Grado en Matemática Computacional

Prácticas externas

y

Proyecto final de grado

Empleo de técnicas de minería de datos para analizar el abandono y la inserción laboral de los estudiantes de la UJI

Autora:
Laura Millán Roures

Supervisor:
Juan Antonio Hernández Rubert

Tutores académicos:
Amelia Simó Vidal
Pablo Gregori Huerta

Fecha de lectura: 25 de Julio de 2016
Curso académico 2015/2016

Resumen

Este documento es el trabajo presentado para la asignatura *MT1030 - Prácticas Externas y Proyecto Final de Grado*, del grado en Matemática Computacional de la Universitat Jaume I (UJI). Incluye una descripción del trabajo realizado durante la estancia en prácticas en el *Gabinete de Planificación y Prospectiva Tecnológica* de la universidad.

El proyecto desarrollado en esta institución ha consistido en la aplicación de técnicas de minería de datos para analizar dos problemas de gran importancia para la Universitat Jaume I, como son el abandono de los estudios y la inserción laboral de los estudiantes de los distintos grados que se imparten en la universidad.

Este documento incluye tanto la fundamentación teórica de las técnicas utilizadas como los resultados obtenidos en los análisis realizados. Para analizar el abandono universitario las técnicas utilizadas han sido: el análisis clúster, la regresión logística, el análisis discriminante, las redes neuronales y los árboles de clasificación. Para analizar la inserción laboral se ha hecho uso de la regresión logística y de las reglas de asociación.

Palabras clave

Abandono de los estudios, Inserción laboral, Regresión logística, Análisis clúster, Reglas de asociación.

Keywords

University dropout, Labour insertion, Logistic regression, Data clustering, Association rules.

Índice general

Capítulo 1

Introdución

En este primer capítulo se describe el contexto en el que el proyecto ha sido desarrollado, así como la motivación que ha permitido llevarlo a cabo y los objetivos generales que se pretenden conseguir durante el desarrollo del proyecto. Además, también se mencionan algunos estudios realizados en otras universidades sobre el abandono universitario que es el principal problema tratado durante la estancia en prácticas.

En el resto de capítulos se expone el trabajo realizado. Se argumenta tanto el motivo por el que se ha decidido realizar cada análisis como las dificultades obtenidas y los resultados finalmente extraídos.

En concreto, en el segundo capítulo se detalla la empresa donde se ha llevado a cabo la estancia en prácticas y el trabajo que se ha realizado para esta institución en líneas generales. En el tercer capítulo se describen teóricamente las técnicas utilizadas, en el cuarto capítulo se muestran los resultados de las técnicas más complejas y, por último, se exponen las conclusiones obtenidas.

1.1. Contexto y motivación del proyecto

El proyecto que se describe en este documento es un proyecto propuesto por el *Gabinete de Planificación y Prospectiva Tecnológica* de la Universitat Jaume I. Esta institución se encarga, principalmente, de dar soporte a nivel técnico y administrativo a la dirección de la universidad. Es por ello que tienen acceso a la información de todos los alumnos de la UJI, en cada curso que han realizado matrícula, para poder, entre otras cosas, elaborar informes dirigidos a los responsables de los centros académicos sobre datos de interés como, por ejemplo, la proporción de alumnos que ha abandonado una titulación, o el número de alumnos que ha cambiado de una titulación a otra.

No obstante, aunque proporcionen esta información, su trabajo no abarca la explicación de cuáles son las causas que provocan que los estudiantes abandonen los estudios, ni qué razones conllevan a que los alumnos consigan adentrarse en el mercado laboral una vez obtenido el

título. Por este motivo han considerado necesaria la incorporación de un estudiante del grado en Matemática Computacional para que, utilizando técnicas de minería de datos _explicadas al alumno por parte de los tutores académicos_, fuera posible analizar parte de la información que disponen, con el objetivo de extraer conclusiones que puedan ser utilizadas posteriormente por la universidad.

Con respecto a la **motivación** del proyecto, se debe considerar tanto la motivación **para la empresa** donde se ha realizado la estancia en prácticas _en este caso el *Gabinete de Planificación y Prospectiva Tecnológica*_ como la motivación **para el alumno** que realiza las prácticas en esta institución.

En lo que respecta a la **motivación para la empresa**, cabe destacar que se ha detectado la **necesidad de analizar y explicar cuáles son las causas que originan el abandono de los estudios**, para poder tomar futuras decisiones que ayuden a paliar el abandono en la medida de lo posible. Por otro lado, también se ha detectado la **necesidad de conocer qué aspectos ayudan a los alumnos a encontrar su primer empleo**, y a determinar si están suficientemente cualificados para ser competentes en el mundo laboral con los conocimientos adquiridos después de haber obtenido el título correspondiente.

Por otro lado, realizar la estancia en prácticas en el *Gabinete de Planificación y Prospectiva Tecnológica* resulta muy interesante **para el alumno** puesto que:

- Permite **familiarizarse con una nueva herramienta de consultas a bases de datos**, como es el software Oracle Discoverer.
- Ofrece la posibilidad de **estudiar el funcionamiento de técnicas de minería de datos y aplicarlas a un problema real**, asumiendo todos los obstáculos que surgen y tratando de resolverlos de la mejor manera posible para obtener un resultado satisfactorio.
- El alumno puede **conocer el funcionamiento de una empresa** y poner en práctica algunos de los conocimientos aprendidos durante la titulación.

Además, cabe remarcar que es un proyecto muy adecuado para poder ser utilizado en la elaboración del Trabajo de Final de Grado, porque permite que tanto la estancia en prácticas como el proyecto de final de grado vayan completamente unidos, de manera que, el proyecto final de grado consista en fundamentar teóricamente las técnicas utilizadas en la realización de la estancia en prácticas.

1.2. Objetivos generales

Los objetivos generales que se pretenden alcanzar durante la realización de este proyecto se pueden resumir en los siguientes:

- **Hacer uso de los conocimientos adquiridos en la asignatura *MT1020 - Bases de datos*** para poder entender cómo está estructurado el modelo de datos de la UJI y realizar consultas utilizando el software Oracle Discoverer.

- **Aprender diferentes técnicas estadísticas multivariantes**, tanto su fundamento teórico como su aplicación, para poder aplicarlas a problemas de la vida real, haciendo frente a todas las dificultades e inconvenientes que puedan ir surgiendo.
- **Ayudar a la universidad a explicar qué factores originan el abandono de los estudios y qué características son las que favorecen la inserción laboral**, para poder tomar futuras medidas al respecto. Para ello, se deben lograr los siguientes objetivos:
  - Describir de manera detallada ambas muestras, para depurar los datos y obtener unas primeras conclusiones acerca de los resultados que se deben lograr al realizar los análisis.
  - Determinar qué variables influyen en el abandono de los estudios.
  - Encontrar tipologías de estudiantes que abandonan los estudios.
  - Crear un modelo que permita predecir _después de haber cursado al menos el primer curso de la titulación_ si el estudiante abandonará, o no, los estudios utilizando para ello diferentes técnicas estadísticas y comparando los resultados.
  - Crear reglas que indiquen qué factores son los que más se repiten en los alumnos que han encontrado un empleo una vez obtenido el título universitario.
- **Integrar al alumno en una empresa real** con el propósito de enseñarle algunas de las vías laborales donde poder aportar sus conocimientos.

1.3. Estudios realizados en otras universidades sobre el abandono universitario

A partir de la década de los '80, ha surgido en las universidades de todo el mundo la preocupación por la calidad del servicio educativo que presentan. Esto ha dado lugar a procesos de evaluación con el fin de detectar las debilidades y fortalezas institucionales para generar acciones correctivas de las deficiencias encontradas [1]. Muchos de estos estudios han sido llevados a cabo en universidades sudamericanas como son algunas universidades de Argentina y México. No obstante, también se han encontrado varios artículos realizados en universidades españolas.

El abandono de los estudios universitarios es un problema cuyos costes son altos tanto para el individuo como para la sociedad. Es por ello que la prevención del mismo es fundamental y cobra especial relevancia en el actual contexto de crisis económica [2]. Diversos autores han desarrollado investigaciones con el objetivo de establecer modelos predictivos de este fenómeno (Castaño, Gallón, Gómez y Vásquez, 2004 [3]; Trevizán, Beltrán y Cosolito, 2009 [4]; Sánchez, 2014 [5]).

Un ejemplo de este tipo de estudio es el desarrollado por Araque, Roldán y Salguero (2009) [6]. Los autores recopilaron información de las bases de datos de la Universidad de Granada de una amplia muestra de estudiantes pertenecientes a 25 titulaciones para identificar las variables relevantes en el abandono. Realizaron el análisis por cada una de las facultades logrando identificar las variables relevantes tanto a nivel global como para cada una de las facultades.

Otro ejemplo es el de Marín, Infante y Troyano (2000) [7] que realizaron una investigación sobre el fracaso académico en la Universidad de Sevilla. A pesar de no contener una amplia muestra, resulta interesante por la cantidad de variables que utilizan en el estudio, dado que cuentan con cuestionarios, entrevistas y tests de inteligencia, motivación y preferencias profesionales.

Por otro lado, Rodrigo, Molina, García-Ros y Pérez-Gonzalez (2012) [8] analizan las variables relevantes en el abandono de los estudios del grado en Psicología, llegando a la conclusión de que las variables que explican el abandono son el sexo, la vía de acceso, la dedicación (a tiempo completo o parcial), el orden de preferencia de la titulación al realizar la preincripción, la residencia familiar y el nivel de estudios de los padres.

De entre los ejemplos de la aplicación de la inteligencia artificial, se debe destacar el trabajo de Sánchez (2014) [5] que consiste en encontrar un modelo capaz de predecir el abandono en el primer año de carrera. Para ello, el autor analiza los datos de tres cohortes de nuevo ingreso (2009-2011) a partir de los datos disponibles en el sistema informático de la universidad y, probando con distintos modelos, el mejor resultado que obtiene es del 80 % de los alumnos predichos correctamente.

También cabe destacar el trabajo de Bernardo, A., Cerezo, R., Núñez, J.C, Tuero, E. y Esteban, M (2015) [2] para estudiar las variables influyentes en el abandono de la universidad de Oviedo. Hacen uso tanto de variables sociodemográficas (nacionalidad, tamaño familiar, nivel de estudios y ocupación de los padres) como otras variables relativas a su ingreso y progreso en la universidad (si los alumnos se han incorporado a los estudios universitarios una vez iniciadas las clases, si asisten a clase, etc.) y comparan los resultados obtenidos con algunos de los estudios mencionados en los párrafos anteriores para corroborar las conclusiones extraídas.

Capítulo 2

Estancia en prácticas

2.1. Introducción

En este capítulo se expone el proyecto realizado durante la estancia en prácticas. En primer lugar, se explica en qué empresa se ha realizado la estancia y se describe la funcionalidad de esta institución. A continuación, se detalla el software del que se ha hecho uso para extraer los datos a analizar y la herramienta utilizada para realizar los análisis. En las siguientes dos secciones se describe, en primer lugar, el plan de trabajo inicial y, en segundo lugar, el trabajo realizado durante la estancia en prácticas en líneas generales. Los resultados de los análisis, y la realización de los mismos, se muestran con más detalle en el capítulo 4.

2.2. Descripción de la empresa

La empresa donde se ha realizado la estancia en prácticas es el *Gabinete de Planificación y Prospectiva Tecnológica* de la Universitat Jaume I. Esta unidad está situada en el rectorado de la universidad y se encarga, principalmente, de dar soporte a nivel técnico y administrativo a la dirección de la UJI. Está constituida por siete trabajadores, uno de los cuales, Juan Antonio Hernández Rubert, ha sido mi supervisor.

Entre las distintas tareas que realizan, se encuentran las siguientes:

- Dan soporte a la ordenación académica de los títulos oficiales de grado y máster. Los centros académicos realizan planes de estudio que ellos revisan y mejoran para su posterior implantación.
- Definen la oferta académica. Determinan los títulos oficiales, las asignaturas de cada título y en cada asignatura cuántos grupos y subgrupos de teoría, de laboratorios y de problemas hay.
- Elaboran los horarios de docencia académica y de los exámenes, y crean los circuitos de matrícula para los estudiantes de nuevo ingreso.

- Dan soporte al plan de ordenación docente. Determinan la docencia que ejerce cada profesor.
- Elaboran presupuestos. Calculan el coste de profesorado anual.
- Distribuyen el presupuesto ordinario de los departamentos y de la biblioteca.
- Dan soporte al plan estratégico. Establecen planes y acciones para lograr los objetivos planteados.
- Dan soporte a la elaboración de estadísticas por parte de instituciones, como pueden ser el ministerio o el INE. Coordinan y pasan datos oficiales de la UJI a estos organismos.
- Estudian la forma de cubrir incidencias del profesorado, bien sea por enfermedad, por baja maternal, etc. Analizan el impacto de estas incidencias en la capacidad docente de las áreas afectadas, y lo comunican al Servicio de Recursos Humanos para poder cubrir las necesidades docentes generadas.
- Realizan algunas predicciones y estadísticas a partir de datos oficiales de la UJI.

2.3. Descripción del software utilizado durante la estancia en prácticas

Oracle Discoverer

La herramienta software, utilizada para la **extracción de los datos** de las bases de datos de la UJI ha sido **Oracle Discoverer**. Este software permite realizar consultas sin necesidad de recurrir a SQL, permitiendo de este modo una gestión más cómoda y sencilla de los datos a analizar.

Cuando se crea una consulta, se muestran todas las tablas de la base de datos en forma de carpetas. Si se despliega una de estas carpetas, aparecen todos los atributos que posee dicha tabla. De este modo, se pueden seleccionar aquellos atributos que se desea que aparezcan en la consulta. Aunque se muestran todas las tablas que tiene la base de datos que se está utilizando, la aplicación sólo permite seleccionar aquellas tablas que tengan alguna relación con las tablas que ya se hayan seleccionado para la consulta. Una vez definidos los campos que se desean mostrar, se deben incluir las condiciones a realizar sobre los datos. Además, también permite crear cálculos para efectuar algún procesamiento sobre la información extraída y obtener totales.

La información se distribuye en libros de trabajo donde cada trabajo puede estar compuesto por distintas hojas de trabajo que se organizan de un modo muy similar a cómo se organizan las hojas en Microsoft Excel. De este modo, si se desean realizar consultas similares se deben organizar en hojas de trabajo dentro de un mismo libro de trabajo. Esto permite tener toda la información extraída bien organizada y acceder a las restricciones que se han realizado en las otras hojas de trabajo para así poder habilitarlas y deshabilitarlas en función de la consulta que se deba realizar.

Microsoft Excel

Para **unir los datos extraídos** a partir de las consultas realizadas a la base de datos con Oracle Discoverer, se ha hecho uso del programa **Microsoft Excel**. Esta herramienta se ha utilizado para realizar la unión entre las distintas consultas y asignar a cada alumno sus datos correspondientes. En concreto, se ha empleado la función *BUSCARV*, que ha permitido unir los datos obtenidos de forma cómoda y eficaz.

RStudio

La herramienta utilizada para **realizar los análisis** ha sido el programa **R**. Es un software libre que permite realizar análisis estadísticos de datos. Se ha escogido este programa debido a la gran variedad de métodos estadísticos que cubre. Estos métodos están contenidos en diferentes bibliotecas o paquetes con funcionalidades de cálculo y graficación [9]. En concreto, se ha utilizado RStudio, una interfaz para el R que contiene una serie de herramientas integradas y diseñadas para ayudar al usuario a hacer uso del programa R de forma más práctica y cómoda.

2.4. Descripción del plan de trabajo inicial

El trabajo a realizar durante la estancia en prácticas ha consistido en analizar el abandono y la inserción laboral de los estudiantes de la Universitat Jaume I.

En un primer momento no se realizó una planificación temporal exacta de las tareas a realizar, porque se trata de un proyecto de utilidad para la institución donde se ha desarrollado la estancia en prácticas, pero se desliga de las actividades que los trabajadores de esta institución deben realizar. Por este motivo, resulta complicado conocer a priori la duración de cada una de las tareas a realizar, y el tiempo que debe dedicar el alumno al aprendizaje de las técnicas posteriormente empleadas en el análisis.

El trabajo inicialmente propuesto consistió en familiarizarse con el modelo de datos de la UJI para poder extraer, mediante Oracle Discoverer, todas aquellas variables que posteriormente se deseaban analizar. En la propuesta inicial, estos análisis podían ser realizados sobre grados, másters y doctorados, pero durante el trascurso del proyecto se decidió que era más conveniente centrarse únicamente en grados, principalmente por falta de tiempo, debido a que la extracción de los datos y su posterior comprobación y corrección conlleva en cada caso una dedicación de tiempo bastante elevada. De este modo, se dispondría de más tiempo para enfatizar sobre cada uno de los dos problemas tratados y se desglosaría el análisis por centros académicos para lograr unas mejores conclusiones, ya que se detectó que la realización de un modelo global para todos los centros no era lo suficientemente precisa, puesto que las diferencias en los centros afectaban al modelo.

Por tanto, el plan de trabajo se fue desarrollando a medida que avanzaba el proyecto. Durante las cinco primeras quincenas se abordó el problema del abandono universitario, y a lo largo de la última quincena el de la inserción laboral, porque en el caso de éste último se dispone de

muy poca información. Ello es debido a que los primeros grados se implantaron en el año 2009, por lo que los primeros egresados finalizaron su titulación en el año 2012. Por ello, se dedicó un tiempo menor a analizar este problema.

Durante las **dos primeras quincenas** se debió obtener la información de los estudiantes de la UJI que iniciaron expediente en alguno de los grados durante los cursos 2009 a 2011, utilizando el programa Oracle Discoverer.

Una vez conocido el funcionamiento del programa Oracle Discoverer, y realizadas las consultas necesarias, en la **tercera quincena** se aprendió a hacer uso de algunas técnicas de minería de datos como son el análisis clúster, la regresión logística, las reglas de asociación, etc., para a continuación seleccionar, de entre todas las técnicas aprendidas, aquellas que se consideraban más adecuadas para aplicar, durante la **cuarta y quinta quincena**, a los datos extraídos.

Finalmente, durante la **última quincena** se debió extraer la información necesaria sobre los alumnos que han finalizado los estudios de grado. No obstante, en este caso, la mayor parte de la información se obtuvo a partir de encuestas que la universidad realiza cada año a los alumnos mediante correo electrónico. Por este motivo, el proceso de extracción de datos fue menor que en el caso del abandono. A continuación, se decidió qué técnicas aplicar para analizar este problema y se llevaron a cabo.

