Alondra Sánchez Molina

Descripción del data set

Nombre del data set: Student Performance Data Set

This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [Cortez and Silva, 2008], the two datasets were modeled under binary/five-level classification and regression tasks. Important note: the target attribute G3 has a strong correlation with attributes G2 and G1.

Por la magnitud del número de atributos, y con el fin de no hacer este documento más extenso de lo necesario, se optó por proporcionar el link para la descripción de los atributos del data set.

Link del data set: https://archive.ics.uci.edu/ml/datasets/Student+Performance

Desarrollo del trabajo

Limpieza y preprocesamiento de los datos

Después de cargar los datos, y comprenderlos, es necesario tratar los datos que se van a utilizar. Para la realización de esta tarea no se utilizará toda la información proporcionada del dataset. Se ha optado por explorar los datos de la materia de matemáticas.

Se ha realizado un filtrado con el objetivo de sólo analizar los datos de los alumnos de la escuela cuyas iniciales son GP, con el fin de disminuir el tamaño de la población a tratar.

Para cumplir con los fines de esta práctica, se han generado dos nuevas variables:

- *alcohol:* Dicha variable, es una variable numérica que se calcula a partir de la media ponderada considerando las dos variables *Dalc* y *Walc* dichas variables representan el consumo de alcohol por parte de los estudiantes entre la semana y los fines de semana respectivamente.

$$alcohol = \frac{5 * Dalc + 2 * Walc}{7}$$

- *pass:* Dicha variable nos retorna un resultado binario, en base a si el estudiante a aprobado el curso o no, esto se calcula haciendo uso de la variable G3. El sistema de calificaciones de Portugal se realiza en una escala de 20 puntos, considerando que necesitan 10 para poder acreditar la materia.

Finalmente, se genera un nuevo data frame el cual contendrá solo los campos que se utilizaran en este estudio junto con nuestros dos nuevos campos creados.

La siguiente captura de pantalla, muestra el código de la limpieza y procesamiento de los datos.

```
#---Importación de datos
mat_class=read.table("RFiles/student/student-mat.csv",sep=";",header=TRUE)

print(nrow(mat_class))

#---Limpieza y preprocesamiento de datos
library(dplyr)

mat_data <- filter(mat_class,school == 'GP')
mat_data$alcohol <- round((((5*mat_data$Dalc) + (2*mat_data$Walc))/7),0)

#se crea la función de asignacion de pasado o no
fun_pass <- function(calif) ifelse(calif >= 10, 1, 0)

mat_data$pass <- fun_pass(mat_data$G3)

mat_data <- select(mat_data,sex,address,Pstatus,Medu,Fedu,failures,schoolsup,famsup,
                   higher,internet,famrel,absences,alcohol,G1,G2,G3,pass)
```

Modelado de los árboles de decisión

Se tomó la decisión, de realizar el modelado de cuatro árboles diferentes, con el fin de examinar tres casos:

1. Si el estudiante de la secundaria GP, pasará su curso de matemáticas
2. Si el estudiante de la secundaria GP, que cursa la materia de matemáticas considera cursar la preparatoria.
3. Los factores que influyen en el nivel de alcoholismo de los estudiantes de la secundaria GP, que cursan la materia de matemáticas.

Para el modelado del primer caso, se realizaron dos árboles, con el fin de compararlos.

Árbol 1. Si el estudiante pasará o no su materia considerando sus dos calificaciones anteriores.

