

ARDACHAM Mahamat Teguene

**Étudiant en Master 1 MIAHS UPVM3
2025-2026**



MEMOIRE DU SEMESTRE 1

**ANALYSE DES DÉTERMINANTS SOCIO-ÉCONOMIQUES
DE LA RÉUSSITE SCOLAIRE PAR DES MÉTHODES DE
DATA SCIENCE**

**Tutrice universitaire :
SOPHIE LEBRE**

Table des matières

Remerciements.....	2
INTRODUCTION.....	3
1. CONTEXTE ET REVUE DE LA LITTERATURE.....	4
1.1 Contexte métier.....	4
1.2 Contexte Data Science.....	4
1.3 Revue bibliographique synthétique.....	5
2. DONNEES ET METHODOLOGIE GENERALE.....	5
2.1 Présentation du jeu de données	5
2.2 Description des variables	6
3. REGRESSION LINEAIRE	7
4. CLASSIFICATION SUPERVISÉE.....	9
4.1 Régression logistique.....	9
5. CLUSTERING.....	11
5.1 K-means.....	11
5.2 Classification Ascendante Hiérarchique (CAH)	12
5.3 Résultats du clustering	12
6. Discussion Générale et Limites.....	15
6.1 Limites méthodologiques.....	15
7. CONCLUSION ET PERSPECTIVES	16
8. BIBLIOGRAPHIE	16
9. ANNEXE.....	16

Remerciements

Je tiens tout d'abord à remercier l'ensemble des enseignants du Master MIASHS de l'Université Paul-Valéry Montpellier pour la qualité des enseignements dispensés tout au long de ce semestre 1 de master. Les cours suivis m'ont permis d'acquérir des bases en statistiques, en data science, indispensables à la réalisation de ce mémoire.

Ce travail de recherche m'a permis de renforcer mes compétences en analyse de données, en statistique et en machine learning appliqué aux sciences sociales. Il m'a également aidé à développer une démarche scientifique structurée, allant de la formulation d'une problématique à l'interprétation critique des résultats.

Enfin, je souhaite remercier toutes les personnes qui, de manière directe ou indirecte, ont contribué à la réalisation de ce mémoire.

INTRODUCTION

La réussite scolaire constitue un enjeu majeur pour les systèmes éducatifs, elle conditionne non seulement l'accès aux études supérieures mais également l'insertion professionnelle et la mobilité des individus. Des nombreuses études ont montré que les performances académiques ne dépendent pas uniquement de leurs capacités individuelles, mais également des facteurs socio-économiques, familiaux et institutionnels (Bourdieu & Passeron, 1970). Dans ce contexte, comprendre les déterminants de la réussite scolaire apparaît comme une problématique centrale pour les acteurs de l'éducation.

Les inégalités socio-économiques, par exemple : le niveau d'éducation des parents, le statut socio-économique ou encore l'accès à des cours de préparation d'examens jouent un rôle important et influencent significativement les performances des élèves.

Face à ces enjeux, la data sciences et les méthodes statistiques (que nous avons vues en cours) offrent des outils pour analyser les données éducatives et mettre en évidence des relations entre les caractéristiques socio-économiques des élèves et leurs résultats scolaires. Pour cela, nous allons utiliser des méthodes telles que la régression linéaire, la classification supervisée ou encore le clustering pour mesurer l'impact des certains facteurs explicatifs, prédire la probabilité de réussite d'un élève et identifier des profils à risque d'échec.

Dans ce cadre, ce mémoire s'inscrit dans une démarche d'analyse quantitative des performances scolaires à partir d'un jeu de données que nous avons téléchargé de la plateforme Kaggle, qui regroupe des informations socio-démographiques et académiques d'élèves.

La problématique de notre travail est alors la suivante :

Dans quelle mesure les facteurs socio-économiques et éducatifs, tels que le niveau d'éducation des parents et la participation à un cours de préparation, influencent-ils la réussite scolaire des élèves, et comment les méthodes de data science permettent-elles de modéliser et de prédire cette réussite ?

Nous avons trois objectifs : premièrement, analyser l'influence de certains déterminants socio-économiques sur les performances scolaires à l'aide de modèles de régression linéaire. Deuxièmement, construire des modèles de classification permettant de prédire la réussite scolaire des élèves et d'identifier ceux présentant un risque d'échec. Et enfin, par le clustering, chercher les groupes d'élèves qui présentent de profils de résultats similaires et examiner leur correspondance avec des caractéristiques sociales.

1. CONTEXTE ET REVUE DE LA LITTÉRATURE

1.1 Contexte métier

Les systèmes éducatifs sont confrontés à des inégalités qui affectent les parcours scolaires des élèves, nous allons montrer dans la partie statistique, que ces inégalités se manifestent par des écarts de performance entre élèves selon leur origine sociale, leur environnement familial ou encore leur accès aux ressources éducatives. Les élèves issus de milieux défavorisés sont en moyenne plus exposés au risque d'échec scolaire, de décrochage ou d'orientation contrainte vers des filières moins valorisés. La réussite scolaire est fortement corrélée aux conditions socio-économiques.

Les travaux fondateurs de la sociologie de l'éducation (Coleman et al., 1966) montrent que l'école, au lieu de neutraliser les inégalités d'origine sociale, peut au contraire les renforcer. C'est parce qu'elle valorise des connaissances et une culture que certaines familles possèdent déjà, mais pas toutes. Le milieu socio-économique (niveau d'études des parents, du revenu familial, ...) joue donc un rôle très important dans la réussite à l'école.

Les enfants dont les parents ont beaucoup étudié ont un avantage car ils bénéficient généralement d'un accompagnement pour leurs devoirs et sont plus encouragés à réussir. Ces différences entre les élèves apparaissent dès les premières classes et elles ont malheureusement tendance à durer tout au long de la scolarité.

