UNIWIN VERSION 10.7.0

FORETS ALEATOIRES

Révision : 26/08/2026

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) 16
Exemple 3 : Fichier GRAISSE (régression) 18
L'option Rapports 20
L'option Graphiques 22
Exemple 4 : Fichier WINES3 (régression) 27
Exemple 5 : Fichier TITANIC (décision) 30
Calculs de la matrice de confusion et des indicateurs 32
Les variables internes créées par la procédure 33
Références 34

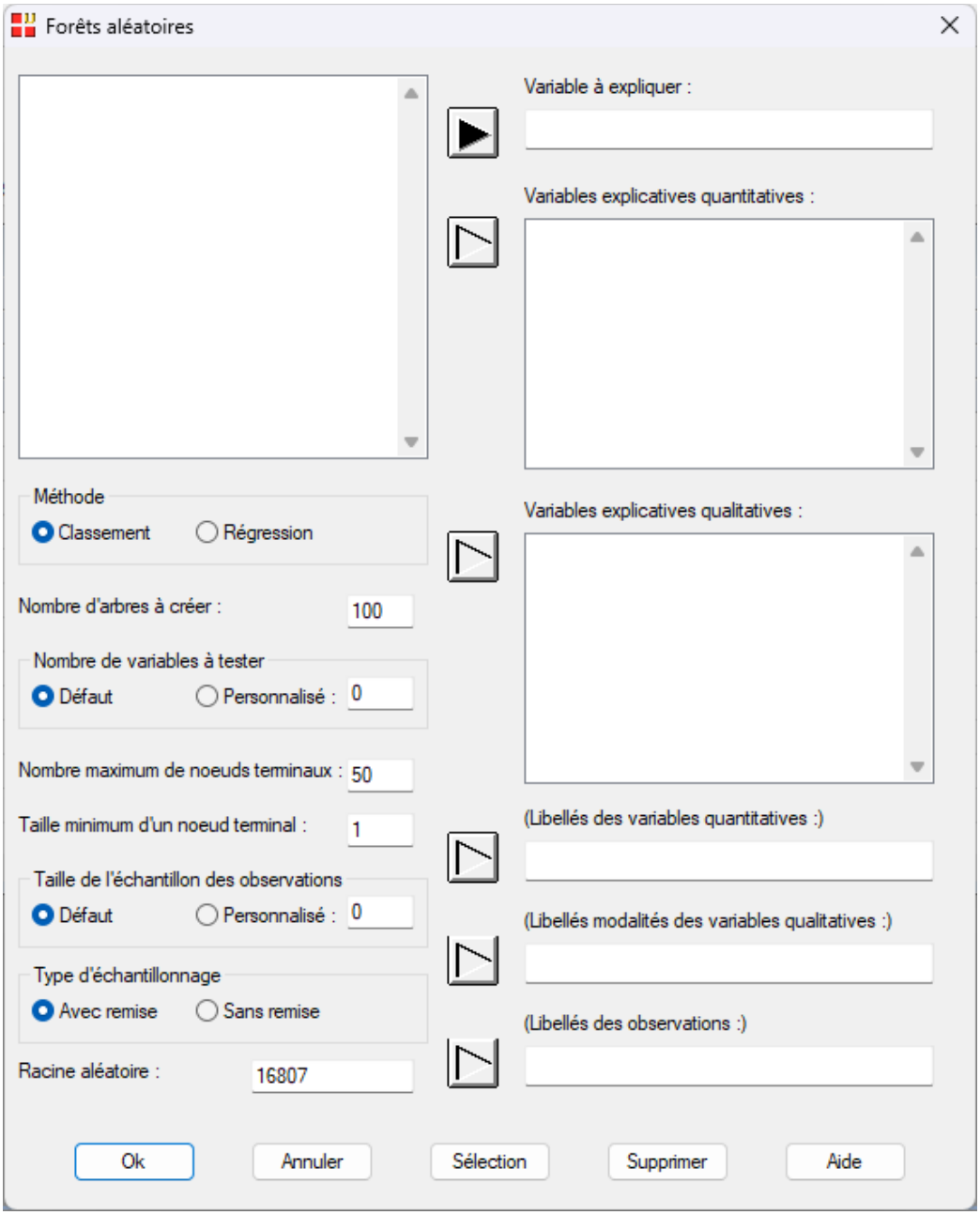
Définition

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 3 jeux : un jeu d'apprentissage utilisé pour construire les arbres, un jeu de validation 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 :



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, les libellés des modalités des variables 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 100.

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 explicatives quantitatives et qualitatives ne sont pas autorisées.
- Les valeurs manquantes de la variable à expliquer définissent le jeu de prévision.
- Les données non sélectionnées définissent le jeu de validation. A noter que les forêts aléatoires ne nécessitent pas de jeu de validation. Elles utilisent une technique appelée évaluation 'out-of-bag' (OOB), ensemble d'observations automatiquement non utilisées pour entrainer le modèle, pour mesurer la qualité du modèle. Toutefois, un jeu de validation peut se révéler utile pour l'optimisation des divers paramètres.

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.

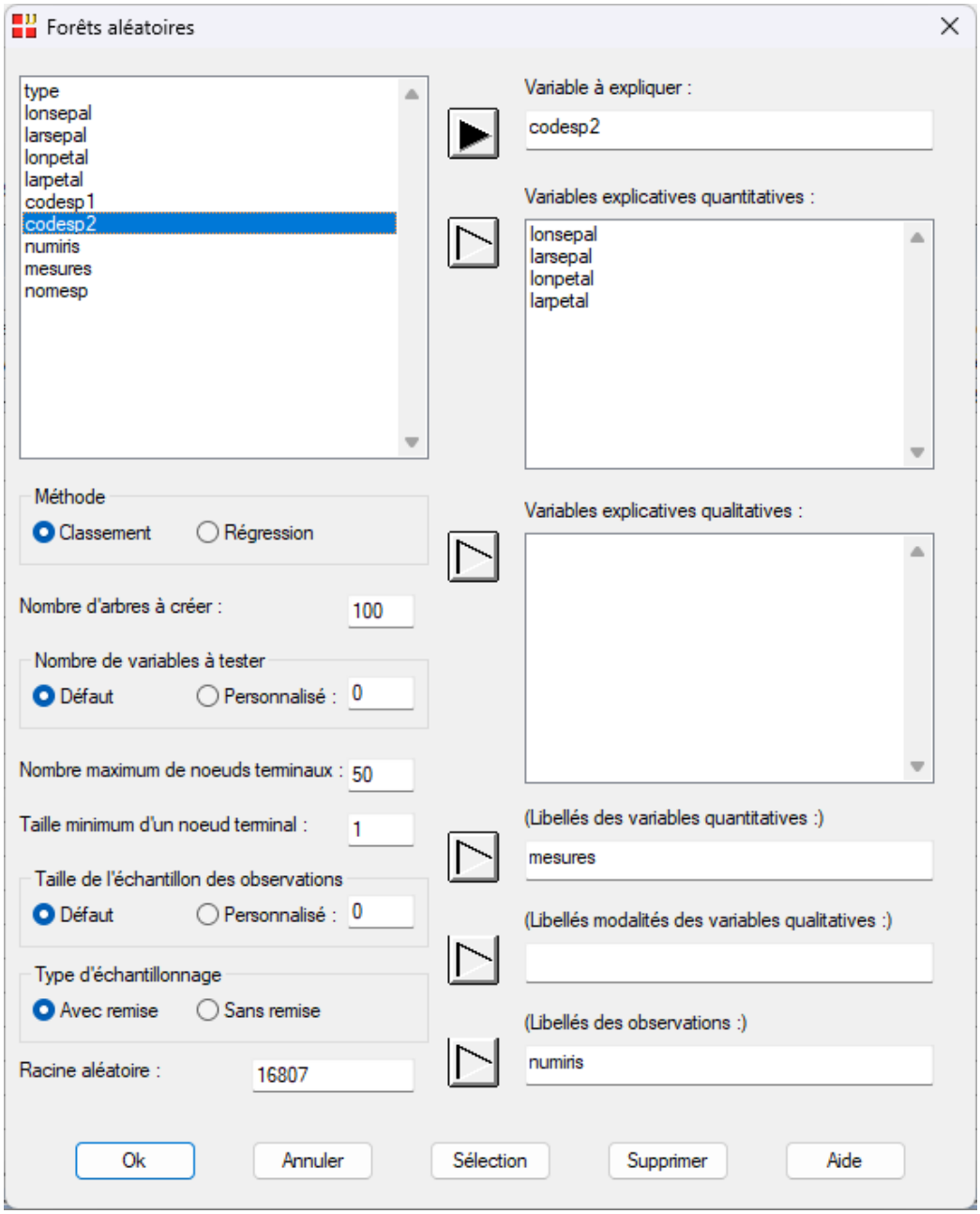
Iris Setosa (1)

Iris Versicolor (2)

Iris Virginica (3)

Cliquons sur l'icône ARBRE dans le ruban Expliquer.

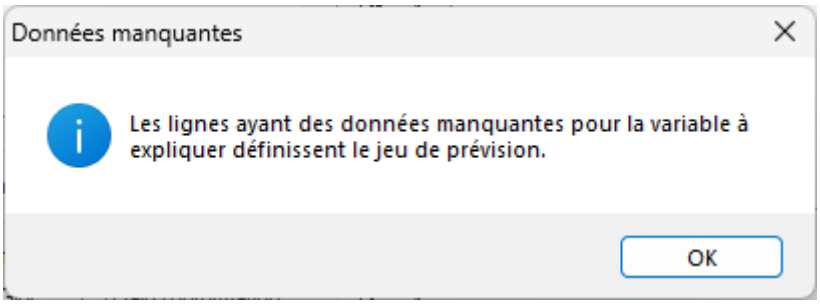
La boîte de dialogue ci-dessous s'affiche.



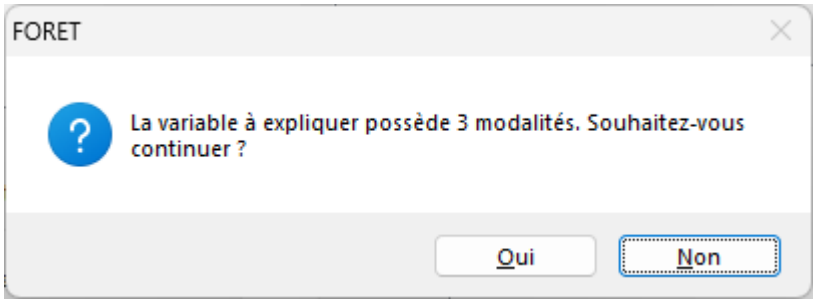
La variable *codesp2* est la variable à expliquer. Elle contient pour chaque observation le libellé de son espèce d'appartenance. Nous choisissons les variables de *lonsepal* à *larpetal* comme variables explicatives quantitatives et laissons les autres paramètres de l'analyse aux valeurs par défaut.

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.

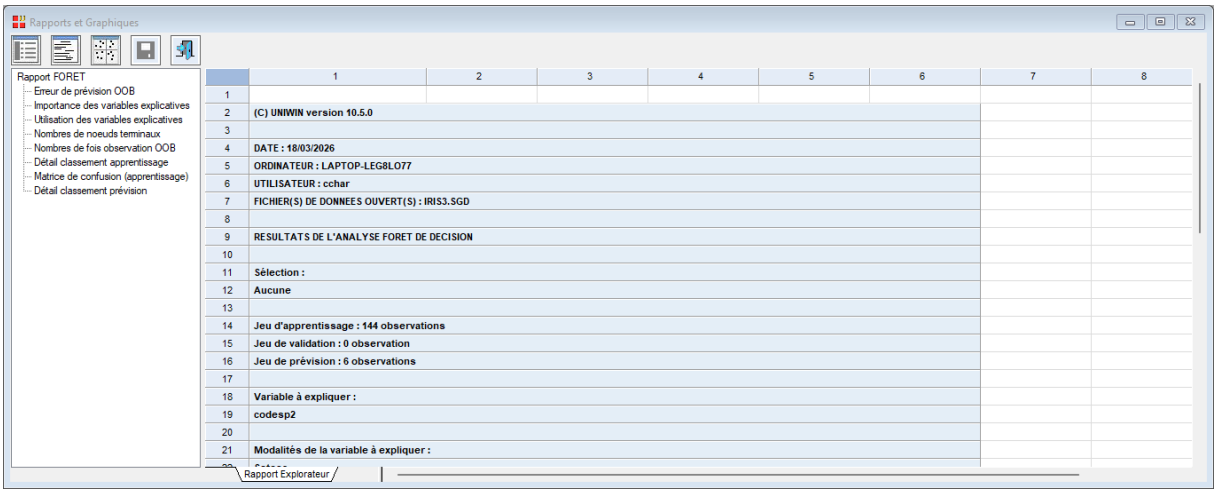


Un second message nous demande de confirmer notre choix d'un arbre de décision en fonction du nombre de modalités de la variable à expliquer :



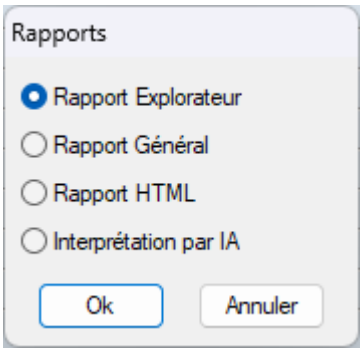
Cliquons sur Oui pour exécuter le traitement de l'analyse.

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

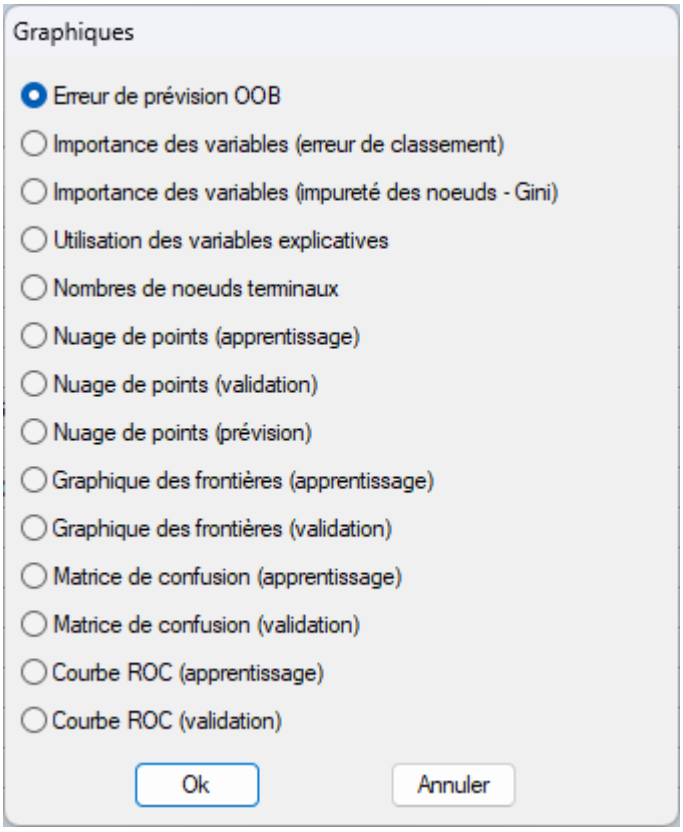


La barre d'outils 'Rapports et Graphiques' permet par l'icône 'Données' de rappeler la boîte de dialogue d'entrée des données.

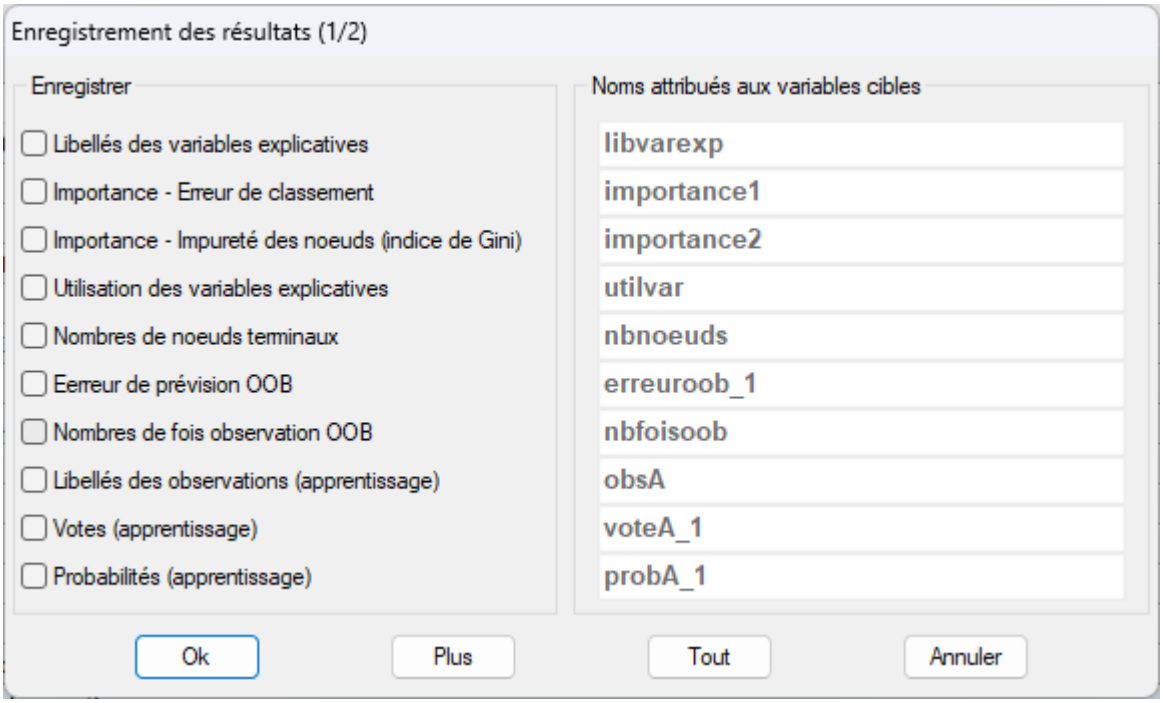
L'icône 'Rapports' affiche la boîte de dialogue des options pour les rapports :



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.



