Compte-rendu Projet SY09

Louis Caron, Waël Hamdan, Zhou Xingjian

June 2019

Abstract

Ce projet porte sur l'analyse du jeu de données "Students Performance in Exams" disponible sur Kaggle. Ces données contiennent les caractéristiques d'étudiants ainsi que leurs scores. Dans le but de prévoir les scores d'un étudiant à partir de ses autres variables, nous avons exploré et analysé le jeu de donnée dans l'espoir de trouve une relation entre les scores et les autres variables. Malheureusement, nos recherches montrèrent que, s'il existe un moyen de prédire un score, celui-ci implique d'utiliser les autres scores qui sont corrélés. Après avoir utilisé différent modèles de régression, classification et une ACM, nous avons trouvé que les scores n'étaient pas prévisibles, et que l'on pouvait juste déterminer l'importance minime de certaines variables sur les scores.

Nom de variable	Description	Intervalle	Remarques
gender	Qualitative : card.2	male female	48.2% male
race/ethnicity	Qualitative : card.5	groupA groupB groupC groupD groupE	8.9% 19.0% 31.9% 26.2% 14.0%
parental level of education	Qualitative : card.6	associate's degree bachelor's degree high school master's degree some college some high school	22.2% 11.8% 19.6% 5.9% 22.6% 17.9%
lunch	Qualitative : card.2	free or reduced standard	35.5% free/reduced
test preparation course	Qualitative : card.2	completed none	35.8% completed
math score	Quantitative	0 - 100	
reading score	Quantitative	0 - 100	
writing score	Quantitative	0 - 100	

TABLE 1 – Tableau récapitulatif des différentes variables

1 Introduction

Dans le cadre de l'UV SY09 de l'UTC, nous avons pris part à une analyse de données provenant de *Kaggle*. Ces données sont nommées "*Students Performance in Exams*". Il s'agit d'un jeu de 1000 individus et de 8 variables, dont 3 quantitatives représentant les scores obtenus par les étudiants aux examens. Les autres variables sont qualitatives. Les données sont bien sûr anonymes. Dans le cadre de notre projet, nous essayerons d'explorer et d'analyser ces données et de voir l'impact des différentes variables sur les scores des étudiants. Idéalement, nous espérons pouvoir prévoir les scores d'un étudiant en connaissant ses autres variables.

2 Analyse Univariée

Afin de mieux appréhender les données que nous avons, nous commençons par analyser les variables indépendamment les unes des autres. On étudiera les distributions et valeurs prises par les différentes variables.

Nous rappelons qu'il n'y a pas de valeur manquante dans ce jeu de données.

2.1 Traitement de données

La Table 1 contient les répartitions des différentes modalités de chacune des variables qualitatives de notre dataset.

2.2 Statistiques descriptives

Avant de modéliser les facteurs d'influence des résultats de test, une analyse descriptive des variables est présentée dans les tables 1 et 2.

TABLE 2 – Statistiques descriptives des performances des étudiants

Taille de l'échantillon (n)	1000	**min**	0
Moyenne (mean)	66.09	**max**	100
Écart type (sd)	15.16	**Etendue (range)**	100
Médiane (median)	66	**Asymétrie (skew)**	-0.27
Erreur type (se)	0.48	**Kurtosis**	0.24

L'étudiant de l'échantillon qui obtenu les meilleures notes en mathématiques est un homme, de race E, dont le niveau d'éducation des parents correspond à un diplôme d'associé, qui n'avait pas de déjeuner et qui avait suivi de test préparatoire.

L'étudiant de l'échantillon aux pires notes en mathématiques est une femme de l'ethnie C, dont le niveau d'éducation des parents est le lycée, n'avait pas de déjeuner et qui n'avait pas suivi de test préparatoire.

On constate que la distribution des scores est biaisée à droite, la plupart des scores étant compris entre 60 et 70.

2.3 Test de normalité

On effectue maintenant un test de normalité sur les trois scores, on dessine l'histogramme de test de normalité (Figure **??**) ainsi que les normal Q-Q plot et P-P plot (Figure 1).

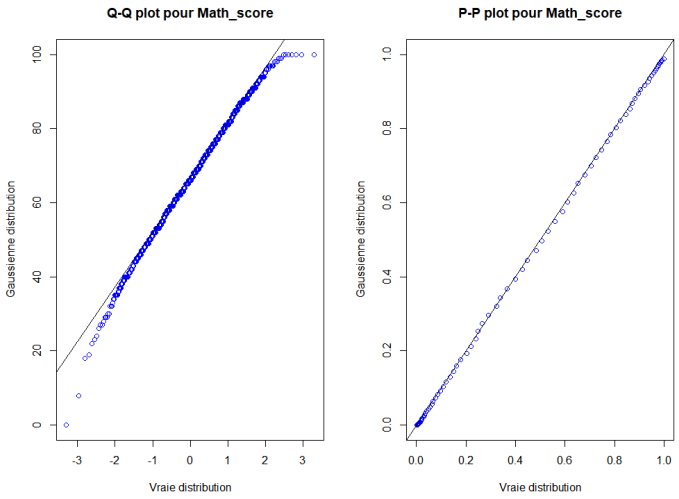

FIGURE 1 – normal Q-Q plot et P-P plot

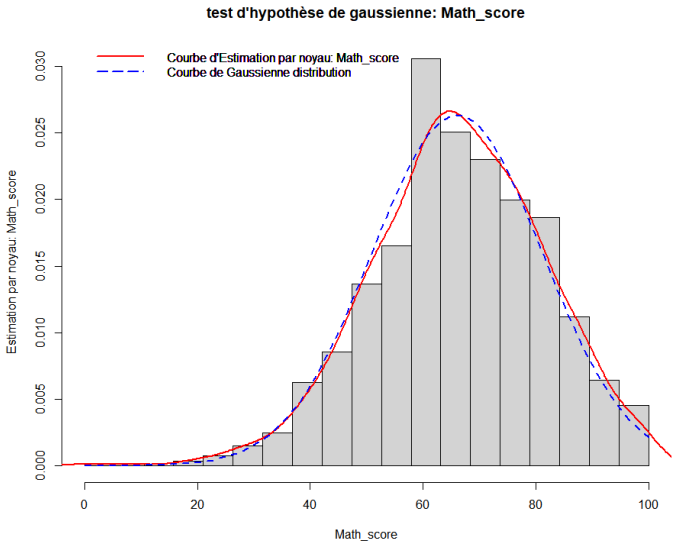

FIGURE 2 – histogramme de test de normalité : Math scores (les autres scores ont des courbes similaires)

Le test de Shapiro-Wilk donne pour *math.score*, *reading.score* et *writing.score* respectivement une p-value de 0.0001455, 0.0001055 et $2.922e-05$. Elles sont toutes inférieures au seuil de signification choisi de 5%. L'hypothèse de normalité est donc rejetée pour chacun des trois scores. Les données ne sont pas normales d'après le test de Shapiro-Wilk. Au vu de la distribution des notes et des résultats du Q-Q et P-P plot, on se permettra d'appliquer des outils normalement réservés aux données gaussiennes tout en gardant à l'esprit que l'hypothèse ne normalité n'est pas vérifiée.

3 Analyse Multivariée

Nous étudions maintenant les liens statistiques entre les différentes variables.

3.1 L'influence des facteurs sur la performance

Nous cherchons d'abord à déterminer quelles variables qualitatives ont une influence sur chacun des trois scores : math.score, reading.score et reading.score.

Pour ce faire, nous effectuons des test ANOVA.

