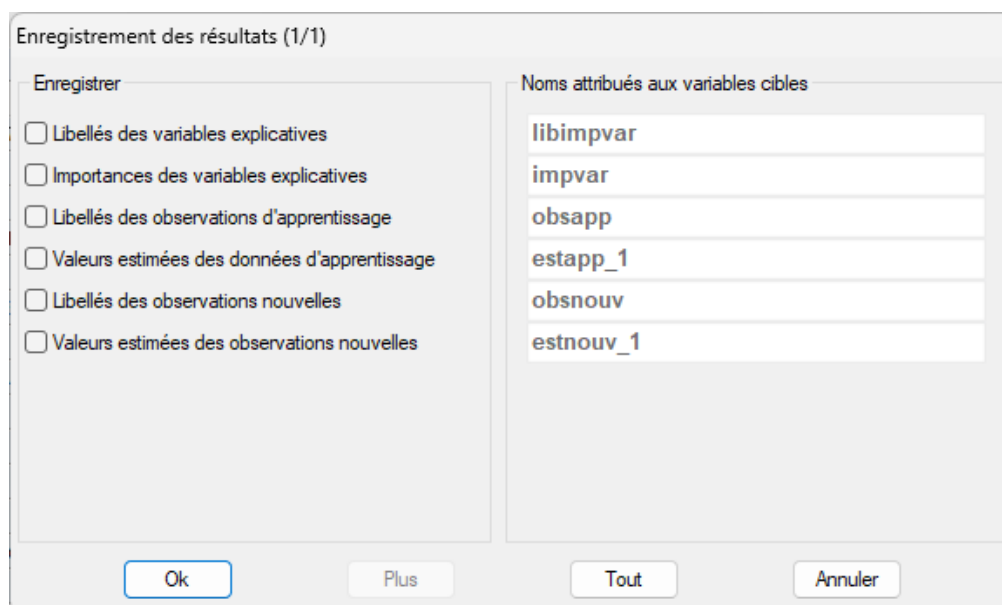
L'icône 'Rapports'  affiche la boîte de dialogue des options pour les rapports :



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.

Voici trois exemples du rapport pour notre analyse : Explorateur, Général, HTML.

Rapports et Graphiques

Rapport ARBRE

1

2 COEFFICIENTS DE COMPLEXITE

3

4 Coefficient de complexité optimal : 0,01000

5 Erreur de validation croisée : 0,07292

6 Nombre de coupures : 4

7

8

9

		Complexité	Nombre de coupures	Erreur apprentissage	Erreur validation	Ecart-type validation
10	1	0,50000	0	1,00000	1,17708	0,05138
11	2	0,43750	1	0,50000	0,68750	0,06228
12	3	0,01563	2	0,06250	0,13642	0,03682
13	4	0,01000	4	0,03125	0,07292	0,02688

14

15

16

17

18

19

20

21

Rapport Explorateur

Rapports et Graphiques

289

290 SYNTHESE DU CLASSEMENT DE LA POPULATION D'APPRENTISSAGE

291

292 En lignes, les classes observées

293 En colonnes, les classes prévues

294

295 Pourcentage de mal classés : 2,083 %

296 Pourcentage de bien classés : 97,917 %

297

Observé \ Prévu	Setosa	Versicolor	Virginica	Total
300 Setosa	48	0	0	48
301 Versicolor	0	46	2	48
302 Virginica	0	1	47	48
303 Total	48	47	49	144

304

305 DETAIL DU CLASSEMENT DE LA POPULATION DE PREVISION

306

307 Observations et probabilités d'affectation aux classes

308

309

Rapport Général

Rapports et Graphiques

RESULTATS POUR L'ARBRE COMPLET

Noeud : numéro du noeud dans l'arbre  
 Coupure : Critère de coupure de l'arbre  
 n : nombre total d'observations dans le noeud  
 Mal classées : nombre d'observations mal classées  
 Prévu : valeur prévue de la variable à expliquer  
 Probabilités : probabilités des classes

n= 144

Noeud), Coupure, n, Mal classées, Prévu, (Proba.)

- racine 144 96 Setosa (0.33333333 0.33333333 0.33333333)
- lonpetal< 2.45 48 0 Setosa (1.00000000 0.00000000 0.00000000) \*
- lonpetal>=2.45 96 48 Versicolor (0.00000000 0.50000000 0.50000000)
- lonpetal< 4.75 44 1 Versicolor (0.00000000 0.97727273 0.02272727) \*
- lonpetal>=4.75 52 5 Virginica (0.00000000 0.09615385 0.90384615)
- larpetal< 1.75 8 4 Versicolor (0.00000000 0.50000000 0.50000000)
- lonpetal< 4.95 3 0 Versicolor (0.00000000 1.00000000 0.00000000) \*
- lonpetal>=4.95 5 1 Virginica (0.00000000 0.20000000 0.80000000) \*
- larpetal>=1.75 44 1 Virginica (0.00000000 0.02272727 0.97727273) \*

REGLES POUR L'ARBRE COMPLET

Classe prévue

Ce rapport contient les informations suivantes :

Coefficients de complexité : une fois l'arbre initial construit en utilisant la valeur du coefficient de complexité précisée dans la boîte de dialogue d'entrée des données, si le nombre de nœuds terminaux est jugé trop grand, on peut le simplifier en élaguant ses branches de bas en haut. Un élagage judicieux s'arrête quand on atteint un bon compromis entre la complexité de l'arbre et la précision de la prévision. Ce compromis se calcule par validation croisée en testant différentes versions élaguées de l'arbre. Un élagage judicieux correspond à une valeur du paramètre de complexité rendant petite l'erreur de validation croisée. La valeur optimale de la complexité est alors utilisée automatiquement comme règle d'arrêt pour créer le nouvel arbre élagué.

COEFFICIENTS DE COMPLEXITE					
Coefficient de complexité optimal : 0,01000					
Erreur de validation croisée : 0,07292					
Nombre de coupures : 4					
	Complexité	Nombre de coupures	Erreur apprentissage	Erreur validation	Ecart-type validation
1	0,50000	0	1,00000	1,17708	0,05138
2	0,43750	1	0,50000	0,68750	0,06228
3	0,01563	2	0,06250	0,13542	0,03582
4	0,01000	4	0,03125	0,07292	0,02688

Dans cet exemple, le coefficient de complexité optimal est égal à 0,01 pour une erreur de validation croisée de 0,07292 et un nombre de coupures égal à 4.

Note : s'il est souhaité utiliser une autre valeur du coefficient de complexité, il suffit de rappeler la boîte de dialogue d'entrée des données (via l'icône 'Données' de la barre d'outils) et de préciser cette valeur dans le champ 'Coefficient de complexité'.

Résultats pour l'arbre complet : ce tableau décrit la construction de l'arbre complet. Il indique pour chaque nœud le numéro du nœud, le critère de coupure, le nombre total d'observations dans le nœud, le nombre d'observations mal classées, la valeur prévue de la variable à expliquer, les probabilités d'affectation aux différentes classes.

Par exemple le nœud 3 est défini par le critère 'lonpetal >= 2,45'. Ce nœud contient 96 observations dont 48 sont mal classées. Les observations de cette classe sont prévues 'Versicolor'. La probabilité d'affectation à la classe 'Setosa' est égale à 0, celle à la classe 'Versicolor' est égale à 0,5 ainsi que celle à la classe 'Virginica'.

Le nœud 15 est défini par le critère 'larpetal >= 1,75'. Ce nœud contient 44 observations dont 1 est mal classée. Les observations de cette classe sont prévues 'Virginica'. La probabilité d'affectation à la classe 'Setosa' est égale à 0, celle à la classe 'Versicolor' est égale à 0,023 et celle à la classe 'Virginica' est égale à 0,977. Le symbole \* indique que ce nœud est un nœud terminal (feuille).



RESULTATS POUR L'ARBRE COMPLET		
Noeud : numéro du noeud dans l'arbre		
Coupure : critère de coupure de l'arbre		
n : nombre total d'observations dans le noeud		
Mal classées : nombre d'observations mal classées		
Prévu : valeur prévue de la variable à expliquer		
Probabilités : probabilités des classes		
n= 144		
* indique un noeud terminal		
Noeud), Coupure, n, Mal classées, Prévu, (Proba.)		
1)	racine 144 96	Setosa (0.33333333 0.33333333 0.33333333)
2)	lonpetal < 2.45 48 0	Setosa (1.00000000 0.00000000 0.00000000) *
3)	lonpetal >= 2.45 96 48	Versicolor (0.00000000 0.50000000 0.50000000)
6)	lonpetal < 4.75 44 1	Versicolor (0.00000000 0.97727273 0.02272727) *
7)	lonpetal >= 4.75 52 5	Virginica (0.00000000 0.09615385 0.90384615)
14)	larpetal < 1.75 8 4	Versicolor (0.00000000 0.50000000 0.50000000)
28)	lonpetal < 4.95 3 0	Versicolor (0.00000000 1.00000000 0.00000000) *
29)	lonpetal >= 4.95 5 1	Virginica (0.00000000 0.20000000 0.80000000) *
15)	larpetal >= 1.75 44 1	Virginica (0.00000000 0.02272727 0.97727273) *

Règles pour l'arbre complet : ce tableau décrit les règles construites par l'arbre complet. Pour chaque règle, il indique la classe prévue et les probabilités d'appartenance des observations de la règle aux différentes classes.

Par exemple, la première règle indique que si 'lonpetal < 2,5', alors la classe prévue est 'Setosa'.

La deuxième règle indique que si la règle est 'lonpetal est compris entre 2,5 et 4,8', alors la classe prévue est 'Versicolor' avec une probabilité de 0,98 et 'Virginica' avec une probabilité de 0,02.

REGLES POUR L'ARBRE COMPLET	
Classe prévue	
Setosa [1.00 .00 .00]	lorsque lonpetal < 2.5
Versicolor [ .00 .98 .02]	lorsque lonpetal est 2.5 à 4.8
Versicolor [ .00 1.00 .00]	lorsque lonpetal est 4.8 à 5.0 & larpetal < 1.8
Virginica [ .00 .20 .80]	lorsque lonpetal >= 5.0 & larpetal < 1.8
Virginica [ .00 .02 .98]	lorsque lonpetal >= 4.8 & larpetal >= 1.8

Toutes ces informations sont affichées graphiquement dans l'arbre complet de décision.

Importances des variables : ce tableau affiche l'importance des variables explicatives dans l'ajustement de l'arbre. A noter qu'il est possible que des variables non utilisées dans l'arbre soient présentes dans ce tableau car l'algorithme 'rpart' gère les variables de substitution, c'est-à-dire des variables qui ne sont pas choisies pour les divisions, mais qui étaient sur le point de remporter la compétition.

IMPORTANCES DES VARIABLES EXPLICATIVES (%)	
	Importance
lonpetal	33,94663
larpetal	31,59906
lonsepal	21,15181
larsepal	13,30250

Résultats arbre élagué :

Même interprétation que pour l'arbre complet.

