

UNIWIN VERSION 10.1.0

FORETS ALEATOIRES

Révision : 19/11/2024

Définition	1
Entrée des données.....	2
Données manquantes ou non sélectionnées	3
Exemple 1 : Fichier IRIS3 (décision).....	3
L'option Rapports.....	7
L'option Graphiques.....	11
Exemple 2 : Fichier DIABETES (décision).....	14
Exemple 3 : Fichier GRAISSE (régression).....	16
L'option Rapports.....	18
L'option Graphiques.....	20
Exemple 4 : Fichier WINES3 (régression)	24
Exemple 5 : Fichier TITANIC (décision)	27
Calculs de la matrice de confusion et des indicateurs	29
Les variables internes créées par la procédure.....	30
Références.....	31

Définition

La procédure Forêts aléatoires crée des modèles de deux formes : modèles décisionnels qui découpent les observations en groupes basés sur les caractéristiques observées et modèles de régression qui prévoient la valeur d'une variable à expliquer. Les modèles sont élaborés en construisant un grand nombre d'arbres et en faisant la moyenne des prévisions obtenues à partir de ces arbres. Les arbres sont construits en utilisant une procédure similaire à celle des arbres de décision et de régression, avec optimisation aléatoire des nœuds et agrégation de bootstrap (bagging). Les données brutes sont utilisées pour les calculs car la structure d'un arbre n'est pas impactée par les habituelles transformations monotones des données. Les observations sont découpées en deux jeux : un jeu d'apprentissage utilisé pour construire les arbres et un jeu de prévision pour lequel les classes ou valeurs de la variable à expliquer ne sont pas connues et doivent être prévues. La variable à expliquer est soit qualitative, soit quantitative, comme c'est également le cas pour les variables explicatives.

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

Entrée des données

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

Forêts aléatoires

Variable à expliquer :

Variables explicatives quantitatives :

Variables explicatives qualitatives :

(Libellés des variables quantitatives :)

(Libellés des variables qualitatives :)

(Libellés des observations :)

Type de forêt
 Classement Régression

Nombre d'arbres à créer : 500

Nombre de variables à tester
 Défaut Personnalisé : 0

Nombre maximum de noeuds terminaux : 50

Taille minimum d'un noeud terminal : 1

Taille de l'échantillon des observations
 Défaut Personnalisé : 0

Type d'échantillonnage
 Avec remise Sans remise

Racine aléatoire : 174835488

Ok Annuler Sélection Supprimer Aide

Cette boîte de dialogue permet de définir la variable à expliquer et les variables explicatives quantitatives et qualitatives.

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

Différents critères pour la construction de la forêt doivent être précisés :

Type de forêt : Classement (décision) pour une variable à expliquer qualitative alphanumérique ou Régression pour une variable à expliquer quantitative.

Nombre d'arbres à créer : le nombre d'arbres pour la forêt. La valeur par défaut est 500.

Nombre de variables à tester : le nombre de variables candidates sélectionnées aléatoirement parmi toutes les variables lors de la division d'un nœud. Les valeurs par défaut sont la racine carrée du nombre de variables pour le classement et le nombre de variables divisé par 3 pour la régression.

Nombre maximum de nœuds terminaux : le nombre maximum de nœuds terminaux dans chaque arbre de la forêt.

Taille minimum d'un nœud terminal : la taille minimum d'un nœud terminal dans chaque arbre de la forêt. Des valeurs usuelles sont 1 pour le classement et 5 pour la régression.

Taille de l'échantillon : le nombre d'observations sélectionnées aléatoirement pour construire chaque arbre. Les valeurs par défaut sont le nombre d'observations non manquantes si l'échantillonnage est avec remise et 0,632 fois le nombre d'observations non manquantes si l'échantillonnage est sans remise.

Type d'échantillonnage : permet de choisir un échantillonnage avec ou sans remise.

Racine aléatoire : la racine utilisée pour l'échantillonnage des observations et des variables.

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 données non sélectionnées ne sont pas utilisées.
- Les forêts aléatoires ne nécessitent pas de jeu de validation. Elles utilisent une technique appelée évaluation out-of-bag (évaluation OOB) pour mesurer la qualité du modèle.

Exemple 1 : Fichier IRIS3 (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)



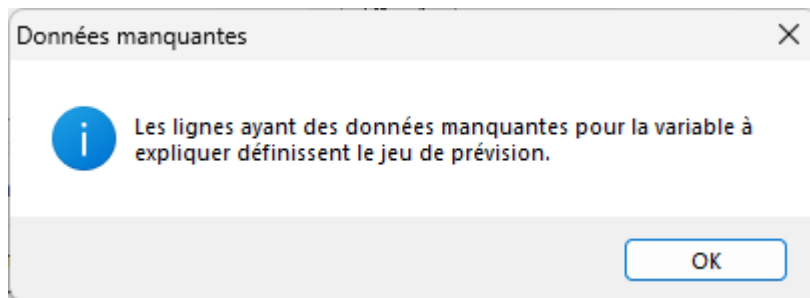
Cliquons sur l'icône ARBRE dans le ruban Expliquer.

La boîte de dialogue ci-dessous s'affiche.

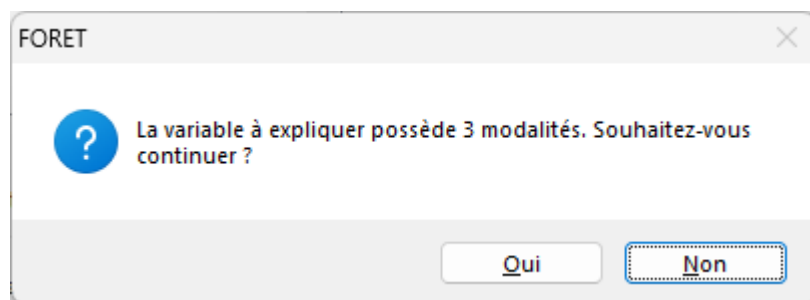
La variable *codesp2* est la variable à expliquer. Elle contient pour chaque observation le libellé de son espèce d'appartenance. Nous choisissons les variables de *lonsepal* à *larpetal* comme variables explicatives quantitatives et laissons les autres paramètres de l'analyse aux valeurs par défaut.

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 :


Rapports et Graphiques

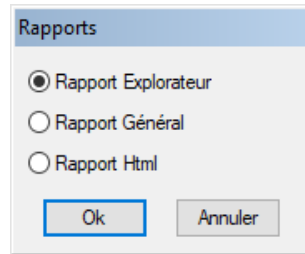
Rapport FORET


	1	2	3	4	5	6	7	8
1								
2	(C) UNIWIN version 10.1.0							
3								
4	DATE : 18/11/2024							
5	ORDINATEUR : LAPTOP-LEGBL077							
6	UTILISATEUR : echar							
7	FICHIER(S) DE DONNEES OUVERT(S) : IRIS3.SGD							
8								
9	RESULTATS DE L'ANALYSE FORET DE DECISION							
10								
11	Sélection :							
12	Aucune							
13								
14	Nombre d'observations : 144							
15								
16	Variable à expliquer :							
17	codesp2							
18								
19	Modalités de la variable à expliquer :							
20	Setosa							
21	Versicolor							

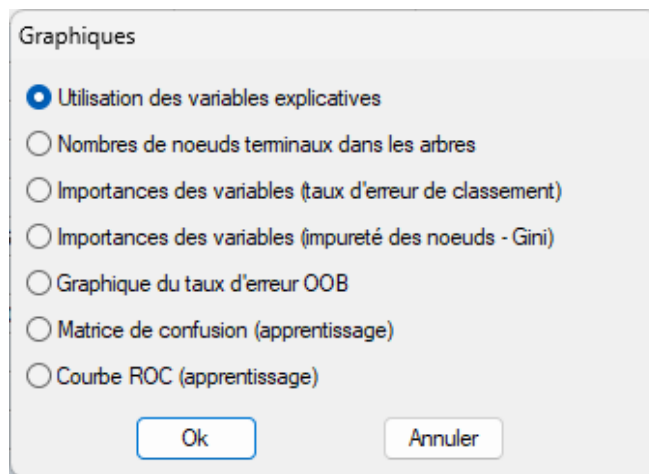
Rapport Explorateur /


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

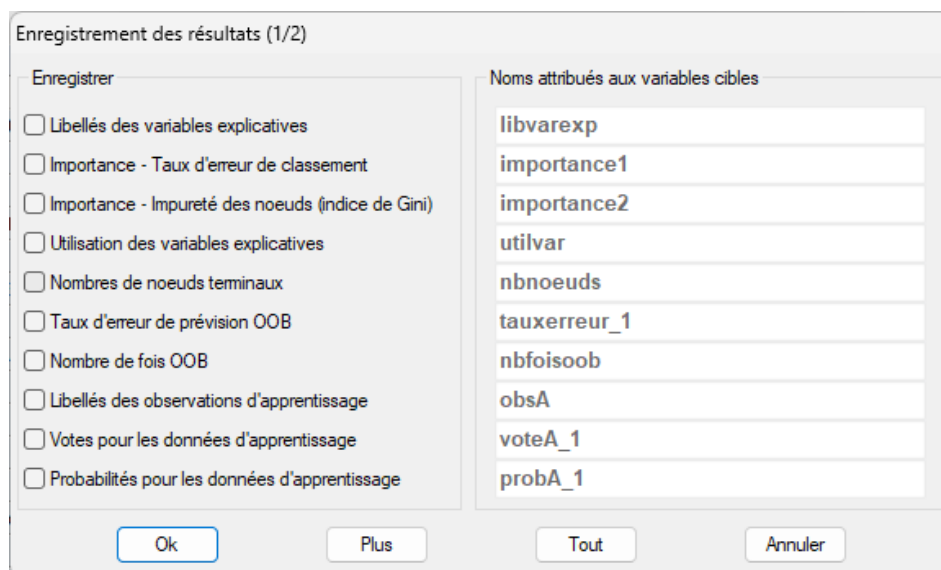
L'icône 'Rapports'  affiche la boîte de dialogue des options pour les rapports :



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.