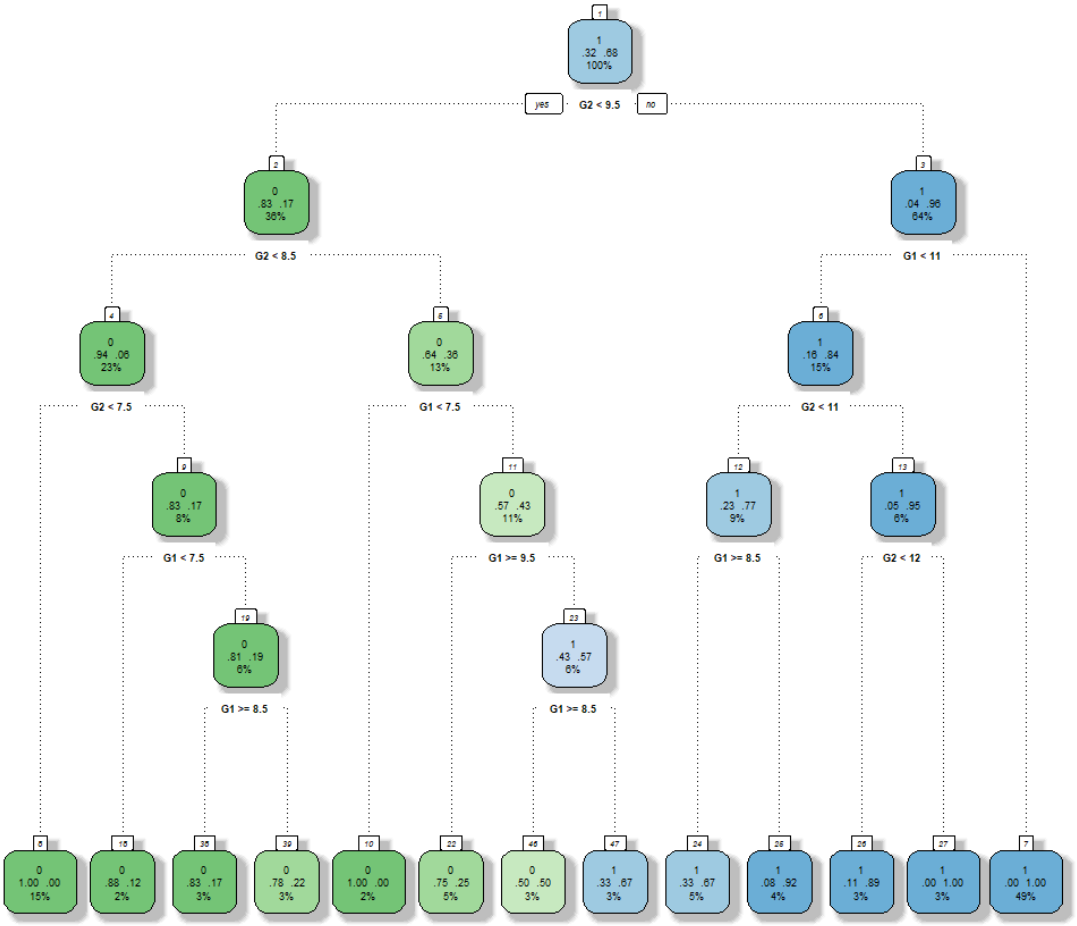
Variable por predecir: pass
Variables a utilizar: G1, G2

Codificación del árbol:

```
arbol_pass_calif <- rpart(
  formula = pass ~ G1 + G2,
  data = mat_data,
  method = 'class',
  cp = -1
)

arbol_pass_calif
fancyRpartPlot(arbol_pass_calif)
```

Árbol graficado:



Árbol 2. Si el estudiante pasará o no su materia considerando sus antecedentes y variables sociales.

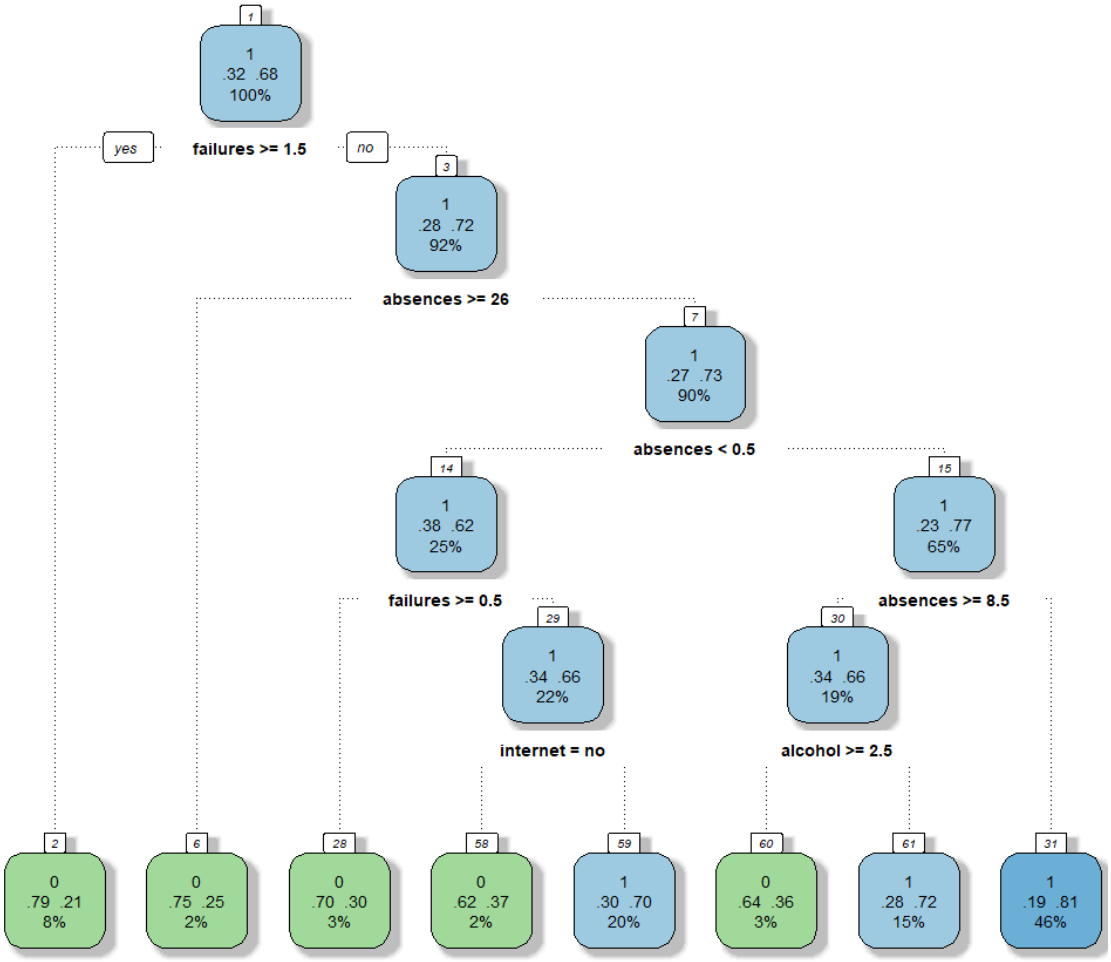
Variable por predecir: pass
Variables a utilizar: failures, internet, absences, alcohol