Règles arbre élagué :

Même interprétation que pour l'arbre complet.

Détail classement apprentissage : ce tableau indique pour chaque observation du jeu d'apprentissage la classe observée et les probabilités d'affectation aux différentes classes. Les observations mal classées sont indiquées par le symbole \*.

DETAIL DU CLASSEMENT DE LA POPULATION D'APPRENTISSAGE			
Observations, classes observées et probabilités d'affectation aux classes			
(*) = observation mal classée.			
Observation - Classe observée	Setosa	Versicolor	Virginica
o1 - Setosa	1	0,00000	0,00000
o2 - Setosa	1	0,00000	0,00000
o4 - Setosa	1	0,00000	0,00000
o5 - Setosa	1	0,00000	0,00000
o6 - Setosa	1	0,00000	0,00000
o7 - Setosa	1	0,00000	0,00000
o8 - Setosa	1	0,00000	0,00000
o9 - Setosa	1	0,00000	0,00000
o10 - Setosa	1	0,00000	0,00000
o11 - Setosa	1	0,00000	0,00000

Synthèse classement apprentissage : ce tableau fait la synthèse du tableau précédent et affiche le pourcentage d'erreur de classement, ici d'environ 2%.

SYNTHESE DU CLASSEMENT DE LA POPULATION D'APPRENTISSAGE				
En lignes, les classes observées				
En colonnes, les classes prévues				
Pourcentage de mal classés : 2,083 %				
Pourcentage de bien classés : 97,917 %				
Observé \ Prévu	Setosa	Versicolor	Virginica	Total
Setosa	48	0	0	48
Versicolor	0	46	2	48
Virginica	0	1	47	48
Total	48	47	49	144

Détail classement validation : ce tableau ne s'affiche pas car il n'y a pas de jeu de validation dans cet exemple.

Synthèse classement validation : ce tableau ne s'affiche pas car il n'y a pas de jeu de validation dans cet exemple.

Détail classement prévision : ce tableau indique les probabilités d'affectation aux différentes classes des six observations dont les classes sont inconnues.

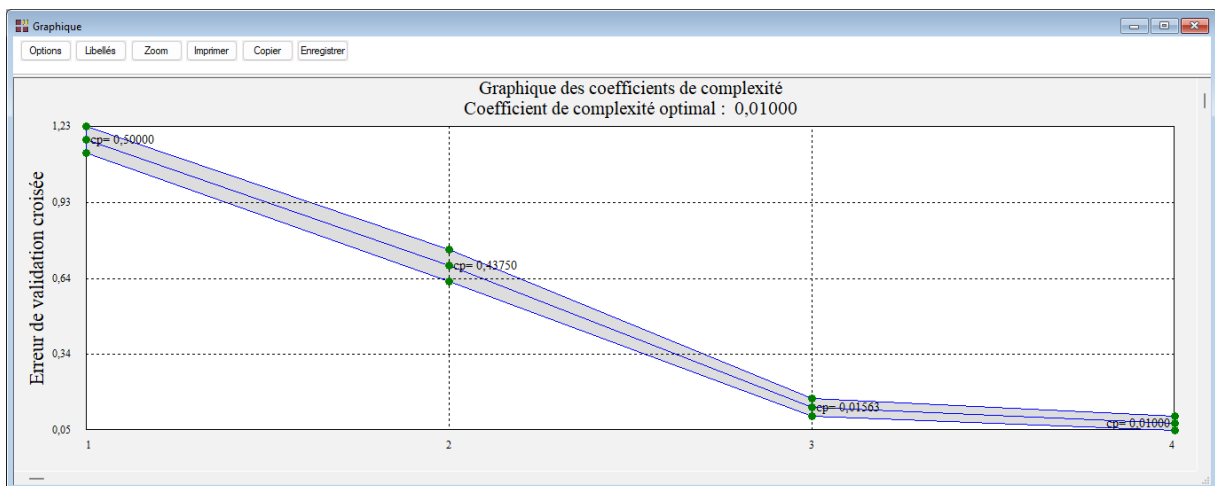
DETAIL DU CLASSEMENT DE LA POPULATION DE PREVISION			
Observations et probabilités d'affectation aux classes			
Observation	Setosa	Versicolor	Virginica
o3	1	0,00000	0,00000
o36	1	0,00000	0,00000
o62	0	0,97727	0,02273
o84	0	0,20000	0,80000
o104	0	0,02273	0,97727
o125	0	0,02273	0,97727

## L'option Graphiques

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

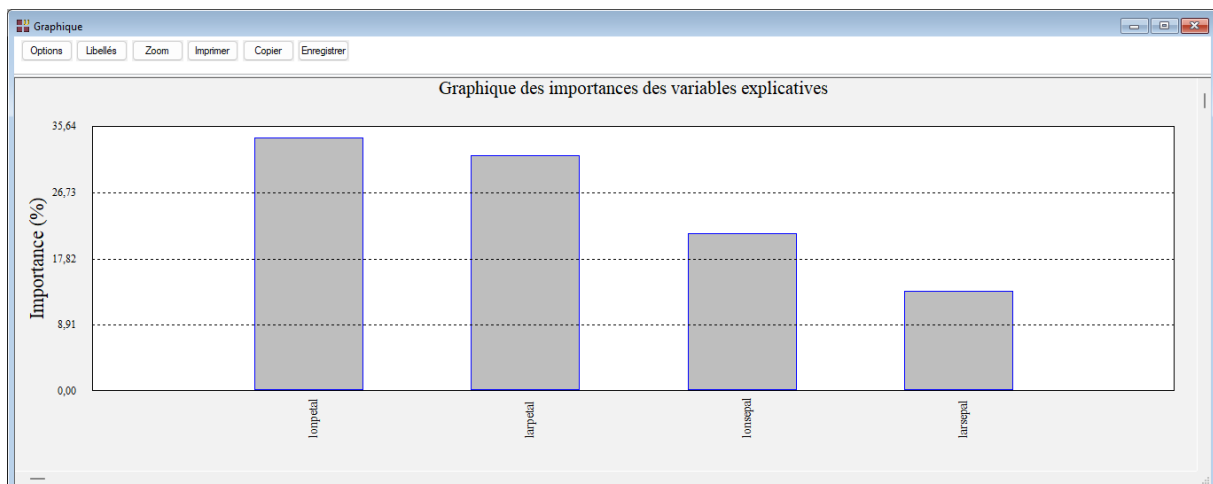
### Graphique des coefficients de complexité

Ce graphique affiche les évolutions de l'erreur de validation croisée (avec son écart-type) en fonction du nombre de la profondeur de l'arbre. Les libellés indiquent les valeurs des coefficients de complexité. Le titre du graphique précise la valeur du coefficient de complexité optimal.



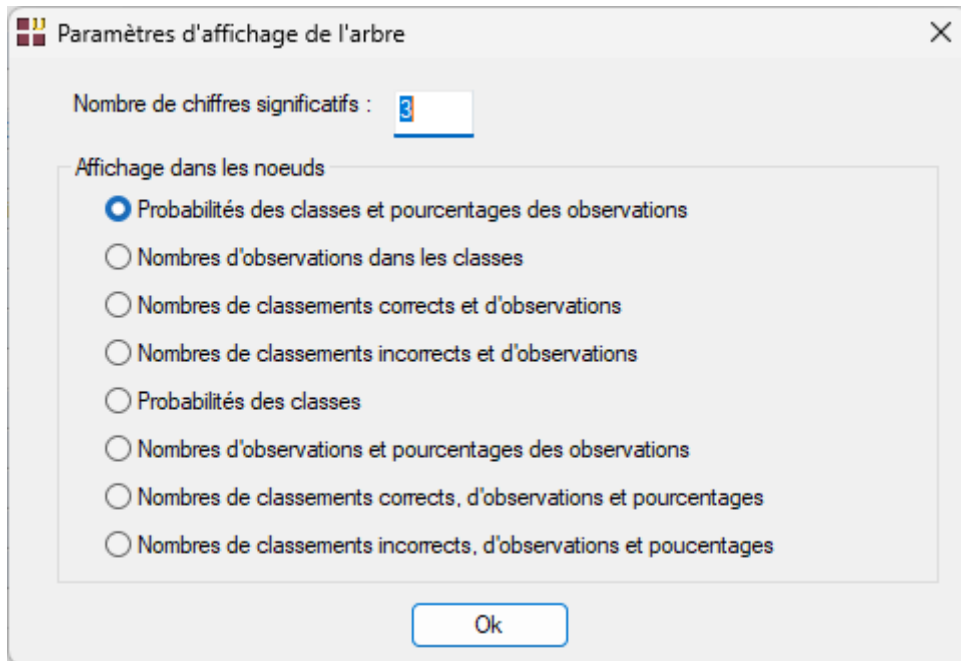
### Graphique de l'importance des variables

Ce graphique affiche les importances des variables explicatives dans l'ajustement de l'arbre.