Pour compenser ces inégalités, les gouvernements ont créé plusieurs politiques éducatives pour soutenir les élèves qui en ont le plus besoin. Cela comprend par exemple les repas gratuits ou moins chers à la cantine, le soutien scolaire gratuit et des cours spécifiques pour s'entraîner aux examens. Ces politiques ont des effets positifs et fonctionnent bien, surtout pour les élèves les plus fragiles. Par exemple, les cours de préparation aux tests permettent souvent d'obtenir de meilleures notes, même si cela dépend beaucoup de la qualité des cours et de la motivation de l'élève (Briggs, 2001).

1.2 Contexte Data Science

La data science et la statistique occupent une place croissante dans le domaine de l'éducation, notamment à travers l'analyse des données scolaires. Elles nous permettent de créer des modèles pour expliquer pourquoi certains élèves réussissent et d'autres non, d'identifier les facteurs explicatifs de la réussite ou de l'échec académique. On utilise souvent une méthode de la régression linéaire car elle est facile à comprendre et permet de mesurer précisément l'influence de chaque détail sur les notes finales.

Par les chiffres, nous arriverons à avoir une vision globale et objective des faits réels plutôt que sur de simples impressions. Cela aidera par exemple les responsables de l'éducation à prendre les meilleures décisions et à savoir exactement où agir pour aider le plus grand nombre d'élèves.

Au-delà de l'analyse explicative, les méthodes de data science permettent également de développer des modèles prédictifs. On utilise des techniques de classification supervisée comme la régression logistique pour estimer la probabilité de réussite ou d'échec d'un élève selon son profil. L'avantage concret est que l'on peut repérer très tôt les élèves qui risquent de

rencontrer des difficultés. En anticipant les difficultés scolaires, les écoles peuvent mettre en place des actions de prévention ciblées afin d'éviter que les concernés n'abandonnent leurs études et les accompagner vers un meilleur parcours scolaire.

1.3 Revue bibliographique synthétique

- Niveau d'éducation des parents

C'est l'un des facteurs les plus déterminants pour expliquer les notes des enfants. De nombreuses études (Sirin, 2005 ; Davis-Kean, 2005), prouvent que plus les parents sont diplômés, mieux les enfants réussissent à l'école car ces derniers partagent leurs connaissances avec leurs enfants et savent mieux les guider dans leur parcours scolaire.

- Cours de préparation aux tests

Suivre des cours spéciaux pour s'entraîner aux examens aide beaucoup à améliorer les notes, surtout en mathématiques et en lecture. Le problème, c'est que ces cours coûtent souvent cher, ce qui peut avantager uniquement les familles favorables. Pour corriger cela, certaines politiques publiques proposent ces cours gratuitement aux élèves défavorisés afin de compenser les désavantages socio-économiques initiaux.

- Performances académiques

Pour savoir si un élève réussit, on regarde ses notes en maths, en lecture et en écriture même si les compétences sont différentes. En utilisant des méthodes de clustering et de classification, on peut regrouper les élèves par profils pour comprendre pourquoi certains réussissent partout et pourquoi d'autres ont plus de mal, de voir le lien direct entre les profils et les notes obtenues.

2. DONNEES ET METHODOLOGIE GENERALE

2.1 Présentation du jeu de données

Les données utilisées dans ce mémoire proviennent de la plateforme Kaggle. Le jeu de données est intitulé *Students Performance Dataset (Cleaned)* et regroupe des informations relatives aux performances scolaires d'élèves ainsi qu'à leurs caractéristiques socio-démographiques et éducatives. Nous tenons à rappeler que le jeu de données est anonyme, librement accessible et couramment utilisé à des fins pédagogiques et de recherche.

Le jeu de données contient des informations sur les résultats académiques d'élèves évalués dans trois disciplines fondamentales : les maths, la lecture et l'écriture. En complément, nous disposons de plusieurs variables telles que le genre, l'origine ethnique, le niveau d'éducation des parents, le type de repas ou encore la participation à un cours de préparation aux tests.

Il est composé de 1000 observations correspondant à 1000 élèves et de 10 variables. Chaque ligne représente un élève et chaque colonne correspond à une caractéristique individuelle ou à un indicateur de performance scolaire.

Dans un souci de lisibilité et de cohérence avec le cadre académique du mémoire, nous avons renommé les variables en français.

Les variables qualitatives codées initialement en binaire, telles que le genre, le type de repas, la participation à un cours de préparation ont été transformées en factor numériques. La variable de niveau d'éducation des parents en ordinale.

Dans le cadre de la classification supervisée, nous avons créé une variable binaire **Succès**, elle prend la valeur 1 si le score moyen d'un élève est supérieur ou égal à 70, seuil que nous avons défini pour caractériser une réussite scolaire satisfaisante, et la valeur 0 dans le cas contraire. Alors pourquoi créer cette variable ? Pour simplement formuler un problème de prédiction de la réussite scolaire et d'identifier les élèves présentant un risque d'échec.

2.2 Description des variables

Nous avons des variables socio-démographiques :

Genre : codée binaire, avec la valeur 0 correspondant au sexe féminin et 1 au sexe masculin.

Race_Ethnique : permet de distinguer différents groupes raciaux, elle nous offre alors la possibilité d'analyser peut-être des disparités de performance entre groupes. On a 5 groupes étiquetés A, ..., E.

Nous n'avons pas en aucun cas le droit d'assimiler ces groupes à des catégories raciales précises. En France, interpréter de données à caractère racial est strictement réglementé, même à titre d'exemple, les conclusions pourraient être perçues stigmatisantes.

Niveau d'Éducation des Parents (NEP) : c'est le diplôme le plus élevé atteint par les parents.

Repas : indique si l'élève bénéficie d'un repas gratuit ou à tarif réduit.

Cours de Préparation aux Tests (CPT) : précise si l'élève a suivi ou non un dispositif de préparation académique.

Nous avons des variables académiques compris entre 0 et 100 : le **Score en Mathématiques, en Lecture et en Écriture** qui mesurent les performances scolaires des élèves.

