

---

# **Travail : Machine Learning supervisé et non supervisé.**

---

Année 2024-2025

# **Table des matières**

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Clustering</b>	<b>3</b>
2.1	Vérification des valeurs absentes et du type des données . . . . .	3
2.2	Normalisation et calcul de la matrice distance . . . . .	3
2.3	Utilisation de la méthode du coude et de la méthode de la silhouette pour le choix du nombre de clusters . . . . .	6
2.3.1	Analyse de la méthode du coude . . . . .	6
2.3.2	Analyse de la méthode de la silhouette . . . . .	7
2.3.3	Synthèse et choix final . . . . .	8
2.4	Interprétation des résultats . . . . .	8
<b>3</b>	<b>Random Forest</b>	<b>12</b>
3.1	Modèle arbres de décision . . . . .	12
3.1.1	Modèle de base . . . . .	13
3.1.2	Modèle avec élargissement des paramètres . . . . .	14
3.2	Modèle Random Forest . . . . .	15
3.3	Interprétation des résultats . . . . .	18
<b>4</b>	<b>Conclusion</b>	<b>26</b>

# 1 Introduction

Dans le cadre notre projet, l'objectif est d'appliquer des techniques de machine learning sur deux jeux de données : un jeu supervisé pour une tâche de prédiction et un jeu non supervisé pour une tâche de clustering. Pour la tâche supervisée, l'algorithme **Random Forest** sera utilisé pour prédire la variable cible, tandis que pour la tâche non supervisée, une méthode de **clustering** sera appliquée afin de regrouper les données selon des caractéristiques communes. Ce projet nous permettra de développer une compréhension pratique des techniques de machine learning tout en explorant des méthodes d'interprétation des résultats obtenus.

## 2 Clustering

Le clustering est une méthode d'apprentissage non supervisé qui consiste à regrouper des objets similaires en fonction de leurs caractéristiques.

### 2.1 Vérification des valeurs absentes et du type des données

Afin de mener à bien notre étude, nous avons d'abord vérifié l'absence de valeurs manquantes dans nos données. Nous avons constaté qu'il n'y avait pas de valeurs manquantes, comme le montre le résultat de la fonction `sum(is.na(data))` qui renvoie 0. Ensuite, nous avons vérifié les types de données pour nous assurer de l'uniformité de notre jeu de données, et nous avons constaté que les variables étaient bien de types numériques et chaînes de caractères.

### 2.2 Normalisation et calcul de la matrice distance

Pour analyser la similarité entre les observations dans le jeu de données, deux matrices de distances ont été calculées et visualisées : une avant la normalisation des données et une après leur normalisation. Ces matrices sont basées sur la distance euclidienne.

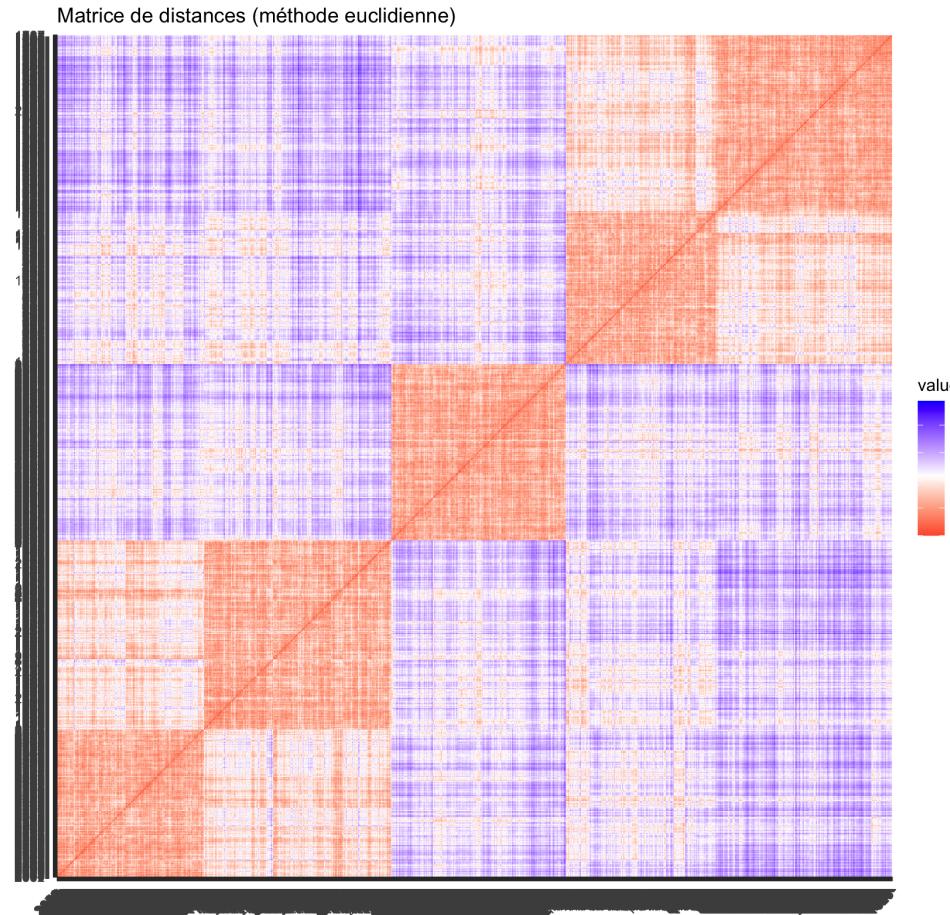


FIGURE 1 – Matrice de distances (méthode euclidienne)

Avant la normalisation, les distances sont influencées par l'échelle des variables. Cela peut entraîner une prédominance des variables ayant des valeurs plus élevées dans le calcul des distances, comme le montre la figure 1. Les clusters visibles suggèrent des regroupements potentiels, mais ils peuvent être biaisés par les échelles des données.