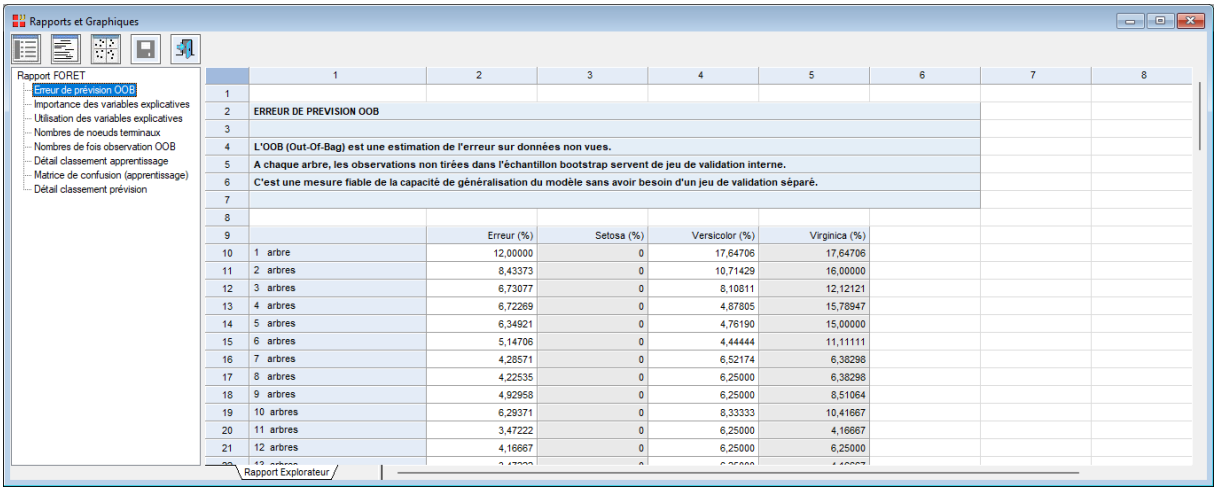
L'option Rapports

Cette option permet d'obtenir le rapport à l'écran sous la forme Explorateur, Général ou HTML et d'accéder à l'aide à l'interprétation par Intelligence Artificielle.

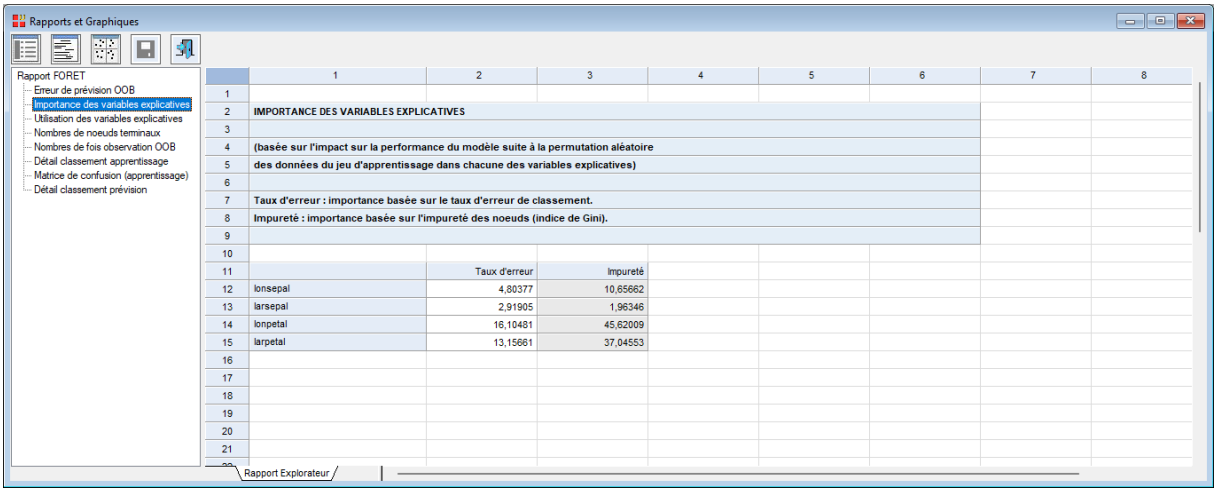
Erreur de prévision OOB

Ce tableau affiche l'évolution de l'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.

Le nombre optimal d'arbres à utiliser est celui à partir duquel l'erreur OOB se stabilise.



Importance des variables explicatives

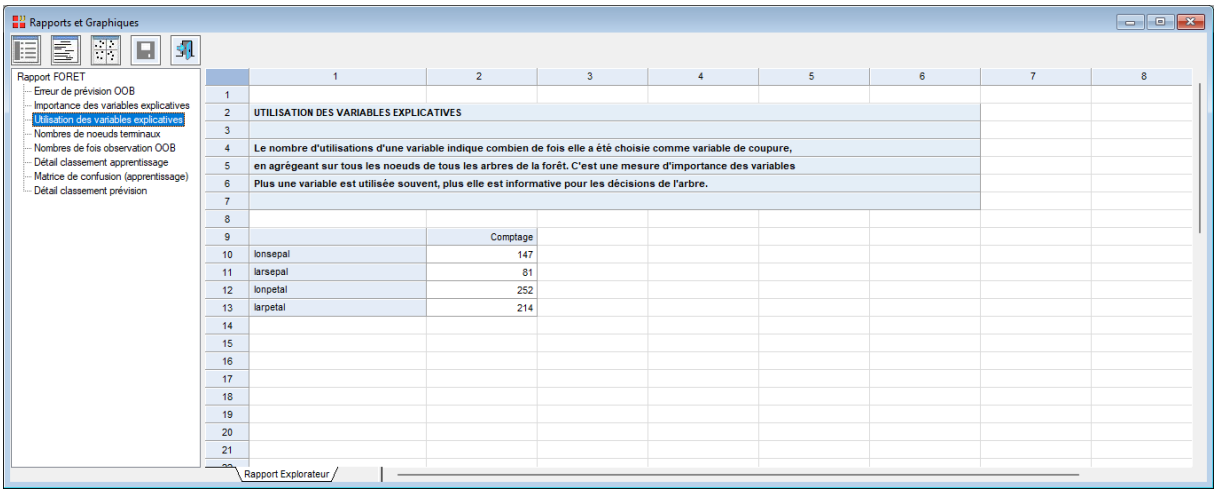


La première mesure est basée sur le taux d'erreur de classement. La deuxième mesure est basée sur l'impureté des nœuds (indice de Gini).

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

Utilisation des variables explicatives

Dans une forêt aléatoire, à chaque nœud de chaque arbre, l'algorithme sélectionne aléatoirement un sous-ensemble de variables (typiquement $\sqrt{p}$ variables pour une classification, p/3 pour une régression) et choisit la meilleure parmi elles pour effectuer la coupure. Le nombre d'utilisations d'une variable indique combien de fois elle a été choisie comme variable de coupure, en agrégeant sur tous les nœuds de tous les arbres de la forêt. C'est une mesure d'importance des variables. Plus une variable est utilisée souvent, plus elle est informative pour les décisions de l'arbre. Cela permet d'identifier les variables clés du modèle, de réduire la dimensionnalité en éliminant les variables peu utilisées et de détecter des redondances entre variables corrélées. Cette mesure peut toutefois surestimer l'importance des variables continues ou à forte cardinalité, car elles offrent plus de points de coupure possibles.

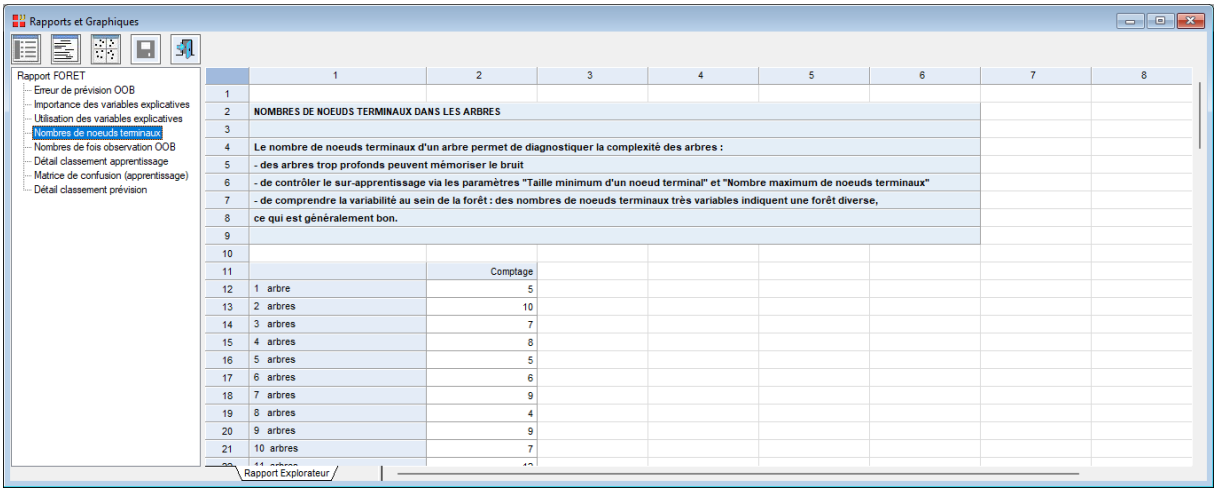


Nombres de nœuds terminaux dans les arbres

Ce tableau indique le nombre de nœuds terminaux dans chaque arbre de la forêt. Un nœud terminal (ou feuille) est un nœud qui n'est plus divisé.

Le nombre de nœuds terminaux d'un arbre permet de :

- Diagnostiquer la complexité des arbres : des arbres trop profonds peuvent mémoriser le bruit
- Contrôler le sur-apprentissage en utilisant les paramètres *Taille minimum d'un nœud terminal* et *Nombre maximum de nœuds terminaux.*
- Comprendre la variabilité au sein de la forêt : si les arbres ont des nombres de nœuds terminaux très variables, la forêt est diverse, ce qui est généralement bon



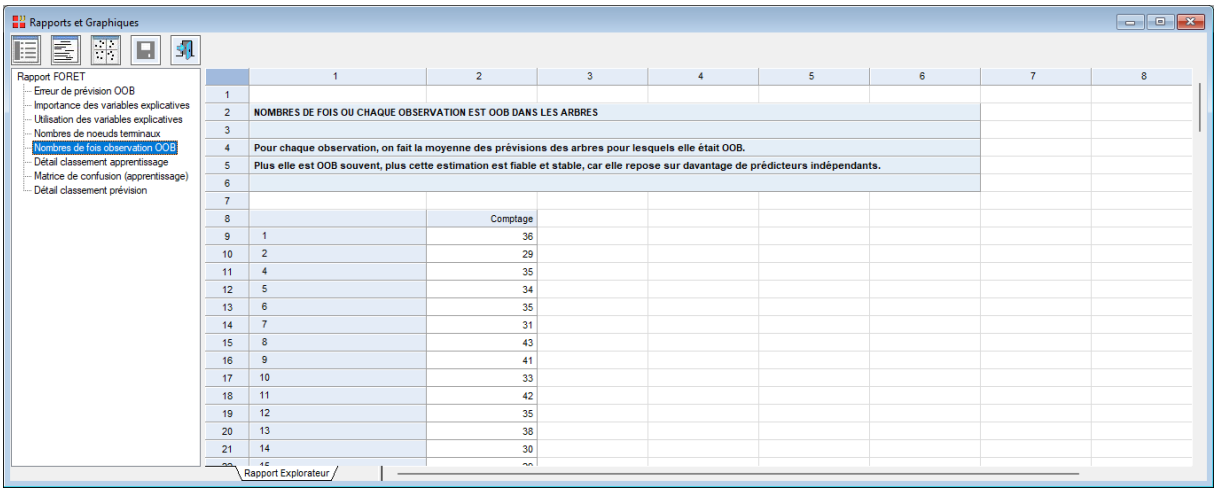
	1	2	3	4	5	6	7	8
1								
2	NOMBRES DE NOEUDS TERMINAUX DANS LES ARBRES							
3								
4	Le nombre de noeuds terminaux d'un arbre permet de diagnostiquer la complexité des arbres :							
5	- des arbres trop profonds peuvent mémoriser le bruit							
6	- de contrôler le sur-apprentissage via les paramètres "Taille minimum d'un noeud terminal" et "Nombre maximum de noeuds terminaux"							
7	- de comprendre la variabilité au sein de la forêt : des nombres de noeuds terminaux très variables indiquent une forêt diverse,							
8	ce qui est généralement bon.							
9								
10								
11		Comptage						
12	1 arbre	5						
13	2 arbres	10						
14	3 arbres	7						
15	4 arbres	8						
16	5 arbres	5						
17	6 arbres	6						
18	7 arbres	9						
19	8 arbres	4						
20	9 arbres	9						
21	10 arbres	7						

Nombre de fois où chaque observation est OOB dans les arbres

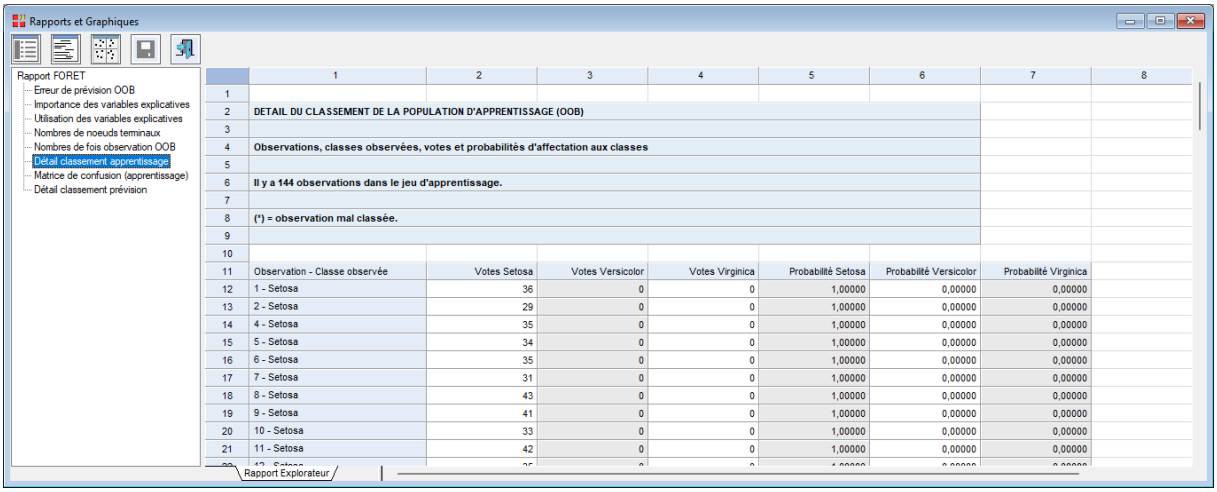
Pour chaque observation du jeu d'apprentissage, ce tableau indique le nombre de fois où cette observation a été OOB (out-of-bag) lors de la construction des arbres.

Ce comptage est crucial pour l'erreur OOB : pour chaque observation, on fait la moyenne des prévisions des arbres pour lesquels elle était OOB. Plus elle est OOB souvent, plus cette estimation est fiable et stable, car elle repose sur davantage de prédicteurs indépendants.

En pratique, avec un nombre d'arbres grand (100 à 500), chaque observation sera OOB suffisamment de fois pour obtenir une bonne estimation de l'erreur de généralisation, sans avoir besoin d'un ensemble de validation séparé.



Détail du classement de la population d'apprentissage



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.

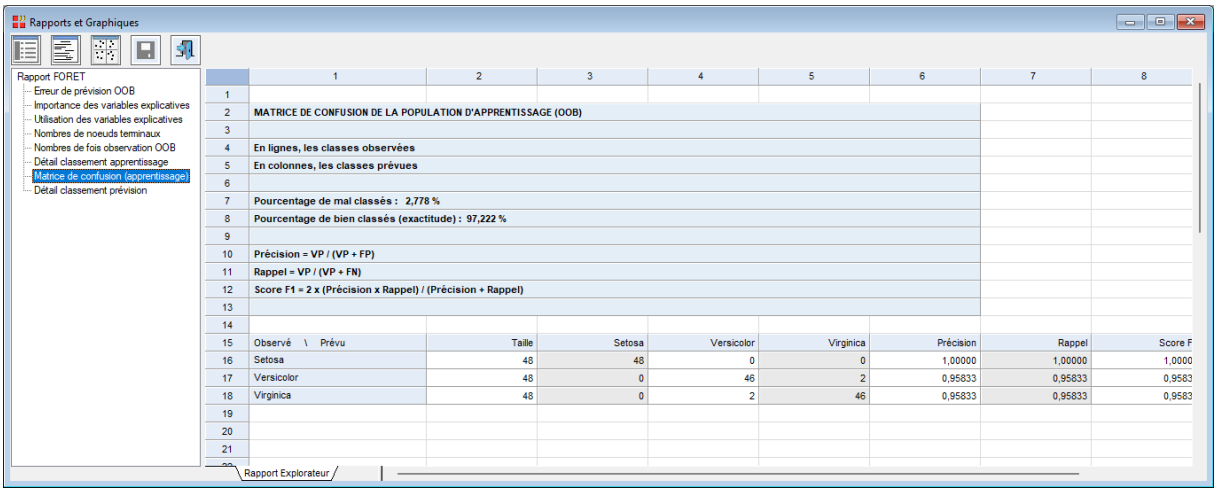
Pour chaque observation, ces votes majoritaires et ces probabilités sont calculés en utilisant uniquement les arbres dans lesquels cette observation est OOB.

