École Nationale Supérieure
d'Informatique et d'Analyse des Systèmes
- Rabat

Rapport du projet Data Driven Decision Making

Prédiction de la réussite scolaire à l'aide du Machine Learning

Élèves :
Imane BENABBOU
Chaymae BOUAZZA

Enseignant :
M. Youness TABII

14 mai 2025

Remerciements

Avant de présenter en détail notre rapport de projet, nous souhaitons exprimer notre profonde gratitude envers toutes les personnes qui ont contribué, de près ou de loin, à la réussite de notre travail. Leur soutien et leur implication ont été essentiels tout au long de ce parcours.

Tout d'abord, nous souhaitons adresser nos remerciements les plus sincères au Pr. Youness TABII, notre encadrant, dont l'expertise, la disponibilité et le soutien indéfectible ont été d'une valeur inestimable. Grâce à lui, nous avons pu mener à bien ce projet et bénéficier de ses précieux conseils qui ont enrichi notre compréhension et notre approche.

Nous tenons également à exprimer notre reconnaissance envers notre école, l'ENSIAS, ainsi que l'ensemble du corps pédagogique, pour avoir mis à notre disposition les ressources nécessaires et pour leur engagement dans notre formation. Leur encadrement de qualité et leur accompagnement bienveillant ont constitué un cadre propice à l'apprentissage et à l'épanouissement de nos compétences.

Résumé

Dans le cadre de notre formation en data-driven decision making, nous avons mené un projet visant à exploiter les techniques de science des données pour analyser et prédire les performances scolaires d'élèves, dans le but de soutenir les décisions pédagogiques et l'orientation académique.

Le projet repose sur l'utilisation de données éducatives comprenant des informations personnelles, sociales et académiques. En appliquant des algorithmes de machine learning (régression, classification, clustering), nous avons développé des modèles capables de prédire la note finale d'un élève, d'estimer sa probabilité de réussite ou d'échec, et de regrouper les élèves en profils d'apprentissage distincts. Ces analyses fournissent des indicateurs objectifs pour guider l'orientation ou le soutien éducatif.

Ce rapport détaille les étapes du processus décisionnel basé sur les données, de la préparation du jeu de données à l'évaluation des performances des modèles, en passant par l'interprétation des résultats. Ce travail illustre la manière dont l'intelligence artificielle et l'analyse prédictive peuvent enrichir les décisions stratégiques dans le domaine de l'éducation.

Mots clés : Data-driven decision making, Machine Learning, Éducation, Prédiction scolaire, Orientation académique, Analyse des données, Clustering, Classification.

Abstract

As part of our training in data-driven decision making, we conducted a project aimed at leveraging data science techniques to analyze and predict students' academic performance, with the goal of supporting educational decisions and academic orientation.

The project is based on the use of educational data containing personal, social, and academic information. By applying machine learning algorithms (regression, classification, and clustering), we developed models capable of predicting students' final exam scores, estimating their probability of passing or failing, and grouping them into distinct learning profiles. These analyses provide objective indicators to guide educational support and orientation.

This report details the steps of the data-driven decision-making process, from data preparation to model evaluation, including result interpretation. It illustrates how artificial intelligence and predictive analytics can enhance strategic decision-making in the field of education.

Keywords : Data-driven decision making, Machine Learning, Education, Academic prediction, Student profiling, Orientation, Clustering, Classification.

Introduction Générale

Dans un contexte où l'éducation évolue vers une personnalisation accrue des parcours, la capacité à exploiter efficacement les données devient un levier stratégique pour les établissements scolaires. L'essor de la data science et du machine learning offre aujourd'hui de nouvelles opportunités pour comprendre, anticiper et améliorer la réussite académique des élèves.

