Análisis de la influencia del consumo de alcohol en los resultados académicos utilizando aprendizaje automático

José Antonio Yébenes Montenegro
Máster en Inteligencia de Negocio y Big Data
Modelos de aprendizaje automático

Directora:
Dra. Laia Subirats Maté

Profesoras responsables de la asignatura:
Dra. Teresa Sancho Vinuesa y Dra. María Pujol Jover

7 de Julio de 2017

FICHA DEL TRABAJO FINAL

Título del trabajo:	Análisis de la influencia del consumo de alcohol en los resultados académicos utilizando aprendizaje automático
Autor:	José Antonio Yébenes Montenegro
Directora:	Dra. Laia Subirats Maté
Profesoras responsables de la asignatura:	Dra. Teresa Sancho Vinuesa y Dra. María Pujol Jover
Fecha de entrega:	07/2017
Titulación:	Máster en Inteligencia de Negocio y Big Data
Área del trabajo final:	Modelos de aprendizaje automático
Idioma:	Castellano
Palabras clave:	Análisis de datos, Alcoholismo, Estudiante

Resumen del trabajo:

Objetivo: El consumo de alcohol en exceso es perjudicial a cualquier edad pero en menores es un problema que nos afecta a todos. El objetivo de este trabajo es estudiar si existe relación entre el consumo de alcohol y los resultados académicos en estudiantes de educación secundaria.

Método: Se ha utilizado un conjunto de datos que contiene información social y académica de un grupo de estudiantes de educación secundaria de edades comprendidas entre los 15 y los 22 años de dos escuelas Portuguesas. Se han realizado diferentes tipos de análisis (univariante, bivariante, segmentación, reglas de asociación, profiling y árboles de decisión) para ver si el consumo de alcohol afecta a los resultados o existen otras variables más influyentes.

Resultados: En el grupo que suspende la proporción de estudiantes que no quiere realizar estudios superiores es mayor (20 %) que en general (8,52 %). En el grupo que suspende la proporción de estudiantes del colegio Mousinho da Silveira es mayor (36,95 %) que en general (26,95 %). El número de suspensos previos es la variable más influyente a la hora de predecir si se aprueba o no, la nota media final no supera el aprobado cuando hay suspensos previos.

Conclusiones: El consumo de alcohol influye en las notas finales pero existen otras variables que influyen más como el colegio al que va, si quiere seguir estudiando, los suspensos previos, si hace refuerzo escolar, los estudios de los padres o el número de ausencias.

Abstract:

Objective: Alcohol consumption in excess is harmful at any age but in minors is a problem that affects us all. The objective of this work is to study the relationship between alcohol consumption and academic achievement in secondary school students.

Method: The dataset used contains social and academic information from a group of high school students aged 15-22 of two Portuguese schools. Different types of analysis (univariate, bivariate, segmentation, association rules, profiling and decision trees) has been performed to see if alcohol consumption affects the results or there are other more influential variables.

Results: In the group that fails the proportion of students who do not want to study higher, it is higher (20 %) than in general (8.52 %). In the group that fails the proportion of students of the school Mousinho da Silveira is higher (36.95 %) than in general (26.95 %). The number of previous fails is the most influential variable in predicting whether to pass or not, the final average grade does not exceed 10 points when there are previous fails.

Conclusions: The consumption of alcohol influences the final grades, but there are other variables that influence more, such as the school to which it is going, if it wants to continue studying, the previous fails, if it makes school reinforcement, the studies of the parents or the number of absences.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Contexto y justificación del trabajo

El conjunto de datos elegido se ha obtenido del repositorio de Machine Learning de la Universidad de California, Irvine (UCI) `https://archive.ics.uci.edu/`. Se ha seleccionado el dataset Student Alcohol Consumption `https://archive.ics.uci.edu/ml/datasets/STUDENT+ALCOHOL+CONSUMPTION` para poder hacer un trabajo relacionado con algún tema médico o de salud. Actualmente hay una campaña de sensibilización sobre este tema `https://www.youtube.com/watch?v=Vv7Qc23aras`, esto hace que sea un tema de actualidad y de preocupación social.

Según datos de la OMS sobre el consumo de alcohol [18]:

- Cada año se producen 3,3 millones de muertes en el mundo debido al consumo nocivo de alcohol, lo que representa un 5,9 % de todas las defunciones.
- El uso nocivo de alcohol es un factor causal en más de 200 enfermedades y trastornos.
- El 5,1 % de la carga mundial de morbilidad y lesiones es atribuible al consumo de alcohol.
- El consumo de alcohol provoca defunción y discapacidad a una edad relativamente temprana. En el grupo etario de 20 a 39 años, un 25 % de las defunciones son atribuibles al consumo de alcohol.
- Existe una relación causal entre el consumo nocivo de alcohol y una serie de trastornos mentales y comportamentales, además de las enfermedades no transmisibles y los traumatismos.
- Recientemente se han determinado relaciones causales entre el consumo nocivo y la incidencia de enfermedades infecciosas tales como la tuberculosis y el VIH/sida.
- Más allá de las consecuencias sanitarias, el consumo nocivo de alcohol provoca pérdidas sociales y económicas importantes, tanto para las personas como para la sociedad en su conjunto.

Estos datos hacen reflexionar sobre las consecuencias del consumo de alcohol, en especial entre los jóvenes. La campaña de la FAD (Fundación de Ayuda contra la Drogadicción) pretende concienciar de que se trata de una situación extendida y que no se le presta la atención necesaria.

El consumo de alcohol cada vez se inicia a una edad más temprana. Se asocia con actividades de ocio y a veces incluso se propicia o tolera en el ambiente familiar y escolar. Por ello se trata de un problema de todos.

Los datos utilizados se recolectaron como parte de un estudio para alertar anticipadamente a los profesores del deterioro en los resultados académicos tiendo el consumo de alcohol como variable implicada.

El conjunto de datos tiene 1044 observaciones correspondientes a variables socio-demográficas y académicas de los estudiantes de secundaria de dos escuelas y dos asignaturas (matemáticas y portugués) en dos institutos de Portugal. Los datos están repartidos en dos ficheros uno para la asignatura de matemáticas (student-mat.csv) y otro para la asignatura de portugués (student-por.csv). El dataset tiene 32 variables la mayoría de ellas (26) son categóricas (sobre su entorno familiar, horas de estudio, consumo de alcohol, etc.) y (6) son numéricas (edad, número de suspensos anteriores, número de ausencias y notas).

Dada la variedad, la cantidad de registros y los tipos de variables se pueden aplicar diferentes técnicas de análisis descriptivo (análisis estadístico, segmentación, reglas de asociación) y predictivo (árboles de decisión) para determinar si el consumo de alcohol influye en los resultados académicos.

1.2. Objetivos del trabajo

El objetivo del trabajo es analizar si existe relación entre el consumo de alcohol y los resultados académicos en estudiantes de educación secundaria. Para ello se trabaja con un conjunto de datos que contiene información social y académica de un grupo de estudiantes de educación secundaria de edades comprendidas entre los 15 y los 22 años de dos escuelas Portuguesas.

El análisis de este conjunto de datos tiene como objetivos:

- **Analizar la calidad de los datos:** es importante asegurar que los datos son completos y coherentes antes de iniciar el estudio. En caso de existir valores faltantes se deberán tomar las medidas necesarias para regularizarlo.
- **Detectar si existen valores outliers:** es importante conocer si existen valores extremos ya que su presencia puede alterar los resultados obtenidos sobre todo si se aplican técnicas basadas en distancias y no se estandarizan los datos.
- **Entender los datos mediante visualizaciones sencillas:** para ver los valores que toma cada una de las variables y su distribución, se mostrarán gráficos de barras que permitirán entender mejor los datos que se están trabajando. Estas visualizaciones permitirán una mejor contextualización a la hora de interpretar los resultados obtenidos.
- **Entender los datos a través del análisis estadístico univariante y bivariante:** para entender los datos y poderlos interpretar correctamente se realizará un análisis univariante sencillo (summary). También se analizará si existen variables altamente correlacionadas y se buscarán relaciones entre las variables mediante visualizaciones sencillas (diagramas de puntos por pares de variables). Dependiendo del tipo de técnica de análisis que se utilice las variables muy correlacionadas habrá que excluirlas, de este modo se puede reducir la dimensionalidad y eliminar redundancias en los datos.
- **Segmentar los datos:** mediante la técnica de aprendizaje automático no supervisado Kmeans, se agruparán los estudiantes en grupos lo más diferentes posible de manera que dentro de cada grupo los estudiantes sean los más parecidos posible. Esto permitirá ver si el consumo de alcohol manifiesta comportamientos distintos en los diferentes grupos de estudiantes o por el contrario es parecido.
- **Detectar reglas de asociación:** a través de esta técnica de aprendizaje automático no supervisado se buscarán relaciones entre las variable y ciertos valores del estilo “cuando pasa A también

pasa B". Este tipo de reglas no denotan una relación de causa-efecto, detectan situaciones que se dan a la vez pero no necesariamente una es la causa de la otra. Interesará detectar relaciones en las que en A aparezcan variables relacionadas con el consumo de alcohol y donde en B aparezca la variable de éxito (superar o no la asignatura).

- **Detectar las variables más influyentes en el hecho de superar o no la asignatura:** usando la técnica de profiling (descripción de características) se estudiará qué variables son las más influyentes sobre la variable de éxito/fracaso (superar o no la asignatura). Esto ayudará a determinar si el consumo del alcohol es la más influyente o existen otros factores más importantes.

- **Construir un modelo predictivo sobre la variable objetivo (superar o no la asignatura):** usando los árboles de decisión (técnica de aprendizaje automático supervisado) se pretende entrenar y evaluar un modelo predictivo. Este modelo alertará cuando un estudiante no va a superar la asignatura, prestando especial atención si uno de los factores influyentes es el consumo de alcohol.

1.3. Enfoque y método seguido

1.3.1. Entorno de trabajo

El entorno de trabajo elegido es RStudio ya que es la interfaz más completa y versátil para trabajar con R y también porque permite interpretar código LaTeX e integrarlo con código R y sus resultados. Después de hacer muchas pruebas con R Markdown se ha descartado ya que el tipo de documento que se puede confeccionar aunque sea pdf es más limitado.

LaTeX permite confeccionar artículos y libros científicos de alta calidad tipográfica permitiendo integrar código R y sus ejecuciones, imágenes externas, links internos dentro del documentos (por ejemplo en el índice y las referencias bibliográficas), links externos... todo ello con una complejidad relativamente baja. Como manual de referencia de LaTeX se ha utilizado [5].

La mayor parte del trabajo se ejecuta en una máquina Linux con sistema operativo Xubuntu, se requiere tener instalado un compilador/editor de LaTeX, en este caso se ha utilizado Texmaker. Además de los paquetes de R básicos, para la edición de textos se requieren los paquetes: knitr, rmarkdown [8, 9] y Sweave [6, 11].

El esqueleto de libro en formato LaTeX para Trabajos de Final de Máster se ha obtenido de la web de la Universidad Santiago de Compostela [10] pero se han requerido muchas adaptaciones para:

- Incluir links internos (índice y referencias bibliográficas), se requiere `\usepackage{hyperref}` [12].

- Incluir código R [6, 11] y que compile correctamente el código LaTeX.

- Se requiere `\usepackage[utf8]{inputenc}` y `\usepackage[spanish]{babel}` para escribir en español sin tener que usar \' para que aparezcan las tildes.

1.3.2. Metodología

La metodología que se ha utilizado es una aproximación a CRISP-DM (Cross Industry Standard Process for Data Mining) que es una metodología enfocada a proyectos de minería de datos y que tiene una componente cíclica/iterativa que permite corregir y mejorar las fases iniciales a medida que se avanza en el proceso y se identifican puntos de mejora en las fases previas.

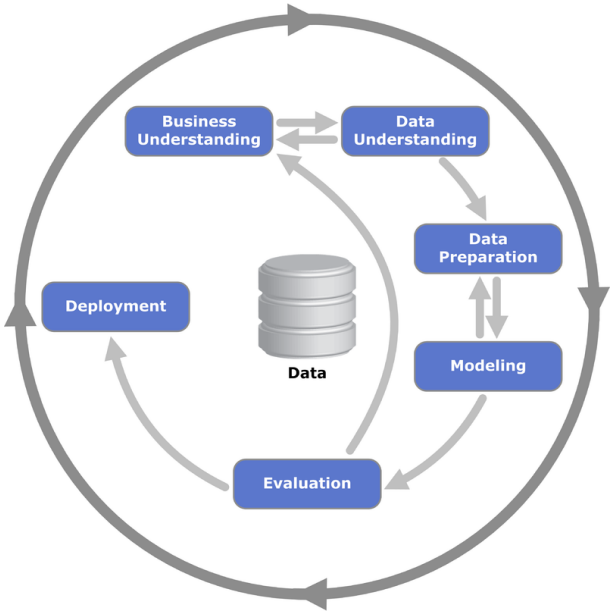

Figura 1.1: Metodología CRISP-DM [14]

Tal como se explica en [14], esta metodología consta de las siguientes fases:

1. **Comprensión del negocio**

   Esta fase inicial se centra en la comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial, y luego convertir este conocimiento en una definición del problema de minería de datos, y un plan preliminar diseñado para alcanzar los objetivos.

   Esta fase se desarrolla a través de la definición de los objetivos del trabajo y en el apartado **Descripción de los ficheros originales** donde se describen los ficheros y el contexto de los datos.

2. **Comprensión de Datos**

   Esta fase comienza con una colección inicial de datos y procesos con actividades con el objetivo de familiarizarse con los datos, identificar la calidad de los problemas, para descubrir las primeras señales dentro de los datos y detectar temas interesantes para poder formular hipótesis de información oculta.

   Esta fase se desarrolla en los apartados **Preparación de los datos**, **Análisis de la calidad de los datos**, **Interpretación de los datos** y **Análisis univariante y bivariante**.

3. **Preparación de datos**

   Esta fase cubre todas las actividades para construir el conjunto de datos. Estas tareas son ejecutadas en múltiples oportunidades y sin orden. Las tareas incluyen selección y transformación de tablas, registros y atributos y limpieza de datos para las herramientas de modelado.

En el apartado **Preparación de los datos** se cargan los datos y se preparan de forma genérica para los tratamientos posteriores. Después, en cada uno de los apartados, se seleccionan las variables más apropiadas y se acaban de hacer los ajustes para preparar los datos según las necesidades específicas de cada técnica.

4. **Modelado**

   En esta fase se seleccionan y aplican varias técnicas de modelado y se calibran los parámetros para obtener óptimos resultados. Hay varias técnicas que tienen requerimientos específicos para la forma de los datos, por lo que frecuentemente es necesario volver a la fase de preparación de datos.

   Esta fase se desarrolla en los apartados **Segmentación**, **Reglas de asociación**, **Profiling** y **Árboles de decisión**. En cada uno de estos apartados se aplica una técnica diferente para analizar los datos y ver si existe relación entre el consumo de alcohol y los resultados académicos. En algunos de ellos, como la segmentación, las reglas de asociación y los árboles de decisión se ejecuta la misma técnica con parámetros distintos para encontrar el mejor modelo.

5. **Evaluación**

   En esta etapa en el proyecto ha construido un modelo (o modelos) que parece tener gran calidad, desde una perspectiva de análisis de datos.

   En el caso de los árboles de decisión esta fase es vital para identificar qué modelo es mejor a los demás. Para ello se evalúan los modelos con la fracción de los datos que no se ha utilizado en el entreno del modelo.

6. **Despliegue**

   Esta fase depende de los requerimientos, pudiendo ser simple como la generación de un reporte o compleja como la implementación de un proceso de explotación de información que atraviese a toda la organización.

   En este caso, no se ejecuta esta fase del proceso en el sentido estricto de "puesta en producción"del modelo obtenido pero sí se resumen las conclusiones y resultados obtenidos en el apartado **Conclusiones**.

1.4. Planificación del trabajo

Esta es la planificación del trabajo ajustada a la duración y a las fechas de entrega establecidas en el plan docente.

Fecha Inicio	Fecha Fin	Días	Entrega / Tarea
15/03/2017	19/03/2017	5	Selección y comunicación de la temática del trabajo final de máster
20/03/2017	26/03/2017	7	Detalle de las tareas a realizar y preparación del entorno de trabajo
27/03/2017	02/04/2017	7	Preparación de la PEC1 y Carga de los datos y ajustes del data-frame
03/04/2017	09/04/2017	7	Calidad de los datos y detección de outliers
-	09/04/2017	-	Entrega Propuesta Inicial (PEC1)
10/04/2017	16/04/2017	7	Entender los datos mediante visualizaciones sencillas
17/04/2017	23/04/2017	7	Análisis univariante y bivariante
24/04/2017	30/04/2017	7	Segmentación y Preparación de la PEC2
01/05/2017	07/05/2017	7	Reglas de asociación
-	07/05/2017	-	Entrega Estructura de la Memoria (PEC2)
08/05/2017	14/05/2017	7	Profiling
15/05/2017	28/05/2017	14	Predicción con árboles de decisión
29/05/2017	18/06/2017	21	Elaboración de la memoria (PEC3)
-	23/06/2017	-	Entrega Primera Versión de la Memoria (PEC3)
19/06/2017	30/06/2017	12	Elaboración de la presentación
-	30/06/2017	-	Calificación Primera versión de la memoria
01/07/2017	06/07/2017	6	Retoques de la memoria en base a la calificación de la PEC3
-	07/07/2017	-	Entrega Final TFM
10/07/2017	14/07/2017	5	Tribunal de evaluación

Tabla 1.1: Planificación: tareas y entregas

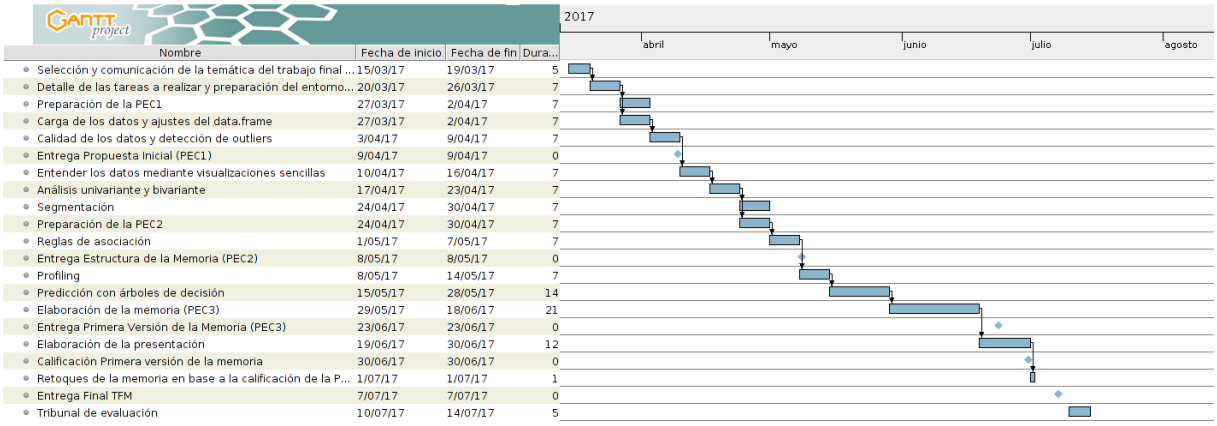

Figura 1.2: Planificación: Gannt

1.5. Breve sumario de productos obtenidos

Al presente documento se anexa el script en R **yebenesja_TFM_Rscript.Rmd** que contiene todo el procesado de datos, visualizaciones y análisis necesarios para desarrollar el trabajo final de máster.

Como los datos utilizados son de origen público se ha compartido este script en Kaggle `https://www.kaggle.com/yebenesja/an-lisis-de-la-influencia-del-consumo-de-alcohol` para fomentar la colaboración y el estudio de este conjunto de datos.

Se adjunta también el documento **yebenesja_TFM_Rscript.pdf** que contiene todo el código R utilizado, comentado y con sus ejecuciones donde se puede obtener cualquier detalle del proceso de análisis ya que muestra todos los cálculos parciales y todas las visualizaciones que se han realizado aunque finalmente no se hayan incluido en la memoria.

Ambos documentos, de forma análoga a la presente memoria, se ha estructurado en los siguientes apartados:

- Configuración del entorno de trabajo
- Preparación de los datos
- Análisis de la calidad de los datos
  - Detección de nulls y na
  - Detección de outliers
- Interpretación de los datos
- Análisis univariante y bivariante
- Segmentación
  - Segmentación óptima
- Reglas de asociación
- Profiling
- Árboles de decisión

1.6. Breve descripción de los otros capítulos de la memoria

En el capítulo 2 (Datos):

- se encuentra el Estado del Arte
- se describen los ficheros originales
- se describe el proceso de carga y transformación para obtener el dataset básico de trabajo
- se analiza la calidad de los datos detectando valores faltantes y valores extremos
- finalmente se interpretan los datos usando visualizaciones sencillas para entenderlos

1.6. BREVE DESCRIPCIÓN DE LOS OTROS CAPÍTULOS DE LA MEMORIA

En el capítulo 3 (Análisis) se utilizan diferentes técnicas de análisis de datos sobre el dataset preparado previamente:

- se realiza un análisis estadístico univariante y bivariante
- se utiliza la técnica de aprendizaje automático no supervisado K-Means para segmentar los estudiantes
- se usa el algoritmo de aprendizaje automático no supervisado 'Apriori' para detectar reglas de asociación
- se realiza un profiling (descripción de características) para detectar las variables más influyentes sobre el hecho de superar o no la asignatura
- finalmente se usan árboles de decisión, técnica de aprendizaje automático supervisado, para crear un modelo para predecir si el estudiante supera la asignatura o no

En el capítulo 4 se presentan las conclusiones del trabajo y en el capítulo 5 se confecciona un glosario de términos y acrónimos utilizados.

Capítulo 2

Datos

En este capítulo se describen los ficheros originales y el proceso de carga y transformación para obtener el dataset básico de trabajo. Se analiza también la calidad de los datos detectando valores faltantes y valores outliers. Además de explicar el contenido de los ficheros originales se pretende entender los datos y empezar a obtener información de los mismos a través de visualizaciones sencillas.