2.5. Trabajo realizado

Por un lado, el trabajo realizado durante la estancia en prácticas radica principalmente en la necesidad de explicar las causas que originan el abandono de los estudios de grado de la Universitat Jaume I, así como la búsqueda de tipologías de alumnos que abandonan los estudios, y de un modelo que permita predecir si un estudiante abandonará, o no, la titulación que está realizando. Por otro lado, también se deben encontrar cuáles son las características que favorecen la inserción laboral de los estudiantes. Se abordan ambos problemas por separado en los dos apartados siguientes.

2.5.1. Abandono de los estudios

El primer problema planteado sobre el **abandono de los estudios** es encontrar un criterio que permita determinar qué alumnos han abandonado los estudios. De entre los diferentes criterios solicitados al *Gabinete de Planificación y Prospectiva Tecnológica* para elaborar informes anteriores, cabe destacar los tres siguientes:

- En el **Real Decreto 1393/2007**, se define la tasa de abandono como la relación porcentual entre el número de estudiantes de una cohorte de nuevo ingreso que debieron obtener el título en el año anterior, y que no se han matriculado, ni en ese año ni en el año precedente. Dicho de otro modo, se buscan aquellos alumnos que no se han matriculado ni en el año de finalización teórica de los estudios, ni en el siguiente, y que no han finalizado la titulación que estaban cursando.

- **Criterio SIUVP** (Sistema de Información de las Universidades Valencianas Públicas): se define la tasa de abandono como el número de alumnos matriculados en un grado en el curso académico $n-2/n-1$, que no se han matriculado en ese grado en el curso $n-1/n$ ni en el curso $n/n+1$, sin ser egresados en el mismo título. Dicho de otro modo, se buscan los estudiantes matriculados en un año, que no se han vuelto a matricular en ese grado en los dos años siguientes, y que no han finalizado la titulación que estaban cursando.
- **Criterio CRUE** (Conferencia de Rectores de las Universidades Españolas): se define la tasa de abandono como el número de alumnos matriculados de nuevo ingreso en el curso académico $n-2/n-1$, y que no han vuelto a matricularse en los cursos académicos $n-1/n$ y $n/n+1$. Dicho de otro modo, se buscan aquellos alumnos de nuevo ingreso que no se han matriculado en ninguno de los dos años siguientes.

Al tener tres criterios diferentes, y no saber cuál es el más adecuado para extraer los datos, se ha decidido comprobar si los criterios obtenidos son equivalentes. Para ello, se ha obtenido, por cada uno de los criterios, el número de alumnos que ha abandonado cada titulación en cada año para, a continuación, calcular la matriz de correlaciones entre los datos obtenidos a partir de los tres criterios. Únicamente se ha debido de comprobar para los años 2009, 2010 y 2011, porque el criterio del *Real Decreto 1393/2007* requiere por definición comprobar que el alumno no se ha matriculado en el año teórico de finalización de los estudios ni en el siguiente. Por tanto, teniendo en cuenta que las titulaciones son de cuatro años se puede calcular como máximo el número de abandonos de los alumnos que iniciaron una titulación en el año 2011.

Las correlaciones obtenidas entre los tres criterios para cada uno de los cursos académicos han sido muy próximas a uno, por lo que se ha conseguido demostrar que **los tres criterios son equivalentes**.

Por tanto, ante la posibilidad de elegir cualquiera de los tres criterios anteriormente mencionados se ha decidido escoger el criterio de *Real Decreto 1393/2007* por ser el más demandado a los trabajadores del *Gabinete de Planificación y Prospectiva Tecnológica*, y por la menor dificultad que requiere a la hora de realizar las consultas.

Las variables extraídas para realizar el estudio sobre el abandono universitario son las que se muestran en la Tabla 2.1.

Tabla 2.1: Variables utilizadas en el estudio del abandono universitario.

Variable	**Definición**
edad	Edad del estudiante a fecha de inicio en la universidad.
nota.de.acceso	Nota que obtuvo el alumno en la prueba de acceso realizada para acceder a la universidad. Se ha dividido por la nota media de acceso de ese año en esa titulación para evitar que la nota de acceso dependa de la titulación.
Continúa en la página siguiente	

Tabla 2.1 – continuación de la página anterior

Variable	Definición
orden.pref	Opción en la que el alumno situó la titulación que finalmente cursó al realizar la preinscripción.
via.acc	Vía de acceso a través de la que el alumno entró en la universidad. Es una variable categórica que toma los valores: Selectividad, Formación Profesional, Titulados Universitarios y Resto (en el caso de entrar por una vía de acceso distinta a las anteriores).
sexo	Sexo del estudiante.
prov	Provincia de la residencia familiar durante el curso académico de inicio. Es una variable categórica que toma los valores: Castellón, Limítrofes (para las provincias de Valencia, Tarragona, Teruel) y Resto.
cred.pres.pri	Número de créditos que el alumno presentó a examen en el primer curso que hizo matrícula.
cred.pres.ultimo	Número de créditos que el alumno presentó a examen en el último curso que hizo matrícula.
num.asi.matric.pri	Número de asignaturas en las que el alumno se matriculó en primer curso. No incluye las asignaturas que reconoció.
num.asi.matric.ultimo	Número de asignaturas en las que el alumno se matriculó en el último curso que hizo matrícula. No incluye las asignaturas que reconoció.
cred.honor.pri	Número de créditos de honor que el alumno obtuvo en el primer año de matrícula.
cred.honor.ultimo	Número de créditos de honor que el alumno obtuvo en el último año de matrícula.
trabPri	Variable binaria que indica si el alumno tenía algún empleo en el primer curso en el que hizo matrícula.
trabUltimo	Variable binaria que indica si el alumno tenía algún empleo en el último curso en el que hizo matrícula.

Continúa en la página siguiente

Tabla 2.1 – continuación de la página anterior

Variable	Definición
abandono	Variable binaria que toma el valor 1 si el alumno ha abandonado los estudios o el valor 0 en caso contrario.
num.asi.rep.ultimo	Número de asignaturas en las que el alumno se matriculó en el último curso de matrícula, pero que ya había cursado anteriormente, y no superó.
cred.rec.media	Promedio de créditos que el alumno reconoció durante los cursos en los que hizo matrícula.
cred.sup.exam.media	Promedio de créditos de los que se examinó el estudiante, y que superó en los cursos que hizo matrícula. No se incluyen los créditos que reconoció.

Nota: Como se observa en la Tabla 2.1, para las variables que se pueden evaluar en los distintos cursos académicos, se ha decidido escoger la variable evaluada en el primer y en el último curso de matrícula. De este modo, al no tomar las variables evaluadas en todos los cursos en los que el alumno ha hecho matrícula, se ha evitado la presencia de valores nulos que resultan un inconveniente en algunos análisis. Sin embargo, en las dos últimas variables que se muestran en la tabla, se ha utilizado un promedio durante los cursos en los que el alumno ha hecho matrícula, en lugar de las variables de primer y último curso por separado porque ambas variables están altamente correlacionadas, y este hecho puede afectar a los resultados obtenidos en los diferentes análisis realizados.

Una vez obtenidas todas las variables descritas en la Tabla 2.1, el siguiente paso ha consistido en realizar una descripción de la muestra representando gráficos de barras o de caja dependiendo del tipo de variable, separando por un lado los alumnos que han abandonado los estudios, y por otro lado los alumnos que no han abandonado, para ver si hay diferencias en la distribución de cada una de las variables entre abandonar y no abandonar los estudios. Además, se han realizado tests de independencia de la *Chi-cuadrado* para las variables categóricas y contrastes de hipótesis de la *t de Student* para las variables numéricas, con el objetivo de comprobar si realmente hay diferencias significativas entre los dos grupos en cada una de las variables. Se puede encontrar la **descripción completa de la muestra** en el Anexo A de este documento.

La primera técnica de minería de datos utilizada para analizar el abandono ha sido el **análisis clúster**. Este análisis ha permitido **encontrar tipologías de alumnos que abandonan los estudios**. Para ello, se ha seleccionado el algoritmo más conveniente para las variables a estudio y, haciendo uso de este algoritmo, se ha clasificado a los alumnos en varios grupos. De esta forma se han podido detectar patrones que se repiten en los alumnos que abandonan los estudios para poder identificar las causas más frecuentes de abandono. El objetivo de este estudio es encontrar patrones que sean obvios, como por ejemplo que un motivo por el que los alumnos abandonan es porque encuentran trabajo, patrones no tan obvios como el anterior, y desmentir patrones que puedan considerarse evidentes en el abandono, pero que en realidad no lo sean.

La segunda técnica utilizada ha sido la **regresión logística**. Esta técnica ha permitido, por un lado, detectar aquellas **variables** de las consideradas en el análisis **que influyen en el abandono** y, por otro lado, crear un modelo que permita predecir, dado un nuevo alumno, si abandonará, o no, los estudios que está cursando. Además, se ha probado que funciona correctamente y que predice de manera adecuada la mayor parte de los alumnos, incluso aunque esos estudiantes no hayan sido incluidos en la muestra utilizada para generar el análisis.

Otra de las técnicas empleadas ha sido el **análisis discriminante**. La funcionalidad de este análisis es bastante similar a la de la anterior pero, en este caso, también es útil si se dispone de más de dos grupos. En realidad no hubiera sido necesario realizar este análisis porque la regresión logística es un método más completo, ya que ofrece también la explicación de qué variables son las más influyentes en el abandono. Esto es debido a que incluye una probabilidad (*p-valor*) que indica si realmente hay diferencias significativas en los valores que toman cada una de las variables, en función de si los alumnos han abandonado, o no, los estudios. No obstante, también se ha realizado este análisis para poder comparar los resultados.

Además de las técnicas mencionadas anteriormente, también se ha hecho uso de las **redes neuronales** para, de nuevo, **clasificar a los alumnos en función de si han abandonado, o no, los estudios** y, de este modo, poder comparar los resultados obtenidos con los producidos en los dos análisis anteriores.

Finalmente, se han utilizado **árboles de clasificación**. Está técnica, a diferencia de las anteriores, permite observar **qué decisiones se evalúan cuándo se desea predecir si un alumno abandonará, o no, los estudios** y en qué orden se realizan dichas evaluaciones.

Los análisis mencionados se han aplicado en primer lugar a todos los alumnos de la muestra creando de este modo un modelo global. No obstante, algunos de estos análisis, en concreto el análisis clúster y la regresión logística, se han aplicado también por centros académicos, pues se ha considerado que los resultados obtenidos no eran lo suficientemente precisos y que se podían obtener mejores conclusiones si se realizaban por centros ya que, como se puede intuir, las causas de abandono son diferentes dependiendo del centro considerado a estudio.

2.5.2. Inserción laboral de los estudiantes

Para analizar el problema de la **inserción laboral**, se ha extraído gran parte de la información necesaria a partir de encuestas enviadas por correo electrónico a los alumnos que han finalizado alguna de las titulaciones implantadas en la UJI. Por este motivo, la cantidad de alumnos que las han contestado ha sido bastante inferior al número de alumnos egresados, ya que una vez finalizados los estudios de grado, gran parte de los alumnos deja de acceder al correo de la universidad o, si lo hacen, en ocasiones no dedican el tiempo necesario a realizar las encuestas.

Como las encuestas se envían cada año a todos los alumnos que han finalizado la titulación, se han tomado únicamente los datos de la encuesta que se realizó en el año 2014, para evitar que los estudiantes pudieran estar duplicados si habían realizado varias de las encuestas enviadas. A esta información se han unido también algunos datos académicos extraídos mediante consultas a las bases de datos de la UJI.

Las variables que se han considerado en este estudio son las que se muestran en la Tabla 2.2.

Tabla 2.2: Variables utilizadas en el estudio de la inserción laboral.

Variable	**Definición**
sexo	Sexo del estudiante.
edad	Edad del estudiante en el año en que realizó la encuesta (2014).
practicas.extracurriculares	Variable binaria indicando si el alumno realizó prácticas extracurriculares.
erasmus.estudios	Variable binaria indicando si el alumno participó en un programa de Erasmus por motivo de estudios.
erasmus.practicas	Variable binaria indicando si el alumno participó en un programa de Erasmus por motivo de prácticas externas.
trabajo.empresa.practicas	Variable binaria indicando si el alumno permaneció trabajando en la empresa donde realizó las prácticas una vez finalizada la estancia en prácticas.
trabajo	Variable binaria indicando si el alumno tiene algún empleo en el momento de realizar la encuesta.
necesita.titulacion	Variable binaria indicando, en caso de que el alumno tenga algún tipo de empleo, si ese trabajo requiere el título universitario obtenido.
trabajo.titulacion	Variable binaria indicando: - 1: El alumno ha obtenido un trabajo que requiere la titulación cursada - 0: El alumno no tiene trabajo, o tiene un trabajo que no requiere el título obtenido
tipo.empresa	Variable categórica indicando el tipo de empresa donde el alumno ha conseguido trabajo: - 1: Pública - 2: Privada - 3: Otro
Continúa en la página siguiente	

Tabla 2.2 – continuación de la página anterior

Variable	Definición
jornada	Variable categórica indicando si el alumno trabaja a jornada completa o a jornada parcial.
nota.expediente	Nota media obtenida en el expediente.
cursos	Variable categórica indicando el tiempo que ha tardado el alumno en obtener el título. Los posibles valores son: más de cuatro años o cuatro años o menos.
situacion.actual	Variable categórica indicando la situación actual del alumno: - 1: Autoempleado - 2: Empleado - 3: Desocupado pero buscando empleo - 4: Desocupado y sin buscar empleo
tiempo.encontrar.empleo .categorica	Variable categórica indicando el tiempo que ha tardado el alumno en encontrar un empleo: - 1: Menos de 3 meses - 2: Entre 4 y 6 meses - 3: Entre 7 y 9 meses - 4: Entre 10 y 12 meses - 5: Más de 12 meses

Una vez extraídas todas las variables mencionadas en la Tabla 2.2, como la inserción laboral en un centro académico es muy diferente a la del resto, se ha decidido separar la muestra original en cuatro para realizar un análisis por cada uno de los centros.

En primer lugar, **se han descrito cada una de las muestras** dibujando diagramas de sectores para mostrar cuántos alumnos han conseguido trabajo, cuántos de los que trabajan necesitan el título obtenido para desempeñar ese puesto de trabajo, cuántos se quedaron trabajando en la empresa donde realizaron la estancia en prácticas, qué tipo de jornada laboral tienen los alumnos que han encontrado un empleo que requiere la titulación cursada, en qué tipo de empresas trabajan, etc. A continuación, se han dibujado diagramas de barras y de cajas y se han realizado contrastes de hipótesis para ver qué variables influyen a la hora de encontrar un trabajo que requiere la titulación cursada. Este estudio se muestra con detalle en el Anexo A.

Una vez descrita cada una de las muestras, el siguiente paso ha consistido en realizar un análisis de **regresión logística** para ver qué **variables explican la inserción laboral** y, por otro lado, también se ha hecho uso de las **reglas de asociación** para **encontrar patrones, en los estudiantes, que impliquen encontrar un trabajo que requiere el título obtenido**. No obstante, no se han podido extraer grandes conclusiones debido a que las muestras son

demasiado pequeñas, por lo que sería conveniente volver a repetir el análisis dentro de unos cuántos años, cuando el número de egresados sea bastante superior.

2.6. Informes realizados

Al finalizar la estancia en prácticas se redactaron **dos informes**, uno sobre el abandono universitario y otro sobre la inserción laboral, para que quedara constancia de los análisis realizados y de las conclusiones obtenidas. Además, se realizó una **presentación al Consejo de Dirección de la UJI** para concienciar de la utilidad de realizar este tipo de análisis y para que pudieran conocer las conclusiones extraídas.

2.7. Grado de consecución de los objetivos propuestos

En general, los objetivos propuestos han sido **satisfechos con éxito**. Por un lado, se ha conseguido aprender a realizar consultas sobre las bases de datos de la UJI y, por otro lado, también se ha aprendido a hacer uso de varias técnicas de minería de datos que han permitido poder extraer varias conclusiones de interés para el *Gabinete de Planificación y Prospectiva Tecnológia*. El mayor éxito se ha conseguido en el análisis del abandono universitario, sin embargo, hubiera sido deseable obtener mejores conclusiones sobre la Facultad de Ciencias de la Salud. No ha sido posible debido a la heterogeneidad que presentan sus diferentes titulaciones. Sería conveniente volver a repetir los análisis realizados para cada uno de los grados de esta facultad.

Por lo que respecta a la inserción laboral, los objetivos conseguidos han sido ligeramente inferiores a los objetivos propuestos, pues se carecía de suficiente información como para poder obtener reglas que pudieran indicar qué características permiten conseguir un empleo que requiere el título obtenido. Aún así, se ha decidido realizar también el estudio permitiendo, de este modo, que la empresa pueda utilizarlo para realizar un estudio similar cuando dispongan de mayor cantidad de información.

2.8. Conclusiones

Mi experiencia en el *Gabinete de Planificación y Prospectiva Tecnológica* ha sido gratamente satisfactoria. He podido tener mi primera experiencia laboral en una empresa en la que me he sentido bastante integrada desde el primer día. Además, me he sentido muy atendida en todo momento por parte de mi supervisor, que me ha enseñado y me ha ayudado siempre que ha sido necesario, para poder familiarizarme con el software que utilizan en la empresa.

Por otro lado, también he conseguido ampliar mis conocimientos en técnicas de minería de datos, gracias a la dedicación de mis tutores, que me han enseñado todas las técnicas necesarias para poder realizar los análisis de datos.

Finalmente, también he podido descubrir una de las posibles salidas laborales del grado en Matemática Computacional, que considero que puede ser bastante valorada en el mundo empresarial.

Capítulo 3

Fundamentación teórica del TFG

3.1. Motivación y objetivos de la minería de datos

Los grandes avances en tecnología de almacenamiento, y la aparición de un gran motor de búsqueda como es Internet, ha ocasionado que se generen y se almacenen datos de manera exponencial.

Actualmente, la mayor parte de la información se almacena en medios digitales como son discos duros, tarjetas de vídeo, etc. El abaratamiento de estos dispositivos y la facilidad de transportar la información de un lugar a otro, ha tenido como consecuencia que todo el mundo tenga al alcance grandes volúmenes de datos. Además, la aparición de e-mails, páginas web, blogs, etc., ha ocasionado que se generen billones de terabytes de nueva información cada día [10]. La necesidad de analizar todo ese conjunto de datos y los grandes avances en procedimientos de cálculo han impulsado la aparición de la minería de datos.