L'analyse des données éducatives permet non seulement de mieux évaluer les performances individuelles, mais aussi d'identifier les facteurs clés de succès et d'échec. Elle constitue un outil précieux pour orienter les élèves de manière plus objective et plus adaptée à leur profil. Dans ce cadre, la mise en œuvre d'un processus de prise de décision fondé sur les données, ou data-driven decision making, représente une approche innovante et pertinente.

Le présent projet s'inscrit dans cette démarche. Il vise à analyser un jeu de données réelles sur les performances scolaires d'élèves afin de construire des modèles prédictifs. À l'aide d'algorithmes de machine learning, nous cherchons à estimer les notes finales, à prédire la réussite ou l'échec, et à regrouper les élèves selon leurs caractéristiques dominantes. Ces résultats permettent d'envisager des recommandations pédagogiques ciblées et, dans certains cas, des pistes d'orientation.

Cette introduction présente le contexte et les enjeux du projet. Les chapitres suivants détaillent successivement les étapes de préparation des données, le choix et l'évaluation des modèles d'apprentissage automatique, ainsi que l'interprétation des résultats obtenus. Ce travail met en lumière le rôle essentiel que peuvent jouer les données dans l'amélioration de la prise de décision éducative.

Table des matières

Table des figures

Chapitre 1

Structure des données

1.1 Introduction

Dans le cadre de ce projet, nous souhaitons analyser et prédire la performance académique d'un ensemble de 708 étudiants à l'aide de leurs caractéristiques personnelles et scolaires. La base de données utilisée, student_performance_dataset.csv, regroupe des informations démographiques (sexe, niveau d'éducation des parents), comportementales (heures d'étude hebdomadaires, taux de présence) et pédagogiques (notes passées, accès à Internet, activités extrascolaires), ainsi que la note finale et le statut « Pass / Fail ». Cette richesse d'attributs permet de mettre en œuvre des méthodes supervisées et non supervisées pour mieux comprendre les facteurs influençant la réussite.

1.2 Description des données

Le tableau suivant présente les principales colonnes de la base de données utilisées dans le projet :

Table 1.1 – Structure des données

Colonne	**Type**	**Description**
Student_ID	Catégoriel	Identifiant unique de l'étudiant
Gender	Catégoriel	Sexe de l'étudiant : Male / Female
Study_Hours_per_Week	Numérique	Nombre d'heures d'étude par semaine
Attendance_Rate	Numérique	Taux de présence en classe (%)
Past_Exam_Scores	Numérique	Moyenne des notes obtenues aux examens précédents
Parental_Education_Level	Catégoriel	Niveau d'éducation des parents : High School, Bachelor, Master, PhD
Internet_Access_at_Home	Catégoriel	Accès à Internet à domicile : Yes / No
Extracurricular_Activities	Catégoriel	Participation à des activités parascolaires : Yes / No
Final_Exam_Score	Numérique	Note obtenue à l'examen final (0–100)
Pass_Fail	Catégoriel	Statut final : Pass (≥ 50) / Fail (< 50)

1.3 Prétraitement des données

Afin de rendre la base exploitable pour les algorithmes de *machine learning*, plusieurs opérations de prétraitement ont été réalisées :

- **Encodage des variables catégorielles** : Les colonnes Gender, Parental_Education_Level, Internet_Access_at_Home et Extracurricular_Activities ont été transformées en variables numériques à l'aide d'un encodage basique (LabelEncoder).
- **Normalisation des variables numériques** : Les caractéristiques Study_Hours_per_Week, Attendance_Rate et Past_Exam_Scores ont été standardisées (mean = 0, std = 1) pour éviter qu'une échelle dominante n'influence la construction des modèles.
- **Contrôle qualité** : Vérification de l'absence de valeurs manquantes et de doublons : la base ne présentait ni trous ni redondances, ce qui a permis de passer immédiatement à la phase d'entraînement.

Ce prétraitement rigoureux a constitué une étape essentielle pour garantir la fiabilité des résultats générés dans les phases suivantes du projet.