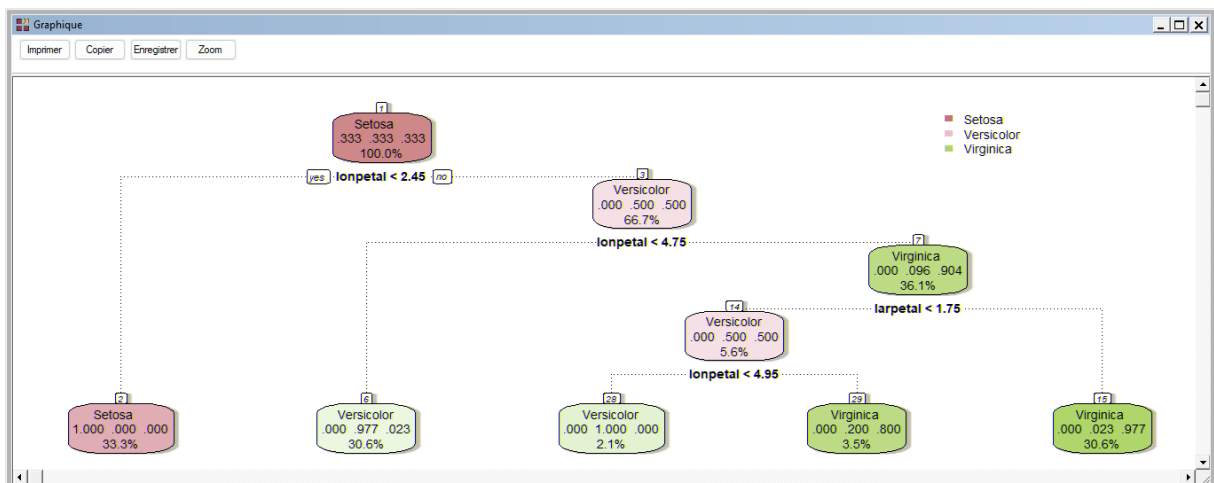


## Graphique de l'arbre complet de décision

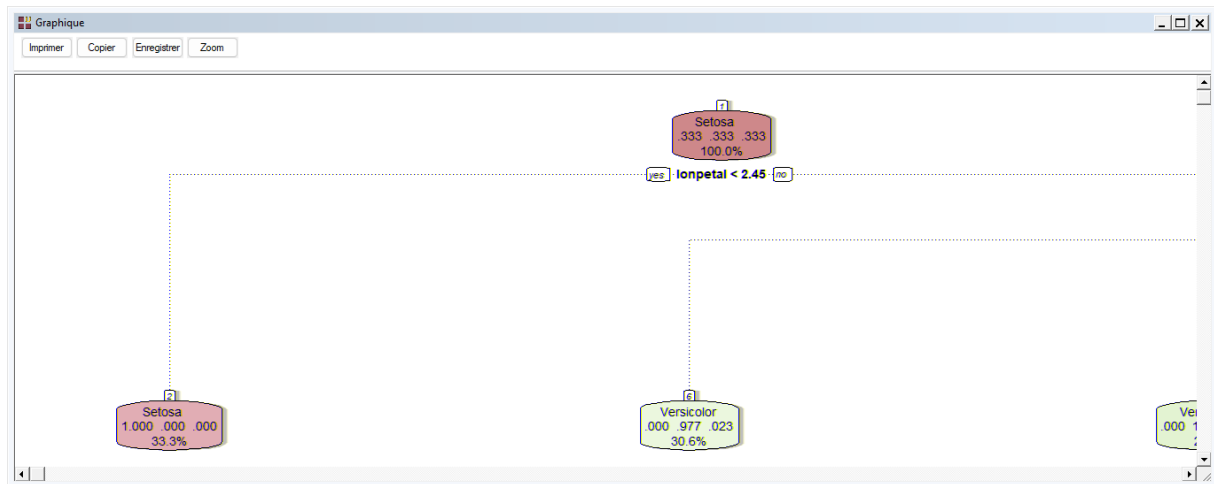
Ce graphique affiche l'arbre complet de décision. Une boîte de dialogue permet de préciser les informations qui sont affichées dans les nœuds :



Voici un exemple affichant dans chaque nœud les probabilités des classes et les pourcentages des observations.



Le bouton 'Zoom' dans la barre d'outils permet d'effectuer divers zooms en X et/ou Y dans l'arbre, ce qui est utile lorsque l'arbre devient complexe.



### Graphique de l'arbre élagué de décision

Dans cet exemple, l'arbre élagué est identique à l'arbre complet.

### Courbe ROC

La courbe ROC est disponible uniquement dans le cas de deux classes. Il y a trois classes dans cet exemple et donc le graphique n'est pas proposé.

### **Exemple 2 : Fichier DIABETES (arbre de décision)**

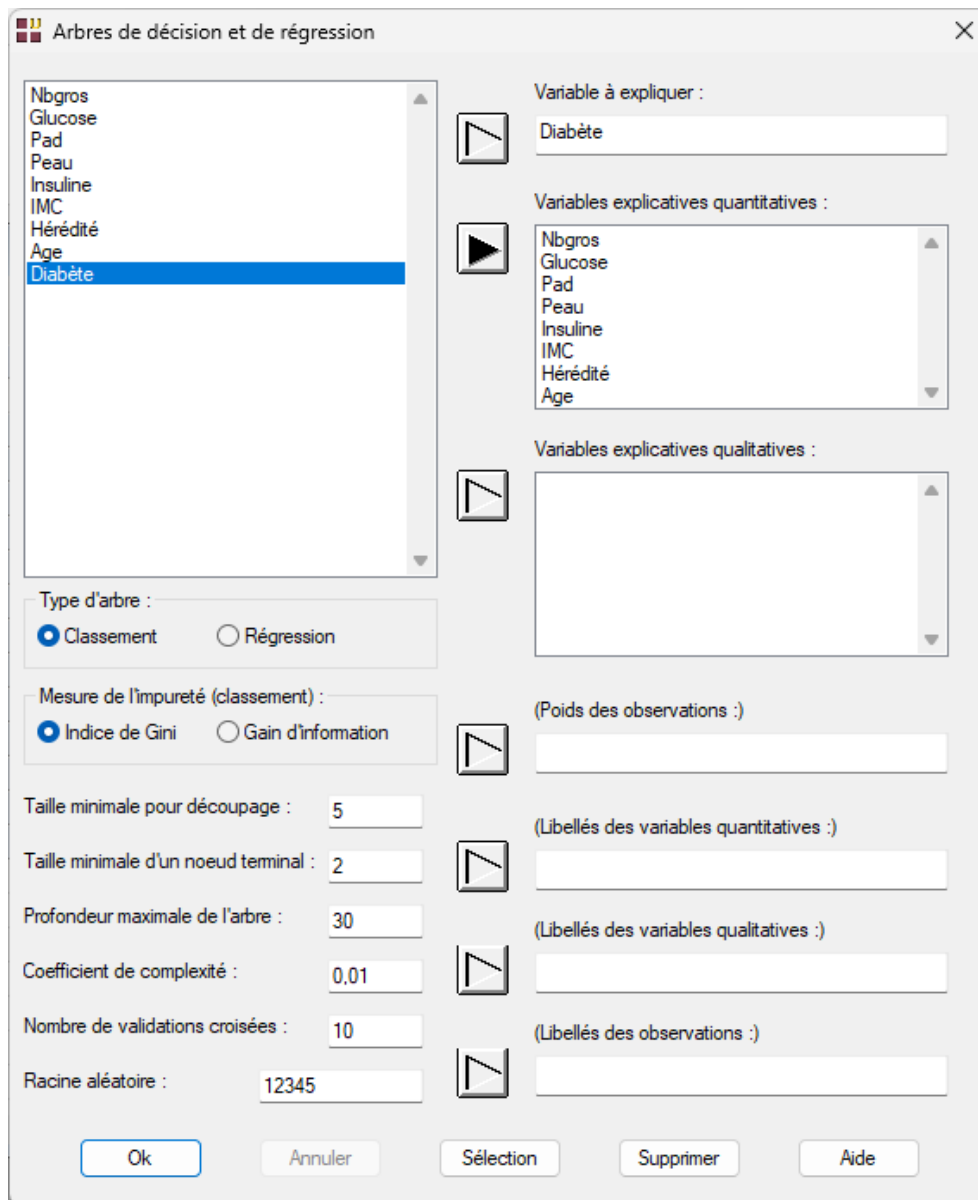
Nous utiliserons le fichier DIABETES pour ce deuxième exemple.

Une population de 768 femmes âgées d'au moins 21 ans, d'origine indienne Pima et vivant près de Phoenix, en Arizona, a été testée pour le diabète selon les critères de l'Organisation Mondiale de la Santé. Les données ont été recueillies par l'Institut national américain du diabète et des maladies digestives et rénales.

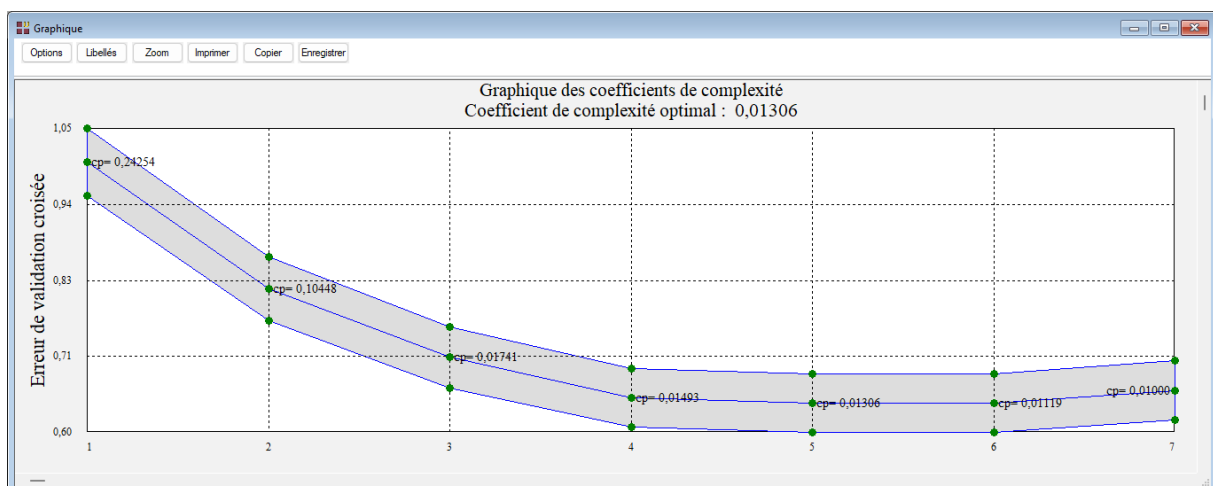
Neuf variables ont été collectées :