La minería de datos es un campo de las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. Muchas de las técnicas de minería de datos se basan en modelos estadísticos conocidos desde hace años, pero ha sido preciso el desarrollo del poder de cálculo de estos últimos años, para que pudieran ser utilizables de manera sencilla.

Las técnicas de minería de datos pueden ser clasificadas en dos conjuntos. Por un lado, se encuentran las *técnicas descriptivas* y, por otro lado, las *técnicas de inferencia*. En las primeras, el usuario que las aplica no tiene una idea pre-especificada de los resultados que pretende obtener; únicamente busca conocer las características y la estructura que tiene la información que está analizando, mientras que en las segundas técnicas, el usuario pretende confirmar la validez de unas hipótesis.

Los principales problemas que se pretenden resolver en minería de datos son:

- **Buscar relaciones o dependencias entre las variables.** Si la muestra dispone de gran cantidad de variables, seguramente muchas de ellas estarán altamente correlacionadas, y

con un número menor de variables se podría tener prácticamente la misma información.

- **Clasificar las observaciones en grupos predeterminados** (clasificación supervisada). Los grupos donde clasificar las observaciones están predefinidos a priori. El objetivo es encontrar una regla de clasificación que permita determinar, dado un nuevo elemento, en qué grupo debe ser clasificado.

- **Construir grupos de observaciones similares según las variables estudiadas** (clasificación no supervisada). No se establece ningún grupo a priori, aunque es necesario determinar el número de grupos que se deben crear.

- **Realizar asociaciones** para determinar hechos que ocurren en común dentro de un determinado conjunto de datos.

Los análisis de minería de datos que se van a detallar en este capítulo son el análisis clúster para construir grupos de observaciones similares, la regresión logística y los árboles de clasificación para clasificar observaciones en grupos predeterminados y las reglas de asociación para realizar asociaciones. Éstos son algunos de los métodos de los que se ha hecho uso durante la estancia en prácticas.

3.2. Análisis clúster

3.2.1. Introducción

El análisis clúster tiene por objeto agrupar elementos en grupos homogéneos en función de las similitudes entre ellos. Este método se conoce también por el nombre de clasificación automática o no supervisada para distinguirlos de los métodos de clasificación supervisada como son la regresión logística y los árboles de clasificación que se describen en las secciones 3.3 y 3.4 donde los grupos para realizar la clasificación ya están determinados a priori.

El desarrollo del análisis clúster ha sido llevado a cabo de modo interdisciplinar. Han contribuido a su desarrollo taxonomistas, psicólogos, sociólogos, matemáticos, ingenieros, médicos, etc. La palabra *data clustering* apareció por primera vez en un artículo de 1954 relacionado con datos antropológicos [10].

A lo largo del tiempo, se han propuesto muchos procedimientos para realizar este tipo de análisis debido a que no existe una definición rigurosa de clúster. De entre los diferentes métodos de agrupación cabe destacar los métodos de partición que dividen los objetos en un número de grupos prefijado, los métodos jerárquicos que construyen una jerarquía a partir de la que se pueden construir los grupos y los métodos basados en modelos que asumen una distribución de probabilidad conocida en la población.

Un libro excelente para seguir este apartado es el libro de Kaufman, L. y Rousseeuw, P.J. [11].

3.2.2. Descripción teórica del análisis clúster

Se considera que se han observado p variables en una muestra aleatoria de tamaño n. La muestra queda organizada en una matriz de datos de dimensiones $n \times p$:

$$X_{n\times p} = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1p} \\ x_{21} & x_{22} & ... & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & ... & x_{np} \end{bmatrix} \quad \text{siendo } x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix},$$

el vector observación del individuo i-ésimo.

El objetivo de este análisis es que, en base a estas variables, se puedan clasificar las n observaciones en grupos de forma que las observaciones pertenecientes a un mismo grupo sean muy similares entre sí, y las pertenecientes a grupos distintos lo más diferentes posible. Como se ha comentado en la introducción, el número de procedimientos que se ha propuesto en la literatura es muy grande. En este documento se detallan únicamente los métodos de partición.

Antes de comenzar a explicar los métodos de partición, se deben mostrar algunos tipos de distancias, o disimilaridades, que pueden utilizarse en los diferentes métodos de partición.

Tipos de distancias y disimilaridades

Dados dos vectores de observación $x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}$ y $x_j = \begin{bmatrix} x_{j1} \\ x_{j2} \\ \vdots \\ x_{jp} \end{bmatrix}$, se define una distancia como una aplicación $d : \mathbb{R}^p \times \mathbb{R}^p \longmapsto \mathbb{R}$ verificando las siguientes propiedades:

1) $d(x_i, x_j) \geq 0$

2) Si $x_i = x_j \Longrightarrow d(x_i, x_j) = 0$

3) $d(x_i, x_j) = d(x_j, x_i)$ *Propiedad simétrica*

4) $d(x_i, x_j) \leq d(x_i, x_k) + d(x_k, x_j)$ *Propiedad triangular*

Si d no verifica la propiedad triangular, se dice que d es una medida de disimilaridad. En algunos algoritmos se puede trabajar con disimilaridades. No obstante, en otros es necesario trabajar con distancias.

A continuación, se muestra qué distancias o disimilaridades se pueden utilizar en función del tipo de variables de la muestra.

★ **Para variables numéricas:**

Si todas las variables de la muestra son cuantitativas, se puede trabajar con distancias. Las más utilizadas son las basadas en las normas L_q *(o de Minkowsky)*: $d_q(x_i, x_j) = \sqrt[q]{\sum_{l=1}^{p} |x_{il} - x_{jl}|^q}$

En particular,

- *Distancia de Manhattan* (L_1): $d_1(x_i, x_j) = \sum_{l=1}^{p} |x_{il} - x_{jl}|$
- *Distancia euclídea* (L_2): $d_2(x_i, x_j) = \sqrt{\sum_{l=1}^{p} (x_{il} - x_{jl})^2}$
- *Distancia de Chebychev* (L_∞): $d_\infty(x_i, x_j) = \max_{l=1,\dots,p} |x_{il} - x_{jl}|$

Cuanto mayor es q más enfásis se da a las diferencias en cada variable y por tanto más influencia tendrán los valores atípicos de la muestra *(outliers)*.

Estas distancias no son invariantes frente a cambios de escala, por lo que si la magnitudes de los datos de las variables no son comparables, habrá variables que influirán mucho más que otras en los resultados. Así pues, se deberán estandarizar previamente los datos si las unidades de medida no son comparables.

Una distancia que no depende de las unidades es:

$d(x_i, x_j) = \sum_{l=1}^{p} |\frac{x_{il} - x_{jl}}{x_{il} + x_{jl}}|$ *Distancia de Camberra*

★ **Para variables binarias:**

Si todas las variables de la muestra son binarias (cada variable sólo puede tomar los valores 0 o 1), se pueden utilizar dos tipos de distancia: la *proporción de no coincidencias*, y el *coeficiente de Jacard*.

Dados dos elementos x_i, x_j se construye una tabla del siguiente modo:

x_i / x_j	1	0
1	a	b
0	c	d

entonces se pueden definir las siguientes dos distancias:

- *Proporción de no coincidencias*: $d(x_i, x_j) = \frac{b+c}{a+b+c+d}$
- *Coeficiente de Jacard*: $d(x_i, x_j) = \frac{b+c}{a+b+c}$

Estas distancias difieren en el papel dado a los acuerdos en el 0. El *coeficiente de Jacard* no los tiene en cuenta, pues trata las variables como asimétricas y considera que un acoplamiento en el 0-0 aporta menos información que un acoplamiento en el 1-1; es decir, que los elementos que coinciden en el 1-1 son más similares entre sí que los que coinciden en el 0-0.

★ **Medidas para datos de tipo mixto:**

Para variables de tipo mixto se debe hacer uso de la distancia de *"Gower"*. Su expresión es la siguiente:

$$d(x_i, x_j) = \frac{\sum_{l=1}^{p} w_{ij}^{(l)} \cdot d_{ij}^{(l)}}{\sum_{l=1}^{p} w_{ij}^{(l)}},$$

donde $w_{ij}^{(l)}$ es siempre 1 excepto si la comparación entre los individuos x_i y x_j en la l-ésima variable no es posible de realizar porque hay valores faltantes o bien si la variable l es binaria asimétrica y se tiene entre los individuos x_i y x_j un acoplamiento en 0-0.

El valor d_{ij}^{l} es la disimilaridad entre los individuos x_i y x_j para la l-ésima variable:

- **Si la variable l es binaria o categórica**: $\begin{cases} d_{ij}^{(l)} = 0 \text{ si } x_{il} = x_{jl} \\ d_{ij}^{(l)} = 1 \text{ si } x_{il} \neq x_{jl} \end{cases}$
- **Si la variable l es numérica**: $d_{ij}^{(l)} = \frac{|x_{il} - x_{jl}|}{R_l}$ donde R_l es la diferencia entre el valor máximo y el valor mínimo de la l-ésima variable.
- **Si la variable l es ordinal** se deben calcular los rangos normalizados y tratar a continuación como si fuese una variable numérica.

  Para calcular los rangos normalizados, si M_l es el conjunto de distintos valores que puede tomar la variable l, se calcula para cada valor x_{il}: $z_{il} = \frac{x_{il}-1}{M_l-1}$.

Métodos de partición

Son los algoritmos más utilizados para realizar agrupaciones entre objetos. Son métodos descriptivos y no se basan en ningún modelo de probabilidad.

Parten de un número K de clústers específico de modo que $K < n$ siendo n el número de individuos de la muestra.

Cada grupo se etiqueta con un valor $k \in \{1, ..., K\}$ y cada observación es asignada a un sólo grupo $C(i) = k \quad \forall i \in \{1, ..., n\}$.

Se busca una asignación $C^*(i)$ que minimice la suma de disimilaridades entre todos los elementos dentro de cada grupo, es decir, que minimice la siguiente función:

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(j)=k} d(x_i, x_j). \tag{3.1}$$

Esta función se divide por dos para que las distancias entre dos pares de elementos sólo se sumen una vez. Recibe el nombre de *función objetivo* y se utiliza también para determinar el número de grupos en que debe dividirse la muestra.

Para buscar la asignación óptima que minimice la suma de disimilaridades entre los elementos de cada grupo se buscan estrategias iterativas que garanticen la convergencia a un óptimo local.

Aunque hay distintos métodos para resolver este problema, todos tienen en común los siguientes pasos:

- Se parte de una asignación inicial.
- En cada paso se cambia una pequeña parte de las asignaciones de forma que el valor de $W(C)$ de la ecuación (3.1) disminuya con respecto al paso anterior.
- Cuando no se puede producir ninguna mejora el algoritmo termina.

De entre los métodos de partición, se debe destacar el método de *k-medias* y el de *k-medoides* que se explican en las siguientes secciones.

■ Método de k-medias

Es el algoritmo de partición más conocido. Tiene una historia muy larga porque fue descubierto de forma independiente en diferentes campos científicos por Steinhaus (1956) [12], Lloyds (1957) [13], Ball y Hall (1965) y MacQueen (1967) [14]. Aunque este algoritmo fue propuesto hace más de 50 años, continúa siendo uno de los más usados por su simplicidad y eficiencia. Sólo puede emplearse cuando todas las variables de la muestra son numéricas. La distancia utilizada es la distancia euclídea al cuadrado. Por tanto, el objetivo de este método es encontrar una partición C que minimice:

$$W(C) = \tfrac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(j)=k} d_2(x_i, x_j)^2 = \tfrac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(j)=k}\sum_{l=1}^{p}(x_{il} - x_{jl})^2.$$

Se muestran tres algoritmos distintos para lograr alcanzar este objetivo:

▲ Método de Lloyds.

Es el método más sencillo y el inicialmente propuesto. Consta de tres pasos que se muestran en la Tabla 3.1.

Paso	Descripción
Paso 0	Se parte de K puntos que se toman como los vectores de medias asociados a los K-grupos. Para elegir estos puntos se puede hacer aleatoriamente, tomando los más alejados, etc.
Paso 1	Se asigna cada observación al clúster cuya media esté más cercana a él.
Paso 2	Una vez realizadas las asignaciones se recalculan los vectores de medias de cada uno de los grupos.

Tabla 3.1: Pasos del método de Lloyds.

Nota: Los pasos 1) y 2) son iterados hasta que las asignaciones no cambien.

▲ Método de MacQueen

El método de MacQueen [14], es muy similar al método de Lloyds. La principal diferencia que presenta este algoritmo con respecto al de Lloyds es que las medias de cada clúster son recalculadas cada vez que una observación cambia de clúster. Esta característica hace que el algoritmo de MacQueen sea más eficiente que el de Lloyds.

Este método consta de los siguientes pasos contenidos en la Tabla 3.2.

Paso	Descripción
Paso 0	Se parte de K puntos que se toman como los vectores de medias asociados a los K-grupos. El método de MacQueen propuesto inicialmente propone considerar los K primeros elementos de la muestra.
Paso 1	Se asigna cada una de las observaciones al clúster cuya media esté más cercana, con la característica de que al efectuar cada asignación se recalcula de nuevo la media del clúster donde ha sido asignada la observación y la media del clúster donde estaba anteriormente clasificada esa observación.
Paso 2	Una vez realizadas todas las asignaciones, se actualizan los centros de cada clúster calculando la media de los elementos contenidos en cada clúster.

Tabla 3.2: Pasos del método de MacQueen.

Nota: Los pasos 1) y 2) son iterados hasta que las asignaciones no cambien.

▲ **Método de Hartigan-Wong**.

El método de Hartigan-Wong [15], es el algoritmo que va incorporado por defecto en el programa R. El objetivo general de este algoritmo es buscar una partición en K grupos que sea localmente óptima moviendo los elementos de un clúster a otro. Este método se muestra en la Tabla 3.3.

Paso	**Descripción**
Paso 0	Se parte de K puntos que se toman como los vectores de medias asociados a los K-grupos. Para elegir estos puntos se puede hacer aleatoriamente, tomando los más alejados, etc.
Paso 1	El primer paso consiste en que, para cada individuo x_i de la muestra $i = 1, ..., n$, se buscan los dos clústers cuyos centros estén más cercanos. Se denomina al clúster más cercano a x_i, $C_1(i)$, y al segundo clúster más cercano, $C_2(i)$. Se asigna x_i al clúster $C_1(i)$.
Paso 2	Una vez realizadas todas las asignaciones, se actualizan los centros de cada clúster calculando la media de los elementos contenidos en cada clúster.
Paso 3	Inicialmente todos los clústers pertenecen a lo que se denomina el *'conjunto activo'* y que posteriormente estará formado por los clústers involucrados en la transferencia de elementos del Paso 6 denominada *'quick-transfer'*.
Paso 4	Esta etapa se denomina *'optimal-transfer'*. Se considera cada vector observación x_i donde $i = 1, ..., n$. Sea x_i perteneciente al clúster k_1. Si el clúster k_1 está en el denominado *'conjunto activo'* se debe realizar el Paso 4a y, en otro caso, el Paso 4b. **Paso 4a:** Se calcula $R_2 = [n_k \cdot d(x_i, k)^2]/[n_k + 1]$ donde $d(x_i, k)$ es la distancia euclidea del elemento x_i al centro del clúster k. Esta cantidad debe calcularse para todos los clústers excepto para el clúster k_1. Notemos que n_k es el número de elementos del clúster k. Sea k_2 el clúster con el menor R_2. Si R_2 es mayor o igual que $R_1 = [n_{k_1} \cdot d(x_i, k_1)^2]/[n_{k_1} - 1]$, no es necesario realizar ninguna reasignación y k_2 es el nuevo $C_2(i)$. Notar que R_1 permanece idéntico para el elemento x_i hasta que el clúster k_1 es actualizado. En otro caso, el elemento x_i es reasignado al clúster k_2 y k_1 es el nuevo $C_2(i)$. Los centros de los clústers son actualizados cuando se produce una nueva reasignación. **Paso 4b:** Es el mismo paso que el Paso 4a pero R_2 se calcula únicamente para los clústers que están en el *'conjunto activo'*.

Paso 5:	El algoritmo termina si el *'conjunto activo'* está vacío. En otro caso, se continúa con el Paso 6.
Paso 6:	Esta etapa se denomina *'quick-transfer'*. Se considera cada vector observación x_i donde $i = 1, ..., n$. Sea el clúster $k_1 = C_1(i)$ y el clúster $k_2 = C_2(i)$. Si los dos clústers no han cambiado en los últimos n pasos no es necesario realizar ninguna comparación. En otro caso, se calcula $R_1 = [n_{k_1} \cdot d(x_i, k_1)^2]/[n_{k_1} - 1]$ y $R_2 = [n_{k_2} \cdot d(x_i, k_2)^2]/[n_{k_2} + 1]$. Si $R_1 < R_2 \Rightarrow x_i$ permanece en el clúster $C_1(i)$. En otro caso, se intercambian $C_1(i)$ y $C_2(i)$, se actualizan los centros de los clústers k_1 y k_2 y estos dos clústers se incluyen en el *'conjunto activo'*.
Paso 7:	Si ninguna transferencia ha tenido lugar en los últimos n pasos se debe ir al Paso 4. En otro caso, se debe ir al Paso 6.

Tabla 3.3: Pasos del método de Hartigan-Wong.

El problema del algoritmo de *k-medias* es que es muy sensible a datos atípicos, que crea grupos esféricos, y que sólo es apropiado para datos cuantitativos.

■ Método de k-medoides

Este método fue propuesto en 1987 por Kaufman, L. y Rousseeuw, P.J. [11]. Es muy similar al algoritmo *k-medias* pero puede usarse con cualquier tipo de datos y no sólo con variables numéricas. El único paso del algoritmo anterior que necesita que la distancia sea euclídea, es el que toma como representante de cada clúster a las medias. Este algoritmo sustituye este paso forzando a que los representantes de cada clúster sean una de las observaciones.

Los pasos de los que consta el algoritmo se muestran en la Tabla 3.4.

Paso	**Descripción**
Paso 0	Se parte de K puntos $\{m_1, ..., m_k\}$ que se toman como los representantes de los K-grupos.
Paso 1	Se asigna cada observación al clúster cuyo representante esté más cercano a él, es decir, $C_{(i)} = \underset{k}{arg\,mín} \quad d(x_i, m_k)$
Paso 2	Una vez realizadas las asignaciones, se recalculan los representantes de cada grupo minimizando la distancia total a los otros puntos del clúster.