1.4 Objectifs du projet

L'objectif principal de ce travail est de :

1. **Modéliser** la note finale d'un étudiant (`Final_Exam_Score`) par régression.
2. **Classer** chaque étudiant entre réussite ou échec (`Pass_Fail`).
3. **Segmenter** les profils étudiants à l'aide d'un algorithme de clustering non supervisé.

Ces trois volets permettront de mieux comprendre les leviers pédagogiques et de proposer, le cas échéant, des recommandations personnalisées.

1.5 Conclusion

Cette première étape de structuration et de prétraitement garantit une base solide pour l'expérimentation des modèles de machine learning. La clarté de la description des variables et la rigueur du nettoyage assurent la fiabilité des résultats présentés dans les chapitres suivants.

Chapitre 2

Méthodologie et algorithmes utilisés

Le projet s'inscrit dans une démarche de *data-driven decision making*, dont l'objectif est d'extraire de la valeur à partir de données éducatives afin de prédire la performance académique des étudiants.

Trois approches principales ont été explorées dans le cadre de ce projet : la régression, la classification et le clustering.

2.1 Régression

La régression sert à prédire la note finale moyenne d'un étudiant à partir de ses caractéristiques. Deux modèles principaux ont été testés :

- **Régression linéaire** : modèle simple basé sur une combinaison linéaire des variables explicatives. Il constitue un bon point de départ pour la comparaison.
- **Random Forest Regressor** : modèle d'ensemble basé sur des arbres de décision. Il permet de capturer des relations non linéaires et d'obtenir de meilleures performances.

Les performances ont été évaluées à l'aide de métriques telles que : l'erreur absolue moyenne (MAE), la racine de l'erreur quadratique moyenne (RMSE) et le coefficient de détermination (R^2).

2.2 Classification

L'objectif ici est de prédire si un étudiant réussit (Pass) ou échoue (Fail), à partir de la variable binaire Pass_Fail. Les algorithmes comparés sont :

- **Régression logistique**
- **K-Nearest Neighbors (KNN)**
- **Decision Tree Classifier**
- **Random Forest Classifier**

Les modèles ont été évalués selon les métriques classiques de classification : *accuracy*, *precision*, *recall*, *f1-score* ainsi que la matrice de confusion.

2.3 Clustering

Une analyse de regroupement non supervisée a été menée avec l'algorithme K-Means, dont l'objectif est d'identifier des profils d'étudiants selon leur engagement et leurs performances :

- **Choix du nombre de clusters** : déterminé via la méthode du coude (Elbow Method), consistant à tracer l'inertie en fonction de k et à sélectionner le point où la diminution d'inertie se stabilise.
- **Visualisation** : réduction de la dimensionnalité à deux axes par analyse en composantes principales (PCA), permettant de représenter graphiquement les clusters formés.
- **Interprétation** : description de chaque cluster selon les moyennes d'heures d'étude hebdomadaires, de taux de présence et de note finale, afin de dégager des profils pédagogiques typiques.

2.4 Conclusion

Ce chapitre a présenté les choix méthodologiques et les principaux algorithmes appliqués pour prédire la performance académique et segmenter les profils étudiants. Après avoir défini la démarche data-driven et détaillé les étapes de prétraitement, nous avons exploré trois approches complémentaires : la régression pour estimer la note finale, la classification pour distinguer réussite et échec, et le clustering pour identifier des groupes homogènes d'étudiants. Les fondements de chaque méthode et les critères d'évaluation associés (métriques de performance, validation croisée, méthode du coude, PCA) préparent le terrain pour les résultats détaillés présentés dans le chapitre suivant.

Chapitre 3

Résultats et interprétation

Dans cette section, nous présentons les résultats obtenus pour les volets « régression », « classification » et « clustering », accompagnés des graphes correspondants.