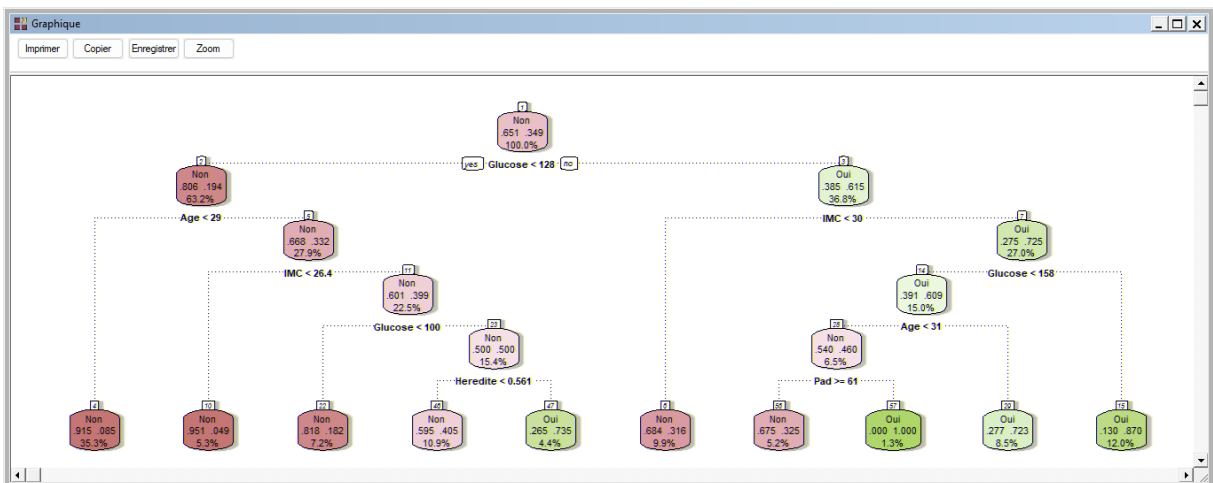
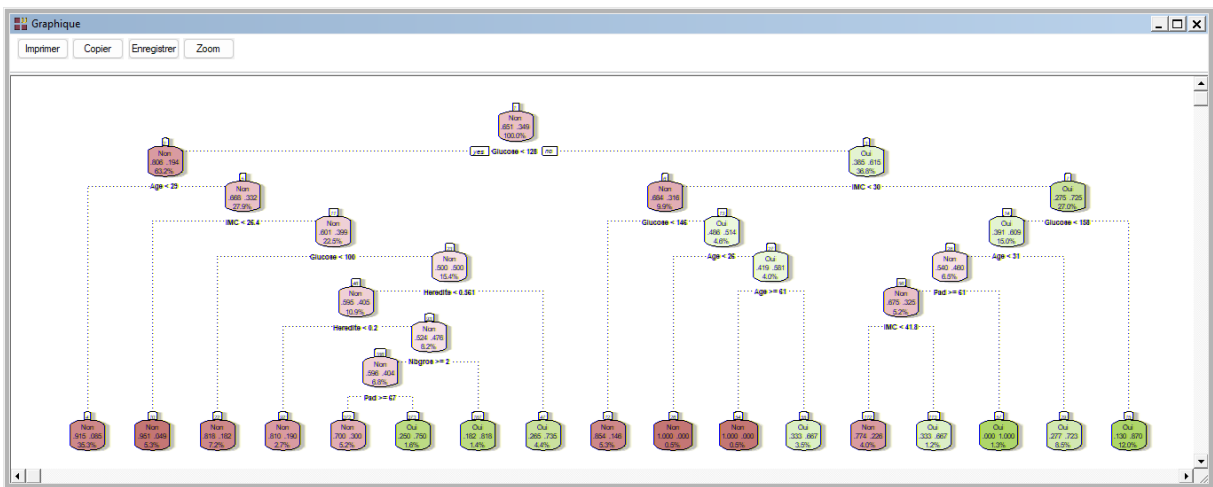
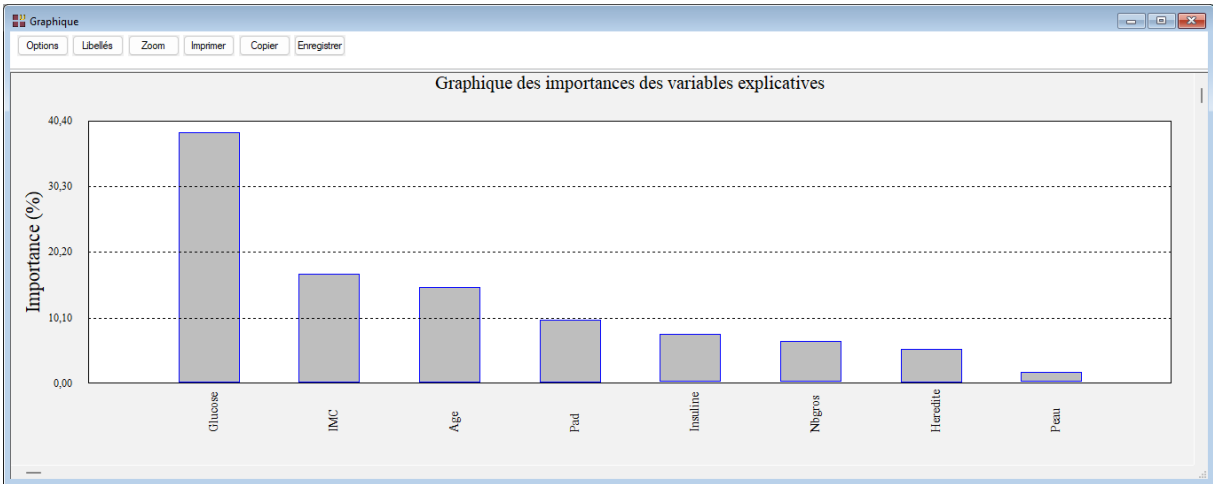
- Nbgros : nombre de grossesses
- Glucose : concentration plasmatique de glucose à 2 heures dans un test oral de tolérance au glucose
- Pad : pression artérielle diastolique (mm Hg)
- Peau : épaisseur du pli cutané du triceps (mm)
- Insuline : insuline sérique 2 heures (mu U/ml)
- IMC : indice de masse corporelle (poids en kg/(taille en m)<sup>2</sup>)
- Hérité : fonction généalogique du diabète
- Âge : âge en années
- Diabète : oui ou non

Renseignons la boîte de dialogue de l'analyse comme montré ci-après, précisons le code de l'événement positif (Oui) et exécutons l'analyse.



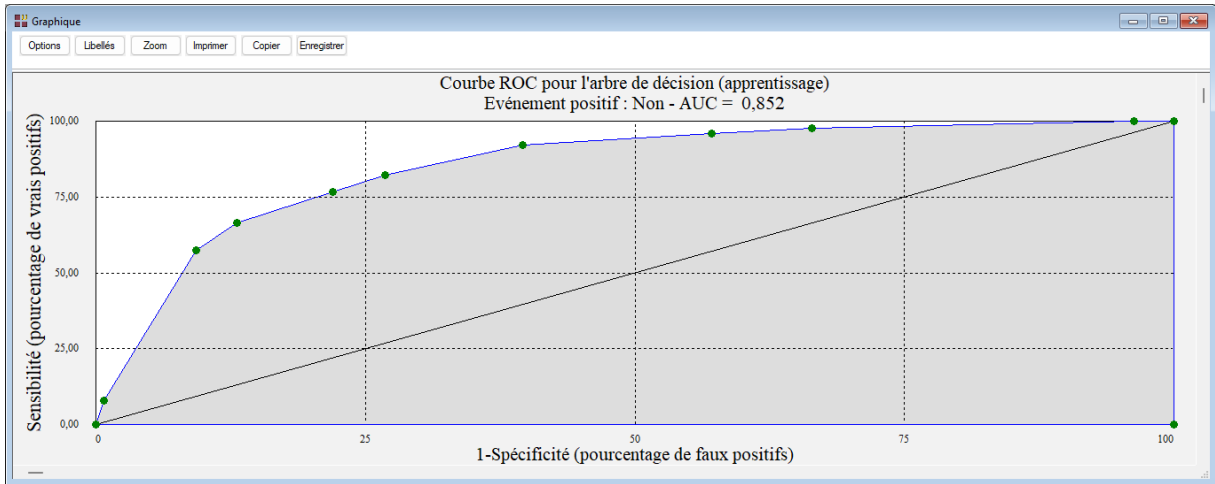
Visualisons les graphiques obtenus :





Visualisons la courbe ROC pour le jeu d'apprentissage, disponible dans cet exemple car la variable à expliquer comporte deux modalités Oui et Non.





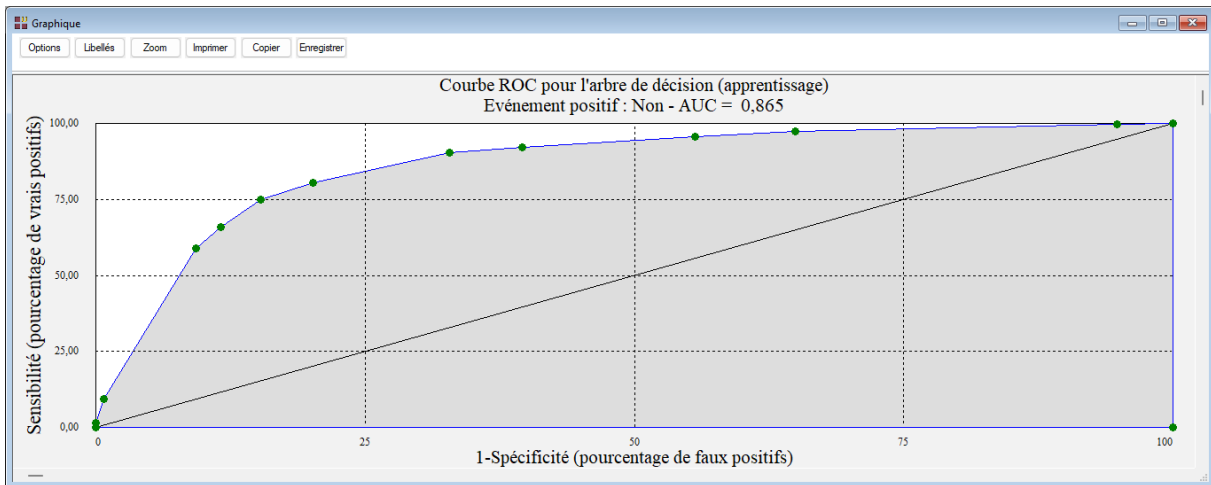
L'aire sous la courbe (AUC) nous indique l'efficacité de l'arbre. Plus la valeur de cette aire est élevée, meilleures sont les performances de l'arbre pour faire la distinction entre les classes Oui et Non.

Visualisons la synthèse du classement (18,88 % des observations sont mal classées) et le tableau des sensibilités et spécificités.

SYNTHESE DU CLASSEMENT DE LA POPULATION D'APPRENTISSAGE				
En lignes, les classes observées				
En colonnes, les classes prévues				
Pourcentage de mal classés : 18,880 %				
Pourcentage de bien classés : 81,120 %				
	Observé \ Prévu	Non	Oui	Total
Non		461	39	500
Oui		106	162	268
Total		567	201	768

Rapports et Graphiques									
<ul style="list-style-type: none"> <li>Coefficients de complexité</li> <li>Résultats arbre complet</li> <li>Règles arbre complet</li> <li>Importances des variables</li> <li>Résultats arbre élagué</li> <li>Règles arbre élagué</li> <li>Détail classement apprentissage</li> <li>Synthèse classement apprentissage</li> <li><b>VP, FN, FP, VN, Sensibilité, Spécificité</b></li> </ul>									
1	1	2	3	4	5	6	7	8	
2	VP, FN, FP, VN, SENSIBILITE, SPECIFICITE POUR LE JEU D'APPRENTISSAGE								
3									
4	Mesures - Mesures uniques								
5	VP = Nombres de vrais positifs								
6	FN = Nombres de faux négatifs								
7	FP = Nombres de faux positifs								
8	VN = Nombres de vrais négatifs								
9	Sensibilité en %								
10	Spécificité en %								
11									
12	Code de l'événement positif : Non								
13	Aire sous la courbe (AUC) = 0,852								
14									
15									
16		Mesures	VP	FN	FP	VN	Sensibilité	Spécificité	
17	1	Infini	500	0	268	0	100,0	0,0000	
18	2	0,00000	500	0	268	0	100,0	0,0000	
19	3	0,13043	500	0	258	10	100,0	3,7313	
20	4	0,26471	488	12	178	90	97,6	33,5820	
21	5	0,27692	479	21	153	115	95,8	42,9104	
22	6	0,27692	479	21	153	115	95,8	42,9104	

Réalisons à nouveau l'analyse en utilisant le gain d'information au lieu de l'indice de Gini.



SYNTHÈSE DU CLASSEMENT DE LA POPULATION D'APPRENTISSAGE				
En lignes, les classes observées				
En colonnes, les classes prévues				
Pourcentage de mal classés : 17,708 %				
Pourcentage de bien classés : 82,292 %				
Observé \ Prévu	Non	Oui	Total	
Non	452	48	500	
Oui	88	180	268	
Total	540	228	768	

L'aire sous la courbe est égale à 0,865 et 17,71 % des observations sont mal classées.