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

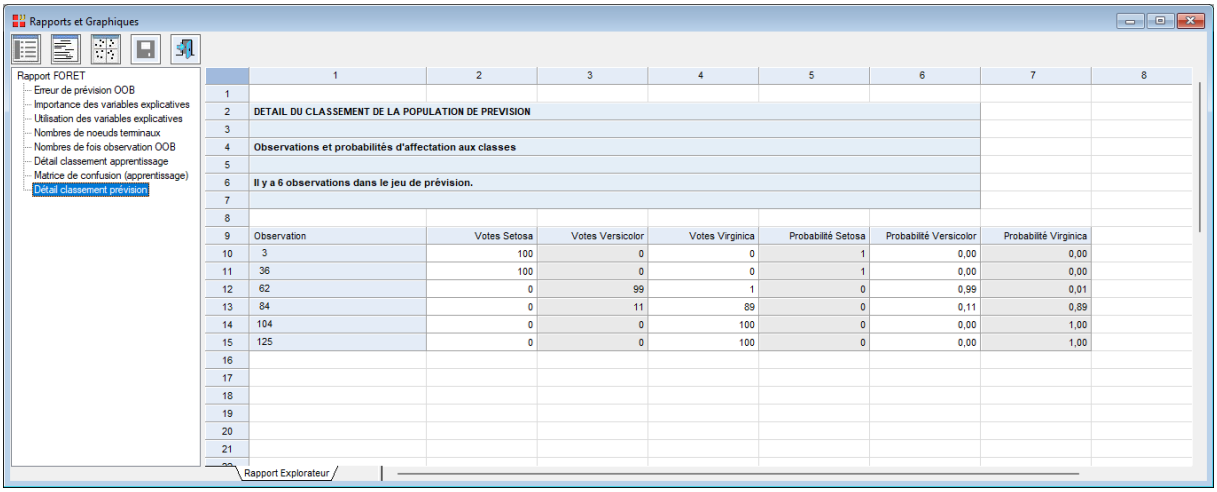
Matrice de confusion du jeu d'apprentissage (OOB)

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



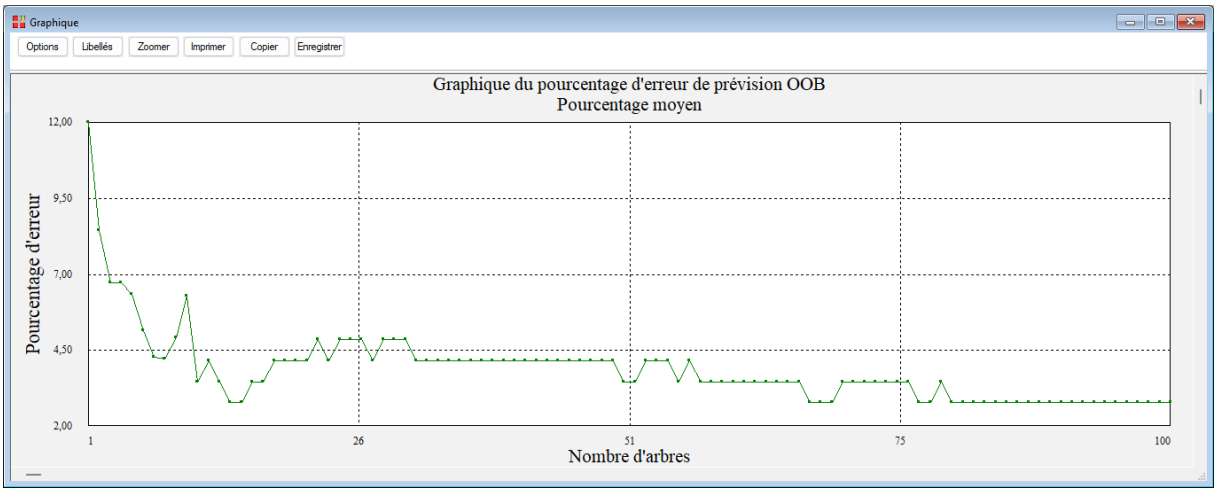
Pour chaque observation de la population de prévision, les votes majoritaires pour chacune des classes et les probabilités d'affectation à chacune des classes sont affichés.

L'option Graphiques

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

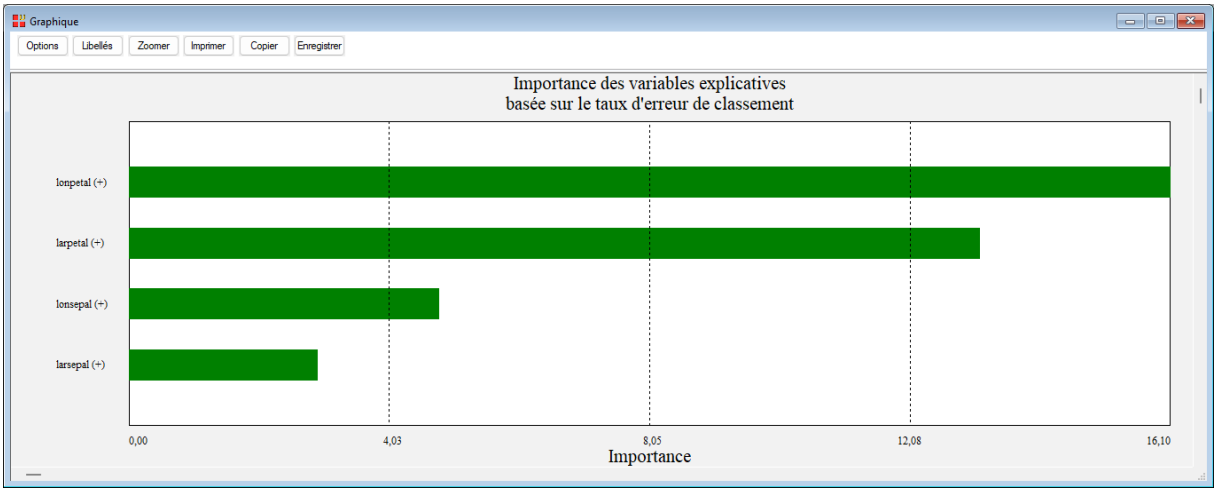
Graphique du pourcentage d'erreur OOB

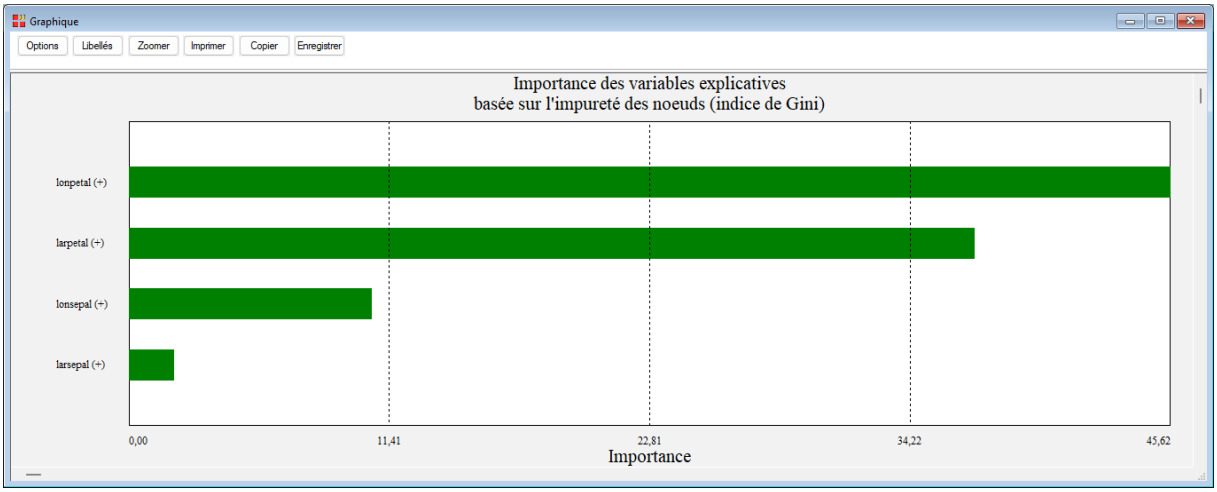
Ce graphique affiche l'évolution du pourcentage moyen d'erreur ou du pourcentage d'erreur pour chacune des classes en fonction du nombre d'arbres dans la forêt.



Graphiques de l'importance des variables explicatives

Ces graphiques affichent l'importance de chaque variable explicative prise en compte lors de la construction des arbres de la forêt. La longueur de chaque barre représente le taux d'erreur de classement ou l'impureté des nœuds (Gini). Les variables sont classées de gauche à droite par ordre décroissant d'importance.



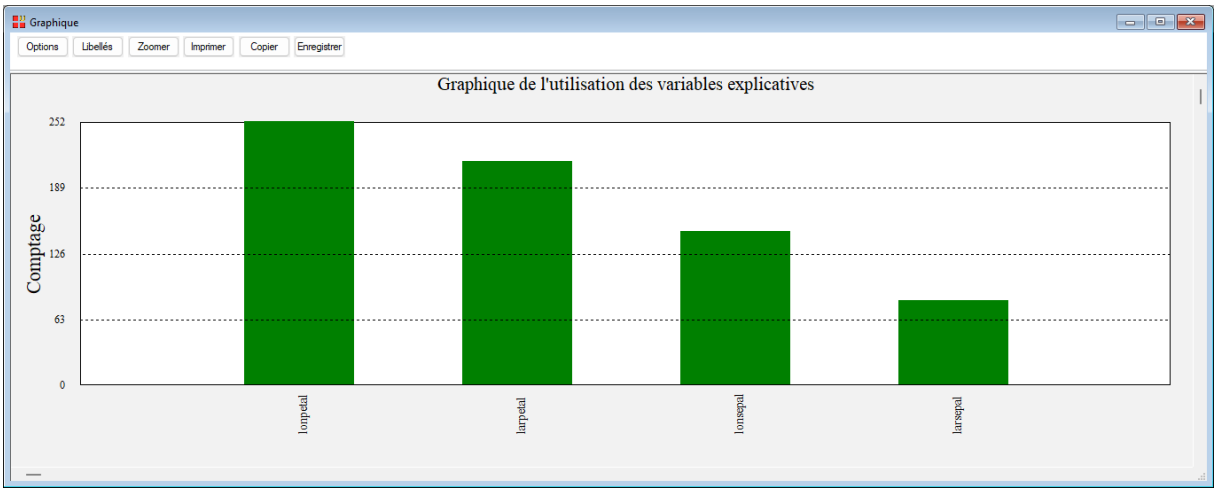


Graphique de l'utilisation des variables explicatives

Le nombre d'utilisations d'une variable indique combien de fois elle a été choisie comme variable de coupure, en agrégeant sur tous les nœuds de tous les arbres de la forêt.

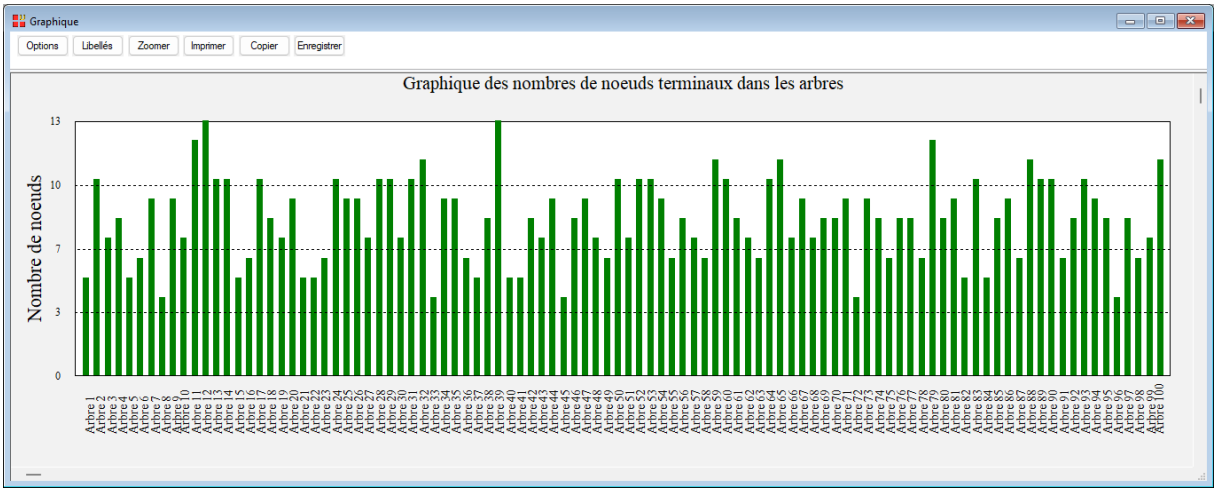
Plus une variable est utilisée souvent, plus elle est informative pour les décisions de l'arbre. Cela permet d'identifier les variables clés du modèle.

Les variables sont classées de gauche à droite par ordre décroissant d'utilisation.



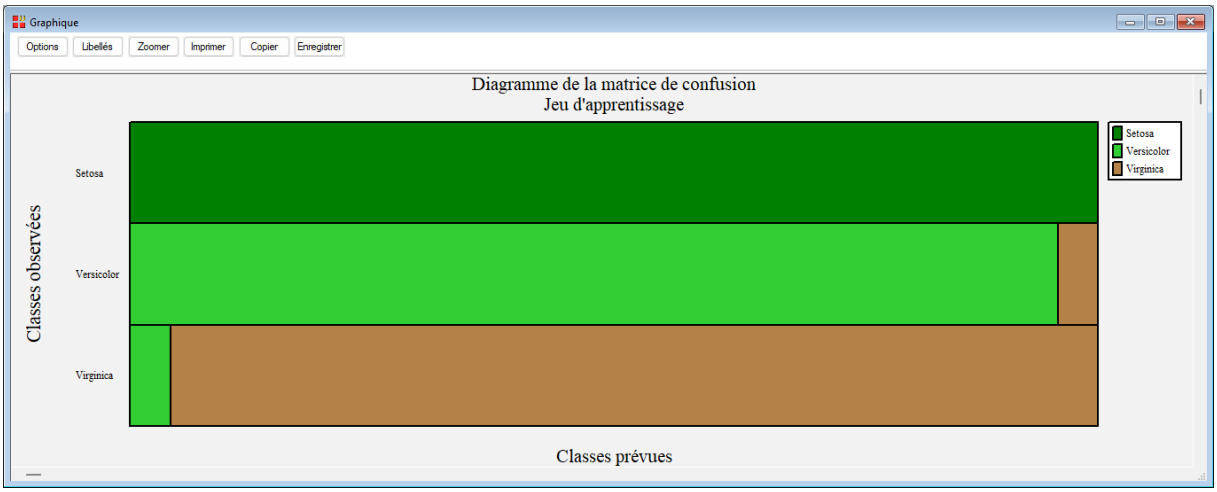
Graphique des nombres de nœuds terminaux dans les arbres

Ce graphique indique les nombres de nœuds terminaux dans chaque arbre de la forêt. Ils permettent de diagnostiquer la complexité des arbres, de contrôler le sur-apprentissage et de comprendre la variabilité au sein de la forêt : si les arbres ont des nombres de nœuds terminaux très variables, la forêt est diverse, ce qui est généralement bon.



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.