Codificación del árbol:

```
arbol_pass_data <- rpart(
  formula = pass ~  failures +  internet + absences + alcohol,
  data = mat_data,
  method = 'class'
)

arbol_pass_data
fancyRpartPlot(arbol_pass_data)
```

Árbol graficado:



Árbol 3. Si el estudiante desea o no estudiar el siguiente nivel educativo a partir de factores sociales y educativos.

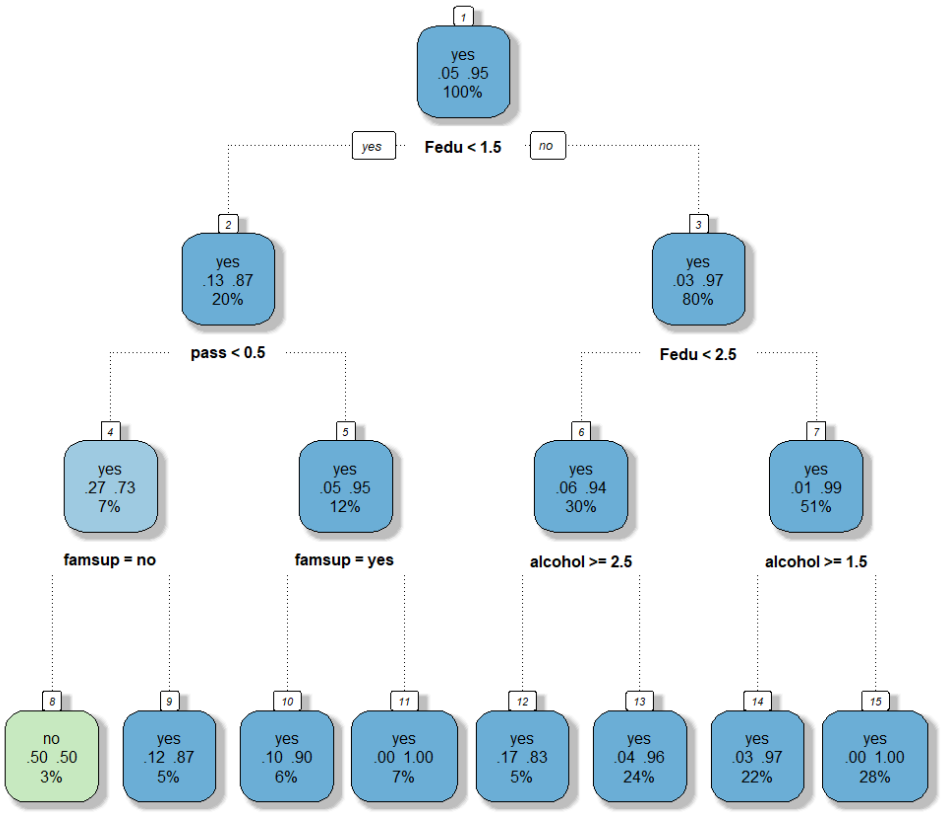
Variable por predecir: higher
Variables a utilizar: sex, address, Pstatus, Medu, Fedu, schoolsup, famsup, famrel, alcohol, pass