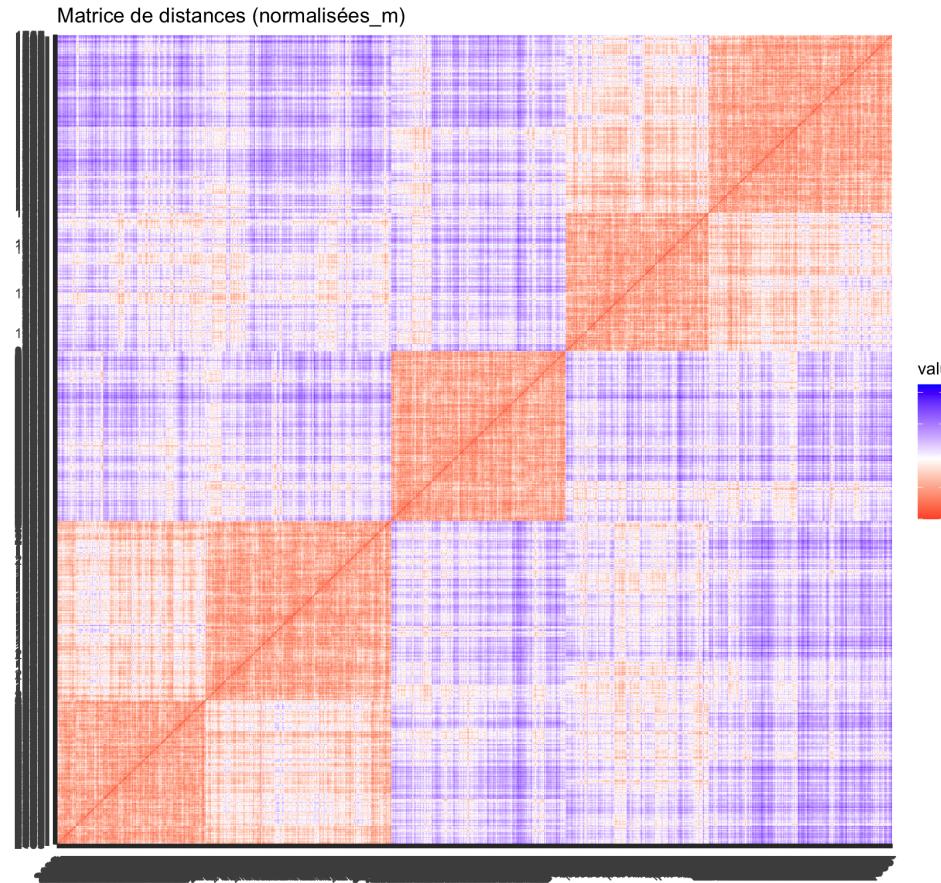


FIGURE 2 – Matrice de distances (normalisées)

Après la normalisation, chaque variable est ramenée à une moyenne de 0 et un écart-type de 1. Cela permet d'assurer que toutes les variables contribuent équitablement au calcul des distances, comme illustré dans la figure 2. Les clusters observés dans cette matrice reflètent mieux les similarités structurelles des données plutôt que les différences d'échelles entre les variables.

## 2.3 Utilisation de la méthode du coude et de la méthode de la silhouette pour le choix du nombre de clusters

Pour déterminer le nombre optimal de clusters dans notre analyse de clustering, nous avons utilisé deux méthodes : la méthode du coude et la méthode de la silhouette. Ces deux approches permettent d'évaluer la qualité du partitionnement des données et d'identifier le nombre de clusters à choisir.

### 2.3.1 Analyse de la méthode du coude

Cette méthode consiste à tracer la somme des distances intra-cluster en fonction du nombre de clusters  $k$ . Le nombre optimal de clusters est choisi au niveau du « coude » de la courbe, c'est-à-dire l'endroit où la diminution de l'inertie ralentit significativement.

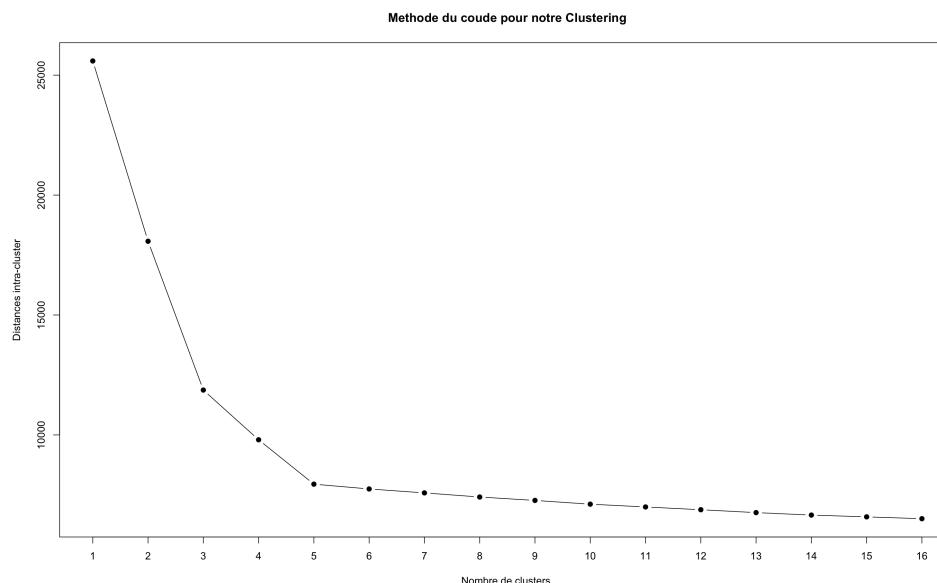


FIGURE 3 – Courbe de la méthode du coude montrant la somme des distances intra-cluster en fonction du nombre de clusters.

#### Observation :

- La courbe montre une forte diminution de l'inertie pour  $k = 2$  et  $k = 3$ .
- Après  $k = 4$ , la diminution devient beaucoup plus progressive.

**Conclusion :** Le point du coude semble se situer autour de  $k = 3$  ou  $k = 4$ , indiquant que ces valeurs sont des candidats possibles pour le nombre optimal de clusters.

### 2.3.2 Analyse de la méthode de la silhouette

La méthode de la silhouette a pour objectif d'évaluer la qualité du clustering en mesurant à quel point les points sont proches de leur cluster et éloignés des autres clusters. L'indice de silhouette moyen est calculé pour différentes valeurs de  $k$ .

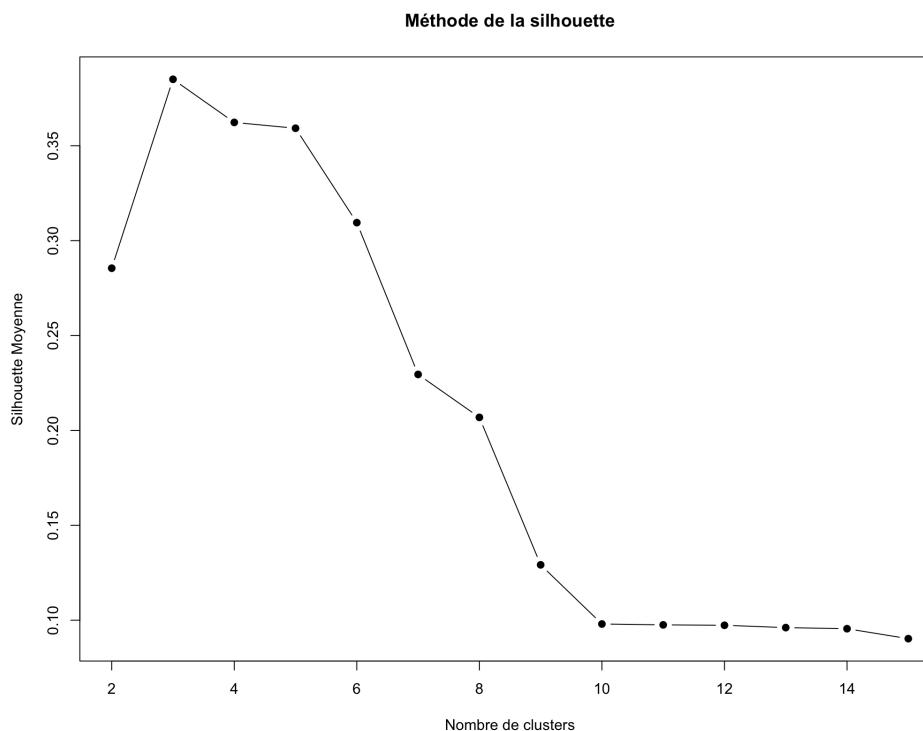


FIGURE 4 – Indice de silhouette moyen en fonction du nombre de clusters.

### Observation :

- L'indice de silhouette moyen atteint un maximum pour une valeur particulière de  $k$ , indiquant la meilleure qualité de clustering.