Tabla 3.4: Pasos del método de k-medoides.

Nota: Los pasos 1) y 2) son iterados hasta que las asignaciones no cambien.

La convergencia de todos estos algoritmos mencionados está asegurada pero el resultado puede ser un mínimo local, por lo que se debería repetir el algoritmo desde distintos puntos iniciales, y elegir la solución con el menor valor de la función objetivo.

Elección del número de grupos

En los métodos de partición es necesario indicar el número de grupos en que se desea dividir la muestra.

Las opciones más utilizadas para resolver este problema son:

- Calcular el valor de $W(C)$ (3.1) para varios clústers, por ejemplo desde 2 hasta 8, y representar un gráfico del valor de $W(C)$ frente al número de grupos. A continuación, se deben buscar los dos valores consecutivos de $W(C)$ que estén más distantes y se debe escoger aquel que tenga un menor valor. Comúnmente esta técnica también recibe el nombre de buscar un 'codo'.
- Calcular alguna medida de agrupamiento y maximizarla.

Médidas de bondad del análisis clúster

Todos los algoritmos comentados dan como resultado una clasificación, pero a veces es posible que no exista una estructura de grupos en la muestra. Es por ello necesario comprobar que la clasificación realizada resulta lo suficientemente aceptable.

A continuación, se detallan dos de las técnicas utilizadas: *la silueta* y *el método Gap.*

♦ Silueta

Para cada elemento x_i de la muestra se define su silueta como $s(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}}$ siendo $a(i)$ la distancia media del elemento x_i a todos los de su clúster y siendo $b(i)$ la distancia media del elemento x_i a todos los elementos del clúster más cercano sin incluir el clúster donde ha sido clasificado. Es decir, $a(i) = promedio\left(\sum_{j:C(j)=C(i)} d(x_i, x_j)\right)$ y $b(i) = promedio\left(\sum_{j:C(j)\neq C(i)} d(x_i, x_j)\right)$ con $C(j)$ el más cercano a x_i en este último caso.

La silueta es un valor que se encuentra entre -1 y 1 para cualquier elemento. Si es muy próxima a 1 quiere decir que el elemento está muy bien clasificado, mientras que si es muy próxima a -1 indica que el elemento está muy mal clasificado.

Como medida de bondad del análisis clúster, se puede calcular la media de los valores de la silueta dentro de cada clúster y, como medida de bondad global, la media de la silueta para todo el conjunto de datos.

♦ **Método Gap**

Tibshirani, Walther y Hastie (2001) propusieron, en [16], un método para estimar el número óptimo de clústers. Este método también puede utilizarse después de aplicar cualquier algoritmo de clustering para analizar la bondad de la clasificación.

Se considera que se ha aplicado un algoritmo de clasificación y que se han obtenido k clústers, C_1, C_2, ..., C_k. La idea propuesta es comparar, para cada k, el valor de $log(W(C_k))$ (siendo $W(C_k)$ el valor de la función objetivo (3.1) evaluada para un número de clústers k) con su esperanza bajo una distribución nula de referencia en la que se asume que no hay clústers:

$$Gap_n(k) = E_n^*[\log(W(C_k))] - \log(W(C_k)), \tag{3.2}$$

donde E_n^* es la esperanza para muestras de tamaño n extraídas de la distribución de referencia. Un primer estimador del número de clústers, en la distribución de la que proviene la muestra, es el valor $\hat{k}$ que maximiza el estimador de $Gap_n(k)$.

Por lo tanto, el problema consiste en proponer una distribución de referencia apropiada y construir la distribución de muestreo del estadístico Gap.

La distribución de referencia:

Se considera el contraste,

$$\begin{cases} H_0 : k = 1 \\ H_1 : k > 1 \end{cases}$$

Es decir, se supone que no existen grupos en la muestra o lo que es lo mismo que todos los individuos pertenecen a un mismo grupo ($k = 1$) y se quiere rechazar este modelo si se encuentra algún $k > 1$ para el que existe suficiente evidencia.

Tibshirani, Walther y Hastie proponen modelar las distribuciones de un sólo clúster como densidades log-cóncavas, es decir, utilizando funciones de densidad cuyo logaritmo es una función cóncava, como por ejemplo la distribución normal. Denotan por S^p al conjunto de tales distribuciones en $\mathbb{R}^p$.

Para buscar una adecuada distribución de referencia, los autores consideran la versión de $Gap_n(k)$ para el caso del algoritmo de *k-medias*:

$$g(k) = \log\left\{\frac{MSE_{X^*}(k)}{MSE_{X^*}(1)}\right\} - \log\left\{\frac{MSE_X(k)}{MSE_X(1)}\right\}, \tag{3.3}$$

donde

$$MSE_X(k) = E(\min_{\mu \in A_k} d_2(X - \mu)^2),$$

con $A_k \subset \mathbb{R}^p$, conjunto de k puntos escogido para minimizar $MSE_X(k)$.

Se divide por MSE(1) en ambos términos para asegurarse de que $g(1) = 0$. Con esto lo que se busca es la distribución de referencia para X^* menos favorable (en el sentido de rechazar H_0), tal que $g(k) \leq 0 \quad \forall X \in S^p$ y todo $k \geq 1$. Es decir, que el primer cociente de (3.3) sea lo más pequeño posible.

Los autores demuestran que en el caso $p = 1$, la distribución de referencia es la uniforme $U[0, 1]$. Sin embargo, para el caso $p > 1$ prueban que no es posible escoger una distribución de referencia que resulte aplicable en general por lo que deben utilizarse métodos de Monte Carlo.

Cálculo del estadístico Gap:

El **estadístico Gap** se define como la expresión obtenida en (3.2) al sustituir $E_n^*[\log(W(C_k))]$ por su estimación según la distribución de referencia:

$$\hat{G}ap_n(k) = \hat{E}_n^*[\log(W(C_k))] - \log(W(C_k)).$$

Hay dos maneras de elegir la distribución de referencia por métodos de Monte Carlo:

- **Generar distribuciones uniformes dentro del rango de valores observados para cada una de las p variables.**
- **Generar distribuciones uniformes a partir de las componentes principales de los datos**. Sea X la matriz de datos de dimensiones $n \times p$. Se asume que las variables tienen media cero ya que de no ser así, debe restarse la media a cada una de las variables para centrar los datos. A continuación, se calcula la descomposición de X en valores singulares $X = UDV^t$ donde:

  - U es una matriz $n \times n$
  - D es una matriz diagonal $n \times p$ con valores no negativos en la diagonal, denominados valores singulares de X.
  - V^t es una matriz $p \times p$.

  Las n columnas de U se denominan vectores singulares izquierdos y las p columnas de V, vectores singulares derechos.

  La descomposición en valores singulares está relacionada con la descomposición en vectores y valores propios. Los vectores singulares izquierdos son los vectores propios de XX^t y los vectores singulares derechos son los vectores propios de X^tX.

  Se calcula $X' = XV$ y se seleccionan datos aleatorios Z' a partir de uniformes sobre los rangos de columnas de X'. Por último, se transforma $Z = Z'V^t$ para obtener una muestra de datos de la distribución de referencia.

En cada caso, se extraen B muestras aleatorias de n observaciones de la distribución de referencia, es decir, se repite B veces alguno de los dos métodos. A continuación, se calcula el valor de $\log(W(C_k))$ en cada muestra y se estima $E_n^*[\log(W(C_k))]$ como una media sobre las B copias.

Para realizar el contraste, interesa controlar la distribución de muestreo del estadístico Gap. Sea $sd(k)$ la desviación típica de $\log(W(C_k)^*)$ en las B copias, y considerando el error en la simulación, la desviación típica de $E_n^*[\log(W(C_k))]$ resulta ser $s_k = \sqrt{1 + \frac{1}{B}} \cdot sd(k)$.

Los autores proponen como criterio elegir $\hat{k}$ como el valor más pequeño de los k que verifique que $\hat{G}ap(k) \geq \hat{G}ap(k+1) - s_{k+1}$ y prueban mediante distintos ejemplos que el criterio funciona correctamente.

3.3. Regresión logística

3.3.1. Introducción

La regresión logística forma parte de los métodos de discriminación o clasificación, al igual que los árboles de clasificación que se explican en la sección 3.4. Se utiliza para predecir el valor de una variable categórica en función de los valores que toman un conjunto de variables denominadas variables explicativas o predictoras. Además, una de las ventajas que presenta con respecto a la mayor parte de los modelos de discriminación es que determina qué variables explican la variable respuesta y cómo influyen cada una de las variables explicativas en la clasificación. El modelo de regresión con respuesta cualitativa más sencillo posible es el modelo binario en el que la variable respuesta sólo puede tomar dos posibles valores; éste es el modelo que se detalla. Para profundizar más en el tema, se puede consultar el libro de Alan Agresti, *Categorical data analysis* [17].

El contenido teórico de la regresión logística ha sido extraído principalmente del libro de Daniel Peña, *Análisis de Datos Multivariantes* [18].

3.3.2. Descripción teórica de la regresión logística

Se considera el problema de discriminación entre dos poblaciones. Una forma de abordar el problema es suponer que un objeto puede pertenecer a uno de dos posibles grupos. Se parte de los valores observados de $X_1, X_2, \ldots, X_k$ variables numéricas en n objetos y una variable respuesta Y indicando el grupo al que pertenece. La variable respuesta es una variable binaria que toma el valor 0 cuando el elemento pertenece a la primera población, y el valor 1 cuando pertenece a la segunda. La muestra queda organizada del siguiente modo:

$$X = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1k} \\ x_{21} & x_{22} & ... & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & ... & x_{nk} \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Los objetivos principales de la regresión logística son los siguientes:

- Obtener una función que permita prever el valor de Y cuando se conocen las variables $X_1, X_2, ..., X_k$.

- Analizar qué variables de las consideradas influyen en la respuesta.
- Valorar la calidad de las predicciones.

No se puede formular este modelo como un modelo de regresión lineal múltiple de la forma $Y = \beta_0 + \beta_1 \cdot X_1 + ... + \beta_k \cdot X_k + \epsilon$ con $\epsilon \sim N(0, \sigma^2)$ porque:

- No está garantizada que la predicción esté entre 0 y 1.
- Como Y sólo puede tomar los valores 0 y 1, ϵ sólo puede tomar los valores $-(\beta_0 + \beta_1 \cdot X_1 + ... + \beta_k \cdot X_k$ y $1 - (\beta_0 + \beta_1 \cdot X_1 + ... + \beta_k \cdot X_k)$, es decir, ϵ es una variable binaria (Bernouilli) y no sigue una distribución normal.

Para garantizar que la variable respuesta esté entre 0 y 1, se debe transformar el modelo de forma que la probabilidad de pertenecer al grupo definido por $Y = 1$ sea una función no lineal de X, que siempre esté entre 0 y 1. Para ello, se pueden utilizar las funciones *logit, probit* y *log-log complementaria*. La más utilizada es la función *logit* luego en este documento se explica únicamente ésta.

Como la variable respuesta sólo puede tomar dos posibles valores, que son 0 y 1, esta variable sigue una distribución de probabilidad de Bernouilli de parámetro p donde p es la probabilidad de pertenecer al grupo 1.

La función que liga la media (en este caso, $\mu = p$ por seguir una distribución de Bernouilli) con los predictores lineales $(\beta_0 + \beta_1 \cdot X_1, ..., \beta_k \cdot X_k)$ se denomina función "$link$". En el caso de la función *logit*, su función "$link$" inversa viene dada por:

$$p = \frac{1}{1+e^{-(\beta_0+\beta_1 \cdot X_1+...+\beta_k \cdot X_k)}} = \frac{1}{1+e^{-(\beta_0+\beta_1' \cdot X)}}, \qquad \text{siendo} \qquad \beta_1' = (\beta_1, ..., \beta_k).$$

La función "$link$" es una función continua cuya expresión es $g(p) = \log(\frac{p}{1-p}) = \beta_0 + \beta_1 \cdot X_1 + ... + \beta_k \cdot X_k$. Es una función lineal de las variables explicativas. Representa en una escala logarítmica la diferencia de las probabilidades de pertenecer a ambas poblaciones y al ser una función lineal facilita la estimación y ayuda a interpretar el modelo.

Se puede observar el gráfico de la función "$link$" inversa en la Figura 3.1.

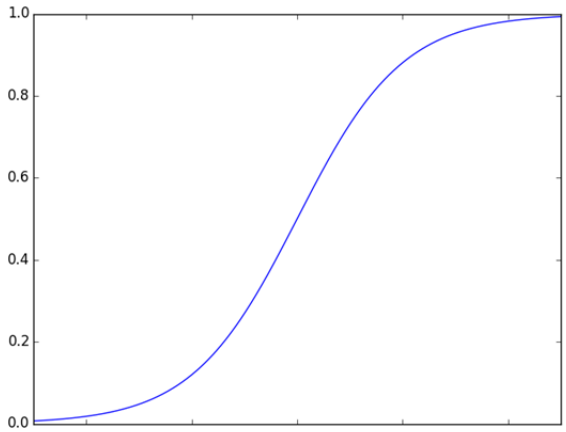


Figura 3.1: Función "*link*" inversa de la función *logit*.

Estimación de parámetros por máxima verosimilitud

Para estimar los parámetros $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$ del modelo se realiza una estimación por el método de máxima verosimilitud.

Como la variable respuesta Y es de tipo Bernouilli (0 ó 1), la función de probabilidad para una respuesta y_i cualquiera es:

$$P(y_i) = p_i^{y_i} \cdot (1-p_i)^{1-y_i} \quad \text{con} \quad p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 \cdot x_{i1}+\ldots+\beta_k \cdot x_{ik})}}; \qquad y_i = 0, 1.$$

Como una muestra aleatoria es un conjunto de n variables aleatorias independientes entre sí e idénticamente distribuidas (con la misma distribución de probabilidad), la distribución de probabilidad conjunta de la muestra es:

$$f(y_1, ..., y_n) = f(y_1) \cdot f(y_2) \cdot \ldots \cdot f(y_n) = \prod_{i=1}^{n} p_i^{y_i} \cdot (1-p_i)^{1-y_i}.$$

Esta función, como función del vector de parámetros $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$ se denomina función de verosimilitud y es la función que se debe maximizar. Se pueden tomar logaritmos pues el logaritmo es una función creciente y, en este caso, es más fácil obtener el estimador máximo verosímil hallando el máximo del logaritmo de la verosimilitud.

$$\log f(y_1,...,y_n) = \log(\prod_{i=1}^{n} p_i^{y_i} \cdot (1-p_i)^{1-y_i}) = \sum_{i=1}^{n} \log\left(p_i^{y_i} \cdot (1-p_i)^{1-y_i}\right) =$$

$$\sum_{i=1}^{n}(\log(p_i^{y_i}) + \log(1-p_i)^{1-y_i}) = \sum_{i=1}^{n}(y_i \log p_i + (1-y_i)\log(1-p_i)) = \sum_{i=1}^{n}(y_i \log p_i + \log(1-p_i)$$

$$-y_i \log(1-p_i)) = \sum_{i=1}^{n}(y_i \log(\frac{p_i}{1-p_i}) + \log(1-p_i)). \tag{3.4}$$

Teniendo en cuenta que,

$$1 - p_i = 1 - \frac{1}{1+e^{-x_i^t \cdot \beta}} = 1 - \frac{e^{x_i^t \cdot \beta}}{1+e^{x_i^t \cdot \beta}} = \frac{1+\cancel{e^{x_i^t \cdot \beta}}}{1+e^{x_i^t \cdot \beta}} - \frac{\cancel{e^{x_i^t \cdot \beta}}}{1+e^{x_i^t \cdot \beta}} = \frac{1}{1+e^{x_i^t \cdot \beta}},$$

la función soporte (el logaritmo de la verosimilitud) es:

$$L(\beta) = \sum_{i=1}^{n}(y_i \log(\tfrac{p_i}{1-p_i}) + \log(1-p_i)) = \sum_{i=1}^{n}(y_i x_i^t \cdot \beta + \log(\tfrac{1}{1+e^{x_i^t \cdot \beta}}) \quad \text{donde} \quad x_i = \begin{bmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{ik} \end{bmatrix}.$$

Como $\log(1/x) = -\log(x)$, entonces se tiene:

$$L(\beta) = \sum_{i=1}^{n} y_i \cdot x_i^t \cdot \beta - \sum_{i=1}^{n} \log(1+e^{x_i^t \cdot \beta}).$$

Para obtener el estimador máximo verosímil de β, se deriva la función soporte para hallar el vector β que maximiza esa función. Escribiendo el resultado como un vector columna:

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{n} y_i \cdot x_i - \sum_{i=1}^{n}\left(\frac{1}{1+e^{x_i^t \cdot \beta}} \cdot (e^{x_i^t \cdot \beta} \cdot x_i)\right) = \sum_{i=1}^{n} x_i \cdot (y_i - \frac{e^{x_i^t \cdot \beta}}{1+e^{x_i^t \cdot \beta}}) = \sum_{i=1}^{n} x_i \cdot (y_i - \frac{1}{e^{-x_i^t \cdot \beta}+1}).$$

Se iguala a cero para buscar el máximo:

$$\sum_{i=1}^{n} x_i \cdot (y_i - \frac{1}{1+e^{-x_i^t \cdot \beta}}) = 0 \Rightarrow \sum_{i=1}^{n} x_i \cdot (y_i - \hat{p_i}) = 0.$$

Se tienen $k+1$ ecuaciones no lineales en los parámetros β. Para obtener el vector β que maximiza la verosimilitud se debe hacer uso de un método de tipo *Newton-Raphson*. Desarrollando el vector $\frac{\partial L(\beta)}{\partial \beta}$ alrededor de un punto β_a, se tiene:

$$\frac{\partial L(\beta)}{\partial \beta} = \frac{\partial L(\beta_a)}{\partial \beta} + \frac{\partial^2 L(\beta_a)}{\partial \beta \partial \beta^t}(\beta - \beta_a).$$

Para que el punto β_a corresponda al máximo de la verosimilitud, su primera derivada debe anularse. Imponiendo la condición $\frac{\partial L(\beta_a)}{\partial \beta} = 0$ se obtiene:

$$\frac{\partial L(\beta)}{\partial \beta} = \cancel{\frac{\partial L(\beta_a)}{\partial \beta}} + \frac{\partial^2 L(\beta_a)}{\partial \beta \partial \beta^t}(\beta - \beta_a).$$

Multiplicando por la inversa de $\frac{\partial^2 L(\beta a)}{\partial \beta \partial \beta^t}$ a ambos lados se tiene:

$$\left(\frac{\partial^2 L(\beta_a)}{\partial \beta \partial \beta^t}\right)^{-1} \cdot \frac{\partial L(\beta)}{\partial \beta} = \beta - \beta_a.$$

Luego,

$$\beta_a = \beta + \left(-\frac{\partial^2 L(\beta_a)}{\partial \beta \partial \beta^t}\right)^{-1} \cdot \frac{\partial L(\beta)}{\partial \beta},$$

donde derivando de nuevo $\frac{\partial L(\beta)}{\partial \beta}$ se tiene:

$$-\frac{\partial^2 L(\beta_a)}{\partial \beta \partial \beta^t} = \sum_{i=1}^{n} x_i \left(\frac{e^{x_i^t \cdot \beta} \cdot x_i^t \cdot (1 + e^{x_i^t \cdot \beta}) - e^{x_i^t \cdot \beta} \cdot e^{x_i^t \cdot \beta} \cdot x_i^t}{(1 + e^{x_i^t \cdot \beta})^2}\right) =$$

$$\sum_{i=1}^{n} x_i \left(\frac{e^{x_i^t \cdot \beta} \cdot x_i^t + \cancel{e^{x_i^t \cdot \beta} \cdot e^{x_i^t \cdot \beta} \cdot x_i^t} - \cancel{e^{x_i^t \cdot \beta} \cdot e^{x_i^t \cdot \beta} \cdot x_i^t}}{(1 + e^{x_i^t \cdot \beta})^2}\right) = \sum_{i=1}^{n} x_i x_i^t w_i,$$

siendo $w_i = \frac{e^{x_i^t \cdot \beta}}{(1+e^{x_i^t \cdot \beta})^2} = \frac{e^{x_i^t \cdot \beta}}{(1+e^{x_i^t \cdot \beta})} \cdot \frac{1}{1+e^{x_i^t \cdot \beta}} = p_i(1 - p_i)$.