Codificación del árbol:

```
arbol_higher <- rpart(
  formula = higher ~ sex + address + Pstatus + Medu + Fedu + schoolsup +
            famsup + famrel + alcohol + pass,
  data = mat_data,
  method = 'class',
  cp = -1,
  maxdepth = 3
)

arbol_higher
fancyRpartPlot(arbol_higher)
```

Árbol graficado:



Árbol 4. El nivel de consumo de alcohol del estudiante considerando factores sociales y educativos.

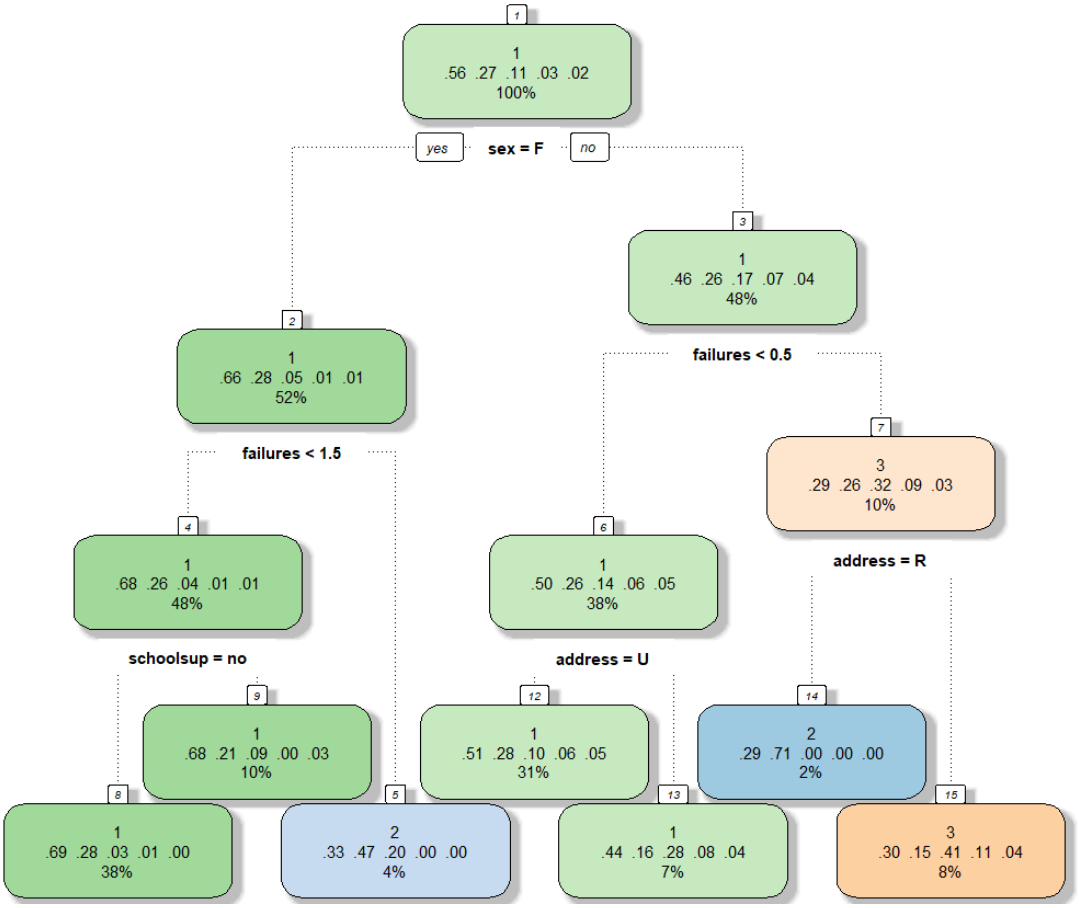
Variable por predecir: alcohol
Variables a utilizar: sex, address , schoolsup, famsup, failures

Codificación del árbol:

```
arbol_alcohol <- rpart(
  formula = alcohol ~ sex + address + schoolsup + famsup + failures,
  data = mat_data,
  method = 'class',
  cp = -1,
  maxdepth = 3
)

arbol_alcohol
fancyRpartPlot(arbol_alcohol)
```

Árbol graficado:



Predicciones con el árbol

Después de la realización de los árboles, se adjuntaron los resultados a un nuevo data frame para comparar las columnas de predicciones con los datos.