3.1 Prédiction de la note finale (Régression)

La Figure 3.1 compare les valeurs prédites par le modèle Random Forest aux valeurs réelles. On observe que la majorité des points se situent autour de la droite idéale (rouge), ce qui témoigne d'une bonne capacité du modèle à suivre la tendance des notes :

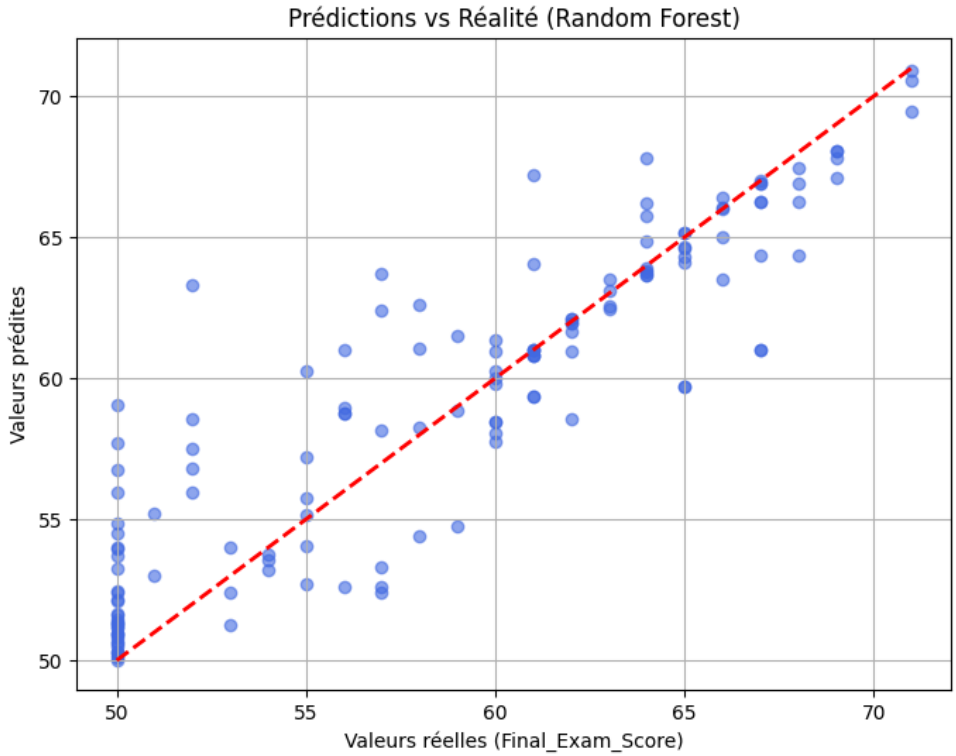

FIGURE 3.1 – Prédictions vs Réalité (Random Forest).

La Figure 3.2 présente l'importance relative (%) des variables explicatives dans la prédiction :

- **Attendance_Rate** et **Past_Exam_Scores** sont les deux variables les plus influentes (= 35 % chacune).
- **Study_Hours_per_Week** apporte également une contribution notable (= 22 %).
- Les autres variables (niveau d'éducation parentale, activités extrascolaires, genre, accès Internet) ont un impact faible (< 5 %).

