UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Techniques Quantitatives de Gestion

Higher Education Students Performance Evaluation

La prévision de la note (Grade)

NGUYET Anh Vu
DUONG Bao Chau

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

INTRODUCTION

Existent-ils aujourd'hui des variables ou des facteurs qui nous permettraient de prédire notre réussite à un examen ou tout simplement notre réussite au niveau scolaire ? Telle est la question que nous nous sommes posés lorsque nous nous sommes mis en groupe.

De nombreuses études indiquent que le milieu social à savoir le fait de naître dans une famille plutôt riche ou plutôt pauvre reste le premier facteur expliquant notre réussite scolaire. En effet, dans une famille riche, l'enfant ou bien l'étudiant est susceptible d'avoir un accès facilité à la culture et ce dès le plus jeune âge, tandis que dans une famille pauvre se serait l'inverse. Au sein d'une famille c'est notamment la CSP des parents, leurs diplômes, la taille de la famille, le sexe de l'enfant ou encore le rang dans la fratrie qui expliqueraient et qui permettraient de prévoir la réussite d'un enfant. D'autres facteurs, tels que la motivation, l'estime de soi ou encore le type d'éducation reçu peuvent venir expliquer la réussite d'un étudiant. Mais qu'en est-il réellement ? De ce fait, nous avons choisi cette thématique là pour construire notre algorithme de prévision. Aussi, nous avons tenté de mettre en évidence des facteurs qui permettraient de prédire notre réussite scolaire.

Comment avons-nous procédé ? Afin de mener à bien notre projet, nous avons récupéré une base de données sur Kaggle pour pouvoir mettre en place nos algorithmes. Cette base contient 146 réponses et comporte 32 variables différentes telles que le métier de la mère et du père, la fréquence de lecture ou encore le fait de se rendre ou non en cours. A savoir que, pour notre projet nous avons choisis de supprimer les variables student ID, siblings, cumulative grade et expected cumulative grade.

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

BASE DE DONNÉES

Les variables de la base de données

1. Student ID	1. STUDENT_ID
2. Student Age (1: 18-21, 2: 22-25, 3: above 26)	2. AGE
3. Sex (1: female, 2: male)	3. GENDER
4. Graduated high-school type: (1: private, 2: state, 3: other)	4. HS_TYPE
5. Scholarship type: (1: None, 2: 25%, 3: 50%, 4: 75%, 5: Full)	5. SCHOLARSHIP
6. Additional work: (1: Yes, 2: No)	6. WORK
7. Regular artistic or sports activity: (1: Yes, 2: No)	7. ACTIVITY
8. Do you have a partner: (1: Yes, 2: No)	8. PARTNER
9. Total salary if available (1: USD 135-200, 2: USD 201-270, 3: USD 271-340, 4: USD 341-410, 5: above 410)	9. SALARY
10. Transportation to the university: (1: Bus, 2: Private car/taxi, 3: bicycle, 4: Other)	10. TRANSPORT
11. Accommodation type in Cyprus: (1: rental, 2: dormitory, 3: with family, 4: Other)	11. LIVING
12. Mother's education: (1: primary school, 2: secondary school, 3: high school, 4: university, 5: MSc., 6: Ph.D.)	12. MOTHER_EDU
13. Father's education: (1: primary school, 2: secondary school, 3: high school, 4: university, 5: MSc., 6: Ph.D.)	13. FATHER_EDU
14. Number of sisters/brothers (if available): (1: 1, 2:, 2, 3: 3, 4: 4, 5: 5 or above)	14. SIBLINGS
15. Parental status: (1: married, 2: divorced, 3: died - one of them or both) ***Listed as "Kids"…woops	15. KIDS
16. Mother's occupation: (1: retired, 2: housewife, 3: government officer, 4: private sector employee, 5: self-employment, 6: other)	16. MOTHER_JOB
17. Father's occupation: (1: retired, 2: government officer, 3: private sector employee, 4: self-employment, 5: other)	17. FATHER_JOB
18. Weekly study hours: (1: None, 2: <5 hours, 3: 6-10 hours, 4: 11-20 hours, 5: more than 20 hours)	18. STUDY_HRS
19. Reading frequency (non-scientific books/journals): (1: None, 2: Sometimes, 3: Often)	19. READ_FREQ
20. Reading frequency (scientific books/journals): (1: None, 2: Sometimes, 3: Often)	20. READ_FREQ_SCI
21. Attendance to the seminars/conferences related to the department: (1: Yes, 2: No)	21. ATTEND_DEPT
22. Impact of your projects/activities on your success: (1: positive, 2: negative, 3: neutral)	22. IMPACT
23. Attendance to classes (1: always, 2: sometimes, 3: never)	23. ATTEND
24. Preparation to midterm exams 1: (1: alone, 2: with friends, 3: not applicable)	24. PREP_STUDY
25. Preparation to midterm exams 2: (1: closest date to the exam, 2: regularly during the semester, 3: never)	25. PREP_EXAM
26. Taking notes in classes: (1: never, 2: sometimes, 3: always)	26. NOTES