2.1. Estado del Arte

El análisis de datos es un disciplina con una larga historia que actualmente está en auge. La era de la computación ha favorecido muchos avances para potenciar la analítica de datos como base del proceso de Knowledge Discovery in Databases (KDD). El KDD es el proceso que permite la transformación de los datos en información, la información en conocimiento o inteligencia y ese conocimiento servirá de palanca para la toma de decisiones y desencadenar acciones concretas que permitan el desarrollo del negocio.

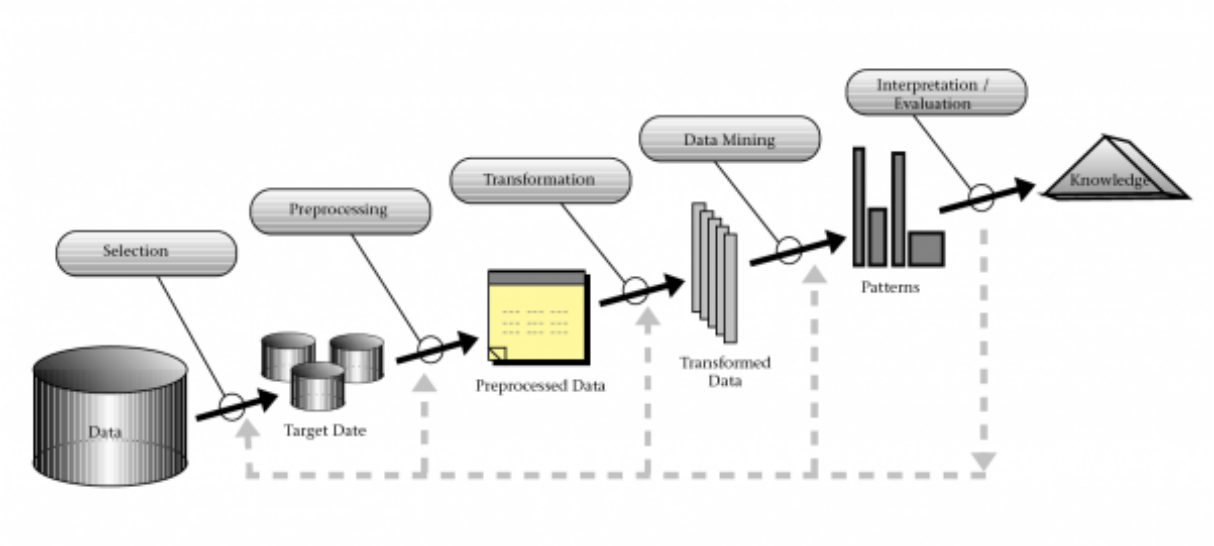

Figura 2.1: Proceso de Knowledge Discovery in Databases [23]

Resulta prácticamente imposible establecer fronteras claras entre la gran cantidad de disciplinas implicadas en este proceso: minería de datos, aprendizaje automático, estadística, matemáticas, computación, inteligencia artificial, etc.

Figura 2.2: Big Data word cloud

En los últimos años gracias a la reducción de costes en almacenamiento de datos, la mejora en las capacidades de cómputo y los avances tecnológicos para el tratamiento de grandes cantidades de información han hecho posible que el análisis de datos esté experimentando un desarrollo nunca vista hasta el momento.

Existen muchos estudios sobre el análisis del consumo de alcohol en estudiantes de educación secundaria. Algunos de los más interesantes:

- Patterns and predictors of alcohol use among 7-12th grade students in New York State [19]. En base a los datos recopilados, se observó que el 71 % de los estudiantes eran bebedores y el 13 % eran bebedores intensos. El consumo elevado de alcohol estaba relacionado con mala conducta escolar y malas calificaciones.
- Parenting Style and Behavior as Longitudinal Predictors of Adolescent Alcohol Use [20]. Este estudio se realizó en Australia. Demostró que los comportamientos familiares influían más en el consumo de alcohol en los adolescentes que el estilo educativo de los padres.
- Alcohol consumption and academic performance in a population of Spanish high school students [21]. El estudio concluyó que el riesgo de fracaso aumenta junto con el consumo de alcohol pero que el rendimiento académico también está influenciado por muchos factores distintos al consumo de alcohol.

- Alcohol consumption among high school students in Barcelona, Spain [22]. Este estudio se realizó para conocer la prevalencia del consumo de alcohol entre los adoscentes en los periodos 1992-93 y 1994-95. A pesar de que la prevalencia del consumo del alcohol pasó del 92,5 % en 1992-93 al 77,0 % en 1994-95, se mostró un aumento de la cantidad de alcohol consumido.

También existen muchos análisis y estudios del dataset seleccionado [13]. Según se explica en la documentación asociada al dataset estos datos fueron recogidos con la finalidad de elaborar un modelo predictivo que permitiera alertar a los docentes de la probabilidad de fracaso de los estudiantes de educación secundaria del sistema educativo portugués. El estudio principal es 'Using data mining to predict secondary school student performance' [7] pero en la web de Kaggle se pueden encontrar algunos más con scripts públicos en R o Python `https://www.kaggle.com/uciml/student-alcohol-consumption`. Algunos son puramente descriptivos, otros tienen una componente predictiva centrada en las notas pero hay muchos otros con enfoques diferentes, por ejemplo, predecir si un estudiante tiene o no una relación sentimental.

El objetivo del presente trabajo es determinar si existe influencia del consumo de alcohol en los resultados académicos desde distintos enfoques de análisis y usando técnicas variadas tanto de aprendizaje automático supervisado como no supervisado.

2.2. Descripción de los ficheros originales

En la documentación del dataset [13, 7] obtenemos la descripción de cada uno de los campos que constituyen los dos ficheros (uno para la clase de matemáticas y otro para la clase de lengua Portuguesa) de dos colegios de Portugal.

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:

1. **school** - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2. **sex** - student's sex (binary: 'F' - female or 'M' - male)
3. **age** - student's age (numeric: from 15 to 22)
4. **address** - student's home address type (binary: 'U' - urban or 'R' - rural)
5. **famsize** - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6. **Pstatus** - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7. **Medu** - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
8. **Fedu** - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
9. **Mjob** - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10. **Fjob** - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11. **reason** - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12. **guardian** - student's guardian (nominal: 'mother', 'father' or 'other')

13. **traveltime** - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

14. **studytime** - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

15. **failures** - number of past class failures (numeric: n if 1<=n<3, else 4)

16. **schoolsup** - extra educational support (binary: yes or no)

17. **famsup** - family educational support (binary: yes or no)

18. **paid** - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

19. **activities** - extra-curricular activities (binary: yes or no)

20. **nursery** - attended nursery school (binary: yes or no)

21. **higher** - wants to take higher education (binary: yes or no)

22. **internet** - Internet access at home (binary: yes or no)

23. **romantic** - with a romantic relationship (binary: yes or no)

24. **famrel** - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

25. **freetime** - free time after school (numeric: from 1 - very low to 5 - very high)

26. **goout** - going out with friends (numeric: from 1 - very low to 5 - very high)

27. **Dalc** - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

28. **Walc** - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

29. **health** - current health status (numeric: from 1 - very bad to 5 - very good)

30. **absences** - number of school absences (numeric: from 0 to 93)

    The grading system in Portugal is a 20-point scale, the minimum grade for passing is 10. These grades are related with the course subject, Math or Portuguese:

31. **G1** - first period grade (numeric: from 0 to 20)

32. **G2** - second period grade (numeric: from 0 to 20)

33. **G3** - final grade (numeric: from 0 to 20, output target)

    Additional note: there are several (382) students that belong to both datasets . These students can be identified by searching for identical attributes that characterize each student, as shown in the annexed R file.

2.3. Preparación de los datos

En la fase de preparación de datos se cargan los dos ficheros originales y se prepara el dataframe para el análisis. También se crean algunas variables adicionales y se configuran todas las variables categóricas con el significado de sus valores para que gráficos y resultados sean más fáciles de interpretar sin tener que recurrir a consultar la documentación del dataset.

A continuación se detalla el proceso:

1. Cargar el fichero 'student-mat.csv' que contiene los datos de los alumnos de matemáticas en un primer dataframe.

2. Cargar el fichero 'student-por.csv' que contiene los datos de los alumnos de lengua portuguesa en un segundo dataframe.

3. En cada uno de los dataframes se crea el campo file que indica si el registro proviene del fichero 'student-mat.csv' que se identifica con 'math' o si proviene del fichero 'student-por.csv' que se identifica con 'portuguese'.

4. En cada uno de los dataframes se crea el campo subject que indica si el estudiante está matriculado en la asignatura de matemáticas 'math' o portugués 'port'. Más adelante los alumnos que están en ambas asignaturas de codifican como 'math&port' en esta variable.

5. Combinar los dos dataframes en uno solo que se llama 'data'.

6. Verificar si hay valores duplicados. No se detectan duplicados.

7. Normalizar los nombres de las variables poniendo todos los nombres en minúsculas.

8. Crear un identificador único de estudiante concatenando los campos: school, sex, age, address, famsize, pstatus, medu, fedu, mjob, fjob, reason, nursery e internet. Esto se explica en el script R adjunto a la documentación del dataset [13]. Este indicador 'student_key' permitirá identificar los estudiantes que se encuentran en las dos asignaturas (en los dos ficheros).

9. Crear una variable binaria 'bing3' en función de las notas finales (g3). Toma el valor 1 si g3 es mayor o igual a 10, toma el valor 0 en caso contrario.

10. Crear una variable que indique si se aprueba una asignatura o la otra. Se construye en base a las variables g3 y subject. Toma los valores: 'fail math', 'fail port', 'pass math' o 'pass port'.

11. Crear una variable numérica que mida el consumo de alcohol global como media ponderada por el número de días.

$$alcohol = \frac{5 * dalc + 2 * walc}{7}$$

12. Crear una función para identificar y clasificar con el valor 'math&port' en la variable 'subject' los alumnos que están en ambos ficheros.

13. Guardar un backup de los datos en el dataframe 'raw_data'.

14. Codificar todas las variables categóricas como factor. Incluyendo las descripciones de cada uno de los factores. De este modo todos los gráficos y análisis que se hagan serán fácilmente interpretables ya que en lugar de ver el valor numérico de la categoría se verá su valor descriptivo.

15. Construir un dataframe de metadatos con el nombre de la variable, su descripción, el tipo de variable, un contador de nulos y un contador de NA. El campo 'variable' se rellena usando la función 'names(data)', el campo 'descripcion' se rellena manualmente y el campo 'tipo' con la función 'sapply(data, class)'.
16. Rellenar el campo 'descripcion' con una breve explicación del contenido de cada una de las variables tanto las originales como las derivadas.
17. Crear una variable 'variables_numericas' para almacenar los nombres de los campos con valores numéricos (integer o numeric), servirá para filtrar el dataset en función del análisis o el tratamiento que se quiera hacer.
18. Crear una variable 'variables_discretas' para almacenar los nombres de los campos con valores finitos (todos menos los campos de tipo character), servirá como lista para hacer un gráfico para cada unos de los campos usando un bucle `for`.
19. Crear una variable 'variables_discretas_names' para almacenar las descripciones de los campos con valores finitos (todos menos los campos de tipo character), servirá como lista para los títulos de los gráficos y los nombres para el índice de figuras.

2.4. Análisis de la calidad de los datos

En esta fase se analiza la calidad de los datos una vez hecha la carga y preparación inicial. Dado que en la documentación del dataset no hay información sobre reglas de coherencia entre los datos no se pueden construir reglas de análisis de calidad más allá de verificar si existen valores faltantes o valores atípicos.

Partiendo del dataframe de metadatos 'variables' construido anteriormente, se completan los campos 'cuenta_null' y 'cuenta_na' usando las funciones `is.null` y `is.na` respectivamente para cada uno de los campos del dataframe 'data'.

Tal como indica la documentación del dataset [13], se verifica que no existen valores faltantes por lo que no resulta necesario realizar ninguna acción adicional.

A continuación se comprueba si existen valores atípicos en el dataframe 'data'.

Visualizando un boxplot de las variables numéricas se observan valores atípicos en todas las variables:

- En la variable edad, 22 años se considera outlier pero es una edad correcta.
- En las variables g1, g2 y g3 considera outlier el valor 0 pero es un valor correcto.
- En la variable failures considera outliers los valores 1, 2 y 3 pero también son valores correctos.
- En la variable alcohol considera outliers los valores superiores a 4 pero son valores correctos ya que se mide en una escala de 1 a 5.

Si se quisieran eliminar todos los valores extremos haría falta iterar el proceso de detección / eliminación tres veces. Se eliminarían 349 registros de los 1044 por lo que se perdería el 33 % de los datos.

Dado que en todas las variables los valores outliers detectados no parecen datos incoherentes se opta por no eliminarlos.

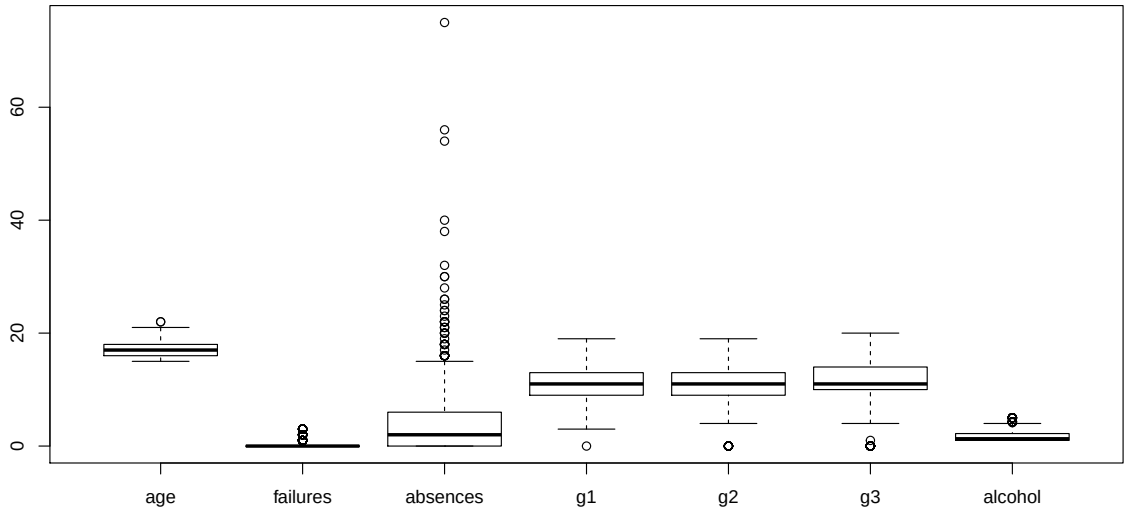

Figura 2.3: Boxplot de las variables númericas sin eliminar los valores atípicos

A continuación se analiza por separado la variable 'absences' que es la que tiene mayor dispersión y se crea un dataframe 'data_clean' en el que se eliminan únicamente los valores outliers correspondientes a esta variable. En esta variable se consideran outliers los valores superiores a 15, con la limpieza solo se pierden 54 registros.

Solo se eliminan los outliers en la variable ausencias ya que presentan valores muy extremos y pueden afectar por ejemplo en el análisis cluster. Los registros que quedan después de eliminar los valores extremos se guardan en el dataframe 'data_clean'.

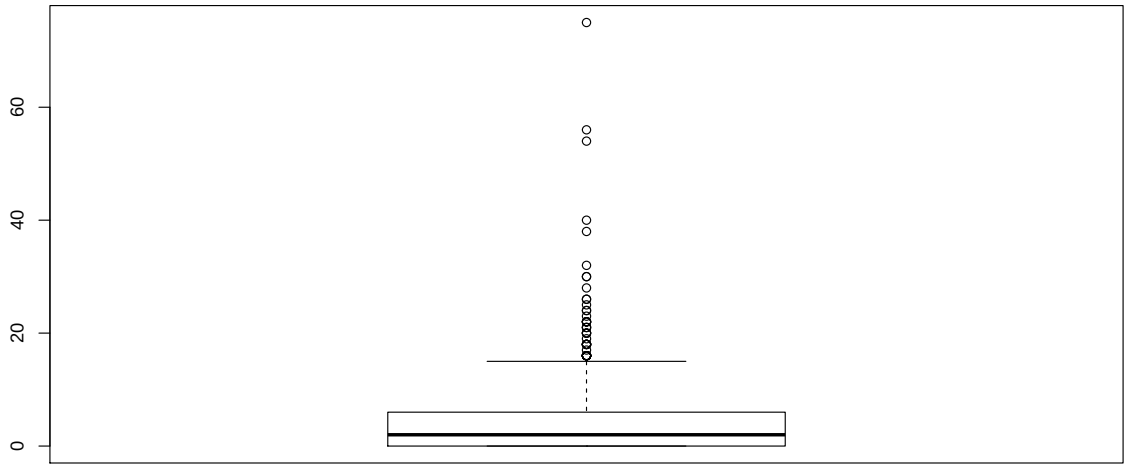

Figura 2.4: Boxplot de la variable número de ausencias (absences) sin eliminar los valores atípicos

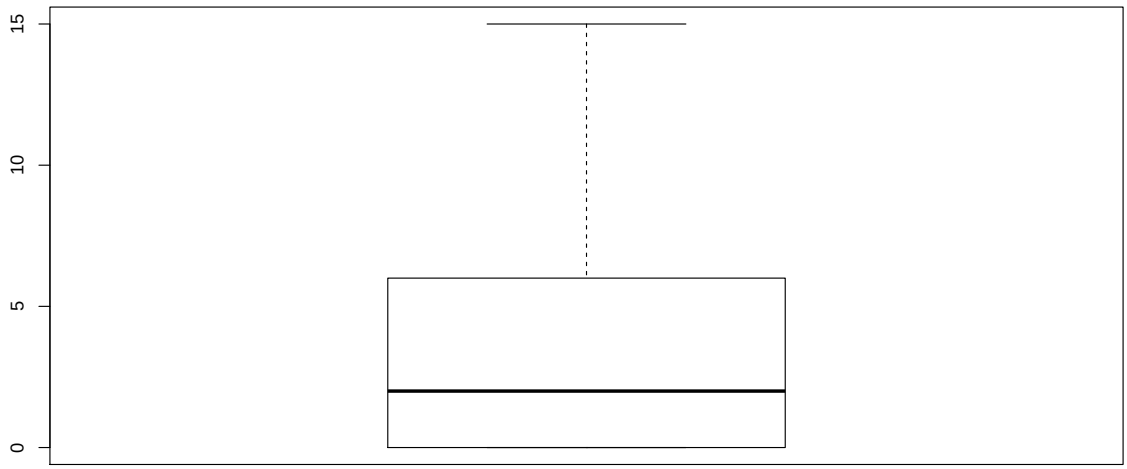

Figura 2.5: Boxplot de la variable número de ausencias una vez eliminados los valores atípicos

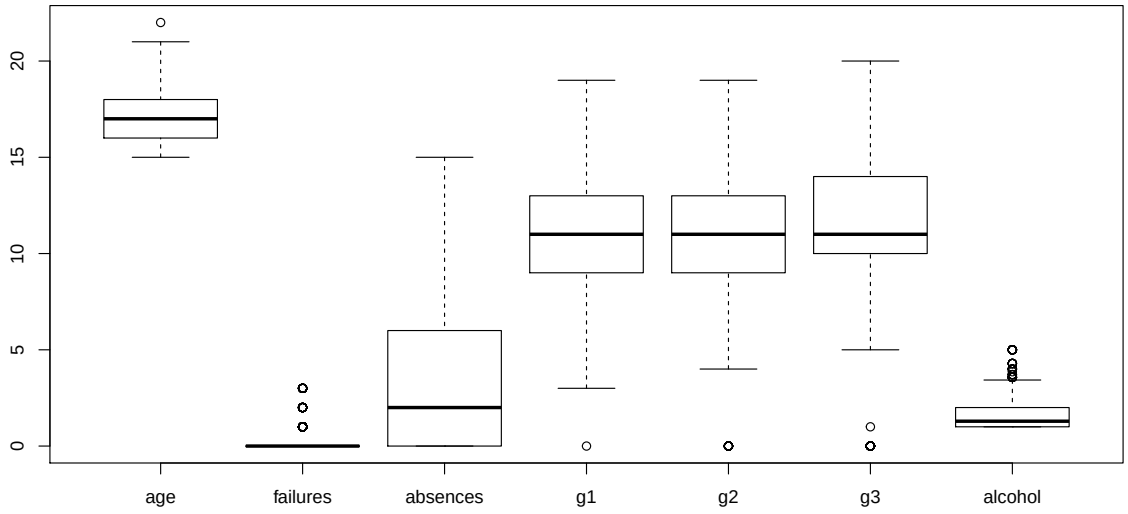

Figura 2.6: Boxplot de las variables numéricas sin los valores atípicos de la variable número de ausencias

2.5. Interpretación de los datos

A continuación se visualiza un gráfico para cada una de las variables. Esto permitirá resumir y obtener información para entender los datos y su distribución:

1. El 73,9 % de los alumnos son de la escuela Gabriel Pereira.
2. El 56,6 % de los estudiantes son del sexo femenino.
3. El 93,3 % de los alumnos tienen una edad comprendida entre los 15 y los 18 años.
4. El 72,7 % de los estudiantes vive en un entorno urbano.
5. La mayoría de estudiantes (70,7 %) pertenecen a una familia de más de 3 miembros.
6. En el 11,6 % de los casos los padres viven separados.
7. La mayoría de las madres (29,3 %) tienen estudios superiores.
8. La mayoría de los padres (31 %) tienen estudios intermedios.
9. El 38,2 % de las madres tienen un empleo del grupo 'otros' y el 18,6 % trabajan en casa.
10. El 55,9 % de los padres tienen un empleo del grupo 'otros' y el 5,9 % trabajan en casa.
11. La mayoría de estudiantes (41,2 %) escogieron la escuela por preferencia de curso.
12. La persona encarga del alumno en la mayoría de los casos (69,7 %) es la madre.
13. Para la mayoría (59,7 %) el trayecto de casa a la escuela es inferior a 15 minutos.