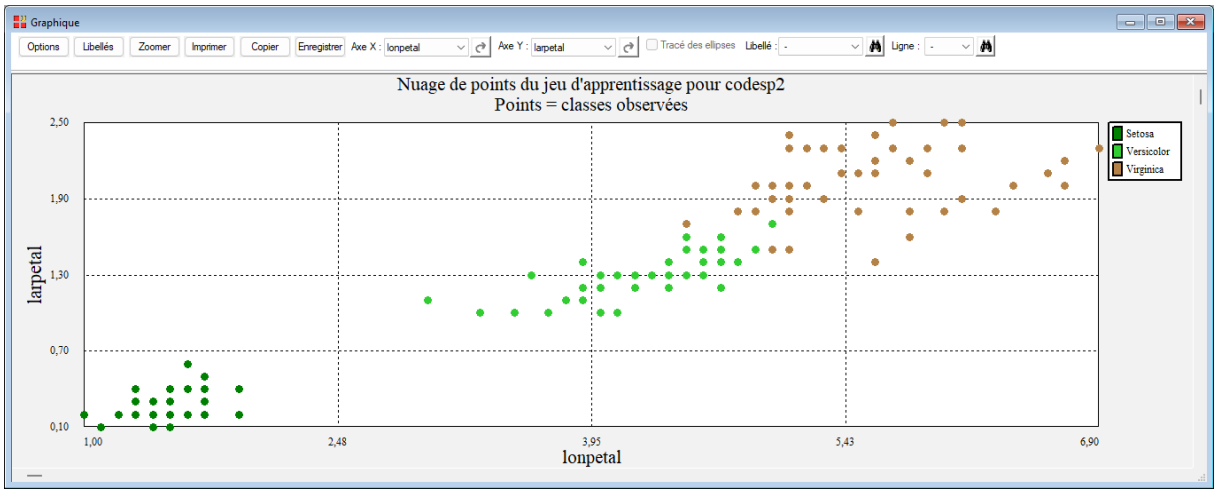
Courbes ROC

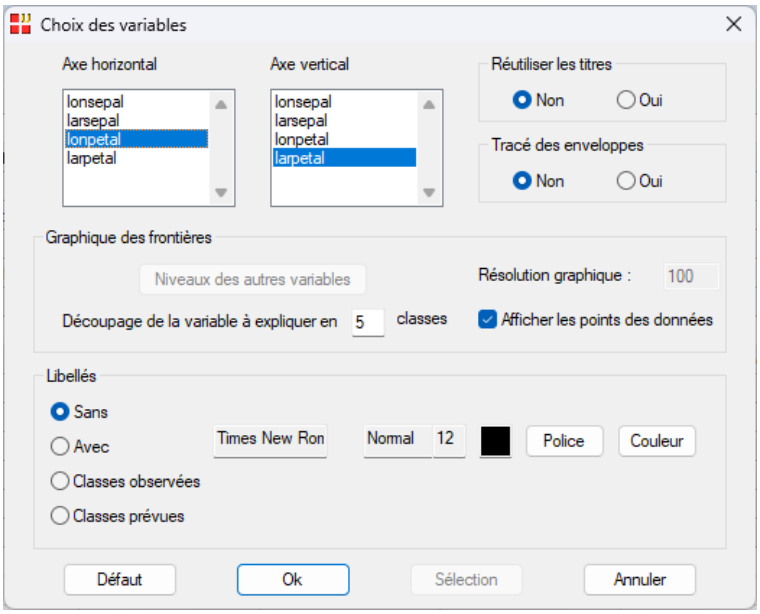
La courbe ROC pour le jeu d'apprentissage n'est disponible que dans le cas de deux classes et pour un arbre de décision. Il y a trois classes dans cet exemple et donc le graphique n'est pas proposé.

Nuages de points

Cette option permet d'afficher un nuage des observations pour deux variables sélectionnées. Il est possible d'afficher les points avec ou sans libellés, avec les codes des classes observées ou avec les codes des classes prévues. Dans le graphique, la barre d'outils permet de modifier les variables affichées.

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



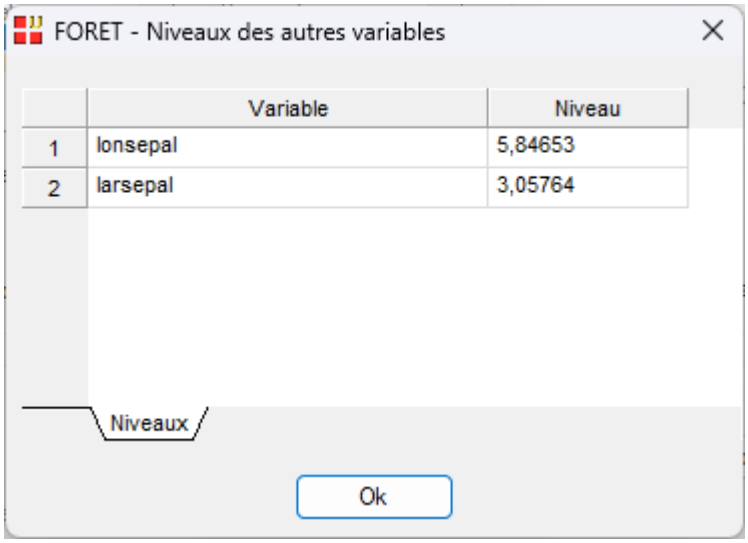


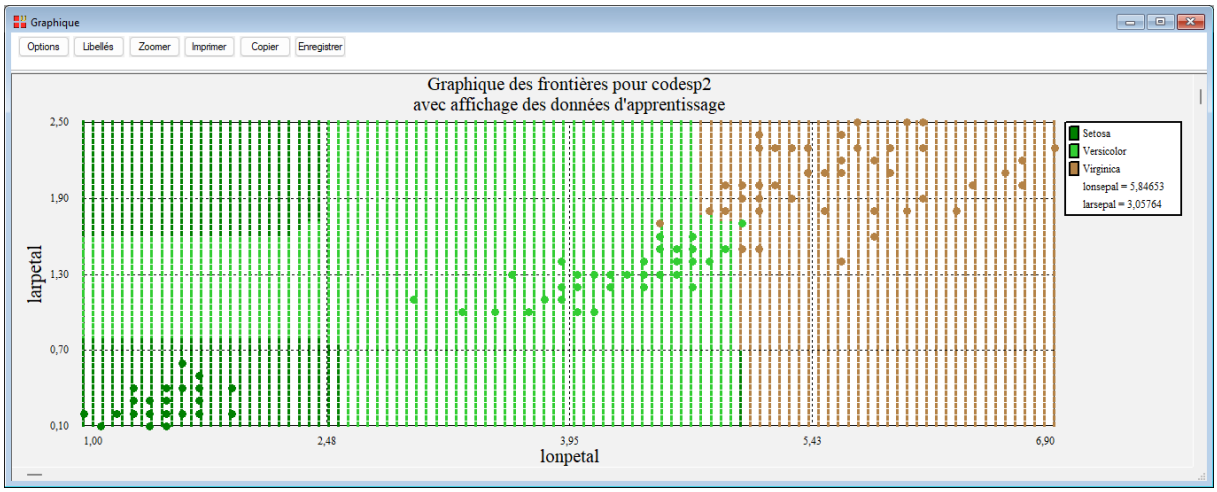
Graphique des frontières

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

Pour préciser les niveaux des autres variables, il faut cliquer sur le bouton 'Niveaux des autres variables' dans la boîte de dialogue 'Choix des variables'. Par défaut les autres variables quantitatives sont maintenues aux valeurs moyennes des données du jeu d'apprentissage.

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





Exemple 2 : Fichier DIABETES (décision)

Nous utiliserons le fichier DIABETES pour ce deuxième exemple.

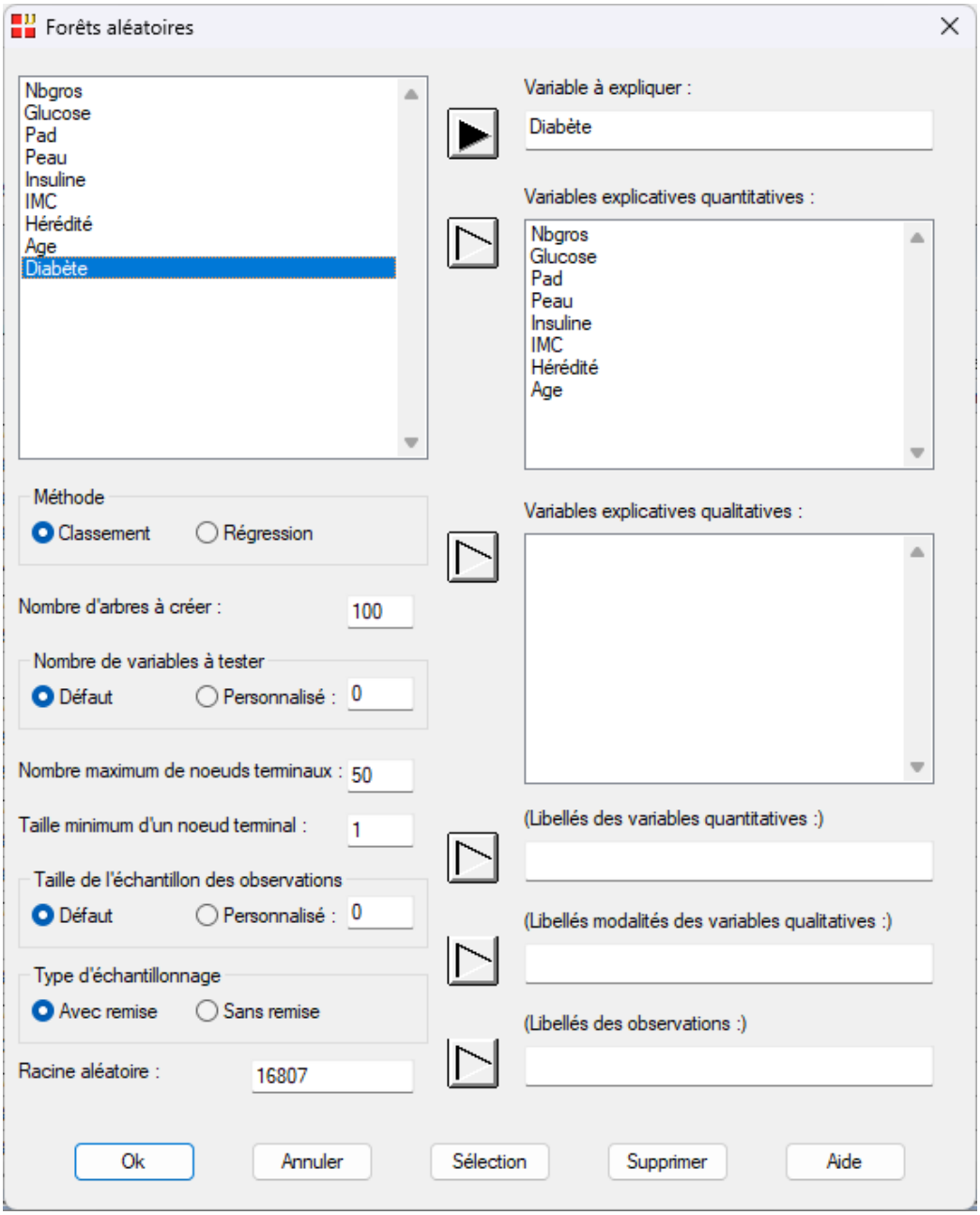
Une population de 768 femmes âgées d'au moins 21 ans, d'origine indienne Pima et vivant près de Phoenix, en Arizona, a été testée pour le diabète selon les critères de l'Organisation Mondiale de la Santé.

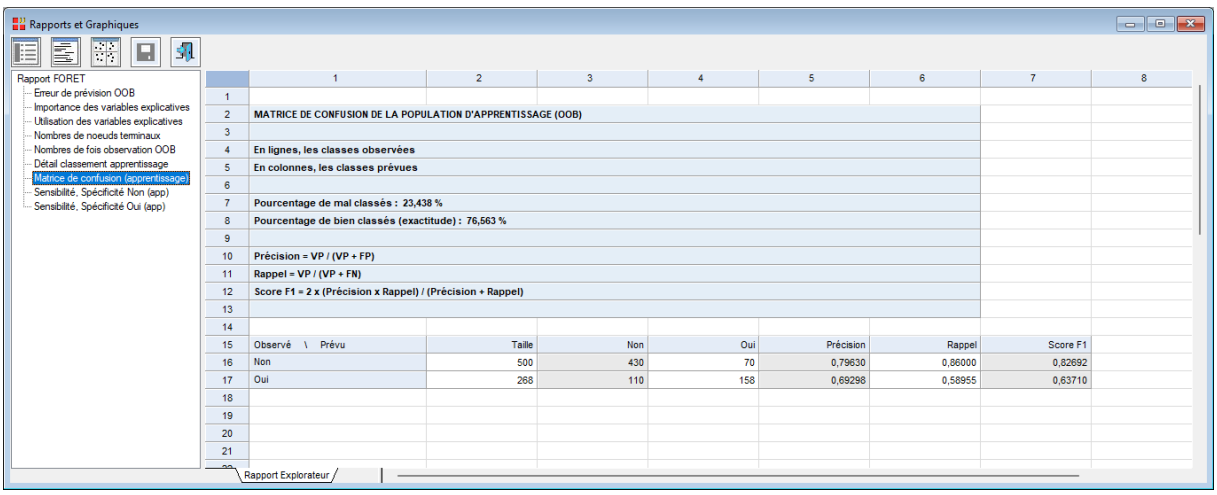
Les données ont été recueillies par l'Institut national américain du diabète et des maladies digestives et rénales.

Neuf variables ont été collectées :

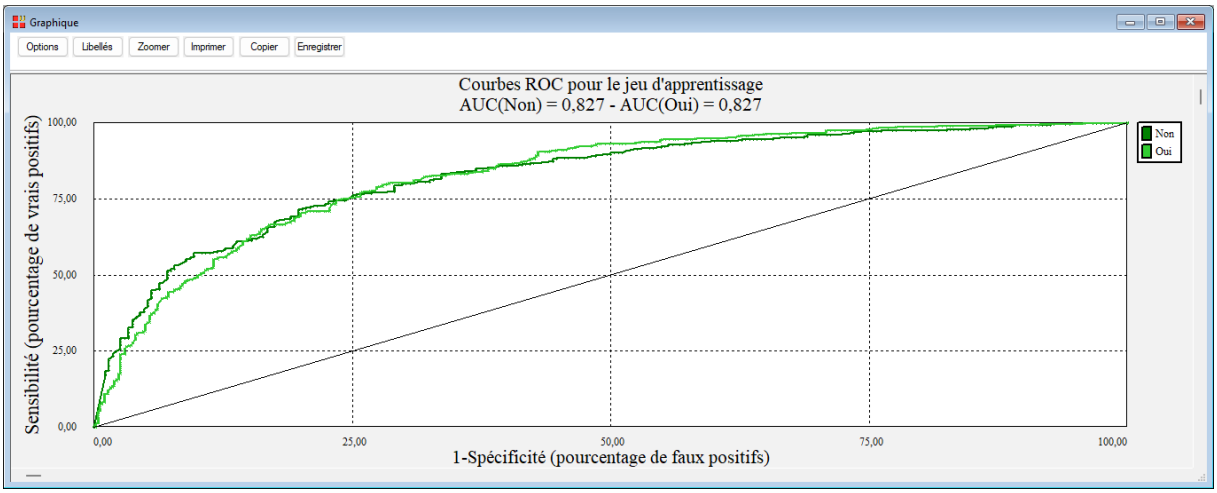
- Nbgros : nombre de grossesses
- Glucose : concentration plasmatique de glucose à 2 heures dans un test oral de tolérance au glucose
- Pad : pression artérielle diastolique (mm Hg)
- Peau : épaisseur du pli cutané du triceps (mm)
- Insuline : insuline sérique 2 heures (mu U/ml)
- IMC : indice de masse corporelle (poids en kg/(taille en m)^2)
- Hérédité : 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).


Forêts aléatoires
Nbgros
Glucose
Pad
Peau
Insuline
IMC
Hérédité
Age
Diabète
Variable à expliquer :
Diabète
Variables explicatives quantitatives :
Nbgros
Glucose
Pad
Peau
Insuline
IMC
Hérédité
Age
Méthode
Classement
Régression
Variables explicatives qualitatives :
Nombre d'arbres à créer :
100
Nombre de variables à tester
Défaut
Personnalisé :
0
Nombre maximum de noeuds terminaux :
50
Taille minimum d'un noeud terminal :
1
(Libellés des variables quantitatives :)
Taille de l'échantillon des observations
Défaut
Personnalisé :
0
(Libellés modalités des variables qualitatives :)
Type d'échantillonnage
Avec remise
Sans remise
(Libellés des observations :)
Racine aléatoire :
16807
Ok
Annuler
Sélection
Supprimer
Aide


Rapports et Graphiques
Rapport FORET
Erreur de prévision OOB
Importance des variables explicatives
Utilisation des variables explicatives
Nombres de noeuds terminaux
Nombres de fois observation OOB
Détail classement apprentissage
Matrice de confusion (apprentissage)
Sensibilité, Spécificité Non (app)
Sensibilité, Spécificité Oui (app)
MATRICE DE CONFUSION DE LA POPULATION D'APPRENTISSAGE (OOB)
En lignes, les classes observées
En colonnes, les classes prévues
Pourcentage de mal classés : 23,438 %
Pourcentage de bien classés (exactitude) : 76,563 %
Précision = VP / (VP + FP)
Rappel = VP / (VP + FN)
Score F1 = 2 x (Précision x Rappel) / (Précision + Rappel)
Observé \ Prévu
Taille
Non
Oui
Précision
Rappel
Score F1
Non 500 430 70 0,79630 0,86000 0,82692
Oui 268 110 158 0,69298 0,58955 0,63710
Rapport Explorateur

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.