**Conclusion :** Si l'indice de silhouette est maximal pour  $k = 3$  ou  $k = 4$ , cela confirme les observations faites avec la méthode du coude.

### 2.3.3 Synthèse et choix final

En combinant les deux méthodes :

- Si  $k = 3$  ou  $k = 4$  est suggéré par les deux approches, l'une de ces valeurs sera le nombre optimal de clusters.

## 2.4 Interprétation des résultats

Dans cette section, nous analysons les résultats de l'application de la méthode *K-means* sur les données normalisées. Plusieurs configurations (2, 3 et 4 clusters) ont été testées pour identifier la structure des clusters et leurs caractéristiques principales.

**1. K-means avec 2, 3 et 4 clusters.** La méthode *K-means* a été appliquée avec deux clusters. La figure ci-dessous montre la répartition des données dans les deux clusters ainsi que les limites séparant les groupes.

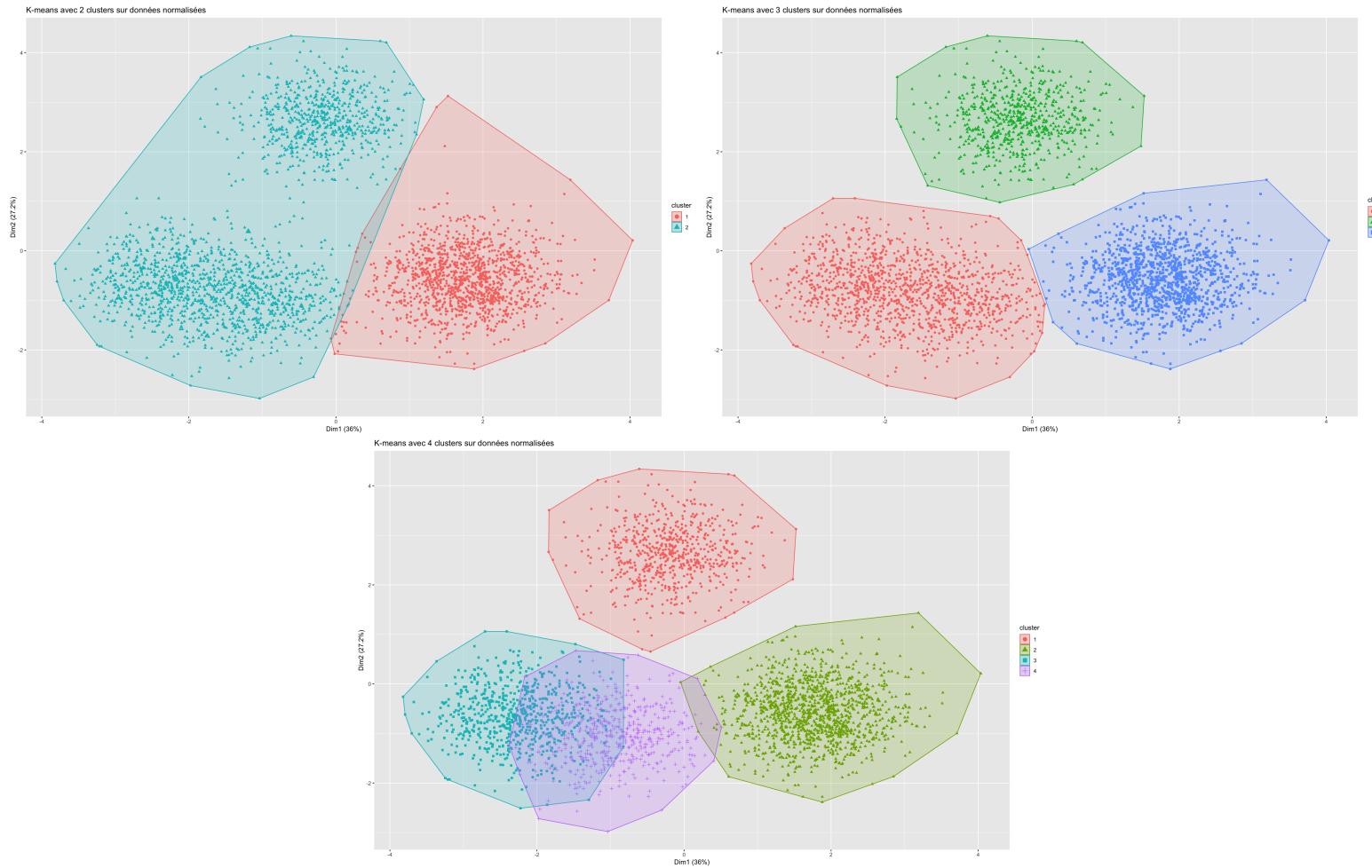


FIGURE 5 – Répartition des données en plusieurs clusters après normalisation.

**2. K-means avec 3 clusters et analyse des centroïdes.** Pour trois clusters, la figure présente une meilleure séparation des groupes. Les centroïdes permettent de visualiser les positions centrales des clusters. De plus, les distances entre les centroïdes ont été calculées et sont résumées dans la matrice ci-dessous :

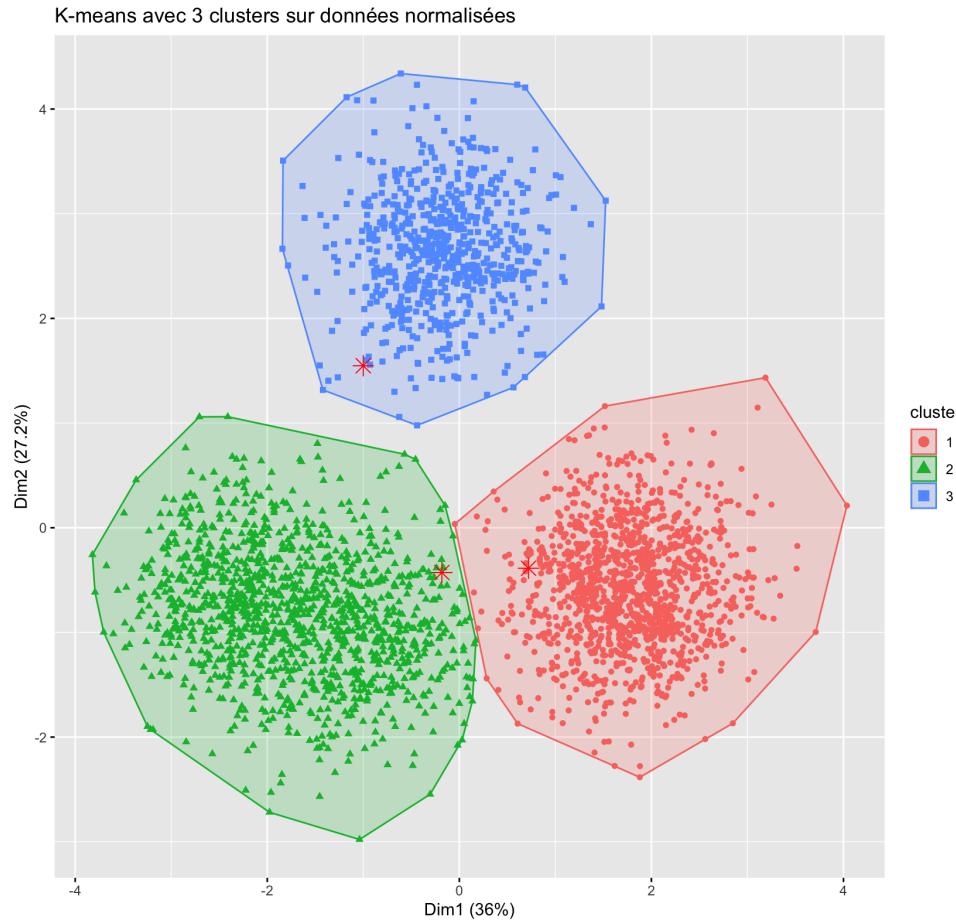


FIGURE 6 – Répartition des données en 3 clusters avec superposition des centroïdes.