Sustituyendo en la ecuación

$$\beta_a = \beta + \left(-\frac{\partial^2 L(\beta_a)}{\partial \beta \partial \beta^t}\right)^{-1} \cdot \frac{\partial L(\beta)}{\partial \beta},$$

las expresiones

$$-\frac{\partial^2 L(\beta_a)}{\partial \beta \partial \beta^t} = \sum_{i=1}^{n} x_i x_i^t w_i \qquad y \qquad \frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{n} x_i \cdot (y_i - \hat{p_i}),$$

y evaluando las derivadas en un estimador inicial $\hat{\beta}$ se obtiene el siguiente método iterativo para obtener un nuevo valor del estimador de $\hat{\beta_h}$:

$$\hat{\beta}_h = \hat{\beta}_{h-1} + \left(\sum_{i=1}^{n} x_i x_i^t \hat{w}_i\right)^{-1} \cdot \left(\sum_{i=1}^{n} x_i \cdot (y_i - \hat{p_i})\right),$$

donde $\hat{p_i}$ y $\hat{w_i}$ se calculan con el valor $\hat{\beta_{h-1}}$.

El algoritmo puede escribirse como:

$$\hat{\beta}_h = \hat{\beta}_{h-1} + \left(X^t \hat{W} X\right)^{-1} X^t \left(Y - \hat{Y}\right),$$

donde $\hat{W}$ es una matriz diagonal con términos $\hat{p_i}(1-\hat{p_i})$ y $\hat{Y}$ es el vector de valores esperados de Y.

Escoger $\hat{\beta}_0 = 0$ es un buen punto de partida, aunque la convergencia no está garantizada.

Desviación

Maximizar la verosimilitud puede expresarse como minimizar una función que mide la desviación entre los datos y el modelo. Esa función es $D(\beta) = -2L(\beta)$ y se conoce como desviación (*deviance* en inglés). Atendiendo a la ecuación (3.4) se tenía que:

$$L(\beta) = \sum_{i=1}^{n} \left(y_i \log p_i + (1-y_i)\log(1-p_i)\right),$$

por tanto,

$$D(\beta) = -2\sum_{i=1}^{n} \left(y_i \log p_i + (1-y_i)\log(1-p_i)\right). \tag{3.5}$$

La desviación de cada dato viene dada por: $d_i = -2\left(y_i \log p_i + (1-y_i)\log(1-p_i)\right)$ y mide el ajuste del modelo al dato $(y_i,\mathbf{x}_i)$.

Como los p_i son menores que uno, sus logaritmos son negativos por lo que la desviación es siempre positiva. Además, en el cálculo de la desviación sólo interviene uno de sus dos términos ya que y_i sólo puede valer cero o uno.

Si $y_i = 1$, el segundo término es nulo y $d_i = -2\log p_i$. La observación tiene una desviación grande si la probabilidad estimada de pertenecer al grupo 1 es pequeña, es decir, cuando esta observación está mal explicada por el modelo. En el caso en que $y_i = 0$, sólo interviene el segundo término, $d_i = -2\log(1-p_i)$. Si la probabilidad de pertenecer al grupo 1, p_i, es grande entonces la probabilidad de pertenecer al grupo 0 es pequeña, la desviación es grande y el modelo ajusta mal dicho dato.

Notar que esta función vale cero cuando el ajuste es perfecto: todas las observaciones con $y_i = 1$ tienen $p_i = 1$ y las que tienen $y_i = 0$ tienen $p_i = 0$.

Contrastar si una variable es significativa

Para contrastar si una variable x_i es significativa se puede construir un contraste de la razón de verosimilitudes, comparando los máximos de la función de verosimilitud para los modelos con y sin estas variables.

Si $\beta = (\beta_1,\, \beta_2)$ donde β_1 tiene dimensión $k - s$ y β_2 dimensión s, el contraste es:

$$\begin{cases} H_0 : \beta_2 = 0 \\ H_1 : \beta_2 \neq 0 \end{cases}$$

Para contrastar, se toma como estadístico de contraste $\lambda = -2(L(H_0) - L(H_1))$ siendo $L(H_0)$ el máximo de la función soporte bajo H_0 y siendo $L(H_1)$ el máximo de la función soporte en el valor del estimador máximo verosímil. El estadístico λ tiene aproximadamente una distribución χ^2_s.

Otra forma alternativa para definir el contraste es llamando $D(H_0) = -2L(\hat{\beta}_1)$ a la desviación cuando el modelo se estima bajo H_0, es decir, suponiendo $\beta_2 = 0$ y $D(H_1) = -2L(\hat{\beta}_1, \hat{\beta}_2)$ la desviación bajo H_1. Si H_0 es cierta, la diferencia de desviaciones $D(H_0) - D(H_1) = 2L(\hat{\beta}_1, \hat{\beta}_2) - 2L(\hat{\beta}_1)$ se distribuye como una χ^2_s.

Medida de la bondad de ajuste

Se puede definir una medida global del ajuste con valores entre 0 y 1 basada en la desviación:

$$R^2 = 1 - \frac{D(\hat{\beta})}{D(\beta_0)} = 1 - \frac{L(\hat{\beta})}{L(\beta_0)}.$$

El numerador es la verosimilitud para el modelo con los parámetros estimados y el denominador para el modelo que sólo incluye la constante β_0. En este último caso, la estimación de la probabilidad p_i es constante para todos los datos e igual a $\frac{m}{n}$ donde m es el número de elementos de la muestra pertenecientes al grupo 1 ($y = 1$). Sustituyendo en la ecuación (3.5), la desviación máxima, que corresponde al modelo más simple posible con sólo β_0, es:

$$D(\beta_0) = -2\sum_{i=1}^{n}\Big((\frac{m}{n}\log(\frac{m}{n})) + (1-\frac{m}{n})\log(1-\frac{m}{n})\Big) = -2\sum_{i=1}^{n}\left(\frac{m}{n}\log(\frac{m}{n}) + (1-\frac{m}{n})\log(\frac{n-m}{n})\right) =$$

$$-2\sum_{i=1}^{n}\Big(\frac{m}{n}\log m - \cancel{\frac{m}{n}\log n} + \log(n-m) - \log n - \frac{m}{n}\log(n-m) + \cancel{\frac{m}{n}\log n}\Big) =$$

$$-2m\log m + -2n(1-\frac{m}{n})\log(n-m) + 2n\log n = -2m\log m + -2n(n-m)\log(n-m) + 2n\log n.$$

Si el ajuste es perfecto, se ha comentado que entonces $D(\hat{\beta}) = 0$ luego $R^2 = 1$. Por el contrario, si las variables explicativas no influyen nada, la desviación con las variables explicativas será igual que sin ellas: $D(\hat{\beta}) = D(\beta_0)$ luego $R^2 = 0$.

Notar que $0 \leq R^2 \leq 1$.

3.4. Árboles de clasificación

3.4.1. Introducción

Los árboles de clasificación son un procedimiento alternativo a la regresión logística para clasificar observaciones en grupos predeterminados. Una de sus principales diferencias con respecto a la regresión logística radica en que no utilizan ningún procedimiento estadístico formal para realizar la clasificación. Este método surgió en 1984 de la mano de Breiman y Friedman [19].

La mayor parte del contenido teórico de los árboles de clasificación, ha sido extraída del libro de Daniel Peña, *Análisis de Datos Multivariantes* [20].

3.4.2. Descripción teórica de los árboles de clasificación

Se parte de una muestra de entrenamiento que incluye los valores observados de las variables explicativas $X_1, ..., X_p$ para n individuos y una variable Y indicando la clase a la que pertenece dicha observación. La muestra queda organizada del siguiente modo:

$$X = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1p} \\ x_{21} & x_{22} & ... & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & ... & x_{np} \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Para crear un árbol de clasificación se comienza con un nodo inicial y se debe preguntar cómo dividir el conjunto de datos en dos partes, de modo que esas partes queden lo más homogéneas posible. Para ello, se debe utilizar una de las variables de la muestra y formular una pregunta acerca de esa variable, que permita dividir en dos el conjunto de datos.

Se considera el primer nodo *(nodo raíz)*, y se supone que se ha seleccionado en primer lugar una variable, por ejemplo, x_1, y un punto de corte, c, de manera que se separen los datos con $x_1 \leq c$ de aquellos con $x_1 > c$. De este nodo inicial salen ahora dos nuevos nodos: uno al que llegan las observaciones que verifican $x_1 \leq c$ y otro al que llegan las que verifican $x_1 > c$. En cada uno de estos nodos se vuelve a repetir el proceso anterior: se selecciona una nueva variable y un punto de corte de forma que permitan dividir esa parte de la muestra en otras dos partes lo más homogéneas posible. El proceso termina cuando se hayan clasificado todas las observaciones (o casi todas) correctamente en su grupo.

La construcción del árbol requiere tomar las siguientes decisiones:

- Seleccionar las variables y sus puntos de corte para realizar las divisiones de la muestra.
- Determinar cuándo un nodo se considera terminal y cuándo se continúa dividiendo.
- Asignar los grupos a los nodos terminales.

Un ejemplo de árbol de clasificación sería el que se representa en la Figura 3.2.

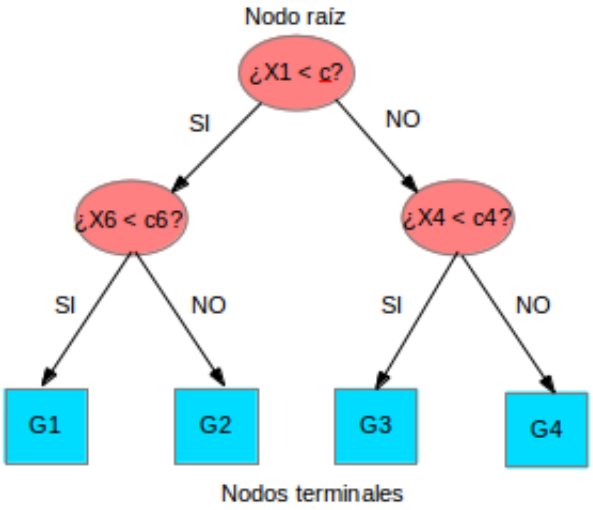


Figura 3.2: Ejemplo de un árbol de clasificación.

Para decidir la variable que va a utilizarse para hacer la partición en un nodo, se calcula en primer lugar la proporción de observaciones que pasan por ese nodo para cada uno de los grupos.

Sea G el número de grupos, $t = 1, ..., T$ cada uno de los nodos y $p(g|t)$ la probabilidad de que las observaciones que llegan al nodo t pertenezcan a cada una de las clases. Se define la **impureza** o la **entropía** de un nodo t como:

$$I(t) = -\sum_{g=1}^{G} p(g|t) \cdot \log p(g|t).$$

Esta medida es no negativa y mide la diversidad. Por ejemplo, si se tienen G grupos y $p(g|t) = 1$ y $p(i|t) = 0$ con $i \neq g$ (todas las observaciones que pasan por el nodo t pertenecen al grupo g) entonces la impureza del nodo t es $I(t) = 0$. En otro caso, la impureza es positiva y alcanza su valor máximo cuando $p(g|t) = \frac{1}{G}$.

Para determinar las variables y las constantes que se deben comparar en cada uno de los nodos, se debe definir un conjunto de preguntas del tipo ¿$x_i < a$? $\quad \forall i = 1, ..., p$ y $a \in (-\infty, \infty)$. Es decir, se deben formular las preguntas para todas las variables y para todas las posibles constantes. Para cada una de las preguntas, se consideran p_S y p_N las proporciones de las observaciones del nodo t que van a los nodos resultantes de responder SI o NO a la pregunta. Las observaciones que respondan SI van a parar al nodo t_S y las que respondan NO al nodo t_N. Si se denota por $I(t_S)$ y $I(t_N)$ a las impurezas resultantes de estos nodos que surgen como respuesta a una de las posibles preguntas, el cambio en la entropía después de esa pregunta es la diferencia entre la entropía del nodo anterior $I(t)$ y la entropía después del nodo que viene dada por $p_S \cdot I(t_S) + p_N \cdot I(t_N)$, luego el cambio en la entropía producido por una pregunta q concreta es:

$$\Delta I(t, q) = I(t) - p_S \cdot I(t_S) - p_N \cdot I(t_N).$$

Por tanto, se escoge para cada nodo, de todas las posibles preguntas aquella que maximice la impureza resultante, pues es esta pregunta la que proporciona los dos grupos más homogéneos posible como resultado de la división.

Una vez alcanzado un nodo terminal, para determinar su valor se escoge aquel del que haya más observaciones que pasen por ese nodo. Es decir, aquel grupo g con máxima $p(g|t)$.

3.5. Reglas de asociación

3.5.1. Introducción

Las reglas de asociación son un método de aprendizaje que se utiliza para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos [23]. Se han investigado ampliamente diversos métodos para aprendizaje de reglas de asociación que han resultado ser muy interesantes para descubrir relaciones entre variables en grandes conjuntos de datos. Piatetsky-Shapiro [24] describe el análisis como la presentación de reglas 'fuertes' descubiertas en bases de datos utilizando diferentes medidas de interés. Basado en el concepto de regla fuerte, Agrawal et al. [25] presentaron un trabajo en el que indicaban las reglas de asociación que descubrían las relaciones entre los datos recopilados a gran escala en los sistemas de terminales de punto de venta de unos supermercados.

La mayor parte del contenido de esta sección ha sido extraída del libro de Robert Tibshirani, *The Elements of Statistical Learning* [22].

3.5.2. Descripción teórica de las reglas de asociación

Para comenzar se debe suponer que se han observado p variables en una muestra aleatoria de tamaño n y que se tiene la muestra organizada en una matriz de datos de dimensiones $n \times p$ del siguiente modo:

$$X_{n\times p} = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1p} \\ x_{21} & x_{22} & ... & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & ... & x_{np} \end{bmatrix}.$$

El objetivo de las reglas de asociación es encontrar valores conjuntos de las variables que aparezcan más frecuentemente en la muestra.

Es una técnica pensada para variables binarias pero también se puede utilizar si la muestra contiene variables categóricas o numéricas. No obstante, estas variables deben ser transformadas previamente logrando convertir la matriz de datos en una matriz de transacciones con tan sólo variables binarias. Es decir,

- Si alguna de las variables es categórica y puede tomar Q posibles valores distintos, se deben crear Q nuevas variables binarias para las cuales el individuo i-ésimo toma el valor 1 en la nueva variable j-ésima si esa categoría definida por la variable j-ésima aparece en la transacción. En otro caso, toma el valor 0.
- Si alguna de las variables es numérica discreta debe ser convertida a variable categórica y se le aplica el mismo procedimiento que a las variables categóricas.
- Si alguna de las variables es numérica continua debe ser troceada en intervalos y se le aplica el mismo procedimiento que a las variables categóricas.

Por tanto, se asume que las p variables de la muestra son binarias ya que si no lo son deben ser transformadas a binarias.

Este análisis también se conoce como el análisis de la cesta de la compra, ya que una de sus aplicaciones más conocidas tiene lugar en supermercados para saber qué objetos se compran a la vez. En este caso, las variables representan los productos a la venta (**items**) y los individuos son **transacciones** (facturas). Si la transacción i-ésima contiene el valor 1 en el producto j-ésimo, esto indica que el elemento j-ésimo ha sido comprado en esa transacción.

Al conjunto de enteros $I \subset \{1, ..., p\}$ se le denota por **itemset** y se define como el conjunto de items que han sido comprados en una transacción, es decir, cuya variable binaria asociada es 1.

Por tanto, se busca un itemset de modo que la probabilidad de que se encuentre ese itemset en la base de datos sea elevada. Es decir, de manera que

$$Pr\left[\bigcap_{j\in I}(X_j = 1)\right] = Pr\left[(\prod_{j\in I} X_j) = 1\right],$$

sea lo suficientemente grande. Se observa que también se puede escribir como un producto, pues que una transacción contenga a un itemset es que todas las variables de ese itemset considerado

tengan el valor 1, luego escribir $(\prod_{j\in I} X_j) = 1$ es lo mismo que decir que se den todos los elementos del itemset en la transacción.

El valor que se toma para estimar esta probabilidad es el número de transacciones de la muestra que contienen ese itemset dividido por el conjunto total de transacciones de la muestra, es decir,

$$\hat{Pr}\left[(\prod_{j\in I} X_j) = 1\right] = \frac{1}{n} \cdot \sum_{i=1}^{n} \prod_{j\in I} x_{ij},$$

donde x_{ij} es el valor de la variable X_j para la i-ésima transacción. Esta expresión se define como el **soporte** del itemset I.