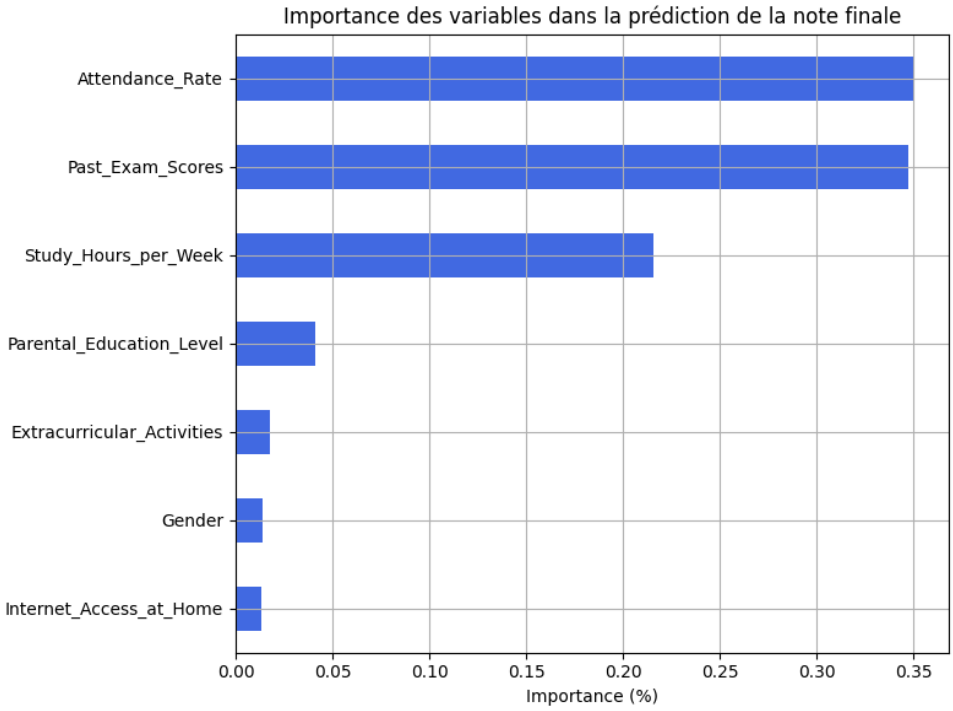

FIGURE 3.2 – Importance des variables dans la prédiction de la note finale.

Interprétation : le modèle s'appuie principalement sur l'assiduité et les performances passées, confirmant l'importance de ces facteurs pour prédire la réussite.

3.2 Prédiction Pass/Fail (Classification)

La matrice de confusion de la Figure 3.3 montre la performance du classifieur Random Forest :

- Sur 71 échecs réels, 61 sont correctement identifiés (10 faux négatifs).
- Sur 71 réussites réelles, 69 sont correctement identifiés (2 faux positifs).

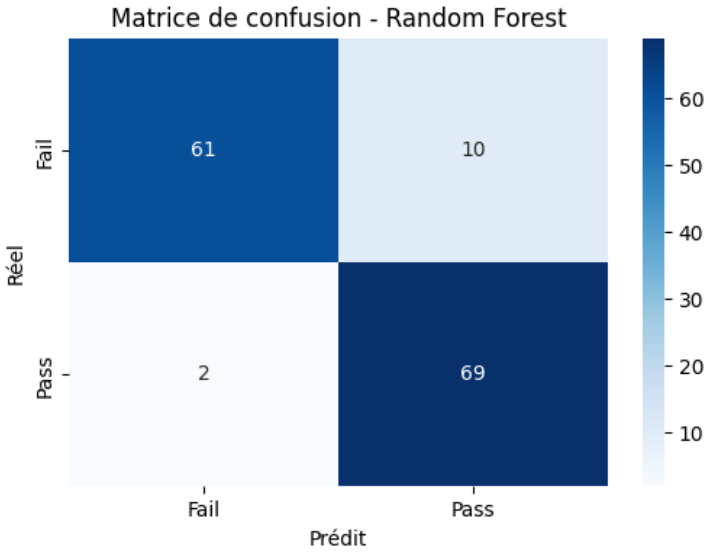

FIGURE 3.3 – Matrice de confusion – Random Forest (Pass/Fail).

La Figure 3.4 indique l'importance des mêmes variables pour la classification :

- **Past_Exam_Scores** et **Attendance_Rate** dominent (respectivement = 33 % et 31 %).
- **Study_Hours_per_Week** reste important (= 22 %).
- Les autres facteurs sont relativement secondaires.