Afin de disposer d'un indicateur qui synthétise la réussite scolaire, une variable **Score Moyen** a été calculée comme la moyenne arithmétique des trois scores. Elle va nous permettre voir la performance globale de chaque élève et constitue la variable dépendante (VD) principale dans les analyses de régression et de classification.

3. REGRESSION LINEAIRE

L'objectif de cette partie est d'étudier l'influence des facteurs socio-économiques sur la réussite scolaire des élèves, mesurée par le **Score_Moyen**. Plus précisément, nous cherchons à mesurer l'effet de la participation à un **cours de préparation aux tests (CPT)** et le **niveau d'éducation des parents (NEP)** tout en contrôlant pour des variables démographiques et socio-économiques telles que le genre, le type de repas (gratuit/réduit vs standard), et la race ethnique.

- La régression linéaire multiple

Elle permet d'exprimer le **Score_Moyen** comme une combinaison linéaire de plusieurs variables explicatives. Le modèle complet est :

```
Call:
lm(formula = Score_Moyen ~ CPT + NEP + Genre + Repas + Race_Ethnique,
    data = JD)

Residuals:
    Min       1Q   Median       3Q      Max
-48.148  -8.298   0.646   8.736  27.522

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    58.6335     1.5279  38.375 < 2e-16 ***
CPTOui         7.6386     0.8302   9.201 < 2e-16 ***
NEP.L          8.0335     1.2523   6.415 2.18e-10 ***
NEP.Q          0.4482     1.1608   0.386 0.699477
NEP.C         -1.0812     1.0946  -0.988 0.323492
NEP^4          1.0598     1.0057   1.054 0.292248
NEP^5         -1.2998     0.8830  -1.472 0.141344
GenreMasculin  -3.7242     0.7955  -4.682 3.24e-06 ***
RepasStandard  8.7751     0.8275  10.605 < 2e-16 ***
Race_Ethniquegroup B  1.5290     1.6116   0.949 0.342983
Race_Ethniquegroup C  2.3855     1.5093   1.581 0.114296
Race_Ethniquegroup D  5.1258     1.5398   3.329 0.000904 ***
Race_Ethniquegroup E  6.9285     1.7081   4.056 5.38e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.49 on 987 degrees of freedom
Multiple R-squared:  0.2423,    Adjusted R-squared:  0.2331
F-statistic: 26.3 on 12 and 987 DF,  p-value: < 2.2e-16
```

Tableau 3.1 : Résumé du modèle de régression linéaire multiple

Le **R² ajusté = 0.233** indique que 23% de la variance des scores moyens est expliquée par les variables incluses. Le modèle est hautement significatif (F-statistic = 26.3, $p < 0.001$).

On remarque que les élèves ayant suivi un cours de préparation obtiennent en moyenne **7,6** points de plus que les autres, un niveau d'éducation parental plus élevé est associé à un score moyen supérieur. Les élèves avec repas standard réussissent en moyenne **8,8** points de plus que ceux avec repas gratuit/réduit et que certains groupes ethniques obtiennent des scores significativement plus élevés.

Aussi, les garçons obtiennent en moyenne **3,7** points de moins que les filles.