3.1.1 L'influence du genre sur la performance

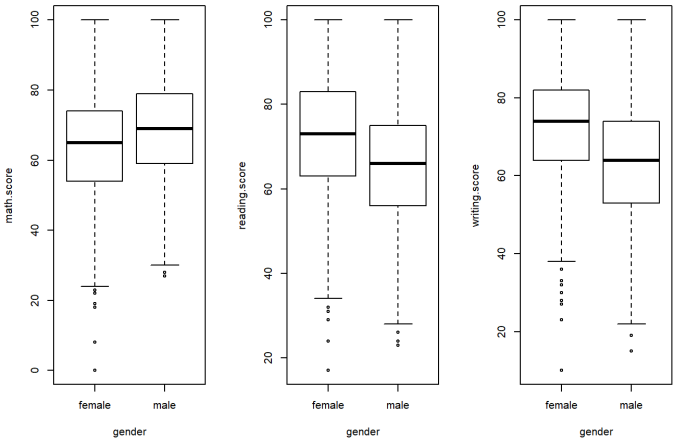

FIGURE 3 – Distribution des scores en fonction du genre

Il semble y avoir une différence notable entre les hommes et les femmes pour les scores en fonction du genre (Figure 3). Il semblerait que les hommes soient légèrement meilleurs en mathématiques que les femmes, qui seraient meilleures en lecture et en écriture.

Cela est encore plus visible sur la figure 4. On peut en effet clairement voir deux clusters correspondant au deux genres. Une frontière linéaire est clairement visible entre les deux genres. Bien que le genre semble voir une influence sur les notes, il ne permet pas de prédire celles-ci. Un genre ne semble en effet pas complètement surpasser l'autre.

Pour confirmer ou infirmer ces résultats, nous effectuons des ANOVA et on constate alors une différence significative entre les hommes et les femmes pour chacun des scores : *math.score*, *reading.score* et *writing.score*.

2

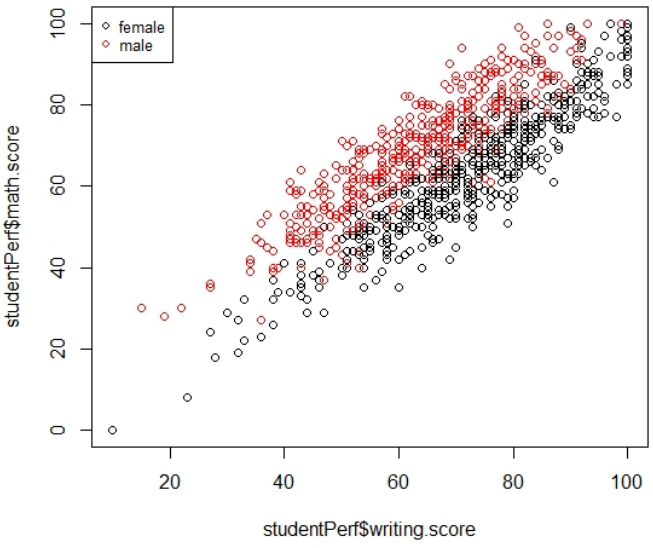

FIGURE 4 – Scores des élèves en maths et écriture en fonction du genre

3.1.2 L'influence de l'ethnicité sur la performance

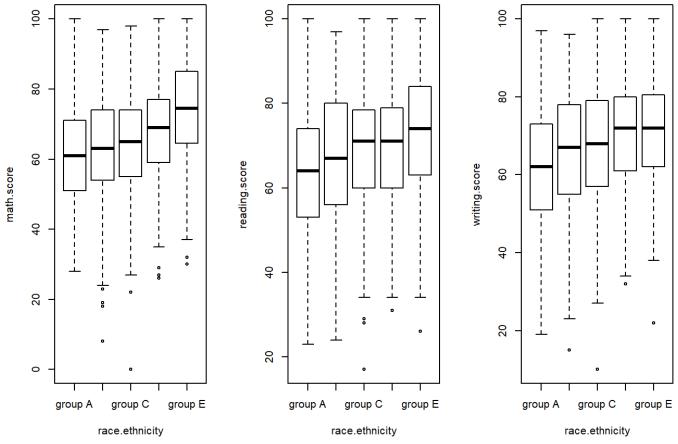

FIGURE 5 – Distribution des scores en fonction de l'ethnicité

Pour chacun des 3 scores, il semble y avoir des différences en fonction de l'ethnicité de l'étudiant (Figure 5). Les scores semblent être d'autant plus élevés que l'étudiant appartienne au groupe E, puis D etc. jusqu'au groupe A dont les scores semblent êtres les moins élevés. Pour confirmer ou infirmer ces résultats, nous effectuons des ANOVA et on constate alors une différence significative en fonction de l'ethnicité pour chacun des scores : *math.score*, *reading.score* et *writing.score*.

Vu les résultats obtenus, il semble cependant très difficile de prévoir les performances d'un étudiant en se basant seulement sur ce critère. Les étudiants de toutes les ethnies ayant des scores étendus sur une très large intervalles.

3.1.3 L'influence du type de déjeuner sur la performance

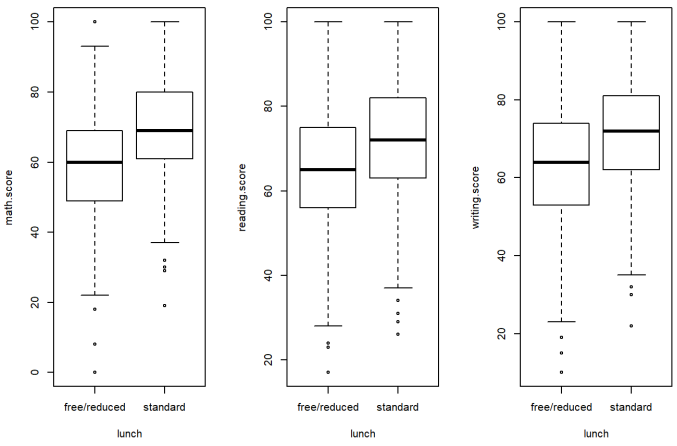

FIGURE 6 – Distribution des scores en fonction du type de déjeuner

Pour chacun des 3 scores, il semble y avoir une différence notable en fonction du déjeuner 6. Ceux ayant un déjeuner standard semblent avoir de meilleurs résultat. L'ANOVA confirme cette intuition.

Là encore, bien que l'avantage du *standard* soit évidente, elle ne semble pas assez marquée pour pouvoir en déduire les performances avec cette variable seule.

3.1.4 L'influence du passage d'un test de préparation sur la performance

Pour chacun des 3 scores, il semble y avoir une différence notable entre les étudiants ayant passé un test préparatoire ou non (Figure 7). Ceux ayant passé un test semblent avoir de meilleurs scores, et l'ANOVA confirme une différence significative selon le passage ou non d'un test préparatoire.

3.1.5 L'influence du niveau d'éducation des parents sur la performance

On dessine le boxplot des notes en math des étudiants (disponible sur la figure 8), on remarque que les étudiants dont les parents ont un niveau d'éducation de type lycée ont des performances plus basses, mais pour les autres, le niveau d'éducation des parents n'influence pas beaucoup sur la performance des étudiants.