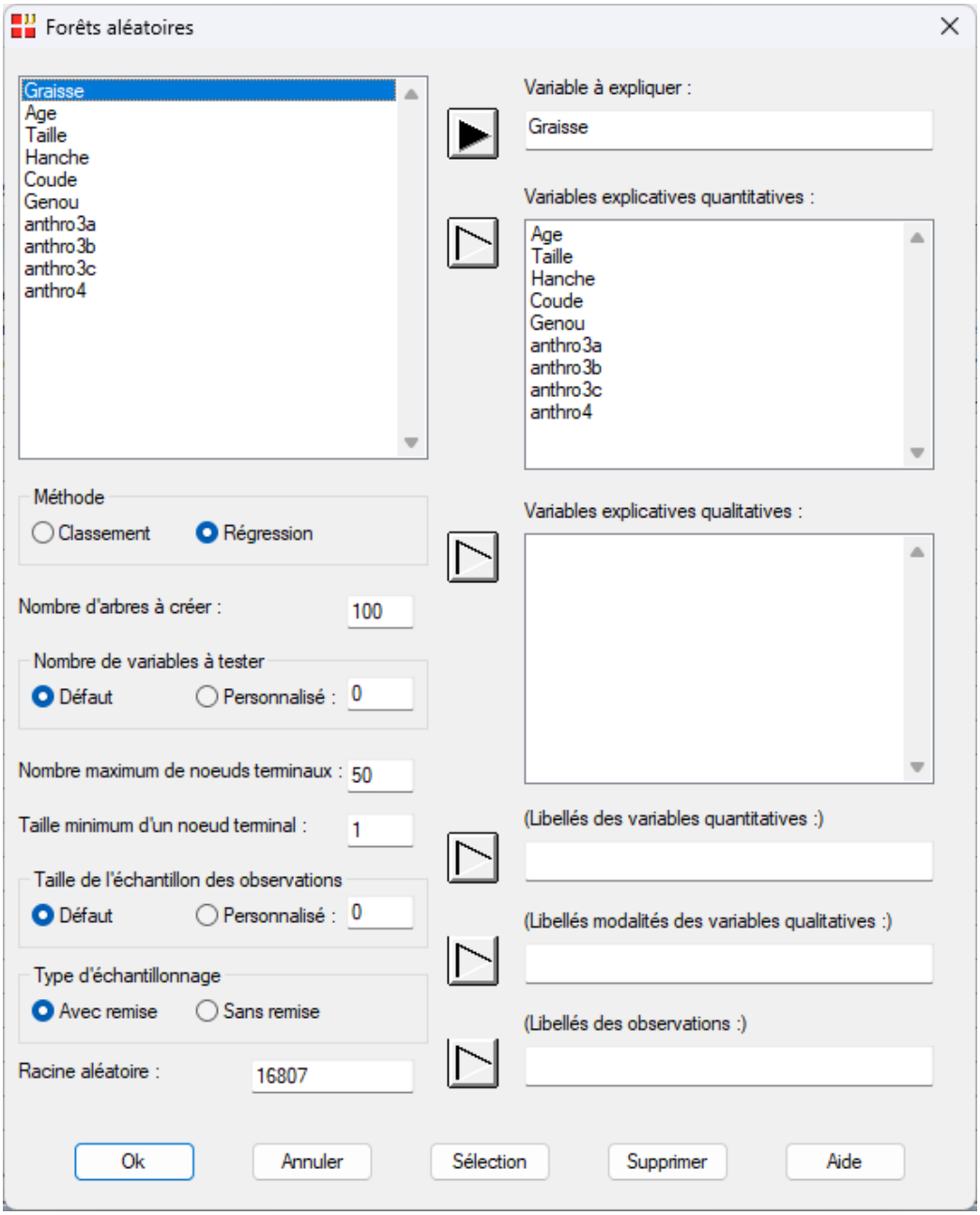
Exemple 3 : Fichier GRAISSE (régression)

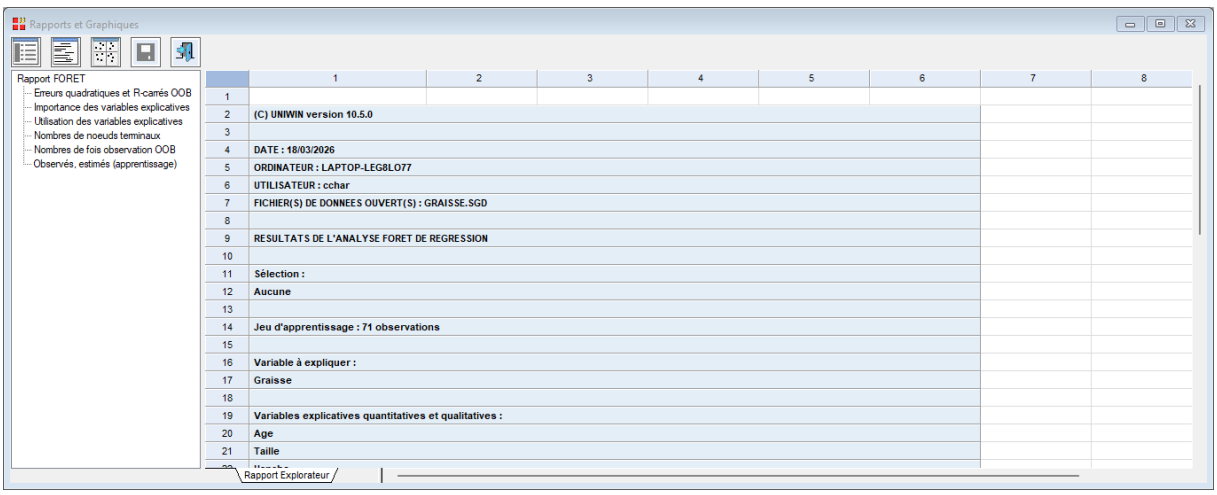
Pour 71 sujets féminins en bonne santé, neuf mesures anthropométriques sont utilisées pour modéliser la graisse corporelle.

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.)

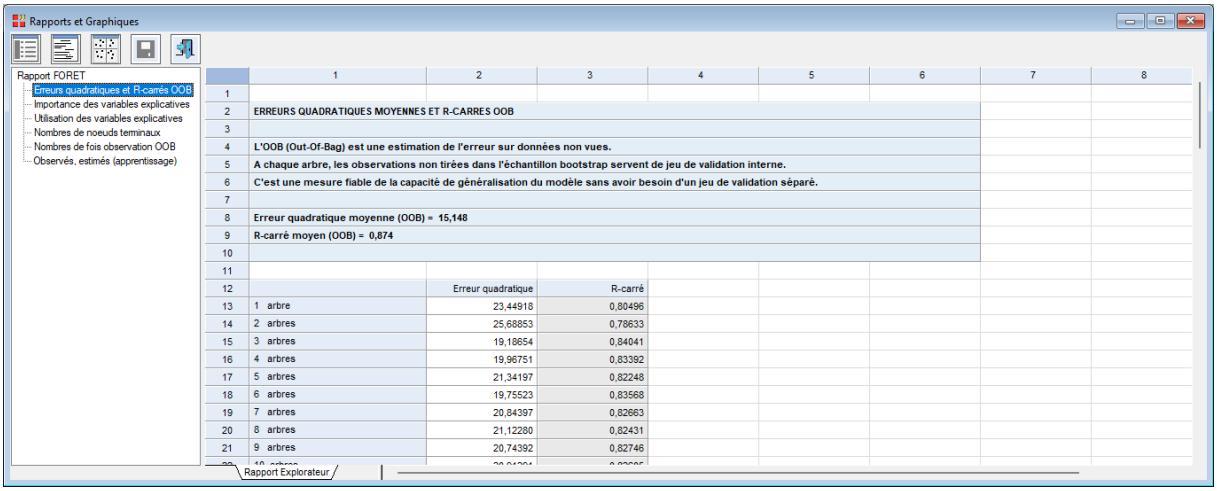
Renseignons la boîte de dialogue comme montré ci-dessous (les variables explicatives sélectionnées sont 'Age' à 'anthro4').


Forêts aléatoires
Graisse
Age
Taille
Hanche
Coude
Genou
anthro3a
anthro3b
anthro3c
anthro4
Variable à expliquer :
Graisse
Variables explicatives quantitatives :
Age
Taille
Hanche
Coude
Genou
anthro3a
anthro3b
anthro3c
anthro4
Méthode
Classement
Régression
Variables explicatives qualitatives :
Nombre d'arbres à créer :
100
Nombre de variables à tester
Défaut
Personnalisé :
0
Nombre maximum de noeuds terminaux :
50
Taille minimum d'un noeud terminal :
1
(Libellés des variables quantitatives :)
Taille de l'échantillon des observations
Défaut
Personnalisé :
0
(Libellés modalités des variables qualitatives :)
Type d'échantillonnage
Avec remise
Sans remise
(Libellés des observations :)
Racine aléatoire :
16807
Ok
Annuler
Sélection
Supprimer
Aide


Rapports et Graphiques
Rapport FORET
Erreurs quadratiques et R-carrés OOB
Importance des variables explicatives
Utilisation des variables explicatives
Nombres de noeuds terminaux
Nombres de fois observation OOB
Observés, estimés (apprentissage)
(C) UNIWIN version 10.5.0
DATE : 18/03/2026
ORDINATEUR : LAPTOP-LEG8LO77
UTILISATEUR : cchar
FICHIER(S) DE DONNEES OUVERT(S) : GRAISSE.SGD
RESULTATS DE L'ANALYSE FORET DE REGRESSION
Sélection :
Aucune
Jeu d'apprentissage : 71 observations
Variable à expliquer :
Graisse
Variables explicatives quantitatives et qualitatives :
Age
Taille
Rapport Explorateur

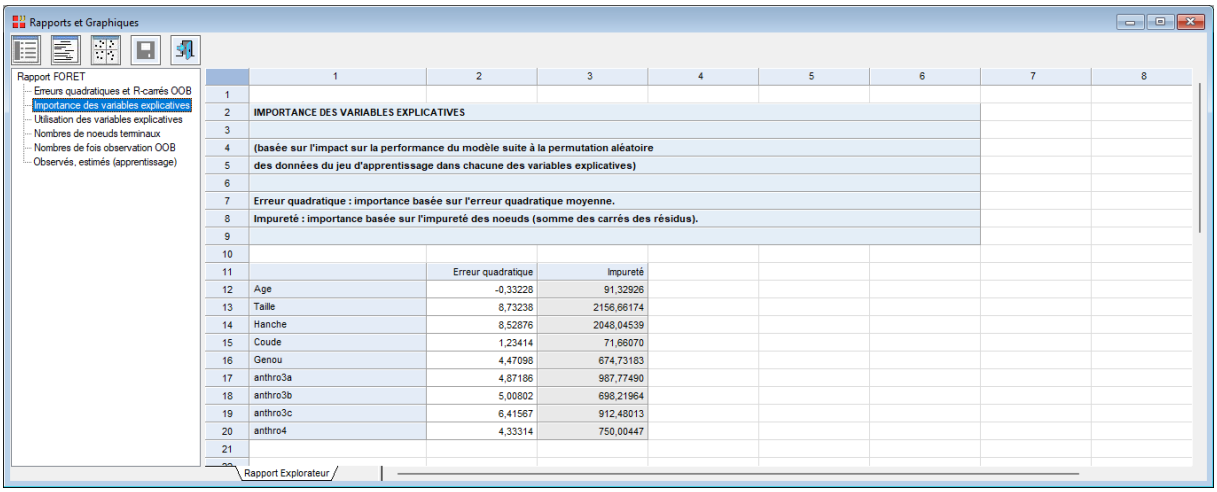
L’option Rapports

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



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.

Importance des variables explicatives

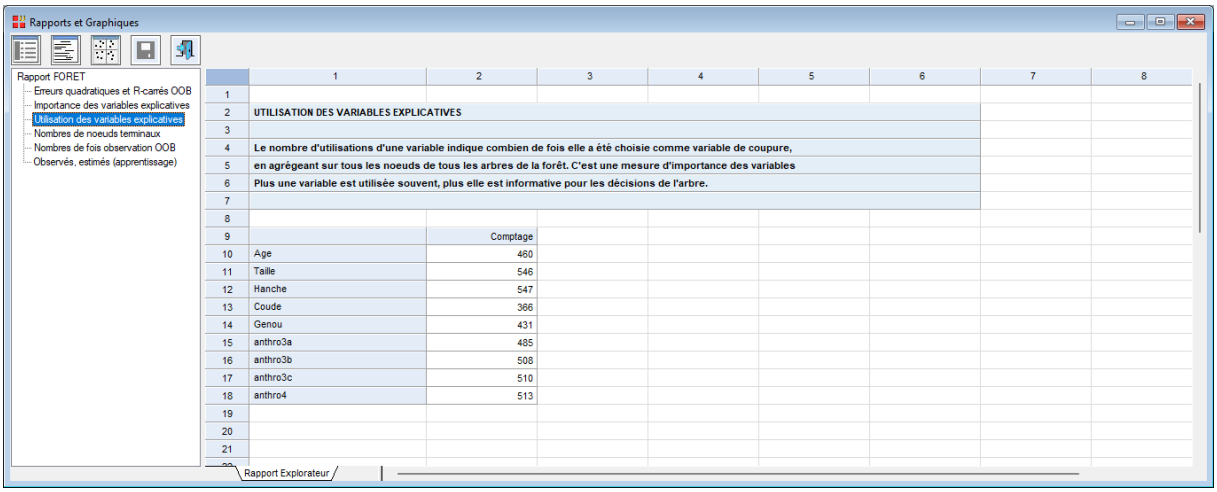


La première mesure est basée sur l’erreur quadratique moyenne. La seconde mesure est basée sur l‘impureté des nœuds (somme des carrés des résidus).

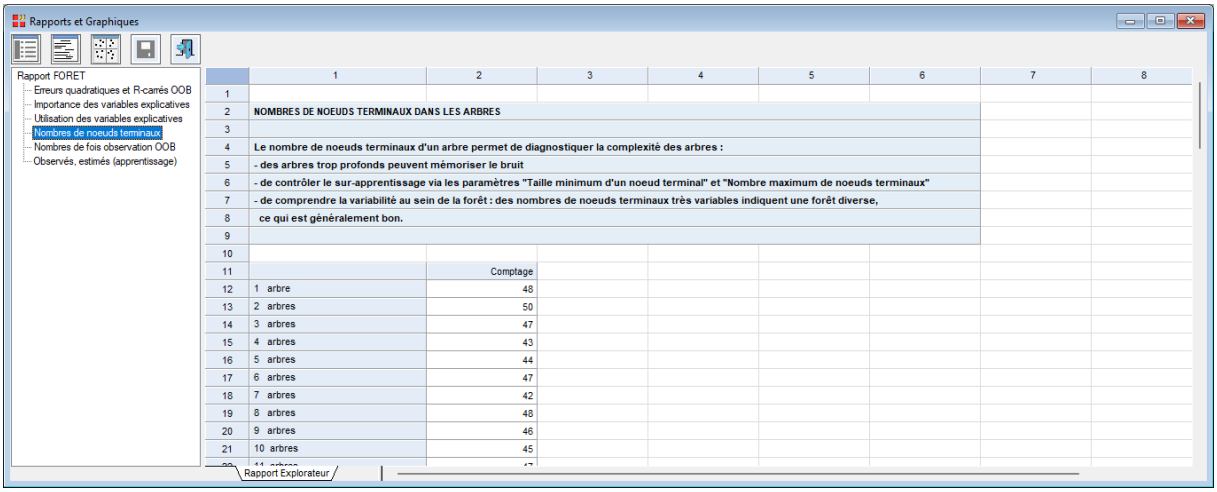
Plus l'importance d'une variable est grande, plus l'exclusion de cette variable impacte la qualité du modèle.

Utilisation des variables explicatives

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

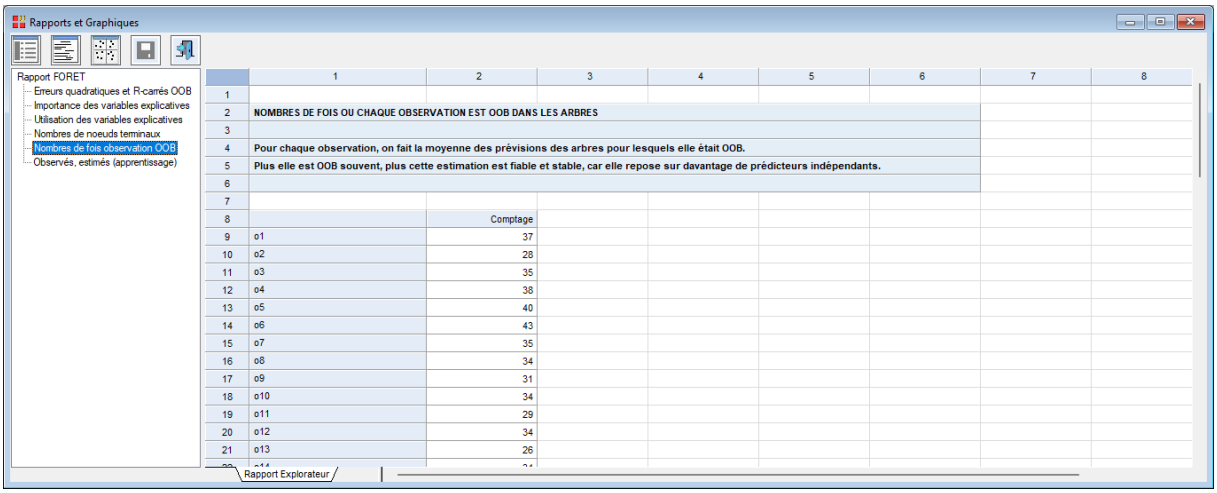


Nombres de nœuds terminaux dans les arbres



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

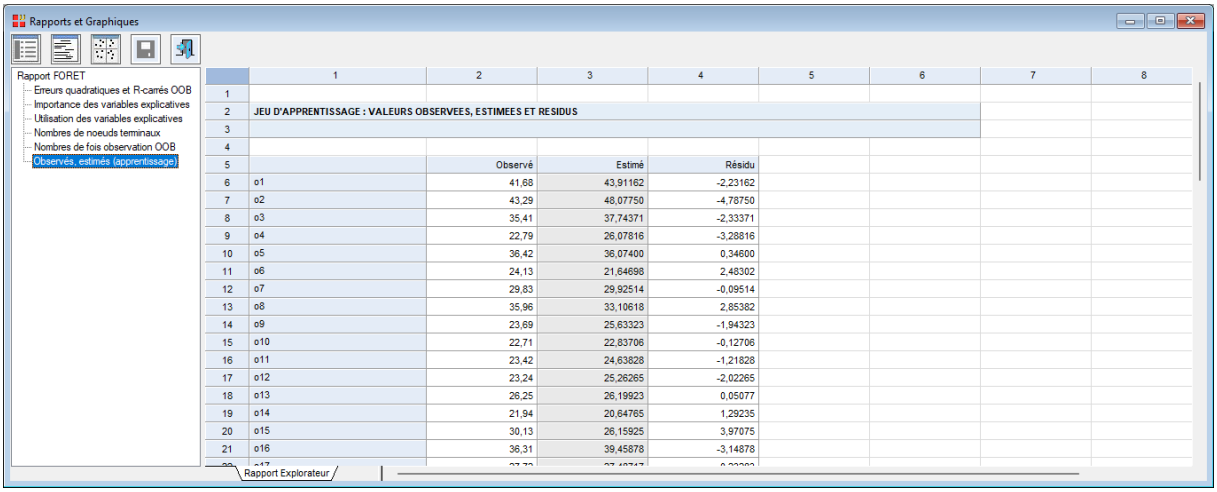
Nombre de fois où chaque observation est OOB dans les arbres



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.