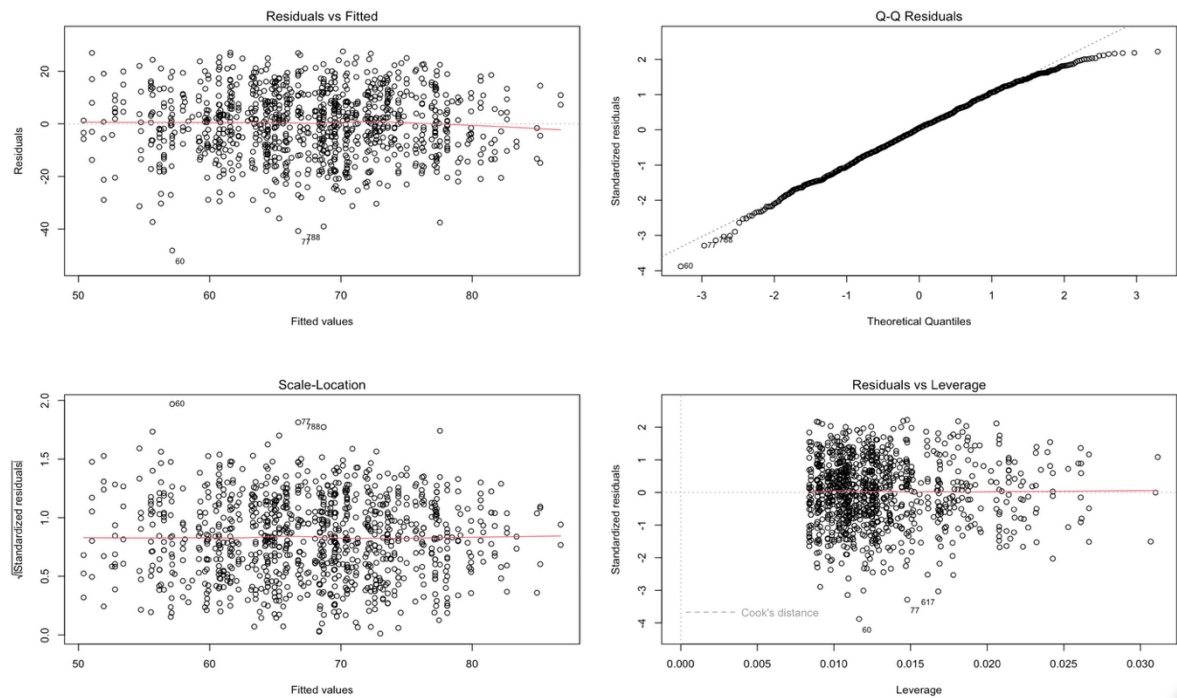


Figure 3.2 : Diagnostics graphiques du modèle de régression linéaire multiple

Score moyen selon la participation à un cours de préparation

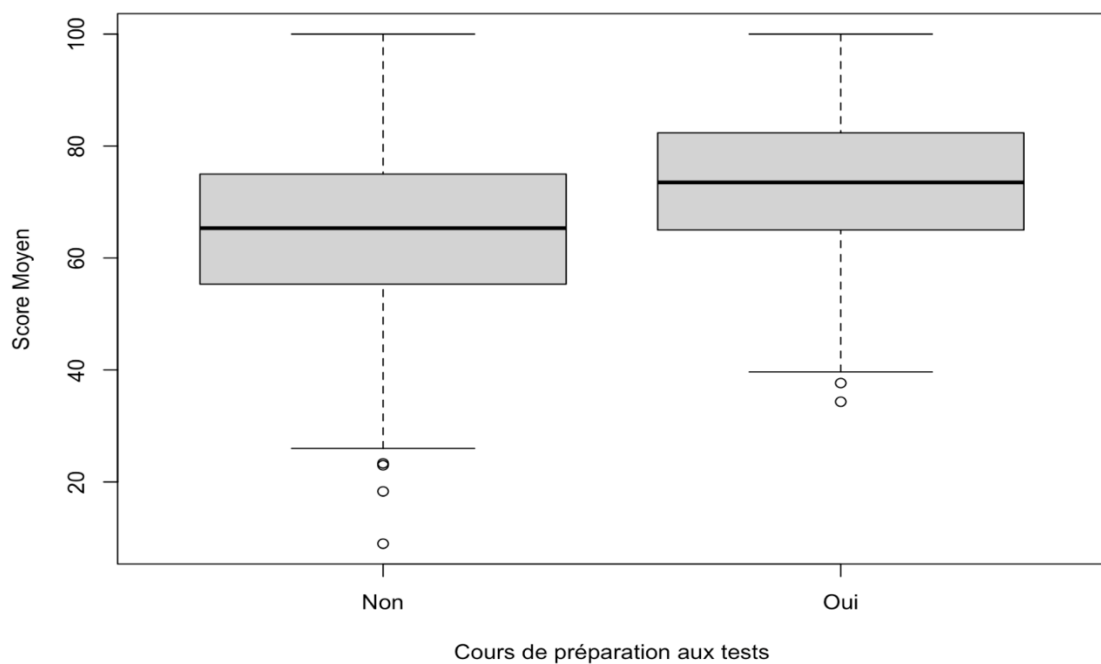


Figure 3.3 : Effet du (CPT) sur le score moyen – Boxplot comparatif

La préparation aux tests améliore la performance.

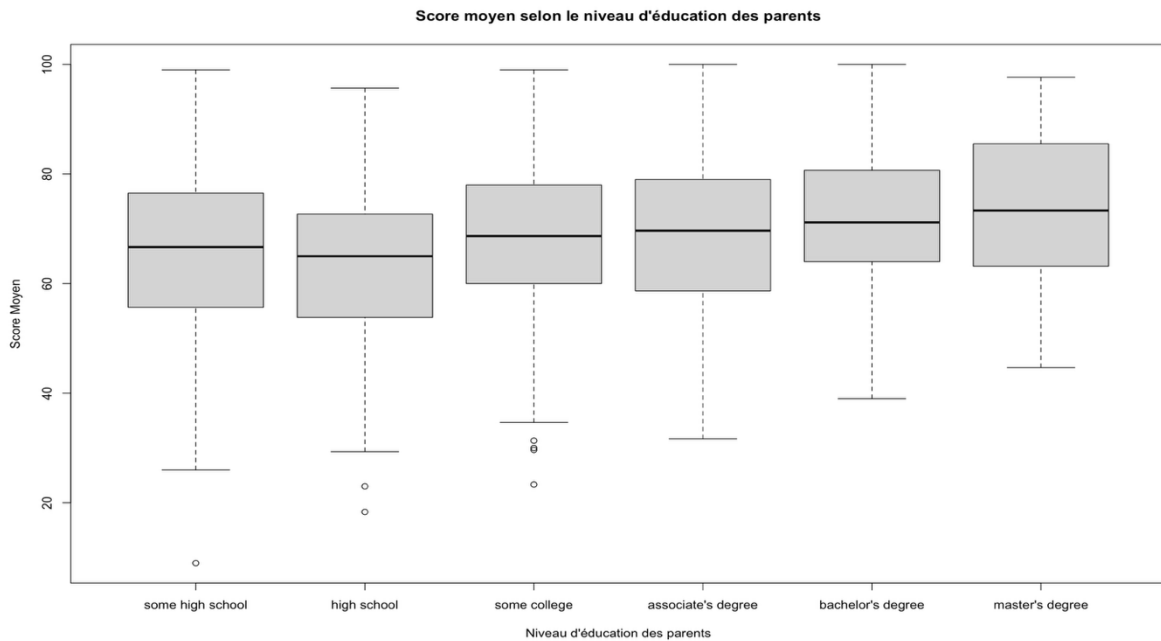


Figure 3.4 : Relation entre le (NEP) et le score moyen – Boxplot par niveau

Les boxplots confirment une progression régulière du score moyen avec le NEP.

- Sélection de variables

Pour vérifier la robustesse du modèle, nous avons appliqué différentes méthodes de sélection :

Backward et Forward : même modèle final que le modèle complet.

AIC/BIC identiques : confirme que le modèle complet est déjà optimal selon ces critères.

Cp de Mallows : Les variables retenues sont (*CPT, NEP, Repas, Race_Ethnique D/E, Genre*) correspondent exactement au modèle backward. Le modèle est bien ajusté.

4. CLASSIFICATION SUPERVISÉE

Notre objectif dans cette partie est de prédire la réussite des élèves en utilisant leurs caractéristiques socio-économiques et académiques. Nous avons défini la variable binaire **Succès** à partir du score moyen des élèves : $Score_Moyen \geq 70$ "Succès" sinon "Échec".