### Exemple 3 : Fichier GRAISSE (arbre de régression)

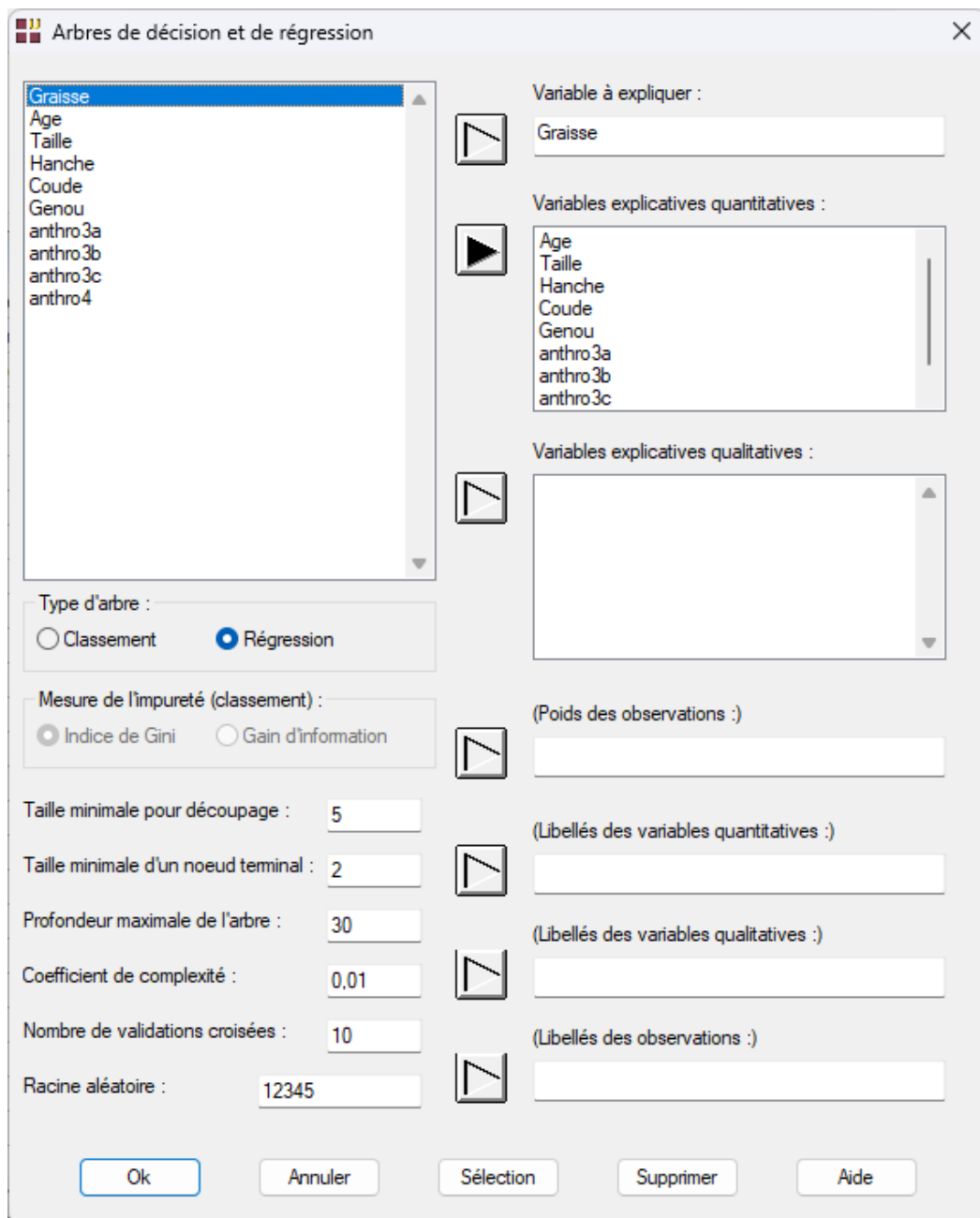
Pour 71 sujets féminins en bonne santé, neuf mesures anthropométriques sont utilisées pour modéliser la graisse corporelle.

Graisse	graisse corporelle mesurée par DXA (Dual X Ray Absorptiometry)
Age	âge (en années)
Taille	tour de taille
Hanche	tour de hanche
Coude	largeur de coude
Genou	largeur du genou

anthro3a somme du logarithme de trois mesures anthropométriques  
 anthro3b somme du logarithme de trois mesures anthropométriques  
 anthro3c somme du logarithme de trois mesures anthropométriques  
 anthro4 somme du logarithme de trois mesures anthropométriques

(source : Ada L. Garcia, Karen Wagner, Torsten Hothorn, Corinna Koebnick, Hans-Joachim F. Zunft and Ulrike Trippo (2005), Improved prediction of body fat by measuring skinfold thickness, circumferences, and bone breadths. *Obesity Research*, **13**(3), 626–634.)

Renseignons la boîte de dialogue comme montré ci-dessous et cliquons sur Ok :



Après quelques instants, la fenêtre suivante s'affiche :

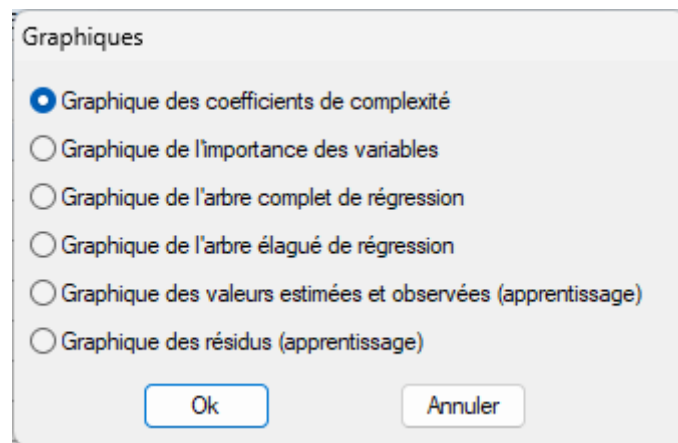
Les règles obtenues pour l'arbre élagué sont :

<b>REGLES POUR L'ARBRE ELAGUE</b>	
Valeur moyenne de la variable à expliquer	
13	lorsque Taille < 88 & anthro3c < 3.2
19	lorsque Taille < 88 & anthro3c est 3.2 à 3.4
23	lorsque Taille < 88 & Hanche < 101 & anthro3c >= 3.4
30	lorsque Taille < 88 & Hanche >= 101 & anthro3c >= 3.4
35	lorsque Taille >= 88 & Hanche < 110 & Genou < 11
43	lorsque Taille >= 88 & Hanche >= 110 & Genou < 11
61	lorsque Taille >= 88 & Genou >= 11

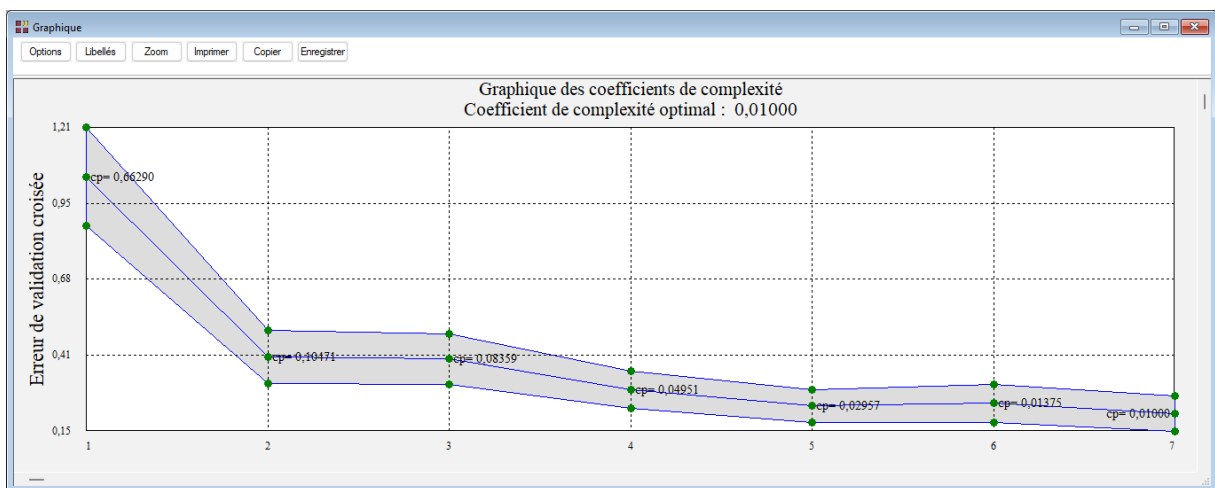
Un tableau affiche les valeurs observées, les valeurs estimées et les résidus :

<b>JEU D'APPRENTISSAGE : VALEURS OBSERVEES, ESTIMEES ET RESIDUS</b>			
	Observé	Estimé	Résidu
o1	41,68	42,95438	-1,27438
o2	43,29	42,95438	0,33563
o3	35,41	35,27846	0,13154
o4	22,79	23,31938	-0,52938
o5	36,42	35,27846	1,14154
o6	24,13	23,31938	0,81063
o7	29,83	29,54182	0,28818
o8	35,96	35,27846	0,68154
o9	23,69	23,31938	0,37063
o10	22,71	23,31938	-0,60938
o11	23,42	23,31938	0,10063

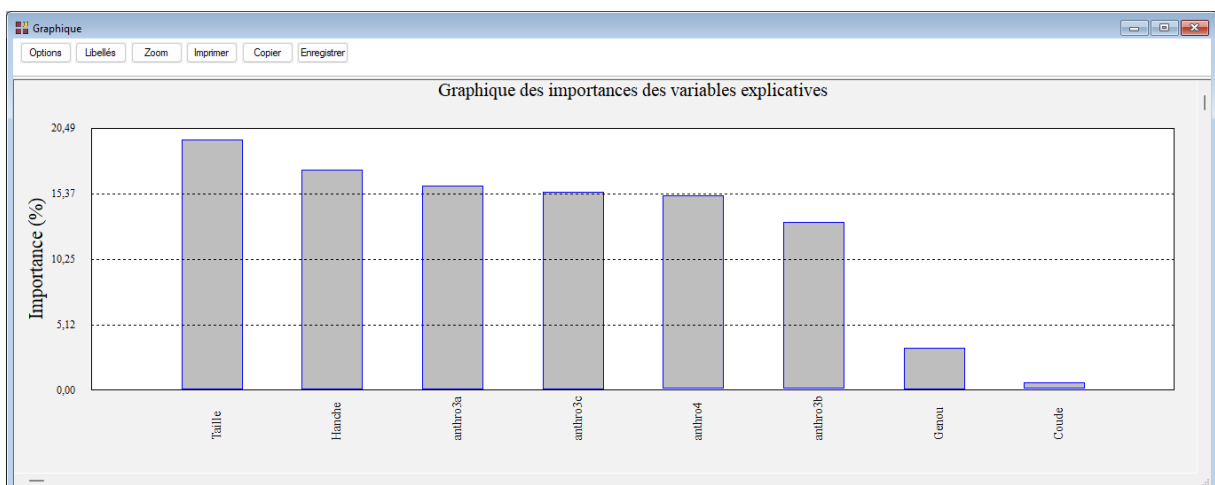
Les graphiques proposés sont :

