# Analyse de la performance scolaire de lycéens - Printemps 2025

COMBELLE Matthieu - JARDRI Rosalie - SCALABRE Matteo*

25/08/2025

**Résumé**

A partir de l'analyse du notre jeu de données, nous avons déterminé et évalué l'influence des facteurs socioculturels sur la performance scolaire. L'analyse en composante principale révèle l'importance de l'absentéisme, du score GPA, du temps de travail hebdomadaire et du support parental dans cette réussite. La classification ascendante hiérarchique permet de rendre compte de groupements spontanés en fonction des notes, une construction obtenue sous l'influence particulière du temps d'étude hebdomadaire et de taux d'absence de l'élève. Plusieurs méthodes supervisées ont ensuite été utilisées pour construire des classifieurs à 5 variables (pour chaque note possible de l'élève : A, B, C, D ou F). Les méthodes d'analyse discriminante, de régression logistique ordinale et de renforcement du gradient fournissent des résultats satisfaisants. Ainsi, nos analyses sur ce jeu de données montre que la performance scolaire est principalement liée à 4 variables : (i) le taux d'absence, (ii) le temps de travail hebdomadaire, (iii) l'existence d'un tutorat, et (iv) le support parental.

## Introduction

Afin de mener à bien ce projet de SY09, nous avons sélectionné un jeu de données sur lequel nous pouvions appliquer des méthodes de visualisation, de clustering et de classification supervisée. Suite à une analyse exploratoire préliminaire, nous avons décidé de travailler avec le jeu de données suivant : Students Performance Dataset - Academic Success Factors in High School Students. Nous nous sommes alors posé la question suivante : Dans quelle mesure les éléments sociodémographiques (ie. support des parents, niveau d'éducation des parents, origine ethnique, genre) influencent-ils la performance scolaire des lycéens ?

Une fois le jeu de données choisi, nous avons affiné notre analyse en nettoyant et en corrigeant certaines données altérées, puis nous avons appliqué des méthodes de classification (partie 2) et d'apprentissage pour répondre à notre problématique et identifier des profils 'types' de lycéens (partie 3).

*Auteurs par ordre alphabétique

## 1 Sélection du jeu de données

Cette première partie précise nos critères de choix du dataset ainsi que les analyses préliminaires effectuées.

### 1.1 Critères de sélections

Après avoir établi des critères qualitatifs (eg. description des variables) et quantitatifs de sélections (eg. taille de l'échantillon, diversité de données), nous avons utilisé des *méthodes de visualisation* (eg. scatterplots, heatmaps) et des *méthodes de classification* (comme l'Analyse en Composantes Principales ou ACP et la Classification Ascendante Hiérarchique ou CAH). Nos critères de sélection étaient les suivants : (i) l'existence de plusieurs corrélations entre variables ; (ii) une ACP avec une variance expliquée répartie sur plus de deux axes ; (iii) une CAH distinguant au minimum 3 classes.

### 1.2 Présentation du jeu de données choisi

Nous avons finalement choisi le jeu de données présenté dans l'introduction [3]. Il s'agit d'un *ensemble de données générées synthétiquement à objectif pédagogique*. Il décrit les résultats académiques de *2392* étudiants, en fonction d'un certain nombre de variables (n=15), toutes numériques.

Plus précisement, chaque individu est caractérisé par un *identifiant* unique (un entier compris entre 1001 et 3392) et des éléments socio-démographiques : *'Age'* en années (entier, entre 15 et 18), *'Gender'* (0 : homme ; 1 : femme), *'Ethnicity'* (0 : caucasien ; 1 : afro-américain ; 2 : asiatique ; 3 : autre). Ces variables catégorielles non ordinales ont été transformées en variables discrètes. Chaque individu est également caractérisé par des variables catégorielles ordinales transformées en variables

quantitatives : le *'Parental Support'* (0 : Aucun ; 1 : Faible ; 2 : Moyen ; 3 : Fort ; 4 : Très fort) et le *'Parental Education'* (0 : Aucun diplôme ; 1 : Lycée (équivalent au baccalauréat) ; 2 : Université ; 3 : Licence ; 4 : niveau supérieur).

Les habitudes de travail de l'étudiant sont également décrites par des variables quantitatives : le *nombre d'heures d'absence ou 'Absences'* pendant une année scolaire (entier entre 0 et 30), le *'Study Time Weekly'* en heures (entier entre 0 et 20), si l'étudiant fait du *tutorat ou 'Tutoring'* (variable binaire, ie. 0 : non ; 1 : oui). D'autres variables binaires, au nombre de 4, permettent de décrire des activités qui sortent du cadre scolaire ( 0 : non ; 1 : oui) : des *activités extra-scolaires générales ou 'Extracurricular'*, des *activités sportives*, la pratique de la *musique* et du *volontariat*.