La distribution montre 541 élèves en échec et 459 en réussite, une répartition relativement équilibrée, et donc adaptée à la classification supervisée.

Nous avons utilisé deux approches :

4.1 Régression logistique

Le modèle logit estime la probabilité de réussite en fonction des variables socio-économiques et démographiques :

```
Call:
glm(formula = Succes ~ CPT + NEP + Genre + Repas + Race_Ethnique,
     family = binomial, data = JD)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.6623     0.2950  -5.635 1.75e-08 ***
CPTOui          1.0828     0.1479   7.322 2.45e-13 ***
NEP.L           0.8789     0.2215   3.968 7.24e-05 ***
NEP.Q           0.1318     0.2040   0.646 0.518202
NEP.C          -0.1539     0.1933  -0.797 0.425741
NEP^4           0.1403     0.1778   0.789 0.430099
NEP^5          -0.1962     0.1546  -1.269 0.204274
GenreMasculin  -0.5786     0.1408  -4.111 3.95e-05 ***
RepasStandard   1.1186     0.1516   7.379 1.59e-13 ***
Race_Ethniquegroup B  0.4365     0.3006   1.452 0.146514
Race_Ethniquegroup C  0.6293     0.2833   2.221 0.026339 *
Race_Ethniquegroup D  0.9962     0.2883   3.456 0.000548 ***
Race_Ethniquegroup E  1.3272     0.3160   4.200 2.67e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1379.6  on 999  degrees of freedom
Residual deviance: 1201.9  on 987  degrees of freedom
AIC: 1227.9

Number of Fisher Scoring iterations: 4
```

Tableau 4.1.1 : Modèle de régression logistique expliquant la probabilité de succès

Les élèves ayant suivi un cours de préparation ont une probabilité de succès élevée, donc plus de chances de réussir que les autres. Ceux avec repas standard ont plus de chance de réussir. On remarque aussi que certains groupes ethniques ont une probabilité de succès significativement plus élevée, par exemple D et E.

Les garçons ont moins de chances de réussir que les filles. Et enfin l'effet linéaire significatif du niveau parental montre qu'un niveau plus élevé augmente les chances de succès.

AIC = 1227.9, résidu déviance = 1201.9 : modèle bien ajusté pour une classification binaire.

- Prédictions, Matrice de confusion et ROC / AUC

On fixe une règle, si la probabilité ≥ 0.5 on classe l'élève dans la catégorie "Succès". Si c'est moins, on le classe en "Échec". La matrice de confusion donne : 384 (TN), 288 (TP), 171 (FN) et 157 (FP).

Accuracy = 0.672, raisonnable pour un modèle utilisant uniquement les variables socio-économiques et démographiques.

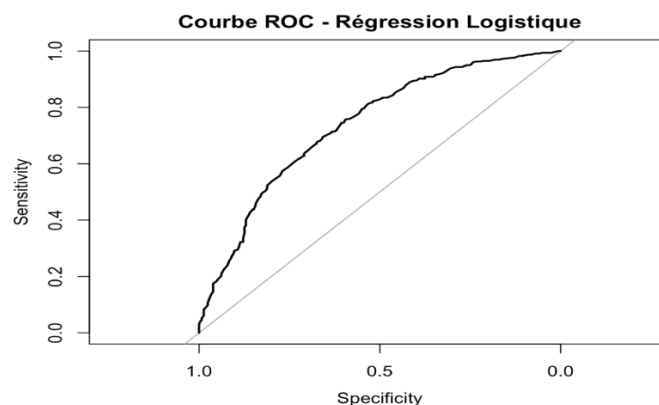


Figure 4.1.2 : Courbe ROC du modèle de régression logistique pour la prédiction du succès

AUC, l'aire sous la courbe mesure la capacité globale du modèle à séparer les classes. Nous avons obtenu $AUC = 0.737$, le modèle distingue correctement succès/échec dans 73.7% des cas, ce qui est bon pour des données réelles en éducation.

Enfin, nous avons réalisé un tableau (tri par probabilité prédite de succès régression logistique) identifiant les 10 élèves les plus à risque (probabilité de succès < 0.09).

Description: df [10 x 7]

	Genre <fctr>	NEP <ord>	Repas <fctr>	CPT <fctr>	Race_Ethnique <fctr>	Score_Moyen <dbl>	Prob_Succes <dbl>
396	Masculin	high school	Gratuit_ou_reduit	Non	group A	44.66667	0.05775186
689	Masculin	high school	Gratuit_ou_reduit	Non	group A	51.66667	0.05775186
812	Masculin	high school	Gratuit_ou_reduit	Non	group A	47.00000	0.05775186
62	Masculin	some high school	Gratuit_ou_reduit	Non	group A	37.33333	0.06926455
229	Masculin	some high school	Gratuit_ou_reduit	Non	group A	68.00000	0.06926455
429	Masculin	some high school	Gratuit_ou_reduit	Non	group A	59.00000	0.06926455
445	Masculin	some high school	Gratuit_ou_reduit	Non	group A	78.00000	0.06926455
732	Masculin	some high school	Gratuit_ou_reduit	Non	group A	48.00000	0.06926455
82	Masculin	high school	Gratuit_ou_reduit	Non	group B	46.33333	0.08662091
219	Masculin	high school	Gratuit_ou_reduit	Non	group B	71.00000	0.08662091

10 rows

Tableau 4.1.3 : Les 10 élèves présentant la plus faible probabilité prédite de succès scolaire

Le tableau montre tous des garçons, repas gratuit ou réduit, faible NEP, n'ayant pas suivi de CPT. Les groupes ethniques A/B sont surreprésentés. Les 10 élèves les plus à risque ont des probabilités de succès entre 5% et 9%, ce qui confirme que la combinaison de faibles scores académiques et contexte socio-économique défavorable identifie les élèves vulnérables.