Por tanto, el objetivo consiste en buscar todos los itemsets de la muestra cuyo soporte sea mayor que un cierto valor umbral. No obstante, teniendo en cuenta que el número de itemsets de la muestra es 2^p, si p es grande es necesario algún algoritmo que simplifique esta búsqueda. El más utilizado es el algoritmo *Apriori*. Para simplificar el problema, este algoritmo tiene en cuenta los siguientes aspectos:

- El número de itemsets cuyo soporte es mayor que un cierto umbral es relativamente pequeño.
- *Propiedad del soporte:* Todo subconjunto de un itemset tiene un soporte mayor o igual que el de ese itemset.

El algoritmo *Apriori* realiza una búsqueda en anchura para encontrar los itemsets ejecutando los siguientes pasos:

Paso 1: Se calcula el soporte para todos los items individuales. Aquellos con un soporte menor que un cierto soporte dado se descartan.

Pasos desde 2 hasta p: Se calculan todos los conjuntos de items de cardinal correspondiente al paso en el que nos encontramos combinando los conjuntos de items que han sobrevivido al paso anterior con los que fueron conservados en el Paso 1.

Paso $p+1$: Para cada itemset I que haya sobrevivido a todos los pasos anteriores se subdivide el itemset en dos conjuntos disjuntos de modo que $A \cup B = I$ y $card(B) = 1$ y se escribe $A \Rightarrow B$ donde A recibe el nombre de antecedente y B el nombre de consecuente.

A continuación, se definen algunas medidas de bondad de la regla:

♦ Se denota por **soporte** ($supp(A \Rightarrow B) = supp(A \cup B)$) de la regla $A \Rightarrow B$ al porcentaje de observaciones de la muestra en las que se da tanto A como B, que es justamente el soporte del itemset I, como se había definido anteriormente. Se puede ver como $Pr(A \cap B)$.

♦ Se define la **confianza** de una regla como la proporción de datos para los que la regla es cierta, es decir,

$$conf(A \Rightarrow B) = \frac{supp(A \cup B)}{supp(A)}.$$

Se puede ver como $Pr(B|A)$.

♦ Se denomina **confianza esperada** al soporte de B que es una estimación de $Pr(B)$

♦ Se llama **lift** a la confianza de la regla dividida por la confianza esperada:

$$lift(A \Rightarrow B) = \frac{conf(A \Rightarrow B)}{supp(B)} = \frac{supp(A \cup B)}{supp(A) \cdot supp(B)}.$$

Es una medida de asociación entre A y B. Se puede interpretar como el porcentaje de mejora. Notar que si A y B son independientes entonces:

$$lift(A \Rightarrow B) = \frac{supp(A \cup B)}{supp(A) \cdot supp(B)} = \frac{\cancel{supp(A)} \cdot \cancel{supp(B)}}{\cancel{supp(A)} \cdot \cancel{supp(B)}} = 1.$$

Por tanto, una vez finalizado el paso $p+1$ del algoritmo *Apriori*, se calcula la confianza de cada una de las reglas y se devuelven únicamente las reglas cuya confianza es mayor que un cierto umbral.

El soporte y la confianza de las reglas que se deseen crear deben ser determinados por el usuario.

Uno de los inconvenientes del algoritmo *Apriori* es que si el soporte escogido es muy bajo, el número de itemsets y su tamaño puede crecer de manera exponencial y puede hacer que el problema sea inviable, por lo que las reglas con una confianza y un lift muy altos nunca podrán ser descubiertas. Por ejemplo la regla *vodka* $\Rightarrow$ *caviar* no podrá ser descubierta debido a las escasas ventas de caviar.

Capítulo 4

Resultados obtenidos

4.1. Análisis realizados sobre el abandono universitario

En primer lugar, se detallan los análisis realizados de manera global (para todos los centros académicos en conjunto) y, a continuación, se muestran algunos de los análisis separados por centros. Los análisis detallados por centros son la regresión logística y el análisis clúster.

4.1.1. Creación de un modelo global

Análisis clúster

La primera técnica de minería de datos utilizada para explicar el abandono ha sido el análisis clúster. Este análisis se ha realizado únicamente sobre los alumnos que han abandonado los estudios para **encontrar tipologías de alumnos que abandonan**.

El algoritmo utilizado para llevar a cabo el análisis clúster ha sido el de **k-medoides** porque la muestra contiene tanto variables numéricas como variables categóricas como se puede observar en la Tabla 2.1. Por este motivo, se ha considerado que este algoritmo con el empleo de la distancia de *"Gower"* era sin duda el más adecuado para aplicar a la muestra.

Para la **elección del número de grupos**, se ha representado el valor de la función objetivo (3.1) variando desde dos hasta ocho grupos y se ha buscado un cambio de pendiente que indica una disminución brusca del valor de la función objetivo. El gráfico obtenido es el que se representa en la Figura 4.1. En este gráfico se observa que la mejor opción es crear tres grupos distintos de estudiantes que abandonan los estudios.

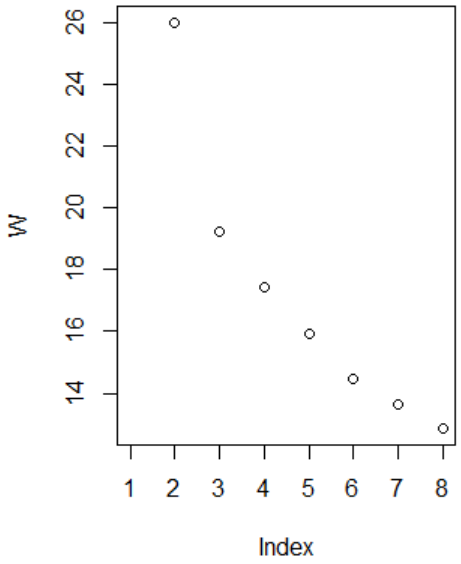


Figura 4.1: Representación de la función objetivo de 2 a 8 grupos.

Una vez elegido el número de grupos se ha procedido a lanzar el análisis de forma reiterada y desde distintos puntos aleatorios, eligiendo la solución con el menor valor de la función objetivo para lograr obtener el mínimo global. A continuación, se han mostrado los valores de los **representantes de cada uno de los grupos**, como se puede ver en la Figura 4.2, que permiten observar en qué variables hay una mayor diferencia entre los tres grupos.

```
     curso.abandono edad nota.de.acceso orden.pref
[1,]              1   23      0.9181641          2
[2,]              3   19      0.9618608          2
[3,]              1   21      1.1527812          2
     via.acc sexo prov cred.pres.pri cred.honor.pri
[1,]       3    2    1             6              0
[2,]       3    1    1            39              0
[3,]       3    2    1            54              0
     cred.pres.ultimo num.asi.rep.ultimo
[1,]                6                  0
[2,]               30                  4
[3,]               54                  0
     cred.honor.ultimo cred.sup.exam.media trabPri
[1,]                 0                   0       1
[2,]                 0                  13       1
[3,]                 0                  42       1
     trabUltimo
[1,]          1
[2,]          1
[3,]          1
```

Figura 4.2: Representantes de cada uno de los grupos.

A simple vista se observan diferencias en la edad, la nota de acceso, el número de créditos presentados a examen en primer y último curso y el número medio de créditos superados en los cursos que los alumnos han hecho matrícula.

Para analizar en detalle cada uno de los grupos obtenidos, se han realizado **gráficos mostrando los valores que toma cada una de las variables utilizadas en el análisis en función de los distintos grupos**. De este modo, observando todos los gráficos conjuntamente para un mismo grupo, se puede saber cuáles son las características de los alumnos que forman cada uno de los grupos.

Observando los gráficos se pueden determinar las características de cada grupo, y ver cuáles son las variables que diferencian a ese grupo del resto, para intentar definir varios patrones que se verifiquen en los alumnos que abandonan.

Para saber en qué variables hay diferencias significativas entre los distintos grupos, se han utilizado ANOVAS para las variables numéricas y contrastes de la *Chi-cuadrado* para las variables categóricas. Aunque el test ANOVA no se puede tomar como exacto, ayuda a afirmar si hay diferencias en cada variable para los distintos grupos.

Los *p-valores* obtenidos a partir de los contrastes se pueden observar en la Tabla 4.1.

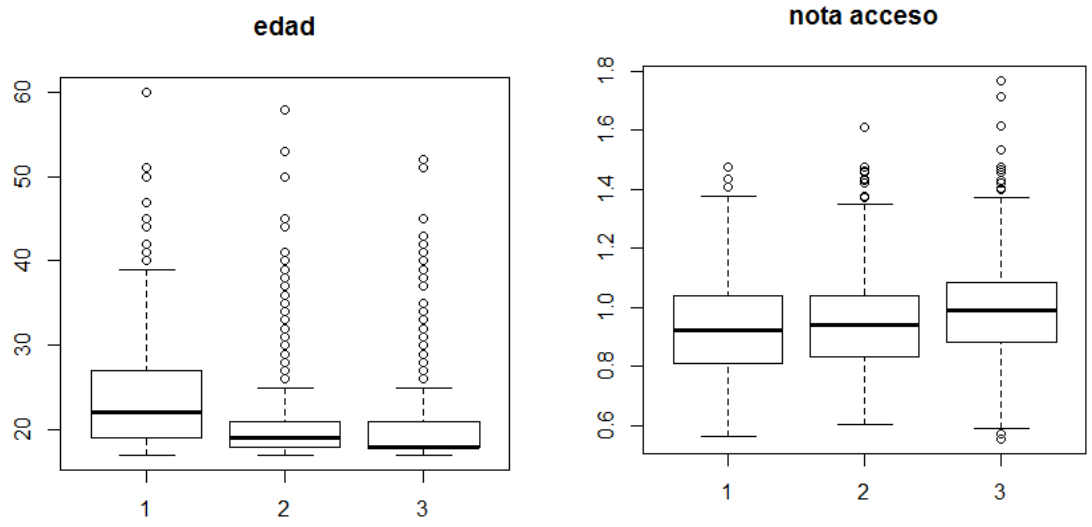
edad	nota.de.acceso	cred.pres.pri	cred.pres.ultimo	cred.honor.pri	cred.sup.exam.media
2.2e-16	7.529e-11	2-2e-16	2.2e-16	7.289e-05	2.2e-16

num.asi.rep.ult	via.acc	curso.abandono	orden.pref	sexo	trabPri	trabUltimo	prov
1.834e-06	1.9e-05	2.2e-16	0.02611	4.7e-08	5.3e-11	1.552e-12	1.06e-06

Tabla 4.1: *p-valores* de los contrastes ANOVA y *Chi-cuadrado*.

Tomando el nivel de significación habitual (0.05), se puede afirmar que hay diferencias significativas en todas las variables.

Se puede observar la distribución de la edad y de la nota de acceso en la Figura 4.3.



(a) Edad

(b) Nota de acceso

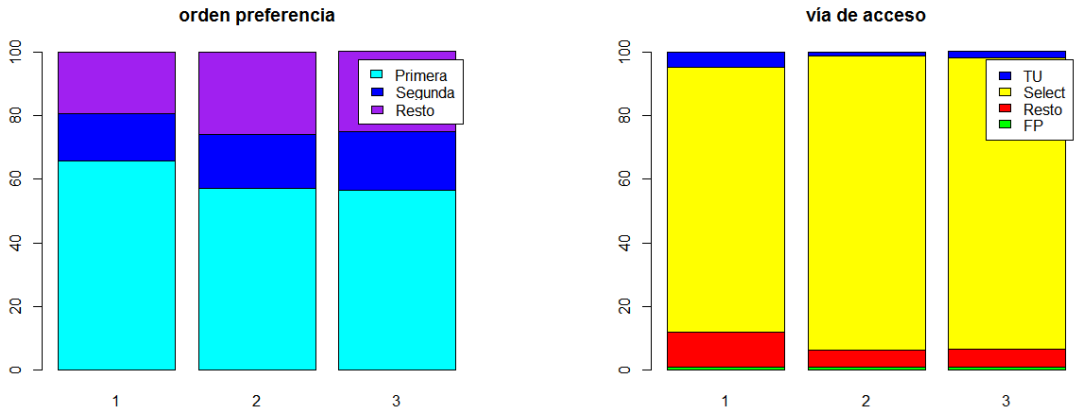
Figura 4.3: Edad y nota de acceso, en función de cada uno de los grupos.

En estos gráficos se muestra para cada uno de los tres grupos cuál es la distribución tanto de la edad como de la nota de acceso.

Se observa que los alumnos del primer grupo son mayores que los del segundo grupo, y éstos a su vez mayores que los del tercer grupo.

La nota de acceso de los alumnos de los distintos grupos es similar, excepto para los del tercer grupo, ya que éstos tienen una nota de acceso ligeramente superior.

Se puede observar la distribución del orden de preferencia y de la vía de acceso en la Figura 4.4. Para interpretar estos gráficos, se debe considerar cada barra como un porcentaje sobre el 100 % de los alumnos de ese grupo. Por ejemplo, en el gráfico del orden de preferencia, el 70 % de los alumnos del primer grupo han entrado en una titulación como primera opción, el 12 % como segunda opción y el 18 % como más de segunda opción.



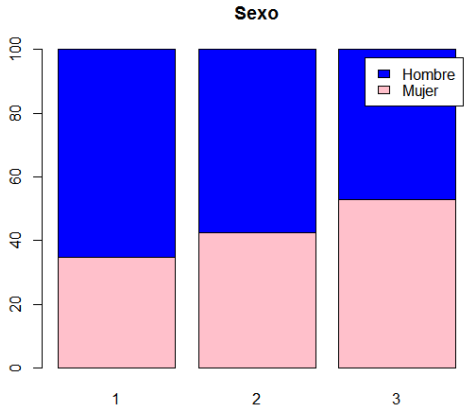
(a) Orden de preferencia

(b) Vía de acceso

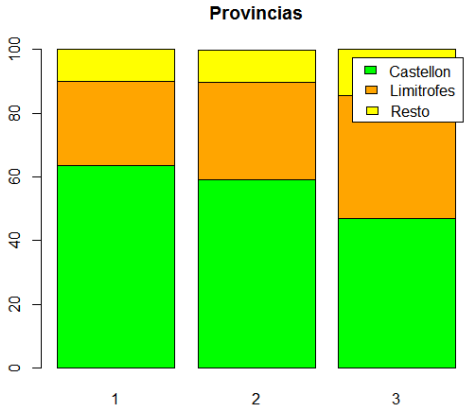
Figura 4.4: Orden de preferencia y vía de acceso, en función de cada uno de los grupos. En el gráfico del orden de preferencia se representa en **azul claro** la primera opción, en **azul oscuro** la segunda opción y en **morado** una opción superior. En el gráfico de la vía de acceso, el color **amarillo** es para los alumnos que han accedido a la universidad a través de Selectividad, el color **azul** para los Titulados Universitarios, el color **verde** para los alumnos de Formación Profesional y el color **rojo** para el Resto.

No se observan prácticamente diferencias en el orden de preferencia entre los distintos grupos aunque en el primer grupo hay más alumnos que entran como primera opción que en el resto de grupos. También se puede apreciar que muchos de los alumnos del primer grupo entran por vías de acceso como son Titulados Universitarios y el resto de vías de acceso que no son Formación Profesional ni Selectividad, en comparación el resto grupos.

Se puede observar la distribución del sexo y de la provincia de residencia familiar en la Figura 4.5.



(a) Sexo



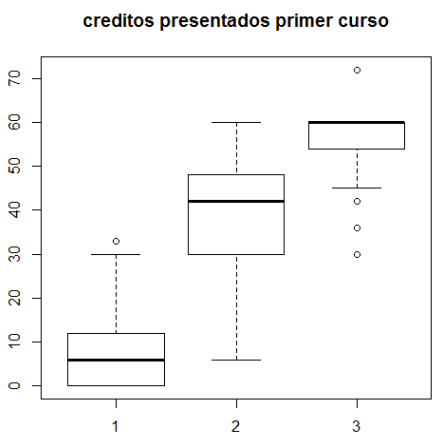
(b) Provincia de residencia familiar

Figura 4.5: Sexo y provincia de residencia familiar en función de cada uno de los grupos. En el gráfico del sexo, se representa en **azul** a los hombres y en **rosa** a las mujeres. En el gráfico de la provincia de residencia familiar, el color **verde** es para los alumnos que pertenecen a la provincia de Castellón, el color **naranja** para los alumnos que provienen de provincias limítrofes a la provincia de Castellón y el color **amarillo** para el resto.

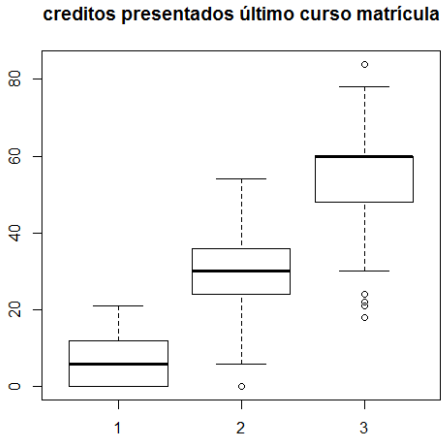
En el primer grupo hay el doble de hombres que de mujeres mientras que en el tercer grupo hay tantas mujeres como hombres.

La provincia de residencia familiar también presenta diferencias en el tercer grupo pues en ese grupo hay más alumnos que abandonan pertenecientes a provincias limítrofes a Castellón que en el resto de grupos.

Se puede observar la representación del número de créditos presentados en el primer y último curso de matrícula en la Figura 4.6.



