SAE 503 : Mise en œuvre d'un processus de Datamining

Romain TROILLARD

January 30, 2025

Table des matières

1 Présentation des données

1.1 Source des données

Le jeu de données utilisé dans cette étude provient de la plateforme Kaggle et est disponible à l'adresse suivante : `https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset`. Ce jeu de données est destiné à la prédiction du diabète et contient des informations sur divers attributs de patients, tels que l'âge, le sexe, l'indice de masse corporelle (IMC), le niveau d'HbA1c, et d'autres indicateurs de santé.

1.2 Description des variables

Le jeu de données comprend les variables suivantes :

- **gender** : Le sexe du patient (Male, Female).
- **age** : L'âge du patient en années.
- **hypertension** : Indicateur binaire (0 ou 1) pour l'hypertension.
- **heart_disease** : Indicateur binaire (0 ou 1) pour les maladies cardiaques.
- **smoking_history** : Historique de tabagisme du patient (never, former, current, etc.).
- **bmi** : Indice de masse corporelle (IMC) du patient.
- **HbA1c_level** : Niveau d'HbA1c (hémoglobine glyquée) du patient.
- **blood_glucose_level** : Niveau de glucose dans le sang du patient.
- **diabetes** : Variable cible binaire (0 ou 1) indiquant si le patient est diabétique.

1.3 Problématique

L'objectif de cette étude est de prédire si un patient est diabétique en fonction des attributs fournis. Le problème est donc un problème de classification supervisée, où la variable cible est binaire (diabétique ou non diabétique). Nous allons comparer trois approches de classification : les k-plus proches voisins (k-NN), l'analyse discriminante linéaire (LDA), et les arbres de décision. Nous analyserons également l'importance des différents prédicteurs dans la prédiction de la variable cible.

1.4 Analyse préliminaire

Le jeu de données contient un mélange de prédicteurs quantitatifs (comme l'âge, l'IMC, le niveau d'HbA1c, etc.) et qualitatifs (comme le sexe et l'historique de tabagisme). Une première analyse sera réalisée en utilisant uniquement les prédicteurs quantitatifs, suivie d'une deuxième analyse incluant tous les prédicteurs après avoir effectué les transformations nécessaires pour appliquer les méthodes k-NN et LDA.

Le jeu de données contient 100 000 individus, dont 8 500 sont diabétiques et 91 500 ne le sont pas. Cette répartition déséquilibrée entre les classes devra être prise en compte lors de l'analyse et de la modélisation.

2 Analyse des données

Tous les résultats, graphiques et analyses présentés dans ce rapport ont été réalisés en utilisant le langage de programmation R. Les codes sources des analyses sont disponibles dans le dossier du rapport.

2.1 Pretraitement des données

Le prétraitement des données est une étape cruciale pour garantir la qualité des résultats obtenus lors de l'analyse. Voici les étapes de prétraitement effectuées sur le jeu de données :

- **Suppression des valeurs manquantes** : Le jeu de données initial a été nettoyé en supprimant les lignes contenant des valeurs manquantes. Cela permet d'éviter les biais dans les modèles de classification.
- **Filtrage des catégories non pertinentes** : Les individus avec un historique de tabagisme marqué comme "No Info" et ceux avec un genre marqué comme "Other" ont été exclus de l'analyse, car ces catégories ne fournissent pas d'informations utiles pour la prédiction.
- **Encodage des variables catégorielles** : La variable `smoking_history` a été encodée en facteur avec des niveaux ordonnés : "never", "former", "current", "ever", et "not current". De même, la variable cible `diabetes` a été transformée en facteur avec les niveaux "no diabete" et "diabete".
- **Échantillonnage équilibré** : Étant donné que le jeu de données est déséquilibré (8 500 cas de diabète contre 91 500 cas sans diabète), un échantillonnage équilibré a été effectué. Un sous-ensemble de 1 000 individus a été créé, comprenant 500 cas de diabète et 500 cas sans diabète, afin de permettre une analyse plus équilibrée.
- **Mélange des données** : Les données ont été mélangées aléatoirement pour éviter tout biais lié à l'ordre des observations.

Ces étapes de prétraitement permettent de préparer le jeu de données pour les analyses ultérieures en garantissant que les données sont propres, équilibrées et prêtes à être utilisées par les algorithmes de classification.

2.2 Analyse avec prédicteurs quantitatifs

Cette section présente les résultats obtenus en utilisant uniquement les prédicteurs quantitatifs suivants :

- `age` : Âge du patient.
- `bmi` : Indice de masse corporelle.
- `HbA1c_level` : Niveau d'HbA1c.
- `blood_glucose_level` : Niveau de glucose dans le sang.

2.2.1 Méthode des k-plus proches voisins (k-NN)

L'algorithme k-NN a été implémenté avec une distance euclidienne, en testant des valeurs de (k) de 3 à 51 (pas de 2). L'évaluation s'est faite par validation croisée sur 500 itérations.

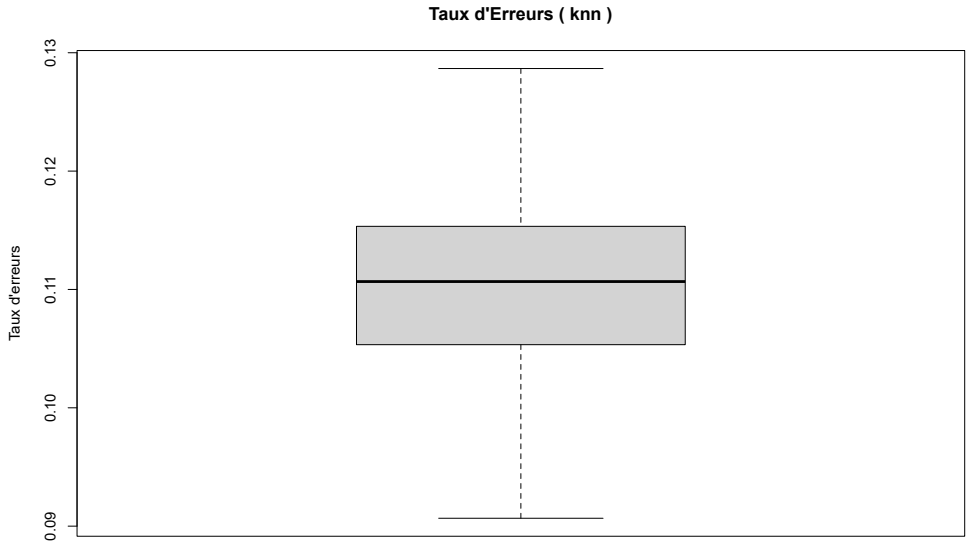

Figure 1: Taux d'erreurs pour k-NN sur les prédicteurs quantitatifs

La distribution des taux d'erreurs présente une médiane de 0.11, avec une répartition symétrique entre 0.09 et 0.13. La rareté des valeurs extrêmes témoigne de la stabilité de l'algorithme sur l'ensemble des itérations.