Nous nous sommes demandé pourquoi l'élève n°219 est classé avec les plus à risque alors que son $Score_Moyen = 71 > 70$ (seuil). Cela semble contradictoire, mais ce ne l'est pas.

La logistique répond à la question : “Compte tenu du profil socio-économique de l'élève, quelle est la probabilité qu'il réussisse ?”. Elle ne regarde pas le $Score_Moyen$, ce dernier n'est pas une variable explicative dans le `modele_logit`, le modèle apprend des profils socio-économiques.

Cet élève a réussi (atypique) malgré un profil à risque, le modèle considère que sa réussite est peu probable, mais pas impossible.

5. CLUSTERING

L'objectif de cette partie est d'identifier des profils similaires d'élèves à partir de leurs performances académiques et de leurs caractéristiques socio-économiques, sans utiliser a priori la variable de réussite. Contrairement aux approches supervisées présentées précédemment, le clustering permet d'explorer la structure intrinsèque des données et de révéler des groupes latents d'élèves.

5.1 K-means

La méthode K-means consiste à partitionner les observations en k groupes disjoints en minimisant l'inertie intra-classe, mesurée par la somme des distances au centroïde de chaque cluster. Les variables académiques continues $Score_Maths$, $Score_Lecture$, $Score_Ecriture$, $Score_Moyen$ ont été préalablement standardisées afin d'éviter qu'une variable ne domine artificiellement la distance euclidienne.

Le choix du nombre optimal de clusters (k) n'a pas été fixé arbitrairement. Il a été déterminé à l'aide de la méthode du coude et l'indice de silhouette moyenne.

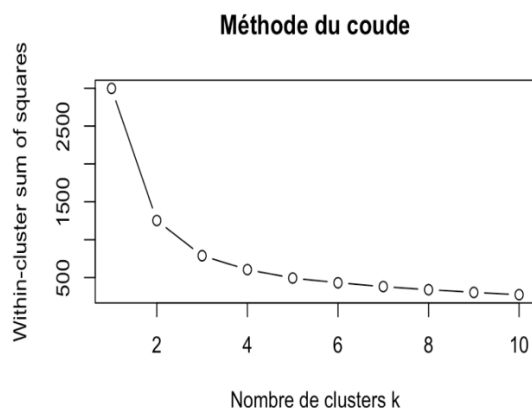


Figure 5.1.1 : Méthode du coude

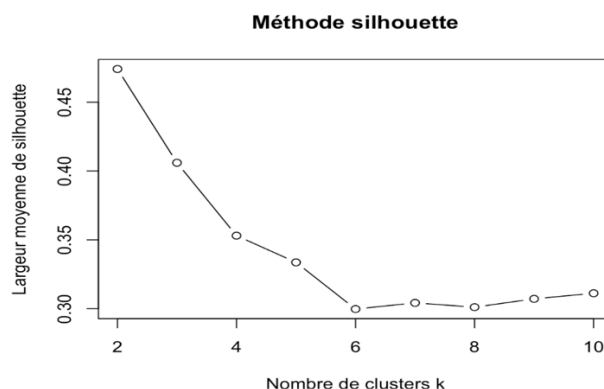


Figure 5.1.2 : Méthode de la silhouette

Nous avons obtenu $k = 2$, indiquant l'existence de deux groupes bien distincts au sein de la population étudiée.

5.2 Classification Ascendante Hiérarchique (CAH)

En complément, une CAH a été réalisée à partir d'une matrice de distances euclidiennes, en utilisant la méthode de Ward, qui vise à minimiser la variance intra-classe à chaque fusion. L'analyse du dendrogramme confirme également une coupure naturelle en deux classes, ce qui renforce la robustesse du choix de $k = 2$ retenu pour l'analyse finale.

Ainsi, les deux méthodes non supervisées aboutissent à une segmentation cohérente et stable des élèves.

5.3 Résultats du clustering

Cluster 1 : Élèves à performance élevée présentant une performance globalement homogène.

Cluster 2 : Élèves à performance plus faible.

Description: df [2 x 5]

Cluster <fctr>	Score_Maths <dbl>	Score_Lecture <dbl>	Score_Ecriture <dbl>	Score_Moyen <dbl>
1	75.81786	79.33929	78.51250	77.88988
2	53.70682	56.22500	54.74318	54.89167

2 rows

Tableau 5.3.1 : Caractéristiques moyennes des deux clusters d'élèves selon les scores

Nous avons réalisé un Scatter plot 3D (Maths, Lecture, Écriture) et Scatter plot matrix, on remarque que le premier cluster (en rouge) occupe une zone de l'espace qui correspond à des scores élevés sur les trois axes, contrairement au second à des scores faibles.