Taux d'erreur de prévision OOB

Ce tableau affiche l'évolution du taux d'erreur de prévision OOB (Out Of Bag) en fonction du nombre d'arbres dans la forêt. Il affiche également l'évolution de l'erreur pour chacune des modalités de la variable à expliquer.

	1	2	3	4	5	6	7	8
TAUX D'ERREUR DE PREVISION (OOB)								
		OOB	Setosa	Versicolor	Virginica			
Arbres 1 à 1		4,08163	0,00000	5,26316	8,33333			
Arbres 1 à 2		4,54545	0,00000	3,44828	10,71429			
Arbres 1 à 3		6,36364	2,50000	6,06061	10,81081			
Arbres 1 à 4		9,01639	4,65116	7,50000	15,38462			
Arbres 1 à 5		6,10687	2,17391	4,65116	11,90476			
Arbres 1 à 6		4,47761	0,00000	4,54545	9,09091			
Arbres 1 à 7		4,44444	0,00000	4,54545	9,09091			
Arbres 1 à 8		4,41176	0,00000	6,66667	6,81818			
Arbres 1 à 9		5,07246	0,00000	4,34783	11,11111			
Arbres 1 à 10		4,96454	0,00000	4,16667	10,86957			
Arbres 1 à 11		4,89510	0,00000	4,16667	10,63830			
Arbres 1 à 12		4,19580	0,00000	4,16667	8,51064			
Arbres 1 à 13		4,19580	0,00000	4,16667	8,51064			
Arbres 1 à 14		4,19580	0,00000	4,16667	8,51064			
Arbres 1 à 15		3,49650	0,00000	4,16667	6,38298			
Arbres 1 à 16		4,19580	0,00000	4,16667	8,51064			

Importances relatives des variables

	1	2	3	4	5	6	7	8
IMPORTANCES DES VARIABLES EXPLICATIVES DANS LE MODELE								
Taux d'erreur : importance basée sur le taux d'erreur de classement.								
Impureté : importance basée sur l'impureté des nœuds (indice de Gini).								
Plus l'importance d'une variable est grande, plus l'exclusion de cette variable impacte la qualité du modèle.								
		Taux d'erreur	Impureté					
lonsepal		11,62093	9,74507					
larsepal		7,39503	2,45054					
lonpetal		35,32771	43,02794					
larpetal		30,79800	39,98005					

La première mesure est basée sur le taux d'erreur de classement.

La deuxième mesure est basée sur l'impureté des nœuds (indice de Gini).

Pour chacune de ces mesures, les deux variables les plus importantes sont *lonpetal* et *larpetal*.

Utilisation des variables explicatives

Ce tableau indique combien de fois chaque variable a été utilisée pour construire la forêt.

	1	2	3	4	5	6	7	8
1								
2	UTILISATION DES VARIABLES EXPLICATIVES							
3								
4								
5		Comptage						
6	lonsepal	691						
7	larsepal	479						
8	lonpetal	1254						
9	larpetal	1061						
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								

Nombres de nœuds terminaux dans les arbres

Ce tableau indique le nombre de nœuds terminaux dans chaque arbre de la forêt.

	1	2	3	4	5	6	7	8
1								
2	NOMBRES DE NOEUDS TERMINAUX DANS LES ARBRES							
3								
4								
5		Comptage						
6	Arbre 1	6						
7	Arbre 2	8						
8	Arbre 3	9						
9	Arbre 4	5						
10	Arbre 5	8						
11	Arbre 6	9						
12	Arbre 7	8						
13	Arbre 8	6						
14	Arbre 9	10						
15	Arbre 10	11						
16	Arbre 11	9						
17	Arbre 12	10						
18	Arbre 13	7						
19	Arbre 14	7						
20	Arbre 15	8						
21	Arbre 16	6						

Nombre de fois où chaque observation est OOB dans les arbres

Pour chaque observation du jeu d'apprentissage, ce tableau indique le nombre de fois où cette observation a été OOB lors de la construction des arbres.

	1	2	3	4	5	6	7	8
1								
2	NOMBRES DE FOIS OU CHAQUE OBSERVATION EST OOB DANS LES ARBRES							
3								
4								
5		Comptage						
6	o1		213					
7	o2		176					
8	o4		181					
9	o5		190					
10	o6		202					
11	o7		191					
12	o8		172					
13	o9		186					
14	o10		196					
15	o11		174					
16	o12		187					
17	o13		177					
18	o14		174					
19	o15		192					
20	o16		183					
21	o17		191					

Détail du classement de la population d'apprentissage

	1	2	3	4	5	6	7	8
1								
2	DETAIL DU CLASSEMENT DE LA POPULATION D'APPRENTISSAGE (OOB)							
3								
4	Observations, classes observées, votes et probabilités d'affectation aux classes							
5								
6	Il y a 144 observations dans le jeu d'apprentissage.							
7								
8	(*) = observation mal classée.							
9								
10								
11	Observation - Classe observée	Votes Setosa	Votes Versicolor	Votes Virginica	Probabilité Setosa	Probabilité Versicolor	Probabilité Virginica	
12	o1 - Setosa	213	0	0	1,00000	0,00000	0,00000	
13	o2 - Setosa	176	0	0	1,00000	0,00000	0,00000	
14	o4 - Setosa	181	0	0	1,00000	0,00000	0,00000	
15	o5 - Setosa	190	0	0	1,00000	0,00000	0,00000	
16	o6 - Setosa	202	0	0	1,00000	0,00000	0,00000	
17	o7 - Setosa	191	0	0	1,00000	0,00000	0,00000	
18	o8 - Setosa	172	0	0	1,00000	0,00000	0,00000	
19	o9 - Setosa	186	0	0	1,00000	0,00000	0,00000	
20	o10 - Setosa	196	0	0	1,00000	0,00000	0,00000	
21	o11 - Setosa	174	0	0	1,00000	0,00000	0,00000	

Pour chaque observation de la population d'apprentissage, les votes majoritaires pour chacune des classes et les probabilités d'affectation à chacune des classes sont affichés.

Pour chaque observation, ces votes majoritaires et ces probabilités sont calculés en utilisant uniquement les arbres dans lesquels cette observation est OOB.

Les observations mal classées sont indiquées par une étoile « * ».

Matrice de confusion du jeu d'apprentissage (OOB)

Rapport FORET

- Taux d'erreur de prévision
- Importances des variables explicatives
- Utilisation des variables explicatives
- Nombres de noeuds terminaux
- Nb. fois observation OOB
- Détail classement apprentissage
- Matrice de confusion (apprentissage)**
- Détail classement prévision

	1	2	3	4	5	6	7	8
1								
2	MATRICE DE CONFUSION DE LA POPULATION D'APPRENTISSAGE (OOB)							
3								
4	En lignes, les classes observées							
5	En colonnes, les classes prévues							
6								
7	Pourcentage de mal classés : 4,167 %							
8	Pourcentage de bien classés (exactitude) : 95,833 %							
9								
10	Précision = $VP / (VP + FP)$							
11	Rappel = $VP / (VP + FN)$							
12	Score F1 = $2 \times (\text{Précision} \times \text{Rappel}) / (\text{Précision} + \text{Rappel})$							
13								
14								
15	Observé \ Prévu	Taille	Setosa	Versicolor	Virginica	Précision	Rappel	Score F
16	Setosa	48	48	0	0	1,0000	1,0000	1,000
17	Versicolor	48	0	45	3	0,9375	0,9375	0,937
18	Virginica	48	0	3	45	0,9375	0,9375	0,937
19								
20								
21								

Pour chaque classe observée, le tableau affiche les effectifs prévus pour chacune des classes, la précision, le rappel et le score F1.

Voir le paragraphe « Calculs de la matrice de confusion et des indicateurs » pour des détails concernant ces indicateurs.

Détail du classement de la population de prévision

Rapport FORET

- Taux d'erreur de prévision
- Importances des variables explicatives
- Utilisation des variables explicatives
- Nombres de noeuds terminaux
- Nb. fois observation OOB
- Détail classement apprentissage
- Matrice de confusion (apprentissage)
- Détail classement prévision**

	1	2	3	4	5	6	7	8
1								
2	DETAIL DU CLASSEMENT DE LA POPULATION DE PREVISION							
3								
4	Observations et probabilités d'affectation aux classes							
5								
6	Il y a 6 observations dans le jeu de prévision.							
7								
8								
9	Observation	Votes Setosa	Votes Versicolor	Votes Virginica	Probabilité Setosa	Probabilité Versicolor	Probabilité Virginica	
10	o3	500	0	0	1	0,000	0,000	
11	o36	500	0	0	1	0,000	0,000	
12	o62	0	497	3	0	0,994	0,006	
13	o84	0	74	426	0	0,148	0,852	
14	o104	0	1	499	0	0,002	0,998	
15	o125	0	0	500	0	0,000	1,000	
16								
17								
18								
19								
20								
21								

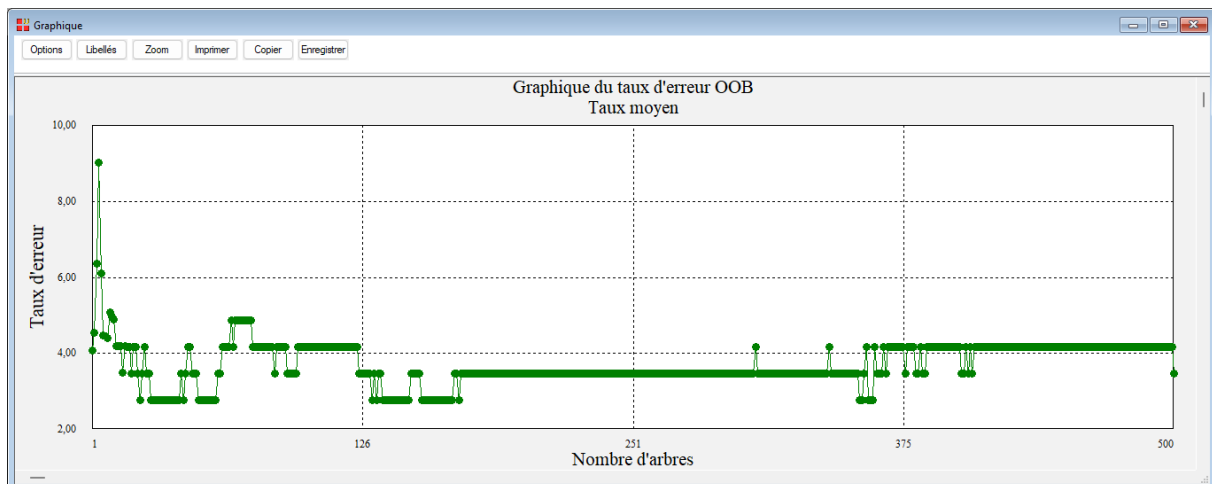
Pour chaque observation de la population de prévision, les votes majoritaires pour chacune des classes et les probabilités d'affectation à chacune des classes sont affichés.