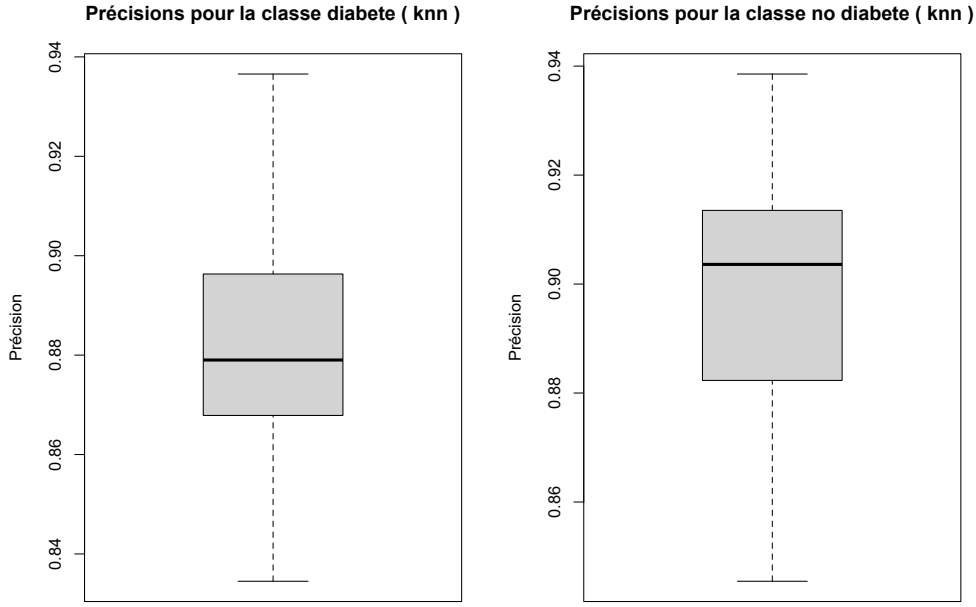

Figure 2: Précisions des classes pour k-NN sur les prédicteurs quantitatifs

Les résultats indiquent une performance équilibrée entre les deux classes, avec un léger avantage pour la prédiction des cas non diabétiques.

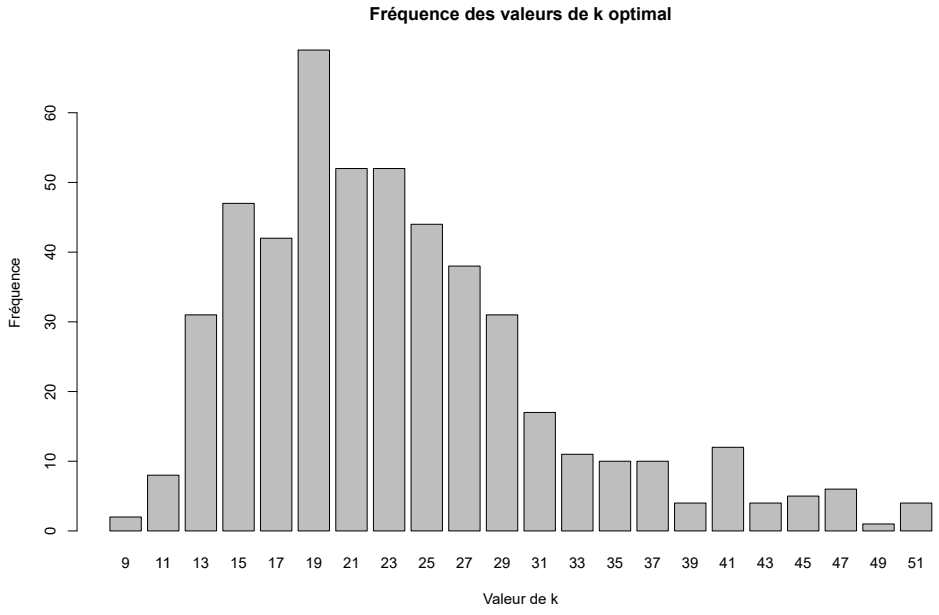

Figure 3: Fréquence des valeurs de k optimales (k-NN)

La valeur optimale de (k) se situe autour de 19, cette configuration apparaissant le plus fréquemment lors de la validation croisée.

2.2.2 Analyse discriminante linéaire (LDA)

L'application de la LDA, basée sur l'hypothèse de normalité des données et d'homogénéité des matrices de covariance, a également fait l'objet de 500 itérations de validation croisée.

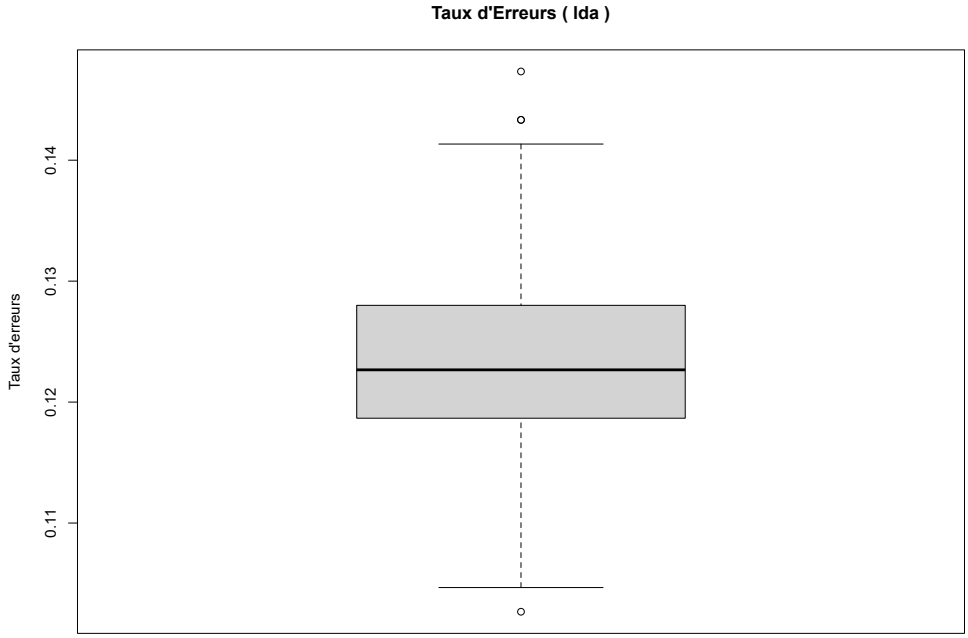

Figure 4: Taux d'erreurs pour l'analyse discriminante linéaire (LDA)

Le taux d'erreurs médian de 0.12 reflète une performance satisfaisante, bien qu'une dispersion plus importante et la présence de valeurs aberrantes suggèrent une stabilité inférieure à celle du k-NN.

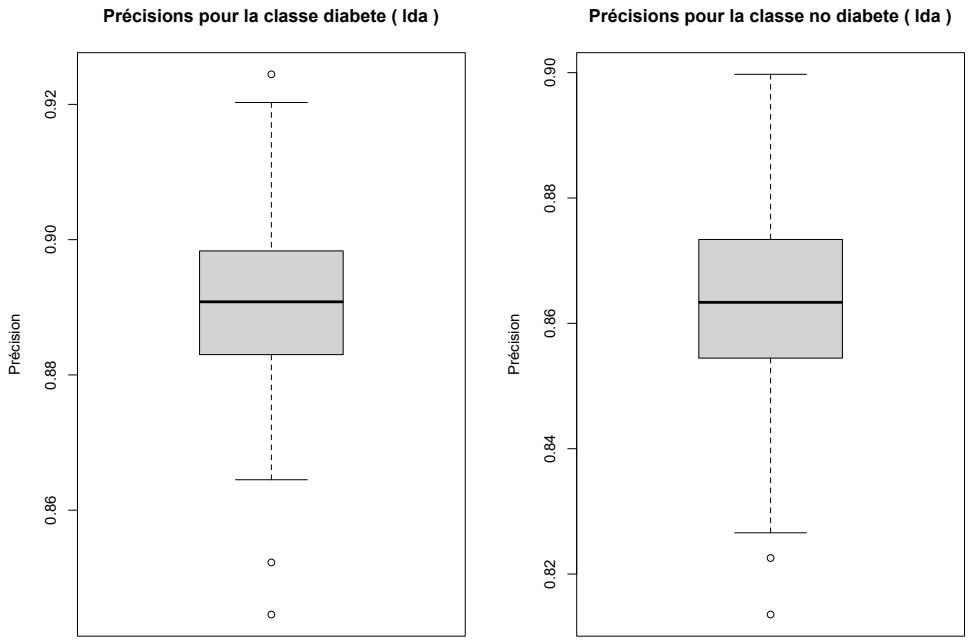

Figure 5: Précisions des classes pour LDA sur les prédicteurs quantitatifs

L'analyse des classes révèle une précision légèrement supérieure pour la détection des cas diabétiques, tout en maintenant une bonne performance globale.