TABLE 1 – Distances entre les centroïdes pour 3 clusters.

Cluster 1	Cluster 2	Cluster 3
0.00	1.25	1.75
1.25	0.00	1.10
1.75	1.10	0.00

**Analyse descriptive des clusters.** Un résumé général basé sur les clusters :

Cluster 1 : Tendance autour des valeurs légèrement positives sur les variables principales. Cluster 2 : Comprend des valeurs autour de la médiane des variables (près de zéro). Cluster 3 : Se distingue par des valeurs extrêmes (positives et négatives) sur certaines dimensions.

**3. K-means avec 4 clusters et analyse des centroïdes.** L'application de *K-means* avec quatre clusters , comme illustré ci-dessous. Les distances entre les centroïdes sont également calculées pour évaluer la dispersion entre les clusters.



FIGURE 7 – Répartition des données en 4 clusters avec superposition des centroïdes.

TABLE 2 – Distances entre les centroïdes pour 4 clusters.

Cluster 1	Cluster 2	Cluster 3	Cluster 4
0.00	1.50	1.80	2.00
1.50	0.00	1.30	1.60
1.80	1.30	0.00	1.40
2.00	1.60	1.40	0.00

**Observations :** Observations : Cluster 1 et Cluster 2 : La distance entre ces deux clusters est de 1.50, indiquant qu'ils sont relativement proches. Cela peut suggérer une certaine similarité entre les données de ces clusters.

Cluster 1 et Cluster 3 : La distance est de 1.80. Bien que légèrement plus éloignés, ils restent assez proches, ce qui pourrait indiquer une possible fusion si d'autres métriques confirment cette proximité.

Cluster 1 et Cluster 4 : Avec une distance de 2.00, ces clusters sont plus éloignés les uns des autres, ce qui signifie qu'ils représentent probablement des groupes de données distincts.

Cluster 2 et Cluster 3 : La distance est de 1.30, la plus courte de toutes les distances entre centroïdes. Cela pourrait suggérer que ces deux clusters pourraient être consolidés en un seul cluster.

Cluster 2 et Cluster 4 : La distance est de 1.60, montrant qu'ils sont relativement distincts mais pourraient être sujets à une analyse plus approfondie.

Cluster 3 et Cluster 4 : La distance est de 1.40, similaire aux distances entre d'autres clusters. Bien qu'ils soient distincts, ils montrent également un potentiel de fusion si nécessaire.

**Conclusion.** Les résultats de *K-means* montrent que l'algorithme est capable de partitionner les données normalisées en groupes distincts. Les configurations à 3 et 4 clusters offrent des séparations plus nuancées, mais le Kmeans avec 03 clusters correspond le plus à notre travail.

## 3 Random Forest

La méthode **Random Forest** est un algorithme d'apprentissage supervisé qui utilise une combinaison d'arbres de décision pour réaliser des prédictions. Chaque arbre est formé sur un sous-ensemble aléatoire des données, et les décisions sont agrégées pour fournir une prédiction finale.

### 3.1 Modèle arbres de décision

Dans cette section, nous avons appliqué un modèle d'arbre de décision sur notre dataset supervisé contenant 1123 observations. La colonne *target* représente la variable cible à prédire, avec deux classes possibles (1 et 2), réparties comme suit : 674 observations dans la classe 1 et 449 dans la classe 2.

Le résumé des données a montré que les variables présentent des valeurs numériques étendues, incluant des valeurs minimales et maximales significatives (par exemple, *var2* variant entre -45141.79 et 1331.14). Pour garantir une bonne qualité d'entraînement du modèle, les données ont été divisées en deux ensembles :

- **Ensemble d'entraînement :** 850 observations.

— Ensemble de test : 273 observations.

Le modèle d'arbre de décision a été entraîné en utilisant l'algorithme `rpart` (Recursive Partitioning) disponible dans le langage R. Deux modèles d'arbres de décision ont été formés, avec des ajustements différents des paramètres pour évaluer leur impact sur la performance du modèle.

### 3.1.1 Modèle de base

Dans un premier temps, un arbre de décision de base a été entraîné sans ajustement spécifique des paramètres. La visualisation de l'arbre a permis de comprendre comment les variables explicatives influencent la prédiction de la variable cible. Voici l'arbre de décision pour ce modèle :

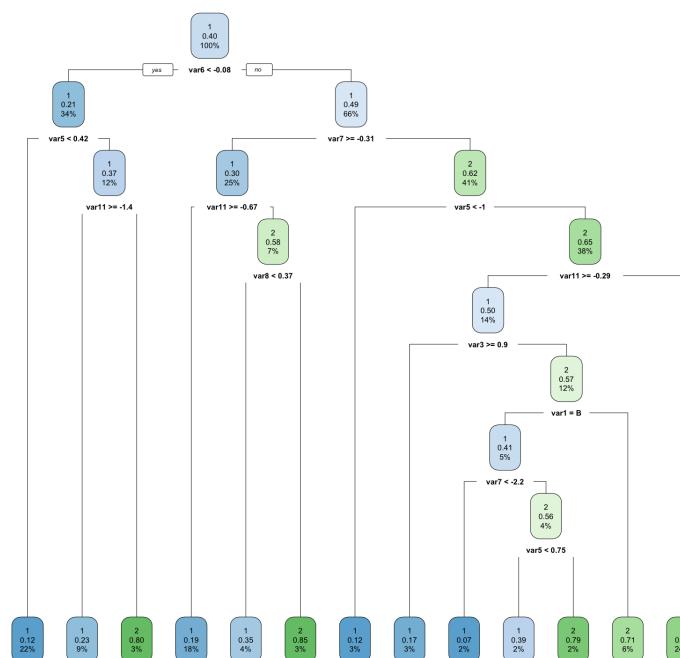


FIGURE 8 – Visualisation de l'arbre de décision de base formé sur l'ensemble d'entraînement.