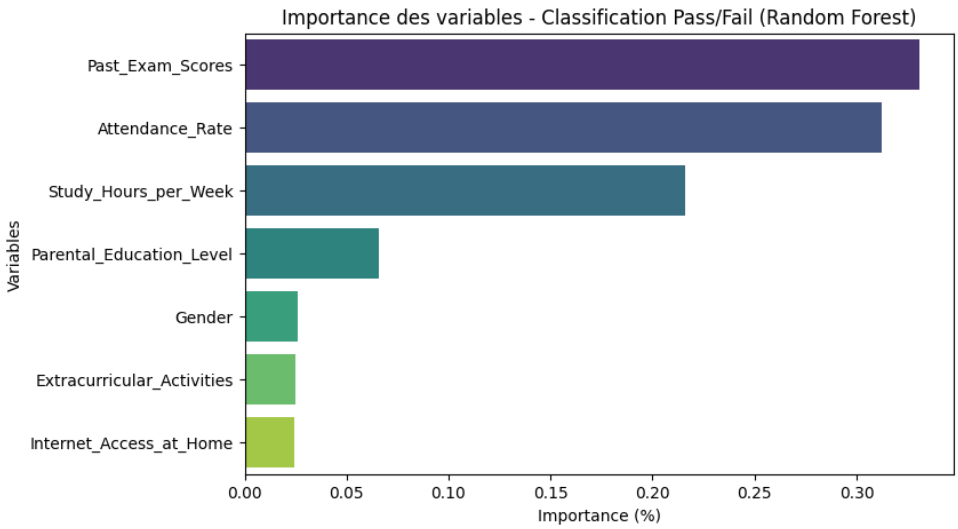

FIGURE 3.4 – Importance des variables – Classification Pass/Fail (Random Forest).

Interprétation : ces résultats confirment que les mêmes variables clés expliquent à la fois la note continue et le statut de réussite.

3.3 Segmentation des profils étudiants (Clustering)

La méthode du coude (Figure 3.5) révèle une cassure à $k = 3$, justifiant le choix de trois clusters.

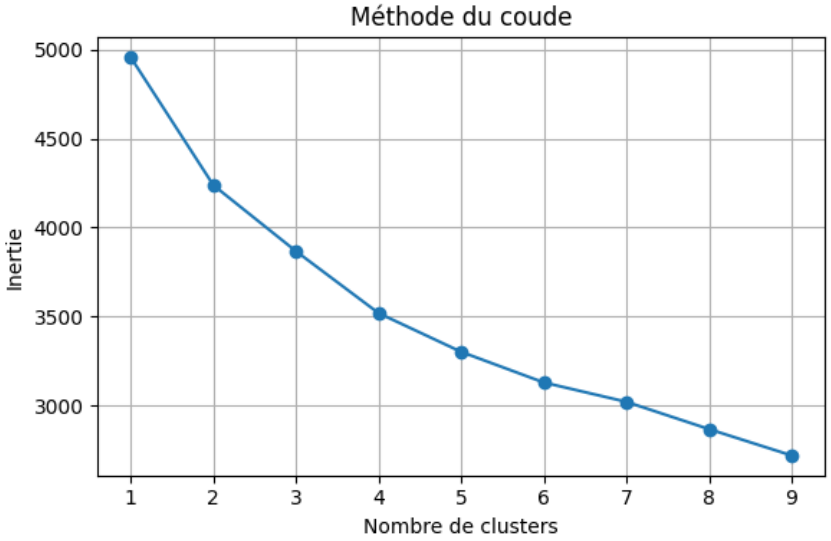

FIGURE 3.5 – Méthode du coude pour déterminer le nombre de clusters.