2.2.3 Arbres de décision

Notre troisième approche a consisté à construire un arbre de décision partitionnant l'espace des prédicteurs en régions homogènes, évalué sur 500 itérations.

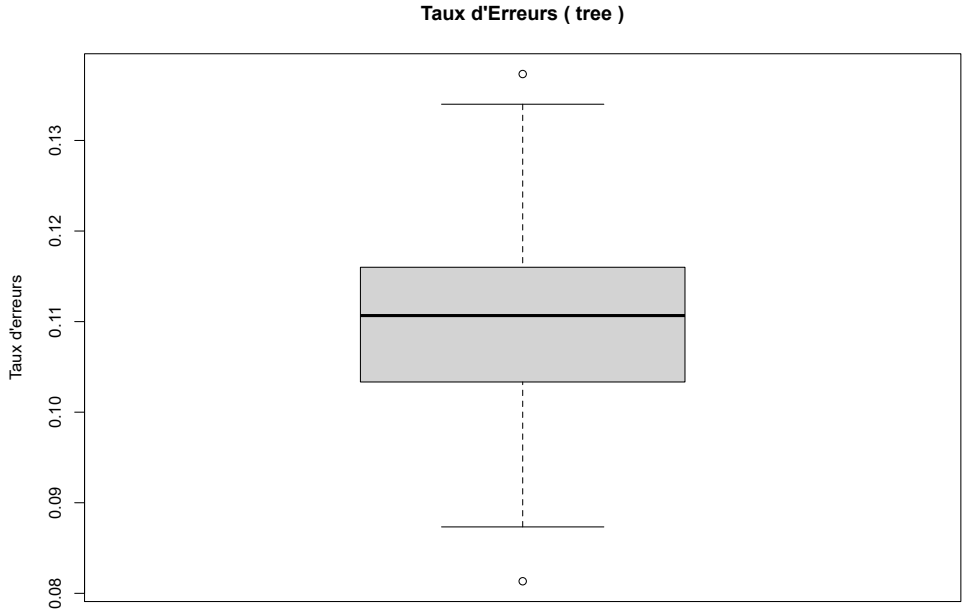

Figure 6: Taux d'erreurs pour l'arbre de décision

Avec un taux d'erreurs médian de 0.11 et une dispersion modérée, ce modèle démontre une stabilité satisfaisante malgré quelques valeurs aberrantes.

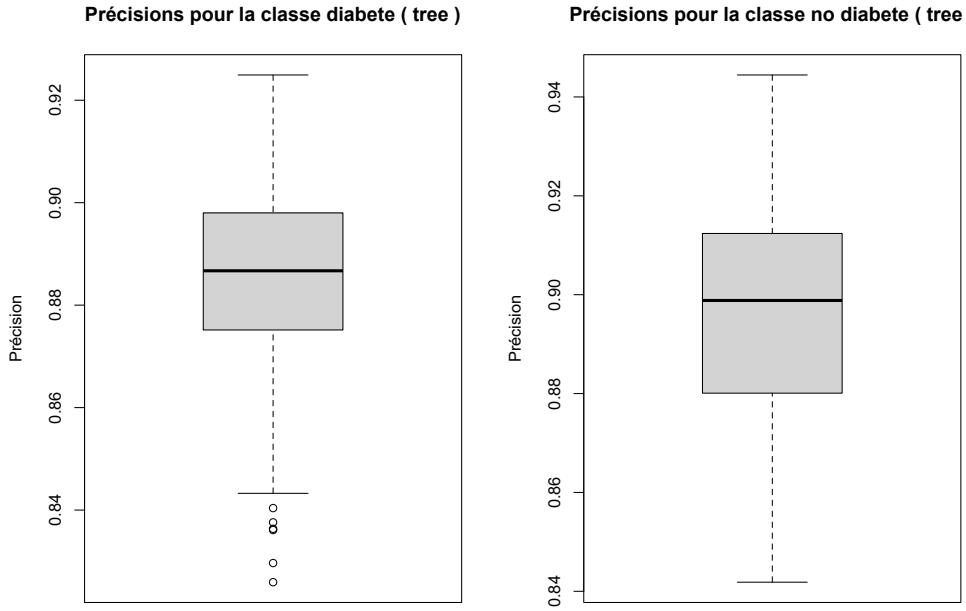

Figure 7: Précisions des classes pour l'arbre de décision sur les prédicteurs quantitatifs

Les précisions obtenues (0.89 pour le diabète, 0.90 pour les cas négatifs) se comparent favorablement aux autres méthodes testées.

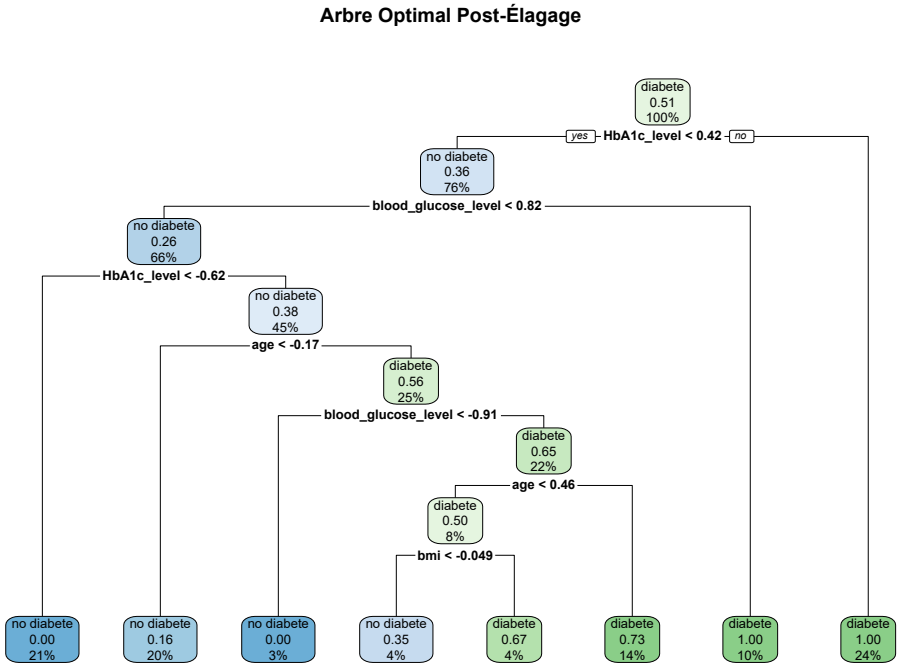

Figure 8: Arbre de décision optimal post-élagage

L'arbre optimal identifie le niveau d'HbA1c et la glycémie comme variables déterminantes, avec quatre feuilles pures regroupant 53% des observations.