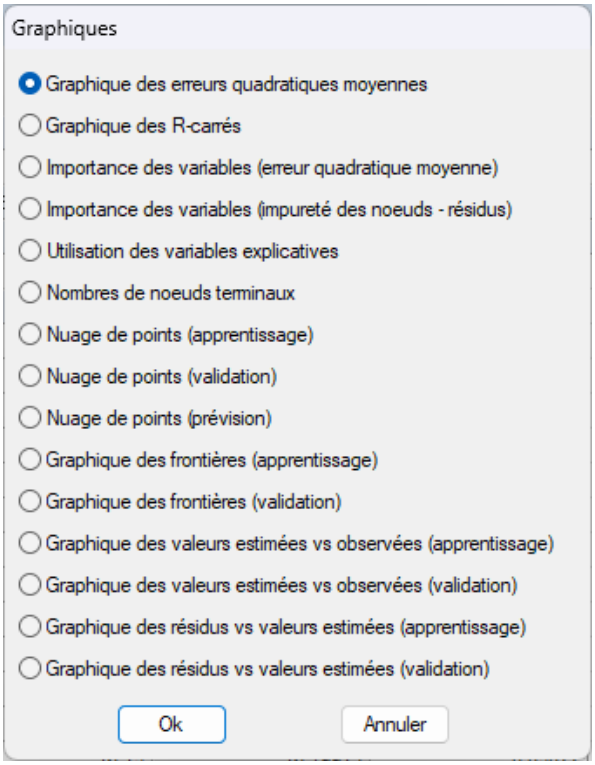
Valeurs observées, estimées et résidus

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

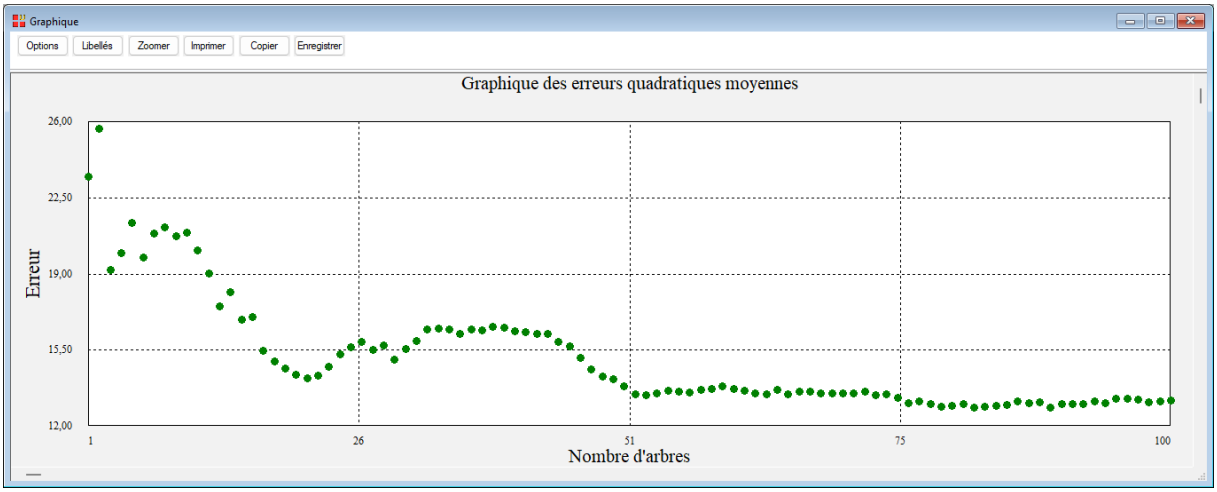


L'option Graphiques

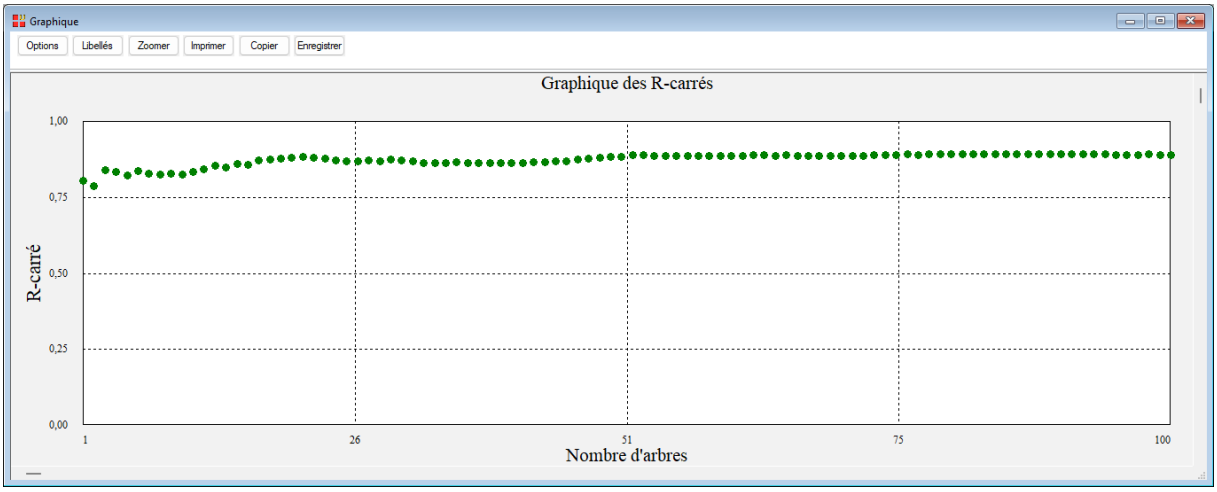
Cette option permet d'obtenir divers graphiques pour l'analyse.



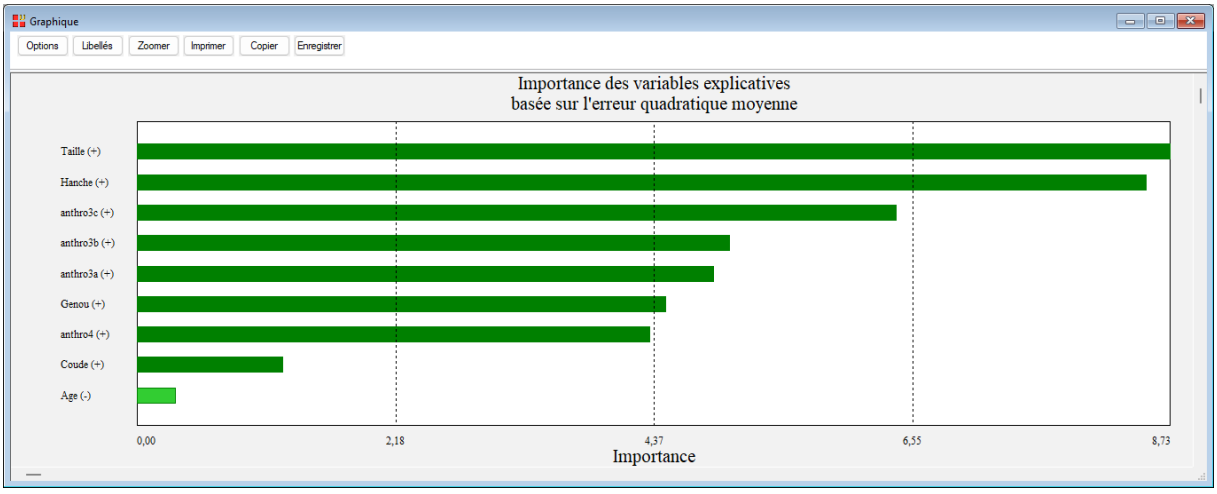
Graphique des erreurs quadratiques moyennes (OOB)



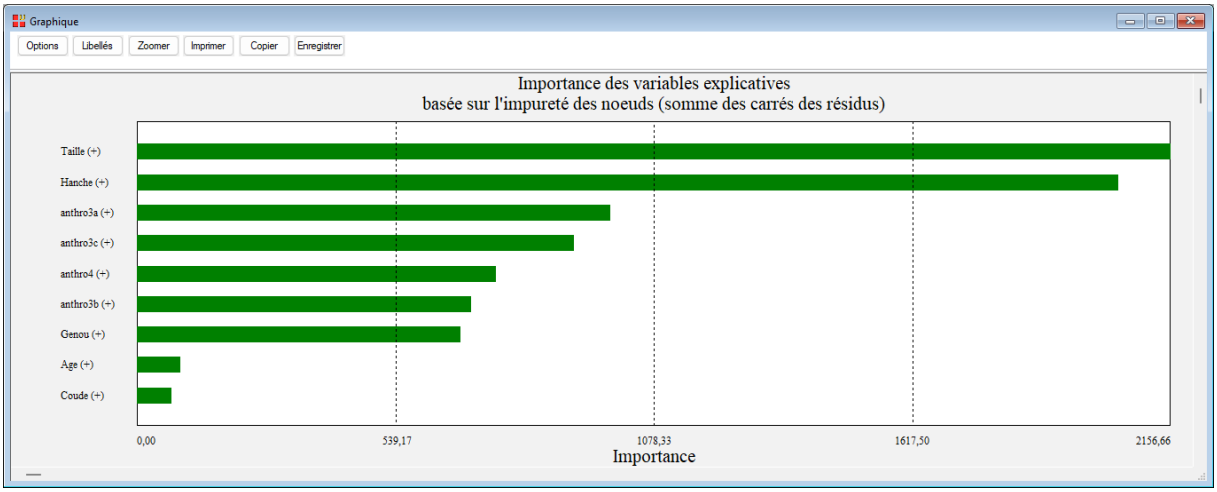
Graphique des R-carrés (OOB)



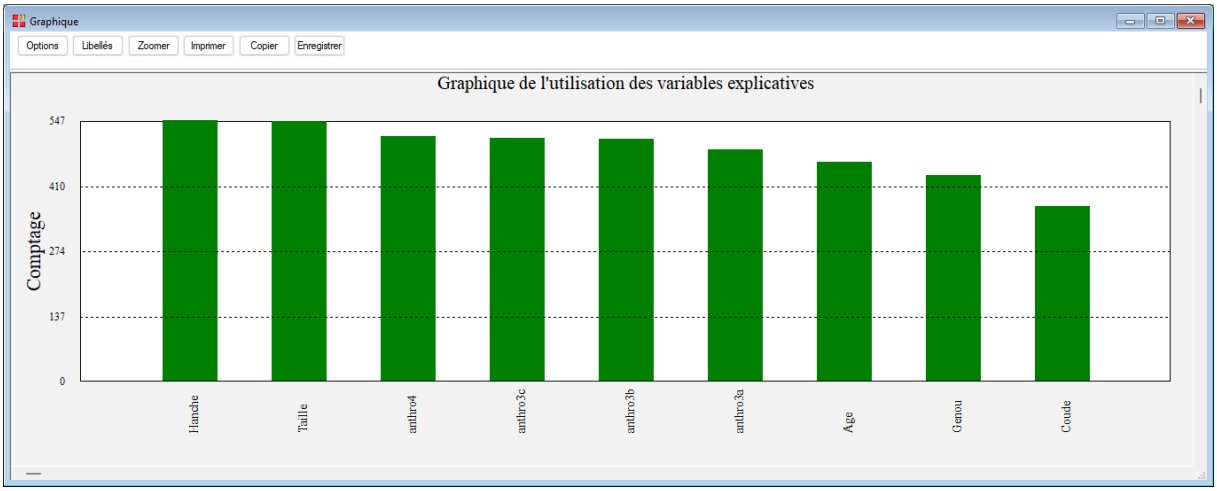
Importance des variables (erreur quadratique moyenne)



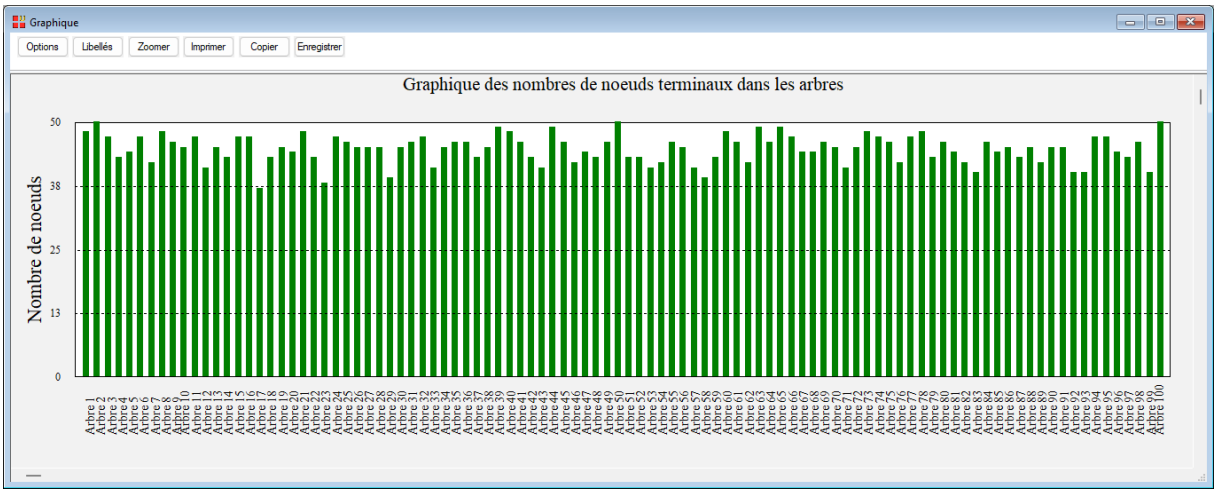
Importance des variables (résidus)



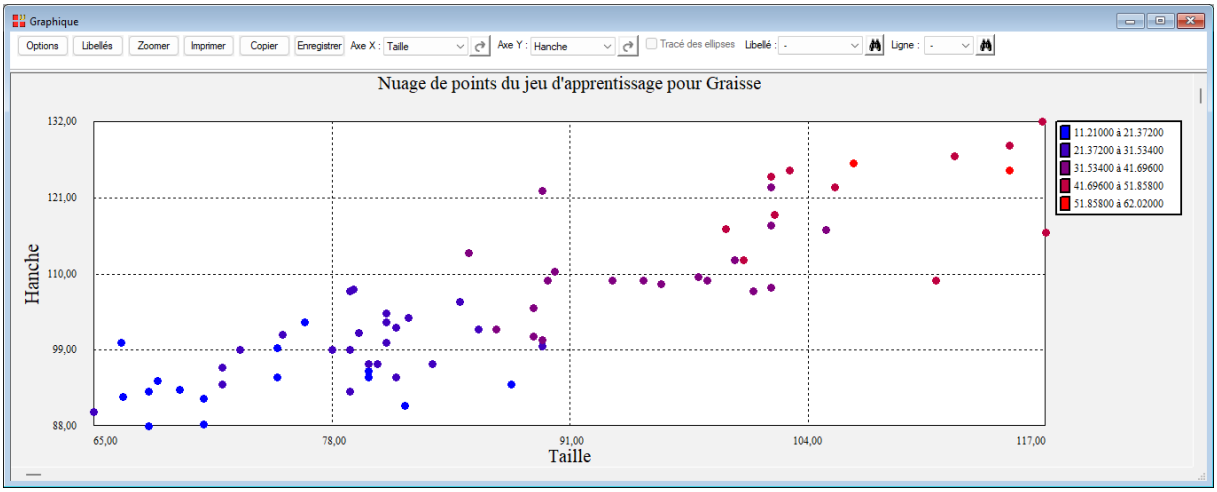
Utilisation des variables explicatives



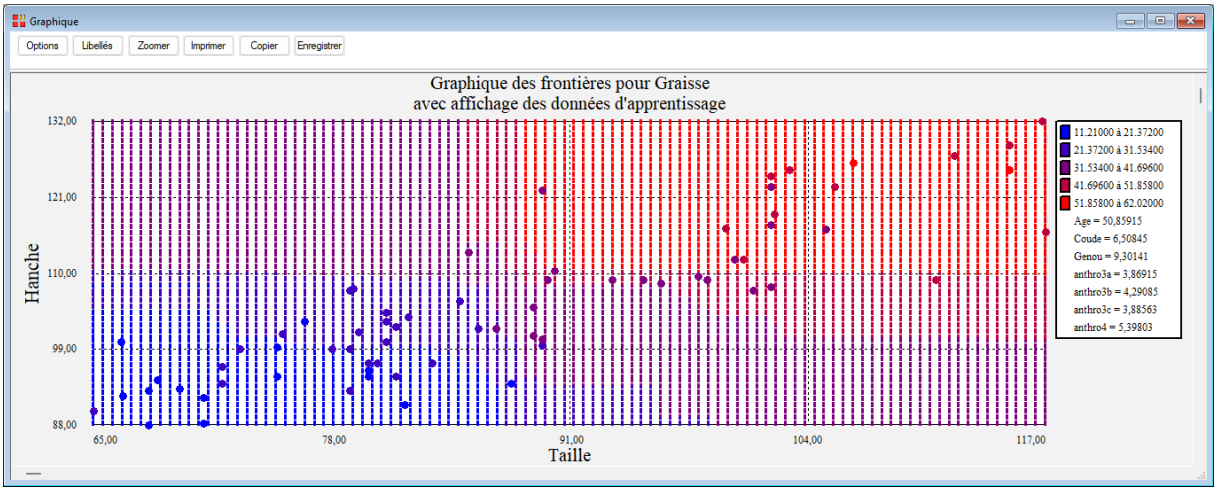
Nombres de nœuds terminaux dans les arbres