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

27. Listening in classes: (1: never, 2: sometimes, 3: always)	27. LISTENS
28. Discussion improves my interest and success in the course: (1: never, 2: sometimes, 3: always)	28. LIKES_DISCUSS
29. Flip-classroom: (1: not useful, 2: useful, 3: not applicable)	29. CLASSROOM
30. Cumulative grade point average in the last semester (/4.00): (1: <2.00, 2: 2.00-2.49, 3: 2.50-2.99, 4: 3.00-3.49, 5: above 3.49)	30. CUML_GPA
31. Expected Cumulative grade point average in the graduation (/4.00): (1: <2.00, 2: 2.00-2.49, 3: 2.50-2.99, 4: 3.00-3.49, 5: above 3.49)	31. EXP_GPA
32. Course ID	32. COURSE ID
33. OUTPUT Grade (0: Fail, 1: DD, 2: DC, 3: CC, 4: CB, 5: BB, 6: BA, 7: AA)	33. GRADE

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

CODES DANS LE LOGICIEL R

- Data featuring

```
library(rio)
eleve=import(file.choose())
eleve=eleve[,-c(1,14,28,30,31)]


 eleve$AGE=as.factor(eleve$AGE)
 eleve$GENDER=as.factor(eleve$GENDER)
 eleve$HS_TYPE=as.factor(eleve$HS_TYPE)
 eleve$SCHOLARSHIP=as.factor(eleve$SCHOLARSHIP)
 eleve$WORK=as.factor(eleve$WORK)
 eleve$ACTIVITY=as.factor(eleve$ACTIVITY)
 eleve$PARTNER=as.factor(eleve$PARTNER)
 eleve$SALARY=as.factor(eleve$SALARY)
 eleve$TRANSPORT=as.factor(eleve$TRANSPORT)
 eleve$LIVING=as.factor(eleve$LIVING)
 eleve$MOTHER_EDU=as.factor(eleve$MOTHER_EDU)
 eleve$FATHER_EDU=as.factor(eleve$FATHER_EDU)
 eleve$MOTHER_JOB=as.factor(eleve$MOTHER_JOB)
 eleve$KIDS=as.factor(eleve$KIDS)
 eleve$MOTHER_JOB=as.factor(eleve$MOTHER_JOB)
 eleve$FATHER_JOB=as.factor(eleve$FATHER_JOB)
 eleve$STUDY_HRS=as.factor(eleve$STUDY_HRS)
 eleve$READ_FREQ=as.factor(eleve$READ_FREQ)
```

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

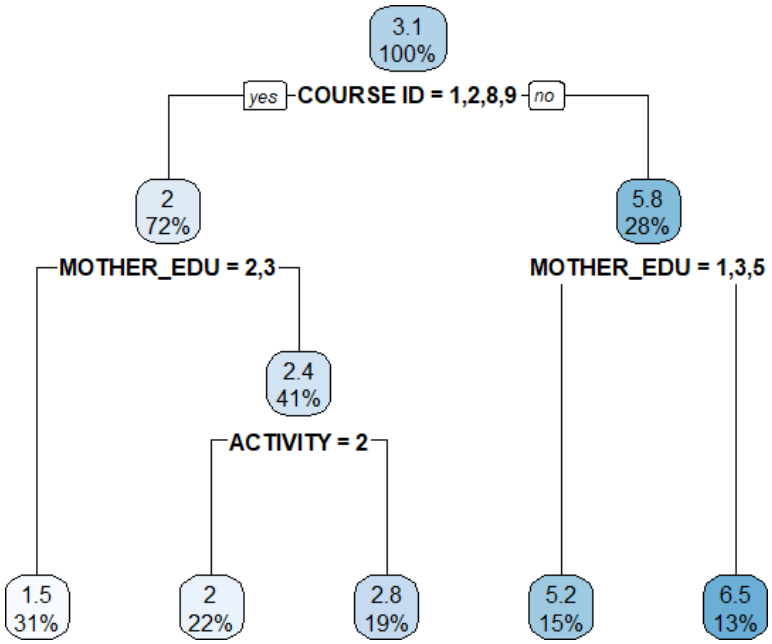
```
eleve$READ_FREQ=as.factor(eleve$READ_FREQ)
eleve$READ_FREQ_SCI=as.factor(eleve$READ_FREQ_SCI)
eleve$ATTEND_DEPT=as.factor(eleve$ATTEND_DEPT)
eleve$ATTEND=as.factor(eleve$ATTEND)
eleve$IMPACT=as.factor(eleve$IMPACT)
eleve$PREP_STUDY=as.factor(eleve$PREP_STUDY)
eleve$PREP_EXAM=as.factor(eleve$PREP_EXAM)
eleve$NOTES=as.factor(eleve$NOTES)
eleve$LISTENS=as.factor(eleve$LISTENS)
eleve$LIKES_DISCUSS=as.factor(eleve$LIKES_DISCUSS)
eleve$CLASSROOM=as.factor(eleve$CLASSROOM)
eleve$`COURSE ID`=as.factor(eleve$`COURSE ID`)
eleve$GRADE=as.numeric(eleve$GRADE)
```