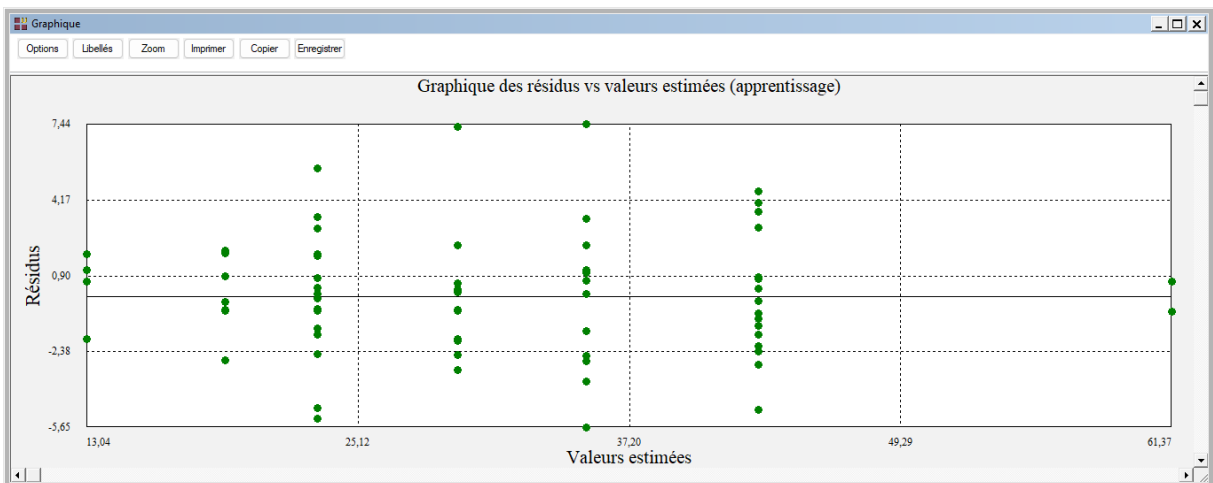
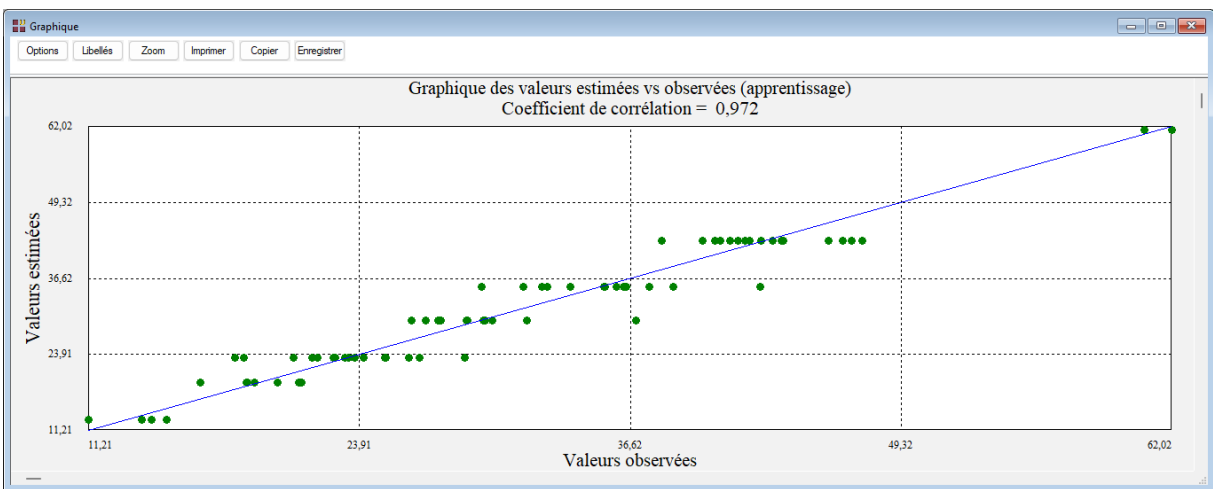
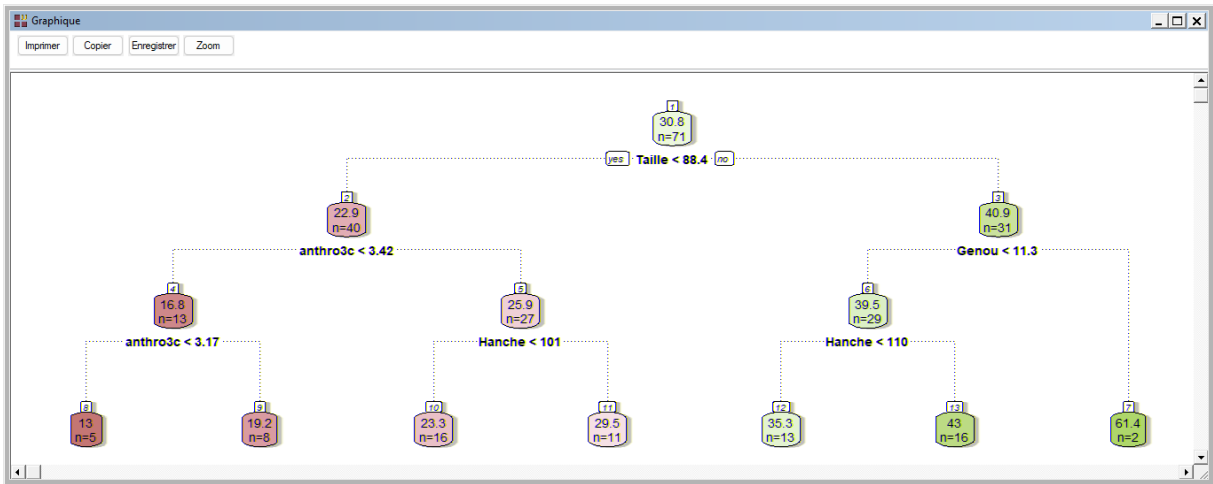


Visualisons les différents graphiques :



L'indice de complexité optimal étant celui pour le nombre maximal de coupures égal à 6, il n'y a pas dans ce cas d'élagage de l'arbre.





Les variables suivantes peuvent être enregistrées :

Enregistrement des résultats (1/1)

Enregistrer

- Libellés des variables explicatives
- Importances des variables explicatives
- Libellés des observations d'apprentissage
- Valeurs estimées des données d'apprentissage
- Résidus pour les données d'apprentissage

Noms attribués aux variables cibles

libimpvar

impvar

obsapp

estapp

residapp

Ok Plus Tout Annuler

#### Exemple 4 : Fichier WINES3 (arbre de régression)

Cet ensemble de données contient des informations concernant des variantes rouges et blanches du vin portugais « Vinho Verde » (source Cortez et al., 2009). Pour des raisons de confidentialité, seules les variables physico-chimiques (entrées) et sensorielles (sortie) sont disponibles :

- Acidité fixe
- Acidité volatile
- Acide citrique
- Sucre résiduel
- Chlorure
- SO<sub>2</sub> (teneur en dioxyde de soufre libre)
- TSO<sub>2</sub> (teneur totale en dioxyde de soufre)
- Densité
- pH
- Sulfate
- Alcool
- Qualité (note entre 0 et 10)

Il y a au total 4898 observations qui ont été aléatoirement réparties en jeu d'apprentissage (3428), jeu de validation (1225) et jeu de prévision (245).

La variable 'jeu' dans le fichier de données indique l'appartenance des observations aux trois jeux.

La variable quantitative à expliquer est la variable 'Alcool'.

Renseignons la boîte de dialogue comme montré ci-dessous et cliquons sur Ok.

Arbres de décision et de régression

Jeu  
Libobs  
Alcool  
Acidité fixe  
Acidité volatile  
Acide citrique  
Sucre résiduel  
Chlorure  
SO2  
TSO2  
Densité  
pH  
Sulfate  
Qualité

Type d'arbre :  
 Classement  Régression

Mesure de l'impureté (classement) :  
 Indice de Gini  Gain d'information

Taille minimale pour découpage : 5  
Taille minimale d'un noeud terminal : 2  
Profondeur maximale de l'arbre : 30  
Coefficient de complexité : 0,01  
Nombre de validations croisées : 10  
Racine aléatoire : 12345

Variable à expliquer :  
Alcool

Variables explicatives quantitatives :  
Acidité fixe  
Acidité volatile  
Acide citrique  
Sucre résiduel  
Chlorure  
SO2  
TSO2  
Densité

Variables explicatives qualitatives :

(Poids des observations :)  
  
(Libellés des variables quantitatives :)  
  
(Libellés des variables qualitatives :)  
  
(Libellés des observations :)