2.2.4 Synthèse comparative

Les performances des trois modèles (k-NN, LDA et arbre de décision) sont résumées dans le tableau 1.

Modèle	**Taux d'erreurs moyen**	**Remarques**
k-NN	0.11	$k = 19$
LDA	0.12	
Arbre de décision	0.11	Variables importantes : HbA1c et glucose

Table 1: Comparaison des performances des modèles

Le tableau 1 montre que les trois modèles sont performants, avec des taux d'erreurs inférieurs à 0.15.

2.3 Extension à l'ensemble des prédicteurs

Cette seconde phase intègre les variables catégorielles et binaires à notre analyse.

2.3.1 Méthode des k-plus proches voisins (k-NN)

L'application du k-NN sur l'ensemble enrichi utilise un encodage dummy pour les variables catégorielles, conservant les variables binaires dans leur forme originale.

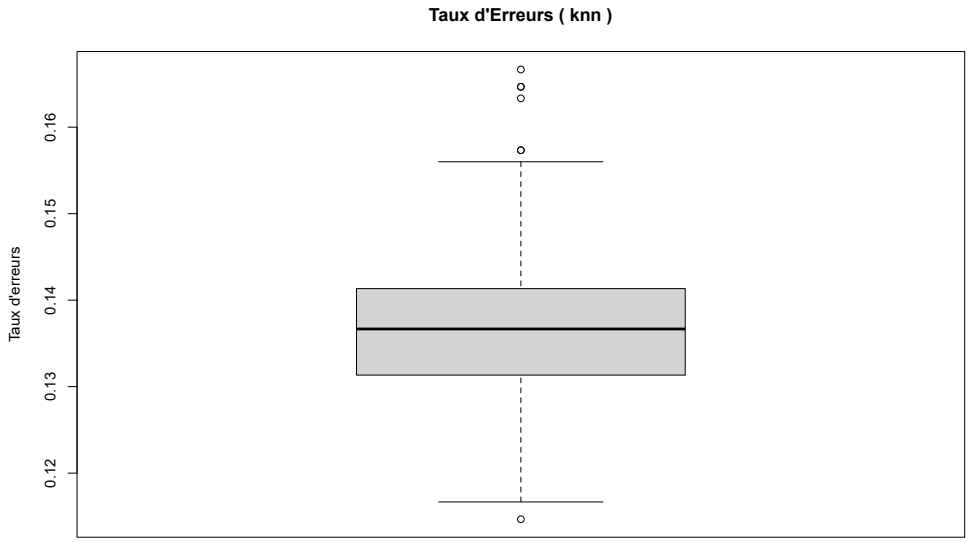

Figure 9: Taux d'erreurs pour k-NN

L'ajout des nouvelles variables augmente légèrement la variabilité, avec une médiane des erreurs à 0.14 et une distribution plus étalée (0.12-0.16).

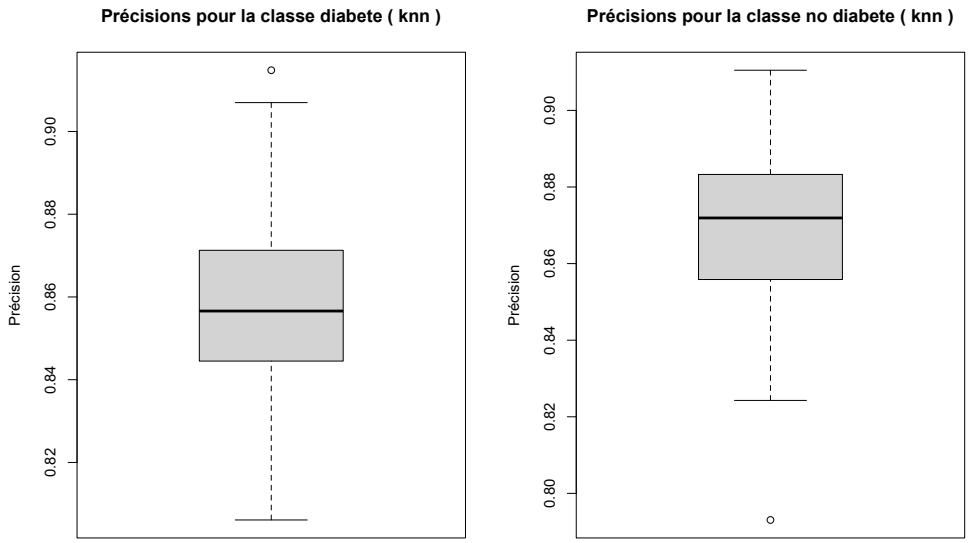

Figure 10: Précisions des classes pour k-NN

Bien que légèrement inférieures aux résultats obtenus avec les seuls prédicteurs quantitatifs, les précisions restent satisfaisantes (0.86 pour le diabète, 0.88 pour les cas négatifs).

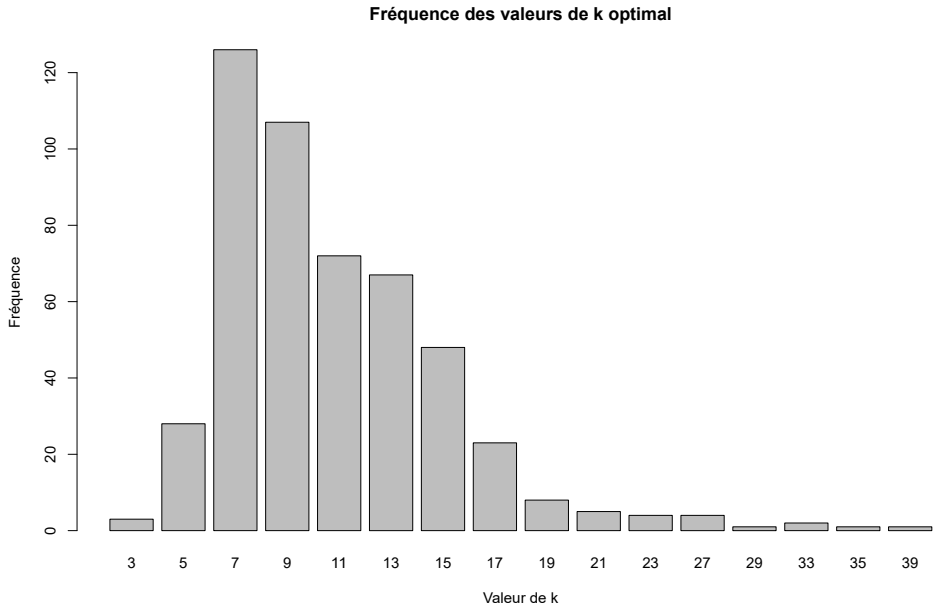

Figure 11: Fréquence des valeurs de k optimales (k-NN)

La valeur optimale de (k) se stabilise autour de 7, différant significativement du modèle précédent.