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

ARBRE

La prévision avec la méthode Arbre

```
library(rpart)
library(rpart.plot)
set.seed(216)
index=sample(145,60)
testeleve=eleve[index,]
traineleve=eleve[-index,]
arbre=rpart(GRADE~.,data=traineleve,method="anova")
rpart.plot(arbre)
```



UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

La corrélation de la méthode Arbre

```
probaArbre=predict(arbre,newdata=testeleve)

library(corrr)

cor.test(probaArbre,testeleve$GRADE)
```

```
        Pearson's product-moment correlation

data:  probaArbre and testeleve$GRADE
t = 7.9223, df = 58, p-value = 8.268e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5716326 0.8240000
sample estimates:
      cor
0.7209153
```

Interprétation du résultat trouvé avec la méthode Arbre

Dans un premier temps, nous avons réalisé un arbre de décision pour prédire la note obtenue par les élèves. Ainsi par exemple, nous pouvons conclure pour les cours autre que 1,2,8,9; si l'élève a une mère avec un niveau d'éducation différent de primaire, lycée ou master of science; alors cet élève a plus de chance de finir avec une bonne note de 6.5 (entre BA et AA). Ces étudiants représentent 13% de la base de données.

A contrario les élèves suivant les cours 1,2,8,9; qui ont une mère avec un niveau collège ou lycée sont ceux qui ont une note 1.5 entre DD et DCl). Ces élèves représentent 31% de l'échantillon. Enfin, nous ne développons pas ici la totalité des prédictions de cet arbre mais il en est que ce raisonnement est applicable à l"ensemble des chemins de l'arbre.

Par la suite, nous avons réalisé un test de corrélation de Pearson afin de déterminer si nos résultats étaient agrégeables. Il en ressort que le coefficient de corrélation est de 0.7 environ donc assez proche de 1. Nous pouvons donc conclure une relation entre les variables Grade, Course ID, Mother_edu et Activity. Mais plus généralement, nous pouvons affirmer que notre arbre de décision permet effectivement de prévoir qu'elle note sera obtenue par un élève en fonction de divers facteurs.

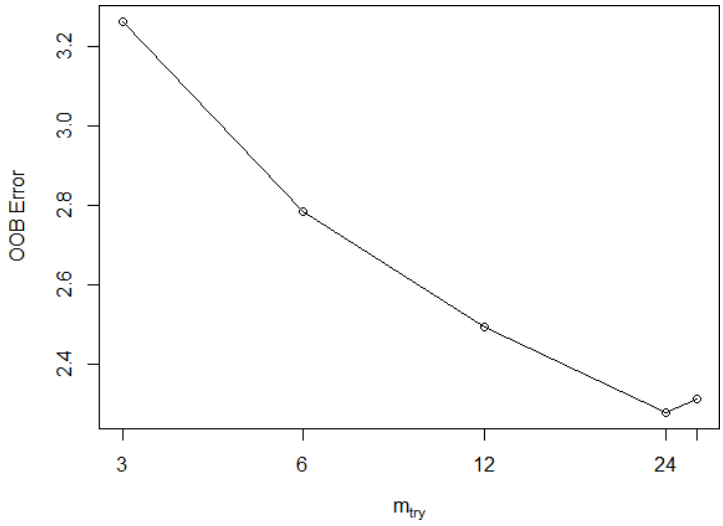
UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

FORÊT

La prévision avec la méthode Forêt

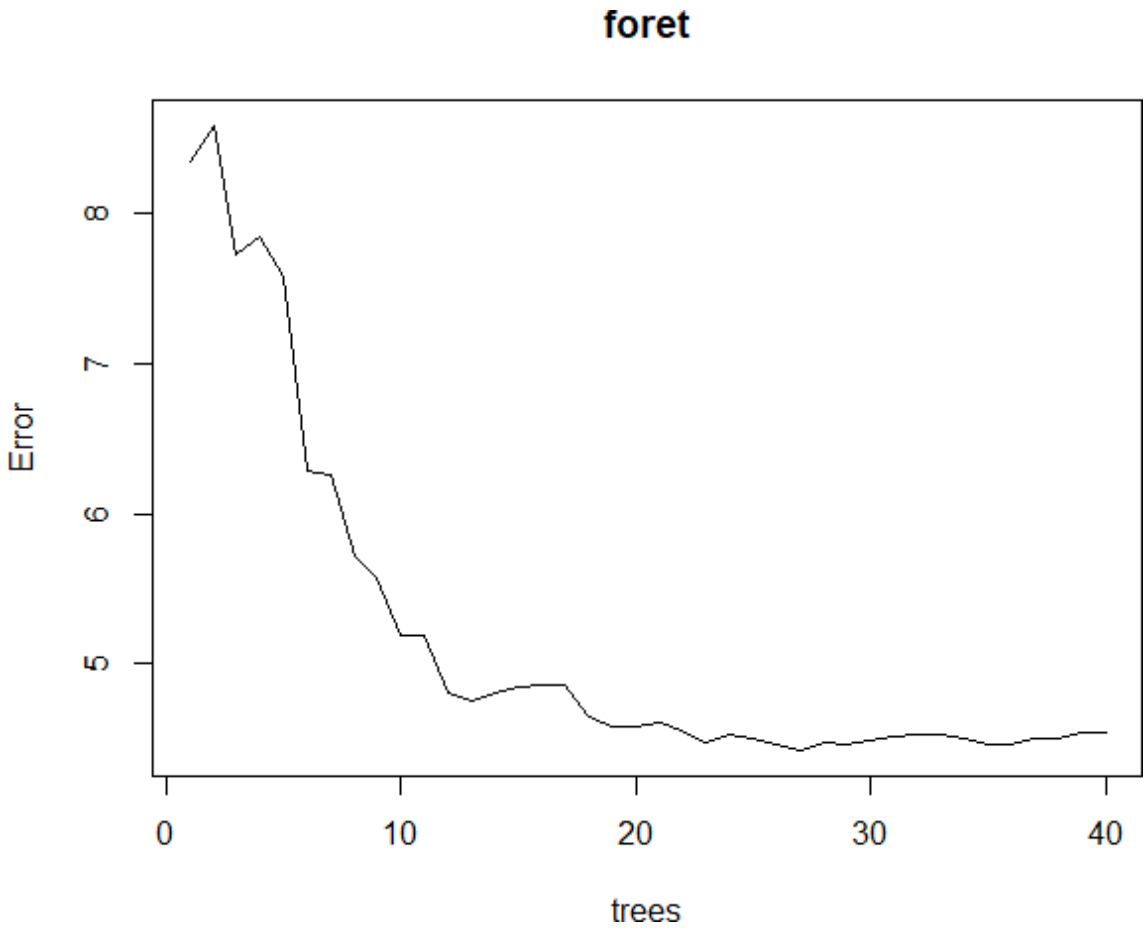
```
library(randomForest)

set.seed(3)

tuneRF(traineleve[,-28],traineleve[,28],mtryStart=6,ntreeTry=300,stepfactor=7)
```

```
mtry = 6  OOB error = 2.783547
Searching left ...
mtry = 3        OOB error = 3.262294
-0.171992 0.05
Searching right ...
mtry = 12       OOB error = 2.493355
0.1042524 0.05
mtry = 24       OOB error = 2.277476
0.08658165 0.05
mtry = 27       OOB error = 2.31135
-0.01487342 0.05
   mtry OOBError
3     3 3.262294
6     6 2.783547
12   12 2.493355
24   24 2.277476
27   27 2.311350
```



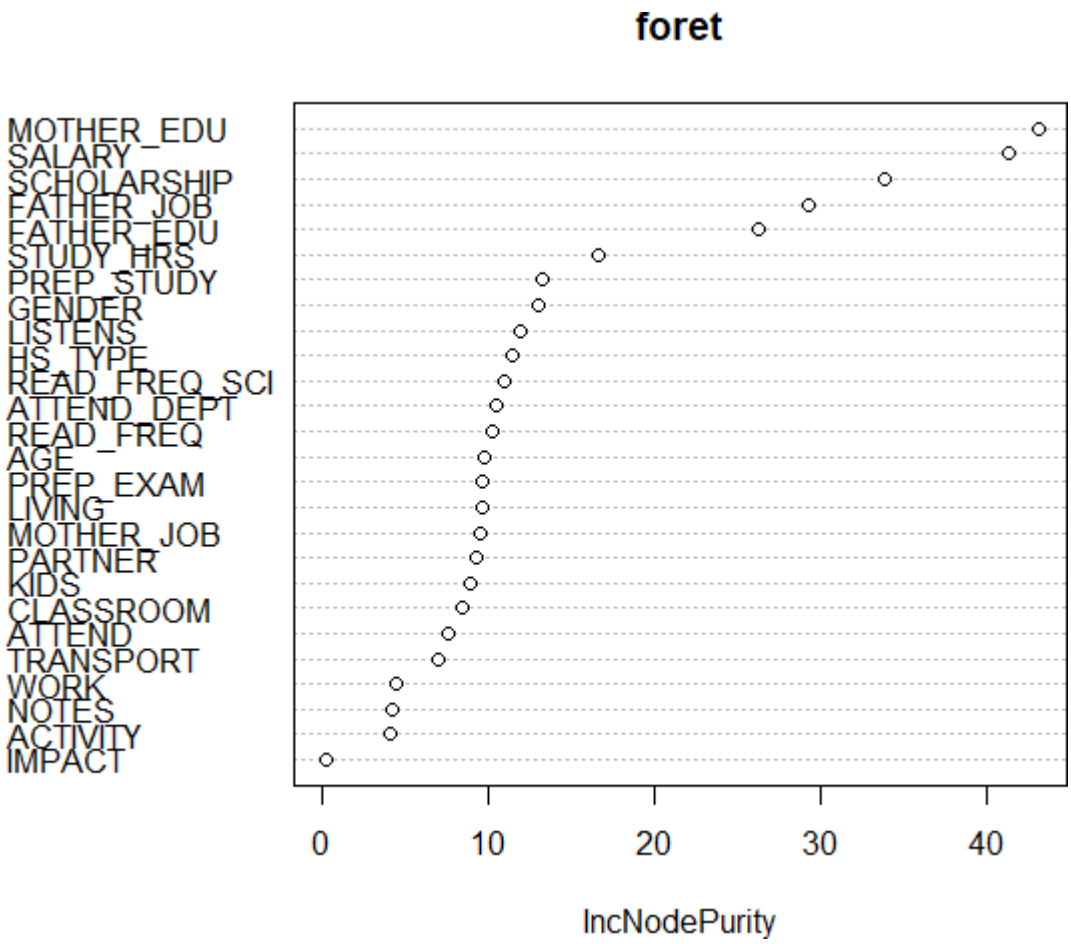

```
set.seed(150)

foret=randomForest(GRADE~.,data=traineleve[,-27],mtry=24,ntree=40)

plot(foret)
```

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE




```
set.seed(3)
foret=randomForest(GRADE~.,data=traineleve[-27],mtry=24,ntree=25)
varImpPlot(foret)
```

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE



La corrélation de la méthode Forêt

```
forestprev=predict(foret,newdata=testeleve)

library(corrr)

cor.test(forestprev,testeleve$GRADE)
```

```
        Pearson's product-moment correlation

data:  forestprev and testeleve$GRADE
t = 1.5531, df = 58, p-value = 0.1258
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.05699482  0.43183530
sample estimates:
      cor
0.1998218
```

Interprétation du résultat trouvé avec la méthode Forêt

Deuxièmement, nous avons réalisé une forêt (voir annexe forêt). Tout d'abord, la fonction tuneRF nous a permis de déterminer le mtry (le nombre de variables sélectionnées pour chaque nœud).Dans notre cas, nous avons ainsi fixé mtry=24 car il correspond au mtry avec le moins de risque d'erreur : en effet OOB error = 2.27. Par la suite, nous avons déterminer le ntree qui minimise le risque d'erreur par tâtonnement : il est ainsi fixé à 25 arbres.

Après avoir réalisé la forêt avec ces constante (ntree=25 et mtry=24), nous pouvons observer sur la modélisation que les variables : scholarship, mother et father edu, father job et salary sont celles qui permettent le mieux de prédire les notes. Toutefois après la réalisation du test de corrélation de Pearson on trouve que ces résultats ne sont pas généralisables car le coefficient est proche de 0. A l'issue de cette méthode de prédiction, nous retenons pour l'heure, l'arbre de décision comme meilleurs outils de prédiction des notes des élèves.

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

LOGIT

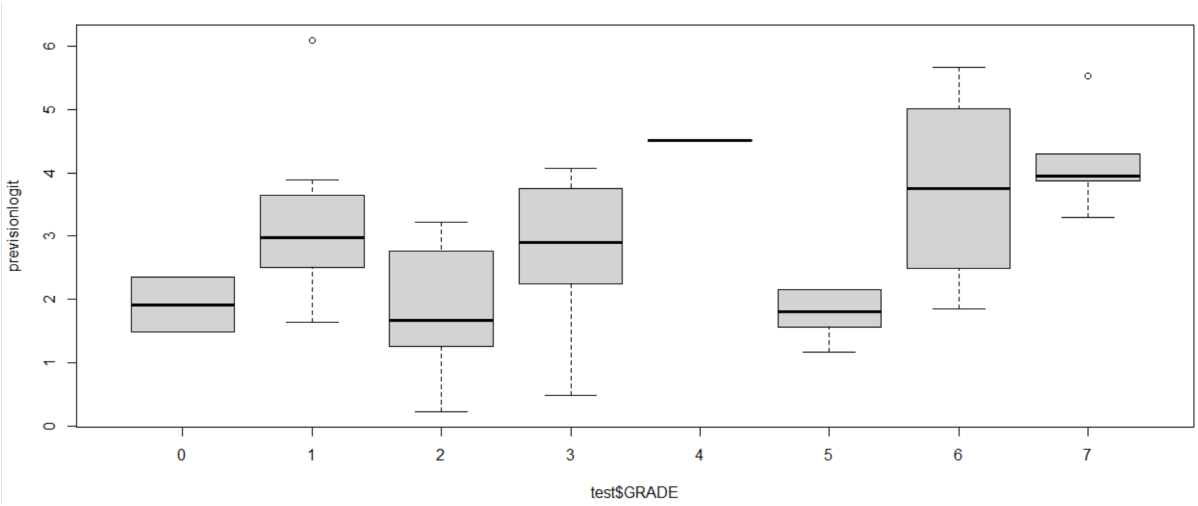
La prévision avec la méthode LOGIT

```
library(ca)
set.seed(31)
index=sample(145,40)
train=eleves[-index,]
test=eleves[index,]
logit=glm(GRADE~.,data=train)
logit=step(logit,direction="both")
previsionlogit=predict(logit,newdata=test)
previsionlogit
```

```
       49        75        43        55        27        62         9
3.6472497 3.2922179 2.9722758 0.4819957 6.0919961 1.5644578 2.1553249
       46       122        84       130       126        14        77
3.7482050 2.3539478 4.2994819 2.2538335 2.8551507 1.6339715 4.5121621
      108       124       142         6       112        66        64
3.1357079 3.4514212 1.8167118 2.6558779 1.6727472 0.9126360 2.1585175
       90        93       103        36        26        54        70
4.3769645 5.5220060 3.8719610 2.5084173 3.4095060 1.8002539 1.1694944
      101       133       129        99        42       117        22
5.6567419 2.3835915 1.4864909 3.9464679 1.7493118 3.8920150 2.5666155
       50        48        11        39        74
1.6003253 4.0744728 0.2262537 3.2195097 1.8570711
```

Troisièmement, nous avons fait une prévision de la variable GRADE avec le LOGIT. Cette méthode consiste à séparer la base de données en deux groupes train, test. Le groupe train composé de 105 individus est choisi au hasard avec une graine de 31. Nous trouvons ci-dessus la note obtenue dans la prévision pour les individus du groupe test.

```
boxplot(previsionlogit~test$GRADE)
```



En Abscisse = le GRADE de test
En Ordonnée = les prévisions de GRADE

Interprétation du résultat trouvé

La prévision est faite à partir des informations du groupe train, puis le résultat trouvé est comparé au groupe test.

Pour le groupe ayant failli, la prévision prévoit un intervalle de [1.5 ; 2.5], avec une médiane à 2. Avec cette méthode, la moitié des individus sont au-dessus de 2, alors que leur vraie note est 0. Pour ceux qui en réalité ont obtenu DD, la prévision a une médiane situé à DC, la prévision a également une meilleure note qu'en réalité.

Dans la prévision pour les individus ayant obtenu DC, la médiane se trouve proche de 2, ce qui semble assez correct. Cependant ce groupe d'individus ayant réellement 2 a dans la prévision une note moins élevée que celle du groupe 1.

L'intervalle des notes prévues pour les individus ayant réellement 3 est grande, quelques individus ont une note prévue très baisse. Mais une majorité des personnes restent dans la boîte à moustaches se trouvant entre [2, 4].

Le groupe d'individus qui ont obtenu 4 dans le test est trop peu nombreux dans la base de données, mais nous pouvons tout de même observer que la médiane est à 5, ce qui n'est pas correct.

La prévision pour le groupe ayant 5 dans le test est trop loin de la réalité, l'intervalle des notes obtenu est petit mais elle se trouve à une note médiane de 2.. La prévision pour les groupes ayant 6 et 7 du test est également incorrecte. L'intervalle pour le groupe 6 est trop grand, et la médiane se trouve à 3,5 pour un groupe ayant réellement 6. Pour les individus appartenant au groupe 7 du test, ils obtiennent une note médiane de 4.

Cette méthode Logit n'a pas été efficace pour prévoir les notes, surement d'autres variables nécessaires n'ont pas été prises en compte dans la prévision. De plus, les intervalles des

différentes boîtes à moustaches se superposent, des individus de groupe différents ont obtenu une même note à la prévision.

summary(logit)

```
Call:
glm(formula = GRADE ~ AGE + GENDER + HS_TYPE + WORK + PARTNER +
    SALARY + TRANSPORT + MOTHER_EDU + FATHER_JOB + READ_FREQ +
    ATTEND_DEPT + IMPACT + PREP_STUDY + PREP_EXAM + NOTES + LISTENS +
    LIKES_DISCUSS + `COURSE ID`, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2283  -1.0182  -0.0914   1.1179   3.4481

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.65004    2.15993  -0.764 0.447000
AGE           -0.42651    0.31375  -1.359 0.177571
GENDER         1.74708    0.39852   4.384 3.28e-05 ***
HS_TYPE        0.49630    0.33187   1.495 0.138446
WORK           0.59487    0.36751   1.619 0.109190
PARTNER       -0.58357    0.34955  -1.669 0.098657 .
SALARY        -0.22175    0.17041  -1.301 0.196639
TRANSPORT     -0.36866    0.16644  -2.215 0.029406 *
MOTHER_EDU     0.22166    0.13972   1.586 0.116314
FATHER_JOB    -0.19578    0.12244  -1.599 0.113485
READ_FREQ      1.16132    0.30440   3.815 0.000256 ***
ATTEND_DEPT   -1.02018    0.46037  -2.216 0.029331 *
IMPACT        -0.38597    0.29419  -1.312 0.193023
PREP_STUDY    -0.40553    0.31384  -1.292 0.199763
PREP_EXAM      0.55616    0.41791   1.331 0.186773
NOTES         -0.58086    0.32410  -1.792 0.076613 .
LISTENS        0.71853    0.27202   2.641 0.009805 **
LIKES_DISCUSS  0.36274    0.27554   1.316 0.191524
`COURSE ID`    0.25940    0.05884   4.409 2.99e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 2.565888)

    Null deviance: 497.56  on 104  degrees of freedom
Residual deviance: 220.67  on  86  degrees of freedom
AIC: 415.96

Number of Fisher Scoring iterations: 2
```

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Interprétation du résultat trouvé avec le modèle AIC

La fonction summary(logit) nous permet de trouver les différentes variables jouant significativement sur la variable GRADE, le degré de significativité est donné par les étoiles à droite des colonnes, mais aussi par les points. De plus, le p value ne doit pas dépasser les 5%, et l'AIC doit être au plus faible possible pour que ça soit significatif.

Cependant, au niveau de notre cas, notre AIC reste tout de même assez grand. Notre modèle logit est donc pas très bon.

Les variables significatives sont rassemblés dans le tableau ci-dessus:

.		*		**		***	
PARTNER NOTES	- 0.58357 - 0,58086	TRANSPORT ATTEND_DEPT	- 0,36866 - 1,02018	LISTENS	0,71853	GENDER READ_FREQ COURSE ID	1,74708 1,16132 0,25940

Seul les variables ayant un degré d'erreur Pr(> | t |) inférieur à 5% sont à prendre en compte. Les variables ayant un '.' ne sont donc pas significatives, les significativité sont à partir de 1 étoile.

*** Le sexe de l'individu aura un impact positive sur la prévision de la note, en faveur des hommes → Cela pourrait s'expliquer par une inégalité dans le pays, les femmes portent moins d'importance à leur étude. Ici, nous avons *** qui signifie que cette variable est significative, la valeur de l'Estimate est à 1,74. Le sexe a donc un impacte significatif sur les notes, un homme aurait une meilleur note de 1,74.

* Le temps de transport quant à lui n'a qu'un léger effet sur la note attendue. On dispose d'un estimate négatif de 0,36. Ce qui représente donc une relation légèrement négative. → Cette relation négative peut être causée par le temps de trajet, plus l'individu passe du temps dans les transports , moins il aura du temps pour réviser. Ce qui impacte ainsi de manière indirecte la note qu'il pourra obtenir.

*** Prendre plus temps dans la lecture augmente la note attendue dans la prévision, même si ce sont des livres pas en rapport avec la formation. En effet, le fait de lire fréquemment est une variable très significative, avec un estimate de 1,16.→ Ce dont nous pouvons en tirer c'est donc : le fait de passer du temps à lire nous apporte des connaissances supplémentaires et que cela peut venir jouer sur la note qu'on pourrait avoir.

* Le fait de ne pas participer aux séminaires, conférences liés au département diminue la note prévue. On a ici une relation négative entre le fait de ne pas participer aux conférences, séminaires etc et les notes obtenues. L'estimate est à -1,02. Ce qui démontre que les étudiants qui sont plus impliqués dans leur formation, qui essaient au maximum de participer à tous les événements, ont plus de chance d'avoir une meilleure note. Ainsi, moins on se

trouve être impliqué, moins la note sera bonne. Ce qui reste tout de même logique car l'étudiant ne s'intéresse pas à sa formation.

** Écouter en cours joue aussi sur le note, plus l'élève écoute en cours plus dans la prévision il aura une bonne note. Ce qui est tout à fait logique car l'élève passe du temps à assimiler des connaissances que le cours lui apporte. Le cours représentant aussi sa base de connaissance pour effectuer ensuite les contrôles → Le travail rapporte, on a ici un estimate égale à 0,72, un étudiant écoutant en cours aura en général une meilleur note de 0,72 dans la prévision.

*** Le ID du cours est également un facteur significatif sur la prévision de Grade. Son estimate est à 0,26. Ainsi, cela démontre qu'il existe une relation positive en fonction de l 'ID du cours et les notes attribuées. Cela a éventuellement un lien entre l'ID du cours donné par le professeur et le degré de difficulté. Moins le degré de difficulté du cours est élevé, plus la note est bonne.

Les facteurs ayant les plus impactés la prévision sur la note sont le Genre et la fréquence de lecture, qui augmente de plus d'1 point par rapport aux autres choix de réponse. Ainsi, certes le genre à un impacte sur les notes cependant elle reste une variable sur laquelle on ne peut pas avoir la main dessus. Il faut donc plus fournir des efforts sur le fait de passer plus de temps dans la lecture pour augmenter sa culture, mais aussi écouter en cours et prendre des notes. Ceci sont des variables qui vont impacter de manière positive les notes de l'étudiant.

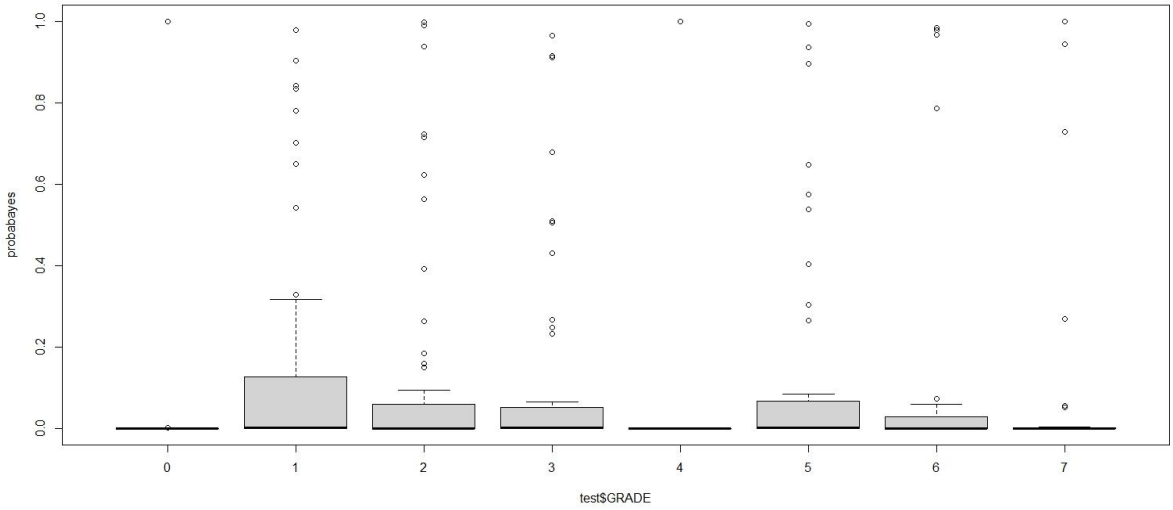
La corrélation de la méthode AIC

```
> library(corrr)
Message d'avis :
le package 'corrr' a été compilé avec la version R 4.1.3
> cor.test(previsionlogit,test$GRADE)

        Pearson's product-moment correlation

data:  previsionlogit and test$GRADE
t = 1.7819, df = 38, p-value = 0.08276
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.03701955  0.54229128
sample estimates:
      cor
0.2776921
```

Après avoir fait un summary de notre logit, nous avons calculer le coefficient corrélation qui va nous renseigner sur l'intensité de la relation linéaire entre notre prévision obtenue et le groupe test. Le coefficient de corrélation est d'environ 0,3, ce qui semble assez faible. Le coefficient est proche de 0, ce qui signifie qu'il n'y a pas de relation linéaire entre les prévisions obtenues et les notes dans le groupe test. Ainsi la méthode Logit n'est pas adaptée pour cette base de données, nous n'avons pas obtenu une très bonne prévision.

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

NAÏF BAYÉSIEN



Ici nous pouvons retrouver, le box plot obtenu grâce à la méthode du Naïf Bayésien.

Voici les lignes de code qui ont permis d'obtenir ce box plot :

```
> bayes=naiveBayes(GRADE~.,data=eleve)
> probabayes=predict(bayes,newdata=test,type="raw")
> summary(probabayes)
       0                   1                   2                   3
 Min.   :0.000000   Min.   :0.000000   Min.   :0.0000000   Min.   :0.0000000
 1st Qu.:0.000000   1st Qu.:0.001686   1st Qu.:0.0003088   1st Qu.:0.0001406
 Median :0.000000   Median :0.064196   Median :0.0193504   Median :0.0068102
 Mean   :0.052350   Mean   :0.236004   Mean   :0.1576836   Mean   :0.1523674
 3rd Qu.:0.000001   3rd Qu.:0.395170   3rd Qu.:0.1603297   3rd Qu.:0.1022173
 Max.   :0.999296   Max.   :0.977688   Max.   :0.9971659   Max.   :0.9652219
       4                   5                   6
 Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000000
 1st Qu.:0.0000001   1st Qu.:0.0000027   1st Qu.:0.0000000
 Median :0.0000073   Median :0.0016489   Median :0.0000005
 Mean   :0.0340157   Mean   :0.1509627   Mean   :0.1111928
 3rd Qu.:0.0013252   3rd Qu.:0.0602922   3rd Qu.:0.0002765
 Max.   :0.9991059   Max.   :0.9939740   Max.   :0.9838692
       7
 Min.   :0.0000000
 1st Qu.:0.0000000
 Median :0.0000020
 Mean   :0.1054243
 3rd Qu.:0.0006229
 Max.   :0.9987680
> boxplot(probabayes[,1]~eleve$GRADE)
Erreur dans stats::model.frame.default(formula = probabayes[, 1] ~ eleve$GRADE) :
  les longueurs des variables diffèrent (trouvé pour 'eleve$GRADE')
> boxplot(probabayes[,1]~test$GRADE)
> boxplot(probabayes~test$GRADE)
>
```

Interprétation du résultat trouvé avec la méthode Naïf Bayésien

Finalement, nous avons utilisé la méthode du Naïf Bayésien. Cette méthode nous a permis notamment d'obtenir les probabilités d'obtention relative à chaque note. Aussi, nous avons pu dessiner des boîtes à moustache afin de voir si cette méthode était bonne et fiable pour prédire les notes. Toutefois, au vu des boîtes à moustache obtenues qui sont quasiment toutes au même niveau, la méthode du Naïf Bayésien n'est pas bonne pour prédire les notes obtenues. L'arbre de décision reste encore le meilleur outil de prévision.

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

CONCLUSION

Pour conclure, il est juste de dire que dans le cadre de notre sujet le meilleur algorithme de prévision est celui de l'arbre de décision. Toutefois, il est important de garder en tête que notre base de données ne comporte que 146 réponses. Le nombre de réponses n'est donc pas représentatif de tous les étudiants. En effet, il est important de se rappeler que la situation de ces étudiants ne représente pas la situation de l'intégralité des étudiants. Enfin et surtout, un algorithme ne remplacera jamais le travail et l'effort fourni par un étudiant. Autrement dit, le travail paie toujours donc ne vous fiez pas toujours aux algorithmes.