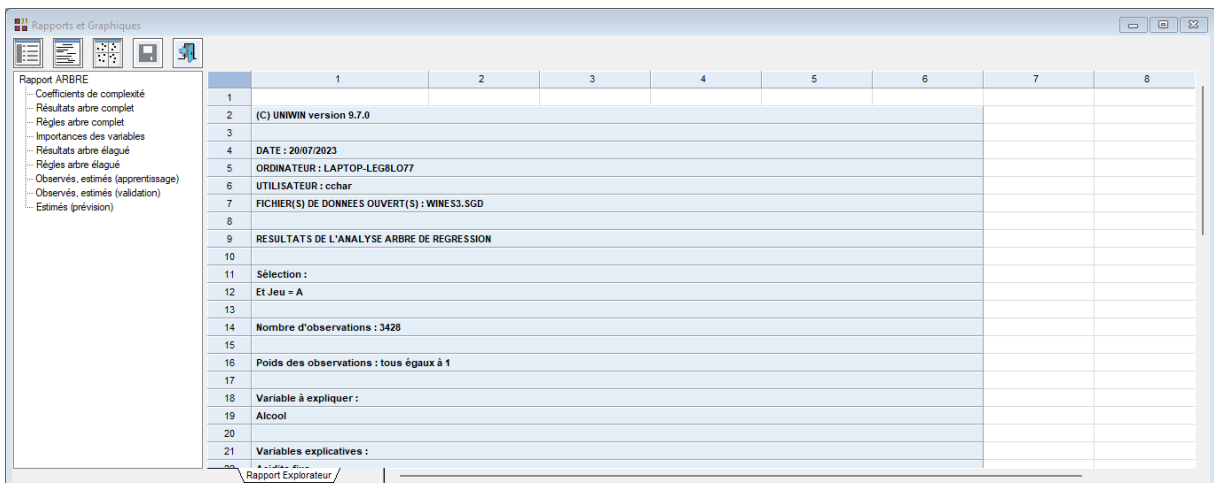
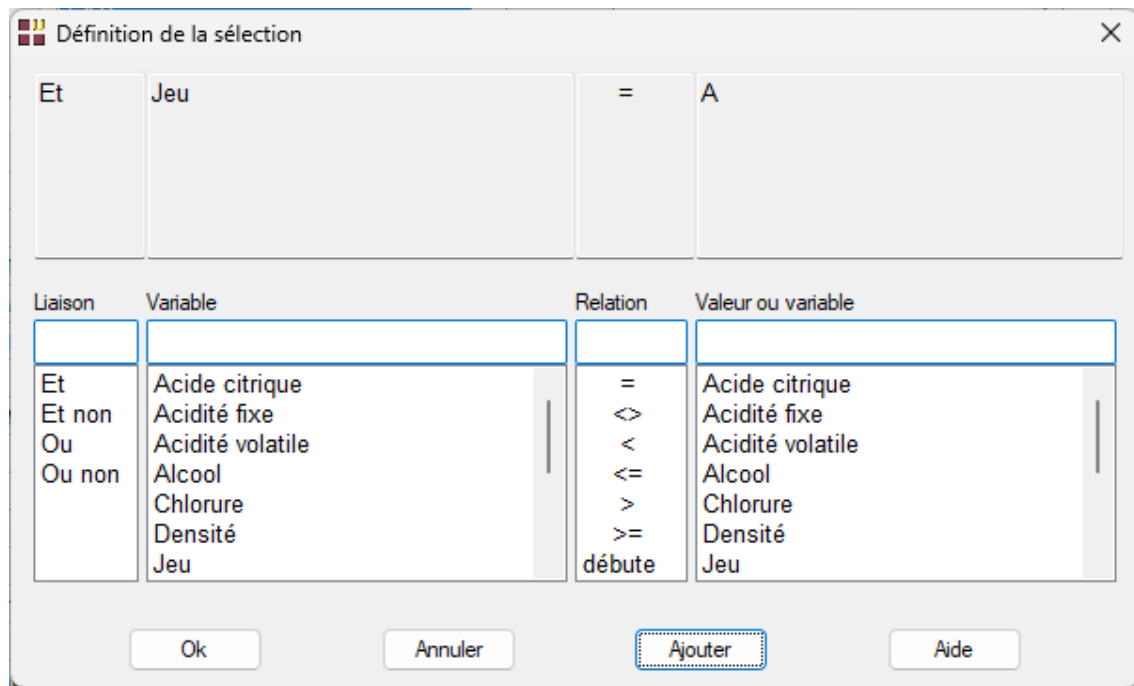
Ok Annuler Sélection Supprimer Aide

Utilisons le bouton 'Sélection' pour définir le jeu d'apprentissage puis cliquons sur Ok.

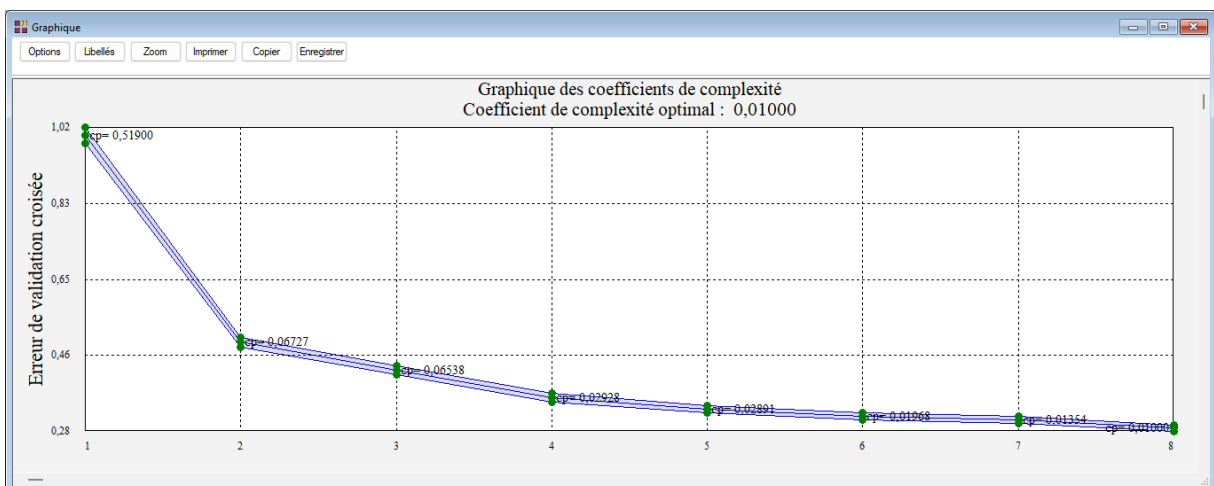
3428 observations seront ainsi utilisées comme jeu d'apprentissage, 1225 comme jeu de validation et 245 comme jeu de prévision.

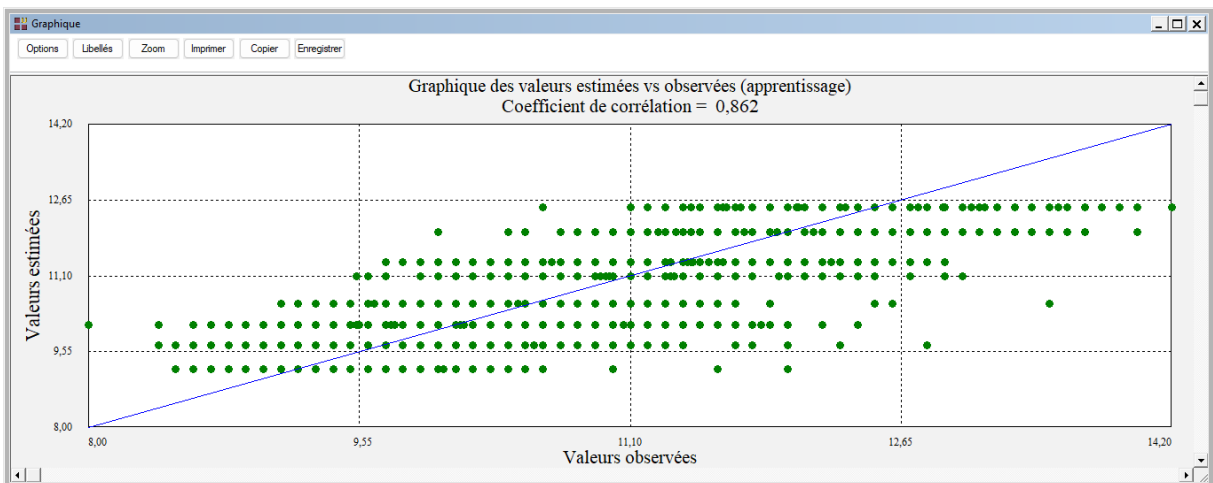
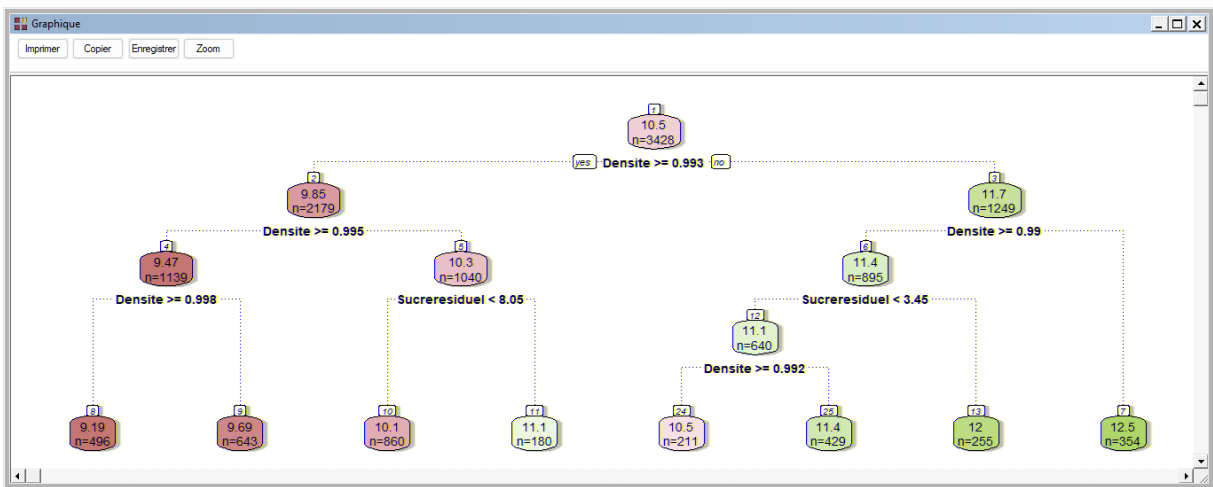
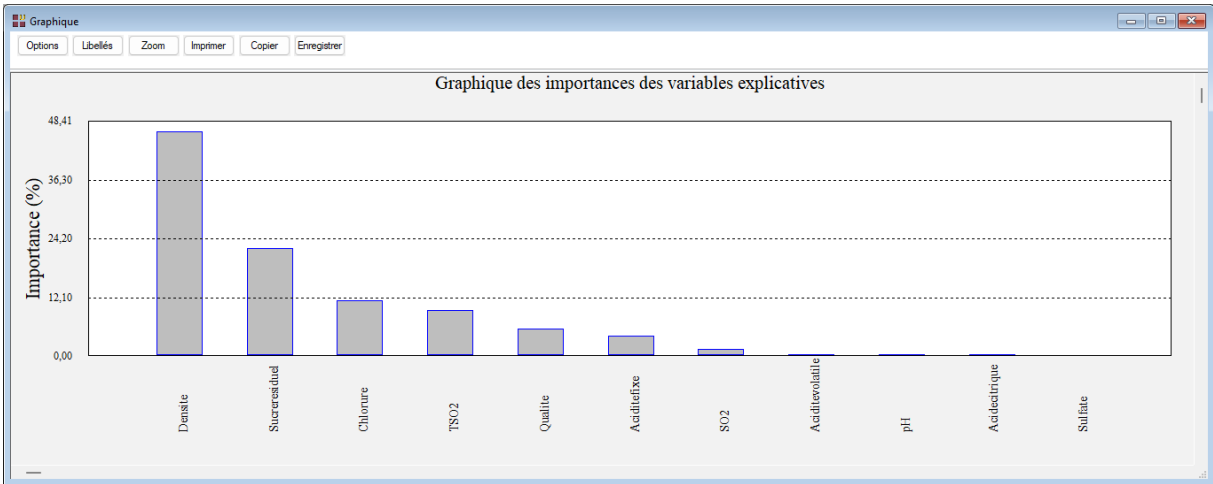
Après quelques instants, la fenêtre 'Rapports et Graphiques' montrée ci-après s'affiche.