L'option Graphiques

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

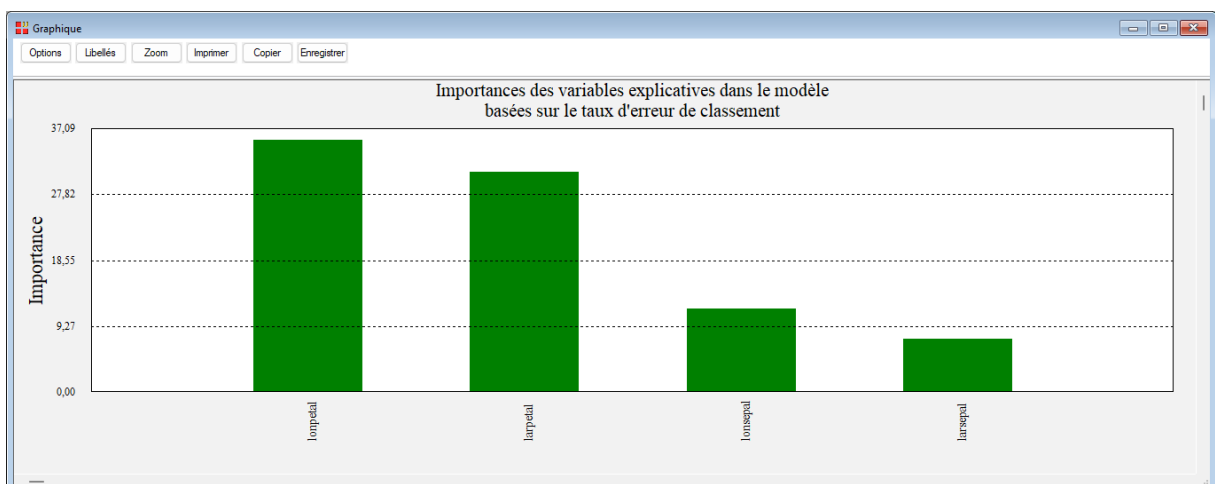
Graphique du taux d'erreur OOB

Ce graphique affiche l'évolution du taux moyen d'erreur ou du taux d'erreur pour chacune des classes en fonction du nombre d'arbres dans la forêt.



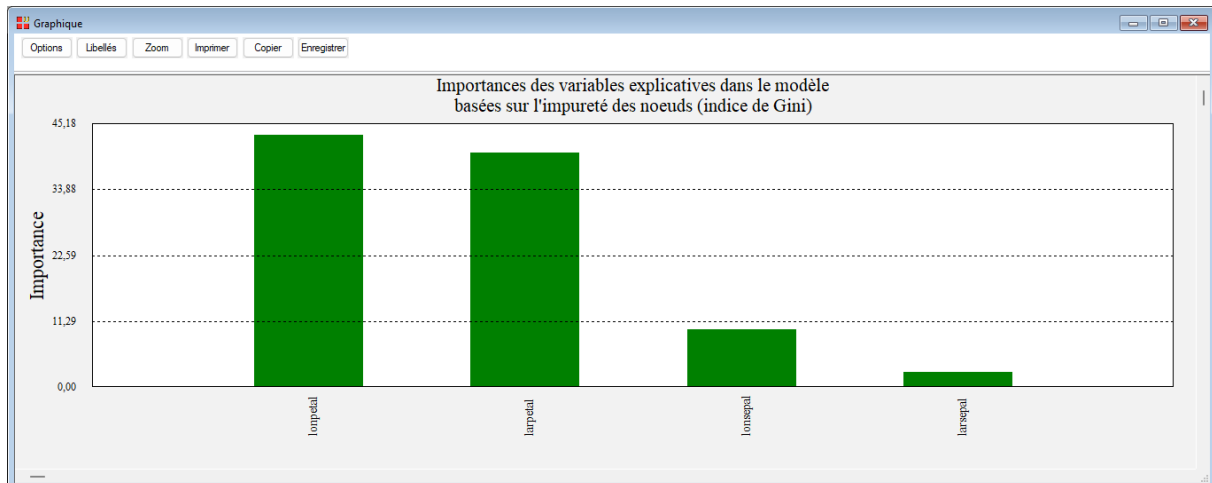
Graphique de l'importance des variables – Taux d'erreur de classement

Ce graphique affiche l'importance de chaque variable prise en compte lors de la construction des arbres de la forêt. La longueur de chaque barre représente le taux d'erreur de classement. Les variables sont classées de gauche à droite par ordre décroissant d'importance.



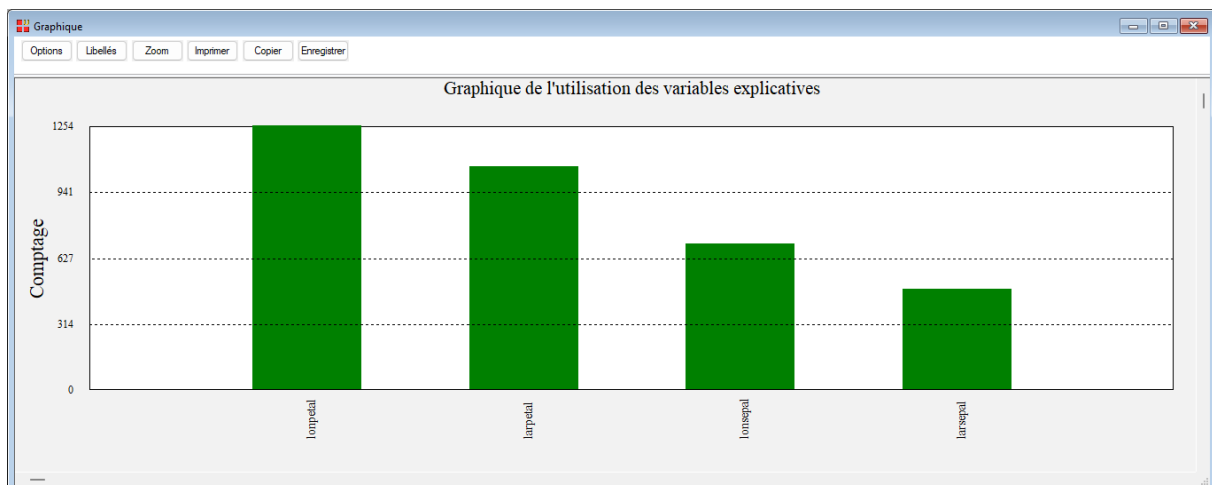
Graphique de l'importance des variables – Impureté des nœuds (Gini)

Ce graphique affiche l'importance de chaque variable prise en compte lors de la construction des arbres de la forêt. La longueur de chaque barre représente l'impureté des nœuds (indice de Gini). Les variables sont classées de gauche à droite par ordre décroissant d'importance.



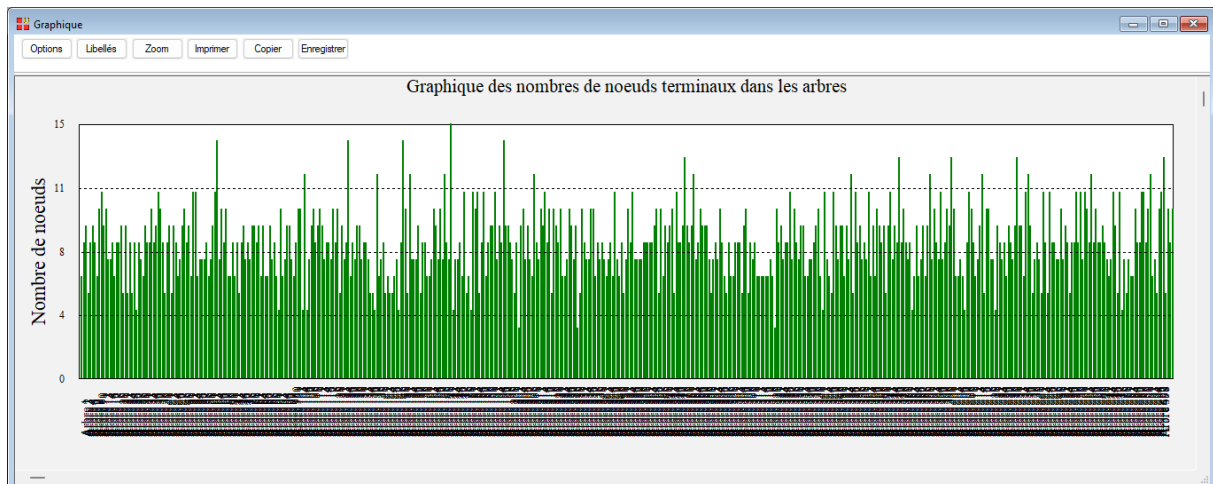
Graphique de l'utilisation des variables explicatives

Ce graphique indique le nombre de fois où chaque variable explicative a été sélectionnée pour la division d'un nœud lors de la construction des arbres de la forêt. Les variables sont classées de gauche à droite par ordre décroissant d'utilisation.