La projection des données sur les deux premières composantes principales, colorée par cluster (Figure 3.6), met en évidence trois groupes distincts :

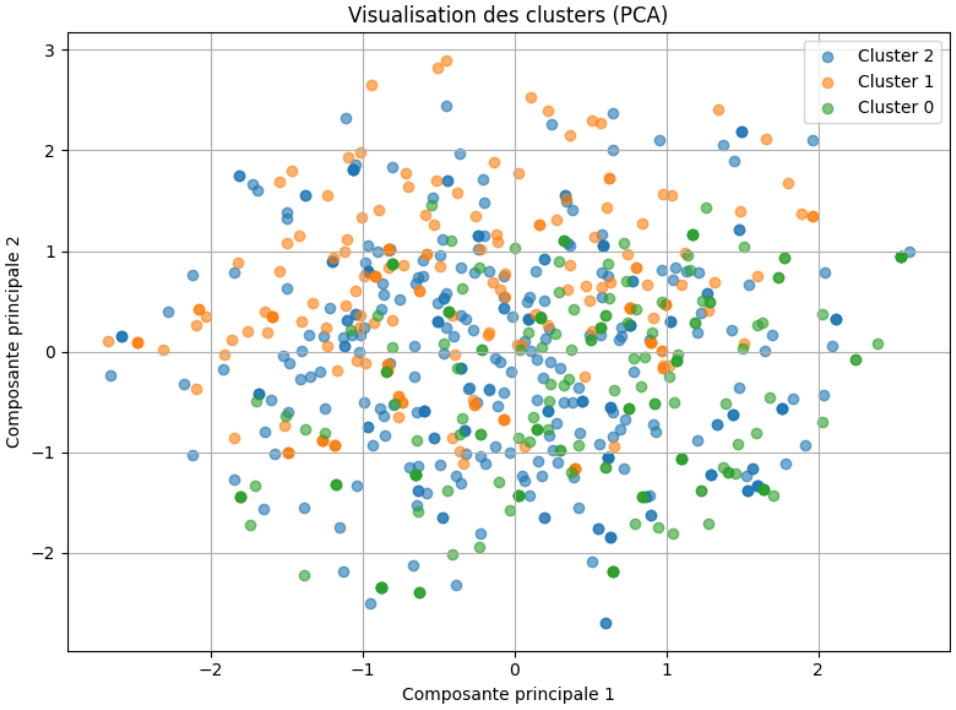

FIGURE 3.6 – Visualisation des clusters (PCA).

Interprétation :

- **Cluster 0** : profils à faibles performances (< 60 %), faible participation extrascolaire et accès Internet limité ;
- **Cluster 1** : performances intermédiaires (60–80 %), variables mixtes selon l'éducation parentale ;
- **Cluster 2** : profils à haut potentiel (> 85 %), forte assiduité et implication extrascolaire.

Ces segments permettent de proposer des actions pédagogiques différenciées (renforcement, soutien ciblé, accompagnement intensif).

Synthèse : Le recours à Random Forest pour la régression et la classification offre une excellente précision (R^2=0,89, accuracy=0,89), tandis que le clustering en trois groupes fournit une base solide pour personnaliser les stratégies d'apprentissage.

3.4 Comparaison des modèles de régression

La Figure 3.7 présente côte-à-côte les performances de la régression linéaire et du Random Forest selon trois métriques : MAE, RMSE et R^2.