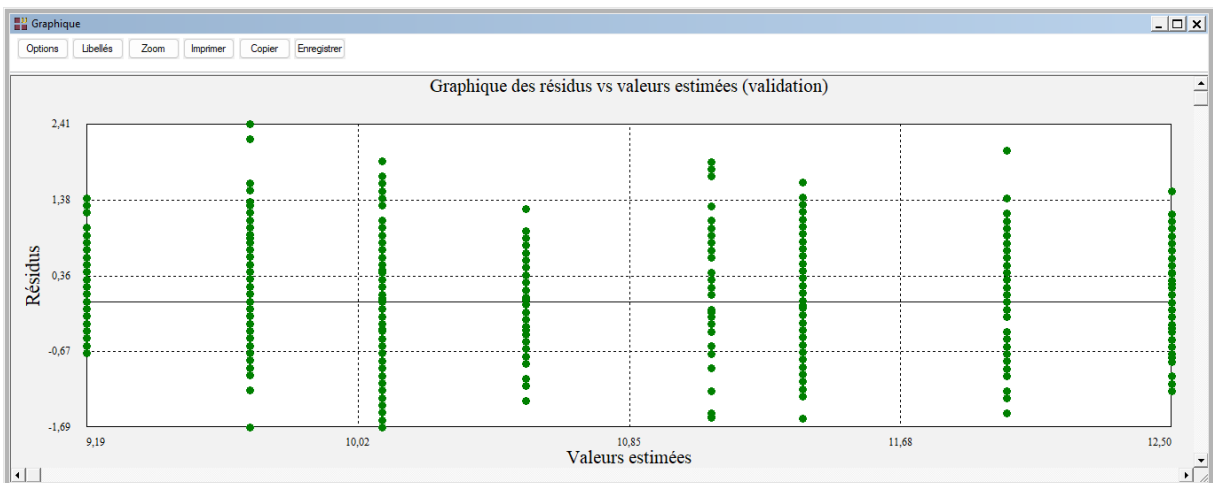
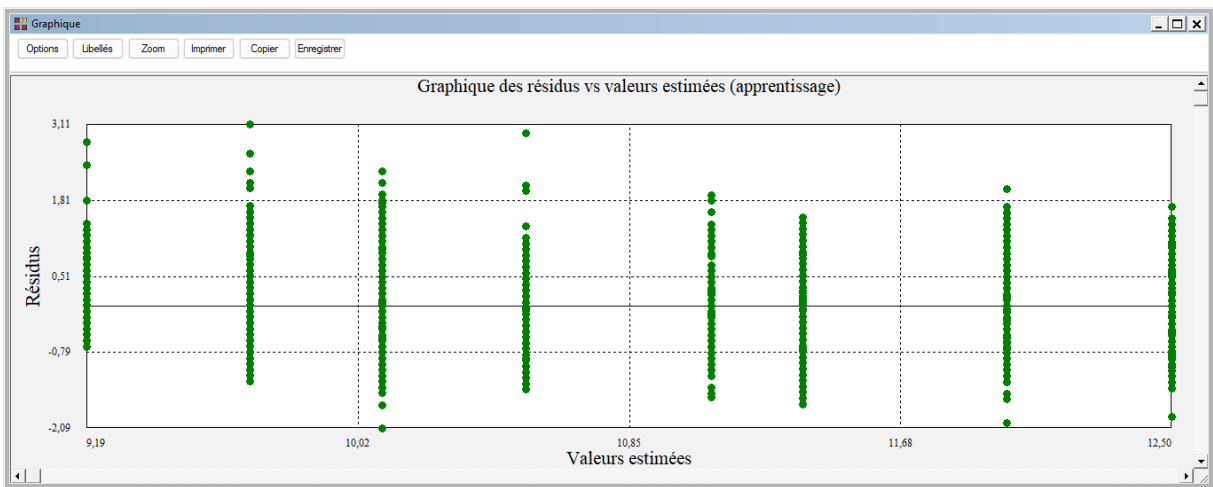
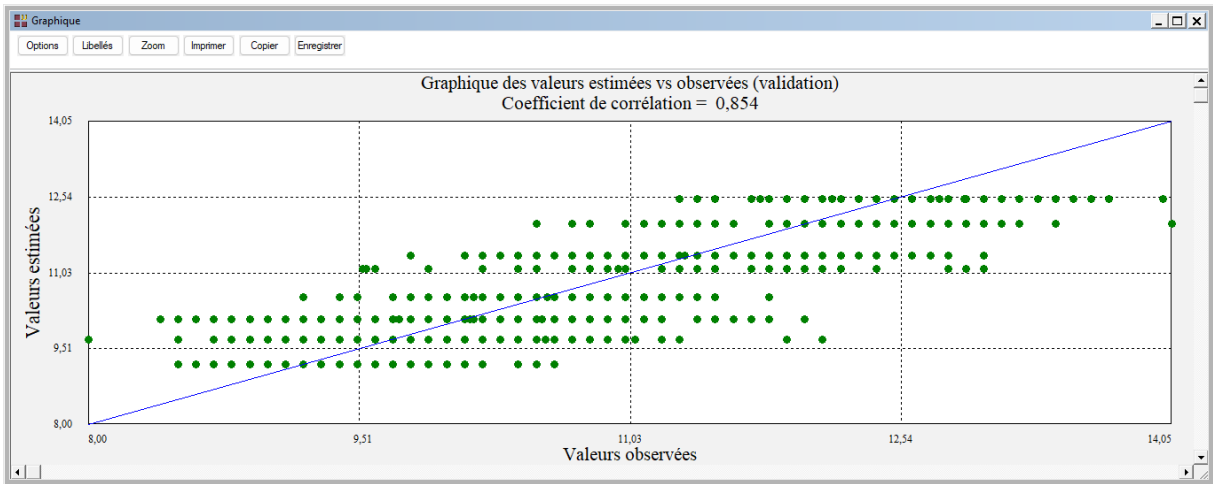




Visualisons les graphiques obtenus.







## Exemple 5 : Fichier TITANIC

Pour ce quatrième exemple, nous utiliserons le fichier TITANIC pour construire un arbre de décision. Ce fichier contient des informations concernant 714 passagers :

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

Cliquons sur l'icône ARBRE dans le ruban Expliquer et renseignons la boîte de dialogue comme montré ci-dessous. Après exécution de la procédure, visualisons le tableau de classement des données d'apprentissage et la courbe ROC associée.

Arbres de décision et de régression

Statut  
Age  
Tarif  
Nbfse  
Nbpe  
Classe  
Sexe  
Poids  
LibVarQuanti  
LibVarQuali  
LibObs

Type d'arbre :  
 Classement  Régression

Mesure de l'impureté (classement) :  
 Indice de Gini  Gain d'information

Taille minimale pour découpage : 5

Taille minimale d'un noeud terminal : 2

Profondeur maximale de l'arbre : 30

Coefficient de complexité : 0,01

Nombre de validations croisées : 10

Racine aléatoire : 12345

Variable à expliquer :  
Statut

Variables explicatives quantitatives :  
Age  
Tarif  
Nbfse  
Nbpe

Variables explicatives qualitatives :  
Classe  
Sexe

(Poids des observations :)

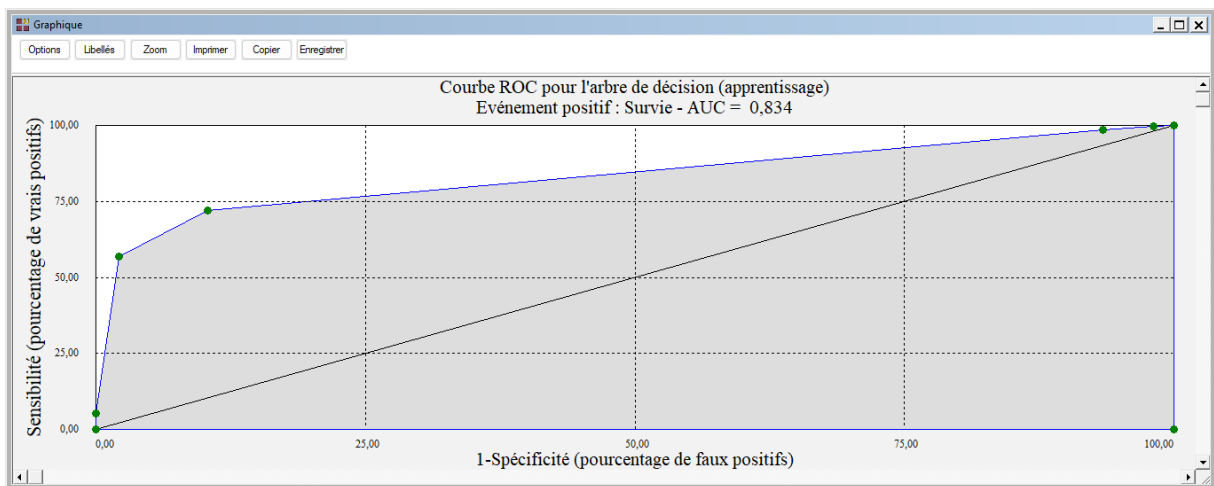
(Libellés des variables quantitatives :)  
LibVarQuanti

(Libellés des variables qualitatives :)  
LibVarQuali

(Libellés des observations :)  
LibObs

Ok Annuler Sélection Supprimer Aide

SYNTHESE DU CLASSEMENT DE LA POPULATION D'APPRENTISSAGE				
En lignes, les classes observées				
En colonnes, les classes prévues				
Pourcentage de mal classés : 17,507 %				
Pourcentage de bien classés : 82,493 %				
	Observé \ Prévu	Deces	Survie	Total
Deces		380	44	424
Survie		81	209	290
Total		461	253	714



Environ 82 % des passagers sont bien classés par cette analyse et l'aire sous la courbe ROC est proche de 0,83.

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 internes créées par la procédure

Voici la liste des variables internes créées par la procédure. A noter que certaines des variables mentionnées ci-dessous peuvent ne pas apparaître, en fonction des options choisies.

<i>Variable</i>	<i>Contenu</i>
libimpvar	Libellés des variables explicatives
impvar	Importances des variables explicatives

obsapp	Libellés des observations d'apprentissage
estapp	Valeurs estimées des données d'apprentissage
residapp	Résidus pour les données d'apprentissage
obsvalid	Libellés des observations de validation
estvalid	Valeurs estimées des données de validation
residvalid	Résidus pour les données de validation
vpA	Vrais positifs (apprentissage)
fnA	Faux négatifs (apprentissage)
fpA	Faux positifs (apprentissage)
vnA	Vrais négatifs (apprentissage)
specificiteA	Spécificité (apprentissage)
sensibiliteA	Sensibilité (apprentissage)
vpV	Vrais positifs (validation)
fnV	Faux négatifs (validation)
fpV	Faux positifs (validation)
vnV	Vrais négatifs (validation)
specificiteV	Spécificité (validation)
sensibiliteV	Sensibilité (validation)
obsnouv	Libellés des observations nouvelles
estnouv	Valeurs estimées des observations nouvelles

## Références

Breiman, L., J. H. Friedman, R. A. Olshen, et C. J. Stone. 1984. *Classification and Regression Trees*. ISBN 978-0412048418. CRC.

Documentation du package R – 'rpart' (2022)

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

Documentation du package R – 'rpart.plot' (2022)

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