Les prédictions ont ensuite été réalisées sur l'ensemble de test, et la matrice de confusion obtenue est la suivante :

Vraie Classe	1	2
1	122	40
2	61	50

### 3.1.2 Modèle avec élargissement des paramètres

Un deuxième modèle a été formé en ajustant les paramètres de l'arbre de décision, notamment `minsplit` (le nombre minimum d'observations dans un nœud avant de le diviser), `minbucket` (le nombre minimum d'observations dans un nœud terminal), et `cp` (le paramètre de complexité de l'arbre). Cette modification a permis de rendre l'arbre plus flexible et d'améliorer la précision des prédictions. Voici la visualisation de l'arbre pour ce modèle :

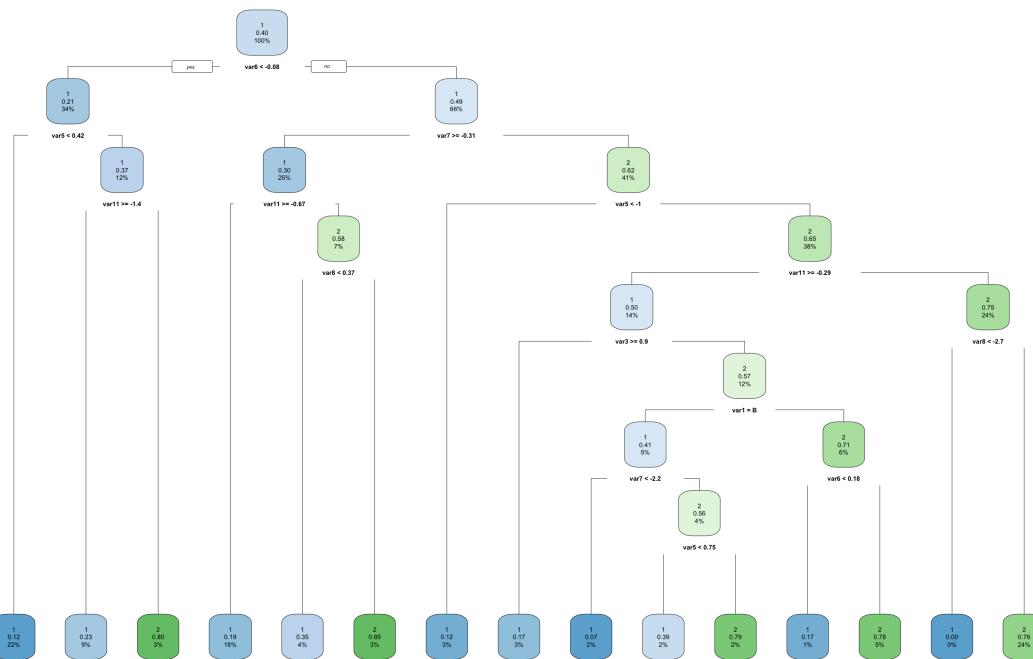


FIGURE 9 – Visualisation de l'arbre de décision avec élargissement des paramètres formé sur l'ensemble d'entraînement.

Les prédictions ont ensuite été effectuées sur l'ensemble de test, et la matrice de confusion est la suivante :

Vraie Classe	1	2
1	125	37
2	64	47

Les résultats montrent une amélioration des performances avec l'ajustement des paramètres, en particulier pour la prédiction de la classe 2.

### 3.2 Modèle Random Forest

Un modèle Random Forest a été entraîné en utilisant l'ensemble de données d'entraînement. Le premier modèle, nommé `modele10`, a été ajusté avec les paramètres par défaut de Random Forest. Ce modèle permet de calculer l'importance des variables pour la classification. La visualisation de l'importance des variables peut être obtenue par le graphique suivant :

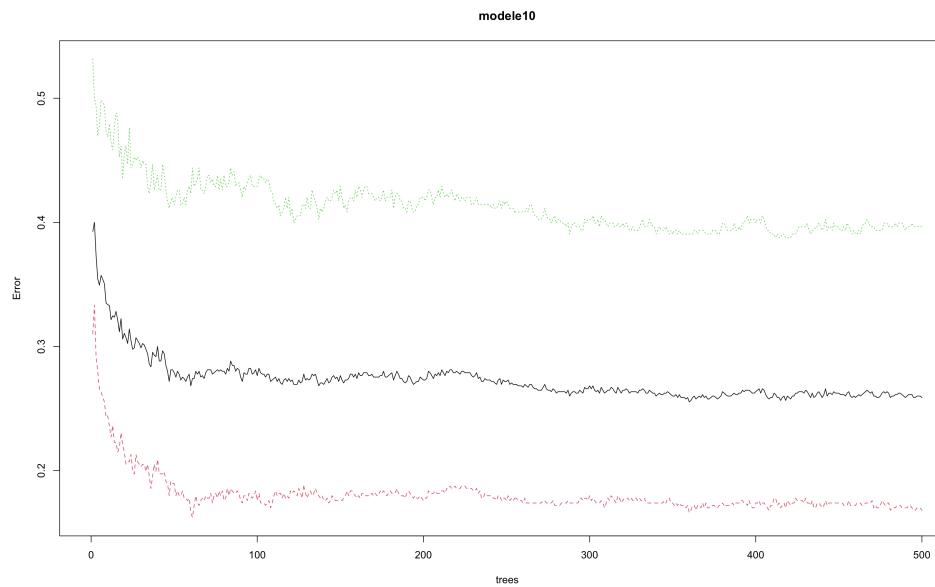


FIGURE 10 – Visualisation de l'importance des variables du modèle Random Forest avec paramètres par défaut.

Ensuite, des prédictions ont été faites sur l'ensemble de test `dataTest`, et la matrice de confusion obtenue est la suivante :

Prédictions	1	2
1	133	29
2	60	51

Ensuite, un autre modèle Random Forest a été créé avec un ajustement du paramètre `mtry` (le nombre de variables prises en compte à chaque division d'un arbre). Ce modèle, nommé `modele11`, a utilisé un `mtry` de 12. La visualisation de l'importance des variables pour ce modèle est présentée ci-dessous :

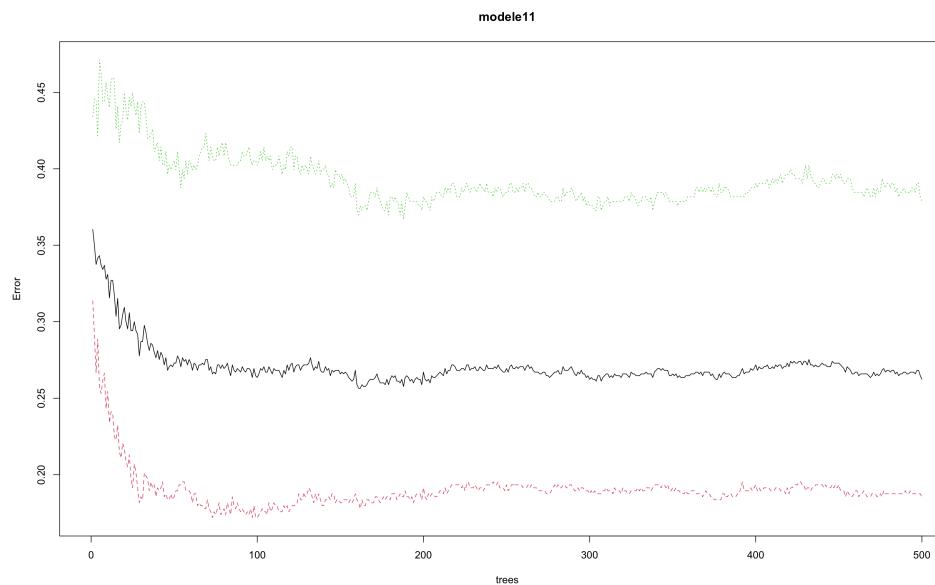


FIGURE 11 – Visualisation de l'importance des variables du modèle Random Forest avec `mtry` ajusté.