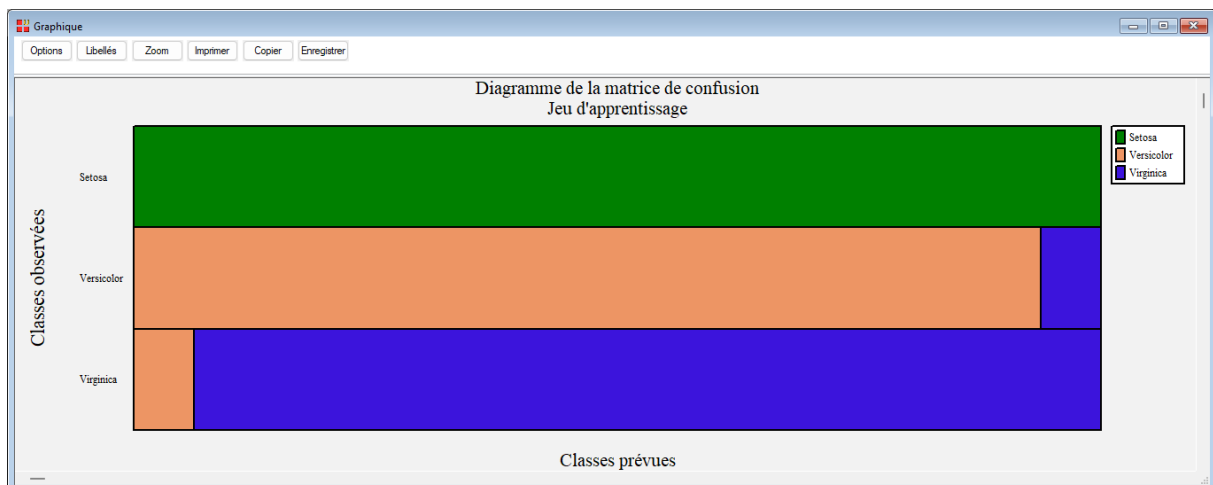
Graphique des nombres de nœuds terminaux dans les arbres

Ce graphique indique les nombres de nœuds terminaux dans chaque arbre de la forêt.



Graphique de la matrice de confusion

Ce graphique affiche sous la forme d'un diagramme en mosaïque les données de la matrice de confusion pour le jeu d'apprentissage dans le cas d'un arbre de décision.



Courbes ROC

La courbe ROC pour le jeu d'apprentissage n'est disponible que dans le cas de deux classes et pour un arbre de décision.

Il y a trois classes dans cet exemple et donc le graphique n'est pas proposé.

Exemple 2 : Fichier DIABETES (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.

Forêts aléatoires

Nbgros
Glucose
Pad
Peau
Insuline
IMC
Hérédité
Age
Diabète

Type de forêt
 Classement Régression

Nombre d'arbres à créer : 500

Nombre de variables à tester
 Défaut Personnalisé : 0

Nombre maximum de noeuds terminaux : 50

Taille minimum d'un noeud terminal : 1

Taille de l'échantillon des observations
 Défaut Personnalisé : 0

Type d'échantillonnage
 Avec remise Sans remise

Racine aléatoire : 12345

Variable à expliquer :
Diabète

Variables explicatives quantitatives :
Nbgros
Glucose
Pad
Peau
Insuline
IMC
Hérédité
Age

Variables explicatives qualitatives :

(Libellés des variables quantitatives :)

(Libellés des variables qualitatives :)

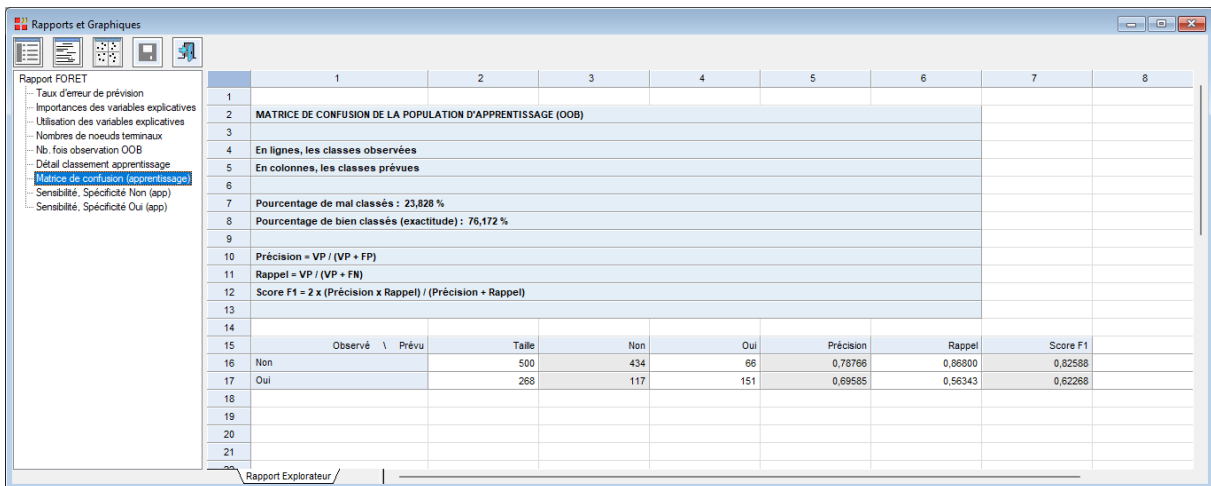
(Libellés des observations :)

Ok Annuler Sélection Supprimer Aide

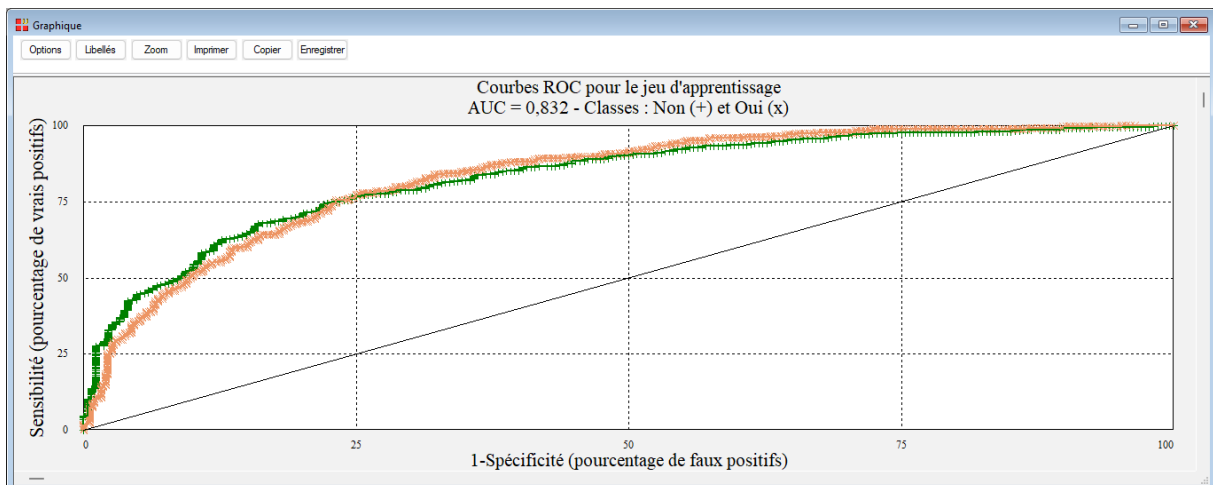
Neuf variables ont été collectées :

- 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)²)
- Hérité : fonction généalogique du diabète
- Âge : âge en années
- Diabète : oui ou non

Renseignons la boîte de dialogue de l'analyse comme montré ci-dessus, exécutons l'analyse et visualisons la matrice de confusion (OOB).



Visualisons les courbes ROC pour le jeu d'apprentissage (OOB), disponibles dans cet exemple car la variable à expliquer possède deux modalités.



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

Exemple 3 : Fichier GRAISSE (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 (les variables explicatives sélectionnées sont 'Age' à 'anthro4').

Cliquons sur Ok.

Après quelques instants, la fenêtre du rapport s'affiche.

Forêts aléatoires

Graisse
 Age
 Taille
 Hanche
 Coude
 Genou
 anthro3a
 anthro3b
 anthro3c
 anthro4

Type de forêt
 Classement Régression

Nombre d'arbres à créer :

Nombre de variables à tester
 Défaut Personnalisé :

Nombre maximum de noeuds terminaux :

Taille minimum d'un noeud terminal :

Taille de l'échantillon des observations
 Défaut Personnalisé :

Type d'échantillonnage
 Avec remise Sans remise

Racine aléatoire :

Variable à expliquer :

Variables explicatives quantitatives :
 Age
 Taille
 Hanche
 Coude
 Genou
 anthro3a
 anthro3b
 anthro3c
 anthro4

Variables explicatives qualitatives :

(Libellés des variables quantitatives :)

(Libellés des variables qualitatives :)

(Libellés des observations :)

Rapports et Graphiques

Rapport FORET

- Erreurs quadratiques et R-carés
- Importances des variables explicatives
- Utilisation des variables explicatives
- Nombres de noeuds terminaux
- Nb. fois observation OOB
- Observés, estimés (apprentissage)

	1	2	3	4	5	6	7	8
1								
2	(C) UNIWIN version 10.1.0							
3								
4	DATE : 18/11/2024							
5	ORDINATEUR : LAPTOP-LEG8L077							
6	UTILISATEUR : cchar							
7	FICHER(S) DE DONNEES OUVERT(S) : GRAISSE.SGD							
8								
9	RESULTATS DE L'ANALYSE FORET DE REGRESSION							
10								
11	Sélection :							
12	Aucune							
13								
14	Nombre d'observations : 71							
15								
16	Variable à expliquer :							
17	Graisse							
18								
19	Variables explicatives quantitatives et qualitatives :							
20	Age							
21	Taille							

Rapport Explorateur /

L'option Rapports

Erreurs quadratiques moyennes (OOB) et R-carrés (OOB)

The screenshot shows a software window titled 'Rapports et Graphiques'. On the left, there is a tree view with 'Rapport FORET' expanded, showing sub-items like 'Erreurs quadratiques et R-carrés', 'Importances des variables explicatives', etc. The main area is a table with 8 columns and 21 rows. The table title is 'ERREURS QUADRATIQUES MOYENNES ET R-CARRÉS (OOB)'. Row 4 shows 'Erreur quadratique moyenne (OOB) = 13,585' and row 5 shows 'R-carré moyen (OOB) = 0,887'. Rows 9-21 show data for 1 to 13 trees, with columns for 'Erreur quadratique' and 'R-carré'.