Código:

```
#---Predicciones
#Predecir con el arbol
pred_arbol_passCalif <- predict(arbol_pass_calif, type = 'class')
pred_arbol_passData <- predict(arbol_pass_data, type = 'class')
pred_arbol_higher <- predict(arbol_higher, type = 'class')
pred_arbol_alcohol <- predict(arbol_alcohol, type = 'class')


mat_predictions <- cbind(mat_data, pred_arbol_passCalif)
mat_predictions <- cbind(mat_predictions, pred_arbol_passData)
mat_predictions <- cbind(mat_predictions, pred_arbol_higher)
mat_predictions <- cbind(mat_predictions, pred_arbol_alcohol)
```

Conclusiones

Análisis árbol 1 y árbol 2.

Como se menciona en la descripción del data set, G1 y G2 tienen un gran peso para la predicción de si el estudiante pasará el curso o no. Puesto que son los antecedentes del rendimiento de dicho alumno en los periodos pasados. Por ello es que para el primer árbol se escogieron los atributos más obvios, con el fin de obtener predicciones más probables.

El aprendizaje por parte del estudiante ya sea guiado por un docente o con ayuda del internet, es un factor importante para saber si pasará o no el curso, es por ello que se consideraron en el segundo árbol, de igual manera, las veces que el estudiante ha repetido el curso y su consumo de alcohol influyen.

Podemos notar que en el primer árbol, el 68% de los estudiantes aprobaron el curso, los estudiantes, cuyo desempeño fue reprobatorio en su G2, no es probable que vayan a pasar la materia, aunque, existe un pequeño grupo que aunque reprobaron el G2, y como pasaron el G1 tienen una pequeña probabilidad de si pasar la materia.

En el segundo árbol, podemos notar que es más probable que alguien que no ha reprobado, vaya a pasar la materia, además, si el alumno asiste a clases va a pasar la materia, entre otra información.

Comparando ambos datos en la tabla se puede notar los siguiente sobre la eficacia de los anteriores árboles de decisión:

pass	pred_arbol_passCalif	pred_arbol_passData
0	0	1
1	1	1
1	1	1
0	0	1
1	1	1

El árbol que utiliza las variables G1 y G2 falla menos que el que utiliza valores sociales.

Análisis árbol 3.

El árbol tres nos muestra que en la mayoría de los casos, los estudiantes independientemente de su género, si viven en una comunidad rural o urbana, factores

sociales como si sus padres viven juntos o no, el nivel de educación de ambos padres, el apoyo por parte de la escuela y de su familia, su nivel de consumo de alcohol o si pasaron la materia o no, ellos quieren continuar con sus estudios. La excepción del .5% sucede cuando el estudiante ha reprobado la materia, su padre no cuenta con estudios altos y no cuenta tampoco con su apoyo.

Análisis árbol 4.

El ultimo árbol, tuvo como fin predecir el consumo de alcohol entre los estudiantes considerando factores como el sexo, su comunidad, el apoyo por parte de la escuela y su familia y el número de veces que han reprobado la materia. En general, se puede notar que los estudiantes de sexo masculino, independientemente si reprueben o no, de comunidades rurales tienden a tener un consumo de alcohol mas elevado.

Una de las predicciones ejemplo realizada fue la siguiente:

```
> predict(object = arbol_alcohol,
+         newdata = data.frame(sex = 'M', address = 'R', schoolsup = 'yes', famsup = 'no', failures = 4
+         type = 'class')
1
2
Levels: 1 2 3 4 5
>
```

En ella podemos notar que un estudiante masculino de una comunidad rural con el apoyo de la escuela, pero sin el apoyo de su familia que ha fallado en clase 4 veces, es probable que se encuentre en un nivel 2 de consumo de alcohol si se considera un rango de 1 a 5 donde 1 es muy bajo y 5 muy alto.

Conclusión general

La función rpart, contiene varios parámetros de control que nos permiten manipular nuestro árbol, por ejemplo, decidiendo el número máximo de nodos a mostrar. Los árboles de decisión son un método de clasificación que cuenta de dos partes, primeramente, debe aprender, para luego predecir, dándonos un enfoque gráfico de arriba hacia abajo, es decir, de un nodo principal hacia sus hojas.

El data set explorado nos proporcionaba un sin fin de datos para ser tratados y explotados de diversas maneras con el fin de obtener una variedad de información, esta vez, se decidió centrar el trabajo en tres planteamientos, pero se podría utilizar el mismo data set para otros más.