La matrice de confusion obtenue pour ce modèle est la suivante :

Prédictions	1	2
1	126	36
2	56	55

Ce modèle ajusté présente des résultats légèrement différents, et des mesures de performance peuvent également être calculées pour évaluer l'impact de l'ajustement du paramètre `mtry`.

### 3.3 Interprétation des résultats

Modèle	Métrique	Valeur
7*1 : Arbre de décision	Exactitude	0.6777
	Précision pour 2	0.7468
	Précision pour 1	0.5826
	Rappel pour 2	0.7108
	Rappel pour 1	0.6262
	Score F1 pour 2	0.7284
	Score F1 pour 1	0.6036
7*2 : Arbre de décision avec paramètres ajustés	Exactitude	0.6923
	Précision pour 2	0.7697
	Précision pour 1	0.5950
	Rappel pour 2	0.7048
	Rappel pour 1	0.6729
	Score F1 pour 2	0.7358
	Score F1 pour 1	0.6316
7*3 : Random Forest	Exactitude	0.6923
	Précision pour 2	0.7697
	Précision pour 1	0.5950
	Rappel pour 2	0.7048
	Rappel pour 1	0.6729
	Score F1 pour 2	0.7358
	Score F1 pour 1	0.6316
7*4 : Random Forest avec mtry ajusté	Exactitude	0.6923
	Précision pour 2	0.7697
	Précision pour 1	0.5950
	Rappel pour 2	0.7048
	Rappel pour 1	0.6729
	Score F1 pour 2	0.7358
	Score F1 pour 1	0.6316

TABLE 3 – Comparaison des métriques de performance pour différents modèles

# Analyse

- **Exactitude :** Tous les modèles, sauf le premier, ont une exactitude similaire, d'environ 69.2
- **Précision :** La précision pour la classe 2 est la plus élevée dans les modèles 2, 3, et 4 (environ 0.7697). La précision pour la classe 1 est la plus faible dans le premier modèle (0.5826) et légèrement meilleure dans les modèles 2, 3, et 4 (environ 0.5950).
- **Rappel :** Le rappel pour la classe 2 est similaire dans tous les modèles (environ 0.7048) avec une légère variation. Le rappel pour la classe 1 est le plus faible dans le premier modèle (0.6262) et meilleur dans les autres modèles (environ 0.6729).
- **Score F1 :** Le score F1 pour la classe 2 est légèrement meilleur dans les modèles 2, 3, et 4 (environ 0.7358). Le score F1 pour la classe 1 est le plus faible dans le premier modèle (0.6036) et légèrement meilleur dans les autres modèles (environ 0.6316).

En résumé, les modèles 2, 3 et 4 montrent une performance légèrement meilleure que le premier modèle sur toutes les métriques évaluées.

**1. Distribution des profondeurs des arbres.** La figure ci-dessous représente la distribution des profondeurs minimales des arbres pour différentes variables dans une forêt aléatoire. Chaque barre horizontale correspond à une variable (var1, var2, etc.), et les couleurs indiquent les différentes profondeurs minimales (de 0 à 9). Les valeurs numériques et les barres noires indiquent la profondeur minimale moyenne pour chaque variable.

Observations : Variables les plus proches de la racine : Les variables var6, var8 et var5 ont des profondeurs minimales moyennes relativement faibles (1.61, 1.95 et 2.06 respectivement). Cela suggère qu'elles apparaissent assez tôt dans les arbres de la forêt, jouant probablement un rôle crucial dans la classification initiale.

Variables de profondeur intermédiaire : Les variables comme var11, var7 et var2 ont des profondeurs minimales moyennes modérées (2.07, 2.31 et 2.47 respectivement). Elles interviennent après les variables les plus importantes mais demeurent significatives.

Variables plus profondes : var9, var4 et var12 apparaissent plus profondément avec des profondeurs minimales moyennes de 3.06, 3.18 et 3.32. Ces variables sont peut-être moins déterminantes dans la classification initiale mais contribuent à des décisions plus détaillées.

Variabilité des profondeurs : La distribution de couleurs sur chaque barre montre la variabilité des profondeurs minimales pour chaque variable. Certaines variables, comme var6 et var8, ont une profondeur minimale relativement constante, tandis que d'autres, comme var4 et var3, montrent une plus grande variation.

Moyennes : Les lignes noires et les valeurs numériques indiquent les profondeurs minimales moyennes pour chaque variable, fournissant un aperçu global de l'importance relative de chaque variable dans la structure des arbres.

Cela permet d'analyser la complexité des arbres générés et d'identifier s'il existe des biais structurels dans la construction de la forêt.

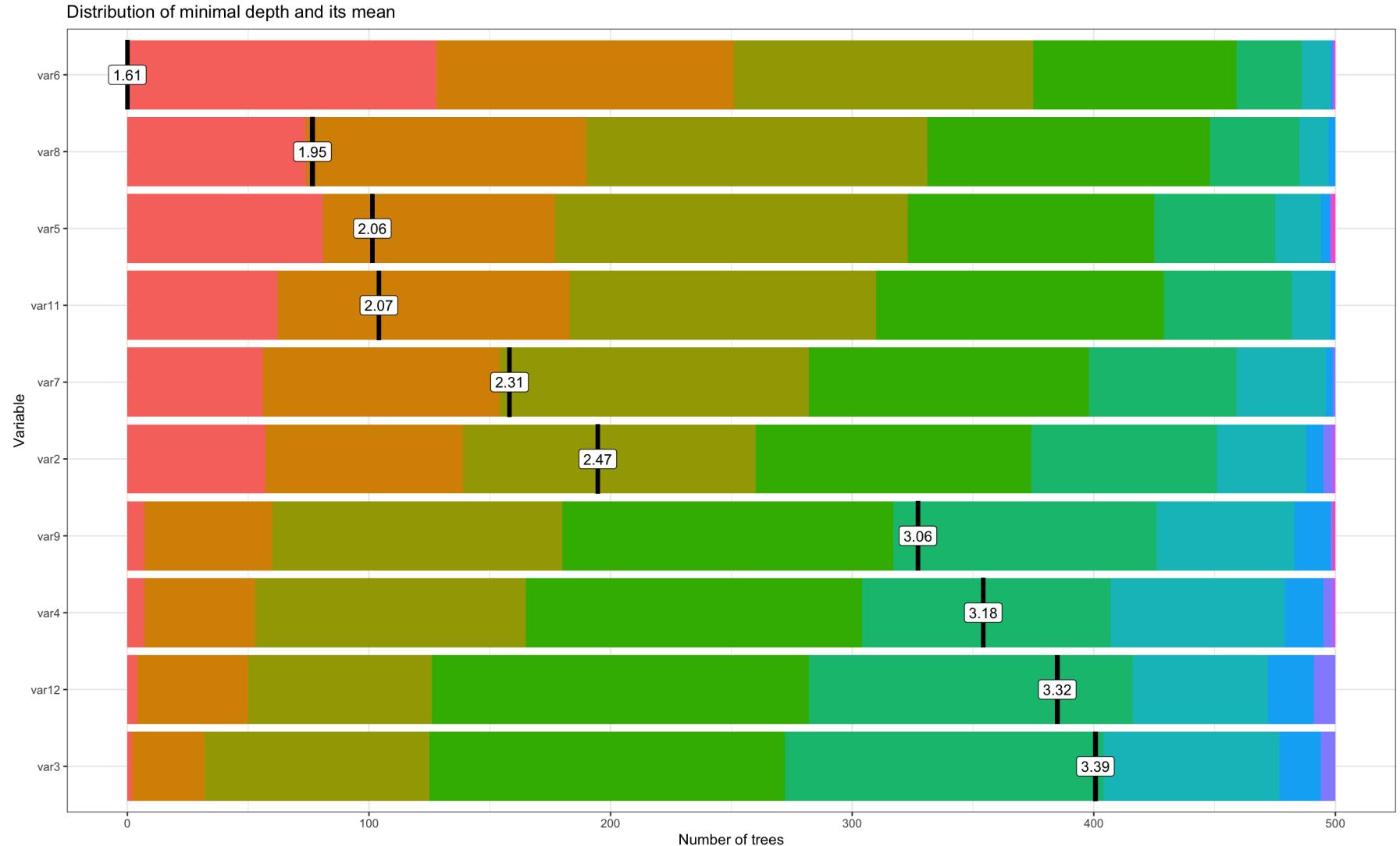


FIGURE 12 – Distribution des profondeurs des arbres dans la forêt aléatoire.