	1	2	3	4	5	6	7	8
1								
2	ERREURS QUADRATIQUES MOYENNES ET R-CARRÉS (OOB)							
3								
4	Erreur quadratique moyenne (OOB) = 13,585							
5	R-carré moyen (OOB) = 0,887							
6								
7								
8		Erreur quadratique		R-carré				
9	Arbres 1 à 1	51,36102		0,57279				
10	Arbres 1 à 2	36,06863		0,69999				
11	Arbres 1 à 3	27,61803		0,77028				
12	Arbres 1 à 4	27,63256		0,76850				
13	Arbres 1 à 5	30,91264		0,74288				
14	Arbres 1 à 6	25,65817		0,78858				
15	Arbres 1 à 7	24,92240		0,79270				
16	Arbres 1 à 8	25,02542		0,79185				
17	Arbres 1 à 9	24,20303		0,79889				
18	Arbres 1 à 10	23,24693		0,80664				
19	Arbres 1 à 11	23,24444		0,80666				
20	Arbres 1 à 12	20,72290		0,82763				
21	Arbres 1 à 13	19,67814		0,83632				

Ce tableau affiche l'évolution de l'erreur quadratique moyenne (OOB) et du R-carré (OOB) en fonction du nombre d'arbres dans la forêt.

Importances des variables explicatives

The screenshot shows the same software window with the tree view expanded to 'Importances des variables explicatives'. The table title is 'IMPORTANCES DES VARIABLES EXPLICATIVES DANS LE MODELE'. Row 4 shows 'Erreur quadratique : importance basée sur l'erreur quadratique moyenne.' and row 5 shows 'Impureté : importance basée sur l'impureté des nœuds (somme des carrés des résidus)'. Row 6 contains a note: 'Plus l'importance d'une variable est grande, plus l'exclusion de cette variable impacte la qualité du modèle.' Rows 9-18 show data for variables like 'Age', 'Taille', 'Hanche', 'Coude', 'Genou', and 'anthro3a' through 'anthro4', with columns for 'Erreur quadratique' and 'Impureté'.

	1	2	3	4	5	6	7	8
1								
2	IMPORTANCES DES VARIABLES EXPLICATIVES DANS LE MODELE							
3								
4	Erreur quadratique : importance basée sur l'erreur quadratique moyenne.							
5	Impureté : importance basée sur l'impureté des nœuds (somme des carrés des résidus).							
6	Plus l'importance d'une variable est grande, plus l'exclusion de cette variable impacte la qualité du modèle.							
7								
8								
9		Erreur quadratique		Impureté				
10	Age	-1,94506		82,04190				
11	Taille	16,20577		2138,31361				
12	Hanche	19,66275		2047,00045				
13	Coude	0,08397		76,90046				
14	Genou	8,93359		418,58086				
15	anthro3a	10,16215		919,78722				
16	anthro3b	11,25089		737,61554				
17	anthro3c	13,82658		1157,52357				
18	anthro4	10,31136		914,81194				
19								
20								
21								

La première mesure est basée sur l'erreur quadratique moyenne. La seconde mesure est basée sur l'impureté des nœuds (somme des carrés des résidus).

Plus l'importance d'une variable est grande, plus l'exclusion de cette variable impacte la qualité du modèle.

Utilisation des variables explicatives

Ce tableau indique combien de fois chaque variable a été utilisée pour construire la forêt.

	1	2	3	4	5	6	7	8
1								
2	UTILISATION DES VARIABLES EXPLICATIVES							
3								
4								
5								
6	Age		Comptage					
7	Taille		2090					
8	Hanche		2851					
9	Coude		2789					
10	Géno		1842					
11	anthro3a		2173					
12	anthro3b		2542					
13	anthro3c		2478					
14	anthro4		2628					
15			2461					
16								
17								
18								
19								
20								
21								

Nombres de nœuds terminaux dans les arbres

	1	2	3	4	5	6	7	8
1								
2	NOMBRES DE NOEUDS TERMINAUX DANS LES ARBRES							
3								
4								
5								
6	Arbre 1		Comptage					
7	Arbre 2		47					
8	Arbre 3		43					
9	Arbre 4		46					
10	Arbre 5		42					
11	Arbre 6		45					
12	Arbre 7		41					
13	Arbre 8		47					
14	Arbre 9		44					
15	Arbre 10		46					
16	Arbre 11		45					
17	Arbre 12		43					
18	Arbre 13		48					
19	Arbre 14		47					
20	Arbre 15		48					
21	Arbre 16		46					
			48					

Ce tableau indique le nombre de nœuds terminaux dans chaque arbre de la forêt.

Nombre de fois où chaque observation est OOB dans les arbres

	1	2	3	4	5	6	7	8
1								
2	NOMBRES DE FOIS OU CHAQUE OBSERVATION EST OOB DANS LES ARBRES							
3								
4								
5								
6	o1		Comptage					
7	o2		187					
8	o3		176					
9	o4		194					
10	o5		183					
11	o6		178					
12	o7		177					
13	o8		193					
14	o9		195					
15	o10		178					
16	o11		191					
17	o12		177					
18	o13		177					
19	o14		180					
20	o15		182					
21	o16		180					

Pour chaque observation du jeu d'apprentissage, ce tableau indique le nombre de fois où cette observation a été OOB lors de la construction des arbres.

Valeurs observées, estimées et résidus

Ce tableau affiche pour chaque observation la valeur observée, la valeur estimée et le résidu.

	1	2	3	4	5	6	7	8
1								
2	JEU D'APPRENTISSAGE : VALEURS OBSERVEES, ESTIMEES ET RESIDUS							
3								
4								
5		Observé	Estimé	Résidu				
6	o1	41,68	42,80710	-1,12710				
7	o2	43,29	50,20398	-6,91398				
8	o3	35,41	37,85309	-2,44309				
9	o4	22,79	26,31377	-3,52377				
10	o5	36,42	35,41298	1,00702				
11	o6	24,13	22,91972	1,21028				
12	o7	29,83	30,32829	-0,49829				
13	o8	35,96	33,43005	2,52995				
14	o9	23,69	24,30146	-0,61146				
15	o10	22,71	22,68247	0,02753				
16	o11	23,42	24,21000	-0,79000				
17	o12	23,24	25,30893	-2,06893				
18	o13	26,25	25,27475	0,97525				
19	o14	21,94	20,69044	1,24956				
20	o15	30,13	27,06027	3,06973				
21	o16	36,31	40,39933	-4,08933				

L'option Graphiques

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

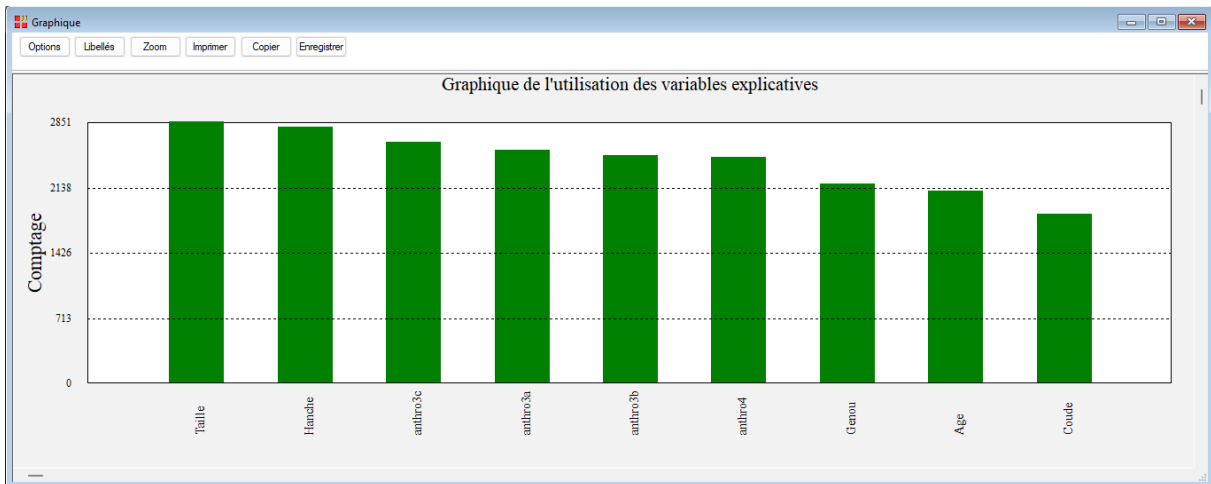
Graphiques

- Utilisation des variables explicatives
- Nombres de noeuds terminaux dans les arbres
- Importances des variables (erreur quadratique moyenne)
- Importances des variables (impureté des noeuds - résidus)
- Graphique des erreurs quadratiques moyennes (apprentissage)
- Graphique des R-carrés (apprentissage)
- Graphique des valeurs estimées vs observées (apprentissage)
- Graphique des résidus vs valeurs estimées (apprentissage)