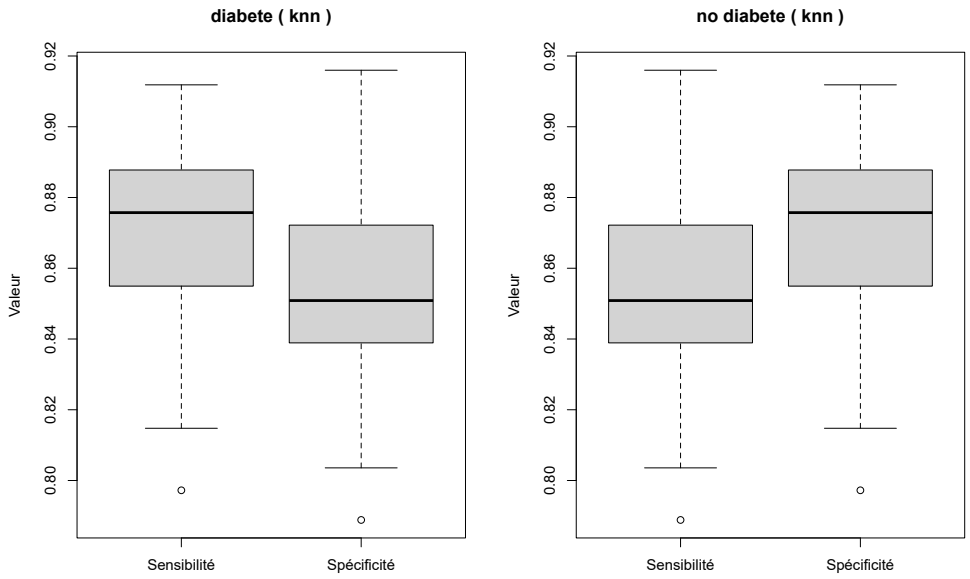

Figure 12: Sensibilité et spécificité pour k-NN

Une sensibilité légèrement supérieure indique une meilleure détection des cas positifs comparée aux cas négatifs.

2.3.2 Analyse discriminante linéaire (LDA)

L'application de la LDA intègre une transformation par ACM des variables catégorielles, conservant les variables binaires intactes.

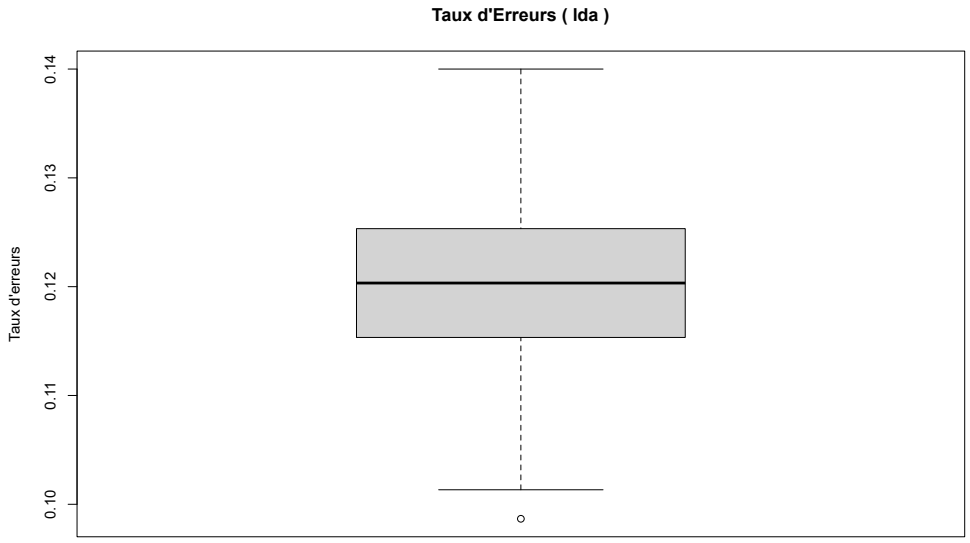

Figure 13: Taux d'erreurs pour l'analyse discriminante linéaire (LDA)

La concentration des erreurs autour d'une médiane de 0.12 suggère une meilleure stabilité que la version précédente du modèle.

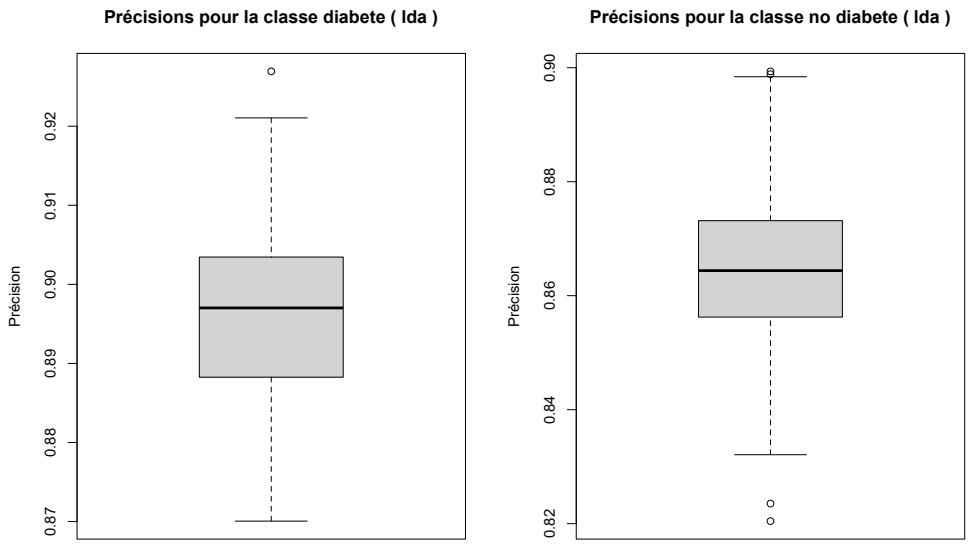

Figure 14: Précisions des classes pour LDA

Les résultats montrent une efficacité accrue dans la détection des cas diabétiques, avec des mesures comme la sensibilité et la précision atteignant des niveaux élevés.

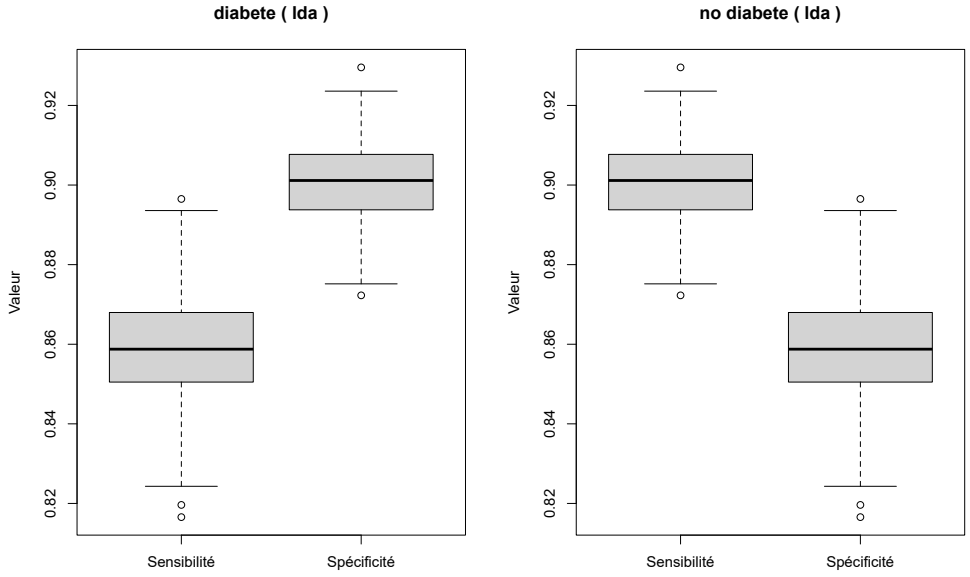

Figure 15: Sensibilité et spécificité pour LDA

Contrairement au k-NN, ce modèle démontre une performance nettement supérieure dans l'identification

des cas négatifs, affichant une précision et une spécificité plus élevées.