**2. Importance des variables.** La figure suivante illustre l'importance relative des variables dans la prédiction de la variable cible. Les variables les plus importantes contribuent significativement à la réduction de l'impureté moyenne dans les arbres.

Observations : Variables les plus importantes : Les points bleus sur le graphique indiquent les variables les plus importantes. Ces variables sont probablement celles qui réduisent le plus l'impureté moyenne dans les arbres de la forêt.

Proximité de la racine : L'axe des abscisses représente la profondeur minimale moyenne (mean\_min\_depth). Les variables proches de l'origine de l'axe des abscisses sont celles qui apparaissent près de la racine de l'arbre.

Fréquence d'utilisation comme racine : L'axe des ordonnées représente le nombre de fois qu'une variable est utilisée comme racine (times\_a\_root). Les variables situées en haut du graphique sont celles qui sont fréquemment utilisées comme racine dans les arbres.

Variables clés : Les variables var1, var2, var4, var5, var6, var7, var8, var9, var11 et var12 sont identifiées dans le graphique. Parmi celles-ci, certaines se distinguent par leur importance et leur apparition fréquente comme racine.

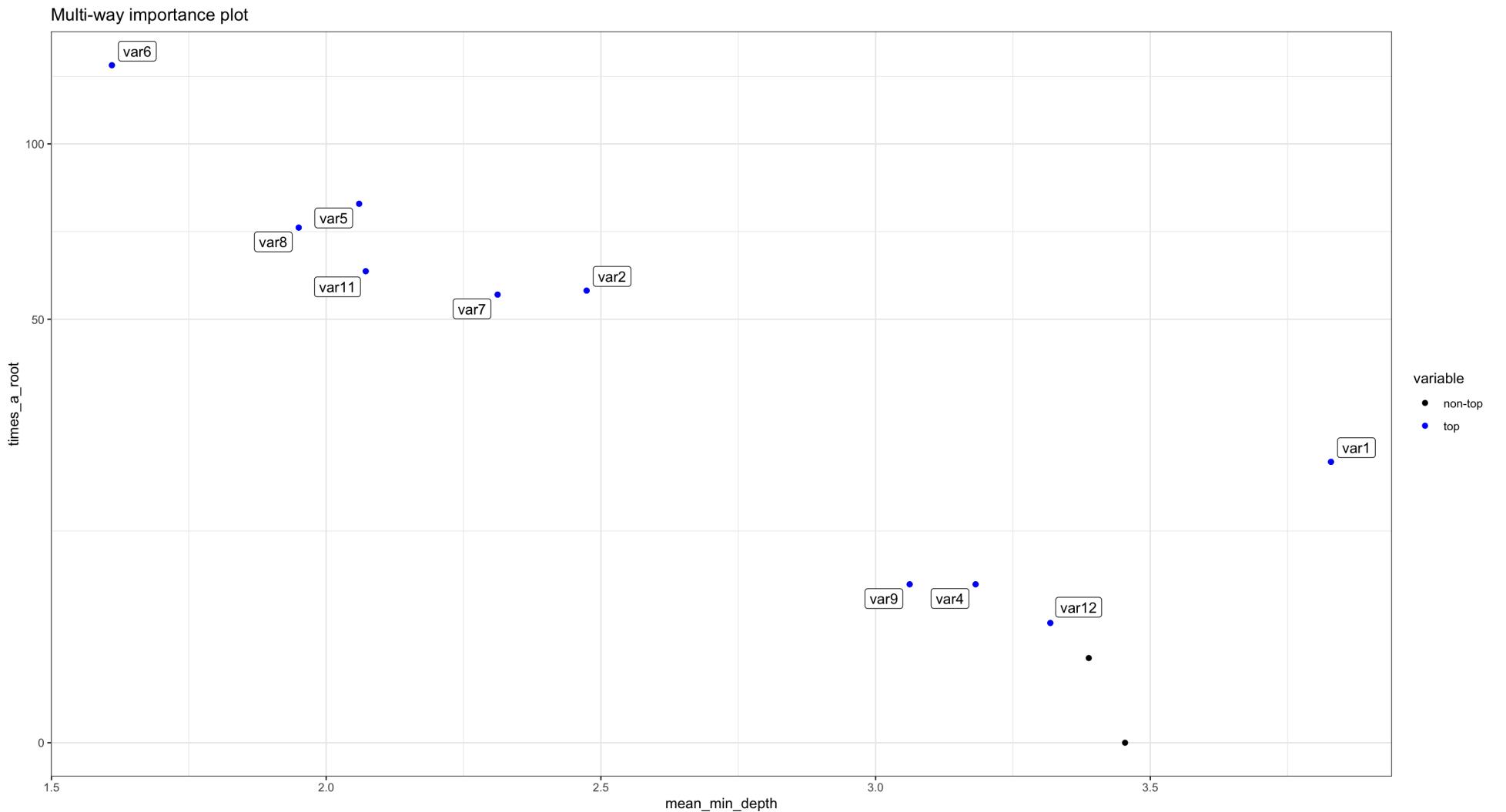


FIGURE 13 – Importance des variables dans le modèle Random Forest.

Ces observations permettent de mieux comprendre l'influence de chaque variable dans le modèle de forêt aléatoire et d'identifier celles qui sont les plus déterminantes pour la prédiction de la variable cible.

**3. Interactions entre variables.** L'analyse des interactions entre variables fournit des informations sur la manière dont deux variables influencent ensemble la variable cible. La figure suivante visualise l'interaction entre var5 et var12.

Observations : Influence de var5 : Le graphique montre un plot de dépendance partielle pour la variable var5. L'axe des abscisses représente les valeurs de var5, allant d'environ -4 à 4, tandis que l'axe des ordonnées représente la dépendance partielle, qui varie de 0,1 à 0,6.

Tendance générale : Initialement, la dépendance partielle reste relativement stable, puis augmente légèrement avant de diminuer significativement à mesure que les valeurs de var5 augmentent.

Interprétation : Cette tendance indique comment var5 influence la variable cible. Une augmentation des valeurs de var5 entraîne une légère augmentation de la dépendance partielle, suivie d'une diminution, suggérant un effet non linéaire de var5 sur la cible.