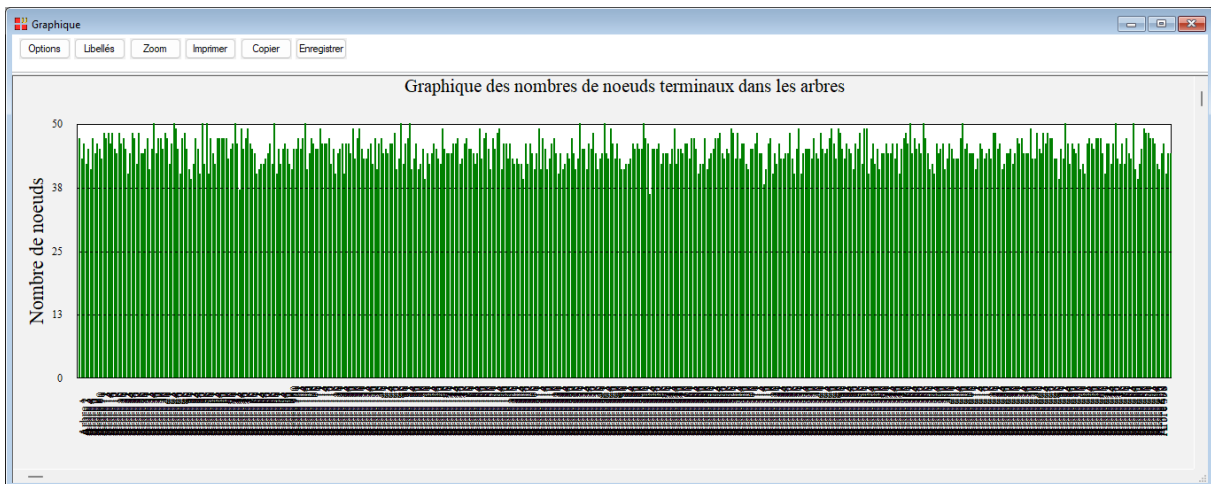
Ok
Annuler

Visualisons les différents graphiques.

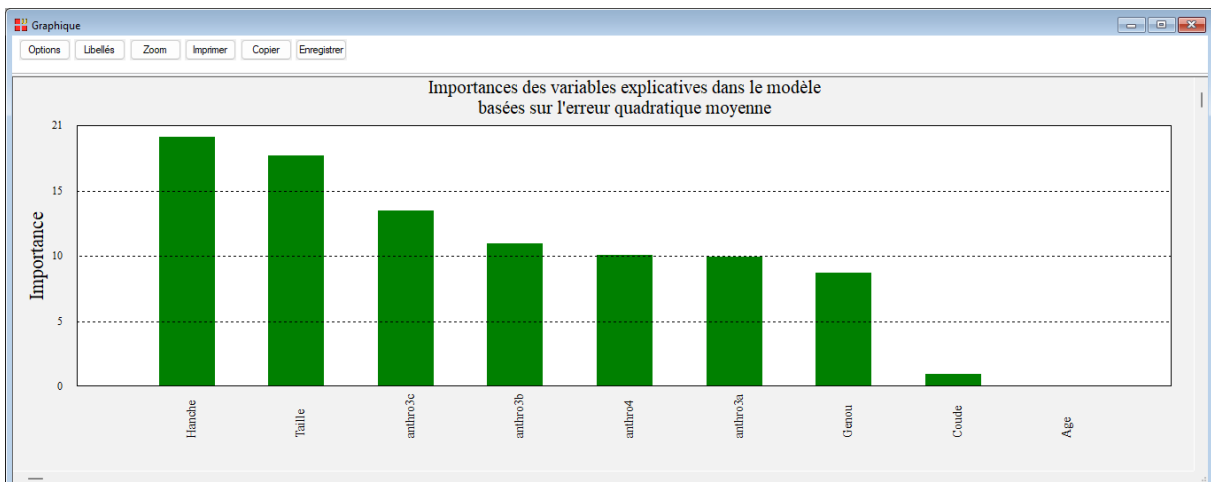
Utilisation des variables explicatives



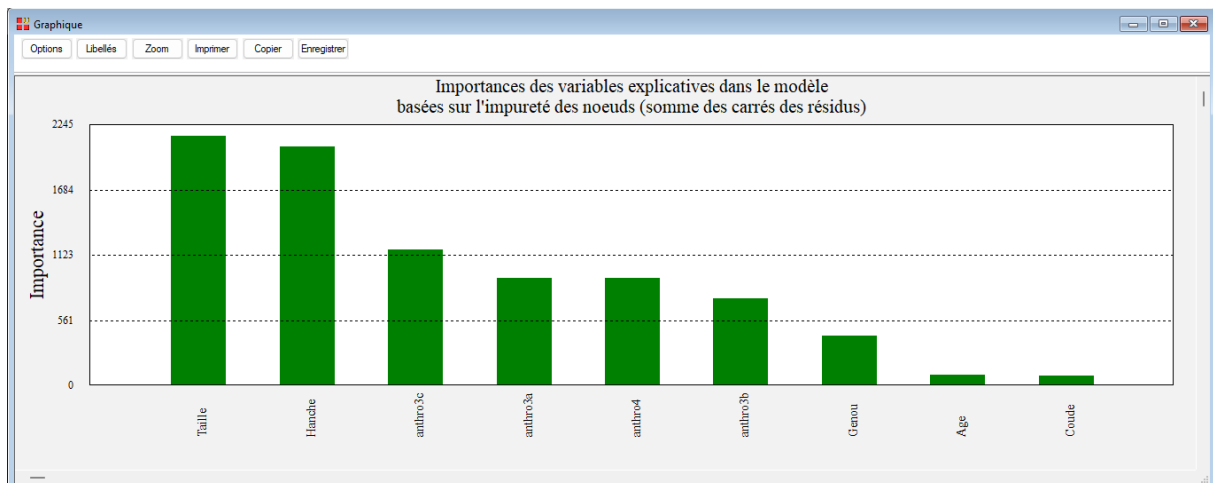
Nombres de nœuds terminaux dans les arbres



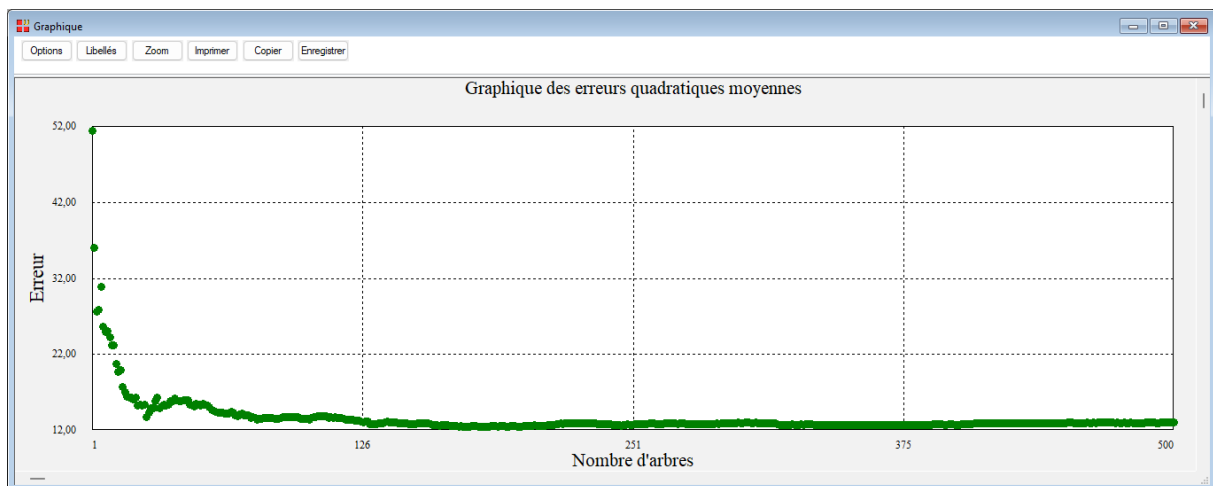
Importances des variables explicatives (erreur quadratique moyenne)



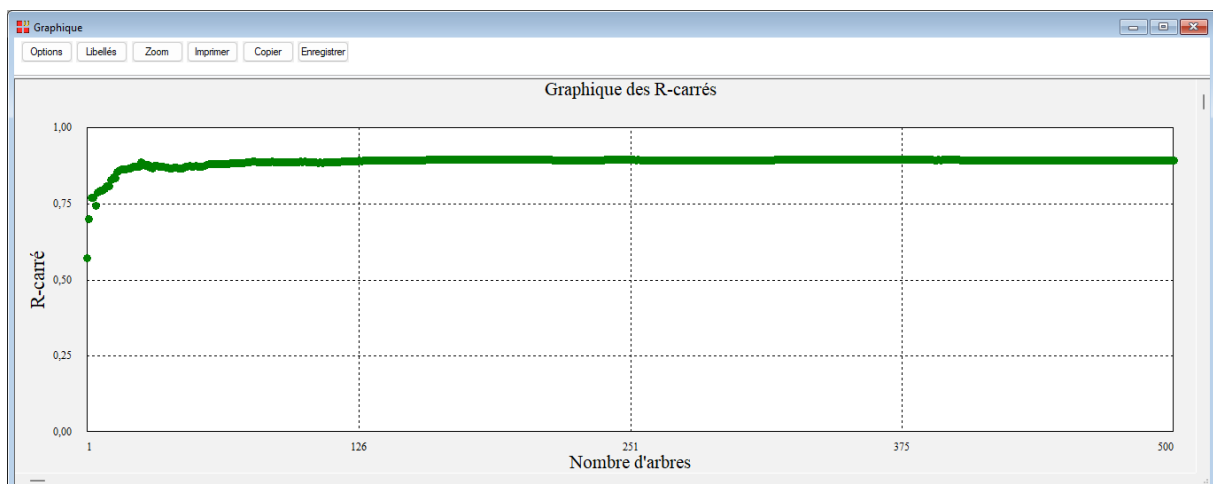
Importances des variables explicatives (impureté des nœuds)