2.3.3 Arbres de décision

L'arbre de décision est capable de traiter naturellement l'ensemble des variables sans nécessiter de transformation préalable, ce qui simplifie son utilisation et réduit les étapes de prétraitement des données. Cette flexibilité lui permet de s'adapter facilement à différents types de données, tout en maintenant une performance optimale.

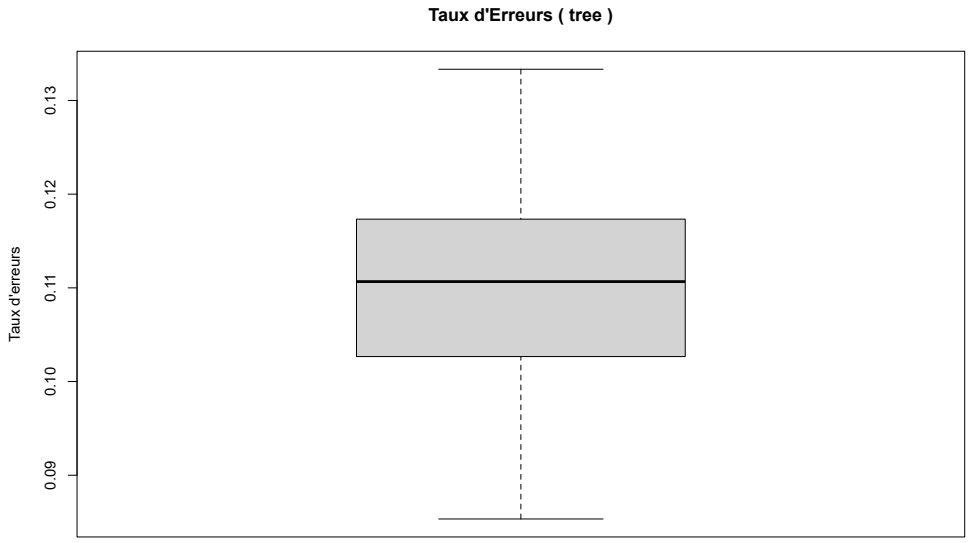

Figure 16: Taux d'erreurs pour l'arbre de décision

La concentration remarquable des erreurs autour de 0.11 met en évidence une performance stable et cohérente du modèle, suggérant une précision élevée et une capacité à maintenir des résultats fiables dans des conditions variées.

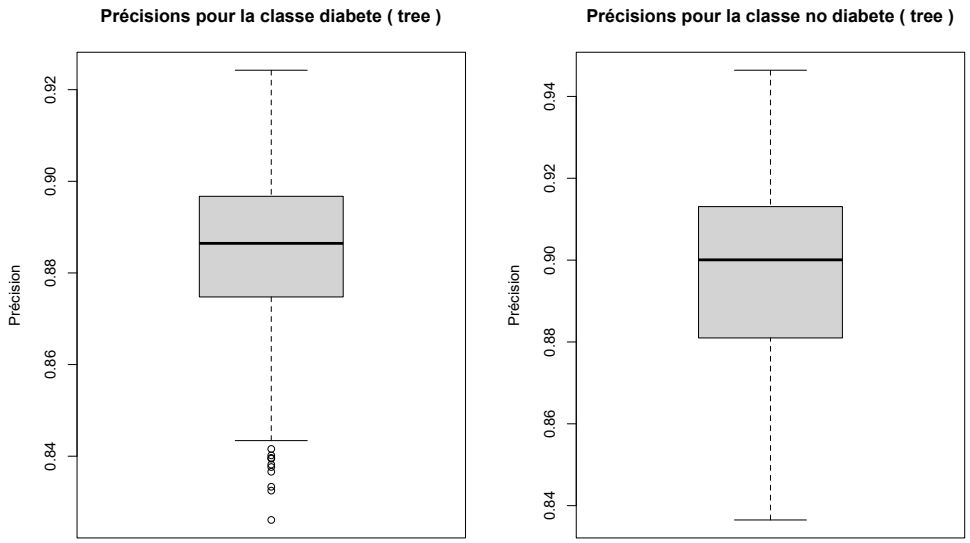

Figure 17: Précisions des classes pour l'arbre de décision

Les précisions obtenues, atteignant respectivement 0.89 et 0.90, démontrent une constance notable dans les performances du modèle, soulignant sa capacité à maintenir un niveau élevé de fiabilité même avec l'introduction de nouvelles variables ou de conditions changeantes. Cette stabilité renforce la confiance dans la cohérence des résultats produits.

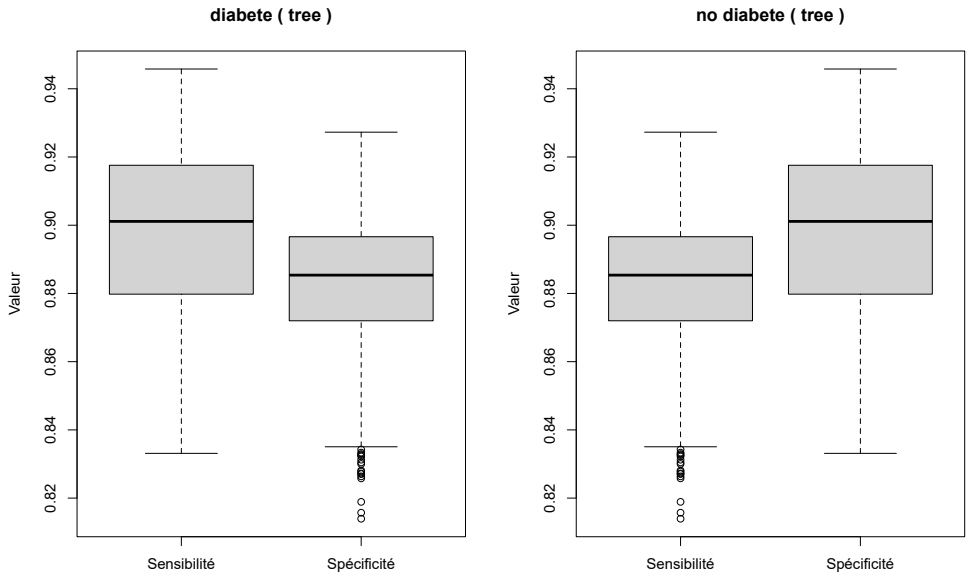

Figure 18: Sensibilité et spécificité pour l'arbre de décision

L'équilibre entre sensibilité et spécificité (≈ 0.90) démontre une performance homogène dans la détection des deux classes.

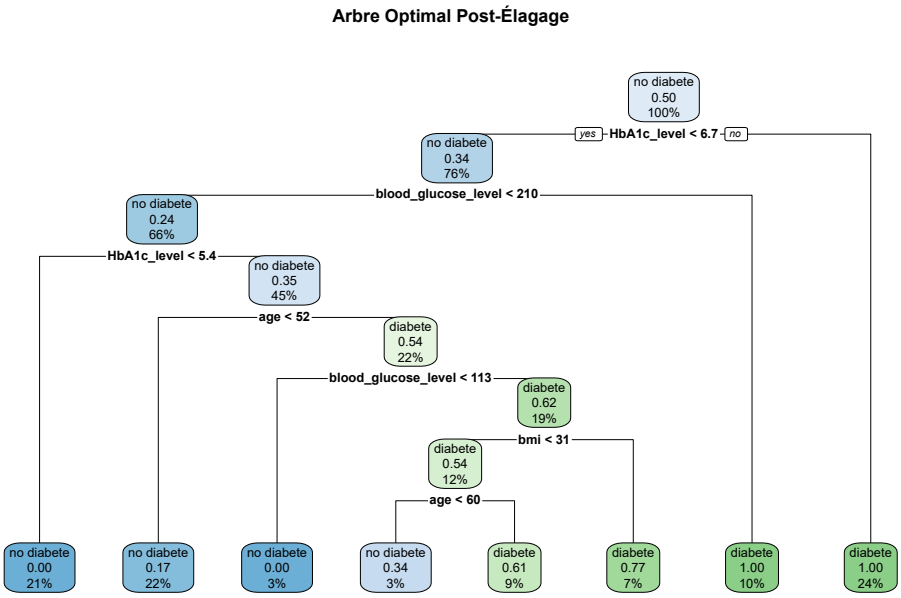

Figure 19: Arbre de décision optimal post-élagage

La structure finale conserve la prédominance des niveaux d'HbA1c et de glucose, avec 58% des observations classées dans des feuilles pures.