(a) Créditos presentados primer curso

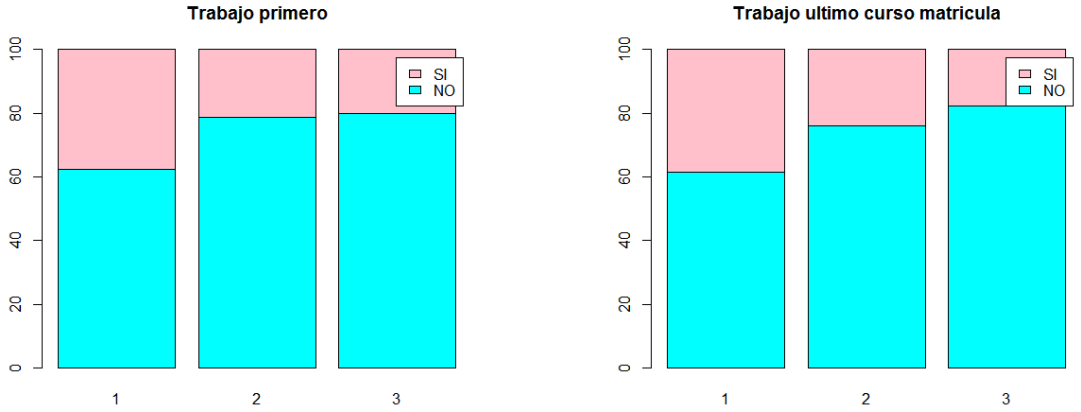


(b) Créditos presentados último curso

Figura 4.6: Créditos presentados en el primer y último curso de matrícula, en función de cada uno de los grupos.

El número de créditos presentados en primer y último curso de matrícula varía enormemente de un grupo a otro. Los estudiantes del primer grupo no se presentan a examen de casi ningún crédito, los estudiantes del segundo grupo se presentan a más exámenes que los anteriores, mientras que los alumnos del tercer grupo se presentan a prácticamente todos los exámenes.

Se puede observar la distribución del trabajo realizado en el primer y último curso de matrícula en la Figura 4.7.



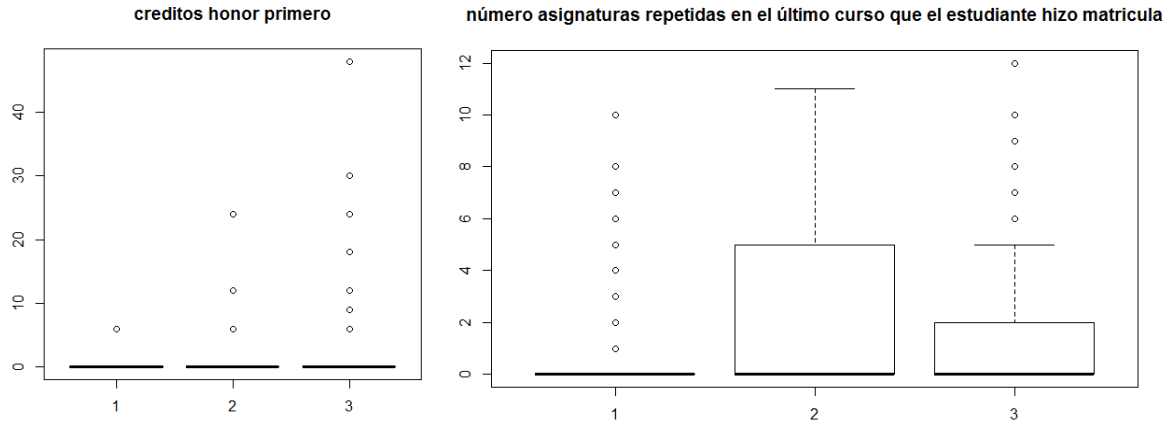
(a) Trabajo primer curso

(b) Trabajo último curso

Figura 4.7: Trabajo realizado en el primer y último curso de matrícula, en función de cada uno de los grupos. De color **rosa** se representa a los alumnos que trabajan y de color **azul** a los alumnos que no trabajan.

El trabajo realizado en primer y último curso también es diferente en función de los grupos, puesto que en el primer grupo hay una mayor proporción de alumnos que trabaja que en el resto de grupos.

Se puede observar la distribución del número de créditos de honor obtenidos en primer curso y del número de asignaturas repetidas en el curso anterior al abandono en la Figura 4.8.



(a) Créditos de honor primer curso

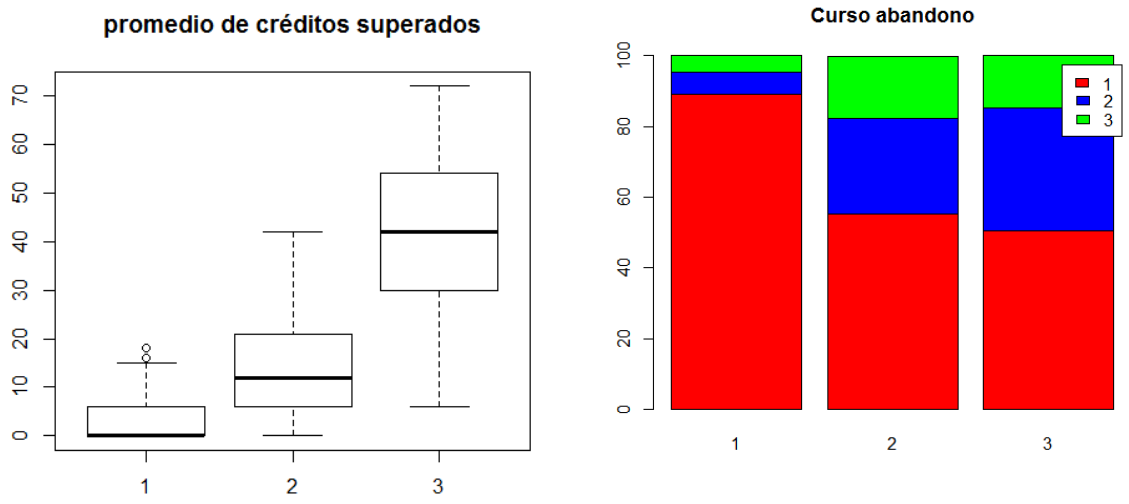
(b) Número de asignaturas repetidas último curso

Figura 4.8: Créditos de honor obtenidos en primer curso y número de asignaturas repetidas en el curso anterior al abandono, en función de cada uno de los grupos.

El número de créditos de honor también varía en función de los grupos. Los alumnos del tercer grupo obtienen más créditos de honor que los alumnos del segundo grupo, y éstos, a su vez, más créditos que los del primer grupo.

Observando el gráfico del número de asignaturas repetidas en el último curso de matrícula, se puede observar que la mayor parte de los alumnos que tienen asignaturas que están repitiendo son clasificados en el segundo grupo.

Se puede observar la distribución del promedio de créditos superados y del curso de abandono en la Figura 4.9.



(a) Créditos superados

(b) Curso abandono

Figura 4.9: Promedio de créditos superados y curso de abandono, en función de cada uno de los grupos. En el gráfico del curso de abandono, se ha representado a los alumnos que abandonan en primer curso de color **rojo**, a los que abandonan en segundo curso de color **azul** y a los que abandonan en tercer curso de color **verde**.

El número de créditos superados en exámenes también varía dependiendo de cada grupo.

Con respecto al curso de abandono, la mayor parte de los alumnos del primer grupo abandona en primer curso mientras que en el resto de grupos también abandonan tanto en segundo como en tercer curso.

En base a todas las características que se han podido determinar para cada grupo, se han encontrado tres tipologías de alumnos que abandonan los estudios:

- *Estudiantes de tipo 1.* Son en **mayor parte hombres**. Muchos de ellos **trabajan** y suelen **abandonar durante el primer año** de matrícula, pues se presentan a examen de muy pocos créditos y no superan prácticamente créditos en exámenes. Evidentemente, no tienen créditos de honor y muchos de ellos, en comparación con los otros grupos, entran por **vías de acceso que no son Selectividad ni Formación Profesional.**

- *Estudiantes de tipo 2*. Siguen siendo en **mayor parte hombres**, pero en este caso también hay muchas mujeres. Son, en general, **más jóvenes que los anteriores**, y la mayor parte no trabaja. Obtienen mejores resultados académicos que los del grupo anterior (**se presentan a un mayor número de exámenes, aunque no superan muchos exámenes**).

- *Estudiantes de tipo 3*. En este grupo hay **tantas mujeres como hombres**. La **nota de acceso** es **ligeramente superior** y **muchos de ellos pertenecen a provincias limítrofes a Castellón**. Además, **se presentan y superan bastantes más créditos** que el resto de grupos. Obtienen más **créditos de honor**, y no suelen realizar **ningún trabajo**. Aunque la mayor parte abandona en primer curso, también hay muchos alumnos de este grupo que abandonan en segundo o tercer curso.

Como medida de bondad de ajuste de este método, se ha utilizado la silueta, que ha proporcionado un valor de 0.4; por tanto, la agrupación realizada es bastante aceptable.

Regresión logística

Para realizar el análisis de regresión logística se ha tomado la información de los alumnos que abandonan y de los que no abandonan.

El objetivo principal de este análisis es encontrar qué variables explican el abandono de los estudios. De todas las variables que se muestran en la Tabla 2.1, únicamente se han incluido en el análisis de regresión logística las variables que, después de realizar un estudio previo mediante gráficos univariantes y contrastes de hipótesis, han resultado significativas en el abandono. No obstante, cabe destacar que aunque utilizando gráficos y contrastes univariantes se puede conocer qué variables influyen en el abandono, siempre es conveniente realizar análisis multivariantes, como es el caso de la regresión logística. Esto es debido a que al considerar todas las variables en conjunto, la influencia de algunas variables será inferior a otras y algunas de ellas, aunque presenten diferencias entre abandonar y no abandonar los estudios, no se considerarán como variables que explican el abandono.

Como la variable *orden.pref* es una variable categórica que contiene nulos, para que el modelo no tome los valores nulos como un nuevo valor, se han creado dos nuevas variables auxiliares que no aparecen en la Tabla 2.1 que son *orden.prefSegunda* y *orden.prefResto* conteniendo un 1 si la opción escogida coincide con el nombre de la variable, o un 0 si la preferencia no coincide con el nombre de la variable. De este modo, se han podido mantener los nulos y que el modelo los interprete como valores faltantes.

Los resultados de la regresión logística se muestran en la Figura 4.10.

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.99211  -0.39804  -0.09245   0.23436   2.79796

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -1.696124   0.710034  -2.389 0.016904 *
edad                             0.038454   0.013560   2.836 0.004570 **
nota.de.acceso                   2.016466   0.335877   6.004 1.93e-09 ***
as.factor(orden.prefSegunda)1    0.195150   0.145866   1.338 0.180938
as.factor(orden.prefResto)1      0.728216   0.141263   5.155 2.54e-07 ***
via.accResto                     1.374293   0.524941   2.618 0.008845 **
via.accSelectividad              1.205834   0.470179   2.565 0.010329 *
via.accTitulados Universitarios  2.187993   0.594102   3.683 0.000231 ***
sexoHome                         0.261242   0.103621   2.521 0.011698 *
provLimitrofes                   0.636504   0.124537   5.111 3.21e-07 ***
provResta                        0.546885   0.169014   3.236 0.001213 **
cred.pres.pri                    0.063871   0.003801  16.804  < 2e-16 ***
cred.honor.pri                  -0.044914   0.020934  -2.146 0.031909 *
cred.pres.ultimo                -0.006540   0.004321  -1.514 0.130119
cred.honor.ultimo               -2.178343  45.605687  -0.048 0.961904
num.asi.rep.ultimo              -0.302436   0.024053 -12.574  < 2e-16 ***
cred.sup.exam.media             -0.148705   0.005838 -25.472  < 2e-16 ***
as.factor(trabPri)1              0.170137   0.160265   1.062 0.288418
as.factor(trabUltimo)1           0.190236   0.160659   1.184 0.236375
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6276.2  on 5039  degrees of freedom
Residual deviance: 2601.7  on 5021  degrees of freedom
  (851 observations deleted due to missingness)
AIC: 2639.7

Number of Fisher Scoring iterations: 15
```

Figura 4.10: Resultado de la regresión logística.

Las variables numéricas que aparecen con un coeficiente positivo *(columna estimate)* indican que, a mayor valor en esa variable, mayor probabilidad de abandono y las que aparecen con signo negativo indican lo contrario, a menor valor en la variable, mayor probabilidad de abandono. Por ejemplo, la edad tiene un coeficiente de 0.038454, luego esto indica que aquellos alumnos que son mayores tienen mayor probabilidad de abandonar los estudios.

Por otro lado, para analizar los coeficientes de las variables categóricas, se deben interpretar los coeficientes con respecto a la categoría que no se muestra de esa variable. Por ejemplo, el trabajo realizado en primer curso, tiene un coeficiente de 0.1701337. Al ser positivo indica que trabajar hace más probable el abandono con respecto al otro valor que puede tomar la variable que en este ejemplo sería no trabajar.

Observando el *p-valor (columna Pr)*, se pueden determinar las variables que resultan significativas en el abandono de los estudios. Si se escoge el nivel de significación habitual (0.05), las variables significativas son las que tienen al menos un asterisco. A mayor número de asteriscos, más relevante resulta esa variable en el abandono de los estudios. Notar que si la variable es categórica y alguna de sus categorías es significativa, también es significativa la variable. Por tanto, las **variables que explican el abandono** son las siguientes:

- **La edad**: a mayor edad, mayor probabilidad de abandono.
- La **nota de acceso**.
- **El orden de preferencia de la titulación**: entrar en una titulación escogida como más de segunda opción hace que la probabilidad de abandono aumente con respecto a los estudiantes que entran en una titulación como primera opción.
- **La vía de acceso**: el abandono también depende de la vía de acceso. A diferencia de lo que se suele pensar, la vía de acceso que tiene menos probabilidad de abandono es Formación Profesional y la que más, Titulado Universitario.
- **El sexo**: ser hombre aumenta la probabilidad de acabar abandonando los estudios.
- **La provincia de residencia familiar**: los estudiantes que menos abandonan son los que pertenecen a la provincia de Castellón, mientras que los que vienen de fuera de Castellón tienen más probabilidad de abandonar los estudios, principalmente los que pertenecen a alguna provincia limítrofe a Castellón.
- **El número de créditos presentados en primer curso**: a mayor número de créditos presentados en primer curso, mayor probabilidad de abandono. En este caso no se obtiene un resultado demasiado lógico. No obstante, si se realiza un contrate de hipótesis de la varianza, se puede afirmar que hay una gran variabilidad en esta variable y este hecho puede estar influyendo bastante en el estudio.
- **Los créditos de honor obtenidos en el primer curso de matrícula**: obtener créditos de honor disminuye la probabilidad de abandono.
- **Repetir asignaturas en el curso anterior al abandono**: repetir asignaturas disminuye la probabilidad de abandono. Este resultado es lógico pues la mayor parte de los alumnos que abandonan lo hacen en primer curso, y al no volverse a matricular no repiten asignaturas.
- **La media de créditos superados en exámenes**: superar créditos disminuye la probabilidad de abandono.

Además de determinar las variables que explican el abandono, también se ha probado este modelo como modelo de predicción para averiguar si un alumno abandonará, o no, los estudios que está cursando. Este dato puede ser interesante para la universidad, porque permite conocer a priori el número de alumnos que se matriculará en el siguiente curso académico. Se ha probado el modelo utilizando una técnica denominada validación cruzada, que consiste en seleccionar una parte de la muestra que se utiliza como entrenamiento para crear el modelo, y predecir la otra parte de la muestra para ver cómo funciona el modelo de predicción creado. La bondad obtenida ha sido de 0.89 utilizando validación cruzada lo que indica que casi el 90 % de los alumnos de la muestra han sido predichos correctamente. No obstante, si se quiere poder predecir el número de alumnos que abandonará alguna de las titulaciones, sin que hayan realizado ningún curso académico todavía, se puede crear otro modelo de regresión logística sin hacer uso de las variables académicas en el grado. La bondad de este modelo utilizando validación cruzada ha sido de 0.71, que no está nada mal considerando que no se tiene ningún dato académico exceptuando la nota de acceso.

Análisis discriminante

Se ha realizado análisis discriminante para comparar los resultados obtenidos con los de la regresión logística. Los resultados han sido idénticos y la bondad del modelo, realizando validación cruzada, ha sido de 0.88.

Redes neuronales

Las redes neuronales también se han empleado para realizar clasificación y determinar si un alumno abandonará, o no, los estudios.

En este caso, se ha realizado un *script* variando el número de neuronas ocultas entre 5 y 41 (el máximo que permite el programa para este problema) y se ha obtenido la red neuronal que mejor predice a los alumnos de la muestra. De nuevo esta bondad ha sido de 0.89.

Comparación entre los métodos de predicción utilizados:

Como la bondad de ajuste obtenida ha sido idéntica con los tres métodos de predicción utilizados, se ha decidido presentar al *Gabinete de Planificación y Prospectiva Tecnológica* los resultados obtenidos mediante regresión logística porque este modelo también ofrece la información de qué variables explican el abandono universitario. Por este motivo, sólo se ha fundamentado teóricamente el modelo de regresión logística.

Árboles de clasificación

Se ha hecho uso de árboles de clasificación para clasificar a los alumnos en función de si han abandonado, o no, los estudios. De este modo se ha podido observar qué variables se comprueban y en qué orden a la hora de clasificar a los alumnos en un grupo u otro. Se puede observar uno de los árboles obtenidos en la Figura 4.11.

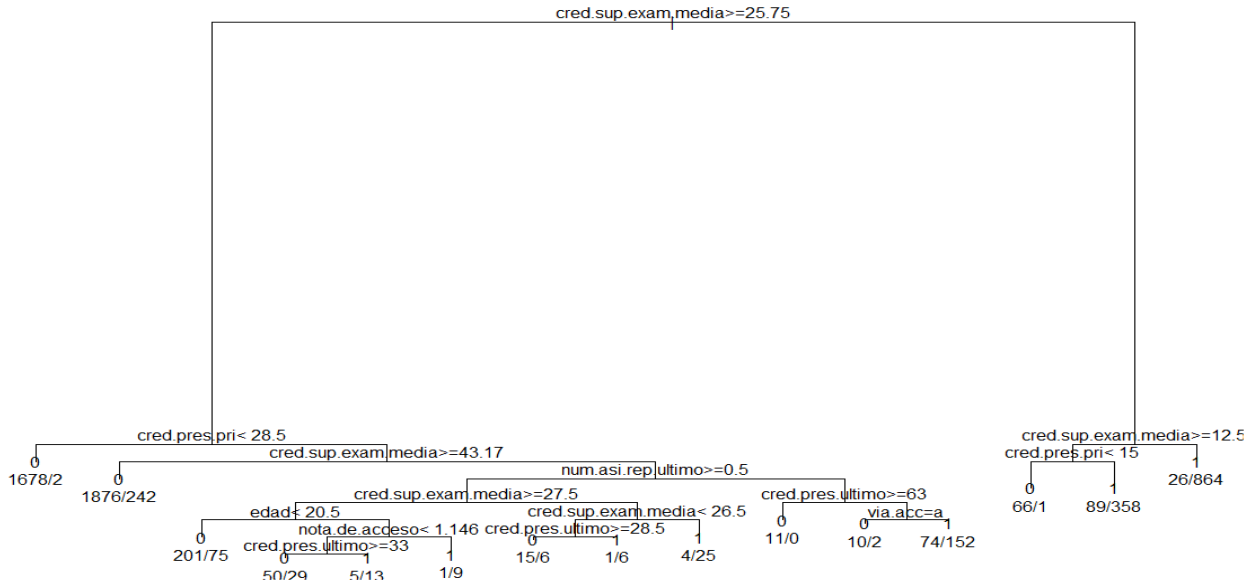


Figura 4.11: Árbol de clasificación que determina si los alumnos han abandonado, o no, los estudios.

Para interpretar el árbol hay que situarse en la parte superior y evaluar la primera condición. Si la condición se verifica se debe descender por la rama izquierda y comprobar la siguiente condición. En otro caso, se debe descender por la rama derecha. Se debe ir descendiendo por el árbol hasta que finalmente se alcance un nodo terminal cuyo valor será 0 ó 1. Si el valor es 0 quiere decir que el estudiante se clasifica en el grupo de los alumnos que no han abandonado los estudios y si el valor es 1 en el grupo de los que sí han abandonado los estudios.

Debajo de cada rama se muestran dos números separados por una barra. El primer valor representa el número de estudiantes de la muestra que han sido bien clasificados siguiendo esa rama mientras que el segundo número pertenece a los estudiantes que han sido mal clasificados. Esto permite calcular la bondad de ajuste de la clasificación.

Observando el árbol obtenido, la primera variable que se evalúa, es decir, la variable en la que hay más diferencias, es la media de créditos superados en exámenes. Si el promedio de créditos superados en exámenes es superior o igual a 25.75, la siguiente condición que se evalúa es si el número de créditos presentados a examen en primer curso es menor que 28.5. En ese caso, se clasifica al alumno en el grupo de los estudiantes que no abandonan los estudios. Si no se verifica dicha condición, se comprueba si la media de créditos superados en exámenes es mayor o igual que 43.17. Si lo es, se clasifica al alumno en el grupo de los estudiantes que no abandonan los estudios mientras que si no lo es, la siguiente variable a evaluar es el número de asignaturas repetidas en el último curso de matrícula. De este modo, se va descendiendo por el árbol hasta llegar a alguno de los nodos terminales donde el alumno queda clasificado en uno de los dos grupos: abandono o no abandono.

En este ejemplo, se puede observar que hay bastantes individuos que no están correctamente clasificados. Para lograr un mejor ajuste se debe hacer crecer el árbol. Sin embargo, en ese caso, las ramas del árbol se solapan al realizar el gráfico y no se puede observar qué condiciones son las que se evalúan para clasificar a un estudiante.

4.1.2. Análisis por centros académicos

Los análisis que se han realizado por cada uno de los centros son la regresión logística y el análisis clúster.

Regresión logística

Para realizar el análisis de regresión logística, por cada centro académico se han escogido únicamente las variables que, realizando contrastes univariantes, son significativas en el abandono de los estudios. El estudio de qué variables son significativas se encuentra en el Anexo A de este documento.

★ Escuela Superior de Tecnología y Ciencias Experimentales

Los resultados obtenidos para los alumnos de la Escuela Superior de Tecnología y Ciencias Experimentales se muestran en la Figura 4.12.

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.00739  -0.29608  -0.06646   0.35699   3.08120

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -0.114168   0.891751  -0.128 0.898128
edad                           0.011423   0.025489   0.448 0.654047
nota.de.acceso                 1.814633   0.677492   2.678 0.007396 **
as.factor(orden.prefSegunda)1 -0.083553   0.307310  -0.272 0.785712
as.factor(orden.prefResto)1    1.260131   0.356678   3.533 0.000411 ***
sexoHome                       0.223067   0.252838   0.882 0.377641
provLimitrofes                 1.429790   0.309569   4.619 3.86e-06 ***
provResta                      0.623715   0.386269   1.615 0.106373
cred.pres.pri                  0.053576   0.006515   8.224  < 2e-16 ***
cred.honor.pri                -0.067226   0.088168  -0.762 0.445776
cred.pres.ultimo              -0.005740   0.010400  -0.552 0.580987
cred.sup.exam.media           -0.166433   0.013732 -12.120  < 2e-16 ***
as.factor(trabPri)1            0.202245   0.363227   0.557 0.577663
as.factor(trabUltimo)1         0.301054   0.350931   0.858 0.390962
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1526.00  on 1130  degrees of freedom
Residual deviance:  579.42  on 1117  degrees of freedom
  (194 observations deleted due to missingness)
AIC: 607.42

Number of Fisher Scoring iterations: 7
```