14. La mayoría de estudiantes (48,2 %) dedican entre 2 y 5 horas de estudio semanales.
15. El 82,5 % de los alumnos no han suspendido la asignatura previamente, un 11,5 % la han suspendido una vez y un 6,1 % la han suspendido dos o tres veces.
16. El 11,4 % tiene apoyo educativo adicional.
17. En la mayoría de los casos (61,3 %) los estudiantes reciben ayuda escolar de la familia.
18. La mayoría de alumnos (78,9 %) no reciben clases extras.
19. Prácticamente la mitad realiza actividades extra escolares (49,4 %).
20. El 80 % de los estudiantes ha sido atendido en la enfermería alguna vez.
21. La gran mayoría (91,5 %) desean realizar estudios superiores.
22. En el 20,8 % de los casos no disponen de conexión a internet en casa.
23. El 35,5 % de los alumnos tienen una relación sentimental.
24. En el 2,9$ de los casos consideran que la relación familiar es muy mala y en el 4,5 % que es mala. La mayoría (49 %) considera que su relación familiar es buena.
25. El 22,5 % de los estudiantes considera que tienen poco o muy poco tiempo libre.
26. En el 30,6 % de los casos consideran que salen poco o muy poco con sus amigos.
27. En consumo de alcohol en días laborables se considera muy bajo en el 69,6 % de los casos.
28. En cambio, el consumo de alcohol en fin de semana se considera muy bajo en el 38,1 % de los casos. En el 20,2 % lo consideran alto o muy alto.
29. La mayoría de casos (37,8 %) considera su estado de salud muy bueno. El 24,9 % lo consideran malo o muy malo.
30. El 34,4 % de los alumnos tiene 0 ausencias.
31. La nota más frecuente el primer semestre es 10. La mediana es 11. Y la media 11,21.
32. La nota más frecuente el segundo semestre es 11. La mediana es 11. Y la media 11,25.
33. Las notas finales más frecuentes son 10 y 11. La mediana es 11. Y la media 11,34.

Las siguientes variables se han construido en base a los campos originales.

34. **file:** La mayoría de los registros (62,2 %) provienen del fichero 'student-por.csv'.
35. **subject:** La mayoría de datos (71,9 %) corresponden a alumnos de ambas asignaturas.
36. **bing3** - binary final grade result: El 22 % suspende la asignatura.
37. **result** - subject and result: El 12,5 % suspende matemáticas y el 9,6 % suspende portugués.
38. **alcohol** - global alcohol consumption: En la mayoría de los casos el consumo global de alcohol es muy bajo. En una escala de 1 a 5, la mediana es 1,29 y la media 1,72.