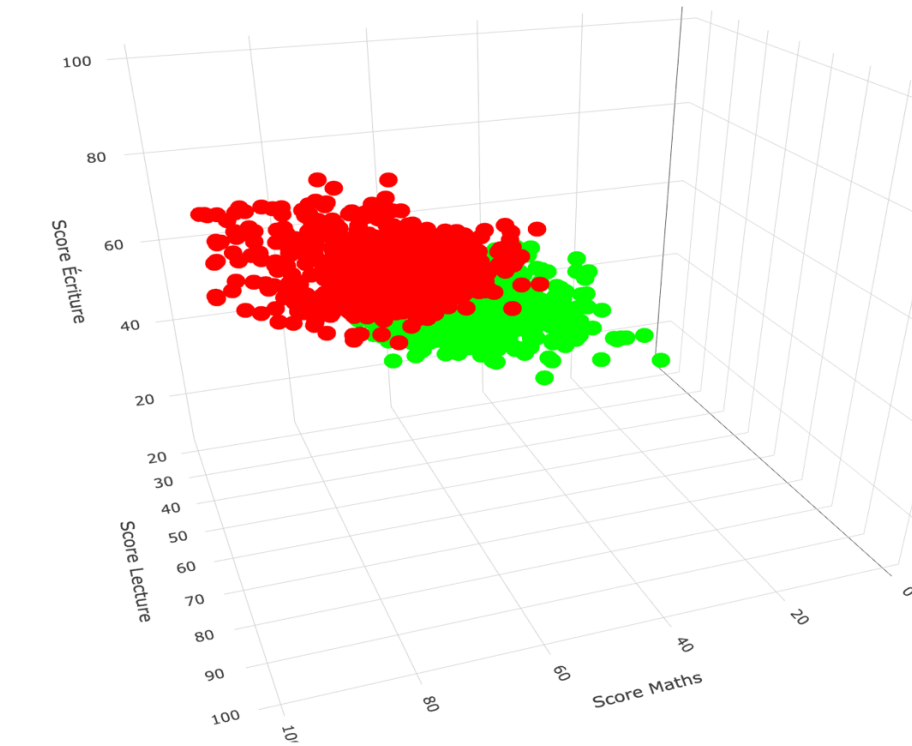


Figure 5.3.2 : Visualisation 3D des clusters d'élèves selon les scores

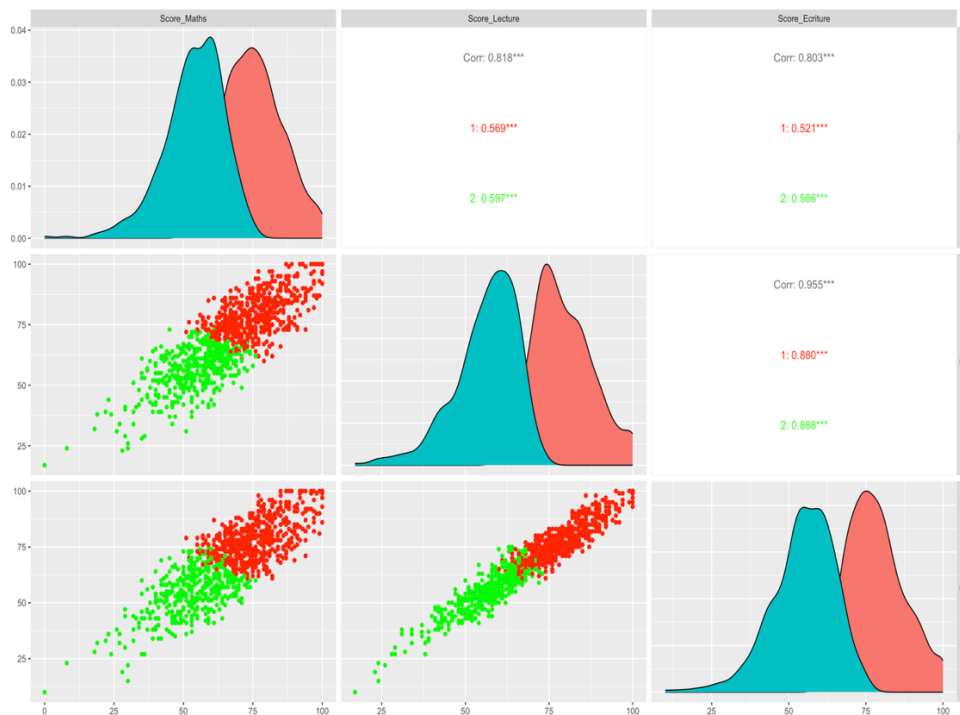


Figure 5.3.3 : Matrice de dispersion des scores académiques par cluster

Ensuite, nous avons ensuite affiché des histogrammes et le dendrogramme :

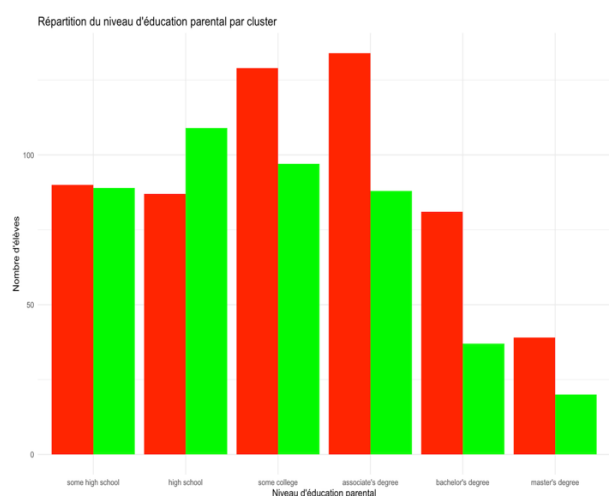


Figure 5.3.4 : NEP selon les clusters

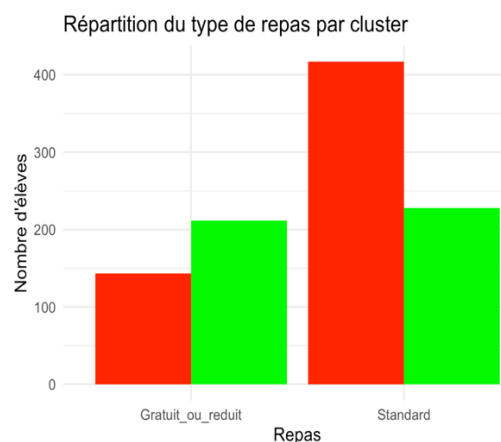


Figure 5.3.5 : Repas selon les clusters

Dans la Figure 5.3.4, le cluster 1 regroupe davantage d'élèves dont les parents possèdent un niveau d'éducation élevé (bachelor, master). Le groupe 2 est surreprésenté par des élèves issus de familles dont le niveau d'éducation est plus faible (some high school, high school).

Dans la Figure 5.3.5, les élèves qui bénéficient de repas standards sont majoritairement concentrés du cluster 1, tandis que le 2 présente une proportion plus élevée d'élèves bénéficiant de repas gratuits ou à tarif réduit, indicateur d'un statut socio-économique plus défavorisé.