Enfin, le *'GPA'* est un réel compris entre 0 et 4, correspondant à la note moyenne des étudiants. Il est associé à un rang en fonction du score GPA, noté *'Grade Class'*, calculé comme suit : un GPA supérieur à 3.5 vaut un A (noté '0') ; B ('1') si compris entre 3 et 3.5 ; C ('2') si compris entre 2.5 et 3 ; D ('3') si compris entre 2 et 2.5 ; et F ('4') sinon. Cette valeur 'Grade Class' correspond à notre variable d'intérêt (ou *target*) pour les méthodes de classification.

Lors de cette phase préliminaire, nous avons également mis en place des stratégies pour nettoyer le jeu de données, détecter les valeurs aberrantes ('outliers'), créer des variables et pondérer des variables déjà existentes. Ces étapes sont détaillées dans l'annexe A.

# 2 Méthodes non supervisées

Cette partie présente les résultats de l'implémentation de méthodes non supervisées sur nos données.

## 2.1 Analyse en composantes principales

Nous avons ensuite appliqué une ACP. Les résultats de variance expliquée par chaque composante sont présentés Figure 1.

Ainsi, *seulement deux composantes principales suffisent à expliquer 94 % de la variance des données*, 95%, si l'on retient également la troisième composante. L'analyse de chaque composante nous montre que les variables explicatives sont les suivantes : (i) 'Absences' et 'GradeClass' pour la première CP, (ii) 'StudyTimeWeekly' pour la deuxième CP, (iii) 'Age', 'Parental Support' et 'GradeClass' pour la troisième CP. Il semblerait donc que certains élements socio-démographiques aient

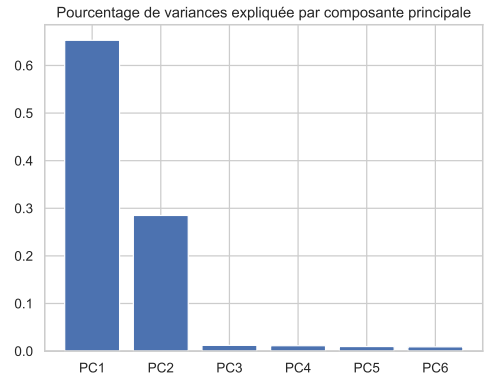


FIGURE 1 – Proportion de la variance expliquée.

effectivement un lien avec la réussite scolaire ('Grade Class').

Nous avons également représenté les individus dans le premier plan factoriel (cf. Figure 2). Ces résultats sont compatibles avec ceux de la heatmap (cf. annexe B) et des corrélations entre variables.

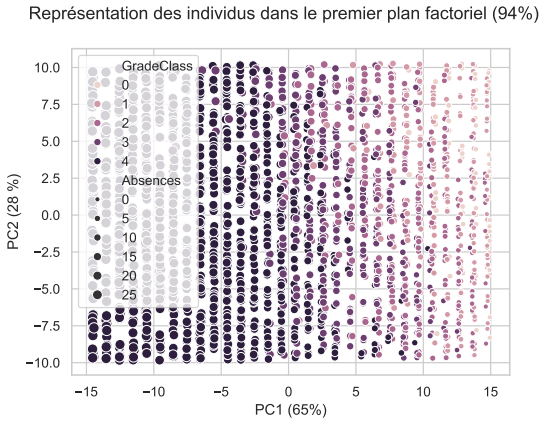


FIGURE 2 – Représentation des individus du jeu de données dans le premier plan factoriel.

## 2.2 Classification ascendante hiérarchique

Enfin, nous avons appliqué une CAH à notre jeu de données. Après avoir tracé le dendrogramme (cf. annexe C) en retirant 'GPA' et 'GradeClass', nous avons repéré à quel indice $r$, nous obtenions une partition de 5 éléments (en référence aux 5 catégories de *GradeClass* : A- E). En effet, nous souhaitions savoir quelles variables avaient joué un rôle significatif sur la classification. Pour

cela, nous avons relié le GPA d'origine aux clusters pour voir si la partition du dendrogramme correspondait au 'GradeClass' (cf. Figure 3). Ces dernières se sont majoritairement formées en fonction de 'Absences' et 'StudyTimeWeekly'. A noter que l'on aurait pu obtenir des résultats plus précis en ne conservant que les variables ayant un impact sur la variance expliquée (cf. partie 2.1) et en retirant les variables binaires par exemple.

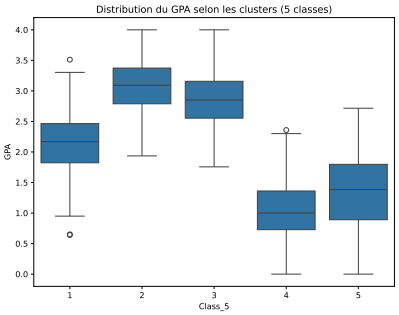


FIGURE 3 – Boxplot comparant la répartition du GPA selon les 5 classes obtenues après CAH.