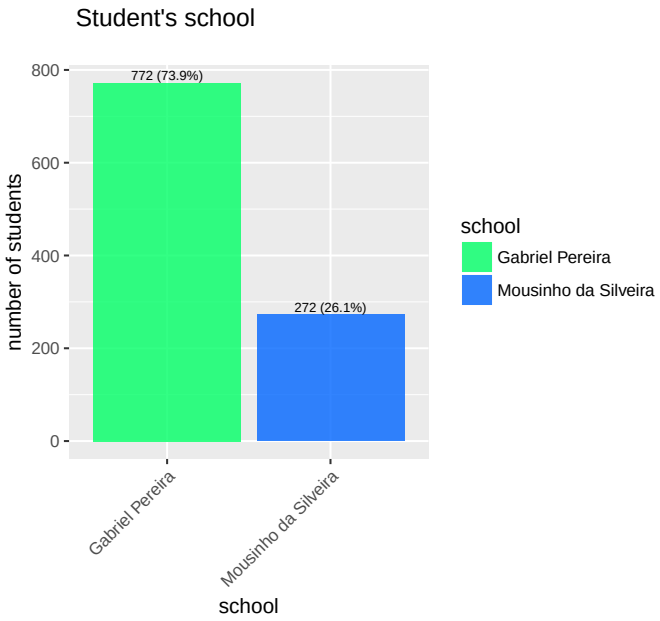

Figura 2.7: Student's school

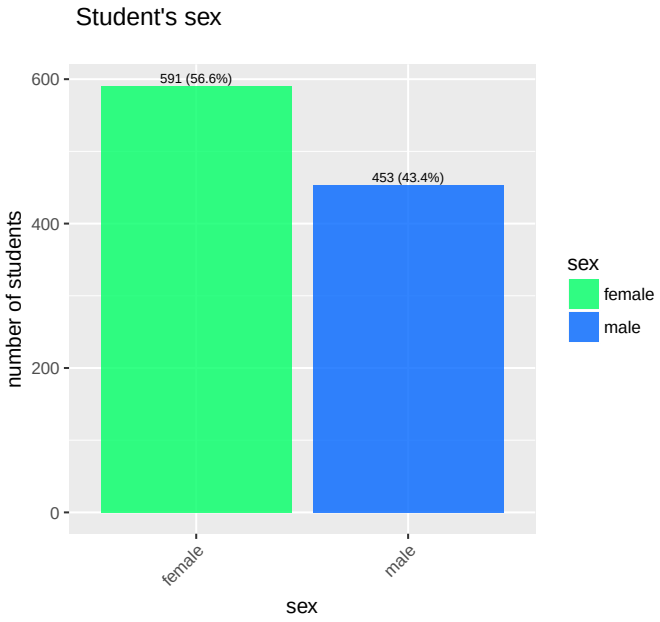

Figura 2.8: Student's sex

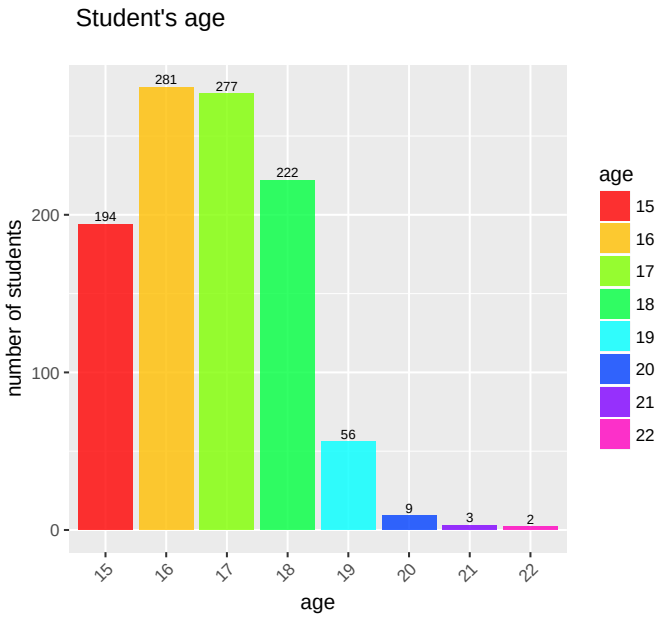

Figura 2.9: Student's age

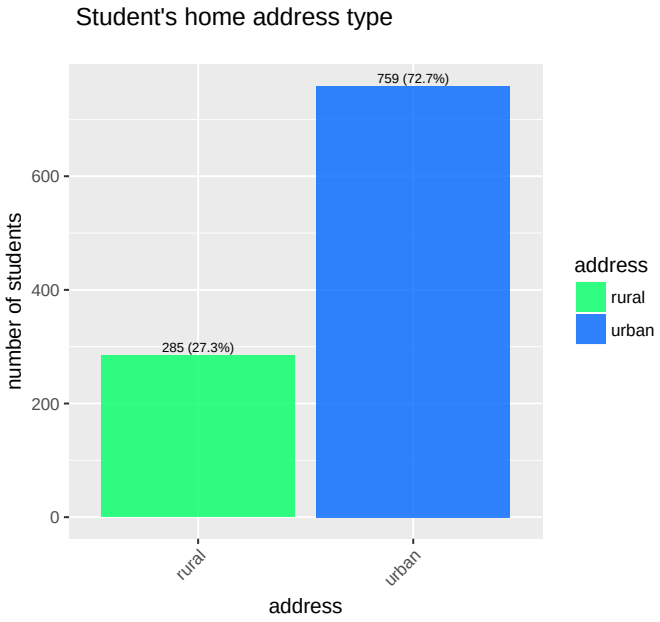

Figura 2.10: Student's home address type

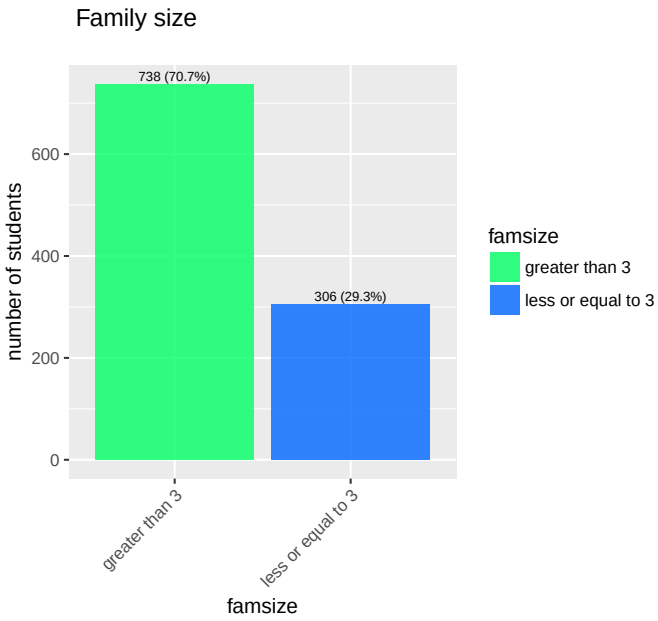

Figura 2.11: Family size

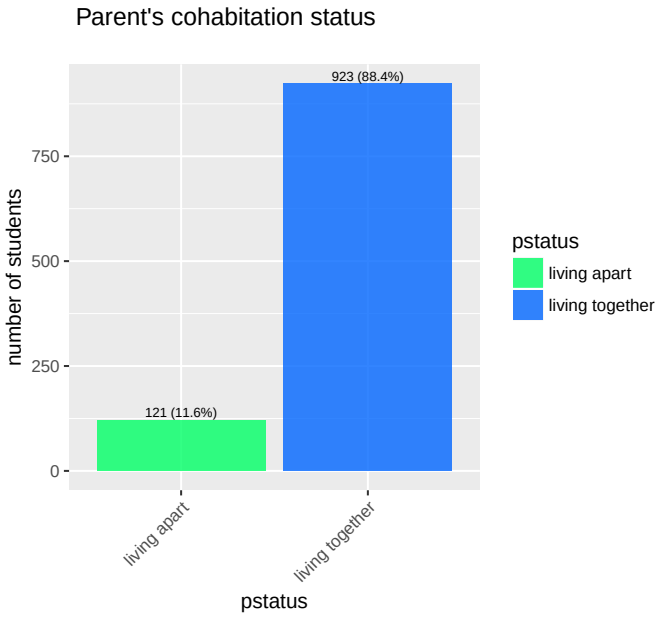

Figura 2.12: Parent's cohabitation status

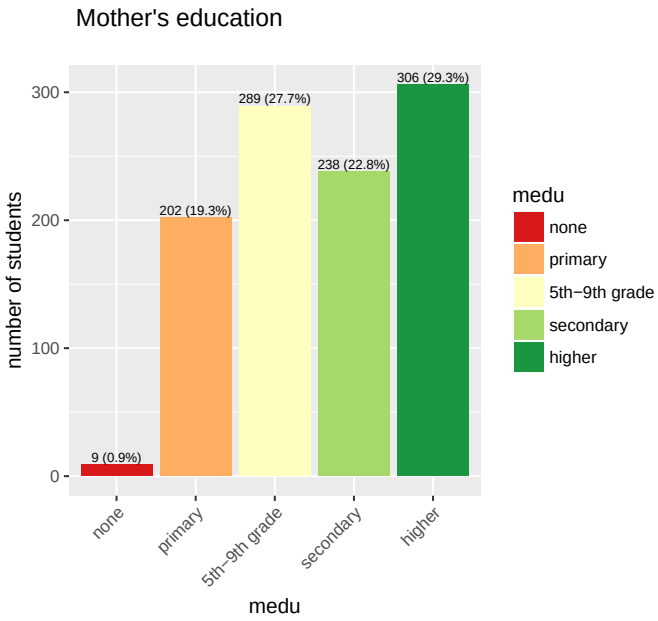

Figura 2.13: Mother's education

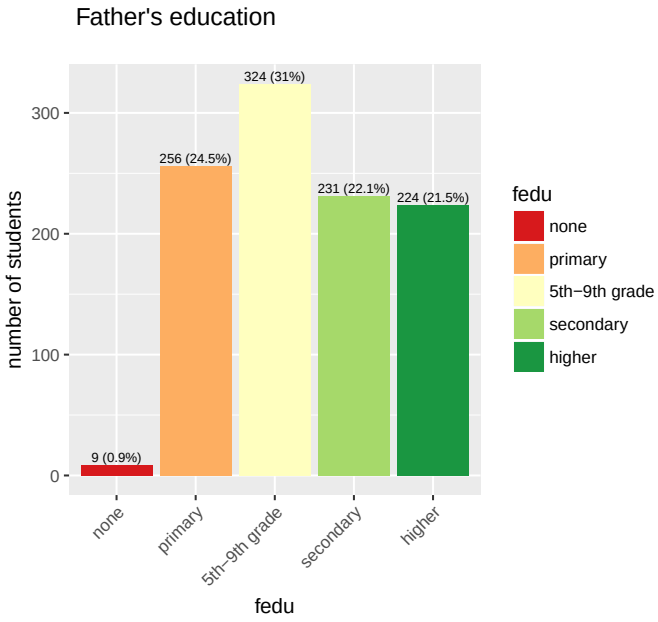

Figura 2.14: Father's education

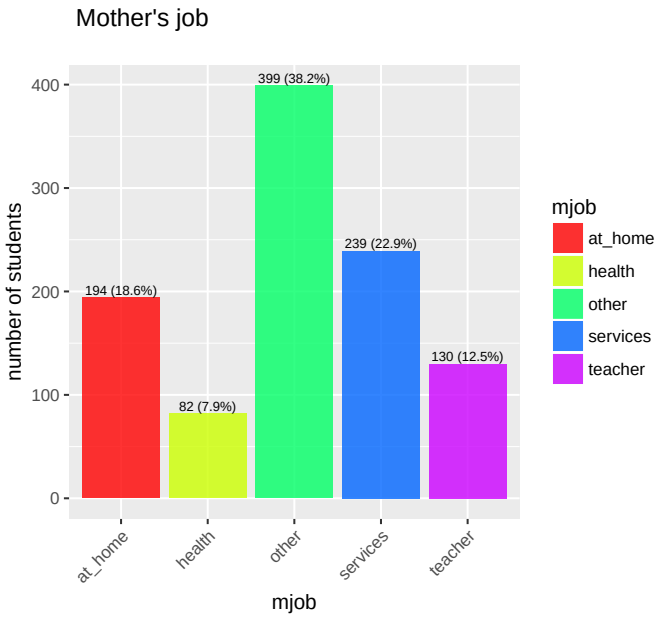

Figura 2.15: Mother's job

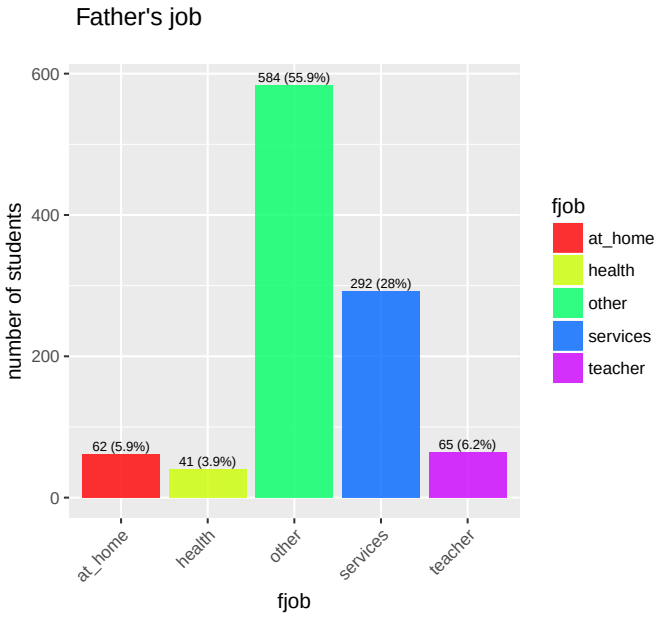

Figura 2.16: Father's job

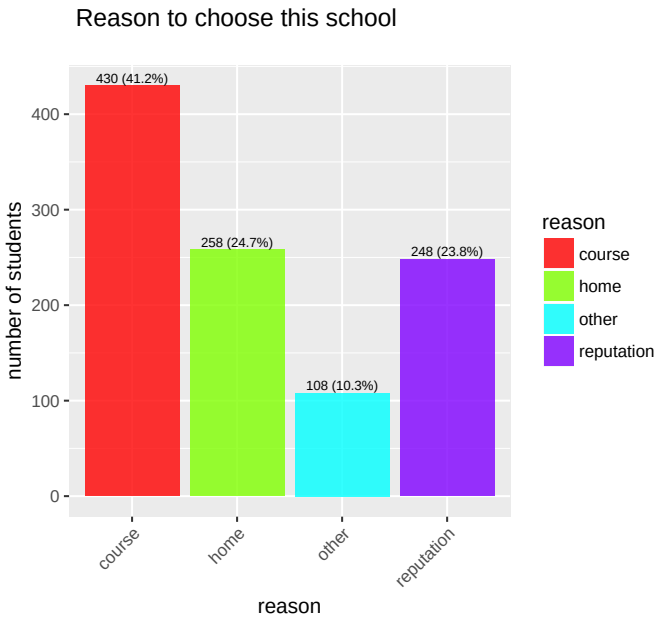

Figura 2.17: Reason to choose this school

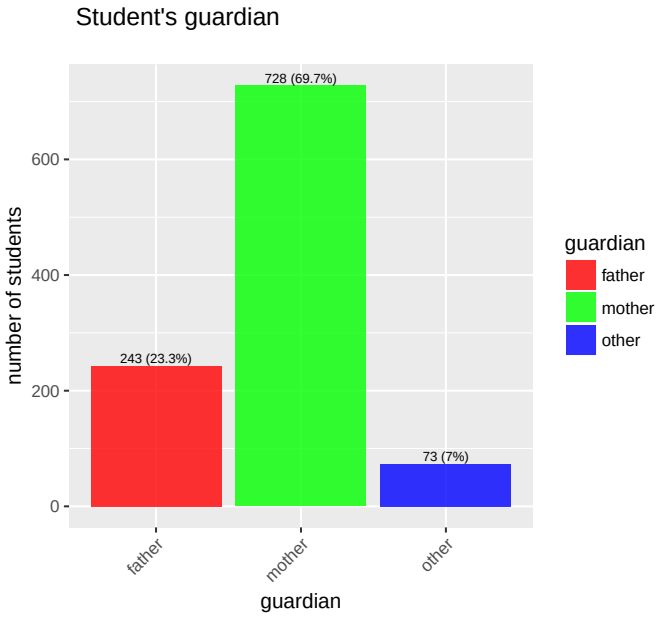

Figura 2.18: Student's guardian

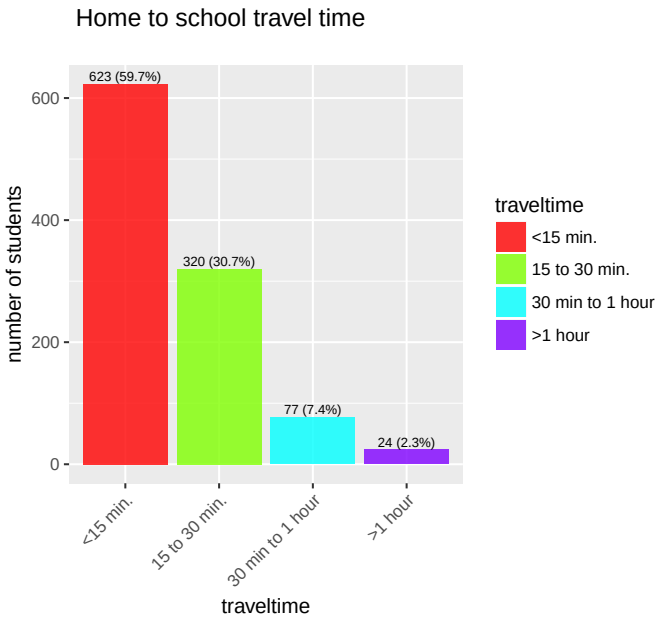

Figura 2.19: Home to school travel time

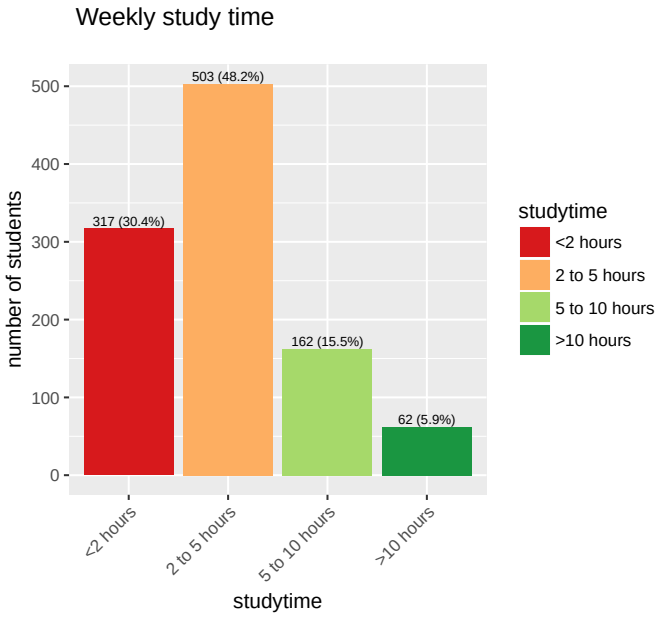

Figura 2.20: Weekly study time

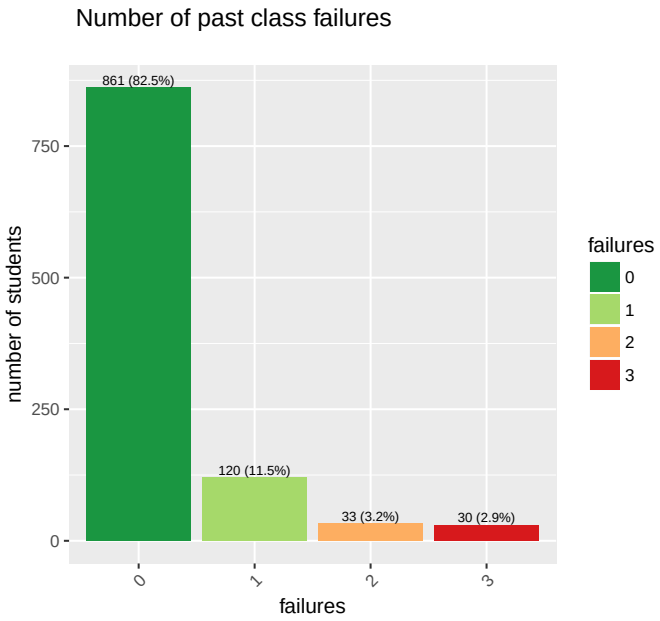

Figura 2.21: Number of past class failures

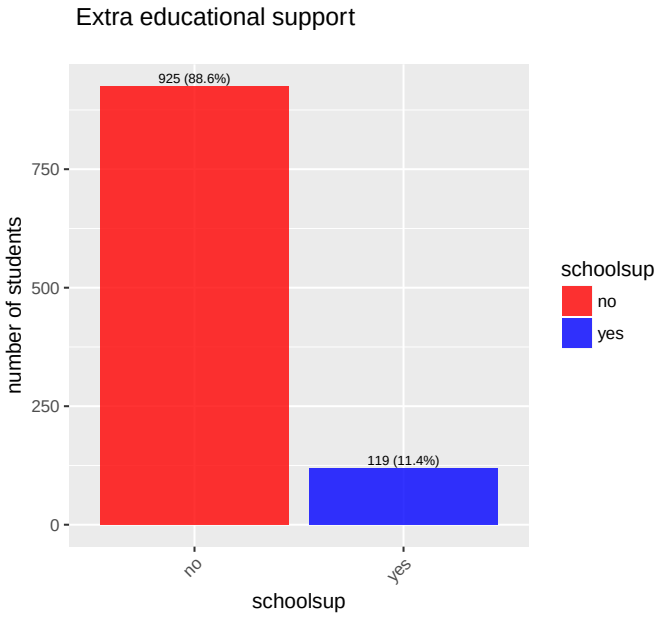

Figura 2.22: Extra educational support

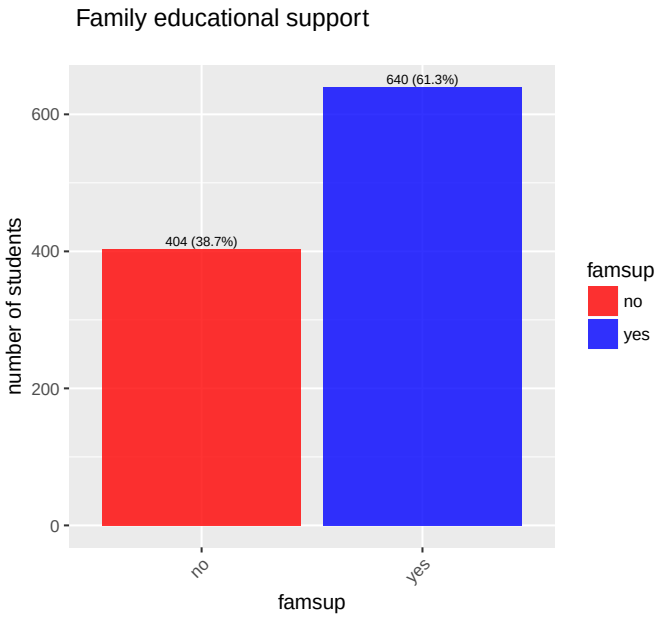

Figura 2.23: Family educational support

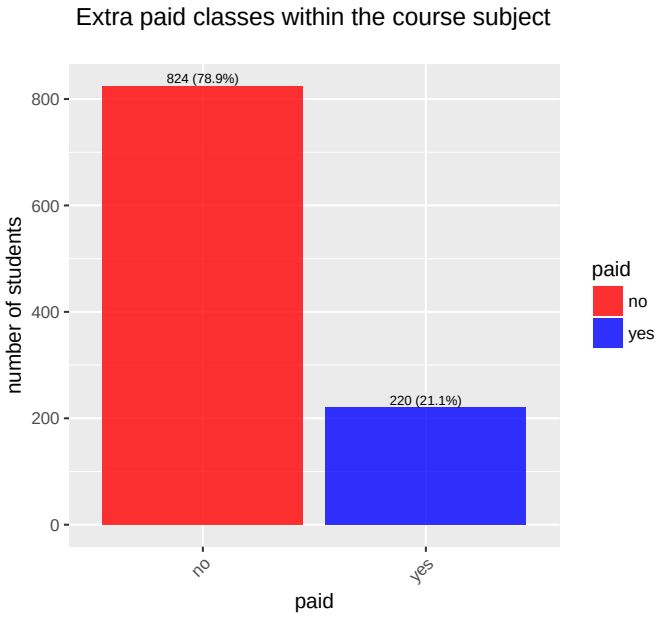

Figura 2.24: Extra paid classes within the course subject

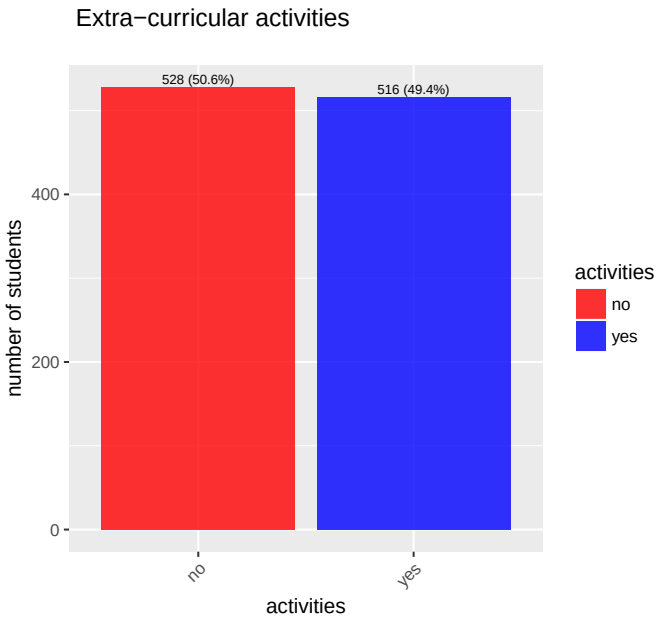

Figura 2.25: Extra-curricular activities

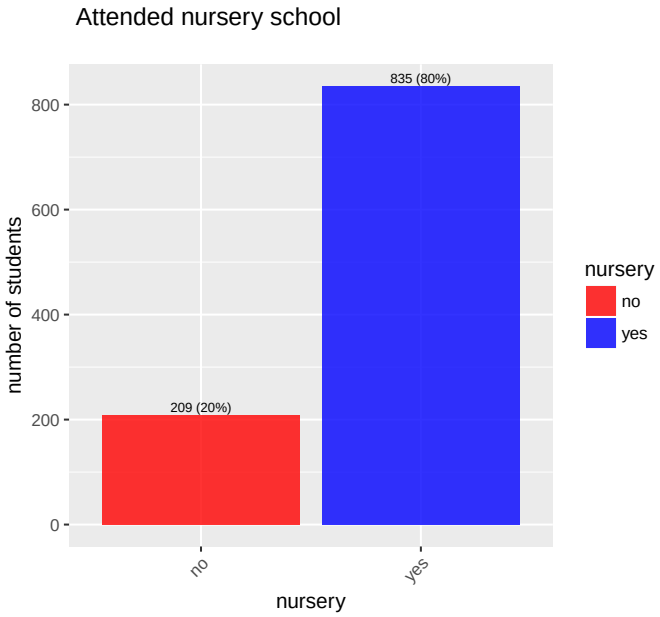

Figura 2.26: Attended nursery school

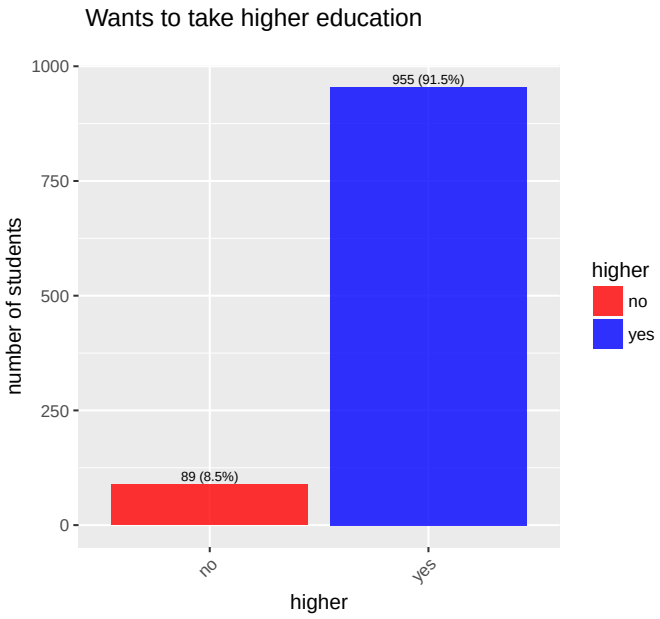

Figura 2.27: Wants to take higher education

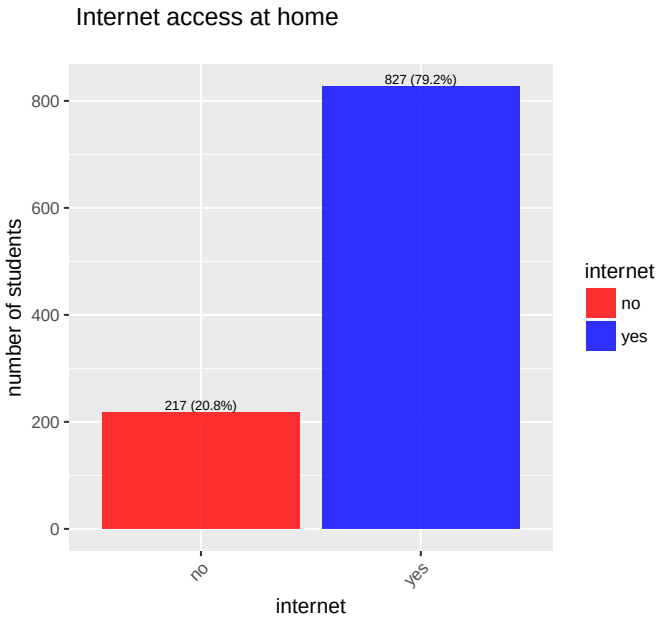

Figura 2.28: Internet access at home

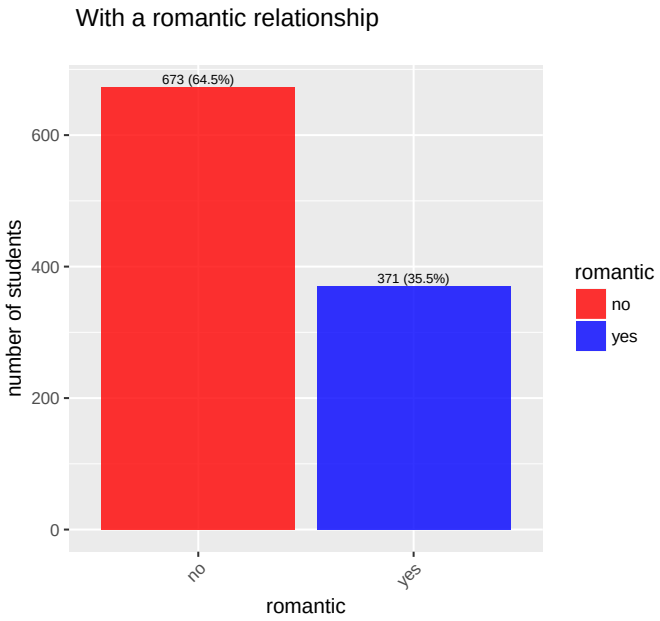

Figura 2.29: With a romantic relationship

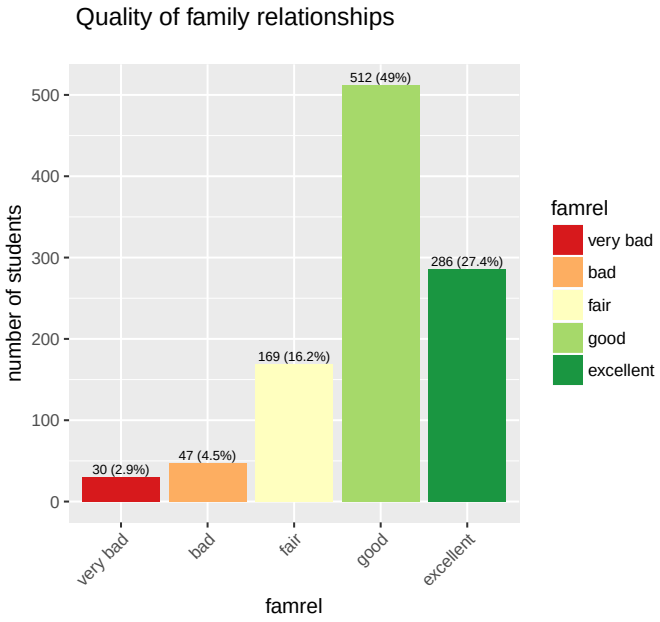

Figura 2.30: Quality of family relationships

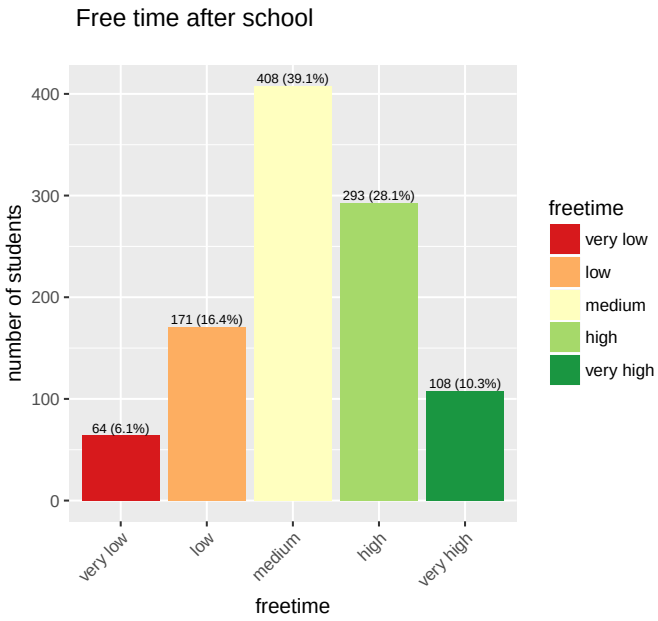

Figura 2.31: Free time after school

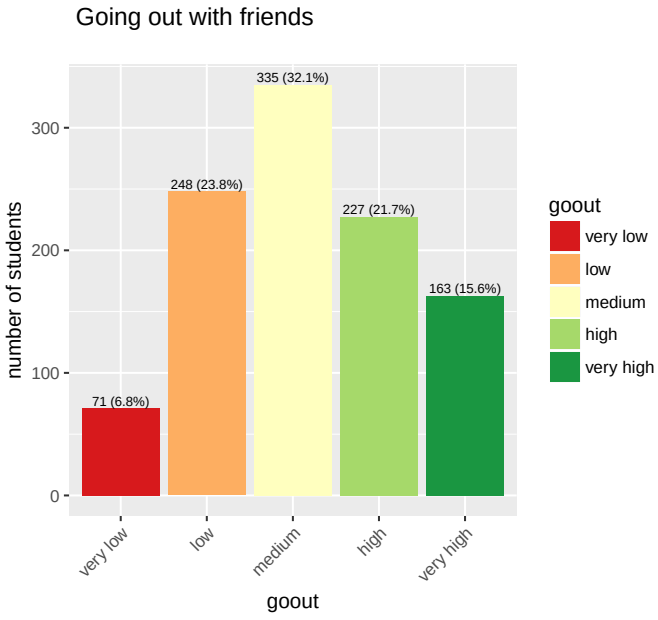

Figura 2.32: Going out with friends

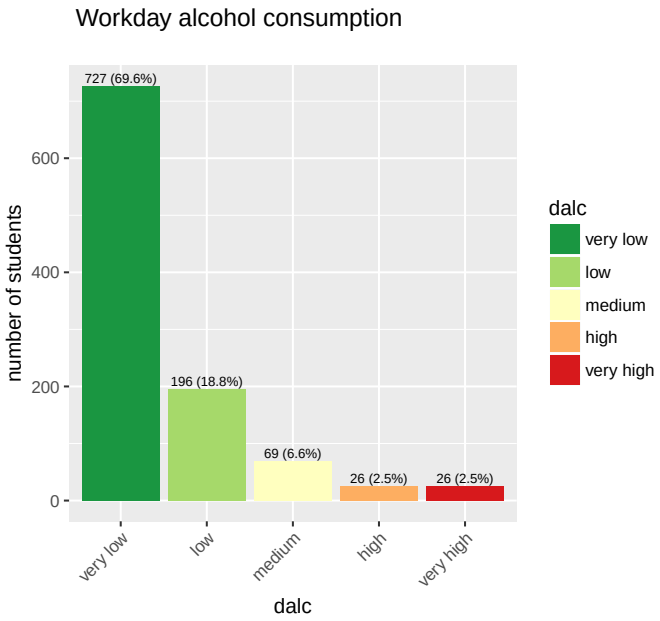

Figura 2.33: Workday alcohol consumption

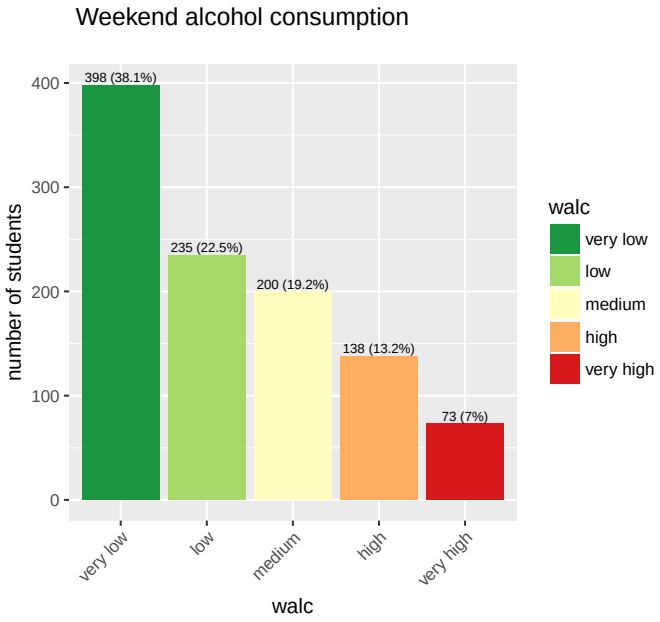

Figura 2.34: Weekend alcohol consumption

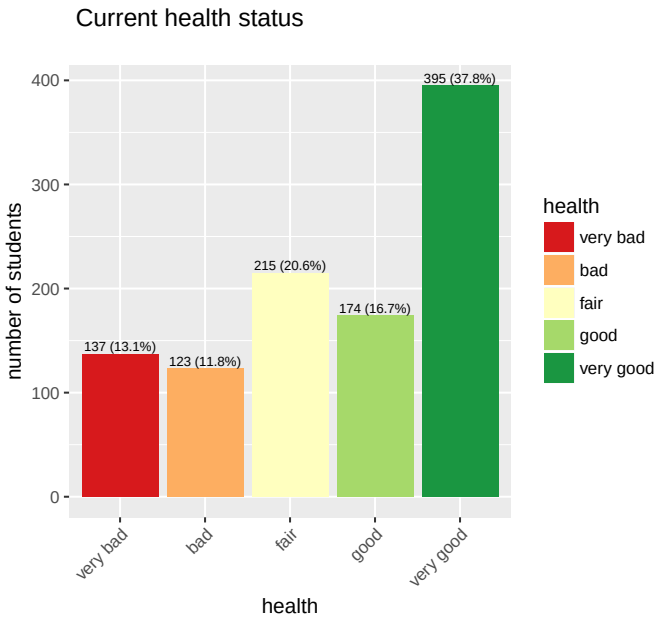

Figura 2.35: Current health status

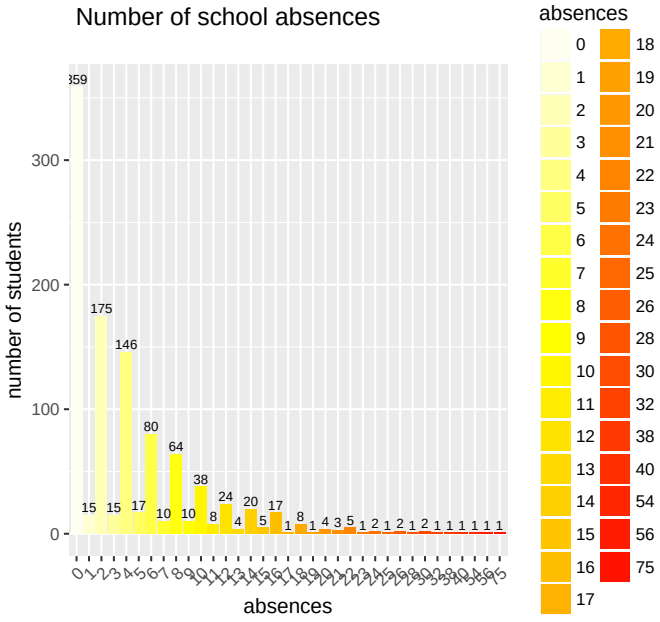

Figura 2.36: Number of school absences

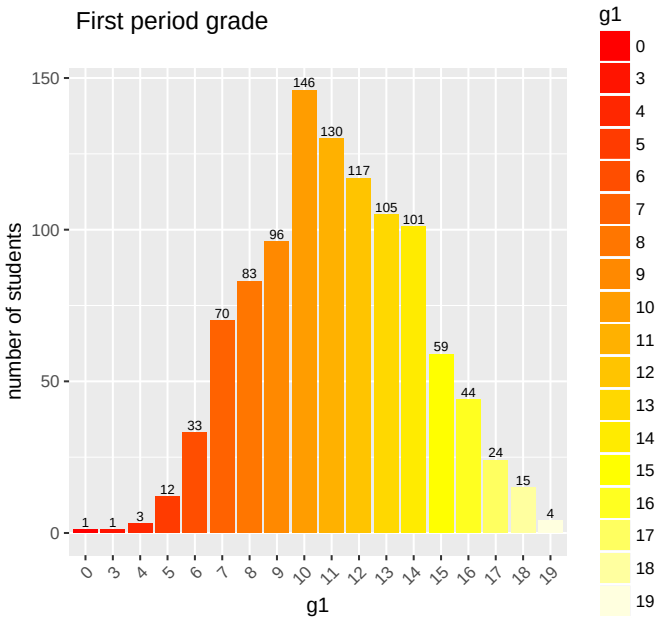

Figura 2.37: First period grade

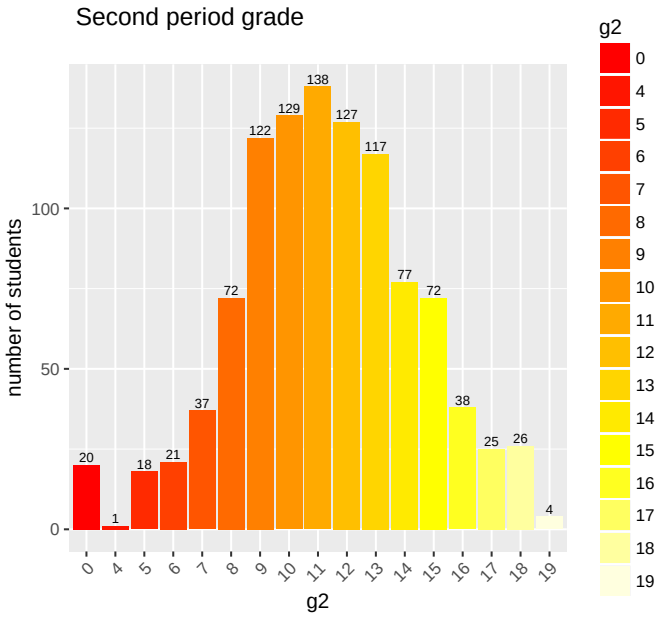

Figura 2.38: Second period grade

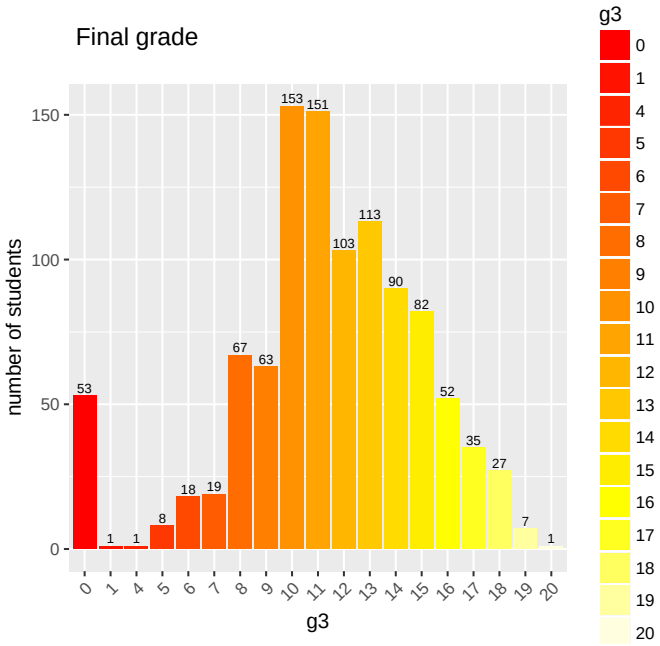

Figura 2.39: Final grade

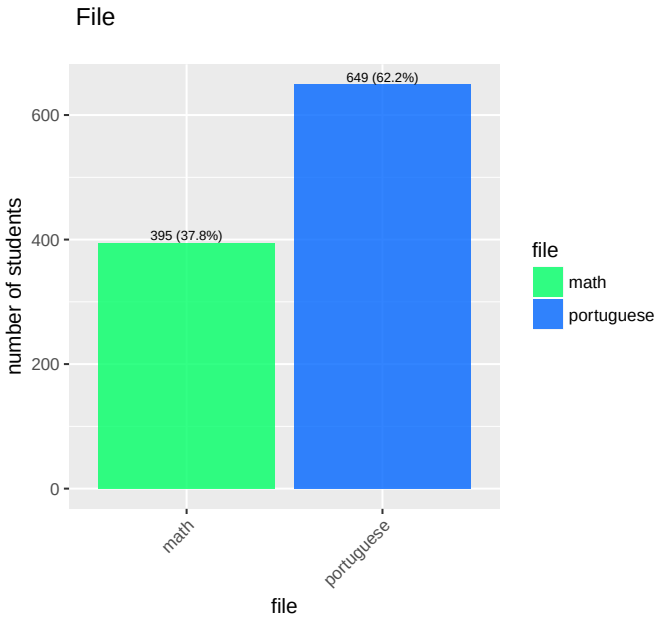

Figura 2.40: File

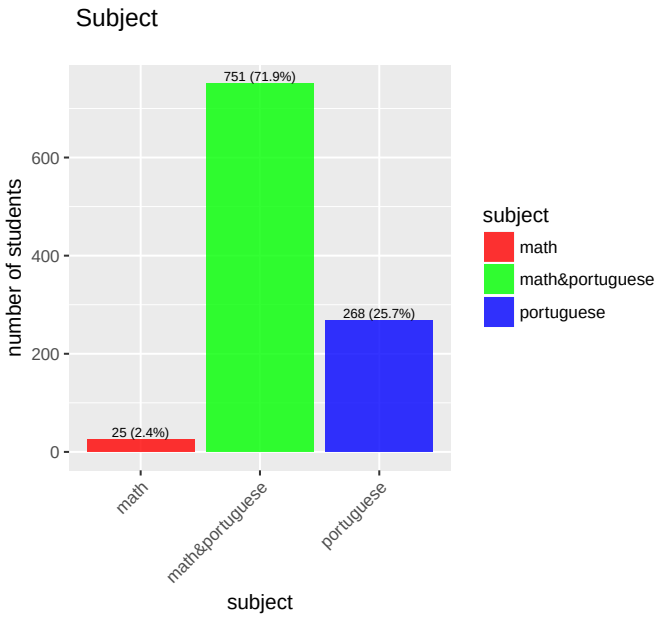

Figura 2.41: Subject

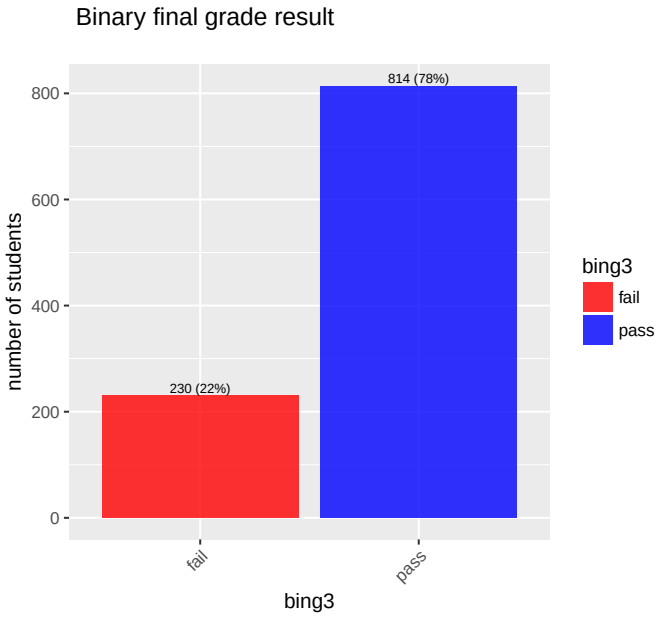

Figura 2.42: Binary final grade result

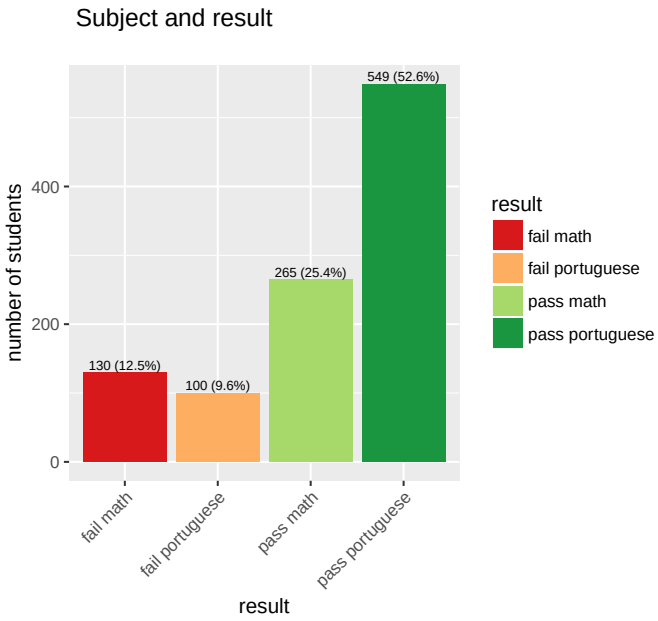

Figura 2.43: Subject and result

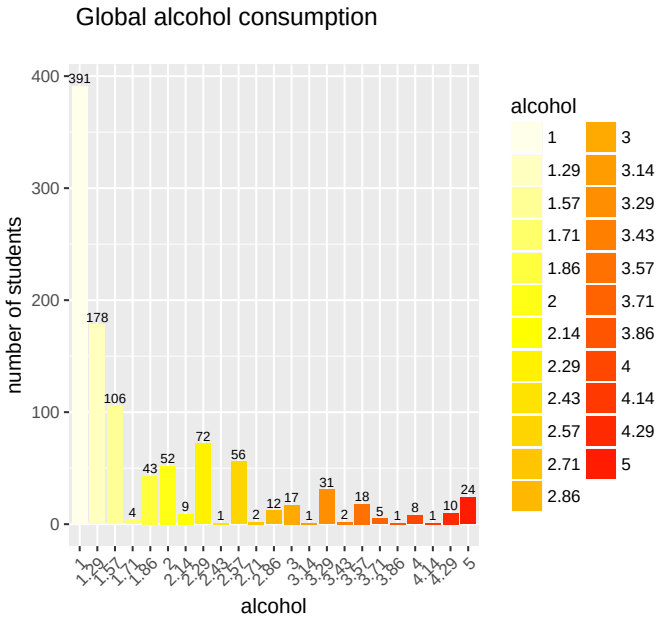

Figura 2.44: Global alcohol consumption

Capítulo 3

Análisis

Una vez preparados los datos, en este capítulo se inician las tareas de análisis. Se recomienda revisar el apartado 'Descripción de los ficheros originales' 2.2 para contextualizar las variables que participan en los diferentes análisis.

Para empezar se realiza un análisis univariante y bivariante para acabar de entender los datos y detectar si existen relaciones entre pares de variables.

La segmentación con Kmeans permitirá agrupar los estudiantes con características parecidas en el mismo grupo y formar grupos lo más diferentes posibles entre sí. Esto permitirá ver si el consumo de alcohol manifiesta comportamientos distintos en los diferentes grupos de estudiantes o por el contrario es parecido.

Mediante la detección de reglas de asociación se pretender estudiar si el consumo de alcohol influye o no en los resultados académicos.

Se realiza un profiling (descripción de características) con la finalidad de detectar las variables que más influyen sobre la variable éxito (superar o no la asignatura). Esto permitirá extraer conclusiones sobre si el consumo de alcohol es la variable más influyente en los resultados académicos o por el contrario existen otras variables que lo son más.

Finalmente, se entrenarán árboles de decisión para predecir el resultado académico. Esto permitirá ver si el consumo de alcohol es una variable influyente en la decisión, es decir, si aparece en los nodos más superiores del árbol o no.

3.1. Análisis univariante y bivariante

El primer paso en el análisis será ver el summary de los datos una vez configurada cada variable con el tipo que le corresponde y añadidas las variables calculadas.

```
                school        sex           age          address
 Gabriel Pereira     :772   female:591   Min.   :15.00   rural:285
 Mousinho da Silveira:272   male  :453   1st Qu.:16.00   urban:759
                                         Median :17.00
                                         Mean   :16.73
                                         3rd Qu.:18.00
```

```
                                      Max.   :22.00
             famsize               pstatus              medu                fedu
greater than 3    :738   living apart   :121   none         :  9   none         :  9
less or equal to 3:306   living together:923   primary      :202   primary      :256
                                               5th-9th grade:289   5th-9th grade:324
                                               secondary    :238   secondary    :231
                                               higher       :306   higher       :224