3

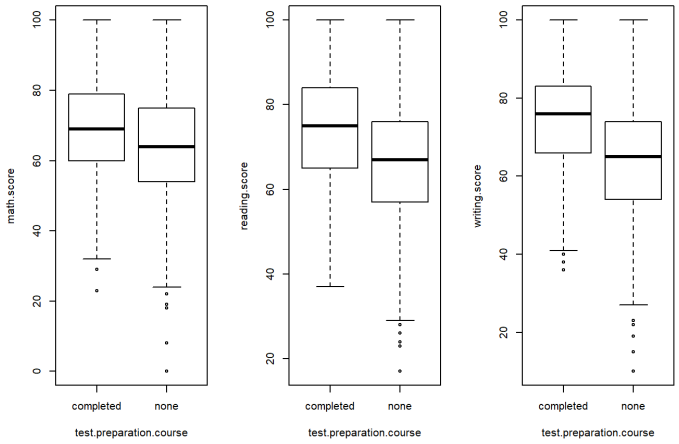

FIGURE 7 – Distribution des scores en fonction du passage ou non d'un test de préparation

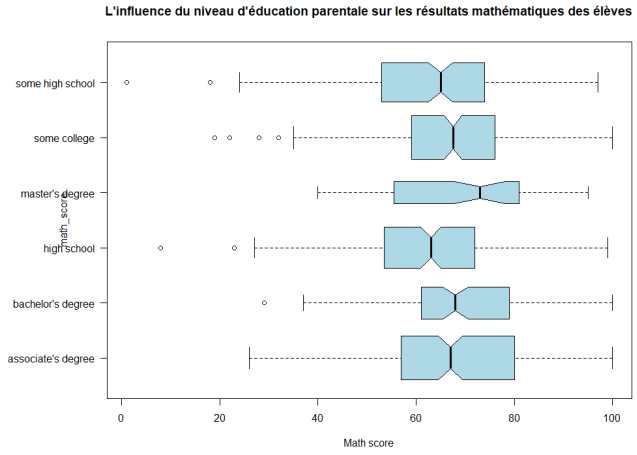

FIGURE 8 – L'influence Niveau d'éducation des parents - Performance

3.2 Influence croisée de plusieurs variables

Vu que nos variables qualitatives prises seules ne semblent pas avoir d'influence suffisante pour prévoir les performances d'un étudiant, nous allons nous intéresser aux relations plus poussées pouvant exister entre les variables.

3.2.1 L'influence ethnie - niveau d'étude des parents

D'après la figure 9, on peut aussi voir que les différentes ethnies ne possèdent pas la même distribution de niveau d'éducation des parents. On peut notamment voir que les groupe A et B ont une éducation moins importante que celle des autres groupes, tandis que les groupes D et E semblent avoir une plus haute proportion de parents ayant obtenus des diplômes universitaires.
Il est intéressant de voir que bien que l'ethnie et le niveau d'étude parental influent sur les performances, le niveau d'étude parental seul ne semble pas avoir une influence significative. Il se peut que l'ethnie influe sur les performances et que cela se traduise par des niveaux d'éducation plus bas au sein de l'ethnie, ou bien que l'influence minime du niveau d'étude parental fasse tendre une ethnie moins "éduquée" vers des performances légèrement plus faibles.

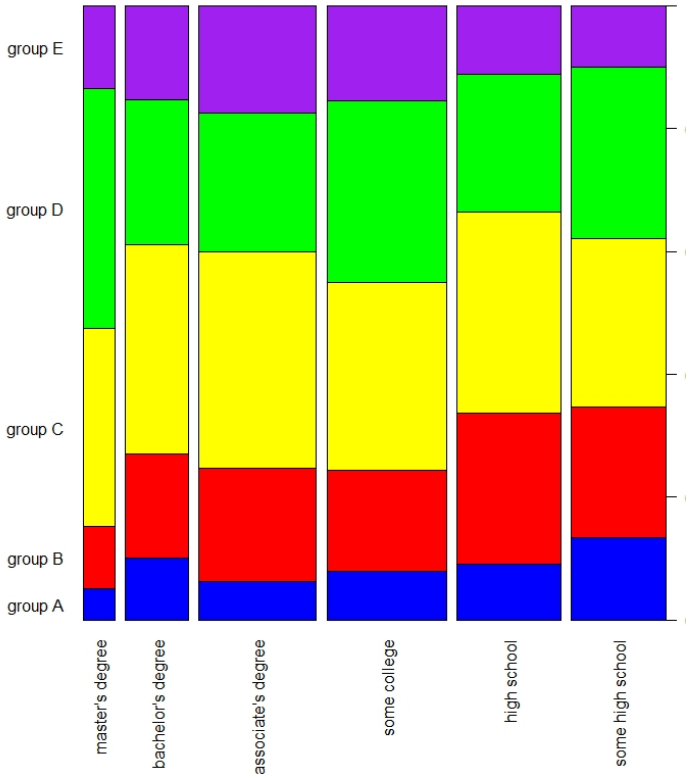

FIGURE 9 – Le niveau d'éducation des parents selon la race pour nos individus

3.2.2 L'influence ethnie - repas - genre - performance

On va essayer d'analyser ici l'influence de 3 variables sur les performances. Nous utiliserons d'autre part ici la somme des 3 scores comme indicateur de performance. On indique dans la figure 10 l'influence de l'ethnie, du genre et du type de déjeuner sur la moyenne de la somme des performances.

On peut voir que les femmes ont de meilleurs résultats globaux comme vu auparavant, et que l'ethnie a un impact ainsi que le type de déjeuner. Il est plus intéressant de constater que la performance est affectée de façon différente lorsqu'on l'étudie avec ces 3 variables simultanément. On se méfiera quand même de ces résultats car bien que la moyenne soit plus simple et lisible, elle ne montre pas toute les informations de la distribution.
Ainsi, l'influence du type de déjeuner est différente selon l'ethnie et le genre. Il semblerait même que chaque ethnie ait des scores assez différents pour un type de déjeuner donné

4

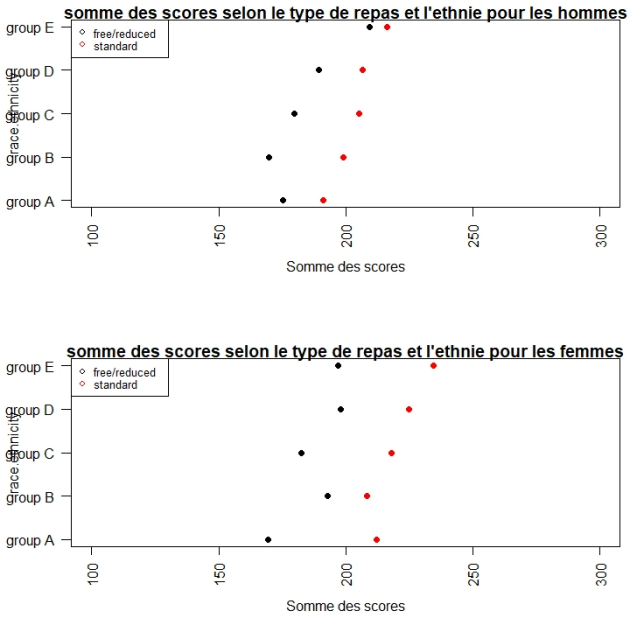

FIGURE 10 – Moyenne de la somme des performances selon l'ethnie, le genre et le type de déjeuner

pour un genre, mais pas pour l'autre. Il est intéressant de voir que cette distribution est inégale à travers les ethnies. Là où le groupe A possède une forte disparité dans les performances selon le déjeuner chez les femmes, la différence semble beaucoup moins significative chez les hommes. Au contraire, le groupe E n'a quasiment aucune différence de performance selon le repas chez les hommes tandis que cette différence est grande chez les femmes.

Au delà de cela, il reste très difficile de tirer des conclusions de ce type de représentation car, bien qu'informative, on ne connaît pas assez les relations entres toutes les variables pour pouvoir donner un avis sûr.

Ce qui semble de plus en plus certain, c'est qu'il n'existe pas de pattern simple entre nos variables qualitatives et les performances, et qu'il semblerait que la prédiction des performances nécessite plus d'informations sur les étudiants que celles que nous avons. Cela sera à confirmer plus tard en testant avec des modèles prédictifs.

3.3 L'influence entre les performances aux différentes épreuves

On fait maintenant une recherche d'une influence entre chacun des trois scores. (math - reading, math - writing, reading - writing).

En effet, la figure 4 indique qu'il existe sûrement une corrélation forte entre les performances aux différents sujets. Plus un élève a un score élevé dans une matière, plus il a de chance d'être bon dans une autre matière. Les bons élèves sont bons partout, et vice-versa.