Importante interaction : Cette visualisation permet de mieux comprendre l'effet de var5 sur la variable cible, ce qui est crucial pour interpréter le comportement du modèle et prendre des décisions informées.

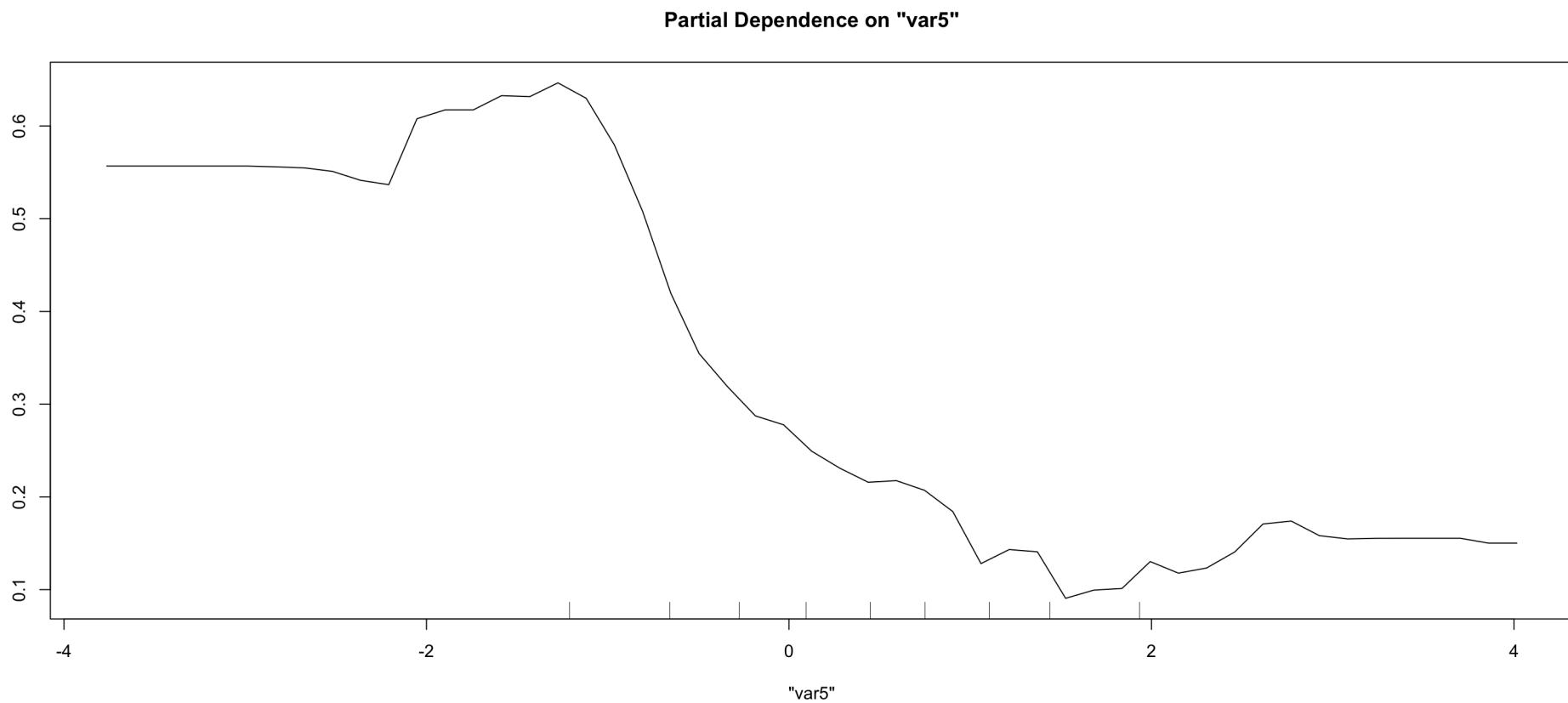
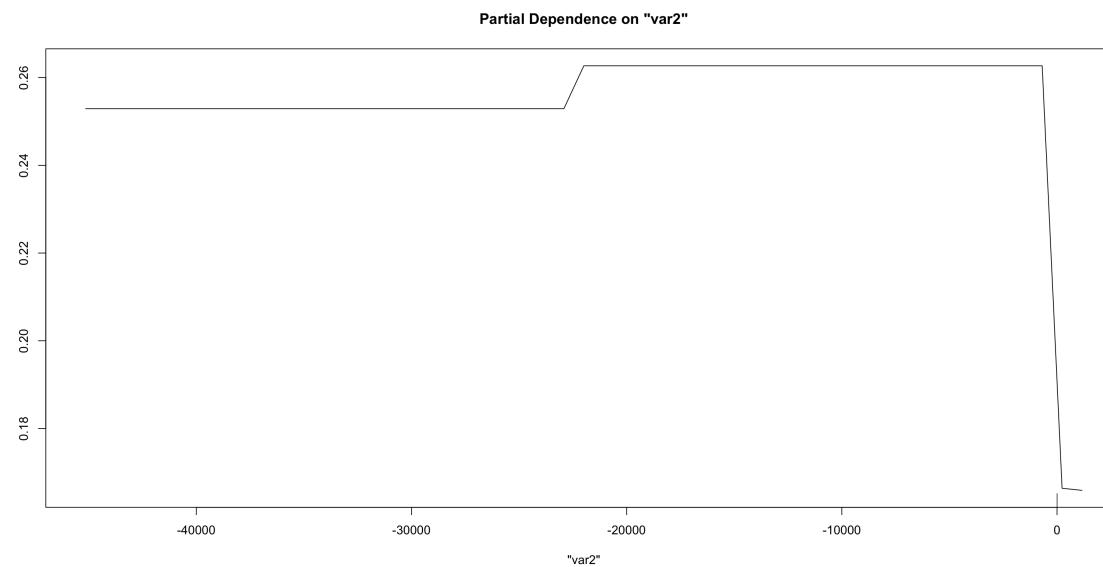


FIGURE 14 – Interaction entre var5 et var12.

**4. Partial plots.** Les *partial plots* permettent d'observer l'effet marginal d'une variable explicative sur la prédiction du modèle tout en maintenant les autres variables constantes. Les figures suivantes montrent l'effet des variables var2 et var3 sur la variable cible.



### Observations :

Effet de var2 : Le graphique indique que la valeur prédictive reste relativement constante autour de 0.24 pour la majorité de la plage de var2, avec une légère augmentation autour de -10000 suivie d'une diminution marquée lorsque var2 approche 0. Cela montre comment var2 influence la prédiction du modèle en maintenant les autres variables constantes.

## 4 Conclusion

En conclusion, ce projet a permis d'appliquer et de mettre en pratique des techniques fondamentales de machine learning sur des jeux de données aux problématiques distinctes. Pour la tâche supervisée, l'algorithme **Random Forest** s'est révélé efficace pour prédire la variable cible, en fournant des résultats interprétables grâce à l'analyse de l'importance des variables. Parallèlement, la tâche non supervisée de **clustering** a permis d'identifier des regroupements naturels dans les données, offrant ainsi une perspective précieuse sur leurs structures sous-jacentes.

Ce travail a non seulement consolidé notre compréhension des approches supervisées et non supervisées, mais a également renforcé notre capacité à interpréter les résultats et à les relier à des objectifs analytiques concrets. Ces compétences sont essentielles pour résoudre des problèmes complexes et extraire des informations pertinentes des données dans des contextes variés.