      mjob           fjob            reason      guardian              traveltime
at_home :194   at_home : 62   course    :430   father:243   <15 min.        :623
health  : 82   health  : 41   home      :258   mother:728   15 to 30 min.   :320
other   :399   other   :584   other     :108   other : 73   30 min to 1 hour: 77
services:239   services:292   reputation:248                >1 hour         : 24
teacher :130   teacher : 65

        studytime      failures      schoolsup famsup     paid     activities nursery
<2 hours     :317   Min.   :0.0000   no :925   no :404   no :824   no :528    no :209
2 to 5 hours :503   1st Qu.:0.0000   yes:119   yes:640   yes:220   yes:516    yes:835
5 to 10 hours:162   Median :0.0000
>10 hours    : 62   Mean   :0.2644
                    3rd Qu.:0.0000
                    Max.   :3.0000
higher    internet  romantic        famrel         freetime         goout
no : 89   no :217   no :673   very bad : 30   very low : 64   very low : 71
yes:955   yes:827   yes:371   bad      : 47   low      :171   low      :248
                              fair     :169   medium   :408   medium   :335
                              good     :512   high     :293   high     :227
                              excellent:286   very high:108   very high:163

       dalc            walc           health       absences            g1
very low :727   very low :398   very bad :137   Min.   : 0.000   Min.   : 0.00
low      :196   low      :235   bad      :123   1st Qu.: 0.000   1st Qu.: 9.00
medium   : 69   medium   :200   fair     :215   Median : 2.000   Median :11.00
high     : 26   high     :138   good     :174   Mean   : 4.435   Mean   :11.21
very high: 26   very high: 73   very good:395   3rd Qu.: 6.000   3rd Qu.:13.00
                                               Max.   :75.000   Max.   :19.00
      g2              g3               file                subject
Min.   : 0.00   Min.   : 0.00   math      :395   math           : 25
1st Qu.: 9.00   1st Qu.:10.00   portuguese:649   math&portuguese:751
Median :11.00   Median :11.00                    portuguese     :268
Mean   :11.25   Mean   :11.34
3rd Qu.:13.00   3rd Qu.:14.00
Max.   :19.00   Max.   :20.00
student_key         bing3                result       alcohol
Length:1044        fail:230   fail math      :130   Min.   :1.000
Class :character   pass:814   fail portuguese:100   1st Qu.:1.000
Mode  :character              pass math      :265   Median :1.290
                              pass portuguese:549   Mean   :1.721
                                                    3rd Qu.:2.178
                                                    Max.   :5.000
```

Para las variables categóricas se muestra para cada valor posible el número de observaciones. Para las variables numéricas se muestran los estadísticos: Min: mínimo, 1st Qu.: primer cuartil, Median: mediana (o segundo cuartil), Mean: media, 3rd Qu.: tercer cuartil y Max.: máximo.

Para completar el análisis univariante se calcula la desviación típica de cada una de las variables numéricas.

age	failures	absences	g1	g2	g3	alcohol
1.24	0.66	6.21	2.98	3.29	3.86	0.93

Tabla 3.1: Desviación típica de las variables numéricas

Tal como se había podido observar con los gráficos y con el análisis de outliers, la variable número de ausencias (absences) es la que mayor dispersión presenta.

Se inicia el análisis bivariante con el cálculo de la matriz de correlaciones.

	age	failures	absences	g1	g2	g3	alcohol
age	1.00	0.28	0.15	-0.12	-0.12	-0.13	0.13
failures	0.28	1.00	0.10	-0.37	-0.38	-0.38	0.12
absences	0.15	0.10	1.00	-0.09	-0.09	-0.05	0.15
g1	-0.12	-0.37	-0.09	1.00	0.86	0.81	-0.16
g2	-0.12	-0.38	-0.09	0.86	1.00	0.91	-0.14
g3	-0.13	-0.38	-0.05	0.81	0.91	1.00	-0.14
alcohol	0.13	0.12	0.15	-0.16	-0.14	-0.14	1.00

Tabla 3.2: Matriz de correlaciones

Se observan altas correlaciones entre las variables g1, g2 y g3. Son las notas obtenidas en el primer periodo en el segundo periodo y la nota final. Tiene sentido que tengan correlación positiva y fuerte. El coeficiente de correlación de las variables g1 y g2 es 0,86. El coeficiente de correlación de las variables g2 y g3 es 0,91. El coeficiente de correlación de las variables g1 y g3 es 0,81.

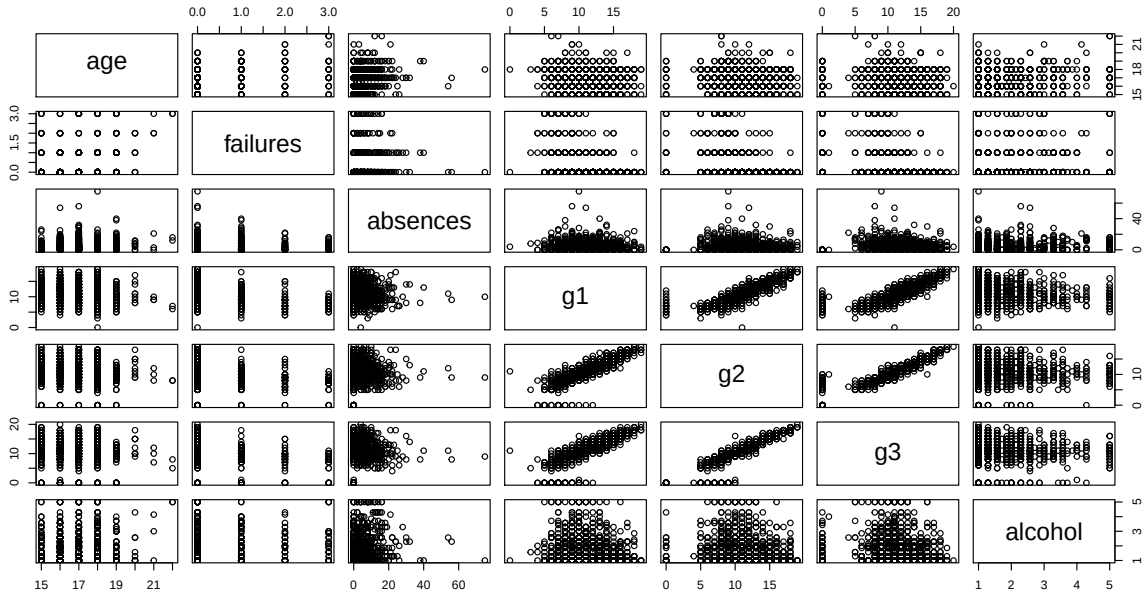

Figura 3.1: Matriz de diagramas de dispersión

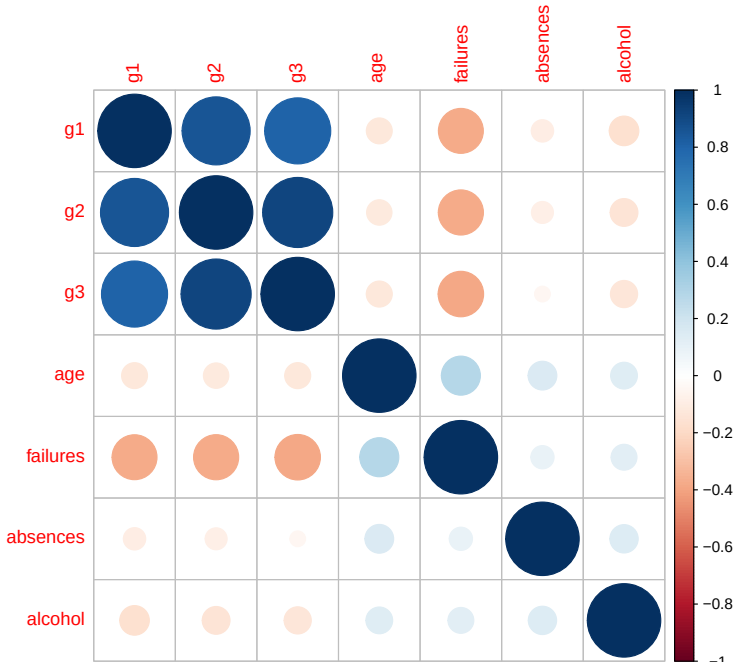

Figura 3.2: Matriz de correlaciones

En el resto de variables numéricas no se observan relaciones ni a través del cálculo de correlaciones ni a través de los diagramas de dispersión.

A continuación se muestran algunos diagramas de dispersión que podrían dar información sobre relaciones entre variables categóricas y numéricas o entre pares de variables categóricas. Para poder visualizar correctamente los datos y que varios puntos no queden solapados se usa la opción 'jitter' en los gráficos elaborados con ggplot2. Esta opción añade cierta dispersión a los datos para evitar el solapamiento y ayuda a ver las agrupaciones de puntos alrededor de un valor.

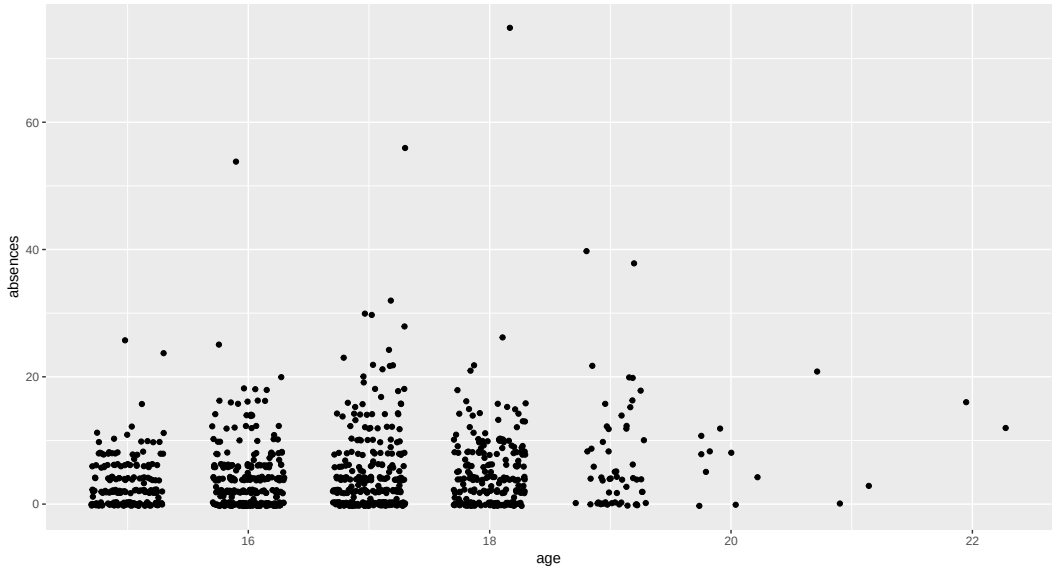

Figura 3.3: Edad (age) vs. número de ausencias (absences)

La media del número de ausencias aumenta a medida que aumenta la edad de los estudiantes:

15	16	17	18	19	20	21	22
3.10	3.88	4.92	4.87	6.88	6.22	8.00	14.00

Tabla 3.3: Media de ausencias por edad

En el gráfico 'Age vs. absences' se observa que a medida que aumenta la edad cada vez hay menos observaciones. Eso hace que la media sea mayor con las edades más altas ya que en algunos casos tienen ausencias elevadas.