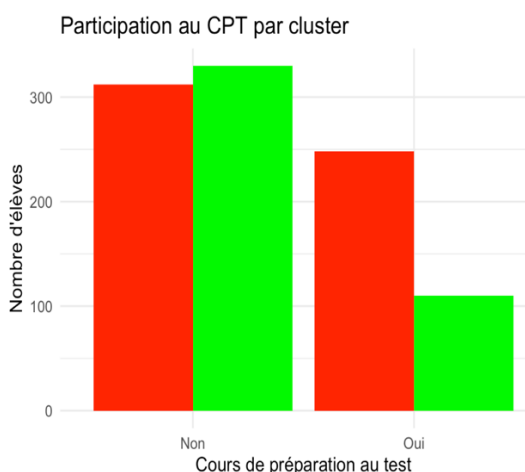


Figure 5.3.6 : CPT selon les clusters

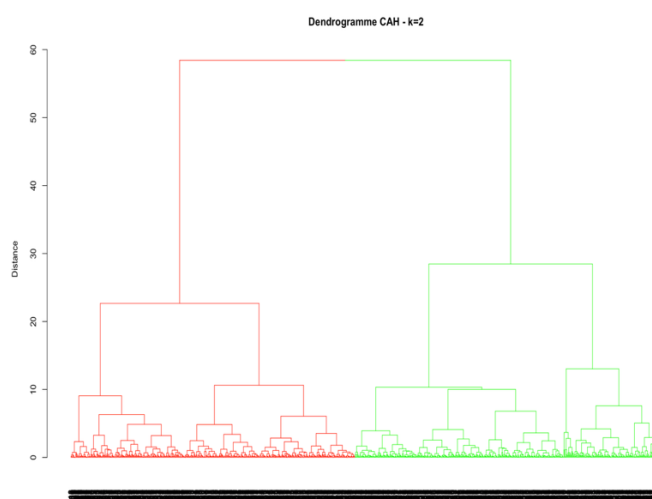


Figure 5.3.7 : Dendrogramme (CAH)

Le recours aux CPT est plus fréquent dans le cluster des élèves performants. Le cluster 2 comprend une majorité d'élèves n'ayant pas suivi de cours de préparation (Figure 5.3.6).

Les résultats sont cohérents avec les analyses supervisées présentées précédemment. D'une part, les variables identifiées comme significatives dans la régression linéaire (le niveau d'éducation des parents et la participation aux cours de préparation) sont également celles qui

discriminent le plus fortement les clusters. D'autre part, le cluster 2 correspond aux profils identifiés comme élèves à risque dans les modèles de classification supervisée.

Ainsi, la classification non supervisée confirme l'existence de deux grands profils d'élèves : d'une part, un profil favorisé, cumulant de meilleures performances académiques et des ressources socio-économiques plus élevées et d'autre part, un profil plus vulnérable, marqué par des difficultés scolaires et un environnement socio-économique moins favorable.

6. Discussion Générale et Limites

Les analyses que nous avons réalisé dans ce mémoire nous ont permis d'étudier l'influence socio-économique et académique sur les performances scolaires des élèves.

La régression linéaire multiple a montré que la participation à un cours de préparation aux tests (CPT), le niveau d'éducation parental (NEP) et d'autres facteurs comme le type de repas influencent les performances.

La classification supervisée (logistique) nous a permis d'identifier les élèves à risque (**AUC = 0,737**). Les élèves à risque partagent des caractéristiques socio-économiques similaires : NEP faible, repas gratuit ou réduit, faible participation au CPT.

Le clustering non supervisé (K-means et CAH) nous a permis enfin de diviser les élèves en deux clusters distincts :

1. Cluster performant : NEP élevé, repas standard, CPT suivi, scores élevés.
2. Cluster à risque : NEP faible, repas gratuit, faible participation CPT, scores faibles.

En somme, la régression explique l'effet des variables, la classification prédit la réussite, et le clustering permet de visualiser les profils types d'élèves.

6.1 Limites méthodologiques

Données : le dataset contient 1000 observations donc on ne généralise pas des résultats.

Variables socio-économiques : d'autres facteurs (revenu familial, contexte scolaire, ...) auraient enrichi l'analyse.

Modèles : la régression suppose une relation linéaire, la logistique un lien logit, et K-means des clusters sphériques. D'autres méthodes pourraient mieux améliorer la précision.

7. CONCLUSION ET PERSPECTIVES

Ce mémoire a démontré que la réussite scolaire des élèves est fortement influencée par leur environnement socio-économique et par leur participation à des cours de préparation. Les méthodes de data science appliquées qu'on a utilisées : régression (a identifié les facteurs significatifs), classification (a permis de prédire la réussite et d'identifier les élèves à risque) et clustering (a mis en évidence les profils types d'élèves et leur distribution socio-économique), se sont révélées complémentaires et cohérentes.

Nos analyses fournissent un outil pratique pour orienter les politiques et les interventions éducatives, nous l'espérons.

Pour les perspectives, inclure d'autres variables socio-économiques, comportementales ou psychologiques pour enrichir l'analyse dans les données. Tester des modèles non linéaires qu'on a vu en cours de Régularisation et Optimisation (Random Forest, ...) pour la prédiction.

8. BIBLIOGRAPHIE

Bourdieu, P., & Passeron, J.-C. (1970). *La reproduction. Éléments pour une théorie du système d'enseignement*. Paris : Éditions de Minuit.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of Educational Opportunity*. Washington, DC: U.S. Government Printing Office.

Briggs, D. C. (2001). The effect of admissions test preparation: Evidence from NELS:88. *Chance*, 14(1), 10–18.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, 75(3), 417–453.

Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement. *Journal of Family Psychology*, 19(2), 294–304.

<https://www.kaggle.com/datasets/muhammadroshaanriaz/students-performance-dataset-cleaned/data>

9. ANNEXE

- Code R complet
- Sorties des modèles
- Graphiques
- Résultats
- Détails méthodologiques