2.3.4 Bilan comparatif

Les performances des trois modèles (k-NN, LDA et arbre de décision) sont résumées dans le tableau 2.

Modèle	**Taux d'erreurs moyen**	**Remarques**
k-NN	0.14	$k = 7$
LDA	0.12	
Arbre de décision	0.11	Variables importantes : HbA1c et glucose

Table 2: Comparaison des performances des modèles avec l'ensemble des prédicteurs

L'intégration des variables supplémentaires n'a pas entraîné de modification significative des performances globales des modèles, ce qui suggère que ces ajouts n'ont pas perturbé sa capacité à maintenir un niveau de précision et de fiabilité constant. Cela nous questionne sur l'importance des prédicteurs non quantitatifs dans la prédiction du diabète.

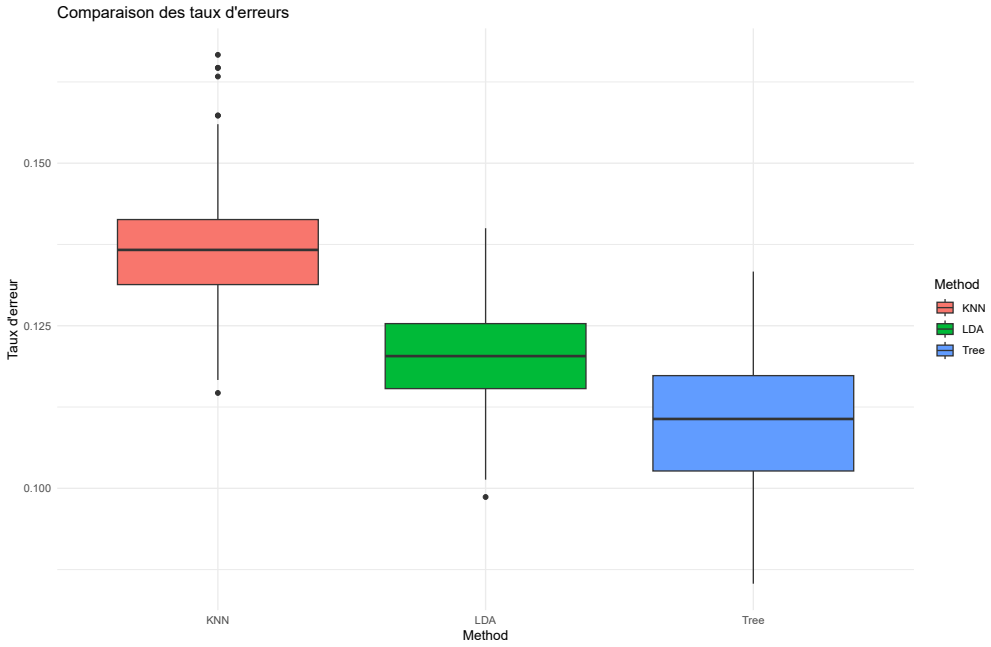

Figure 20: Taux d'erreurs par modèle

La distribution plus compacte des erreurs observée pour l'arbre de décision, par rapport aux autres modèles, confirme sa supériorité en termes de précision et de cohérence. Cette concentration étroite des erreurs autour d'une valeur faible témoigne d'une capacité à produire des résultats plus fiables et moins dispersés, renforçant ainsi sa position comme méthode privilégiée dans ce contexte.

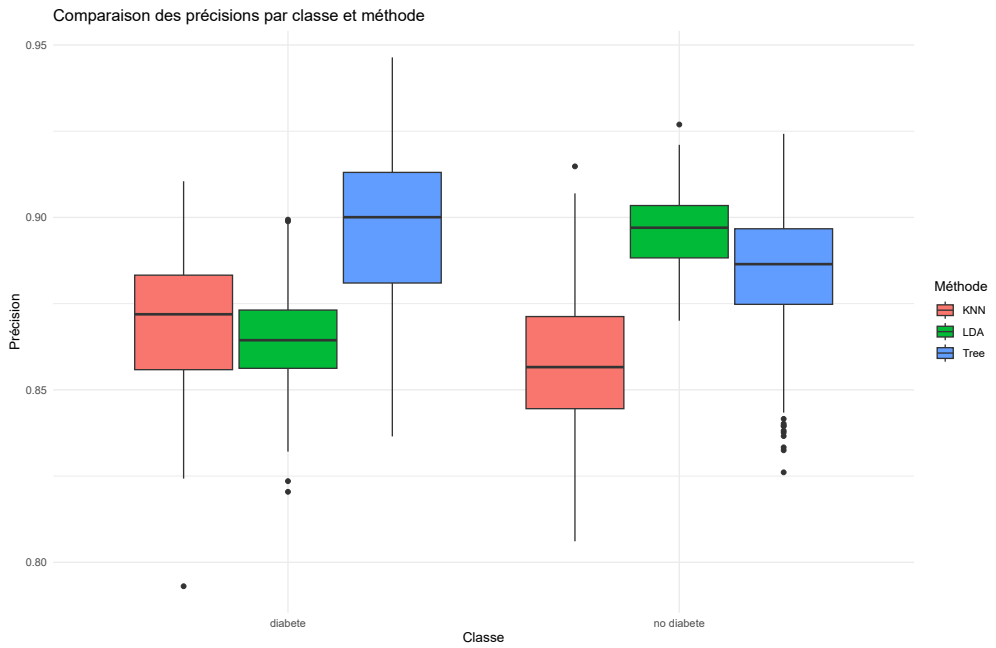

Figure 21: Précisions par classe et par modèle

L'arbre de décision excelle dans l'identification des cas diabétiques, offrant une précision et une fiabilité supérieures. Cet avantage est crucial pour l'application clinique, où une détection exacte des cas positifs est essentielle pour une prise en charge médicale efficace.