...

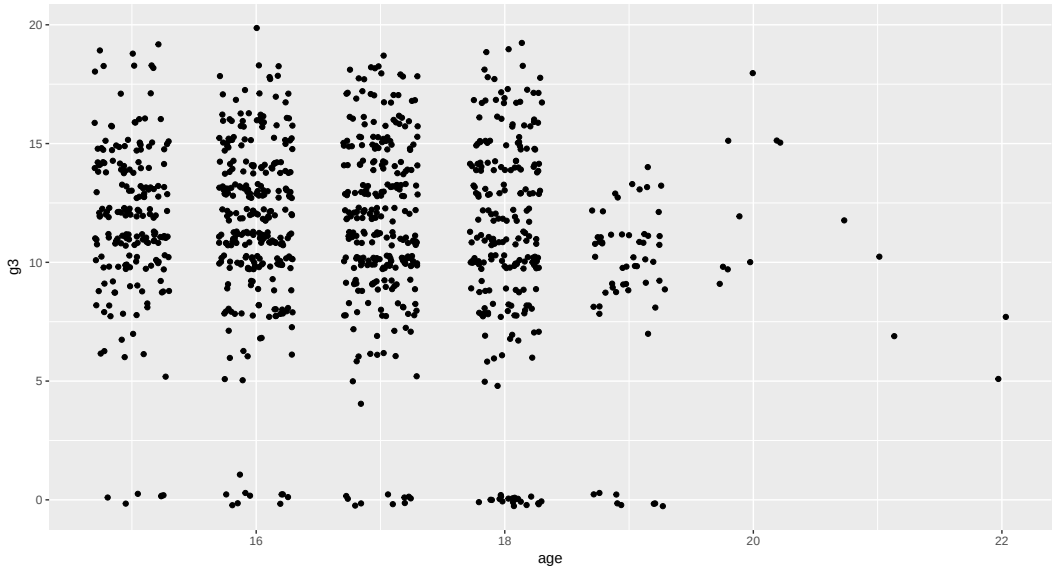

Figura 3.4: Edad (age) vs. nota final (g3)

La nota media final (de 0 a 20) está por debajo de aprobado (menos de 10) en las edades 19, 21 y 22. En el gráfico 'Edad (age) vs. nota final (g3)' que hay pocas observaciones a partir de 19 años por lo que la media es muy sensible a valores concretos.

15	16	17	18	19	20	21	22
11.75	11.64	11.56	10.95	8.96	12.67	9.67	6.50

Tabla 3.4: Media de notas finales por edad

...

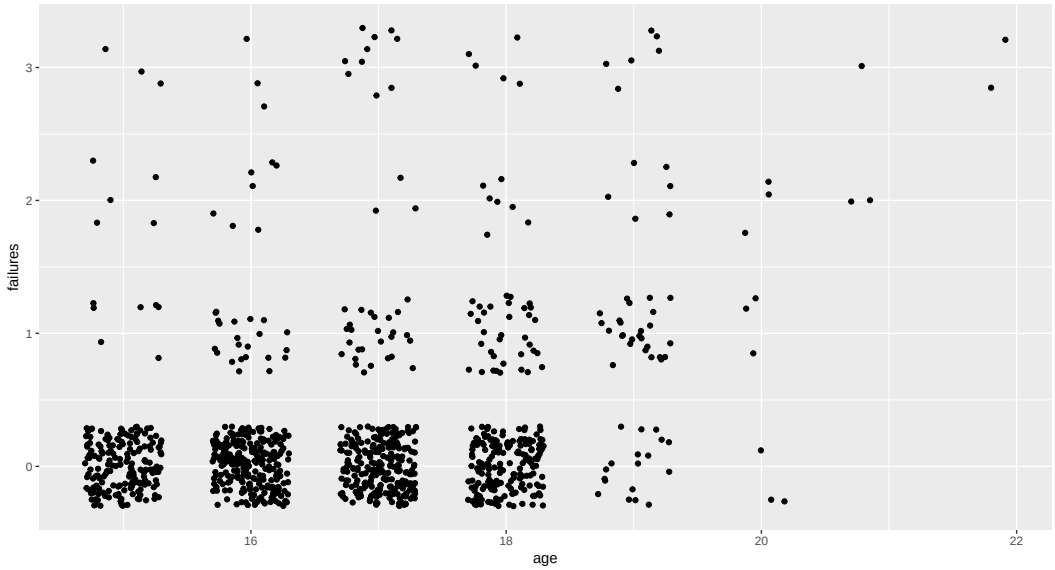

Figura 3.5: Edad (age) vs. suspensos previos (failures)

El número medio de suspensos es prácticamente 0 de los 15 a los 18 años, a partir de esa edad, a medida que aumenta la edad también aumenta el número de suspensos.

15	16	17	18	19	20	21	22
0.13	0.16	0.23	0.29	1.00	1.00	2.33	3.00

Tabla 3.5: Media de suspensos anteriores por edad

...

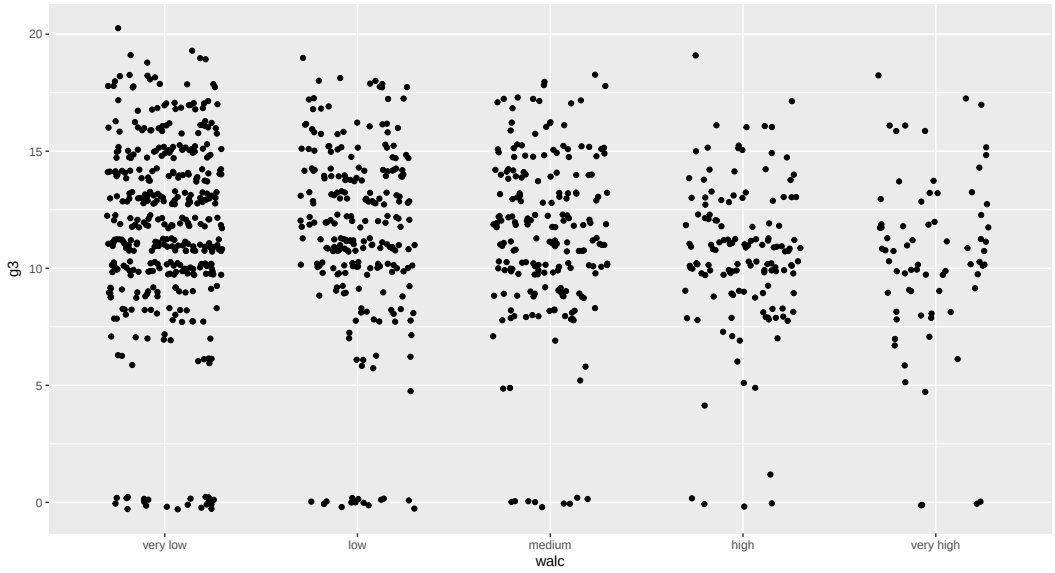

Figura 3.6: Consumo de alcohol en fin de semana (walc) vs. nota final (g3)

La nota media final desciende ligeramente a medida que el consumo de alcohol en fin de semana es más elevado. En el gráfico 'Consumo de alcohol en fin de semana (walc) vs. nota final (g3)' no se observa ningún patrón.

very low	low	medium	high	very high
11.74	11.47	11.29	10.54	10.40

Tabla 3.6: Notas medias finales según el consumo de alcohol en fin de semana

...

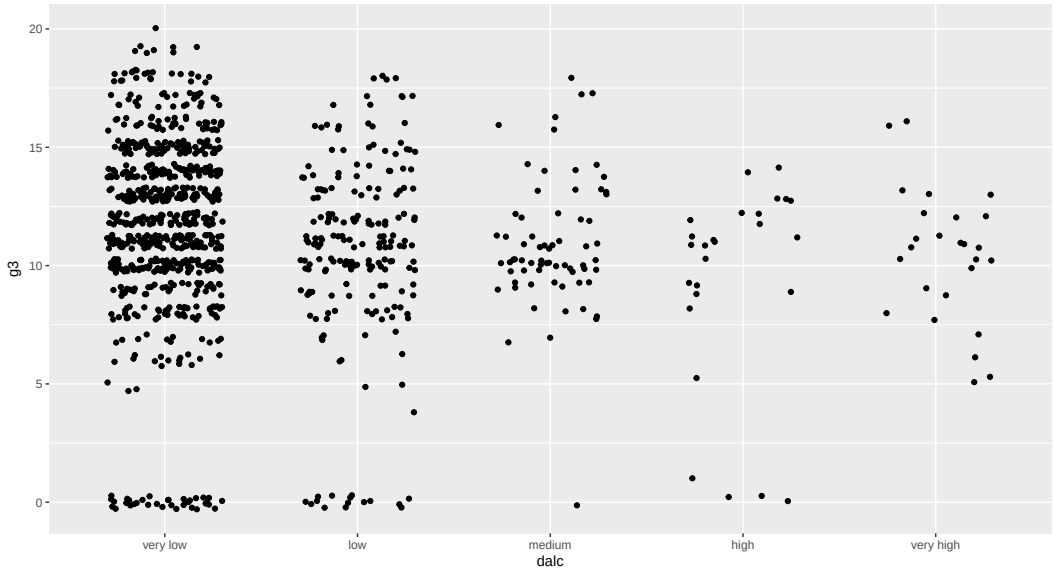

Figura 3.7: Consumo de alcohol en días laborables (dalc) vs. nota final (g3)

Desde el punto de vista de consumo de alcohol en días laborables la nota media es fluctuante, no presenta una tendencia tan claramente descendiente como en el caso del consumo en fin de semana. En el gráfico 'Consumo de alcohol en días laborables (dalc) vs. nota final (g3)' se observa que los consumos elevados de alcohol entre semana son menos frecuentes que en fin de semana.

very low	low	medium	high	very high
11.70	10.56	10.90	9.27	10.38

Tabla 3.7: Notas medias finales según el consumo de alcohol entre semana

...

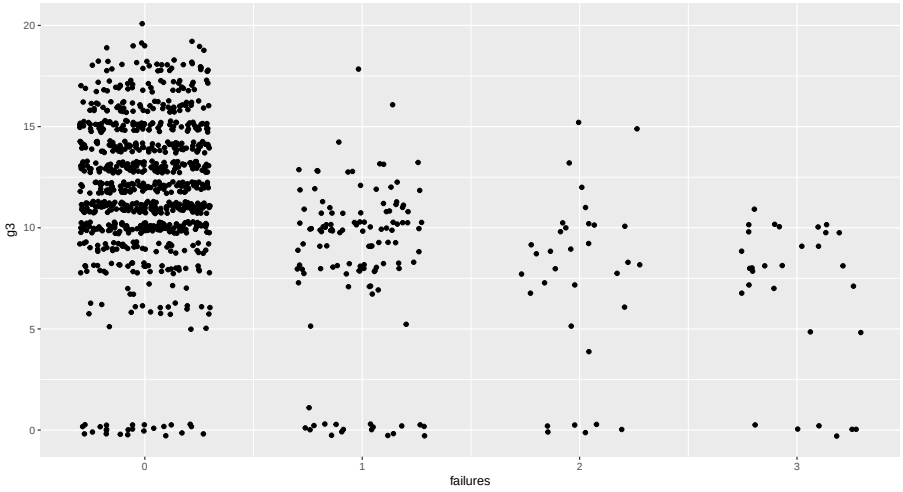

Figura 3.8: Suspensos previos (failures) vs. nota final (g3)

La nota media final no supera el aprobado cuando hay suspensos previos. En el gráfico 'Suspensos previos (failures) vs. nota final (g3)' se puede ver que hay pocas observaciones por encima de 10 cuando failures es mayor o igual a 1.

0	1	2	3
12.05	8.43	7.48	6.80

Tabla 3.8: Notas medias finales según el número de suspensos anteriores

3.2. Segmentación

Para la segmentación se usan únicamente las variables numéricas. Se reserva la variable construida 'alcohol' para analizar si existen diferencias en los grupos resultantes de la segmentación. La variables númericas son: 'age', 'failures', 'absences', 'g1', 'g2' y 'g3'.

Dado que K-means se basa en la distancia euclídea, es muy importante normalizar los datos y eliminar los valores atípicos, por ello, en esta caso se usa el dataset en el que se eliminaron los valores atípicos de la variable 'absences' y cada una de las variables se normaliza restando la media y dividiendo por la desviación típica. A la hora de interpretar los clusters habrá que tener esto en cuenta.

Para elegir el número de clusters más apropiado se busca el valor de k que haga que los individuos pertenecientes a un mismo grupo sean lo más homogéneos posible y los individuos pertenecientes a distintos grupos sean lo más heterogéneos posible teniendo en cuenta que cuanto más grande sea k más cálculos debe hacer el algoritmo.

Los parámetros utilizados son las sumas de cuadrados dentro de los grupos (tot.withinss) y la suma de cuadrados entre grupos (betweenss).

El objetivo es maximizar la suma de cuadrados entre grupos y minimizar la suma de cuadrados dentro de los grupos. Para ello se ejecuta el algoritmo K-means para diferentes valores de k entre 2 y 10 y se visualizan en un gráfico las dos métricas.

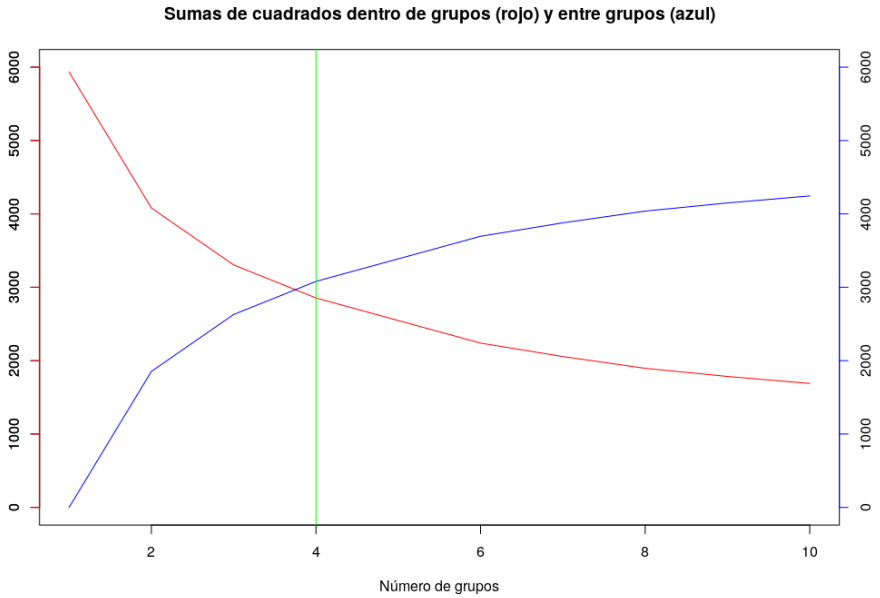

Figura 3.9: Selección del parámetro k para K-means

Para k=4 se consigue un punto de equilibrio entre el número de grupos y el objetivo de minimizar la suma de cuadrados dentro de grupos y maximizar la suma de cuadrados entre grupos. A partir de ese valor de k, la disminución y el aumento respectivamente son muy inferiores a los conseguidos para los valores anteriores.

Si se visualizan los centroides directamente, al estar los datos normalizados, no son fácilmente interpretables. Para solucionarlo se calculan los centroides sin normalizar haciendo la media de cada una de las variables agrupadas por el centroide: 'age' es la edad del estudiante, 'failures' es el número de suspensos previos, 'absences' es el número de ausencias, 'g1' es la nota del primer periodo, 'g2' es la nota del segundo periodo y 'g3' es la nota final. El sufijo "_center", indica que se ha calculado la media de esa variable para determinar el centroide.

cluster	age_center	failures_center	absences_center	g1_center	g2_center	g3_center	n
1	17.45	1.40	2.26	7.52	6.50	4.92	128
2	16.59	0.03	2.42	14.56	14.75	15.11	314
3	16.18	0.05	2.09	10.28	10.54	10.85	378
4	17.48	0.25	9.14	10.09	10.26	10.64	170

Tabla 3.9: Centroides

A continuación se describen los distintos grupos obtenidos en base a las características de los centroides:

1. Cluster 1 (rojo): son alumnos de unos 17 años y medio, con notas bajas (por debajo del aprobado), con más de un suspenso previo y 2 ausencias de media. En este grupo hay 128 alumnos.
2. Cluster 2 (azul): son alumnos de unos 16 años y medio, con notas altas (15 puntos de media), sin suspensos previos y 2 ausencias de media. En este grupo hay 314 alumnos.
3. Cluster 3 (verde): son alumnos de unos 16 años, con notas medias de aprobado, sin suspensos previos y 2 ausencias de media. En este grupo hay 378 alumnos.
4. Cluster 4 (lila): son alumnos de unos 17 años y medio, con notas medias de aprobado, sin suspensos previos y un nivel de ausencias elevado (unas 9 ausencias de media). En este grupo hay 170 alumnos.

Se han realizado varias visualizaciones tipo scatterplot cogiendo varios pares de variables y usando el color para diferenciar los clusters. En general no se consigue una visualización en la que los grupos del mismo cluster se vean claramente diferenciados.

cluster	alcohol_avg	n
1	1.85	128
2	1.47	314
3	1.70	378
4	2.02	170

Tabla 3.10: Consumo de alcohol por cluster

Calculando la media 'alcohol_avg' de la variable 'alcohol' en los diferentes grupos, se observa que el grupo 4 es el que presenta un valor medio más elevado. Este grupo se caracteriza por un elevado número de ausencias. El segundo grupo con un valor mayor de consumo de alcohol es el cluster 1 que se caracteriza por notas muy bajas por debajo de aprobado. En los grupos 2 y 3 el consumo medio de alcohol es parecido, ligeramente inferior en el cluster 2 que es el que tiene notas medias más altas.

3.3. Reglas de asociación

A través de esta técnica de aprendizaje automático no supervisado se buscan relaciones entre las variables y los valores que toman del estilo "cuando pasa A también pasa B". Este tipo de reglas no denotan una relación de causa-efecto, detectan situaciones que se dan a la vez pero no necesariamente una es la causa de la otra. Interesará detectar relaciones en las que en A aparezcan variables relacionadas con el consumo de alcohol y donde en B aparezca la variable de éxito (superar o no la asignatura).

Una regla de asociación está formada por uno o más antecedentes y una consecuencia y está caracterizada por su soporte (porcentaje de casos en los que se dan los antecedentes conjuntamente) y su confianza (porcentaje de casos en los que se da la consecuencia junto con los antecedentes respecto a las veces que se dan los antecedentes conjuntamente). El parámetro lift o apalancamiento es el cociente entre la confianza de la regla y el soporte de la consecuencia, este parámetro permite valorar si dicha consecuencia tiene más probabilidad cuando se da el antecedente o en general.

Para empezar se seleccionan las variables que van a participar en el análisis que son todas las variables categóricas: 'school', 'sex', 'address', 'famsize', 'pstatus', 'medu', 'fedu', 'mjob', 'fjob', 'reason', 'guardian', 'traveltime', 'studytime', 'schoolsup', 'famsup', 'paid', 'activities', 'nursery', 'higher', 'internet', 'romantic', 'famrel', 'freetime', 'goout', 'dalc', 'walc', 'health' y 'bing3'.

Se utiliza la función 'apriori' de la librería 'arules'. En una primera ejecución sin ninguna restricción se obtienen 456.831 reglas. Por defecto, el algoritmo apriori busca reglas de asociación con un soporte del 10 % o superior y una confianza del 80 % o superior.

Aplicando una restricción para filtrar las reglas que tengan un soporte mínimo del 30 % y una confianza superior al 95 % se obtienen 537 reglas pero todas ellas tienen como consecuencia higher=yes. Esto se debe a que en el 91,5 % de los casos la variable higher presenta el valor sí.