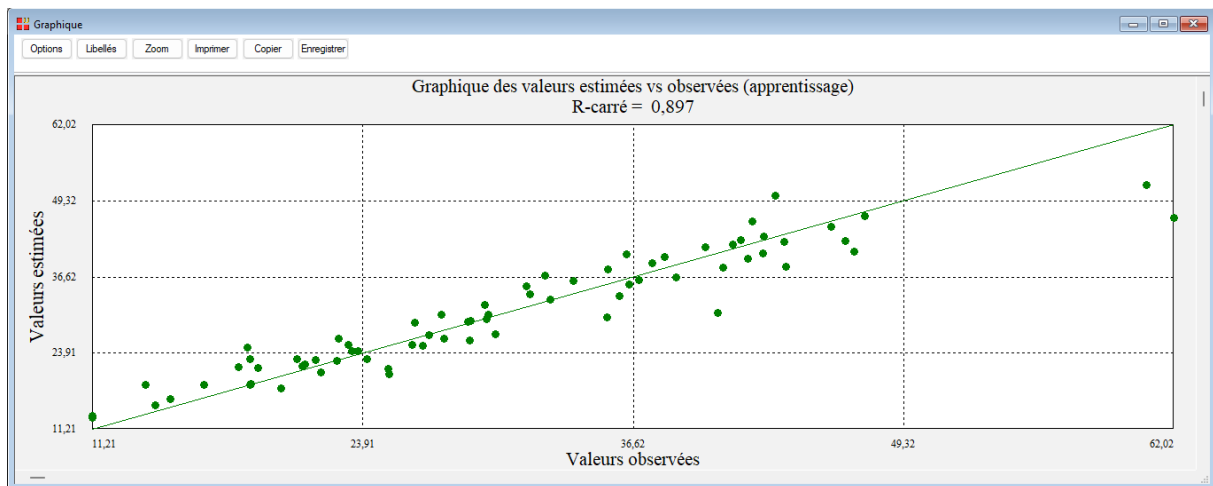
Graphique des erreurs quadratiques moyennes (OOB)



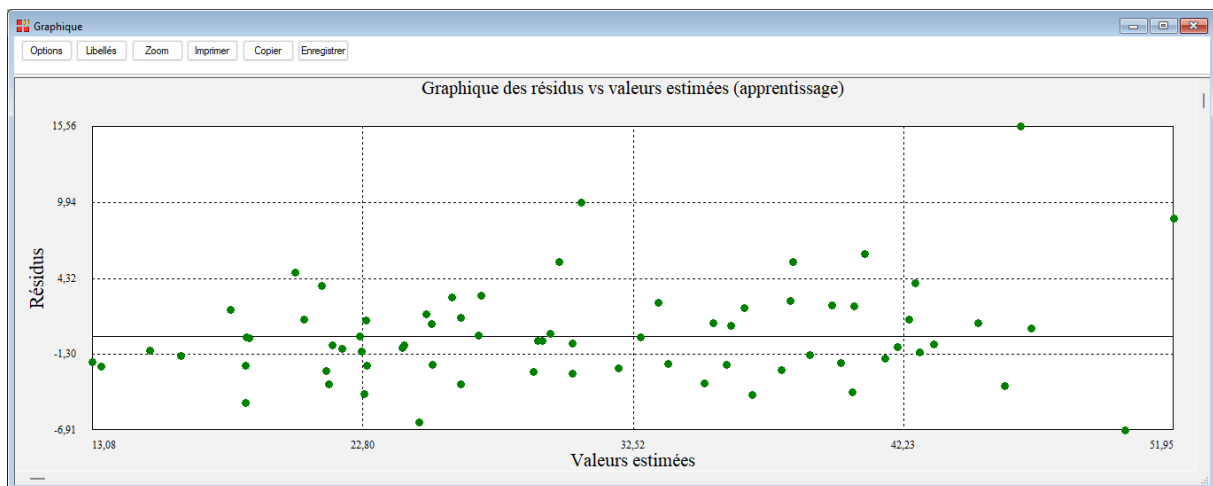
Graphique des R-carrés (OOB)



Graphique des valeurs estimées vs observées



Graphique des résidus vs valeurs observées



Les résultats suivants peuvent être enregistrés :

Enregistrement des résultats (1/1)

Enregistrer

- Libellés des variables explicatives
- Importance - Erreur quadratique moyenne
- Importance - Impureté (somme carrés des résidus)
- Utilisation des variables explicatives
- Nombres de noeuds terminaux des arbres
- Erreurs quadratiques moyennes des arbres
- R-carrés des arbres
- Nombre de fois OOB
- Libellés des observations du jeu d'apprentissage
- Valeurs prévues pour le jeu d'apprentissage

Noms attribués aux variables cibles

libvarexp
importance1
importance2
utilvar
nbnoeuds
errquad
R2
nbfoisob
obsA
prevA

Ok Plus Tout Annuler

Exemple 4 : Fichier WINES3 (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
- SO₂ (teneur en dioxyde de soufre libre)
- TSO₂ (teneur totale en dioxyde de soufre)
- Densité
- pH
- Sulfate
- Alcool
- Qualité (note entre 0 et 10)

Il y a au total 4898 observations dans ce fichier de données.

La variable 'Jeu' dans le fichier de données indique l'appartenance des observations au jeux d'apprentissage.

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

Les variables explicatives sélectionnées sont 'Acidité fixe' à 'Qualité'.

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

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

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

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

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

Définition de la sélection

Et Jeu = A

Liaison	Variable	Relation	Valeur ou variable
Et	Acidité fixe	=	Acide citrique
Et non	Acidité volatile	<>	Acidité fixe
Ou	Alcool	<	Acidité volatile
Ou non	Chlorure	<=	Alcool
	Densité	>	Chlorure
	Jeu	>=	Densité
	Libobs	début	Jeu

Ok Annuler Ajouter Aide

Voici quelques résultats obtenus par cette analyse.

Rapports et Graphiques

Rapport FORET

- Ereurs quadratiques et R-carrés
- Importances des variables explicatives
- Utilisation des variables explicatives
- Nombres de noeuds terminaux
- Nb. fois observation OOB
- Observés, estimés (apprentissage)
- Estimés (prévision)

	1	2	3	4	5	6	7	8
1								
2	JEU D'APPRENTISSAGE : VALEURS OBSERVEES, ESTIMEES ET RESIDUS							
3								
4								
5			Observé	Estimé	Résidu			
6	o3		10,100	10,15994	-0,05994			
7	o4		9,900	9,63566	0,26434			
8	o5		9,900	9,63393	0,26607			
9	o6		10,100	10,13450	-0,03450			
10	o7		9,600	10,01975	-0,41975			
11	o9		9,500	10,17180	-0,67180			
12	o10		11,000	10,25002	0,74998			
13	o11		12,000	11,42390	0,57610			
14	o12		9,700	10,47105	-0,77105			
15	o13		10,800	10,83896	-0,03896			
16	o14		12,400	11,13503	1,26497			
17	o16		11,400	11,47667	-0,07667			
18	o17		9,600	10,14783	-0,54783			
19	o19		11,300	11,38162	-0,08162			
20	o21		12,800	12,43834	0,36166			
21	o22		11,000	11,92396	-0,92396			

Rapport Explorateur /

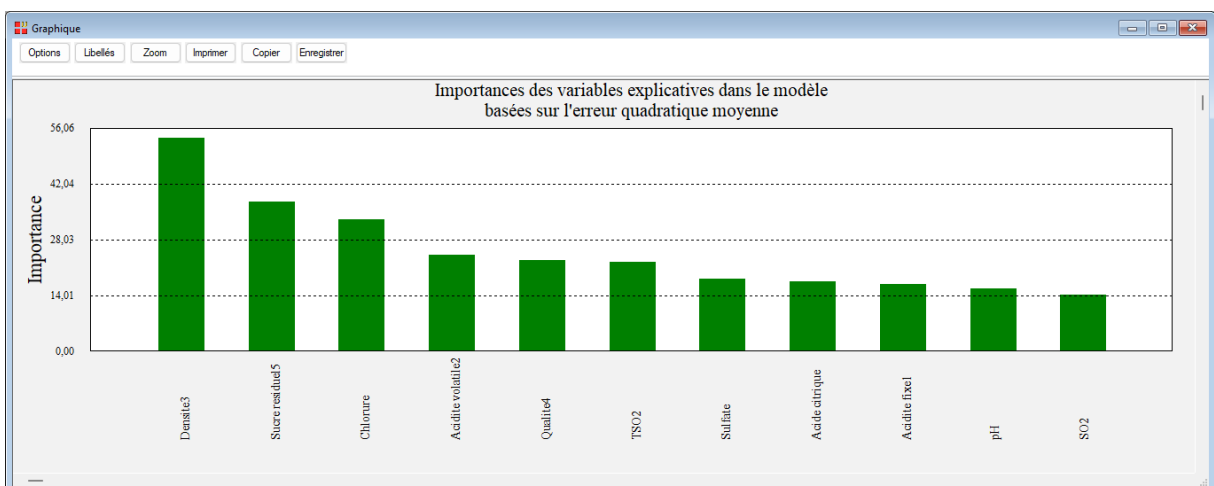
Rapports et Graphiques

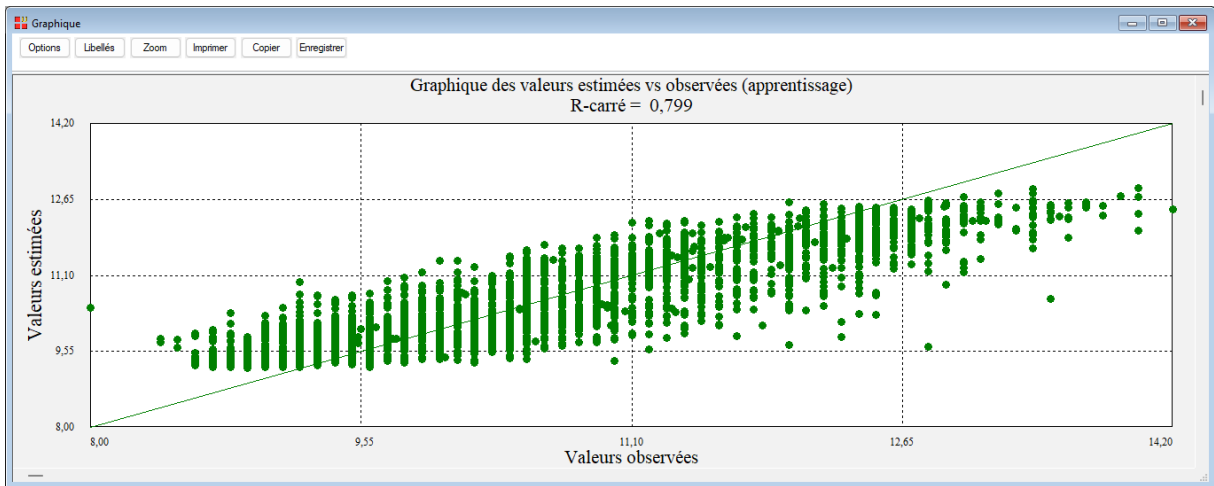
Rapport FORET

- Ereurs quadratiques et R-carrés
- Importances des variables explicatives
- Utilisation des variables explicatives
- Nombres de noeuds terminaux
- Nb. fois observation OOB
- Observés, estimés (apprentissage)
- Estimés (prévision)