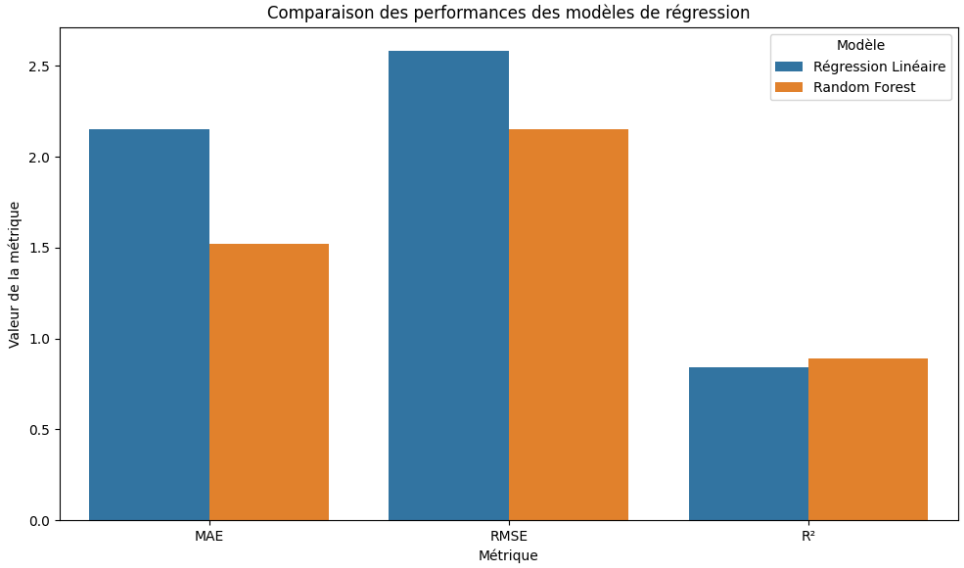

FIGURE 3.7 – Comparaison des performances des modèles de régression.

Interprétation :

- Le Random Forest affiche une MAE plus faible (= 1,52 vs 2,15) et une RMSE plus basse (= 2,15 vs 2,60),
- Le R^2 est légèrement supérieur pour Random Forest (= 0,89 vs 0,84),
- Ces écarts confirment l'avantage des méthodes d'ensemble sur la simple régression linéaire après prétraitement.

3.5 Comparaison des modèles de classification

La Figure 3.8 illustre l'accuracy et le F1-score pour la régression logistique, l'arbre de décision et le Random Forest.

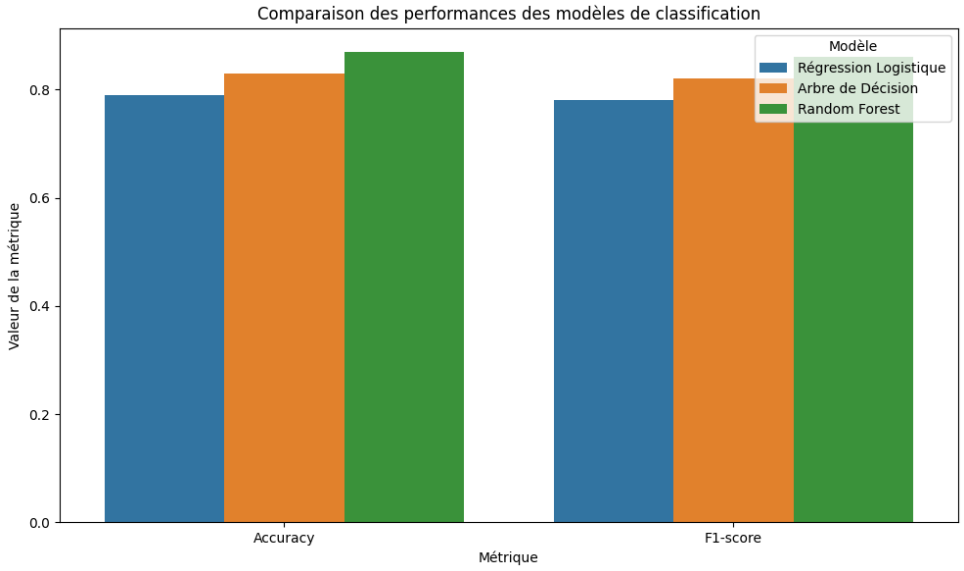

FIGURE 3.8 – Comparaison des performances des modèles de classification.

Interprétation :

- Le Random Forest atteint la meilleure accuracy (= 0,88) et le F1-score le plus élevé (= 0,86),
- L'arbre de décision se situe en intermédiaire (accuracy = 0,83, F1 = 0,82),
- La régression logistique reste performante (accuracy = 0,79, F1 = 0,78), mais moins que les méthodes basées sur des arbres.