Filtrando las reglas que contienen en la consecuencia la variable bing3, con soporte mínimo del 25 % y confianza mínima del 90 % se obtienen 18 reglas. Left-hand-side (lhs) es el antecedente de la regla y right-hand-side (rhs) es la consecuencia.

```
     lhs                          rhs              support confidence     lift
[1]  {school=Gabriel Pereira,
      address=urban,
      guardian=mother,
      paid=no,
      higher=yes}              => {bin_g3=pass} 0.2595785  0.9063545 1.162450
[2]  {address=urban,
      guardian=mother,
      paid=no,
      higher=yes,
      internet=yes}            => {bin_g3=pass} 0.2557471  0.9050847 1.160821
[3]  {school=Gabriel Pereira,
      address=urban,
      paid=no,
      higher=yes,
      internet=yes}            => {bin_g3=pass} 0.3247126  0.9015957 1.156346
[4]  {school=Gabriel Pereira,
      address=urban,
      paid=no,
```

```
      nursery=yes,
      higher=yes}              => {bin_g3=pass} 0.3026820  0.9028571 1.157964
[5]  {school=Gabriel Pereira,
      address=urban,
      schoolsup=no,
      paid=no,
      higher=yes}              => {bin_g3=pass} 0.3218391  0.9008043 1.155331
[6]  {address=urban,
      paid=no,
      nursery=yes,
      higher=yes,
      internet=yes}            => {bin_g3=pass} 0.2988506  0.9017341 1.156524
[7]  {address=urban,
      schoolsup=no,
      paid=no,
      higher=yes,
      internet=yes}            => {bin_g3=pass} 0.3304598  0.9007833 1.155304
[8]  {school=Gabriel Pereira,
      address=urban,
      schoolsup=no,
      nursery=yes,
      higher=yes,
      dalc=very low}           => {bin_g3=pass} 0.2681992  0.9032258 1.158437
[9]  {address=urban,
      schoolsup=no,
      paid=no,
      higher=yes,
      internet=yes,
      dalc=very low}           => {bin_g3=pass} 0.2538314  0.9106529 1.167963
[10] {address=urban,
      schoolsup=no,
      nursery=yes,
      higher=yes,
      internet=yes,
      dalc=very low}           => {bin_g3=pass} 0.2739464  0.9022082 1.157132
[11] {school=Gabriel Pereira,
      address=urban,
      paid=no,
      nursery=yes,
      higher=yes,
      internet=yes}            => {bin_g3=pass} 0.2624521  0.9194631 1.179262
[12] {school=Gabriel Pereira,
      address=urban,
      schoolsup=no,
      paid=no,
      higher=yes,
      internet=yes}            => {bin_g3=pass} 0.2863985  0.9171779 1.176331
[13] {school=Gabriel Pereira,
      address=urban,
      pstatus=living together,
      paid=no,
      nursery=yes,
```

```
      higher=yes}              => {bin_g3=pass} 0.2557471  0.9020270 1.156900
[14] {school=Gabriel Pereira,
      address=urban,
      schoolsup=no,
      paid=no,
      nursery=yes,
      higher=yes}              => {bin_g3=pass} 0.2614943  0.9222973 1.182897
[15] {school=Gabriel Pereira,
      address=urban,
      pstatus=living together,
      schoolsup=no,
      paid=no,
      higher=yes}              => {bin_g3=pass} 0.2777778  0.9062500 1.162316
[16] {address=urban,
      schoolsup=no,
      paid=no,
      nursery=yes,
      higher=yes,
      internet=yes}            => {bin_g3=pass} 0.2653257  0.9172185 1.176383
[17] {address=urban,
      pstatus=living together,
      schoolsup=no,
      paid=no,
      higher=yes,
      internet=yes}            => {bin_g3=pass} 0.2873563  0.9009009 1.155455
[18] {school=Gabriel Pereira,
      schoolsup=no,
      paid=no,
      nursery=yes,
      higher=yes,
      internet=yes}            => {bin_g3=pass} 0.2662835  0.9055375 1.161402
```

Las reglas [8], [9] y [10] son las únicas que contienen en el antecedente alguna de las variables sobre el consumo de alcohol, en los tres casos la consecuencia es que se aprueba la asignatura. En los tres casos el valor de alcohol consumido es muy bajo. Ninguna de las reglas tiene en la consecuencia que se suspende la asignatura por lo que no se puede concluir nada sobre los suspensos.

3.4. Profiling

El Profiling es un técnica para la detección de patrones. Permite reconocer de forma automática qué características inciden más que otras en la diferenciación de los individuos de dos grupos dicotómicos.

En este caso, el uso que se hace de esta técnica es separar los alumnos entre los que aprueban y los que suspenden y ver qué diferencias tienen. En particular se analizará si el consumo de alcohol es una de características diferenciadoras de los dos grupos de alumnos.

La variable éxito/fracaso se define en base a la variable 'g3' (nota final) del siguiente modo:

$$nbing3 = \begin{cases} 1 & \text{si } g3 >= 10 \\ 0 & \text{si } g3 < 10 \end{cases}$$

Para este análisis se excluyen las variables g1, g2, g3, file, subject, student_key, result y bing3 ya que o bien están directamente relacionadas con la variable éxito/fracaso o bien porque son campos de identificación del individuo.

Como paso previo se ha intentado visualizar algunos gráficos para ver si se pueden detectar características que diferencien los dos grupos, pero no hay ninguno que muestre diferencias entre los dos grupos de estudiantes.

Para realizar el Profiling se utiliza la función 'catdes' del paquete 'FactoMiner', como parámetro 'prob' (nivel de significación) se utiliza el 5 %.

```
$test.chi2
               p.value df
higher    1.693028e-12  1
school    1.987478e-05  1
goout     2.761440e-04  4
fedu      4.500810e-04  4
medu      1.292969e-03  4
studytime 4.155114e-03  3
guardian  6.064421e-03  2
romantic  1.115311e-02  1
dalc      1.469728e-02  4
reason    1.686899e-02  3
address   1.720238e-02  1
paid      3.471492e-02  1

$category
$category$`0`
                             Cla/Mod  Mod/Cla    Global      p.value    v.test
higher=no                   51.68539 20.00000  8.524904 1.278896e-10  6.429660
school=Mousinho da Silveira 31.25000 36.95652 26.053640 3.313029e-05  4.150809
fedu=primary                31.25000 34.78261 24.521073 6.872376e-05  3.980661
goout=very high             32.51534 23.04348 15.613027 7.204126e-04  3.381691
medu=primary                30.69307 26.95652 19.348659 1.339017e-03  3.207484
studytime=<2 hours          27.44479 37.82609 30.363985 6.054407e-03  2.744820
romantic=yes                26.41509 42.60870 35.536398 1.202256e-02  2.511482
guardian=other              34.24658 10.86957  6.992337 1.299985e-02  2.483773
reason=course               25.81395 48.26087 41.187739 1.421787e-02  2.451711
address=rural               27.01754 33.47826 27.298851 1.898112e-02  2.345902
dalc=low                    28.06122 23.91304 18.773946 2.699476e-02  2.211594
paid=yes                    27.27273 26.08696 21.072797 3.810241e-02  2.073751
paid=no                     20.63107 73.91304 78.927203 3.810241e-02 -2.073751
fedu=higher                 16.96429 16.52174 21.455939 3.647090e-02 -2.091638
freetime=low                15.78947 11.73913 16.379310 2.810519e-02 -2.195815
famrel=good                 19.14062 42.60870 49.042146 2.730465e-02 -2.207135
```

```
goout=medium                17.91045 26.08696 32.088123 2.611151e-02 -2.224549
guardian=father             16.87243 17.82609 23.275862 2.473245e-02 -2.245556
address=urban               20.15810 66.52174 72.701149 1.898112e-02 -2.345902
romantic=no                 19.61367 57.39130 64.463602 1.202256e-02 -2.511482
goout=low                   16.12903 17.39130 23.754789 8.938933e-03 -2.614381
studytime=5 to 10 hours     14.19753 10.00000 15.517241 6.993573e-03 -2.697150
reason=reputation           15.72581 16.95652 23.754789 5.120815e-03 -2.799334
dalc=very low               19.39477 61.30435 69.636015 2.213793e-03 -3.059943
medu=higher                 15.35948 20.43478 29.310345 6.329086e-04 -3.417105
school=Gabriel Pereira      18.78238 63.04348 73.946360 3.313029e-05 -4.150809
higher=yes                  19.26702 80.00000 91.475096 1.278896e-10 -6.429660

$category$`1`
                             Cla/Mod   Mod/Cla    Global      p.value    v.test
higher=yes                  80.73298 94.717445 91.475096 1.278896e-10  6.429660
school=Gabriel Pereira      81.21762 77.027027 73.946360 3.313029e-05  4.150809
medu=higher                 84.64052 31.818182 29.310345 6.329086e-04  3.417105
dalc=very low               80.60523 71.990172 69.636015 2.213793e-03  3.059943
reason=reputation           84.27419 25.675676 23.754789 5.120815e-03  2.799334
studytime=5 to 10 hours     85.80247 17.076167 15.517241 6.993573e-03  2.697150
goout=low                   83.87097 25.552826 23.754789 8.938933e-03  2.614381
romantic=no                 80.38633 66.461916 64.463602 1.202256e-02  2.511482
address=urban               79.84190 74.447174 72.701149 1.898112e-02  2.345902
guardian=father             83.12757 24.815725 23.275862 2.473245e-02  2.245556
goout=medium                82.08955 33.783784 32.088123 2.611151e-02  2.224549
famrel=good                 80.85938 50.859951 49.042146 2.730465e-02  2.207135
freetime=low                84.21053 17.690418 16.379310 2.810519e-02  2.195815
fedu=higher                 83.03571 22.850123 21.455939 3.647090e-02  2.091638
paid=no                     79.36893 80.343980 78.927203 3.810241e-02  2.073751
paid=yes                    72.72727 19.656020 21.072797 3.810241e-02 -2.073751
dalc=low                    71.93878 17.321867 18.773946 2.699476e-02 -2.211594
address=rural               72.98246 25.552826 27.298851 1.898112e-02 -2.345902
reason=course               74.18605 39.189189 41.187739 1.421787e-02 -2.451711
guardian=other              65.75342  5.896806  6.992337 1.299985e-02 -2.483773
romantic=yes                73.58491 33.538084 35.536398 1.202256e-02 -2.511482
studytime=<2 hours          72.55521 28.255528 30.363985 6.054407e-03 -2.744820
medu=primary                69.30693 17.199017 19.348659 1.339017e-03 -3.207484
goout=very high             67.48466 13.513514 15.613027 7.204126e-04 -3.381691
fedu=primary                68.75000 21.621622 24.521073 6.872376e-05 -3.980661
school=Mousinho da Silveira 68.75000 22.972973 26.053640 3.313029e-05 -4.150809
higher=no                   48.31461  5.282555  8.524904 1.278896e-10 -6.429660


$quanti.var
                Eta2      P-value
failures 0.134823525 1.140091e-34
age      0.018030000 1.342167e-05
absences 0.014194706 1.139821e-04
alcohol  0.008673156 2.594825e-03

$quanti
$quanti$`0`
```

```
           v.test Mean in category Overall mean sd in category Overall sd      p.value
failures 11.858370        0.7173913    0.2643678      0.9705006  0.6558275 1.947294e-32
age       4.336507       17.0391304   16.7260536      1.3329402  1.2393807 1.447650e-05
absences  3.847737        5.8260870    4.4348659      8.7000858  6.2070417 1.192141e-04
alcohol   3.007674        1.8833913    1.7210632      1.0154608  0.9265243 2.632555e-03

$quanti$`1`
             v.test Mean in category Overall mean sd in category Overall sd      p.value
alcohol   -3.007674        1.6751966    1.7210632      0.8944814  0.9265243 2.632555e-03
absences  -3.847737        4.0417690    4.4348659      5.2273355  6.2070417 1.192141e-04
age       -4.336507       16.6375921   16.7260536      1.1968887  1.2393807 1.447650e-05
failures -11.858370        0.1363636    0.2643678      0.4594939  0.6558275 1.947294e-32


attr(,"class")
[1] "catdes" "list "
```

De forma global, las variables más influyentes en el éxito/fracaso, son en orden de más a menos: higher, school, goout, fedu y medu. La única variable relacionada con el alcohol que aparece es dalc y lo hace en la novena posición. Se muestran a continuación algunos gráficos que muestran las diferencias en estas variables sobre los dos grupos.

En los resultados de 'catdes' la métrica 'Global' muestra el porcentaje de población que de forma global presenta esa característica y 'Mod/Cla' muestra el porcentaje de población que presenta esa característica en el grupo.

En el grupo que suspende la proporción de estudiantes que no quiere realizar estudios superiores es mayor (20 %) que en general (8,52 %).

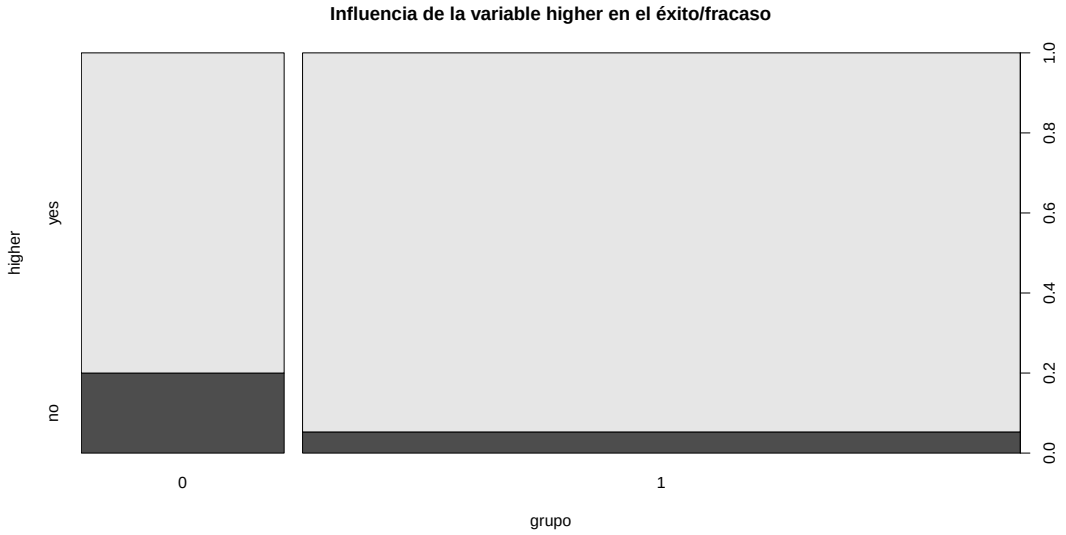

Figura 3.10: Influencia de la intención de continuar con estudios superiores

En el grupo que suspende la proporción de estudiantes del colegio Mousinho da Silveira es mayor (36,95 %) que en general (26,95 %).

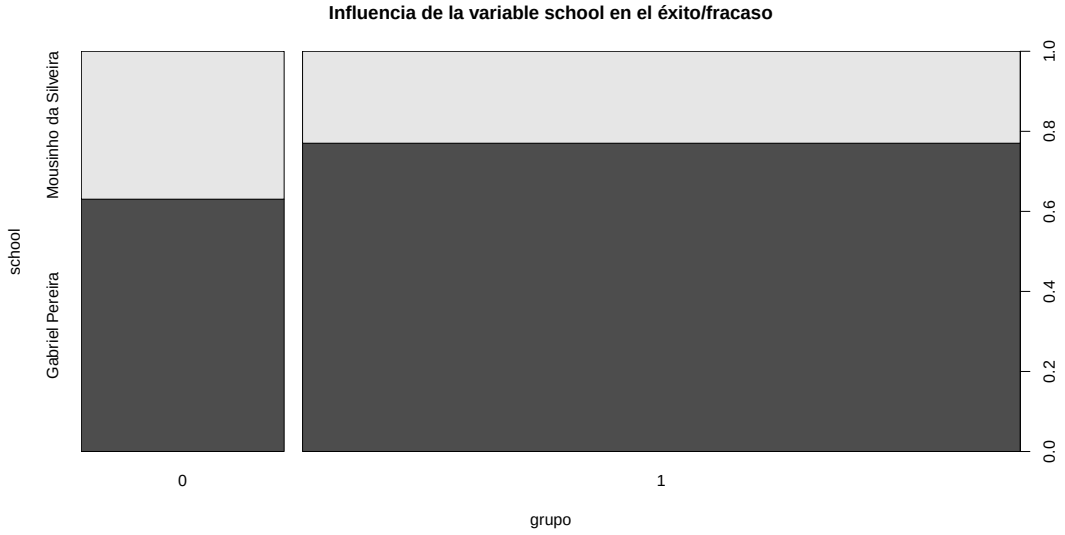

Figura 3.11: Influencia de la escuela

En el grupo que suspende la proporción de estudiantes que sale muy a menudo con los amigos es mayor (23,04 %) que en general (15,61 %).

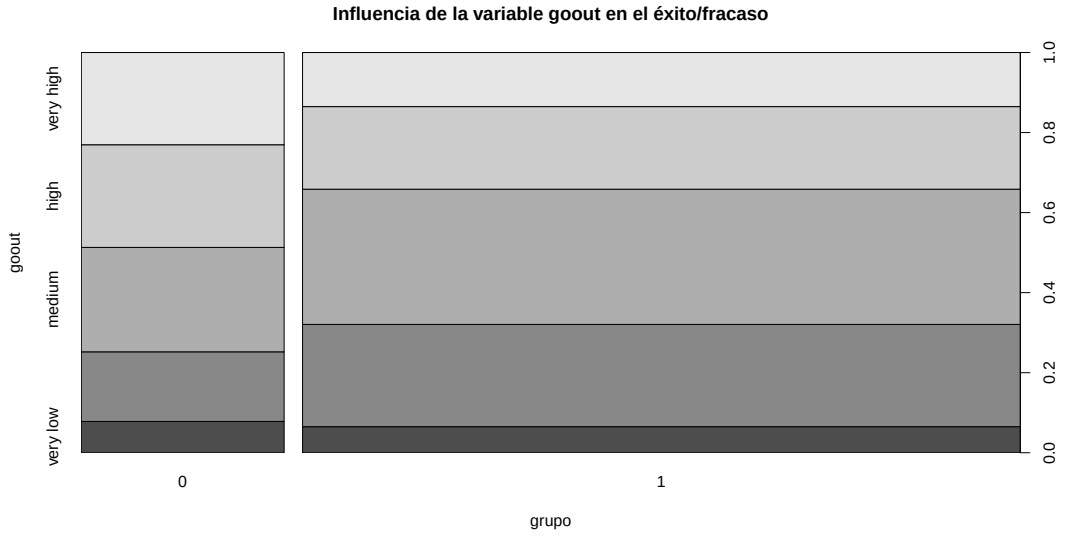

Figura 3.12: Influencia de salir con los amigos

En el grupo que aprueba la proporción de estudiantes cuya madre tiene estudios superiores es mayor (31,81 %) que en general (29,31 %). En el grupo que suspende la proporción de estudiantes cuya madre tiene estudios primarios es mayor (26,95 %) que en general (19,34 %).

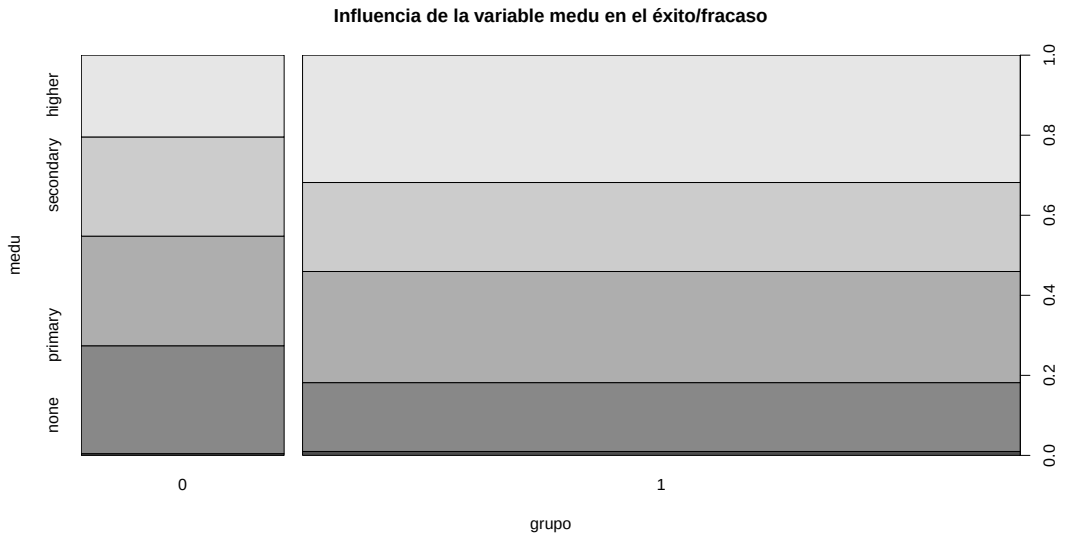

Figura 3.13: Influencia del nivel de estudios de la madre

En el grupo que suspende la proporción de estudiantes cuyo padre tiene estudios primarios es mayor (34,78 %) que en general (24,52 %).

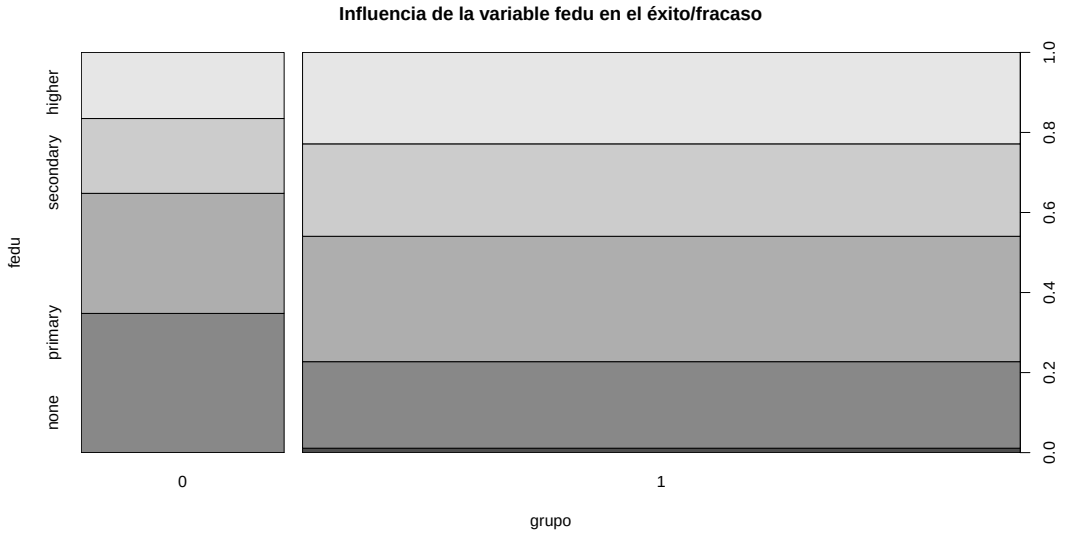

Figura 3.14: Influencia del nivel de estudios del padre

3.5. Árboles de decisión

Usando esta técnica de aprendizaje automático supervisado se pretende entrenar y evaluar un modelo predictivo que alertará cuando un estudiante no va a superar la asignatura, prestando especial atención si uno de los factores influyentes es el consumo de alcohol.

La variable objetivo (variable a predecir) es 'nbing3'. Para este análisis se excluyen las variables g1, g2, g3, file, subject, student key, result y bing3 ya que o bien están directamente relacionadas con la variable objetivo o bien porque son campos de identificación del individuo.

Se prepara el dataset y se divide en dos partes 70 % para entrenar y 30 % para evaluar el modelo. En este caso se utiliza el algoritmo 'C5.0' y se requiere un dataframe para las variables predictoras y otro para la variable objetivo.