	1	2	3	4	5	6	7	8
1								
2	JEU DE PREVISION							
3								
4								
5			Estimé					
6	o15		9,42852					
7	o18		12,46011					
8	o34		10,23248					
9	o57		9,64450					
10	o86		9,50604					
11	o90		9,50604					
12	o108		9,40253					
13	o144		10,58561					
14	o148		9,78697					
15	o207		9,81395					
16	o211		10,79348					
17	o220		10,62235					
18	o232		9,72270					
19	o236		9,30090					
20	o279		10,94116					
21	o337		11,46137					

Rapport Explorateur /





Exemple 5 : Fichier TITANIC (décision)

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 ^{ère} , 2 ^{ème} ou 3 ^{ème})
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 FORET dans le ruban Expliquer et renseignons la boîte de dialogue comme montré ci-dessous.

Après exécution de la procédure, visualisons la matrice de confusion (OOB) des données d'apprentissage et les courbes ROC associées.

Forêts aléatoires

Statut

Age
Tarif
Nbfse
Nbpe
Classe
Sexe
Poids
LibVarQuanti
LibVarQuali
LibObs

Variable à expliquer :
Statut

Variables explicatives quantitatives :
Age
Tarif
Nbfse
Nbpe

Variables explicatives qualitatives :
Classe
Sexe

Type de forêt
 Classement Régression

Nombre d'arbres à créer : 500

Nombre de variables à tester
 Défaut Personnalisé : 0

Nombre maximum de noeuds terminaux : 50

Taille minimum d'un noeud terminal : 1

Taille de l'échantillon des observations
 Défaut Personnalisé : 0

Type d'échantillonnage
 Avec remise Sans remise

Racine aléatoire : 12345

(Libellés des variables quantitatives :)
LibVarQuanti

(Libellés des variables qualitatives :)
LibVarQuali

(Libellés des observations :)
LibObs

Ok Annuler Sélection Supprimer Aide

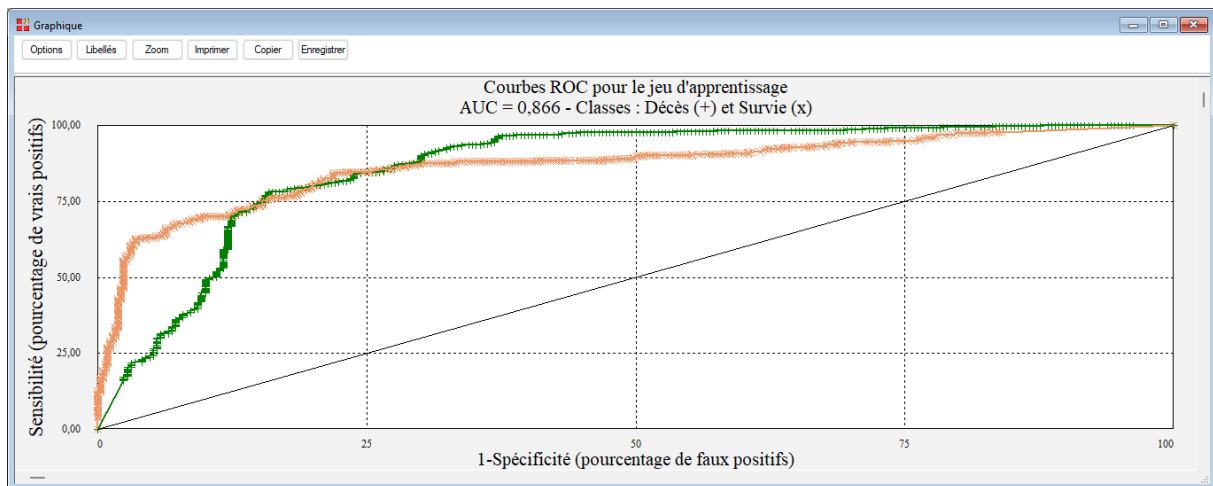
Rapports et Graphiques

Rapport FORET

- Taux d'erreur de prévision
- Importances relatives des variables
- Utilisation des variables explicatives
- Nombres de noeuds terminaux
- Nb. fois observation OOB
- Détail classement apprentissage
- Matrice de confusion (apprentissage)
- Sensibilité, Spécificité Décès (app)
- Sensibilité, Spécificité Survie (app)

	1	2	3	4	5	6	7	8
1								
2	MATRICE DE CONFUSION DE LA POPULATION D'APPRENTISSAGE (OOB)							
3								
4	En lignes, les classes observées							
5	En colonnes, les classes prévues							
6								
7	Pourcentage de mal classés : 17,787 %							
8	Pourcentage de bien classés (exactitude) : 82,213 %							
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	424	391	33	0,80619	0,92217	0,86029	
17	Survie	290	94	196	0,85590	0,67586	0,75530	
18								
19								
20								
21								

Rapport Explorateur /



Environ 82 % des passagers sont bien classés par cette analyse et l'aire sous la courbe ROC (AUC) est proche de 0,87.

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.

Calculs de la matrice de confusion et des indicateurs

Dans le cas de deux classes A et B, nous avons le tableau suivant :

	Prévu A	Prévu B	Total	% correct
Observé A	VP	FN	VP + FN	$\frac{100 * VP}{(VP + FN)}$
Observé B	FP	VN	FP + VN	$\frac{100 * VN}{(VN + FP)}$
Total	VP + FP	FN + VN	VP + FP + VN + FN	
% correct	$\frac{100 * VP}{(VP + FP)}$	$\frac{100 * VN}{(FN + VN)}$		$\frac{100 * (VP + VN)}{(VP + VN + FP + FN)}$
				% total correctement prévu

Dans le cas multi-classes (plus de 2 classes), chaque classe est étudiée par rapport une classe virtuelle réunissant l'ensemble des autres classes.

Définition des indicateurs :

- la sensibilité $VP / (VP+FN)$
- la spécificité $VN / (VN+FP)$
- l'exactitude $(VP+VN) / (VP+VN+FP+FN)$
- la précision $VP / (VP+FP)$
- le rappel $VP / (VP+FN)$
- le score F1 $2 \times (\text{précision} \times \text{rappel}) / (\text{précision} + \text{rappel})$

La sensibilité (ou rappel) indique la capacité du modèle à prévoir les vrais positifs.

La spécificité (ou taux de vrais négatifs) permet de mesurer la capacité du modèle à prévoir les vrais négatifs.

L'exactitude mesure le pourcentage de prévisions correctes par rapport à toutes les prévisions positives et négatives. Elle varie entre 0 et 1 et est sensible aux données déséquilibrées. Plus elle est proche de 1, meilleure est la prévision globale.

Le rappel (ou sensibilité ou taux de vrais positifs) varie entre 0 et 1 et n'est pas sensible aux données déséquilibrées. Un rappel égal à 1 indique une prévision parfaite des positifs.

La précision mesure le pourcentage de prévisions positives correctes. Elle varie entre 0 et 1 et n'est pas sensible aux données déséquilibrées. Une précision égale à 1 indique que tous les positifs sont prédits positifs.

Le score F1 combine la précision et le rappel en utilisant les moyennes harmoniques. Il varie entre 0 et 1. Maximiser ce score revient à maximiser la précision et le rappel. Il n'est pas sensible aux données déséquilibrées.

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>
libvarexp	Libellés des variables explicatives
importance1	Importances des variables explicatives
importance2	Importances des variables explicatives
utilvar	Utilisation des variables explicatives
nbnoeuds	Nombres de nœuds terminaux
nbfoisoob	Nombres de fois OOB

tauxerreur	Taux d'erreur de prévision (décision)
errquad	Erreurs quadratiques moyennes des arbres (régression)
R2	R-carrés des arbres (régression)
obsA	Libellés des observations d'apprentissage
voteA	Votes pour les données d'apprentissage (décision)
probA	Probabilités pour les données d'apprentissage (décision)
classeA	Classes prévues pour les données d'apprentissage (décision)
prevA	Valeurs prévues pour les données d'apprentissage (régression)
obsP	Libellés des observations de prévision
voteP	Votes pour les données de prévision (décision)
probP	Probabilités pour les données de prévision (décision)
classeP	Classes prévues pour les données de prévision (décision)
prevP	Valeurs prévues pour les données de prévision (régression)
seuilA	Seuils ROC (décision)
specificiteA	Spécificités (décision)
sensibiliteA	Sensibilités (décision)
aireA	Aires sous les courbes (décision)

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 – 'randomForest' \(2024\)](https://cran.r-project.org/web/packages/randomForest/randomForest.pdf)

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