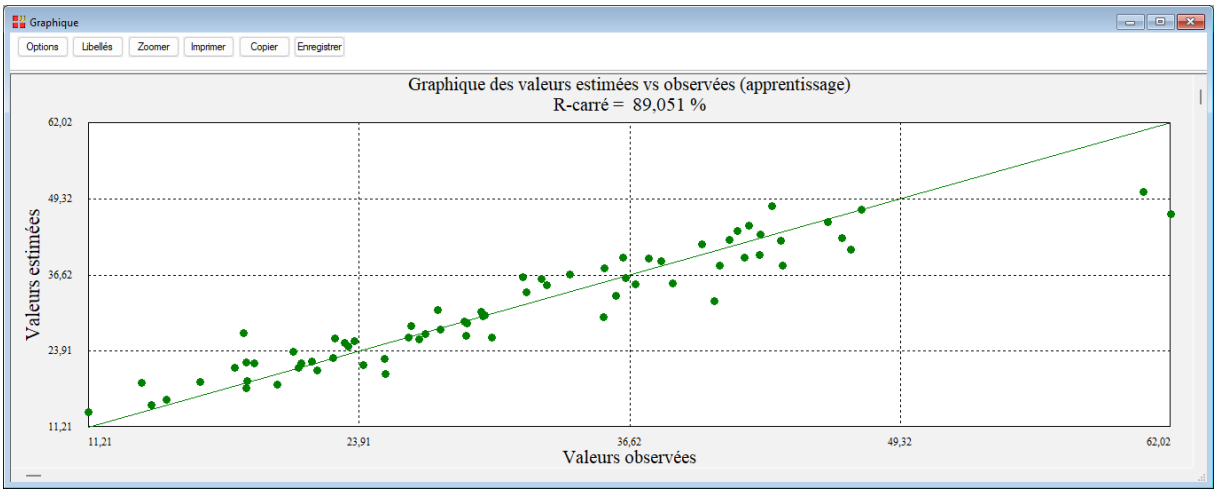
Nuage de points (apprentissage, validation, prévision)



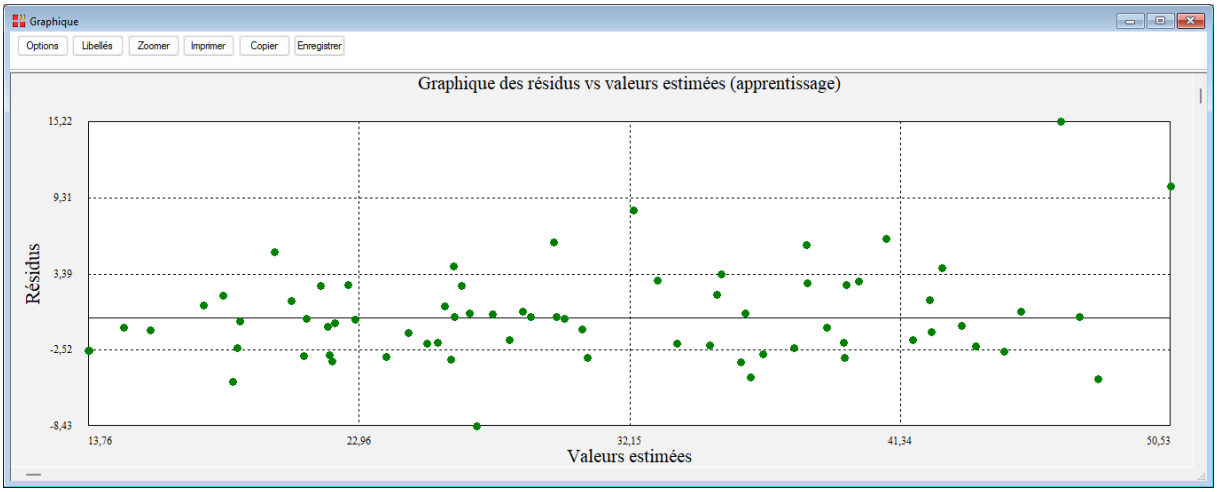
Graphique des frontières (apprentissage, validation)



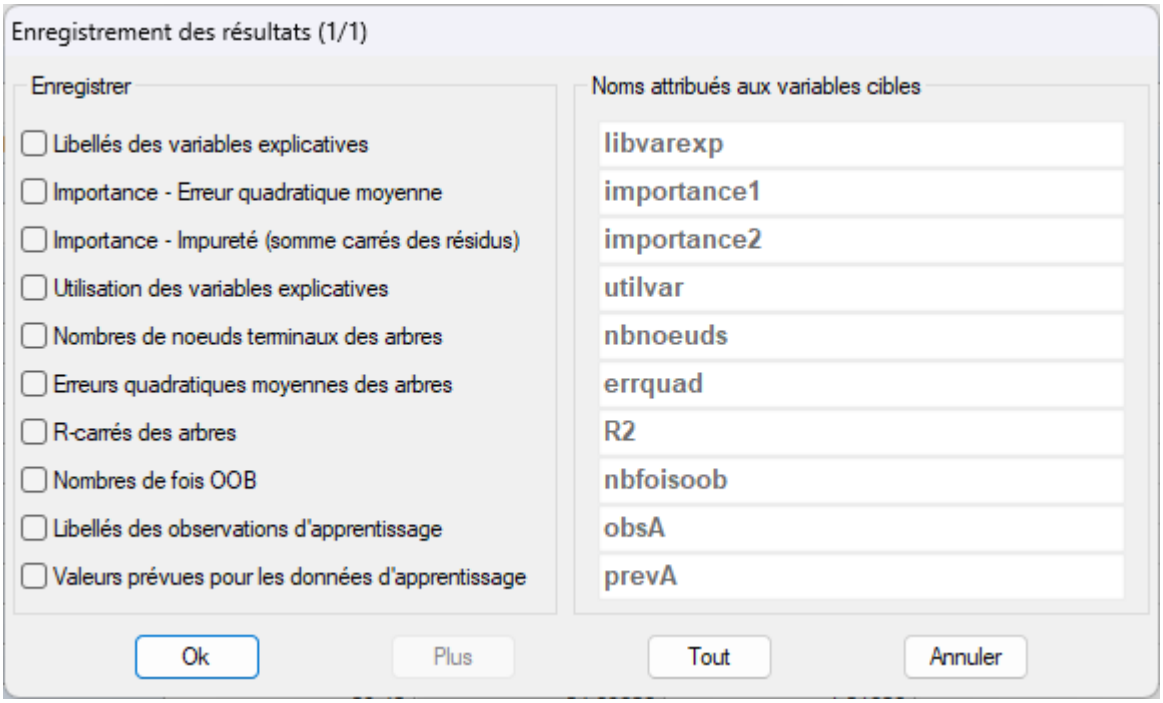
Graphique des valeurs estimées vs observées (apprentissage, validation)



Graphique des résidus vs valeurs observées (apprentissage, validation)



Les résultats suivants peuvent être enregistrés :



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’.

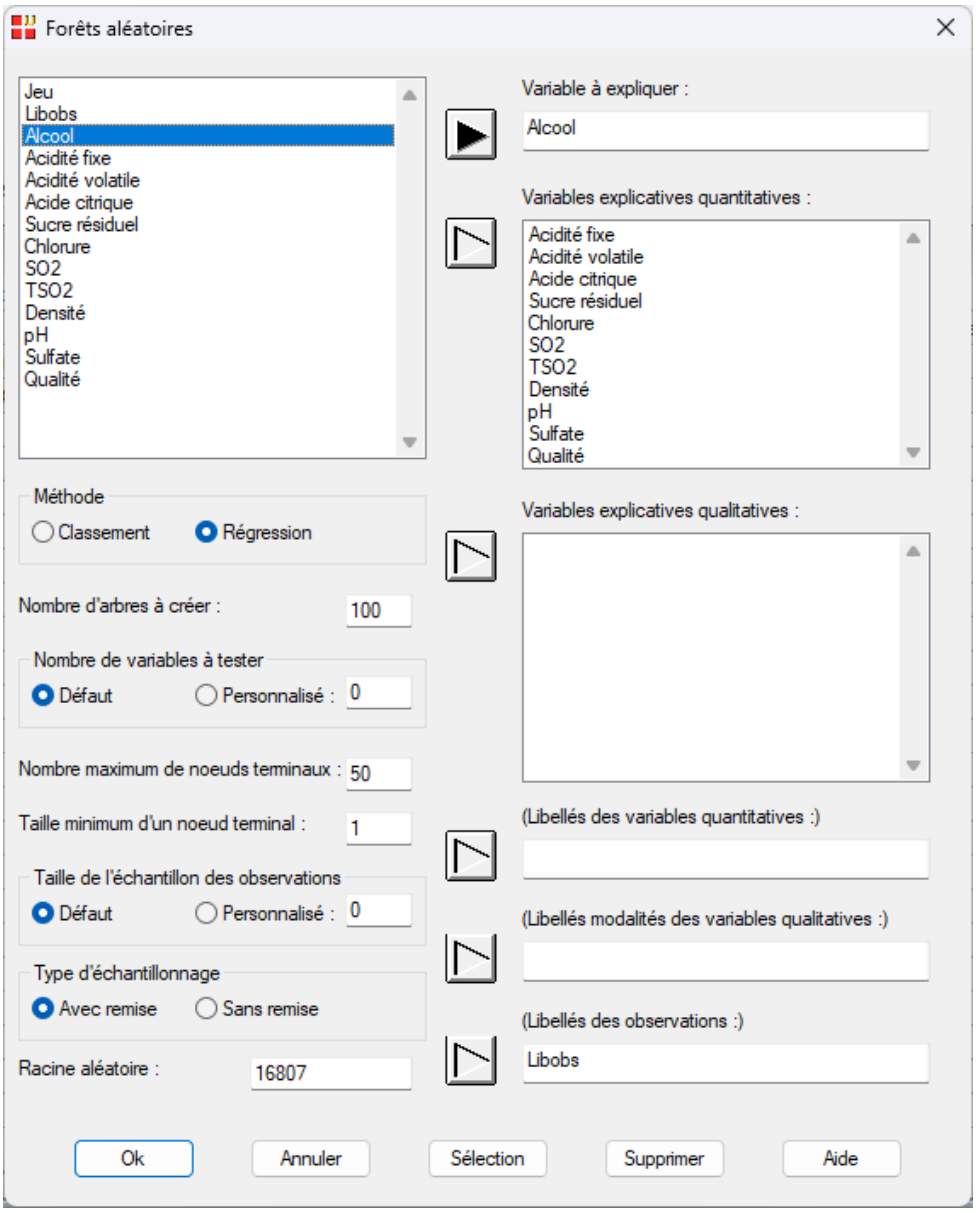
Les variables explicatives sélectionnées sont ‘Acidité fixe’ à ‘Qualité’.

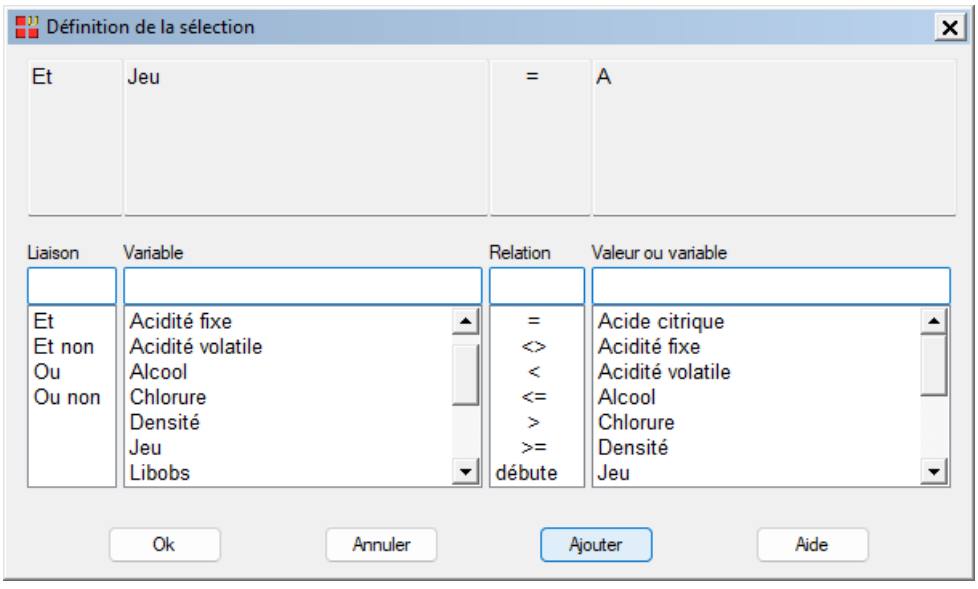
Renseignons la boîte de dialogue comme montré ci-dessous et cliquons sur Ok.

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

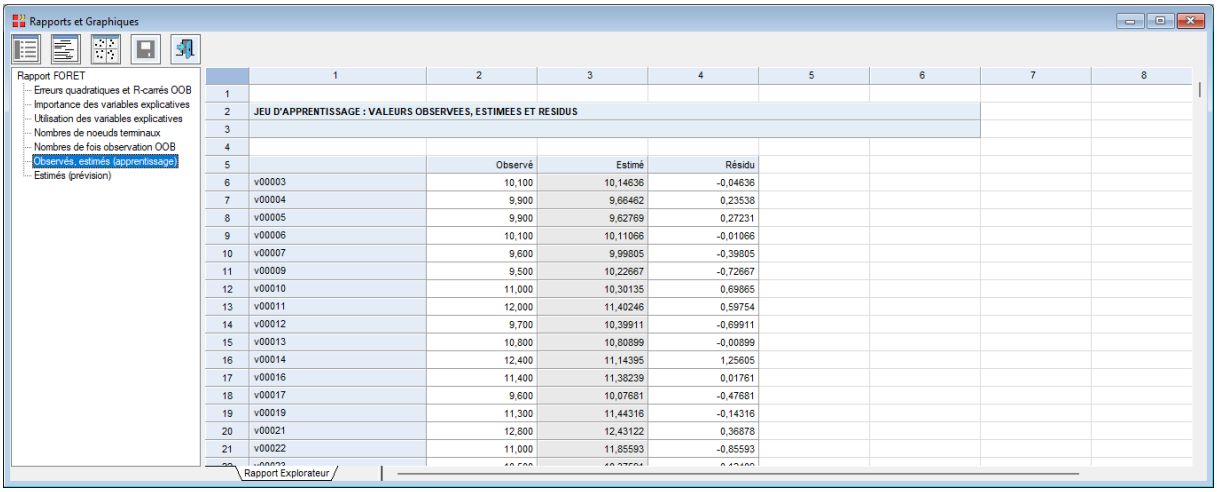
3428 observations seront ainsi utilisées comme jeu d’apprentissage et 245 comme jeu de prévision.

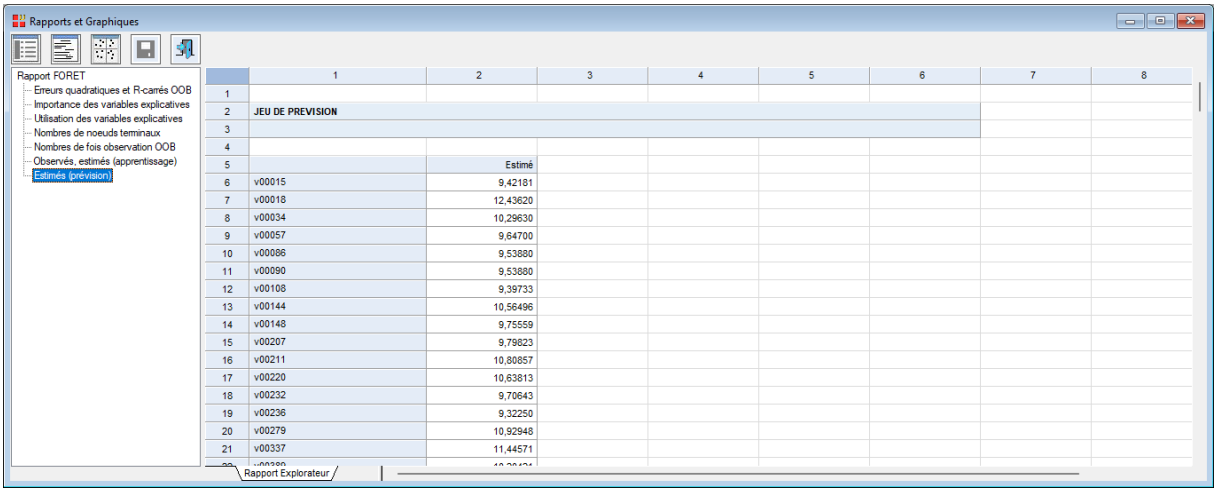
Après quelques instants, la fenêtre ‘Rapports et Graphiques’ montrée ci-après s’affiche.