Se comprueba que la proporción de las dos clases es similar en los dos grupos:

Grupo	Proporción clase 0	Proporción clase 1
train	0.2208505	0.7791495
test	0.2190476	0.7809524

Tabla 3.11: Proporción de las dos clases en los grupos de entreno y validación

El algoritmo C5.0 muestra el siguiente resultado:

```
Call:
C5.0.default(x = trainX, y = trainY)

C5.0 [Release 2.07 GPL Edition]   Sat Jun 10 10:39:10 2017
-------------------------------

Class specified by attribute `outcome'

Read 729 cases (32 attributes) from undefined.data

Decision tree:

failures <= 0:
:...higher = no:
:   :...nursery = no: 1 (6)
:   :   nursery = yes:
:   :   :...goout in {low,high}: 0 (6)
:   :       goout in {very low,medium,very high}:
:   :       :...guardian in {father,other}: 1 (11)
:   :           guardian = mother:
:   :           :...alcohol <= 1.29: 1 (3)
:   :               alcohol > 1.29: 0 (5)
:   higher = yes:
:   :...paid = no: 1 (436/49)
:       paid = yes:
:       :...guardian = other: 0 (5/1)
```

```
:           guardian in {father,mother}:
:           :...schoolsup = no: 1 (111/21)
:               schoolsup = yes:
:               :...famsup = no: 1 (2)
:                   famsup = yes: 0 (16/7)
failures > 0:
:...famrel = very bad: 0 (3)
    famrel = bad:
    :...health in {very bad,bad,fair,good}: 0 (8/1)
    :   health = very good: 1 (3)
    famrel = excellent:
    :...schoolsup = yes: 1 (2)
    :   schoolsup = no:
    :   :...fjob = at_home: 1 (2)
    :       fjob in {health,other,services,teacher}: 0 (29/7)
    famrel = fair:
    :...reason = home: 1 (5)
    :   reason in {course,other,reputation}:
    :   :...studytime = >10 hours: 0 (0)
    :       studytime = 5 to 10 hours: 1 (2)
    :       studytime in {<2 hours,2 to 5 hours}:
    :       :...health in {very bad,bad,good,very good}: 0 (13/1)
    :           health = fair: 1 (2)
    famrel = good:
    :...freetime in {very low,low,very high}: 1 (18/3)
        freetime = medium:
        :...famsize = greater than 3: 0 (9)
        :   famsize = less or equal to 3: 1 (6/2)
        freetime = high:
        :...sex = female: 1 (5)
            sex = male:
            :...fjob = at_home: 0 (2)
                fjob in {health,services,teacher}: 1 (2)
                fjob = other:
                :...fedu = primary: 1 (11/2)
                    fedu in {none,5th-9th grade,secondary,higher}: 0 (6/1)


Evaluation on training data (729 cases):

    Decision Tree
  ----------------
  Size      Errors

    28   95(13.0%)   <<


   (a)   (b)    <-classified as
  ----  ----
    84    77    (a): class 0
    18   550    (b): class 1
```

```
Attribute usage:

100.00% failures
 82.44% higher
 78.19% paid
 22.22% schoolsup
 20.99% guardian
 17.56% famrel
  8.09% freetime
  7.13% fjob
  4.25% nursery
  3.57% sex
  3.57% health
  3.43% goout
  3.02% reason
  2.47% famsup
  2.33% fedu
  2.33% studytime
  2.06% famsize
  1.10% alcohol
```

Se evalúa el modelo usando el conjunto de test, la accuracy del modelo obtenido es de un 79,36 %.

Analizando el uso de las variables en las decisiones, se observa que no aparece ninguna de las variables originales del dataset relacionadas con el consumo de alcohol ('dalc' o 'walc'). En cambio sí aparece la variable numérica construida a partir de ellas como media ponderada del consumo diario y en fin de semana. Esta variable ('alcohol') aparece en última posición, observando el árbol construido, se puede ver que esta variable únicamente participa en la clasificación de 8 observaciones por ello calcula una participación del 1.10 %:

$$\frac{8}{729} * 100 = 1{,}0973.$$

La variable que más participa en las decisiones es 'failures', el porcentaje de participación es el 100 % porque se encuentra en el nodo más alto del árbol y por tanto todas las reglas incluyen esta variable.

Con la técnica de boosting se consigue una pequeña mejora del modelo.

```
Evaluation on training data (729 cases):

Trial     Decision Tree
-----  ----------------
  Size      Errors

   0     28   95(13.0%)
   1     10  151(20.7%)
   2     34  177(24.3%)
   3     25  135(18.5%)
```

```
   4     28  163(22.4%)
   5     19  181(24.8%)
   6     14  136(18.7%)
   7     33  154(21.1%)
   8     34  148(20.3%)
   9     22  134(18.4%)
boost          52( 7.1%)   <<


   (a)   (b)    <-classified as
  ----  ----
   112    49    (a): class 0
     3   565    (b): class 1

Attribute usage:

100.00% school
100.00% failures
100.00% higher
 99.04% traveltime
 97.53% schoolsup
 95.47% fedu
 94.10% absences
 93.96% paid
 91.77% famrel
 84.22% pstatus
 83.26% fjob
 83.26% guardian
 82.44% goout
 78.88% studytime
 78.88% freetime
 73.25% mjob
 69.68% dalc
 69.14% health
 64.75% sex
 62.55% nursery
 62.00% reason
 58.02% age
 51.17% activities
 46.91% alcohol
 41.98% address
 33.33% internet
 32.24% medu
 28.40% famsup
 21.12% walc
 20.30% famsize
 11.25% romantic
```

En este caso sí aparecen todas las variables relacionadas con el consumo de alcohol, pero la primera lo hace en la posición 17. Existen muchas otras variables que participan más activamente en la decisión que el consumo de alcohol. Las variables más influyentes son 'school', 'failures' y 'higher'.

Haciendo una combinación de 10 árboles usando boosting se llega un porcentaje de acierto (accuracy) sobre el conjunto de test del 80.32 %. A continuación se muestra la matriz de confusión y una tabla con las medidas de evaluación más frecuentes en modelos de clasificación binaria.

	Predicción 0 (no)	**Predicción 1 (yes)**	**Total**
Real 0 (no)	TN = 17	FP = 52	**69**
Real 1 (yes)	FN = 10	TP = 236	**246**
Total	**27**	**288**	**315**

Tabla 3.12: Matriz de confusión del árbol de decisión entrenado con boosting

Medida	**Fórmula**	**Valor**
total	TP+FP+FN+TN	315
Accuracy	(TP+TN)/total	0.8031746
Misclassification Rate (1 - Accuracy)	(FP+FN)/total	0.1968254
True Positive Rate (Sensitivity o Recall)	TP/actual_yes	0.9593496
False Positive Rate	FP/actual_no	0.7536232
Specificity (1 - False Positive Rate)	TN/actual_no	0.2463768
Precision	TP/predicted_yes	0.8194444
Prevalence	actual_yes/total	0.7809524

Tabla 3.13: Medidas de evaluación del árbol de decisión entrenado con boosting

Para verificar que la validación del modelo no está condicionada por la selección de los datos concretos que han servido para entrenar y evaluar se repite la evaluación realizando 10-fold cross-validation y la accuracy media es del 78,45 %.

Capítulo 4

Conclusiones

A continuación se muestra un resumen de las conclusiones a las que se llega a través de los distintos análisis realizados.

A través del análisis bivariante: La media del número de ausencias aumenta a medida que aumenta la edad de los estudiantes. La nota media final está por debajo de aprobado en las edades 19, 21 y 22. El número medio de ausencias es prácticamente 0 de los 15 a los 18 años, a partir de esa edad, a medida que aumenta la edad también aumenta el número de ausencias. La nota media final desciende ligeramente a medida que el consumo de alcohol en fin de semana es más elevado. La nota media es fluctuante en función del consumo de alcohol en días laborables, no presenta una tendencia tan claramente descendiente como en el caso del consumo en fin de semana. La nota media final no supera el aprobado cuando hay suspensos previos.

A través de la segmentación: Calculando la media de la variable consumo global de alcohol en los diferentes grupos, se observa que el grupo 4 es el que presenta un valor medio más elevado. Este grupo se caracteriza por un elevado número de ausencias. El segundo grupo con un valor mayor de consumo de alcohol es el cluster 1 que se caracteriza por notas muy bajas por debajo de aprobado. En los grupos 2 y 3 el consumo medio de alcohol es parecido, ligeramente inferior en el cluster 2 que es el que tiene notas medias más altas.

A través de las reglas de asociación: Filtrando las reglas que contienen en la consecuencia si se aprueba o no la asignatura (variable 'bing3'), con soporte mínimo del 25 % y confianza mínima del 90 % se obtienen 18 reglas, de ellas, solo 3 contienen en el antecedente alguna de las variables sobre el consumo de alcohol, en los tres casos la consecuencia es que se aprueba la asignatura. En los tres casos el valor de alcohol consumido es muy bajo. Ninguna de las reglas tiene en la consecuencia que se suspende la asignatura por lo que no se puede concluir nada sobre los suspensos.

A través del profiling: En el grupo que suspende la proporción de estudiantes que no quiere realizar estudios superiores es mayor (20 %) que en general (8,52 %). En el grupo que suspende la proporción de estudiantes del colegio Mousinho da Silveira es mayor (36,95 %) que en general (26,95 %). En el grupo que suspende la proporción de estudiantes que sale muy a menudo con los amigos es mayor (23,04 %) que en general (15,61 %). En el grupo que aprueba la proporción de estudiantes cuya madre tiene estudios superiores es mayor (31,81 %) que en general (29,31 %). En el grupo que suspende la proporción de estudiantes cuya madre tiene estudios primarios es mayor (26,95 %) que en general (19,34 %). En el grupo que suspende la proporción de estudiantes cuyo padre tiene estudios primarios es mayor (34,78 %) que en general (24,52 %).

A través de los árboles de decisión: Usando un único árbol de decisión, se observa que no aparece ninguna de las variables originales del dataset relacionadas con el consumo de alcohol: consumo en días laborables ('dalc') o consumo en fin de semana ('walc'). En cambio sí aparece la variable numérica construida a partir de ellas como media ponderada del consumo diario y en fin de semana. Esta variable ('alcohol') aparece en última posición, por delante aparecen 17 variables, las más influyentes a la hora de decidir son: el número de suspensos previos, si desea realizar estudios superiores y si realiza clases de refuerzo, estas variables participan como mínimo en el 78 % de las decisiones. Con la técnica de boosting se consigue una pequeña mejora del modelo predictivo anterior. En este caso sí aparecen todas las variables relacionadas con el consumo de alcohol, pero la primera lo hace en la posición 17. Existen muchas otras variables que participan más activamente en la decisión que el consumo de alcohol. Las variables más influyentes son el colegio al que van, el número de suspensos previos y si desean realizar estudios superiores.

En resumen, el consumo de alcohol influye en las notas finales pero existen otras variables que influyen más como el colegio al que va, si quiere seguir estudiando, los suspensos previos, si hace refuerzo escolar, los estudios de los padres o el número de ausencias.

Se han cumplido los objetivos iniciales planteados. Se han podido realizar todas las tareas de preparación de datos, detección de valores faltantes y outliers, análisis y visualización de datos planificadas. Los análisis que han sido más concluyentes para ver si el consumo de alcohol influye en las notas finales de los estudiantes han sido el profiling y los árboles de decisión, algunas de las visualizaciones han ayudado o han dado algunas pistas que se han confirmado después.

La planificación inicial se ha cumplido, incluso se han anticipado algunas entregas parciales y ello ha permitido poder incorporar todas las recomendaciones que la Dra. Laia Subirats ha aportado para mejorar la calidad del trabajo.

La metodología utilizada ha sido adecuada y ha permitido iterar y mejorar el dataset inicial por ejemplo incluyendo la variable 'alcohol' que se ha calculado como media ponderada de las variables originales relativas al consumo en días laborables y en fin de semana. Esta variable no estaba prevista inicialmente, al añadirla todo el proceso previo se ha tenido que revisar. Eso es precisamente lo que contempla la metodología CRISP-DM.

Como continuación del trabajo actual se podría profundizar en la mejora del modelo predictivo usando técnicas de extracción de características como PCA o SVD que puedan complementar el dataset original y participar en la mejora de su capacidad predictiva. También sería interesante hacer selección de variables y hacer pruebas añadiendo o quitando algunas variables, esto podría ayudar a mejorar el modelo.

Por otra parte, la proporción de la clase 0 es bastante inferior a la de la clase 1. Para mejorar el modelo se podría intentar también "balancear", es decir, equilibrar la cantidad de casos de ambas clases. Seguramente por ese motivo el modelo predice muy bien los 1 (Sensitivity = 0.9593496) y en cambio predice correctamente muy pocos 0 (Specificity = 0.2463768).

Capítulo 5

Glosario

- **Accuracy:** en un modelo de clasificación es la proporción de casos correctamente clasificados respecto el total de casos.
- **Boosting:** técnica para mejorar la calidad de los modelos de clasificación, consiste en crear un proceso que genere varios modelos (en este caso árboles de decisión) de manera que los nuevos modelos tienen en cuenta los errores cometidos por los anteriores. Con esta técnica se consigue reducir el error global [3].
- **Boxplot:** también conocido como diagrama de caja y bigotes, es un gráfico que está basado en cuartiles y mediante el cual se visualiza la distribución de un conjunto de datos. Está compuesto por un rectángulo (la 'caja') y dos brazos (los 'bigotes'). Es un gráfico que suministra información sobre los valores mínimo y máximo, los cuartiles Q1, Q2 o mediana y Q3, y sobre la existencia de valores atípicos y la simetría de la distribución.
- **CRISP-DM:** acrónimo de Cross Industry Standard Process for Data Mining [14].
- **False negatives (FN):** en la matriz de confusión es la cantidad de casos que el modelo predice como clase negativa (0/no) y que realmente son de la clase positiva (1/yes). También se conoce como el error de tipo II.
- **False positives (FP):** en la matriz de confusión es la cantidad de casos que el modelo predice como clase positiva (1/yes) y que realmente son de la clase negativa (0/no). También se conoce como el error de tipo I.
- **False positives rate (FPR):** en la matriz de confusión es la proporción de FP con respecto al total de casos reales de la clase negativa. Representa la proporción de casos que el modelo predice sí cuando realmente son no.
- **KDD:** acrónimo de Knowledge Discovery in Databases [23].
- **LaTeX:** es un sistema de composición de textos, orientado a la creación de documentos escritos que presenten una alta calidad tipográfica. Por sus características y posibilidades, es usado de forma especialmente intensa en la generación de artículos y libros científicos que incluyen, entre otros elementos, expresiones matemáticas.
- **Matriz de confusión**: tabla de contigencia para modelos de clasificación que resume el desempeño de un algoritmo al comparar los resultados obtenidos en la predicción con los valores reales de la variable objetivo.
- **Misclassification Rate**: en un modelo de clasificación es la proporción de casos incorrectamente clasificados respecto el total de casos. Equivale a 1 - Accuracy y también se conoce como ratio de error.

- **PCA:** el análisis de componentes principales (Principal Component Analysis) es un algoritmo que ayuda a solucionar problemas de reducción de dimensionalidad y extracción de características. Se trata de una rotación de los ejes de coordenadas y la proyección de los datos sobre esos nuevos ejes [2].
- **Precision:** en la matriz de confusión es la proporción de TP con respecto al total de casos que predice de la clase positiva. Representa la proporción de casos que el modelo predice correctamente como sí con respecto a los que predice como sí.
- **Prevalence:** en la matriz de confusión es la proporción de casos reales de la clase positiva con respecto el total de casos de la muestra.
- **Prevalencia:** es el número total de los individuos que presentan un atributo o enfermedad en un momento o durante un periodo dividido por la población en ese punto en el tiempo o en la mitad del periodo. Cuantifica la proporción de personas en una población que tienen una enfermedad (o cualquier otro suceso) en un determinado momento y proporciona una estimación de la proporción de sujetos de esa población que tenga la enfermedad en ese momento.
- **Python:** es un lenguaje de programación interpretado cuya filosofía hace hincapié en una sintaxis que favorezca un código legible. Se trata de un lenguaje de programación multiparadigma, ya que soporta orientación a objetos, programación imperativa y, en menor medida, programación funcional. Es un lenguaje interpretado, usa tipado dinámico y es multiplataforma.
- **R:** entorno y lenguaje de programación enfocado al análisis estadístico. R proporciona un amplio abanico de herramientas estadísticas (modelos lineales y no lineales, tests estadísticos, análisis de series temporales, algoritmos de clasificación y agrupamiento, etc.) y gráficas.
- **Scatterplot:** también conocido como gráfico de dispersión, es un gráfico que muestra los valores de dos variables para un conjunto de datos como un conjunto de puntos sobre un sistema de coordenadas cartesianas. Los gráficos de dispersión permiten explorar la relación potencial entre pares de variables.
- **Specificity:** en la matriz de confusión es la proporción de TN con respecto al total de casos reales de la clase negativa. Representa la proporción de casos que el modelo predice correctamente como no con respecto a los que realmente son no. Es equivalente a 1 - False Positive Rate.
- **SVD:** la descomposición en valores singulares (Singular Value Descomposition) igual que PCA permite reducir la dimensionalidad y extraer características de un conjunto de datos. [2].
- **True negatives (TN):** en la matriz de confusión es la cantidad de casos que el modelo predice como clase negativa (0/no) y que realmente son de la clase negativa.
- **True positives (TP):** en la matriz de confusión es la cantidad de casos que el modelo predice como clase positiva (1/yes) y que realmente son de la clase positiva.
- **True positives rate (TPR):** en la matriz de confusión es la proporción de TP con respecto al total de casos reales de la clase positiva. Representa la proporción de casos que el modelo predice correctamente sí con respecto a los que realmente son sí. Este parámetro también se conoce como Sensitivity o Recall.

Bibliografía

[1] Gironés Roig, J. (2010) Business Analytics - Análisis de datos para organizaciones. FUOC (PID_00233434), Barcelona.

[2] Caihuelas Quiles, R. (2016) Extracción de características. FUOC (B2.332 - M1), Barcelona.

[3] Caihuelas Quiles, R. (2016) Clasificación con árboles de decisión. FUOC (B2.332 - M2), Barcelona.

[4] Caihuelas Quiles, R. (2016) Combinación de clasificadores. FUOC (B2.332 - M5), Barcelona.

[5] Rández, L. (2016) Introducción a LaTeX. IUMA. Departamento de Matemática Aplicada. Universidad de Zaragoza.

[6] Carmona, F. (2012) Generación automática de informes con Sweave y LaTeX. Departamento de Estadística. Universidad de Barcelona.

[7] Cortez, P., Silva. A. (2008) Using Data Mining to Predict Secondary School Student Performance. Portugal. `http://www3.dsi.uminho.pt/pcortez/student.pdf`

[8] Guía R Markdown. `http://fobos.inf.um.es/R/taller5j/30-markdown/guiabreve.pdf`. Accedido el 10 de febrero de 2017.

[9] Hoja de referencia R Markdown. `https://www.rstudio.com/wp-content/uploads/2015/03/rmarkdown-spanish.pdf`. Accedido el 10 de febrero de 2017.

[10] Plantilla LaTeX para trabajos de final de Máster de la Universidad Santiago de Compostela. `http://eio.usc.es/pub/mte/index.php/es/divulgacion/13?task=view`. Accedido el 12 de febrero de 2017.

[11] Tutorial Sweave. `https://cran.r-project.org/doc/contrib/Rivera-Tutorial_Sweave.pdf`. Accedido el 12 de febrero de 2017.

[12] Paquete LaTeX para insertar links en el índice. `http://minisconlatex.blogspot.com.es/2012/09/hyperlinks-con-latex.html`. Accedido el 12 de febrero de 2017.

[13] Conjunto de datos utilizado como base para el Trabajon Final de Máster: Using Data Mining To Predict Secondary School Student Alcohol Consumption. Fabio Pagnotta, Hossain Mohammad Amran. Department of Computer Science, University of Camerino. `https://archive.ics.uci.edu/ml/datasets/STUDENT+ALCOHOL+CONSUMPTION`. Accedido el 13 de febrero de 2017.

[14] Metodología CRISP-DM (Cross Industry Standard Process for Data Mining). `https://es.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining`. Accedido el 14 de febrero de 2017.

[15] Using data mining to predict secondary school student performance. `http://hdl.handle.net/1822/8024`. Accedido el 18 de febrero de 2017.

[16] Eliminar duplicados. https://www.kaggle.com/jhuno137/d/uciml/student-alcohol-consumption/machine-learning-on-student-alcohol-consumption. Accedido el 18 de febrero de 2017.

[17] Más información sobre el dataset STUDENT ALCOHOL CONSUMPTION. https://www.kaggle.com/uciml/student-alcohol-consumption. Accedido el 18 de febrero de 2017.

[18] Datos de la OMS sobre el consumo de alcohol http://www.who.int/mediacentre/factsheets/fs349/es/. Accedido el 02 de abril de 2017.

[19] Barnes,G.M., Welte,J.W. (1986) Patterns and predictors of alcohol use among 7-12th grade students in New York State. Journal of Studies on Alcohol, 47(1), 53–62. http://www.jsad.com/doi/10.15288/jsa.1986.47.53. Accedido el 22 de abril de 2017.

[20] Matin Ghayour Minaie, Ph.D., Ka Kit Hui, B.A., Rachel K. Leung, PH.D., John W. Toumbourou, PH.D., & Ross M. King, Ph.D. Parenting Style and Behavior as Longitudinal Predictors of Adolescent Alcohol Use. http://www.jsad.com/doi/pdf/10.15288/jsad.2015.76.671. Accedido el 22 de abril de 2017.

[21] López-Frías,M., Fernandez,M., Planells,E., Miranda,M.T., Mataix,J., Llopis,J. (2001) Alcohol consumption and academic performance in a population of Spanish high school students. Journal of Studies on Alcohol, 62(6), 741–744. http://www.jsad.com/doi/abs/10.15288/jsa.2001.62.741. Accedido el 22 de abril de 2017.

[22] Fuentes-Almendras,M., Mora-Ripoll,R., Dijk,A., Domínguez-García,A., Salleras-Sanmartí,L. (1999) Alcohol consumption among high school students in Barcelona, Spain. Journal of Studies on Alcohol, 60(2), 228–233. http://www.jsad.com/doi/abs/10.15288/jsa.1999.60.228. Accedido el 22 de abril de 2017.

[23] U. Fayyad, G. P.-Shapiro, and P. Smyth. (1996) From data mining to knowledge discovery in databases. AI Magazine, 17(3):37-54.