Enfin, nous avons également tenté d'implémenter une autre méthode de classification hiérarchique : l'algorithme de clustering *Kmeans*. Cependant, cet essai ne s'est pas avéré concluant (cf. annexe D).

# 3 Méthodes supervisées

Dans cette partie, nous présenterons différents modèles de prédictions de la variable 'GradeClass' et les résultats obtenus. A noter que 'GradeClass' étant une variable ordinale (A > B > C > D > F), des méthodes de classifications conventionnelles telles que le KNN ou la régression logistique ne sont pas adaptées ('une mauvaise prédiction de A en F vaut autant qu'une prédiction fausse de A en B'). Tous les résultats présentés ont été obtenus par *validation croisée* avec $K = 5$ *folds*.

## 3.1 Analyse Discriminante

Ainsi, nous avons cherché à prédire la variable 'GradeClass' en utilisant différentes méthodes de classification discriminante (LDA, QDA, Naive Bayes), en excluant la variable 'GPA' car elle est directement corrélée à 'GradeClass' (celle-ci étant calculée à partir du GPA). Il serait normal d'obtenir une très bonne prédiction de 'GradeClass' avec 'GPA' (près de 98% de précision), bien que cela soit peu pertinent.

### 3.1.1 Prédiction initiale avec toutes les variables et réduction de dimensions par ACP

L'application de LDA, QDA et Naive Bayes sur toutes les variables disponibles, à l'exception de 'GPA', aboutit à des précisions comprises entre 74% et 79% dans le cadre d'une classification en 5 catégories (A, B, C, D, F). Ce résultat paraît satisfaisant compte tenu du recouvrement conséquent des classes.

Nous avons également appliqué une AD sur les données réduites par ACP (cf. Figure 4). Les résultats sont les suivants pour une classification en $g = 5$ catégories (A, B, C, D, F) : une précision d'environ 70%. Cette précision s'améliore lorsque l'on s'intéresse à $g = 2$ et $g = 3$ (cf. annexe E).

### 3.1.2 Sélection de variables pertinentes

Nous avons testé des sous-ensembles de variables considérées arbitrairement comme influentes sur la réussite scolaire :

1. *'Absences', 'StudyTimeWeekly', 'TotalActivities'* (modèle à 5 classes) : (i) environ 74% de précision pour LDA et QDA et (ii) environ 70% pour Naive Bayes. Ces variables ont donc bien un impact significatif, mais insuffisant pour expliquer la réussite dans sa globalité.
2. *'ParentalEducation', 'ParentalSupport', 'Tutoring'* : environ 53% de précision, ce qui correspond à la proportion de la classe F (53%). En réalité, tous les individus sont affectés à une unique classe, expliquant cette faible performance. Considérées ensemble, ces variables ne fournissent donc pas d'information discriminante.

### 3.1.3 Sélection exhaustive de combinaisons

Nous avons ensuite exploré l'ensemble des combinaisons possibles de 2 à 9 variables parmi : *'Age', 'Gender', 'Ethnicity', 'ParentalEducation', 'StudyTimeWeekly', 'Absences', 'Tutoring', 'ParentalSupport', 'totalActivities'*.

Dans le cas $g = 5$, les meilleurs résultats sont obtenus avec LDA ou QDA et une combinaison de 5 variables : 'StudyTimeWeekly', 'Absences', 'Tutoring', 'ParentalSupport', 'TotalActivities'. La meilleure précision est de *81.7%* sur la base d'une QDA.

L'ajout de variables telles que 'Gender' et 'Ethnicity' améliore la précision du système binaire de moins de 0.1% (ajouter une seule de ces variables peut même diminuer légèrement la précision). Le recueil de cette

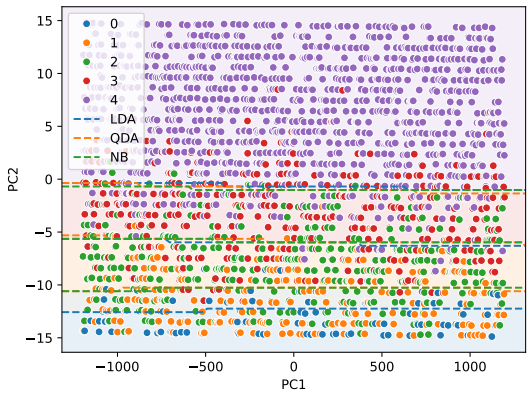


FIGURE 4 – Modèles d'analyse discriminante sur les données réduites par ACP.

dernière variable peut d'ailleurs soulever des questionnements éthiques.

La classe A, représentant seulement 3% des individus, est difficile à identifier. Forcer sa détection augmente l'erreur d'environ 2%. Ces différences de qualité de prédiction entre les classes peuvent s'observer sur la matrice de confusion obtenue (cf. Figure 5). En effet, un modèle est jugé bon lorsque les valeurs diagonales de sa matrice de confusion sont grandes et que celles des extrémités sont petites. On constate ici que F est donc bien mieux prédite que A.

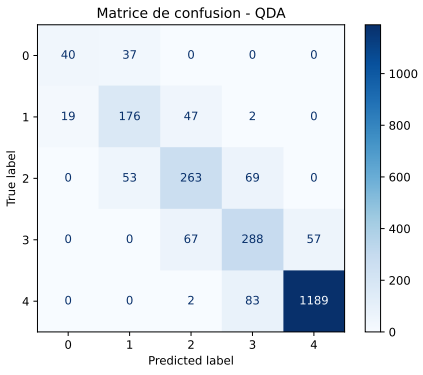


FIGURE 5 – Matrice de confusion de l'analyse discriminate (5 variables).

#### 3.1.4 Évaluation avec des métriques adaptées à la variable ordinale (5 classes)

Puisque 'GradeClass' est une variable ordinale, nous avons évalué les performances du modèle avec des métriques prenant en compte l'ordre des classes.

Ainsi, nous avons obtenu un *Quadratic Weighted Cohen's Kappa (QWK)* de *0.931*, reflétant une excellente performance, la plupart des erreurs étant limitées à une classe d'écart.Nous avons également obtenue une *Mean Absolute Error (MAE)* de *0.184*. Le nombre d'erreur semble donc acceptable en terme de gravité. Enfin, La corrélation de *Spearman's Rho* est mesurée à *0.929*, ce qui confirme l'excellente conservation de l'ordre des notes. Ces métriques et leurs calculs sont explicités dans l'annexe F.

### 3.2 Régression logistique ordinale

Nous avons ensuite décidé d'implémenter une *régression logistique ordinale* pour comparer les deux prédictions obtenues. En effet, contrairement à l'analyse discriminante, qui nécéssitaient des mesures d'évaluation a posteriori (ie. *ex post*), le modèle appliqué ici est d'emblé adapté à des variables ordinales. De plus, la QDA et la LDA sont des modèles génératifs alors que la régression est un modèle discriminant (plus performant en classification, mais nécessitant plus de données pour atteindre une bonne performance).

#### 3.2.1 Sélection des variables pertinentes

Après avoir appliqué ce modèle à nos données [2], nous obtenons une précision de *83%*. Nous observons également que les variables les plus significatives sont les suivantes [1] :

1. 'StudyTimeWeekly' (-0.2615), ie. plus un étudiant passe de temps à étudier, plus la probabilité d'être dans une bonne classe augmente (A ou B).
2. 'Absences' (0.9220), ie. plus 'Absences' est grand, plus le risque d'avoir F, D, C augmente.
3. 'Tutoring' (-2.3446), ie. avoir du tutorat réduit fortement les chances d'avoir une mauvaise note (effet très significatif).
4. 'ParentalSupport' (-1.3510), ie. un bon soutien parental est un fort prédicteur de réussite.

Les variables non significatives sont identifiables à leur p-value élevées au dessus de 0.05 : 'Age' (0.938), 'Gender' (0.197), 'Ethnicity' (0.214), 'ParentalEducation' (0.696).

1. A noter que plus la valeur du coefficient est basse et négative, et plus la note sera élevée et positive.

#### 3.2.2 Précision du modèle pour la prédiction de chaque note

En terme de performance du modèle, F est la note la mieux prédite, suivie de D. Pour B, C et D, le modèle reste raisonable avec une performance moyenne. Pour F, la détection est excellente, ce qui est probablement lié au fait que c'est la note la plus courante (249 F dans l'ensemble de validation).

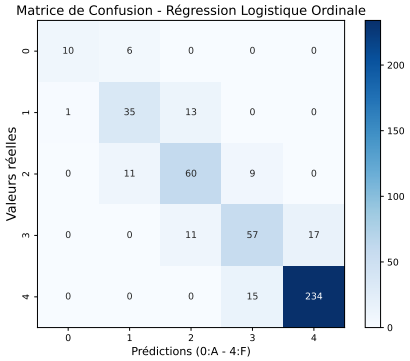


FIGURE 6 – Matrice de confusion du modèle de régression logistique ordinal.

Sur la matrice de confusion (cf. Figure 9), nous observons que F est particulièrement bien prédite (234 individus) et que l'on a une majorité de 0 sur les côtés. Le modèle a cependant tendance à inverser D et F, ou B avec C et D. Pour conclure, il s'agit d'un modèle satisfaisant, possédant de bonnes performances. Le *recall* est légèrement plus faible en raison de la surreprésentation des classes rares, telles que A et F.

### 3.3 Autres modèles

D'autres méthodes de prédiction ont également pu être testés, tels que le *Gradient Boosting* [4] en régression sur la variable 'GPA' (variable quantitative continue que nous avons traduit en 'Grade Class' une fois la prédiction terminée). Pour ce faire, nous avons recherché les hyperparamètres offrant les meilleurs résultats et évalué chaque modèle par validation croisée. L'application de cet algorithme donne de bons résultats (légèrement plus satisfaisants que pour l'AD) en termes de précision. Toutes les combinaisons de 2 à 5 variables ont aussi été testées : les 5 mêmes variables que la QDA offrent le meilleur résultat avec une précision d'environ *0.83*. Ainsi, comme pour les analyses précédentes, les classes les plus faciles et plus difficiles à prédire restent respectivement les classes F et A (cf. annexe G).

Enfin, nous avons également testé des forêts aléatoires afin de pondérer les valeurs de 'Grade Class' et palier au problème de déséquilibre entre classes. En effet, comme mentionné section 3.2.2, certaines classes sont majoritaires (ie. A et F) ce qui pourrait expliquer les meilleures prédictions pour ces notes (cf. annexe H).

## 4 Conclusion

Ce travail nous a permis d'explorer comment les éléments socio-démographiques influençaient les résultats scolaires d'un large échantillon d'étudiants. Nous avons pu montrer que le taux d'absence, le temps de travail hebdomadaire, le tutorat, les activités extrascolaires et le support parental sont les facteurs qui déterminent le plus la réussite scolaire, en accord avec les conclusions de nombreuses rapports et articles sur le sujet de ces dernières années [7] et [6]. A l'inverse, il semblerait que le niveau d'étude des parents et l'origine ethnique en fonction du genre aient peu joué dans la réussite scolaire au sein de notre échantillon, ce qui va à l'encontre de certains résultats publiés [5] et [1]. Une première explication quant à cette discordance pourrait être la nature synthétique du jeu de données utilisé et le manque de certaines informations spécifiques. Il manque en effet un élément de contexte essentiel à l'analyse : *Où, Quand et Comment* ces données ont-elles été recueillies ? Le jeu de données ne permet pas de tirer de conclusions précises quant aux stratégies d'apprentissage à adopter pour réussir : les variables significatives étant très générales, de nombreuses informations restent inconnues, par exemple, la répartition du temps de travail hebdomadaire, des absences sur le semestre ou encore les notes dans chaque matière. En outre, un certain nombre de facteurs empêchent l'analyse des étudiants ayant obtenu des résultats intermédiaires ('GradeClass' entre D et B). Certains facteurs socio-culturels se manifestent dans des matières spécifiques plutôt qu'en terme de réussite scolaire globale. Nous pouvons citer l'exemple bien connu des différences hommes/femmes mises en évidence dans le rapport PISA. En mathématiques, les garçons obtiennent de meilleurs résultats que les filles, tandis qu'en lecture ce sont les filles qui surpassent les garçons, avec une différence encore plus marquée. Cela ne signifie pas pour autant qu'il existe des écarts de performance scolaire en général. De manière plus globale, la nature "synthétique" de note jeu de données se traduit par de faibles corrélations entre variables, sauf avec l'indicateur de performance scolaire.

# A Tentatives d'amélioration du jeu de données

### A.0.1 Nettoyage des données

Lors de notre contrôle qualité, nous nous sommes rendu compte que les valeurs de *'Grade Class'* ne correspondaient pas au 'GPA' associé pour 85 individus sur 1274 (soit environ 6.5% de l'échantillon), ie. GPA inférieur à 2 mais 'Grade' différent de 4.

Or, le 'Grade Class' est calculé à partir du 'GPA'. Plusieurs solutions s'offraient à nous pour résoudre cette incohérence : (i) soit re-calculer les 'Grade Class', (ii) soit retirer les individus présentant des 'Grades' aberrants. Nous avons opté pour de la *data curation* (option 1), ie. nous avons recalculé le classement. Nous nous sommes cependant assurés que l'ACP et la CAH avec et sans ces corrections n'étaient pas modifiées de façon substantielle.

### A.0.2 Analyse des outliers

Nous avons analysé les cas dit *extrêmes* : (i) le cas d'élèves très studieux (nombre d'heures travaillées élevé, ie. $> 19$) mais avec un taux d'absentéisme élevé (heures d'absences égales à 29) ; (ii) le cas d'élèves avec un 'GPA' très élevé (supérieur à 3.9) et peu ou pas de 'soutien parental' et (iii) le cas d'élèves avec un 'GPA' très élevé et un nombre d'heures travaillées très faible.

Ces outliers correspondent à des cas atypiques, extrêmes mais plausibles et sont en réalité des cas isolés possibles (entre 1 à 4 sur 2392). Nous avons donc décidé de les conserver.

### A.0.3 Focus sur les activités extrascolaires

Nous avons créé une nouvelle variable appelée *'TotalActivities'*, pouvant prendre les valeurs suivantes : 0, 1, 2, 3 ou 4. Il s'agit de la somme des activités extrascolaires auxquelles participent les lycéens (1 correspondant à la pratique de l'activité, 0 à sa non-pratique).

Nous avons également essayé de pondérer les variables activités : (i) en utilisant les corrélations entre ces variables et le 'GPA' - mais cela introduisait un biais dans nos résultats (nous avons donc abandonné cette méthode) et (ii) en testant des combinaisons arbitraires, e.g., "la musique est plus demandeuse en temps que le sport" (ces dernières ne changeant pas les résultats de classification, nous avons également abandonné cette méthode).

Finalement, en appliquant une pondération de 0.25 à chaque variable, nous avons réussi à identifier de profils "types" préliminaires de lycéens :

1. un lycéen avec $A$ ne fait *jamais plus de 3 activités*. Plus de la moitié des lycéens ayant A font *une ou deux activités* (résultats semblables pour B) ;
2. le nombre d'activités est beaucoup *plus variable pour les lycéens avec $C$ et $D$* (résultats plus éparses de 0 à 4 activités) ;
3. beaucoup des élèves avec $E$ ne font *aucune activité* (hypothèse plausible de décrochage scolaire).

## A.1 Partitionnement du jeu de données

Nous avons également essayé de partitionner notre jeu de données en fonction des variables catégorielles comme 'Parental Support' ou 'Parental Education' pour observer les corrélations entre variables et les résultats de prédictions pour des éléments socio-démographiques précis. Cependant, les résultats ne se sont pas avérés concluants.

# B Corrélations entre variables

Lors de l'analyse préliminaire, nous avons calculé la corrélation de Pearson entre nos variables quantitatives et représenté ces associations par une carte thermique (heatmap) (7) :

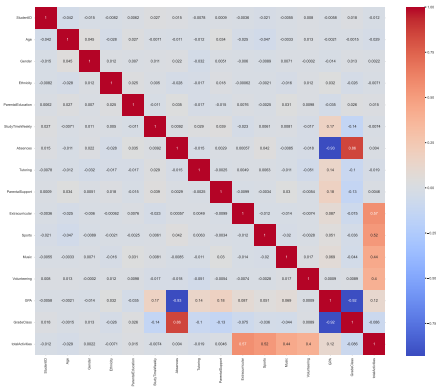

FIGURE 7 – Corrélations de Pearson entre les variables de notre jeu de données

Il existe une corrélation de *-0.92* entre le nombre d'absences et le GPA, et une corrélation de *+0.83* entre le nombre d'absences et le rang dans la classe. Ces corrélations extrêmes peuvent être expliquées par la nature synthétique de notre jeu de données. Nous obtenions

également une corrélation de *+0.18* entre le GPA et le nombre d'heures de travail hebdomadaire (et de *-0.15* entre le rang dans la classe et le nombre d'heures travaillées).

# C Dendrogramme après CAH

Nous avons tracé le dendrogramme ainsi que 3 droites :

1. x = 275 (on obtient 2 classes) ;
2. x = 250 (on obtient 3 classes) ;
3. x = 150 (on obtient 5 classes).

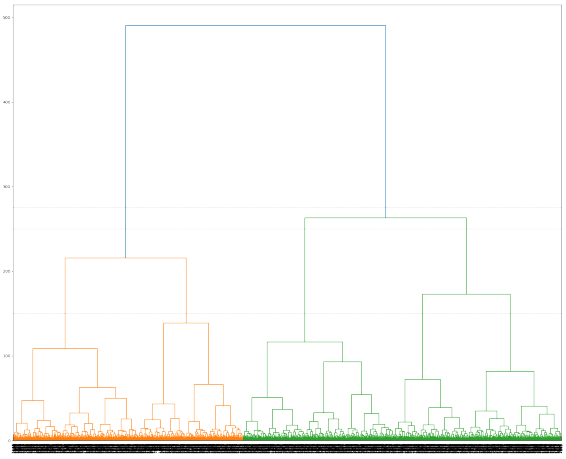

FIGURE 8 – Dendrogramme obtenu après une classification ascendante hiérarchique.

# D Algorithme de clustering Kmeans

L'implémentation d'un algorithme de clustering de type Kmeans sur nos données n'a pas donné de résultat concluant. En effet, cette méthode repose sur des hypothèses fortes concernant la forme et la distribution des données, qui ne sont pas respectées ici. En particulier :

1. *K-means cherche à créer des regroupements compacts et bien isolés.* L'algorithme essaie de minimiser la distance intra-cluster, ce qui suppose que chaque groupe forme une zone dense, bien distincte des autres. Or, nos données présentent des *zones de recouvrement*, ce qui rend cette approche non efficiente.
2. *K-means suppose des clusters de forme circulaire (ou sphérique).* L'algorithme se base sur la distance euclidienne et tend à créer des regroupements de forme arrondie, ce qui n'est pas le cas de nos clusters.