3 Importance des prédicteurs

Cette section analyse l'importance relative des prédicteurs dans le diagnostic du diabète. La méthode de permutation a été employée pour cette évaluation : nous avons comparé le taux d'erreurs du modèle original avec celui obtenu après permutation aléatoire des valeurs de chaque prédicteur. L'écart entre ces taux permet de quantifier l'influence de chaque variable. Les résultats obtenus pour les trois modèles révèlent des tendances similaires :

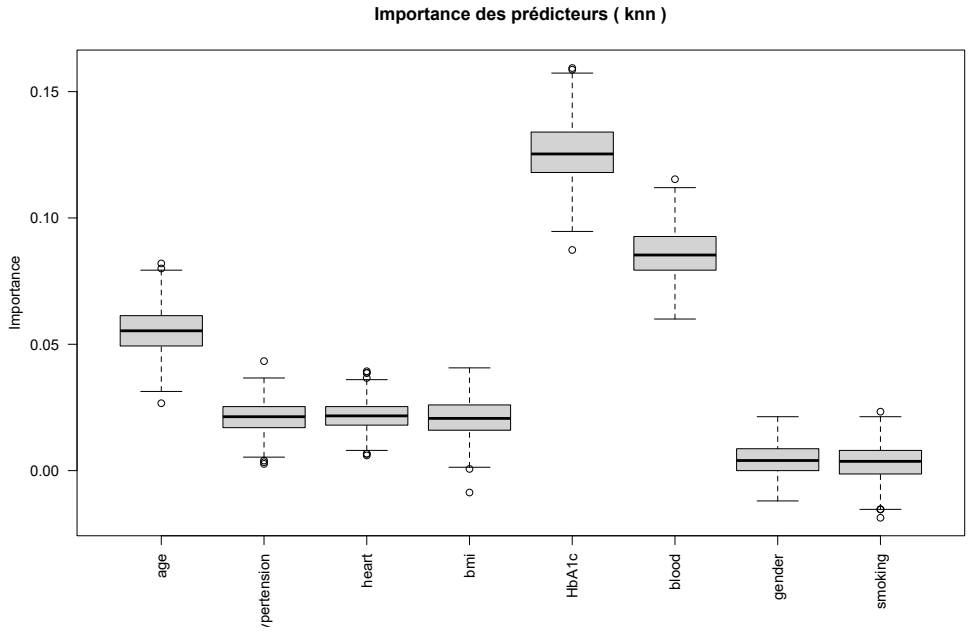

Figure 22: Importance des prédicteurs pour k-NN

Pour le modèle k-NN (figure 22), le niveau de HbA1c et la glycémie émergent comme les facteurs déterminants, tandis que l'âge conserve une influence notable mais secondaire.

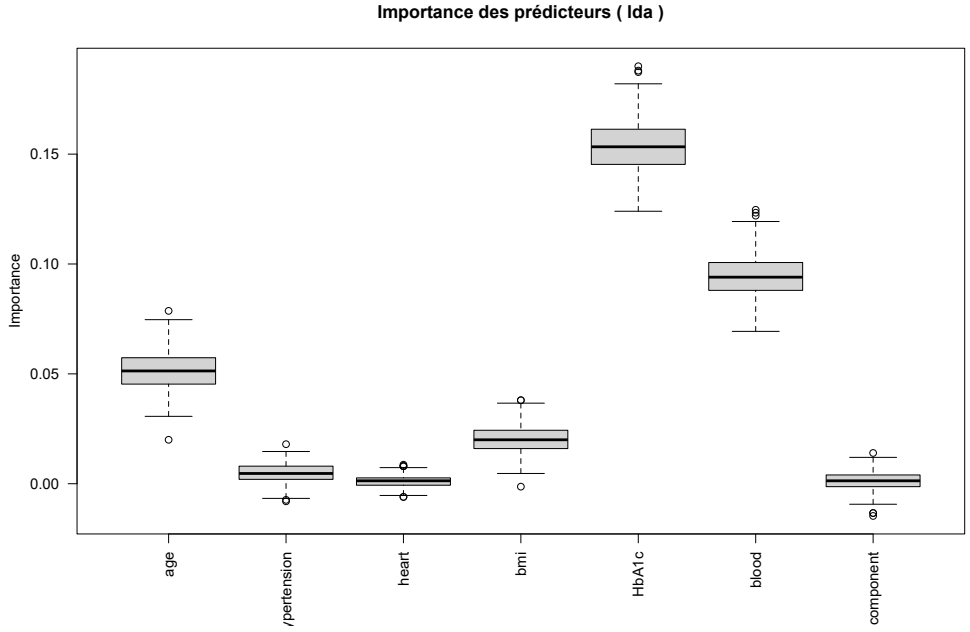

Figure 23: Importance des prédicteurs pour LDA

L'analyse LDA (figure 23) confirme la prépondérance de la HbA1c et de la glycémie. L'âge et l'IMC jouent également un rôle significatif, quoique moins marqué.

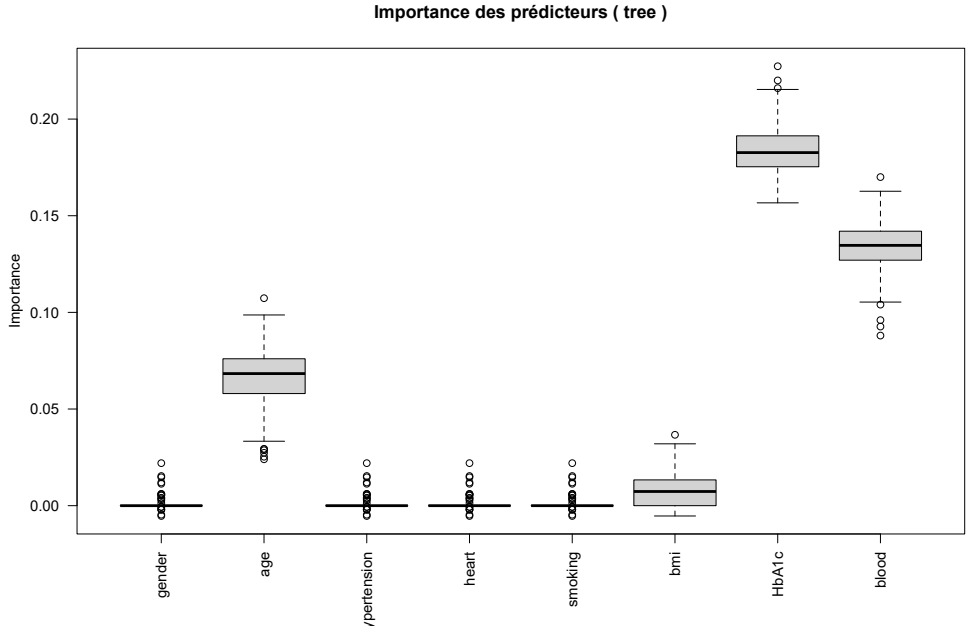

Figure 24: Importance des prédicteurs pour l'arbre de décision

L'arbre de décision (figure 24) corrobore ces observations, avec une hiérarchie similaire : HbA1c et glycémie en tête, suivies par l'âge et l'IMC. Les autres variables exercent une influence marginale. Cette convergence des trois modèles souligne la prédominance des variables quantitatives, particulièrement la HbA1c et la glycémie, dans la prédiction du diabète. Les autres facteurs présentent une importance négligeable.

4 Conclusion

Notre étude a comparé l'efficacité de trois modèles de classification (k-NN, LDA et arbre de décision) dans la prédiction du diabète. L'analyse s'est portée sur deux configurations : l'une limitée aux prédicteurs quantitatifs, l'autre incluant l'ensemble des variables (quantitatives, catégorielles et binaires). Bien que les trois approches affichent des performances satisfaisantes avec des taux d'erreurs moyens inférieurs à 0.15, l'arbre de décision se distingue par sa précision légèrement supérieure et sa meilleure identification des cas positifs. L'étude nous a permis d'identifier que deux facteurs sont particulièrement importants : le niveau de HbA1c et la glycémie, avec l'âge et l'IMC qui jouent aussi un rôle, mais moins important. En combinant ces résultats avec le fait que l'arbre de décision donne les meilleurs résultats parmi nos modèles, cela pourrait aider à améliorer le diagnostic du diabète et à mieux comprendre quels facteurs sont les plus importants pour détecter cette maladie.

Annexes

Liste des Figures

Liste des Tableaux

Liste des Références

- `https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset`