Pràctica 2

Sol Balsells Mejía

6/2/2020

Introducció

El joc de dades seleccionat per a realitzar aquesta pràctica és 'Student Grade Prediction' [1].

Aquest joc de dades conté informació sobre 395 estudiants de dues escoles de secundària portugueses. Aquesta informació inclou dades descriptives dels alumnes (edat, sexe...), dades relacionades amb la seva formació (notes obtingudes, hores d'estudi...) i, finalment, dades relacionades amb aspectes socials (situació familiar, problemes de salut...). En total, el joc de dades conté 33 atributs.

Des del meu punt de vista, es tracta d'un joc de dades amb informació molt interessant per als instituts, ja que permet estudiar la influència de diferents factors al rendiment dels estudiants. A partir d'aquest estudi, es podria determinar en quines situacions els alumnes es poden trobar en situacions on la seva educació es pot trobar en perill (ja sigui perquè no s'aprovaran determinades assignatures o, en alguns casos, perquè aquesta situació pot donar lloc a abandonament escolar) i, per tant, l'institut podria valorar d'una forma més analítica a quins estudiants ha de destinar els recursos disponibles per evitar el fracàs escolar.

En aquest treball, realitzarem primer un anàlisi descriptiu del joc de dades. Posteriorment, estudiarem si determinats aspectes de la vida dels alumnes tenen influència en les notes que aquests treuen i construirem dos models que ens permetin predir com seran els resultats que obtindran els alumnes a final de curs.

Descripció del joc de dades

En primer lloc, llegim les dades de l'arxiu 'student-mat.csv' (adjuntat amb aquest document, descarregat del link del primer punt de la bibliografia) i les emmagatzem a la variable 'data'.

```
data <- read.csv('student-mat.csv')
```

Com s'ha dit anteriorment, el joc de dades conté un total de 33 columnes. Aquestes són:

- School. Indica l'escola a la que pertany l'alumne. Prèn dos valors: 'GP' (Gabriel Pereira) o 'MS' (Mousinho da Silveira).
- Sex. Indica el sexe de l'alumne. Prèn dos valors: 'F' i 'M'.
- Age. Indica l'edat de l'alumne. Prèn valors enters des de 15 fins a 22.
- Adress. Indica si l'alumne viu en una ciutat ('U') o en un poble ('R').
- Famsize. Indica la mida de la família. Prèn dos valors: 'LE3' (família amb 3 o menys membres) i 'GT3' (més de 3 membres).
- Pstatus. Indica si els pares viuen junts ('T') o separats ('A').
- Medu. Indica el nivell d'educació de la mare, prenent valors des de 0 (sense educació primària) fins a 4 (estudis superiors).
- Fedu. Igual que 'Medu', però amb el pare.
- Mjob. Indica la feina de la mare. Prèn 5 valors: 'teacher', 'health', 'services', 'at_home' i 'other'.
- Fjob. Igual que Mjob, però amb el pare.
- Reason. Indica el motiu pel qual l'alumne acudeix a l'escola. Prèn els valors: 'home' (escola propera a la seva casa), 'reputation', 'course' i 'other'.
- Guardian. Indica qui és el tutor de l'alumne. Prèn els valors 'mother', 'father' i 'other'.
- Traveltime. Indica el temps (en hores) que triga cada alumne per arribar a l'escola. Prèn valors enters, de manera que un trajecte de 20 minuts es registra com un trajecte d'1 hora, per exemple.
- Studytime. Indica el temps que l'alumne destina a estudiar al llarg de la setmana. Prèn valors enters.
- Failures. Indica el nombre d'assignatures suspeses l'anterior curs. Prèn valors enters.
- Schoolsup. Indica si l'alumne assisteix a classes de reforç ofertes per l'escola. Prèn valors 'yes' i 'no'.
- Famsup. Indica si l'alumne rep ajuda amb els estudis per part de la seva família. Prèn valors 'yes' i 'no'.
- Paid. Indica si l'alumne assisteix a classes de reforç finançades per la seva família. Prèn valors 'yes' i 'no'.
- Activities. Indica si l'alumne realitza activitats extra-escolars. Prèn valors 'yes' i 'no'.
- Nursery. Indica si l'alumne va realitzar parvulari. Prèn valors 'yes' i 'no'.
- Higher. Indica si l'alumne vol realitzar estudis superiors. Prèn valors 'yes' i 'no'.
- Internet. Indica si l'alumne té accés a internet a casa. Prèn valors 'yes' i 'no'.
- Romantic. Indica si l'alumne té una relació. Prèn valors 'yes' i 'no'.
- Famrel. Indica la qualitat de la relació de l'alumne amb la seva família. Prèn valors enters entre 1 (relació molt dolenta) i 5 (relació excel · lent).
- Freetime. Indica el temps lliure que té l'alumne després de l'escola. Prèn valors enters entre 1 (molt poc temps lliure) i 5 (molt temps lliure).
- Goout. Indica el temps que destina l'alumne a sortir amb els seus amics. Prèn valors enters entre 1 (molt poc) i 5 (molt de temps).
- Dalc. Indica el consum d'alcohol que realitza l'alumne entre setmana. Prèn valors enters entre 1 (molt poc) i 5 (molt elevat).
- Walc. Igual que 'Dalc', però amb els caps de setmana.
- Health. Indica com es troba de salut l'alumne. Prèn valors enters entre 1 (mala condició) i 5 (salut excel · lent).
- Absences. Indica el nombre d'absències de l'alumne. Prèn valors enters dins el rang (0, 93).
- G1. Nota al final del primer trimestre. Prèn valors enters entre 0 i 20.
- G2. Nota al final del segon trimestre.
- G3. Nota a final de curs.

Utilitzem la funció summary() per a tenir una breu descripció dels valors de cada variable, que alhora ens permetrà veure com interpreta RStudio cada variable i identificar alguns canvis que s'haurien de realitzar.

```
##  school   sex          age       address famsize   Pstatus
##  GP:349   F:208   Min.   :15.0   R: 88   GT3:281   A: 41
##  MS: 46   M:187   1st Qu.:16.0   U:307   LE3:114   T:354
##                   Median :17.0
##                   Mean   :16.7
##                   3rd Qu.:18.0
##                   Max.   :22.0
```

```
##       Medu            Fedu             Mjob           Fjob
##  Min.   :0.000   Min.   :0.000   at_home : 59   at_home : 20
##  1st Qu.:2.000   1st Qu.:2.000   health  : 34   health  : 18
##  Median :3.000   Median :2.000   other   :141   other   :217
##  Mean   :2.749   Mean   :2.522   services:103   services:111
##  3rd Qu.:4.000   3rd Qu.:3.000   teacher : 58   teacher : 29
##  Max.   :4.000   Max.   :4.000
```

```
##         reason      guardian     traveltime      studytime
##  course     :145   father: 90   Min.   :1.000   Min.   :1.000
##  home       :109   mother:273   1st Qu.:1.000   1st Qu.:1.000
##  other      : 36   other : 32   Median :1.000   Median :2.000
##  reputation:105                 Mean   :1.448   Mean   :2.035
##                                 3rd Qu.:2.000   3rd Qu.:2.000
##                                 Max.   :4.000   Max.   :4.000
```

```
##  schoolsup famsup     paid     activities nursery   higher
##  no :344   no :153   no :214   no :194    no : 81   no : 20
##  yes: 51   yes:242   yes:181   yes:201    yes:314   yes:375
```

```
##  internet  romantic      famrel         freetime         goout
##  no : 66   no :263   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  yes:329   yes:132   1st Qu.:4.000   1st Qu.:3.000   1st Qu.:2.000
##                      Median :4.000   Median :3.000   Median :3.000
##                      Mean   :3.944   Mean   :3.235   Mean   :3.109
##                      3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000
##                      Max.   :5.000   Max.   :5.000   Max.   :5.000
```

```
##       Dalc            Walc           health         absences
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 0.000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:3.000   1st Qu.: 0.000
##  Median :1.000   Median :2.000   Median :4.000   Median : 4.000
##  Mean   :1.481   Mean   :2.291   Mean   :3.554   Mean   : 5.709
##  3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:5.000   3rd Qu.: 8.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :75.000
```

```
##        G1              G2              G3
##  Min.   : 3.00   Min.   : 0.00   Min.   : 0.00
##  1st Qu.: 8.00   1st Qu.: 9.00   1st Qu.: 8.00
##  Median :11.00   Median :11.00   Median :11.00
##  Mean   :10.91   Mean   :10.71   Mean   :10.42
##  3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:14.00
##  Max.   :19.00   Max.   :19.00   Max.   :20.00
```

Realitzarem dues modificacions a les dades.

En primer lloc, eliminarem les columnes qualitatives que presenten més de 2 opcions, ja que aquest format de dades no s'ajusta gaire bé als anàlisis que realitzarem posteriorment i, per tant, són dades que no utilitzarem. Aquestes columnes són 'Mjob', 'Fjob', 'Reason' i 'Guardian'.

- Altres columnes, com 'Famrel', també expressen mesures qualitatives, però en aquest cas s'han expressat en una escala numèrica que és lògica (menor valor, pitjor relació familiar, per exemple). Aquestes columnes les conservarem.

```
data <- data[ ,-which(names(data) %in% c("Mjob","Fjob", 'reason', 'guardian'))]
```

En segon lloc, homogenitzarem la forma en què s'expressen les columnes qualitatives que conservem. Convertirem les columnes binàries amb valors de tipus 'string' en columnes '0'/'1', de manera que totes les columnes amb valors qualitatius s'expressaran numèricament.

```
data$school <- as.character(data$school)
data$sex <- as.character(data$sex)
data$address <- as.character(data$address)
data$famsize <- as.character(data$famsize)
data$Pstatus <- as.character(data$Pstatus)
data$schoolsup <- as.character(data$schoolsup)
data$famsup <- as.character(data$famsup)
data$paid <- as.character(data$paid)
data$activities <- as.character(data$activities)
data$nursery <- as.character(data$nursery)
data$higher <- as.character(data$higher)
data$internet <- as.character(data$internet)
data$romantic <- as.character(data$romantic)

data$school[data$school=='GP'] <- as.numeric(0)
data$school[data$school=='MS'] <- as.numeric(1)
data$sex[data$sex=='F'] <- as.numeric(0)
data$sex[data$sex=='M'] <- as.numeric(1)
data$address[data$address=='R'] <- as.numeric(0)
data$address[data$address=='U'] <- as.numeric(1)
data$famsize[data$famsize=='LE3'] <- as.numeric(0)
data$famsize[data$famsize=='GT3'] <- as.numeric(1)
data$Pstatus[data$Pstatus=='T'] <- as.numeric(0)
data$Pstatus[data$Pstatus=='A'] <- as.numeric(1)
data$schoolsup[data$schoolsup=='no'] <- as.numeric(0)
data$schoolsup[data$schoolsup=='yes'] <- as.numeric(1)
data$famsup[data$famsup=='no'] <- as.numeric(0)
data$famsup[data$famsup=='yes'] <- as.numeric(1)
data$paid[data$paid=='no'] <- as.numeric(0)
data$paid[data$paid=='yes'] <-as.numeric(1)
data$activities[data$activities=='no'] <- as.numeric(0)
data$activities[data$activities=='yes'] <- as.numeric(1)
data$nursery[data$nursery=='no'] <- as.numeric(0)
data$nursery[data$nursery=='yes'] <- as.numeric(1)
data$higher[data$higher=='no'] <- as.numeric(0)
data$higher[data$higher=='yes'] <- as.numeric(1)
data$internet[data$internet=='no'] <- as.numeric(0)
data$internet[data$internet=='yes'] <- as.numeric(1)
data$romantic[data$romantic=='no'] <- as.numeric(0)
data$romantic[data$romantic=='yes'] <- as.numeric(1)
```

Valors nuls

Comprovem l'existència de valors nuls al joc de dades. Els resultats anteriors de la funció summary ens permeten veure que no hi ha valors nuls a les variables qualitatives, però podria haver-n'hi a les altres columnes.

```
colSums(is.na(data))
```

```
##     school        sex        age    address    famsize    Pstatus
##          0          0          0          0          0          0
##       Medu       Fedu traveltime  studytime   failures  schoolsup
##          0          0          0          0          0          0
##     famsup       paid activities    nursery     higher   internet
##          0          0          0          0          0          0
##   romantic     famrel   freetime      goout       Dalc       Walc
##          0          0          0          0          0          0
##     health   absences         G1         G2         G3
##          0          0          0          0          0
```

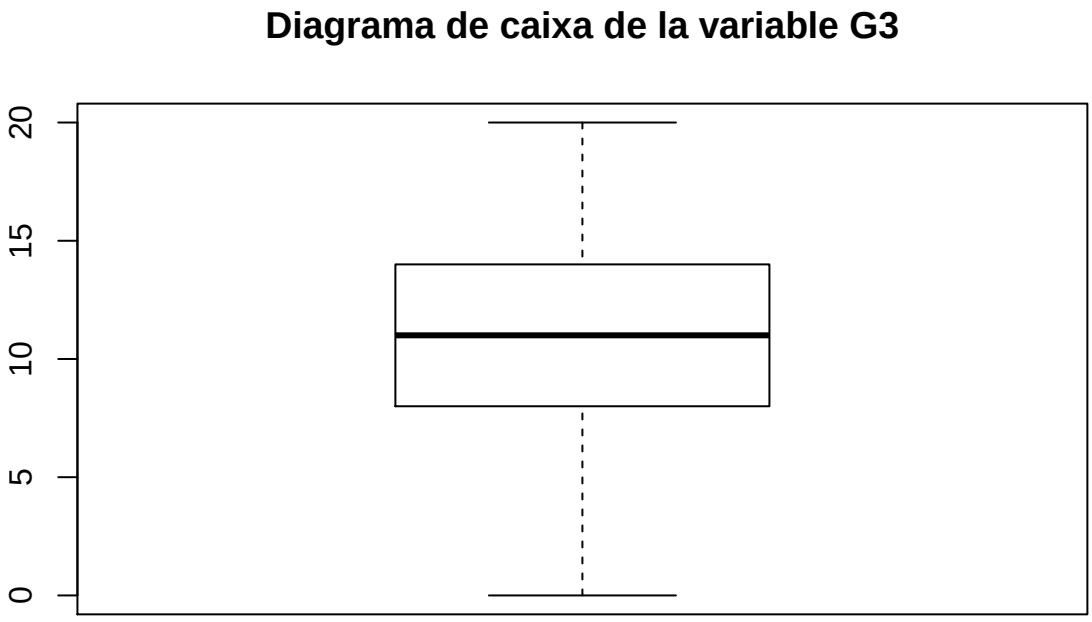
Com podem veure, a cap de les columnes hi ha cap valor 'NA', així que el joc de dades no conté dades nules.

Outliers

Analitzem també la presència d'outliers al joc de dades.

Utilitzem la funció boxplot(), que construeix diagrames de caixa a partir de la distribució de valors de la columna seleccionada. Aquests diagrames de caixa mostren quina és la regió interquartil i, a partir d'aquest valor, es defineix el rang on es trobarien els valors que no són outliers. Això permet, lògicament, poder visualitzar si hi ha valors outliers.

Com a exemple, creem el diagrama de caixa de la variable 'G3'.



Com podem veure, el diagrama ens indica que la regió interquartil ('IQ') es troba al rang de valors (8,14). A partir de l'abast d'aquesta regió, es pot definir a quina regió pertanyen els valors no outliers, que sol definir-se com:

$$(Q1 - 1.5IQ, Q3 + 1.5IQ)$$

En aquest cas, la regió interquartil és de longitud 6 i això dóna lloc a un una regió de valors no outliers (-1, 23), que obviament cobreix tots els valors de la variable 'G3' (prèn valors entre 0 i 20) i, per tant, no mostra l'existència d'outliers.

Pel que fa a les altres variables, analitzem en quines tenim valors extrems. No estudiarem les columnes binàries, ja que anteriorment ja hem vist com es distribuïen els valors dins d'aquestes variables i la informació que obtindríem ara seria molt similar.

```
data_outliers <- data[ ,-which(names(data) %in% c("school","sex", 'address', 'famsize', 'Pstatus', 'sch
                                                  'famsup', 'paid', 'activities', 'nursery', 'higher',
                                                  'internet', 'romantic'))]

for(i in colnames(data_outliers)){
  print(i)
  print(sort(unique(boxplot.stats(data[,i])$out)))}
```

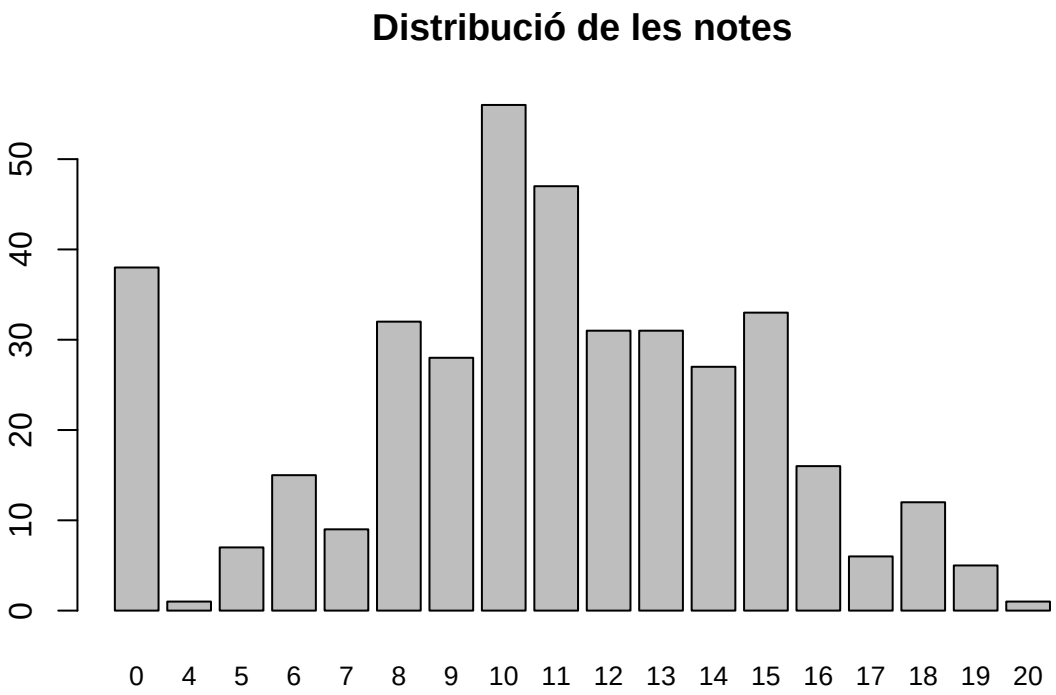
```
## [1] "age"
## [1] 22
## [1] "Medu"
## integer(0)
## [1] "Fedu"
## [1] 0
## [1] "traveltime"
## [1] 4
## [1] "studytime"
## [1] 4
## [1] "failures"
## [1] 1 2 3
## [1] "famrel"
## [1] 1 2
## [1] "freetime"
## [1] 1
## [1] "goout"
## integer(0)
## [1] "Dalc"
## [1] 4 5
## [1] "Walc"
## integer(0)
## [1] "health"
## integer(0)
## [1] "absences"
##  [1] 21 22 23 24 25 26 28 30 38 40 54 56 75
## [1] "G1"
## integer(0)
## [1] "G2"
## [1] 0
## [1] "G3"
## integer(0)
```

Com podem veure, moltes de les columnes presenten l'existència de valors extrems. En tots els casos, però, mantindrem aquests valors, ja que, si ve estadísticament són outliers, són casos que poden donar-se a la realitat i possiblement contindran informació rellevant per als anàlisis que fem de les dades.

Comprovació de la normalitat

Avaluem també si la distribució dels valors anteriors s'ajusta a una distribució normal.

En primer lloc, com hem fet amb els outliers, estudiem amb major detall el cas de la variable 'G3'. Ens interessarà veure com es distribueixen els valors d'aquesta variable.



Aquest diagrama ens mostra l'existència d'un nombre anòmal d'alumnes que s'han avaluat amb una nota de 0. Pel que fa a la resta de valors, sí que s'observa una distribució de valors que recorda a una gaussiana, tot i que la distribució obtinguda no és del tot simètrica.

Per avaluar amb major rigor si podem afirmar que la distribució de valors de la variable 'G3' és pot aproximar a una distribució normal, utilitzem la funció ad.test(), de la llibreria 'nortest'. Es tracta d'una funció que realitza el test d'Anderson-Darling, que és un test que obté, a partir de la distribució de valors, un valor (p-value) que indica la probabilitat que la distribució de valors sigui una distribució normal. El p-valor obtingut es compara, llavors, amb el nivell de significació, que en el nostre cas considerarem que és 0.05 (corresponent a un nivell de confiança del 95%). Així doncs, considerarem que la distribució normal si p-valor és superior a 0.05.

```
library(nortest)

ad.test(data$G3)
```

```
##
##  Anderson-Darling normality test
##
## data:  data$G3
## A = 8.3032, p-value < 2.2e-16
```

En aquest cas, com podíem esperar veient la distribució dels valors de 'G3' que hem vist abans, obtenim un p-valor molt petit i, per tant, inferior al valor 0.05. És a dir, la distribució dels valors de la variable 'G3' no s'ajusta a una gaussiana.

Com hem observat anteriorment que hi havia una gran quantitat de valors 0, repetim el càlcul sense aquests valors.

```
ad.test(data$G3[data$G3>0])
```

```
##
##  Anderson-Darling normality test
##
## data:  data$G3[data$G3 > 0]
## A = 2.5143, p-value = 2.323e-06
```

En aquest cas, el p-valor augmenta 10 ordres de magnitud, de manera que la semblança de la distribució amb una gaussiana ha augmentat. Tot i això, el p-valor segueix sent bastant inferior a 0.05 i, per tant, no podem afirmar que la distribució de valors de 'G3' sigui normal.

Realitzem el test d'Anderson-Darling per a les variables no binàries (les mateixes que s'han analitzat en busca de valors extrems) i ens fixem en els p-valors obtinguts, ja que ens indicaran quines columnes presenten distribucions normals.

```
for(i in colnames(data_outliers)[-16]){
  print(i)
  print(ad.test(data[,i])$p.value)}
```

```
## [1] "age"
## [1] 3.7e-24
## [1] "Medu"
## [1] 3.7e-24
## [1] "Fedu"
## [1] 3.7e-24
## [1] "traveltime"
## [1] 3.7e-24
## [1] "studytime"
## [1] 3.7e-24
## [1] "failures"
## [1] 3.7e-24
## [1] "famrel"
## [1] 3.7e-24
## [1] "freetime"
## [1] 3.7e-24
## [1] "goout"
## [1] 3.7e-24
## [1] "Dalc"
## [1] 3.7e-24
## [1] "Walc"
## [1] 3.7e-24
## [1] "health"
## [1] 3.7e-24
## [1] "absences"
## [1] 3.7e-24
## [1] "G1"
## [1] 4.271907e-08
## [1] "G2"
## [1] 2.165869e-06
```

Com podem veure, cap de les variables analitzades presenta una distribució normal.

Anàlisi de les dades

Realitzarem 3 anàlisis diferents amb les dades disponibles:

- Comparació de la nota mitjana segons valor de les columnes binàries.
- Model d'arbre de decisió per a la predicció d'alumnes que suspendran.
- Model de regressió lineal per a la predicció de la nota que s'obtindrà a final de curs.

Anàlisi 1: influència en la nota mitjana de les columnes binàries

A partir de les columnes binàries, podem dividir les dades en dos conjunts. Per exemple, la columna 'Sex' ens permet distingir entre els alumnes segons si són nois o noies. Podem calcular, en aquests casos, quina és la nota ('G3') mitjana per a cada un d'aquests grups.

```
## [1] "Nota mitjana de les noies: 9.96634615384615"
```

```
## [1] "Nota mitjana dels nois: 10.9144385026738"
```

Com podem veure, els nois tenen una nota mitjana de quasi 11, mentre la de les noies es queda lleugerament per sota dels 10. El que volem fer en aquest anàlisi és valorar si aquesta diferència és significativa estadísticament i, per tant, les dades que tenim ens indiquen que els nois són millors estudiants que les noies.

Per a fer-ho, calcularem en quin interval de valors es troba, amb un nivell de confiança del 95%, el valor 'Mitjana noies - Mitjana nois' [3]. En cas que el sexe dels alumnes no influenciés la nota d'aquests, aquest interval de valors inclourà el valor 0. Calculem, doncs, aquest interval de confiança.

```
## [1] "Interval de confiança: (-1.85205631632208 ,-0.0441283813332076)."
```

Com podem veure, l'interval de confiança no inclou el 0, així que els resultats ens indiquen, amb un nivell de confiança del 95%, que la nota mitjana de les noies és inferior a la dels nois. És possible, però, que si s'ampliés el nombre de dades disponibles sí que s'obtingués un interval de confiança que inclogués el 0, ja que l'interval obtingut té un extrem molt proper a aquest valor.

Repetim el procés, doncs, per les altres variables de tipus binari, per analitzar quines influeixen en la nota mitjana dels alumnes.

```
## [1] "Variable 'School'."
```

```
## [1] "Interval de confiança: (-0.771069386502703 ,2.0553599059994)."
```

```
## [1] "Variable 'Address'."
```

```
## [1] "Interval de confiança: (-2.2472982059527 ,-0.0785087232744235)."
```

```
## [1] "Variable 'Famsize'."
```

```
## [1] "Interval de confiança: (-0.176074482044105 ,1.82020259592311)."
```

```
## [1] "Variable 'Pstatus'."
```

```
## [1] "Interval de confiança: (-2.35556516957112 ,0.615038781256389)."
```

```
## [1] "Variable 'Schoolsup'."
```

```
## [1] "Interval de confiança: (-0.218924076308093 ,2.47827200152469)."
```

```
## [1] "Variable 'Famsup'."
```

```
## [1] "Interval de confiança: (-0.562998574577116 ,1.29858978075656)."
```

```
## [1] "Variable 'Paid'."
```

```
## [1] "Interval de confiança: (-1.84266024046236 ,-0.0306810101185245)."
```

```
## [1] "Variable 'Activities'."
## [1] "Interval de confiança: (-1.05493633227434 ,0.760224325298911)."
## [1] "Variable 'Nursery'."
## [1] "Interval de confiança: (-1.70683278551796 ,0.538003659151676)."
## [1] "Variable 'Higher'."
## [1] "Interval de confiança: (-5.8429651881937 ,-1.7730348118063)."
## [1] "Variable 'Internet'."
## [1] "Interval de confiança: (-2.41840016427949 ,0.00253942926981887)."
## [1] "Variable 'Romantic'."
## [1] "Interval de confiança: (0.306903309241852 ,2.21458534152436)."
```

En resum, hem trobat que les variables 'Sex', 'Address', 'Paid', 'Higher' i 'Romantic' presenten diferències significants entre la nota mitjana dels corresponents grups d'alumnes.

Entre els resultats obtinguts, tenim alguns casos sorprenents i altres que són més comprensibles. Per exemple, és esperable que els alumnes que volen realitzar estudis superiors o que assisteixin a classes extra tinguin notes majors, com també ho és que el fet d'haver anat a una guarderia influenciï en les notes dels alumnes de secundària.

Per altra banda, són resultats més sorprenents els de la variable 'Sex', ja que s'esperaria que no fós un factor important. També podria ser esperable que la variable 'Activities', que indica si un alumne realitza activitats extraescolars, fós un factor important en la nota mitjana dels alumnes, per exemple.

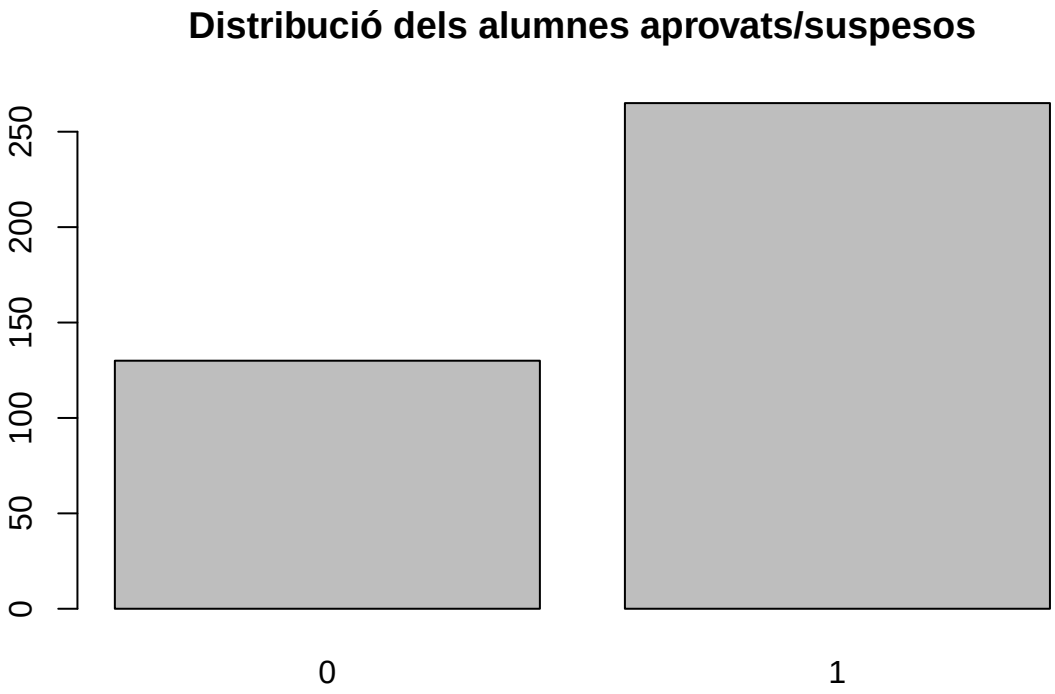
Anàlisi 2: predicció d'alumnes aprovats

El segon anàlisi que plantegem consisteix en estudiar si podem crear un model que ens permeti predir si un alumne suspendrà el curs o no.

El primer que necessitem, doncs, és crear una nova variable que ens indiqui si un alumne ha aprovat o no. Considerarem que els alumnes que treuen una nota de 10 o superior han aprovat.

```
data$aprovat <- 0
data$aprovat[data$G3>9] <- 1
data$aprovat <- as.factor(data$aprovat)
```

Estudiem com es distribueixen els alumnes segons si han aprovat o no.



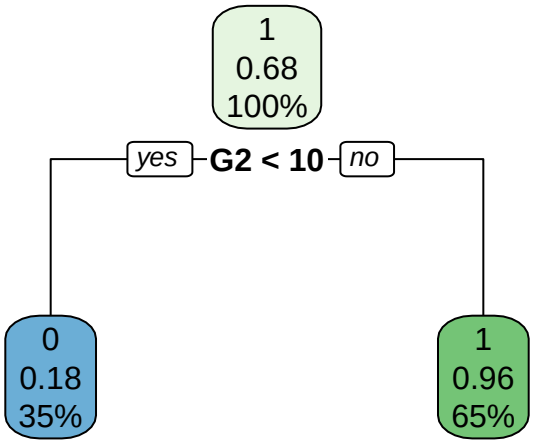
La gràfica ens mostra que, aproximadament, 2/3 dels alumnes han aprovat el curs i el terç restant l'ha suspès.

Construirem un model d'arbre de decisió que ens permeti predir a quin grup pertany cada alumne. Necessitarem, doncs, dividir les dades en un conjunt d'entrenament, per construir el model, i un conjunt de prova, per avaluar-lo.

```
set.seed(55)
sample <- sample.int(n = nrow(data), size = floor(.75*nrow(data)), replace = F)
train <- data[sample, -29]
test  <- data[-sample, -29]
```

Un cop tenim els dos conjunts, utilitzem la funció rpart() de la llibreria 'rpart' per a construir un model d'arbres de decisió a partir del conjunt d'entrenament. Amb la funció rpart.plot() de la llibreria 'rpart.plot', visualitzem aquest arbre de decisió.

```
library(rpart)
library(rpart.plot)
AD <- rpart(aprovat~., data =train, method = 'class')
rpart.plot(AD)
```



Com podem veure, l'arbre de decisió obtingut és un arbre molt simple, ja que únicament utilitza la nota del segon trimestre com a criteri per classificar les dades. El model ens indica que els alumnes que van treure una nota superior a 10 han aprovat el curs en el 96% dels casos, mentre que els alumnes que van treure menys nota l'han suspès en el 82% dels casos (el 0.18 indica la proporció d'alumnes de la classe 1).

Si analitzem la correlació de les columnes 'G1' i 'G2' amb la columna 'G3', veurem que aquestes columnes estan fortament correlacionades.

```
cor(data$G1, data$G3)
```

```
## [1] 0.8014679
```

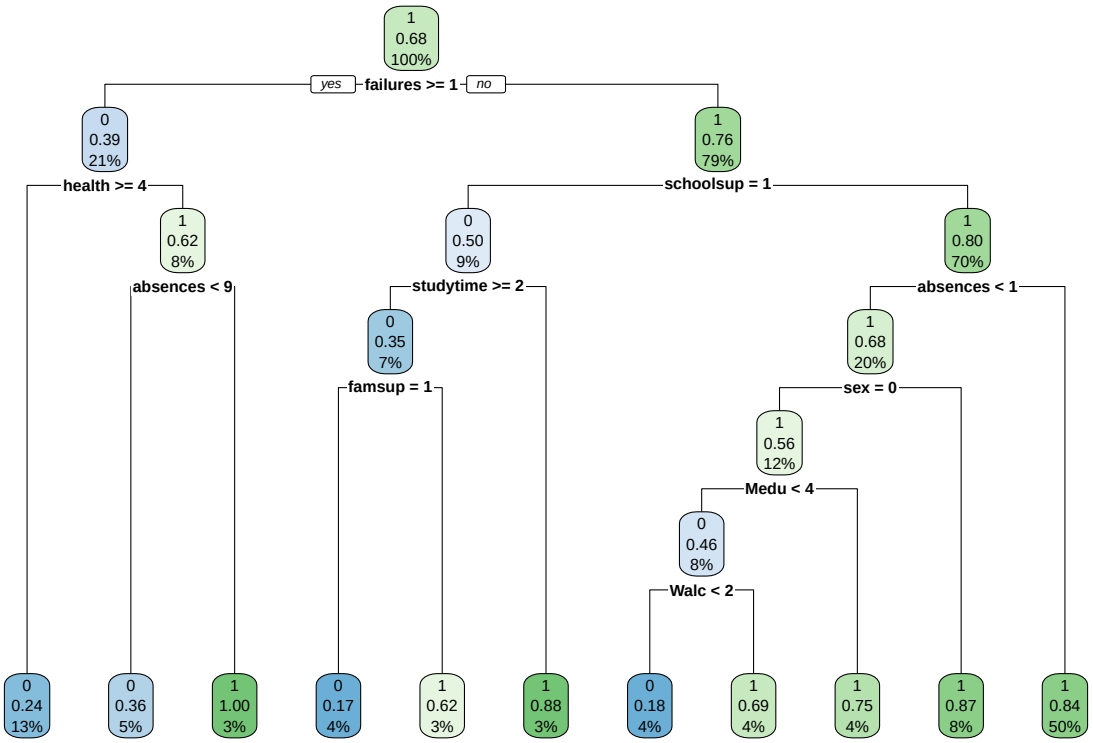
```
cor(data$G2, data$G3)
```

```
## [1] 0.904868
```

Aquesta alta correlació és lògica, ja que la nota de final de curs es calcularà a partir de les notes dels anteriors dos trimestres. El problema amb el que ens trobem, però, és que l'alta correlació d'aquestes dues variables amb 'G3' fa que siguin les úniques que s'utilitzen, eclipsant així altres variables que ens podrien oferir informació més interessant.

Així doncs, construïm un nou arbre de decisió, ara sense utilitzar 'G1' i 'G2'.

```
train <- train[, -(27:28)]
test <- test[, -(27:28)]
AD2 <- rpart(aprovat~., data =train, method = 'class')
rpart.plot(AD2)
```



En aquest arbre de decisió, el primer aspecte que es valora és si l'alumne va suspendre alguna assignatura el curs anterior. En cas que fós així, el model prediu, principalment, que l'estudiant suspendrà. Per altra banda, el model utilitza com a segon criteri les classes de reforç que ofereix l'escola per als alumnes que no van suspendre cap assignatura el curs anterior. Els alumnes que no reben cap reforç a l'escola, generalment, aproven. Pot semblar contradictori però aquesta absència de reforç és un indicatiu de què els alumnes porten bé el curs, així que des d'aquest punt de vista és lògic. Per als alumnes als que l'escola ofereix classes de

reforç, curiosament són els que estudien més i reben ajuda a casa els que es prediu que suspendran. Una possible explicació és que els alumnes que reben més ajuda són els que tenen més dificultats per aprovar els seus cursos i, per tant, el model prediu que suspendran.

Arribat a aquest punt, podem utilitzar el model per a predir si els alumnes del conjunt de test aprovaran o suspendran el curs. Això ens permetrà avaluar l'acert del model.

```
table(test$aprovat, predict(AD2, test, type='class'))
```

```
##
##      0  1
##   0 18 18
##   1 12 51
```

La taula obtinguda ens permet veure com ha predit el model la classe del conjunt de test. Per als alumnes que van suspendre el curs, el model classifica correctament 18 de 36, que és un acert d'un 50%, un valor molt pobre. Per als alumnes aprovats, en canvi, la classificació és correcta en 51 de 63 casos, que és un acert d'un 80%. Això ens indica, doncs, que

En total, el model presenta un 70% d'acert.

Els arbres de decisió són un tipus d'algorisme que s'utilitza habitualment perquè els seus resultats són molt fàcils d'interpretar. En el nostre cas, podem veure, a grans trets, que aprovaran els alumnes que van aprovar totes les assignatures del curs anterior i que durant aquest no han necessitat reforç, mentre que en cas contrari les possibilitats de suspendre augmentaran. No obstant, pot interessar-nos més l'obtenció d'un model que sigui més difícil d'interpretar però que realitzi millors prediccions.

Una possible forma de millorar el rendiment dels arbres de decisió és l'ús d'un algorisme Random Forest. Aquest algorisme es basa en la construcció d'un nombre elevat d'arbres de decisió, a partir dels quals s'extreu informació per la classificació de les dades i es construeix un model únic que l'englobi. En altres paraules, en comptes d'utilitzar un únic arbre de decisió per a realitzar les prediccions, se n'utilitzen molts de diferents.

Construïm el model i n'avaluem la precisió.

```
library(randomForest)
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
rf <- randomForest(x=train[,-1], y=train$aprovat, ntree=100)

table(test$aprovat, predict(rf, test, type='class'))
```

```
##
##      0  1
##   0 36  0
##   1  0 63
```

Com podem veure, el model obtingut a partir d'un algorisme Random Forest amb 100 arbres de decisió té un 100% d'acert en les prediccions. No es tracta d'un resultat realista (si ampliem el conjunt de test, el model acabaria donant errors) però la millora en les prediccions respecte al model d'arbre de decisió és evident.

Anàlisi 3: predicció de la nota

Hem construït un model per a intentar predir si un alumne suspendria o no; intentem ara construir un model que ens permeti predir quin és el valor de la nota que treurà cada alumne.

El model que construirem serà un model de regressió lineal múltiple, però abans d'això necessitem avaluar la correlació entre la nota final i les altres variables.

```
##    failures         age       goout    romantic  traveltime   schoolsup
## -0.36041494 -0.16157944 -0.13279147 -0.12996995 -0.11714205 -0.08278821
##     famsize      health        Dalc        Walc      school      famsup
## -0.08140711 -0.06133460 -0.05466004 -0.05193932 -0.04501694 -0.03915715
##    freetime  activities    absences      famrel     nursery     Pstatus
##  0.01130724  0.01609970  0.03424732  0.05136343  0.05156790  0.05800898
##   studytime    internet        paid         sex     address        Fedu
##  0.09781969  0.09848337  0.10199624  0.10345565  0.10575606  0.15245694
##      higher        Medu     aprovat          G1          G2          G3
##  0.18246462  0.21714750  0.77004217  0.80146793  0.90486799  1.00000000
```

Com hem vist anteriorment, les notes al final dels 2 primers trimestres són les dues variables més correlacionades amb la nota del final del curs. També hi ha una gran correlació, com és lògic, entre les variable 'aprovat' i 'G3'.

Pel que fa a les demés variables, la corelació amb 'G3' d'aquestes és menys forta. La variable 'failures', que anteriorment ja hem vist que era un indicador important per predir si un alumne aprovarà o suspendrà, és la variable que presenta major correlació (tot i que, en aquest cas, és una correlació negativa deguda a que un augment en el valor de 'failures' implica una disminució en la nota predita). També influeixen negativament l'edat de l'alumne i el temps que passa fora de casa, mentre que en influències positives trobarem el nivell educatiu dels pares i si l'alumne té intenció de fer estudis superiors.

Creem, en primer lloc, un model de regressió lineal múltiple que utilitzi les notes dels 2 primers trimestres.

```r
set.seed(55)
sample <- sample.int(n = nrow(data), size = floor(.75*nrow(data)), replace = F)
train <- data[sample, ]
test  <- data[-sample, ]

regresion <- lm(G3 ~ G1 + G2, data = train)
summary(regresion)
```

```
##
## Call:
## lm(formula = G3 ~ G1 + G2, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6550 -0.3084  0.2374  0.8928  3.5601
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.42066    0.38911  -3.651 0.000309 ***
## G1           0.11559    0.06480   1.784 0.075520 .
## G2           0.99198    0.05729  17.316  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.922 on 293 degrees of freedom
## Multiple R-squared:  0.8179, Adjusted R-squared:  0.8167
## F-statistic:   658 on 2 and 293 DF,  p-value: < 2.2e-16
```

El model obtingut ens indica que podem predir 'G3' a partir de:

$$G3 = -1.42066 + 0.11559 \text{ů} G1 + 0.99198 \text{ů} G2$$

Aquesta recta de regressió presenta un coeficient R ajustat de 0.8167. Aquest coeficient ens indica quin percentatge de la variança de G3 s'explica amb aquesta recta de regressió, així que tenim un valor prou bo.

A partir dels coeficients de la recta de regressió podem veure que 'G2' és un paràmetre molt més influent en 'G3' que no pas 'G1'. Si observem els resultats de la taula, podem veure també (ho indiquen els asteriscs) que 'G2' és molt més significativa per al model que 'G1'. Aquesta menor importància de 'G1' pot ser deguda a què 'G1' i 'G2' estan fortament correlacionades, així que 'G1' no pot ampliar gaire la informació que ja conté 'G2'.

Com passava amb els abres de decisió, però, l'ús de 'G1' i 'G2' no és gaire interessant, ja que el seu poder per predir 'G3' és molt evident. Així doncs, podem utilitzar les altres variables que anteriorment han mostrat major correlació amb 'G3' per a construir un model.

```r
regresion2 <- lm(G3 ~ failures + age + goout + Medu + Fedu + higher, data = train)
summary(regresion2)
```

```
##
## Call:
## lm(formula = G3 ~ failures + age + goout + Medu + Fedu + higher,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.3810  -2.0009   0.4107   2.7913   8.1855
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.00295    3.79235   3.165  0.00172 **
## failures    -1.84036    0.34423  -5.346 1.83e-07 ***
## age         -0.13393    0.20027  -0.669  0.50421
## goout       -0.24809    0.22212  -1.117  0.26497
## Medu         0.43029    0.29421   1.463  0.14469
## Fedu         0.05606    0.29956   0.187  0.85169
## higher       0.99380    1.26305   0.787  0.43203
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.141 on 289 degrees of freedom
## Multiple R-squared:  0.1662, Adjusted R-squared:  0.1489
## F-statistic:   9.6 on 6 and 289 DF,  p-value: 1.239e-09
```

El model obtingut té un coeficient R pobre, de només 0.1489. Si ens hi fixem, podem veure que l'única variable rellevant estadísticament és 'failures', mentre que variables com 'Fedu', 'age' i 'higher' tenen probabilitats de ser irrellevants molt elevades (possiblement perquè estan fortament correlacionades amb altres variables més rellevents).

```
regresion3 <- lm(G3 ~ failures + goout + Medu, data = train)
summary(regresion3)
```

```
##
## Call:
## lm(formula = G3 ~ failures + goout + Medu, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.2452  -1.9831   0.4099   2.8458   8.2325
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.7213     0.8995  11.919  < 2e-16 ***
## failures     -1.9535     0.3282  -5.952  7.6e-09 ***
## goout        -0.2621     0.2195  -1.194   0.2334
## Medu          0.5120     0.2276   2.249   0.0252 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.129 on 292 degrees of freedom
## Multiple R-squared:  0.1624, Adjusted R-squared:  0.1538
## F-statistic: 18.87 on 3 and 292 DF,  p-value: 3.27e-11
```

Eliminant aquestes variables, 'Medu' guanya en importància i el coeficient R ajustat millora lleugerament, tot i que segueix estant lluny d'un valor que ens indiqui que el model té poder predictiu.

Conclusions

En aquesta pràctica, hem analitzat un joc de dades amb informació sobre diferents alumnes i l'hem utilitzat per analitzar quins aspectes educatius i personals tenen influència en les notes que treurà a final de curs.

Pel que fa al primer anàlisi, hem estudiat quins aspectes afecten a la nota dels alumnes. Per determinar la influència d'aquests aspectes en les notes, hem dividit les dades en els 2 grups que formen aquestes columnes i hem estudiat si la diferència entre la nota mitjana de cada un d'aquests grups era significativa. Això ens ha permès veure que la nota dels estudiants serà major si són nois, viuen a les ciutats, realitzen classes de reforç fora de l'escola, tenen intenció d'anar a la universitat i no tenen parella. Per altra banda, també és interessant que el tamany de la família o el fet que un alumne realitzi activitats extraescolars no influencien a les seves notes.

Respecte al segon anàlisi, s'ha construït un model que permetés predir si un alumne aprovaria o no. En primer lloc, el model construït ha estat un arbre de decisió. S'ha vist que les variables 'G1' i 'G2' tenien molt pes en aquests models, mentre que si s'excloïen del model guanyaven importància 'failures' i 'schoolsup', que són dues variables que ajuden a entendre com han anat el curs actual i el curs anterior a l'alumne i, per tant, tenen un pes lògic en la predicció de si l'alumne aprovarà.

El model d'arbre de decisió sense 'G1' i 'G2' permetia fer prediccions amb un 70% d'acert, però aquest poder predictiu s'ha intentat millorar amb un model Random Forest. Aquest ha mostrat un acert del 100% amb les dades de prova, que indica que es poden fer bones prediccions per determinar si un alumne aprovarà o no.

Finalment, el darrer anàlisi ha consistit en crear models de regressió lineal múltiple per predir la nota que treurien els alumnes. De nou, s'ha obtingut un model prou bo amb 'G1' i 'G2'. En aquest cas, però, les prediccions que s'obtenien amb el model construït excloent aquestes dues variables han estat molt pobres. Això ens indica que, si ve hi ha factors que ajuden a predir si un alumne podrà aprovar o no, la predicció de les notes és molt més complexa i hi intervenen molts més factors dels que s'inclouen a aquest joc de dades.

Bibliografia

1) https://www.kaggle.com/dipam7/student-grade-prediction
2) https://rpubs.com/MSiguenas/122473
3) 'Contrast de dues mostres', de Josep Gibergans Bàguena. Apunts de l'assignatura Estadística Avançada
4) https://www.guru99.com/r-decision-trees.html
5) https://towardsdatascience.com/random-forest-in-r-f66adf80ec9
6) https://rpubs.com/Joaquin_AR/226291