# E Comparaison des résultats de l'analyse discriminante avec $g = 2$ et $g = 3$

La précision du modèle est meilleure quand on s'intéresse à 2 classes (F vs non-F) ou 3 classes (A, F ou Autre) : respectivement 89.5% et 86.5% de précision.

Cependant, on observe que dans le dernier cas, seuls deux clusters distincts émergent vraiment des modèles. La classe *A*, peu représentée (3%), n'est pas distinguable de la classe *Autre*, ce qui explique la différence d'erreur d'environ 3%. La différence de précision avec le système à 5 classes montre qu'il existe également des difficultés à différencier les classes B, C et D.

Les meilleurs résultats obtenus avec $g = 2$ et $g = 3$ sont les suivants :

1. 91.5% pour les 3 classes : A/F/Autre (QDA)
2. 93.9% pour les 2 classes : F/non-F (LDA).

Dans les deux cas, ces résultats sont obtenus avec une combinaison de 5 variables : 'StudyTimeWeekly', 'Absences', 'Tutoring', 'ParentalSupport', 'TotalActivities'.

# F Métriques supplémentaires

Comme l'analyse discriminante ne prenait pas en compte le caractère ordinal des classes, nous avons utilisé des métriques supplémentaires pour évaluer la prédiction obtenue.

**Quadratic Weighted Cohen's Kappa (QWK)**

Le QWK mesure l'accord pondéré entre classes réelles et classes prédites, en pénalisant plus fortement les erreurs importantes (cf. Tableau 1).

TABLE 1 – Tableau répertoriant les valeurs attribuées en fonction des erreurs de prédictions. Une valeur proche de 1 indique un excellent accord pondéré.

| Valeur | 1 | 0.94 | 0.75 | 0.44 | 0 |
|---|---|---|---|---|---|
| Ecart entre 2 classes | 0 | ± 1 | ± 2 | ± 3 | ± 4 |

**Mean Absolute Error (MAE)**

La MAE mesure l'erreur moyenne absolue entre classes réelles et classes prédites, considérant la distance ordinale (l'erreur maximale étant de 4, entre A et F). Une MAE supérieur à 0.2 signifie que la prédiction moyenne est à moins d'un quart de classe de la vraie classe.

**Spearman's Rho**

Cette métrique mesure la corrélation monotone entre les rangs des classes prédites et réelles. Une valeur proche de 1 signifie que l'ordre des classes est très bien respecté (ie. si un élève a une meilleure note réelle par rapport à un autre élève, il aura aussi une meilleure note prédite).

# G Gradient Boosting : concepts et matrice de confusion

La méthode de renforcement du gradient, ou *Gradient Boosting*, consiste à construire un prédicteur de manière incrémentale. Dans cette approche, à chaque itération, l'erreur du modèle précédent (définie par diverses fonctions de perte) est corrigée par l'ajout d'un apprenant faible associé à un coefficient et des paramètres spécifiques. A noter que le modèle initial est lui même un apprenant faible. Un apprenant faible est un prédicteur meilleur que le hasard (dont la précision dépasse au moins les 50%).

Le nombre d'itération de l'algorithme, l'apprenant faible (noté $h$) et la fonction de perte (notée $L$), sont des paramètres à définir. Le *Gradient Boosting* tire son nom de la manière dont on choisit le poids et les paramètres des nouveaux apprenants faibles. Le gradient d'une fonction représente la direction d'augmentation maximale. Dans le cadre des fonctions de perte, cela représente donc la direction dans laquelle l'erreur augmente le plus rapidement. L'objectif du gradient boosting est donc, dans un premier temps, de trouver : (i) les paramètres, et (ii) l'apprenant faible, permettant la meilleure correction de cette direction. Puis, dans un second temps, (iii) de choisir la pondération réduisant au maximum l'erreur. Ces étapes sont répétées jusqu'à obtention d'un prédicteur suffisamment précis.

Cela peut être résumé par les formules suivantes, où $a_m$ représente les paramètres du $m^{ime}$ apprenant et $\beta_m$, le poids du $m^{ime}$ apprenant faible. $Y_i$ correspond à la valeur à prédire de l'individu $i$. Les $X_i$ correspondent aux variables explicatives et $F_{m-1}$ au $(m-1)^{ime}$ modèle. Pour tout $m$ compris entre 1 et le nombre d'itération de l'algorithme, on a :

$$a_m = \text{argmin}_a \sum_{i=1}^{n} (\frac{\partial L(Y_i, F_{m-1}(X_i)}{\partial F_{m-1}(X_i)} - h(X_i, a))^2$$

$$\beta_m = \text{argmin}_\beta \sum_{i=1}^{n} (L(Y_i, F_{m-1}(X_i) + \beta h(X_i, a))$$

La matrice de confusion obtenue avec *Gradient Boosting* appliquée au jeu de données sélectionné est présentée ci-dessous. Les résultats sont semblables à ceux obtenus avec d'autres modèles de prédiction.