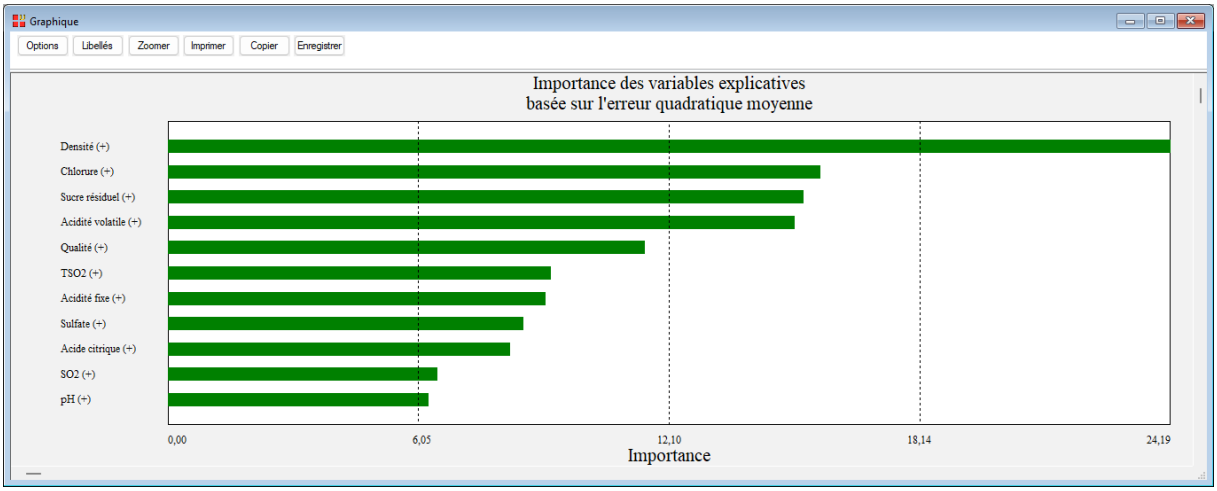

Forêts aléatoires
Jeu
Libobs
Alcool
Acidité fixe
Acidité volatile
Acide citrique
Sucre résiduel
Chlorure
SO2
TSO2
Densité
pH
Sulfate
Qualité
Variable à expliquer :
Alcool
Variables explicatives quantitatives :
Acidité fixe
Acidité volatile
Acide citrique
Sucre résiduel
Chlorure
SO2
TSO2
Densité
pH
Sulfate
Qualité
Méthode
Classement
Régression
Variables explicatives qualitatives :
Nombre d'arbres à créer :
100
Nombre de variables à tester
Défaut
Personnalisé :
0
Nombre maximum de noeuds terminaux :
50
Taille minimum d'un noeud terminal :
1
(Libellés des variables quantitatives :)
Taille de l'échantillon des observations
Défaut
Personnalisé :
0
(Libellés modalités des variables qualitatives :)
Type d'échantillonnage
Avec remise
Sans remise
(Libellés des observations :)
Racine aléatoire :
16807
Libobs
Ok
Annuler
Sélection
Supprimer
Aide

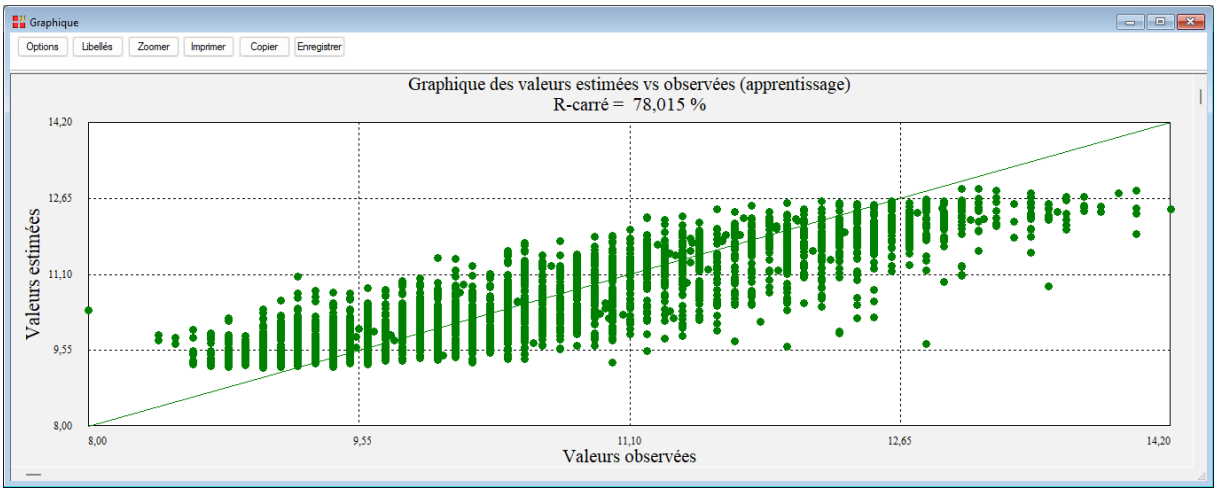

Définition de la sélection
Et
Jeu
=
A
Liaison
Variable
Relation
Valeur ou variable
Et
Et non
Ou
Ou non
Acidité fixe
Acidité volatile
Alcool
Chlorure
Densité
Jeu
Libobs
=
<>
<
<=
>
>=
débute
Acide citrique
Acidité fixe
Acidité volatile
Alcool
Chlorure
Densité
Jeu
Ok
Annuler
Ajouter
Aide

Voici quelques résultats obtenus par cette analyse.









Exemple 5 : Fichier TITANIC (décision)

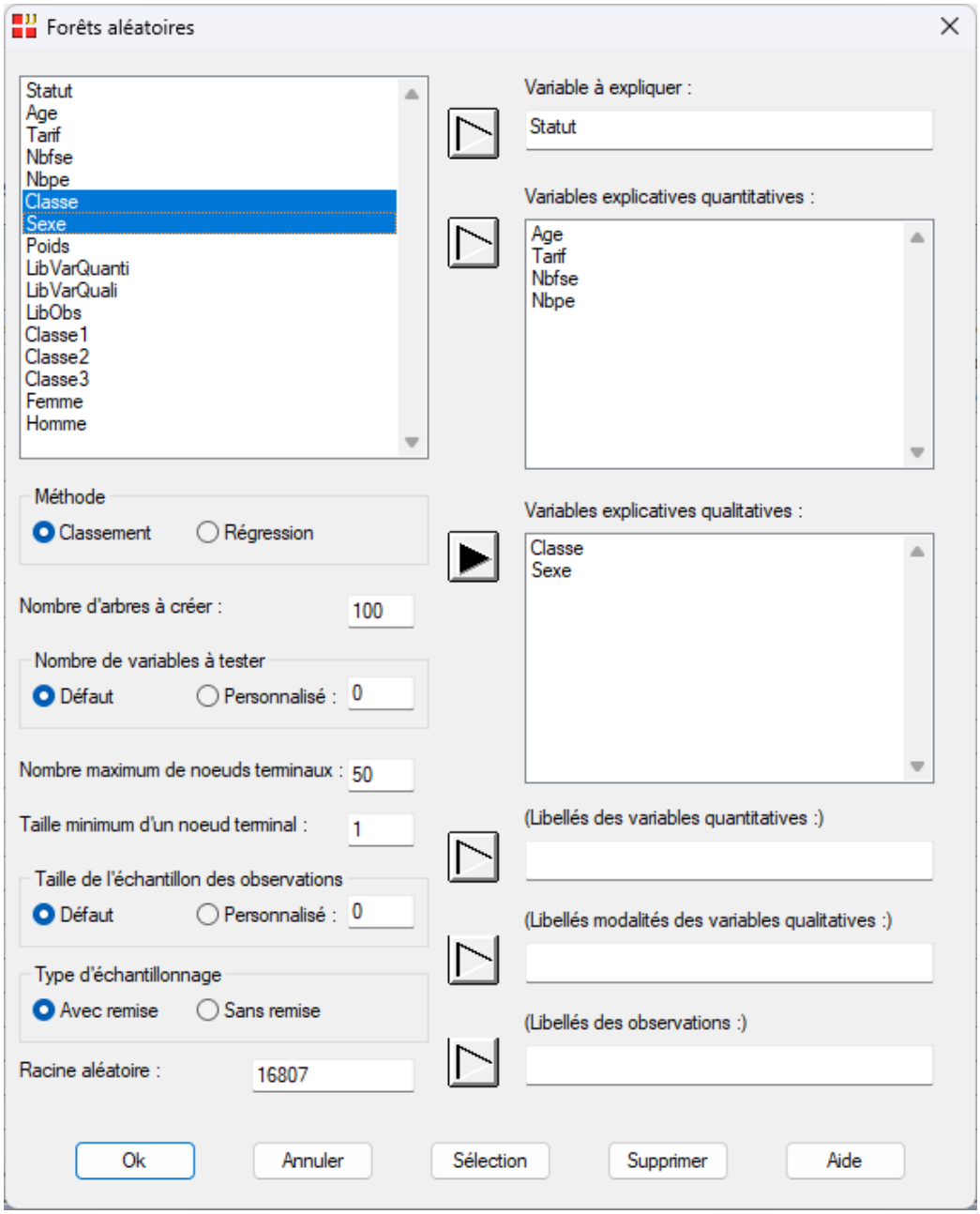
Pour ce quatrième exemple, nous utiliserons le fichier TITANIC pour construire un arbre de décision.

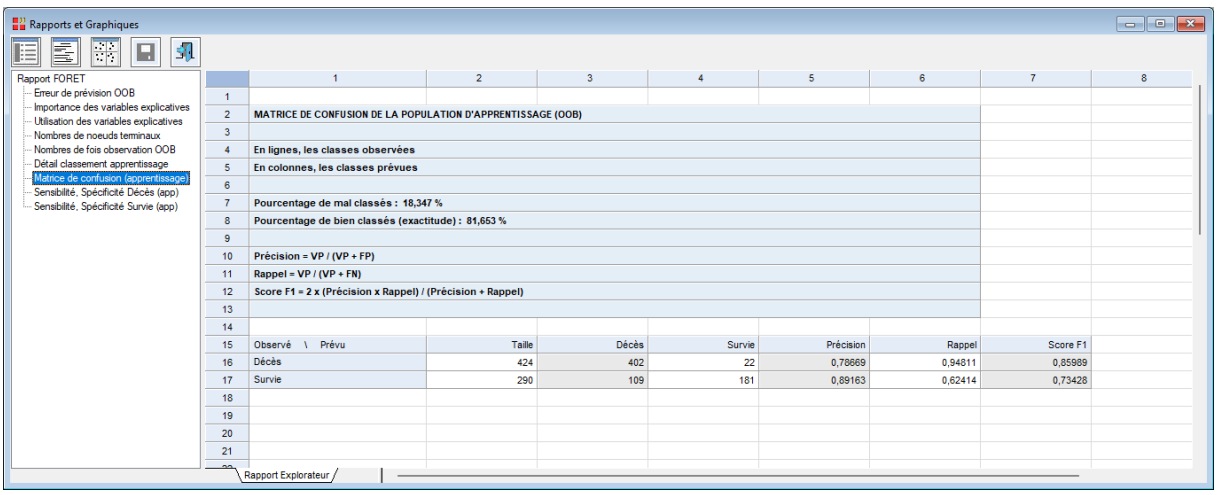
Ce fichier contient des informations concernant 714 passagers :

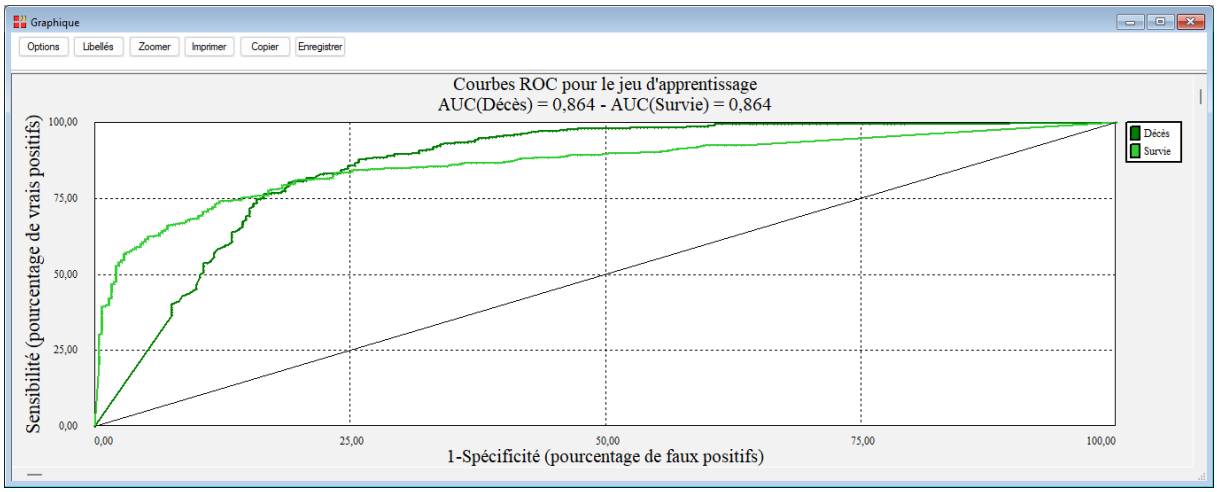
Statut	Survie ou Décès
Classe	Classe du passager (1ère, 2ème ou 3ème)
Sexe	Homme ou Femme
Age	Age du passager
Nbfse	Nombre de frères, sœurs ou époux, épouses à bord
Nbpe	Nombre de parents ou enfants à bord
Tarif	Tarif passager (en £)

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
Classe1
Classe2
Classe3
Femme
Homme
Variable à expliquer :
Statut
Variables explicatives quantitatives :
Age
Tarif
Nbfse
Nbpe
Méthode
Classement
Régression
Variables explicatives qualitatives :
Classe
Sexe
Nombre d'arbres à créer :
100
Nombre de variables à tester
Défaut
Personnalisé :
0
Nombre maximum de noeuds terminaux :
50
Taille minimum d'un noeud terminal :
1
(Libellés des variables quantitatives :)
Taille de l'échantillon des observations
Défaut
Personnalisé :
0
(Libellés modalités des variables qualitatives :)
Type d'échantillonnage
Avec remise
Sans remise
(Libellés des observations :)
Racine aléatoire :
16807
Ok
Annuler
Sélection
Supprimer
Aide


Rapports et Graphiques
Rapport FORET
Erreur de prévision OOB
Importance des variables explicatives
Utilisation des variables explicatives
Nombres de noeuds terminaux
Nombres de fois observation OOB
Détail classement apprentissage
Matrice de confusion (apprentissage)
Sensibilité, Spécificité Décès (app)
Sensibilité, Spécificité Survie (app)
MATRICE DE CONFUSION DE LA POPULATION D'APPRENTISSAGE (OOB)
En lignes, les classes observées
En colonnes, les classes prévues
Pourcentage de mal classés : 18,347 %
Pourcentage de bien classés (exactitude) : 81,653 %
Précision = VP / (VP + FP)
Rappel = VP / (VP + FN)
Score F1 = 2 x (Précision x Rappel) / (Précision + Rappel)
Observé \ Prévu
Taille
Décès
Survie
Précision
Rappel
Score F1
Décès 424 402 22 0,78669 0,94811 0,85989
Survie 290 109 181 0,89163 0,62414 0,73428
Rapport Explorateur



Environ 82 % des passagers sont bien classés par cette analyse et l'aire sous la courbe ROC (AUC) est proche de 0,86.

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	$\frac{100 * \mathrm{VP}}{(\mathrm{VP} + \mathrm{FN})}$
Observé B	FP	VN	FP + VN	$\frac{100 * \mathrm{VN}}{(\mathrm{VN} + \mathrm{FP})}$
Total	VP + FP	FN + VN	VP + FP + VN + FN	
% correct	$\frac{100 * \mathrm{VP}}{(\mathrm{VP} + \mathrm{FP})}$	$\frac{100 * \mathrm{VN}}{(\mathrm{FN} + \mathrm{VN})}$		$\frac{100 * (\mathrm{VP} + \mathrm{VN})}{(\mathrm{VP} + \mathrm{VN} + \mathrm{FP} + \mathrm{FN})}$
				% 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 x (précision x rappel) / (précision + 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.

Variable	*Contenu*
libvarexp	Libellés des variables explicatives
importance1	Importance des variables explicatives
importance2	Importance des variables explicatives
utilvar	Utilisation des variables explicatives
nbnoeuds	Nombres de nœuds terminaux
erreuroob	Erreur de prévision OOB (décision)
nbfoisoob	Nombres de fois OOB

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)
obsV	Libellés des observations de validation
voteV	Votes pour les données de validation (décision)
probV	Probabilités pour les données de validation (décision)
classeV	Classes prévues pour les données de validation (décision)
prevV	Valeurs prévues pour les données de validation (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)
seuilV	Seuils ROC (validation)
specificiteV	Spécificités (validation)
sensibiliteV	Sensibilités (validation)
aireV	Aires sous les courbes (validation)

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' (2026)

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