Cette intuition peut être renforcée par une ACP sur les variables quantitatives (sur les scores uniquement). L'axe principal le plus important représente alors 91% de la variance, et cet axe est une combinaison linéaire des 3 axes scores où les scores ont tous presque la même influence (voir la figure 11). Cela implique que la performance dans une matière indique aussi la performance dans les 2 autres.

```
> studentPerf.acp$loadings

Loadings:
              Comp.1 Comp.2 Comp.3
math.score     0.563  0.826
reading.score  0.574 -0.353 -0.739
writing.score  0.595 -0.440  0.673

               Comp.1 Comp.2 Comp.3
SS loadings     1.000  1.000  1.000
Proportion Var  0.333  0.333  0.333
Cumulative Var  0.333  0.667  1.000
> summary(studentPerf.acp)
Importance of components:
                          Comp.1     Comp.2     Comp.3
Standard deviation     24.6882795 7.33923987 3.15325449
Proportion of Variance  0.9052344 0.07999845 0.01476718
Cumulative Proportion   0.9052344 0.98523282 1.00000000
```

FIGURE 11 – Résultats de l'ACP sur les variables quantitatives

On peut valider cette impression en faisant un test de Pearson ou Kendall afin de vérifier que les 3 variables sont corrélées. D'après les tests, c'est bien le cas.

3.4 test de multicolinéarité

La multicolinéarité fait référence au fait que les estimations du modèle sont faussées ou difficiles à estimer avec précision en raison de l'existence de corrélations exactes ou de corrélations élevées entre les variables explicatives dans les modèles de régression linéaire. Si la variable $X_1, X_2..X_n$ est multicolinéaire, il faut avoir $n+1$ valeurs $C_0, C_1, C_2..C_n$ qui ne sont pas tous des zéros, tel que

$$c_0 + c_1x_1 + c_2x_2 + .. + c_nx_n \approx 0$$

D'autre part, s'il n'y a pas $n+1$ valeurs $C_0, C_1, C_2..C_n$ qui ne sont pas toutes nulles, de sorte que la formule ci-dessus soit vérifiée, il n'y a pas de multicolinéarité entre les variables. Si les variables sont indépendantes les unes des autres, elles ne peuvent pas être décrites par des combinaisons linéaires d'autres variables et la multicolinéarité n'existe pas.

3.4.1 Modèle 1

Ce modèle est construit sous forme $math\_score \sim gender + race + lunch + parental\_level\_edu + test\_prep\_course$

On applique deux méthodes pour vérifier l'existence de la multicolinéarité :

5

— **Calcul du nombre de conditions kappa (X)**. La multicolinéarité est mesurée par le nombre de conditions, communément exprimées par κ, et le nombre de conditions peut être défini par

$$\kappa = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

Ici λ est le valeur propre de $X^T X$ avec X la matrice de variables indépendantes.Le nombre de conditions κ peut être calculé à l'aide de la fonction intégrée $kappa()$. $k < 100$, indiquant que le degré de colinéarité est faible ; si $100 < k < 1000$, il y a plus de multicolinéarité, si $k > 1000$, il existe une multicolinéarité grave. Selon la valeur de retour de la fonction kappa, $k = 79.46793 < 100$, indiquant que le degré de colinéarité entre les variables est faible.

— **En utilisant le facteur d'inflation de la variance (VIF)**.Le VIF est similaire à la matrice de coefficients de corrélation : à travers la matrice de coefficients de corrélation, on ne peut que constater grossièrement qu'il n'y a pas de multicolinéarité, VIF permet de mesurer la gravité de la multicolinéarité. On obtient le facteur d'inflation de la variance de chaque coefficient. On considère généralement que lorsque $0 < VIF < 10$, il n'y a pas de multicolinéarité, lorsque $10VIF < 100$, il existe une forte multicolinéarité, lorsque $VIF >= 100$, la multicolinéarité est très importante. C'est une méthode plus courante pour juger la multicolinéarité. Selon la valeur de retour de la fonction vif, les valeurs vif font varier entre 1 et $1.05 < 10$, indiquant que le degré de colinéarité est faible.

Selon les deux méthodes d'évaluation de la multicolinéarité, il n'y a pas de multicolinéarité dans les variables du modèle, ce qui permet de construire le modèle 1 avec confiance.

3.4.2 Modèle 2

On rajoute donc les notes en reading et writing dans notre modèle : $math\_score - - - gender + race + lunch + parental\_level\_edu + test_prep_course + reading + writing$

```
> vif(lm_sol_math)
                    GVIF Df GVIF^(1/(2*Df))
gender              1.20  1         1.09
race                1.12  4         1.01
lunch               1.11  1         1.05
parental_level_edu  1.16  5         1.01
test_prep_course    1.25  1         1.12
writing_score      15.19  1         3.89
reading_score      13.09  1         3.61
```

À partir de la valeur de VIF, le modèle présente une multicolinéarité : $VIF_{writing} = 15.19 > 10$,$VIF_{reading} = 13.09 > 10$. Il faudra donc faire attention dans la modélisation suivante.

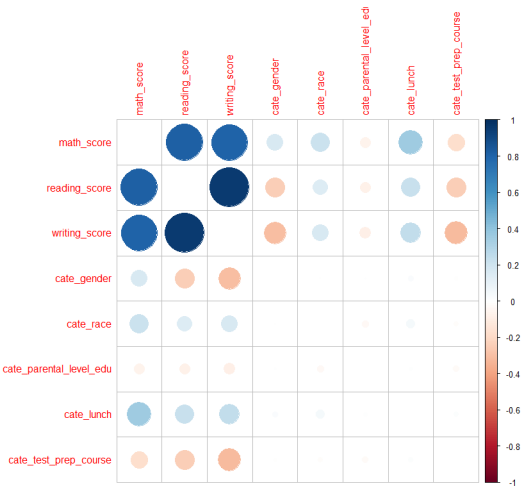

FIGURE 12 – Corrélation des variables

4 Régression

On construit alors un modèle de régression pour les variables d'étudiants.

4.1 Modèle 1

On construit un modèle de régression exprimant le score en math selon les variables qualitatives.

4.1.1 Construction du modèle

```
Call:
lm(formula = math_score ~ gender + race + lunch + parental_level_edu +
    test_prep_course, data = raw_data)

Residuals:
    Min      1Q  Median      3Q     Max
-50.357  -8.744   0.166   9.001  30.655

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       57.6305     1.8721  30.784  < 2e-16 ***
gendermale                         4.9953     0.8390   5.954 3.63e-09 ***
racegroup B                        2.0408     1.6998   1.201 0.230181
racegroup C                        2.4700     1.5918   1.552 0.121060
racegroup D                        5.3410     1.6241   3.289 0.001042 **
racegroup E                       10.1347     1.8015   5.626 2.41e-08 ***
lunchstandard                     10.8768     0.8727  12.463  < 2e-16 ***
parental_level_edubachelor's degree  1.9661     1.5020   1.309 0.190831
parental_level_eduhigh school     -4.8027     1.2971  -3.703 0.000225 ***
parental_level_edumaster's degree  2.8884     1.9382   1.490 0.136490
parental_level_edusome college    -0.5827     1.2470  -0.467 0.640431
parental_level_edusome high school -4.2487     1.3331  -3.187 0.001482 **
test_prep_coursenone              -5.4947     0.8756  -6.275 5.22e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.17 on 987 degrees of freedom
Multiple R-squared:  0.2548,    Adjusted R-squared:  0.2457
F-statistic: 28.12 on 12 and 987 DF,  p-value: < 2.2e-16
```

FIGURE 13 – 1er modèle de régression

D'après la figure 13,la p-value n'est pas toujours inférieure à 0,001, on en déduit que certaines des variables ne sont pas significatives. Maintenant analysons les résultats de la régression.

La facteur gender est significative car sa valeur p $<$ 0,001.

6

Si un élève est un homme, son score en mathématiques devrait augmenter de 5 points.

Pour le facteur *race*, seulement le groupe E est significatif : Si cet étudiant est de race E, son score en mathématiques devrait augmenter de 10 points. Mais pour les autre races, l'influence est très faible.

Pour le facteur *lunch* et *test_prep_course*, elles sont significatives : un étudiant qui prend un déjeuner standard a 10 points de plus, et un étudiant qui passe un test de préparation a 5.5 points de plus.

Pour le facteur niveau d'éducation des parents : les étudiants qui ont des parents de niveau d'éducation High School ont 4 points de moins que les autres. Cette conclusion est conforme aux conclusions tirées dans la partie analyse multivariable.

Ces conclusions sont cohérentes avec les hypothèses précédentes. Mais la valeur R^2 est 24.5%, cela veut dire que le modèle explique 24.5% de la variabilité des données de réponse autour de sa moyenne, ce qui est très faible.

4.2 Modèle 2

On construit un modèle de prédiction du score de math à partir de toutes les autres variables.

4.2.1 Construction du modèle

Dans ce modèle, la valeur p-value $= 2.2e-16 < 0.05$, on constate donc que le modèle est significatif. La valeur R_2 ajusté est 0.875, le modèle est beaucoup plus performant pour décrire la relation entre les différentes variables et le score de math.

4.2.2 Estimation d'intervalle pour les coefficients β

L'équation du modèle est :

$y_{ij} = \beta_0 + \beta_1 Gender + \beta_2 RaceGroup_A + \beta_3 RaceGroup_B + \beta_4 RaceGroup_C + \beta_5 RaceGroup_D + \beta_6 RaceGroup_E \beta_7 Lunch + \beta_8 EduParent_{bachelor} + \beta_9 EduParent_{hs} + \beta_{10} EduParent_{mst} + \beta_{11} EduParent_{somec} + \beta_{12} EduParent_{somehs} + \beta_{13} CoursPrepa + \beta_{14} Writing + \beta_{15} Reading + \epsilon_{ij}$

Nous lançons la fonction beta.int() (voir l'annexe pour le code) et obtenons l'estimation de l'intervalle pour les paramètres β comme suit :

```
> beta.int(lm_sol_math)
                     Estimate   Left  Right
(Intercept)           -11.60 -14.04 -9.16
gendermale             13.24  12.51 13.97
racegroup B             0.83  -0.52  2.19
racegroup C             0.17  -1.09  1.45
racegroup D             0.09  -1.21  1.41
racegroup E             5.07   3.63  6.52
lunchstandard           3.21   2.47  3.94
pa_bachelor            -1.04  -2.25  0.16
pa_high school          0.56  -0.48  1.61
pa_master              -1.85  -3.41 -0.29
pa_some college         0.40  -0.59  1.39
pa_some high school     0.55  -0.52  1.63
prep_course_none        3.50   2.72  4.28
writing_score           0.70   0.61  0.78
reading_score           0.26   0.18  0.34
```

Certains des intervalles de valeurs β contiennent 0 : Race Group B,C et D ; Niveau d'éducation parental bachelor, high school et college ; Cela veut dire que l'influence des variables ci-dessus sur math_score est faible.

4.2.3 Test d'hypothèses

Selon les résultats ci-dessus, certains facteurs ont moins d'impact sur les notes en mathématiques. Nous allons essayer d'expliquer en détail les résultats observés avec le moins de variables significatives possibles.

Nous essayons d'expliquer l'évolution des math_scores uniquement par des variables : gender, race group E, lunch, niveau d'éducation parental Master, cours de préparation, et writing reading scores. L'hypothèse nulle correspondante est :

$$H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_8 = \beta_9 = \beta_{11} = \beta_{12} = 0$$

Ajustons les résultats de sortie du modèle simplifié, y compris le tableau ANOVA et le tableau de coefficients.

La valeur R^2 du modèle simplifié est 0.875, le modèle interprète donc bien la variable dépendante. La somme au carré résiduelle du modèle simplifié $SSE(RM) = 28629.5$, et celle du modèle complet est $SSE(FM) = 28319.15$

$$F = \frac{[SSE(RM) - SSE(FM)]/(p+1-k)}{SSE(FM)/(n-p-1)}$$

Selon la formule ci-dessus, la valeur de la F-mesure est 1.349329 et le degré de liberté de la distribution F correspondante est 7 et 14.

La valeur critique du test est $F_{(7,14;0.05)} = 2.77$. Étant donné que la valeur $F = 1.349329$ est inférieure à 2.77, elle n'est pas significative. Par conséquent, en acceptant l'hypothèse nulle, le modèle simplifié peut remplacer le modèle complet. Cela est dû au fait que notre modèle fait appel à des variables corrélées à la variable que nous voulons prédire. Le modèle peut alors être performant en utilisant principalement ces variables corrélées et se passer des autres. L'utilisation de la régression nous montre surtout que ce type de modèle ne marche pas. On ne peut prédire un score qu'avec un autre score, ce qui n'est pas intéressant pour nous. Les variables qualitatives conservées dans le modèle simplifié sont cependant intéressantes, ce sont celles ayant le plus de potentiel pour prédire le score. La faible performance du premier modèle laisse cependant à penser que la prédiction d'un score n'est pas faisable sans un autre score.

7

5 Prédiction des notes à l'aide d'arbres de décision

Ces méthodes consistent à partitionner de manière récursive l'espace des caractéristiques en régions homogènes, c'est-à-dire correspondant à une valeur particulière de la variable à expliquer Z. Nous nous sommes basés sur l'algorithme CART implémenté sous R.

Cet algorithme est adapté à notre jeu de données puisqu'il peut être aussi appliqué sur des variables descriptives qualitatives. Ici, chaque test sera réalisé sur une variable descriptive qualitative. Cet algorithme permet aussi de prédire aussi bien des notes continues avec un arbre de régression que des notes catégorielles à l'aide d'un arbre de classification.

Nous avons commencé par séparer l'ensemble des données en un ensemble d'apprentissage (70% des données initiales) et un ensemble de test (30% des données initiales) qui servira à évaluer la performance du classifieur retenu à l'issue de l'algorithme de forêt aléatoire. Ceci nous permet d'obtenir un estimateur sans biais du risque du classifieur.

5.1 Arbre de régression

Nous obtenons un arbre complet suite à des divisions successives sur les variables explicatives (qualitatives). Nous déterminons ensuite le paramètre λ de coût-complexité correspondant à l'erreur de validation la plus faible et nous construisons le sous-arbre optimal par élagage. Nous prédisons ensuite les notes de reading.score et nous essayons alors de mesurer la performance de notre classifieur : L'ensemble des notes réelles sont comprises entre 29 et 100 tandis que les notes prédites sont seulement comprises entre 59.35 et 80.31. Le modèle ne semble donc pas adapté pour prédire les valeurs extrêmes. Cela dit, notre modèle semble adapté pour la prédiction des valeurs proches de la moyenne pour des notes comprises entre 60 et 70, où l'on a le plus d'observations. Les quartiles q1, q2 et q3 des notes prédites sont en effet quasiment identiques des quartiles observés à 5% près.

Afin d'évaluer la performance de notre classifieur, nous nous intéressons à des indicateurs de la qualité de la régression. Observons tout d'abord s'il y a une forte corrélation entre les notes prédites et les notes observées. Le coefficient de corrélation est d'environ 29% ce qui est plutôt faible. Calculons alors le MAE (Mean Absolute Error) qui donne une indication de l'écart moyen des prédictions par rapport aux valeurs observées. Ici, le MAE vaut environ 11 ce qui signifie que les notes prédites sont éloignées en moyennes de 11 points par rapport à la vraie note ce qui est relativement élevé mais pas tant que ça au vu du coefficient de corrélation obtenu. Remarque : Nous avons privilégié le MAE par rapport au RMSE car ce dernier est moins facilement interprétable que le MAE.

5.2 Arbre binaire de discrimination

Afin de pouvoir prédire dans quelle catégorie de scores les scores de l'étudiant pourraient être rangés, nous avons tenté de déterminer des classes en prenant en compte les 3 scores uniquement.

5.2.1 Identification de clusters

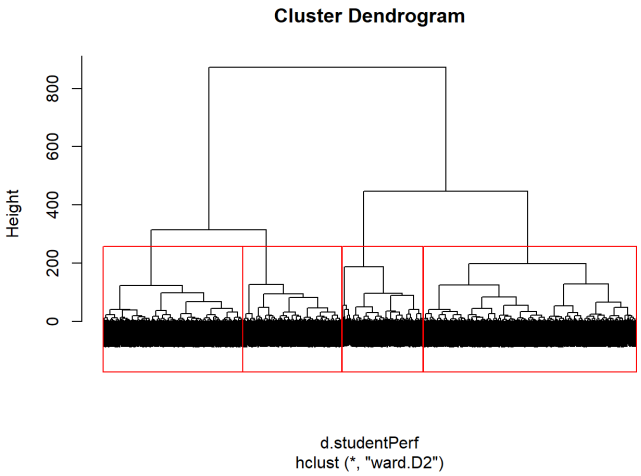

FIGURE 14 – CAH - 4 clusters

Étant donné le dendogramme 14, il parait judicieux de prendre un niveau de regroupement tel qu'il y ait 2 ou 4 classes distinctes.

Nous appliquons alors l'algorithme des k-means avec k=2 puis k=4 avec 25 configurations initiales différentes et on garde celle qui a permis d'obtenir le meilleur clustering.

La projection des centres obtenus par l'algorithme des k-means sur les axes math.score et reading.score (cf. Figure 15) indique clairement que nos scores sont très corrélés. En effet, quelque soit le nombre de classes choisies, on retrouve des centres de clusters alignés sur la diagonale. On observe le même phénomène en projetant sur l'axe writing.score. Nos 3 scores sont donc très corrélés. On constate par la même occasion qu'il ne semble pas y avoir de clusters intéressants qui se dégagent. En effet, l'algorithme se contente ici de séparer les notes en k groupes par ordre croissant des notes comme le montre la figure 15. Il sera donc difficile de savoir comment discrétiser correctement les scores en intervalles qui nous serviront pour la discrimination.

5.2.2 Arbre de classification

Etant donné les résultats peu satisfaisant des clusters obtenus par l'algorithme des k-means, nous avons choisi de discrétiser les notes en intervalles habituels avec des labels de A à F comme suit : A pour les notes appartenant à [90,100] B pour les notes appartenant à [80,90[C pour les notes appartenant à [70,80[D pour les notes appartenant à [60,70[E pour

8

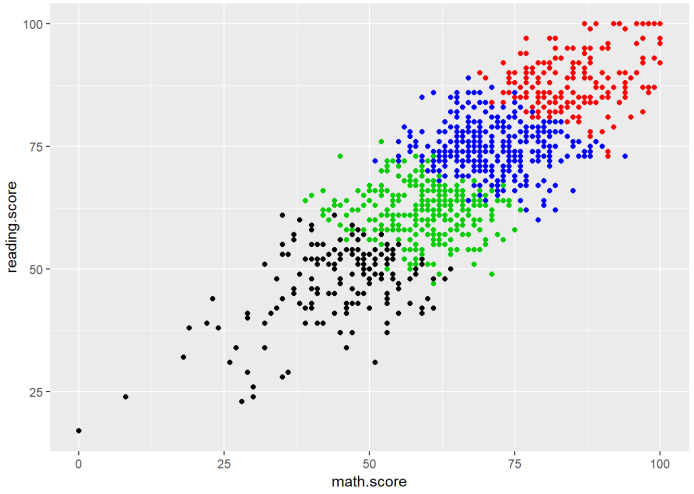

FIGURE 15 – K-means - 4 clusters

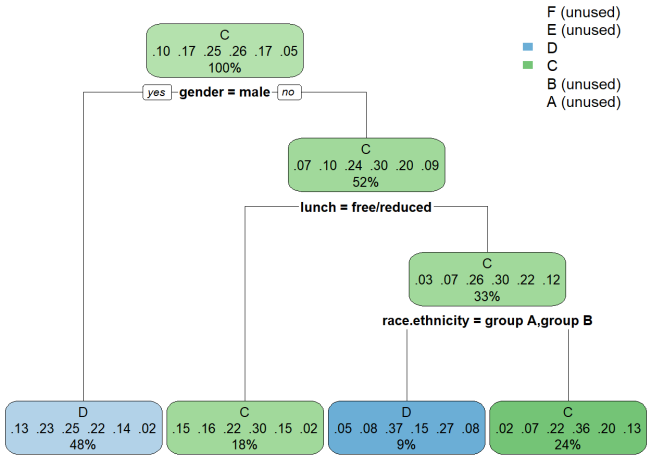

FIGURE 17 – Arbre de décision - A à F

les notes appartenant à [50,60[F pour les notes appartenant à [0,50[

Nous appliquons dans un premier temps l'algorithme de CART sur les notes reading.score.

Les arbres complets sont construits en divisant à chaque étape sur les variables descriptives de façon à minimiser : — l'inertie intra-classe I_W des nœuds fils en régression avec des notes continues. — l'indice de Gini des nœuds fils en discrimination avec des classes de notes de A à F.

Afin d'obtenir de meilleures performances des classifieurs obtenus, nous procédons ensuite à l'élagage des arbres en choisissant le paramètre λ de coût-complexité correspondant à l'erreur de validation la plus faible (voir Figure 16). On fixe le paramètre de coût-complexité dans le cas des notes "Pass"/"Fail" à 0.004.

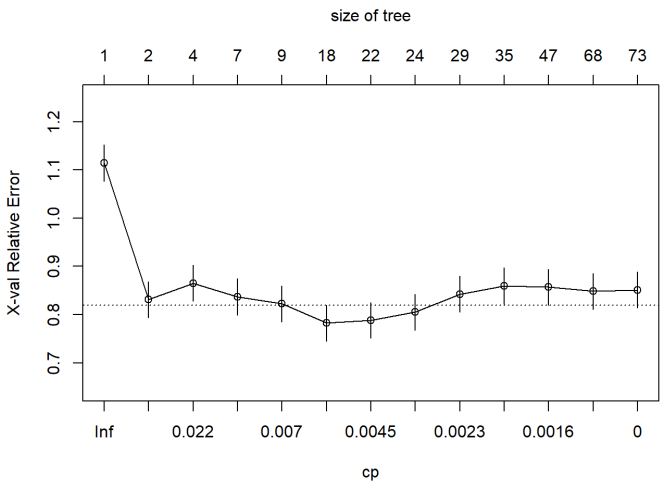

FIGURE 16 – Erreur de validation en fonction du paramètre de coût-complexité

Les sous-arbres optimaux obtenus après élagage dans chaque modèle sont présentés dans les figures 17 et 18.

La matrice de confusion nous permet d'obtenir la performance du classifieur d'arbre binaire pour le modèle associé. Il suffit de sommer les vrais positifs et les vrais négatifs et de diviser par le nombre d'observations. Avec les notes classées de A à F, on obtient 23% de notes bien classées. Avec le modèle simplifié (Pass/Fail), on obtient 61% de notes bien classées.

En observant l'arbre (voir Figure 17), on remarque que l'arbre ne classe même pas des individus dans certains clusters (A, B, E et F). De plus, à chaque division la probabilité d'être dans telle ou telle classe de note (de A à F par exemple) sont souvent très proches. Par exemple, si l'étudiant est un homme, la probabilité que sa note soit dans l'intervalle D est de 25% contre 23% dans E et 22% dans C). Ceci explique pourquoi la prédiction n'a pas été satisfaisante ce qui est probablement dû à la mauvaise qualité des clusters chosisis (notes de A à F ou bien Passage/Echec).

Nous obtenons des résultats du même ordre de grandeur avec les scores en math et en writing (voir Table 3). Remarque : Nous avons appris 20 arbres dans chaque cas et nous avons conservé le pourcentage moyen d'éléments bien classés.

	Notes de A à F	Pass/Fail
Reading	23%	61%
Writing	27%	66%
Math	28%	65%

TABLE 3 – Performance du classifieur obtenu par CART en fonction des cas

Nous décidons ensuite de simplifier le problème en attribuant les valeurs "Pass" et "Fail" si l'étudiant a un score de respectivement plus ou moins de 70 comme suit : Pass pour les notes appartenant à [70 ;100] Fail pour les notes appartenant à [0 ;70]

Les taux d'erreur de prédiction ne sont pas satisfaisants.

9

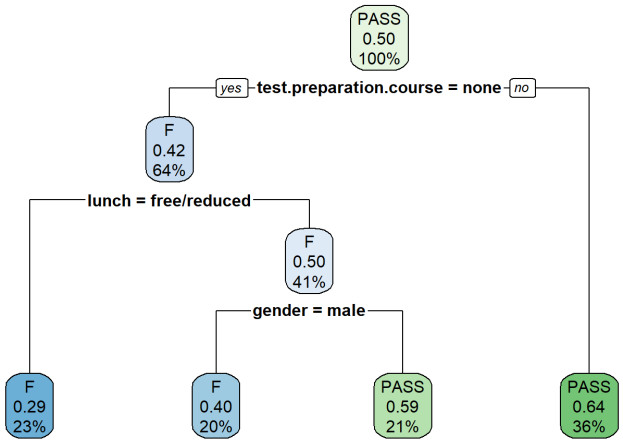

FIGURE 18 – Arbre de décision - Pass/Fail

Ceci pourrait être dû au désavantage des arbres de décision qui sont très instables en cas de légers changements dans les données de l'ensemble d'apprentissage. Nous tentons alors de prédire les notes à l'aide d'une méthode ensembliste afin de palier à ce problème.

5.2.3 Forêt aléatoire

Le principe de cet algorithme est assez simple : Après avoir appris un grand nombre d'arbres simples sur des données légèrement différentes, les différents classifieurs obtenus sont agrégés au moment de la prédiction comme suit : - Cas de la régression : La moyenne des valeurs renvoyées par les arbres pour l'observation en question est retenue. - Cas du classement : La classe majoritaire i.e la plus souvent renvoyée par les arbres pour l'observation en question est retenue. L'objectif ici est d'obtenir des arbres les plus décorrélés possibles.

Nous cherchons ensuite à optimiser l'hyperparamètre correspondant "nombre d'arbres à apprendre". Pour ce faire, on détermine le nombre d'arbre à partir duquel l'erreur semble se stabiliser. Pour prédire la note en reading avec des notes "Pass"/"Fail", il parait suffisant d'apprendre 200 arbres (voir la courbe noire sur la Figure 19). Nous avons procédé de la même façon pour apprendre les modèles en vue de prédire les autres notes en math et writing (classées de A à F ainsi que entre Pass/Fail). Les performances des classifieurs obtenus dans les différents cas sont illustrées dans la Table 4.

	Notes de A à F	Pass/Fail
Reading	25%	64%
Writing	27%	67%
Math	29%	64%

TABLE 4 – Performance du classifieur obtenu par forêt aléatoire en fonction des cas

On constate une très légère amélioration de prédiction à l'aide de forêt aléatoire par rapport aux arbres classiques

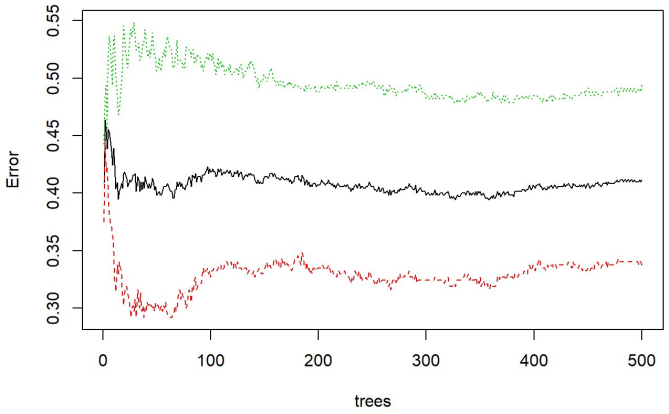

FIGURE 19 – Erreur en fonction du nombre d'arbres appris

(CART) dans certains cas. Cela dit, la différence est trop petite pour pouvoir en tirer un conclusion. La prédiction donne des résultats peu satisfaisants même avec les forêts aléatoires. Les seuls résultats globalement intéressants étants ceux portant sur le problème très simplifié. La difficulté à prédire les notes est sans doute dûe à l'absence de clusters intéressants qui se distinguent clairement dans notre jeu de données.

6 ACM

Vu que nos modèles ne permettent pas de prédire efficacement le score d'un étudiant en se basant sur les variables qualitatives dont nous disposons, nous allons essayer d'analyser les outils avec un outil plus poussé : l'analyse des correspondances multiples.

6.1 Principe de l'ACM

Nous présenterons ici succinctement le principe de l'ACM et de l'AFC en se basant sur (BAC) et (BES). L'ACM est en réalité une généralisation de l'AFC (analyse factorielle des correspondances), elle-même inspirée de l'ACP. L'AFC permet de comparer 2 variables qualitatives et de déterminer leurs liaisons. On peut dire qu'il s'agit de l'équivalent de l'ACP pour les variables qualitatives. Pour parvenir à cela, l'AFC se base sur le tableau de contingence des modalités des deux variables et en déduit les tableaux de *profils-lignes* et *profils-colonnes*. Ces profils représentent la proportion de chaque combinaison du tableau de contingence dans la somme de la ligne ou colonne.
La liaison entre les deux variables est d'autant plus grande que les profils (lignes ou colonnes) sont différents.
L'AFC réalise une ACP sur les tableaux des profils-lignes et colonnes avec une distance "du khi-deux". On obtient alors les nuages de points selon les axes factoriels.
Les ACP sont déductibles l'une de l'autre et se correspondent. On peut superposer les représentations graphiques obtenus et alors voir des relations émerger entres des moda-

10

lités.
Notons que les positions dans l'AFC sont à utiliser de façon relative. La position d'une modalité seule ne veut dire quelque chose que par rapport aux autres modalités. De plus, il faut aussi se méfier de la représentation en 2D des différentes modalités car certaines pourraient être mieux représenter selon les axes utilisés, tandis que d'autres ne se projettent peut-être pas assez dans le plan utilisé. On peut utiliser le carré du cosinus entre le vecteur de la modalité et le plan du graphique utilisé afin d'avoir une idée de la pertinence de la représentation.
L'ACM généralise le principe de l'AFC à des classes multiples.
Pour cela, on réalise une AFC sur un tableau de *Burt* des données qualitatives à tester.

Un tableau de *Burt* est une sorte de tableau de contingence dans lequel on utilise le *one-hot encoding*. Cela consiste à représenter sur chaque axe du tableau les modalités de toutes les variables à analyser.
Les résultats d'une ACM sont moins faciles à interpréter. La symétrie du tableau de *Burt* provoque des redondances qui réduisent l'impact des composants principaux obtenus. Une ACM renvoie ainsi la borne inférieure du pourcentage d'inertie expliquée par chaque axe principal obtenu. Par symétrie, les résultats obtenus de l'ACP pour les colonnes sont les mêmes que ceux obtenus pour les lignes.
Il faut aussi préciser que les variables jouent toutes le même rôle dans l'ACM. Si l'une des variables a une importance particulière, une "classe", celle-ci ne peut éventuellement apparaître qu'à posteriori dans l'interprétation. Si c'est le cas, cela signifie que l'on a bien analysé les données et trouvé la relation permettant de déterminer la variable-classe voulue.

6.2 Notre ACM

Étant donné que notre jeu de données est majoritairement constitué de variables qualitatives, il semble intéressant de faire une ACM afin de confirmer nos résultats et intuitions. On fait l'ACM sous R en utilisant le package "ade4". On divisera les scores en 5 factors "A","B"... etc comme précédemment.

6.2.1 Sans les scores

Afin d'analyser nos données, on retire les scores de notre ACM. On obtient plusieurs axes factoriels qui s'expliquent surtout par les anciennes variables parental.level.of.education et race.ethnicity.

On peut remarquer dès le début que les axes principaux obtenus par ACM représentent tous l'inertie de façon assez bien répartie (environ 8-9% et 12 axes obtenus). Il s'agit d'une borne inférieure dans le cas d'une ACM. Il faut donc faire preuve de prudence quant à nos interprétations mais cela reste normal.

Le cercle de corrélation obtenu ne nous apprend rien d'intéressant pour prédire les scores et nous donne quelques informations sur les corrélations entres nos variables ce qui valide certaines de nos intuitions de l'analyse exploratoire.

La projection de nos individus dans les plans factoriels de l'ACM 20 montre qu'il y a bien une séparation légère entre les différents scores selon nos axes factoriels. On peut voir que certaines variables (ethnie, niveau d'étude des parents...) ont bel et bien une influence sur les scores.

Les barycentres des différentes modalités des scores sont alignés de façon ordonnée, on a donc une augmentation graduelle du score moyen selon la position dans le plan factoriel.

Cependant, on remarque que nos modalités de scores sont très proches et se superposent beaucoup. On ne semble alors pas pouvoir trouver de frontière satisfaisantes qui permette de classer les individus, mais seulement certaines variables qui augmentent la probabilité d'avoir un bon ou mauvais score.

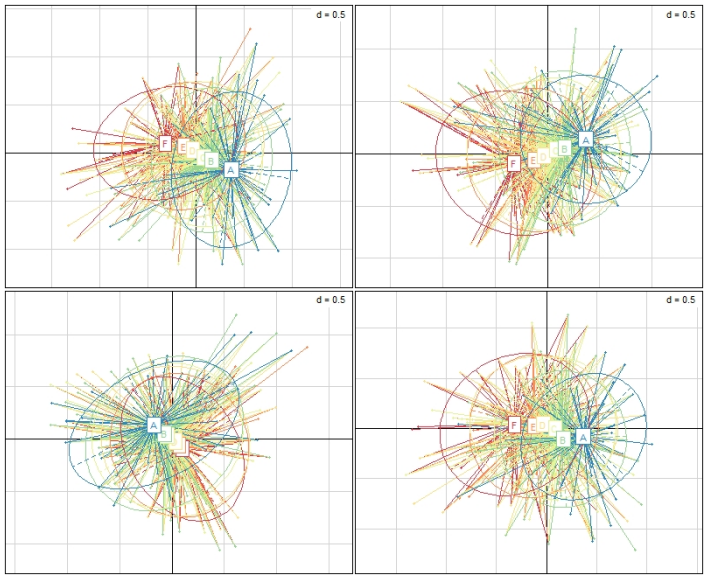

FIGURE 20 – Projection des individus dans différents plans factoriels (1-2, 1-3, 2-3, 1-4)

6.2.2 Avec les scores

Afin de vérifier si certaines variables pourraient être corrélées avec les scores, on fait l'ACM avec les scores cette fois-ci. Les axes factoriels ne semblent pas donner d'informations intéressantes, si ce n'est que les scores semblent être peu affectés par les autres variables. Les 4 premiers axes sont en effet presque exclusivement expliqués par les scores.

Afin de vérifier si un cluster pourrait se dégager et aider à prédire les scores, on affiche les individus sur nos 2 premiers axes factoriels (voir figure 21). On affiche ensuite leurs modalités selon plusieurs variables. On remarque qu'aucun cluster spécifique ne semble se dégager. On peut cependant retrouver les intuitions vus lors de l'analyse multivariée : certaines variables ont un léger impact sur les scores mais celui-ci est

11

tellement minime qu'on ne peut pas l'utiliser pour prédire les scores (les modalités des 3 scores sont disponibles sur les 2 plots du bas et celui au centre à droite de la figure 21).

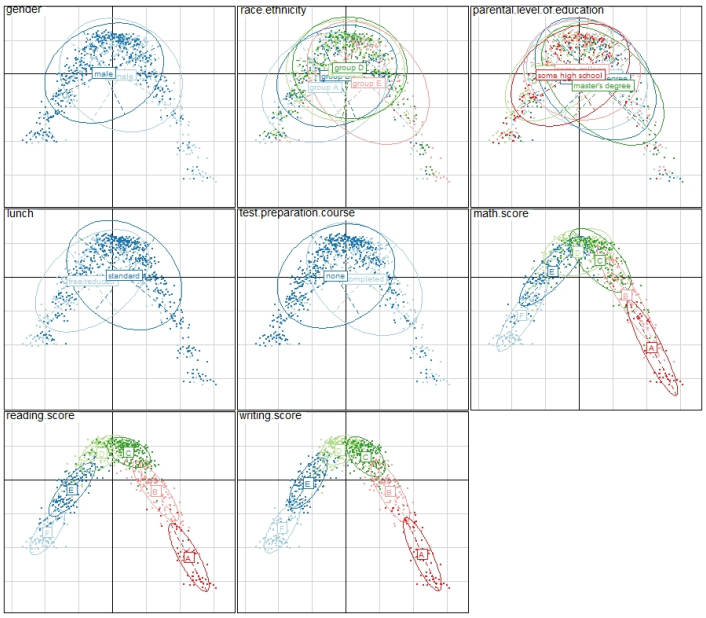

FIGURE 21 – Individus selon nos axes factoriels 1 et 2 (ACM avec scores)

Il semblerait qu'on ne puisse pas, avec nos données, trouver des axes factoriels pour séparer les individus obtenus selon les scores. Les scores semblent être trop peu corrélés aux variables dont nous disposons pour être capable de les prédire efficacement.

7 Conclusion

Au cours de nos recherches, nous avons pu voir l'influence des différentes variables qualitatives de notre jeu de donnée sur les scores. Plus que cela, nous avons pu voir que l'impact de ces données sur le score était minime et ne suffisait pas du tout à expliquer ni prédire les scores. L'analyse exploratoire révéla que les scores étaient tous corrélés, ce qui fut confirmé par différents tests, une ACP et même une régression.

Les modèles que nous avons utilisés ont montrés par leurs faibles performances que les variables qualitatives ne pouvaient pas prédire efficacement les scores des étudiants. Une randomForest ne parvenait en effet pas à trouver de pattern entre les élève, ni une régression, et l'ACM montra bien que les données qualitatives ne pouvaient pas obtenir de résultats correct car ses axes d'inerties n'expliquaient que mal les différents scores. Si nous disposions de plus d'informations, nous pourrions peut-être trouver un pattern, mais cela nécessiterait plus de dimensions et donc d'individus. Il est très probable qu'il soit juste impossible de prédire les scores seulement avec nos variables.

8 Annexe

8.1 Fonction : beta.int

```r
beta.int <- function(fm,alpha = 0.05){
  A <- summary(fm)$coefficients
  df <- fm$df.residual
  left <- A[,1]-A[,2]*qt(1-alpha/2,df)
  right <- A[,1]+A[,2]*qt(1-alpha/2,df)
  rowname <- dimnames(A)[[1]]
  colname <- c("Estimate","Left","Right")
  matrix(c(A[,1],left,right),ncol = 3,
         dimnames = list(rowname,colname))
}
```

Références

[BAC] Alain Baccini. *Statistique Descriptive Multidimensionnelle.* www.math.univ-toulouse.fr/~baccini/zpedago/asdm.pdf, Institut de Mathématiques de Toulouse, Toulouse, 2010.

[BES] Philippe Besse. *Cours sur l'AFC.* www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-explo-afc.pdf, Institut de Mathématiques de Toulouse, Toulouse, 2014.

12