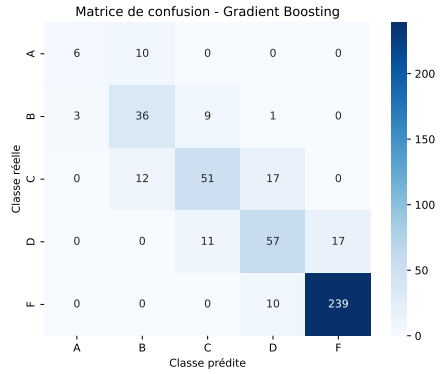


FIGURE 9 – Matrice de confusion du modèle de Gradient Boosting.

# H Forêt aléatoire et pondération de classes

Nous avons expérimenté l'utilisation d'un modèle de forêt aléatoire dans le but de prédire la variable cible 'GradeClass'. Deux approches distinctes ont été testées afin de répondre aux contraintes particulières de notre jeu de données : (i) le déséquilibre important des classes, et (ii) le caractère ordinal de la variable cible.

**Prise en compte du déséquilibre des classes**

Dans un premier test, nous avons appliqué une pondération des classes afin de limiter l'impact du déséquilibre entre les effectifs. Les poids ont été définis selon la fréquence des classes dans l'échantillon d'entraînement ce qui a conduit aux pondérations suivantes : les classes les moins représentées, comme A, ont reçu un poids élevé (6.27), tandis que la classe F, très fréquente, a un poids faible (0.37). Les classes B, C et D ont respectivement reçu 1.96, 1.25 et 1.17. Cette stratégie vise à accentuer la prise en compte des classes rares (notamment A) dans la fonction de perte, et à

réduire l'importance accordée à la classe majoritaire (F). Le modèle entrainé avec ces pondérations obtient les résultats suivants : Accuracy = *75%* ; F1-score macro (non pondéré) = *0.57* ; F1-score pondéré = *0.74*. On constate également que le modèle parvient à bien détecter la classe majoritaire F. Cependant, la classe A reste très peu détectée, malgré une bonne précision lorsqu'elle l'est. Ce déséquilibre reste un point critique de cette approche.

### Prise en compte de la structure ordinale

Une seconde approche a consisté à tester une implémentation de la forêt aléatoire adaptée à la nature ordinale de la variable cible. Pour cela, nous avons utilisé le modèle *OrdinalRidge* (de la bibliothèque *mord*), qui nous a permis d'obtenir les résultats suivants : Accuracy = *67%* ; F1-score macro = *0.48* ; F1-score pondéré = *0.67*. Bien que cette méthode respecte mieux la nature ordinale des variables (moindre coût des erreurs proches), les performances globales sont en baisse, et la classe A n'est jamais détectée.

Ainsi, la version pondérée donne de meilleurs résultats globaux et permet une prédiction plus équilibrée entre les classes. L'approche ordinale, bien que plus rigoureuse et pertinente sur le plan théorie, reste limitée par son incapacité à gérer le déséquilibre des classes, affectant négativement la précision globale.

## Références


[1] Yaël Brinbaum. "Trajectoires scolaires des enfants d'immigrés jusqu'au baccalauréat : rôle de l'origine et du genre". français. In : *Éducation Formations* 100 (2019). Consulté le 5 juin 2025, p. 73-95. URL : https://www.education.gouv.fr/sites/default/files/imported_files/document/depp-2019-EF100-article-04_1221890.pdf.

[2] Jumbong Junior. *Ordinal Logistic Regression in Python and R*. Consulté le 6 juin 2025. 2024. URL : https://medium.com/@jumbongjunior1999/ordinal-logistic-regression-in-python-and-r-f6ee05d48d16.

[3] Rabie El Kharoua. *Students Performance Dataset*. Consulté le 10 avril 2025. 2022. URL : https://www.kaggle.com/datasets/rabieelkharoua/students-performance-dataset/data.

[4] Paul Liautaud et Laure Ferraris. *Méthode de Gradient Boosting*. Consulté le 9 juin 2025. 2022. URL : https://perso.lpsm.paris/~liautaud/projects/gradient_boosting.pdf.

[5] Fabrice Murat. *Évolution des inégalités sociales de compétences au fil du temps et de la scolarité*. français. Consulté le 5 juin 2025. Juin 2024. DOI : 10.48464/ni-24-21.

[6] Organisation de Coopération et de Développement Économiques (OCDE). *PISA 2022 – Résultats (Volume V) : Compétences, motivation et stratégies d'apprentissage*. français. Consulté le 5 juin 2025. Paris : OCDE, 2024.

[7] U.S. Department of Education. *Chronic Absenteeism*. anglais. Consulté le 5 juin 2025. 2024. URL : https://www.ed.gov/teaching-and-administration/supporting-students/chronic-absenteeism.