Figura 4.12: Análisis de regresión logística para los alumnos de la Escuela Superior de Tecnología y Ciencias Experimentales.

Las **variables que explican el abandono** de los estudios en la Escuela Superior de Tecnología y Ciencias Experimentales son:

- La **nota de acceso**.
- Cursar una **titulación escogida como más de segunda opción**: aumenta la probabilidad de abandono con respecto a escoger una titulación como primera opción.
- Pertenecer a una **provincia limítrofe a Castellón**: aumenta la probabilidad de abandono con respecto a pertenecer al resto de provincias.
- El número de **créditos presentados en primer curso**.
- La **media de créditos superados en exámenes**: superar exámenes disminuye la probabilidad de abandono.

★ Facultad de Ciencias Humanas y Sociales

Los resultados obtenidos para la Facultad de Ciencias Humanas y Sociales se muestran en la Figura 4.13.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3287  -0.3795  -0.1044  -0.0294   2.7279

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -4.072570   1.381479  -2.948 0.003199 **
edad                            0.081174   0.023817   3.408 0.000654 ***
nota.de.acceso                  3.735088   0.713050   5.238 1.62e-07 ***
as.factor(orden.prefSegunda)1   0.341489   0.249289   1.370 0.170733
as.factor(orden.prefResto)1     1.112620   0.264318   4.209 2.56e-05 ***
via.accResto                   -0.140943   0.887881  -0.159 0.873873
via.accSelectividad            -0.294637   0.790215  -0.373 0.709255
via.accTitulados Universitarios 0.827523   0.942423   0.878 0.379900
sexoHome                        0.183211   0.194018   0.944 0.345016
provLimitrofes                  0.523282   0.210710   2.483 0.013012 *
provResta                       0.328897   0.322931   1.018 0.308453
cred.pres.pri                   0.067362   0.007562   8.908  < 2e-16 ***
cred.honor.pri                 -0.035316   0.032815  -1.076 0.281823
cred.pres.ultimo               -0.026048   0.009657  -2.697 0.006991 **
num.asi.matric.ultimo           0.210584   0.057697   3.650 0.000262 ***
cred.honor.ultimo              -2.369497  88.651322  -0.027 0.978676
num.asi.rep.ultimo             -0.379785   0.060382  -6.290 3.18e-10 ***
cred.sup.exam.media            -0.142621   0.011924 -11.961  < 2e-16 ***
as.factor(trabPri)1             0.214136   0.282633   0.758 0.448662
as.factor(trabUltimo)1          0.072635   0.285351   0.255 0.799075
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1959.36  on 1924  degrees of freedom
Residual deviance:  843.99  on 1905  degrees of freedom
  (367 observations deleted due to missingness)
AIC: 883.99

Number of Fisher Scoring iterations: 16
```

Figura 4.13: Análisis de regresión logística para los alumnos de la Facultad de Ciencias Humanas y Sociales.

Las **variables que explican el abandono** de los estudios en la Facultad de Ciencias Humanas y Sociales son:

- La **edad**: a mayor edad mayor probabilidad de abandono.
- La **nota de acceso**.
- El **orden de preferencia**: cursar una titulación escogida como más de segunda opción aumenta la probabilidad de abandono.
- Pertenecer a una **provincia limítrofe a Castellón**: aumenta la probabilidad de abandono.

- El número de **créditos presentados en primer curso**.
- El número de **créditos presentados en el último curso de matrícula**: cuantos más créditos presentados menor es la probabilidad de abandono.
- El **número de asignaturas matriculadas en el último curso de matrícula**: matricularse de muchas asignaturas aumenta la probabilidad de abandono.
- El **número de asignaturas repetidas en el último curso de matrícula**: tener más asignaturas repetidas disminuye la probabilidad de abandono.
- La media de **créditos superados en exámenes**: superar créditos en exámenes disminuye la probabilidad de abandono.

★ Facultad de Ciencias Jurídicas y Económicas

Los resultados obtenidos para los alumnos de la Facultad de Ciencias Jurídicas y Económicas se muestran en la Figura 4.14.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3961  -0.2716  -0.0521   0.2101   2.9589

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -5.74270    1.73449  -3.311  0.00093 ***
edad                             0.04458    0.03068   1.453  0.14627
nota.de.acceso                   2.03838    0.75420   2.703  0.00688 **
as.factor(orden.prefSegunda)1    0.03232    0.33488   0.097  0.92311
as.factor(orden.prefResto)1      1.10867    0.37777   2.935  0.00334 **
via.accResto                     2.09984    1.10467   1.901  0.05732 .
via.accSelectividad              1.79827    0.82175   2.188  0.02865 *
via.accTitulados Universitarios  3.02347    1.56757   1.929  0.05376 .
sexoHome                         0.40678    0.27306   1.490  0.13630
provLimitrofes                   1.55448    0.32239   4.822 1.42e-06 ***
provResta                        0.47004    0.41099   1.144  0.25276
cred.pres.pri                    0.06345    0.00806   7.872 3.50e-15 ***
cred.honor.pri                  -0.11626    0.09398  -1.237  0.21603
cred.pres.ultimo                -0.02968    0.01246  -2.383  0.01717 *
num.asi.matric.ultimo            0.45778    0.08120   5.638 1.72e-08 ***
num.asi.rep.ultimo              -0.18979    0.04463  -4.253 2.11e-05 ***
cred.sup.exam.media             -0.17327    0.01555 -11.143  < 2e-16 ***
as.factor(trabPri)1             -0.05026    0.38424  -0.131  0.89593
as.factor(trabUltimo)1           0.75636    0.36988   2.045  0.04087 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1526.00  on 1130  degrees of freedom
Residual deviance:  500.57  on 1112  degrees of freedom
  (194 observations deleted due to missingness)
AIC: 538.57

Number of Fisher Scoring iterations: 7
```

Figura 4.14: Análisis de regresión logística para los alumnos de la Facultad de Ciencias Jurídicas y Económicas.

Las **variables que explican el abandono** de los estudios en la Facultad de Ciencias Jurídicas y Económicas son:

- La **nota de acceso**.
- Entrar en una titulación cuyo **orden de preferencia** es **superior a la segunda opción**: aumenta la probabilidad de abandono.
- La **vía de acceso**: la vía de acceso que menos probable hace el abandono es Formación Profesional, seguida de Selectividad, a continuación, el resto de vías de acceso y, por último, Titulado Universitario.
- Pertenecer a una **provincia limítrofe a Castellón**: aumenta la probabilidad de abandono.
- El número de **créditos presentados en primer curso**: a mayor número de créditos presentados mayor probabilidad de abandono.
- El número de **créditos presentados en el último curso de matrícula**.
- El número de **asignaturas matriculadas en el último curso de matrícula**: matricularse de muchas asignaturas aumenta la probabilidad de abandono.
- El número de **asignaturas repetidas en el último curso de matrícula**: repetir asignaturas disminuye la probabilidad de abandono.
- La media de **créditos superados en exámenes**: superar exámenes disminuye la probabilidad de abandono.
- El **trabajo realizado en el último curso de matrícula**: trabajar aumenta la probabilidad de abandono.

★ Facultad de Ciencias de la Salud

Los resultados obtenidos para la Facultad de Ciencias de la Salud se muestran en la Figura 4.15.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3696  -0.2955  -0.1873  -0.1447   3.0130

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                     14.657582 882.746394   0.017   0.9868
edad                             0.077474   0.046338   1.672   0.0945 .
nota.de.acceso                   3.220807   1.410823   2.283   0.0224 *
via.accResto                   -17.425037 882.743825  -0.020   0.9843
via.accSelectividad            -18.056252 882.743460  -0.020   0.9837
via.accTitulados Universitarios -16.205070 882.743937  -0.018   0.9854
cred.honor.pri                   -0.006512   0.079164  -0.082   0.9344
cred.pres.ultimo                 -0.020316   0.017712  -1.147   0.2514
num.asi.matric.ultimo             0.299784   0.146201   2.050   0.0403 *
cred.sup.exam.media              -0.121821   0.018202  -6.693 2.19e-11 ***
as.factor(trabPri)1               0.504957   0.582275   0.867   0.3858
as.factor(trabUltimo)1           -0.078678   0.631361  -0.125   0.9008
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 391.85  on 421  degrees of freedom
Residual deviance: 175.35  on 410  degrees of freedom
  (42 observations deleted due to missingness)
AIC: 199.35

Number of Fisher Scoring iterations: 13
```

Figura 4.15: Análisis de regresión logística para los alumnos de la Facultad de Ciencias de la Salud.

Las **variables que explican el abandono** de los estudios en la Facultad de Ciencias de la Salud son:

- La **nota de acceso**.
- El **número de asignaturas matriculadas en el último curso de matrícula**: matricularse de bastantes asignaturas aumenta la probabilidad de abandonar los estudios.
- La **media de créditos superados en exámenes**: superar créditos en exámenes disminuye la probabilidad de abandono.

En esta facultad, se observa que prácticamente ninguna variable es significativa en el abandono en comparación con el resto de centros. Los resultados obtenidos no son satisfactorios debido a las diferencias entre sus titulaciones por lo que sería conveniente repetir de nuevo el estudio por cada uno de los grados.

Análisis clúster

Para encontrar tipologías de alumnos que abandonan los estudios, se ha realizado el análisis clúster para cada uno de los centros. En todos los centros se ha obtenido que la mejor opción es crear tres grupos de alumnos. A continuación, se detallan los resultados obtenidos por cada uno de los centros.

★ Escuela Superior de Tecnología y Ciencias Experimentales

Se ha comprobado qué variables influyen en la clasificación realizando contrates de ANOVA y de la *Chi-cuadrado*. Se muestran los *p-valores* obtenidos en la Tabla 4.2.

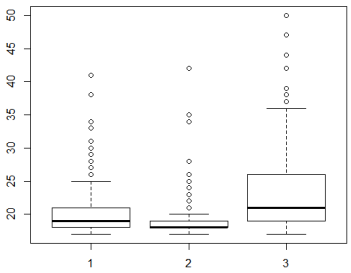
edad	nota.de.acceso	cred.pres.pri	cred.pres.ultimo	cred.honor.pri	cred.sup.exam.media
8.906e-10	0.3699	2-2e-16	1.224e-15	0.2435	1.9e-03

num.asi.rep.ult	curso.abandono	via.acc	orden.pref	sexo	trabPri	trabUltimo	prov
8.721e-05	5.402e-11	8.4e-03	3.929e-03	7.7e-04	6.09e-03	3.269e-05	0.03943

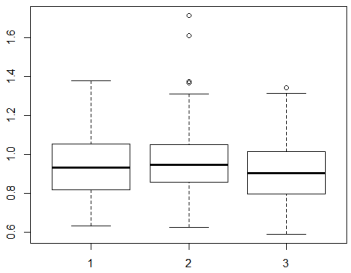
Tabla 4.2: *p-valores* de los contrastes ANOVA y *Chi-cuadrado*.

Tomando un nivel de significación de 0.05, las variables que no influyen en la clasificación son la nota de acceso y el número de créditos de honor obtenidos en primer curso.

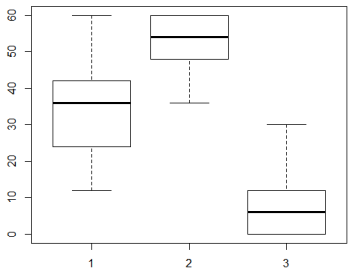
Se puede observar la descripción de los grupos en la Figura 4.16.



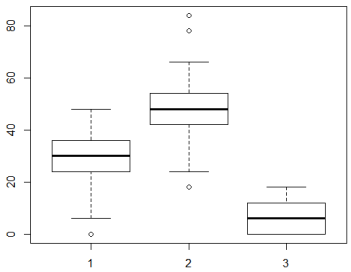
(a) Edad



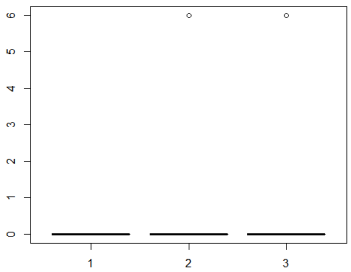
(b) Nota de acceso



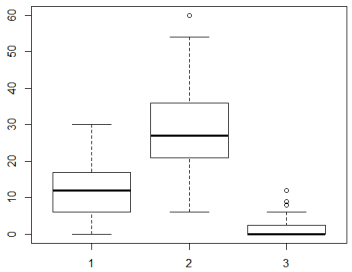
(c) Créditos presentados primer curso



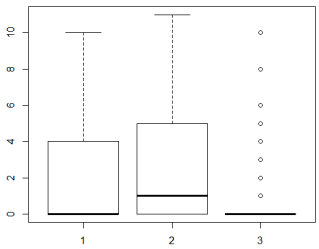
(d) Créditos presentados último curso



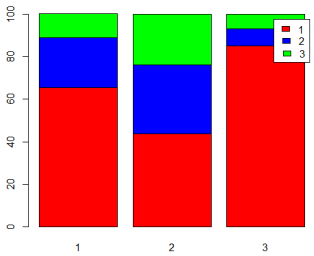
(e) Créditos de honor primer curso



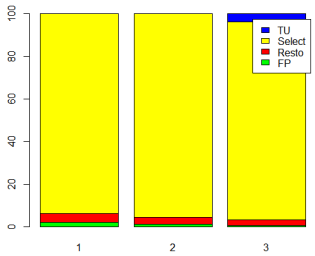
(f) Promedio de créditos superados



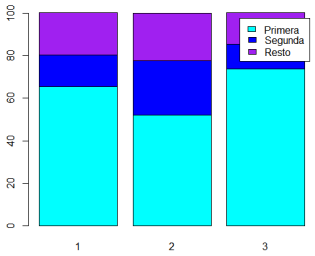
(g) Número de asignaturas repetidas último curso de matrícula



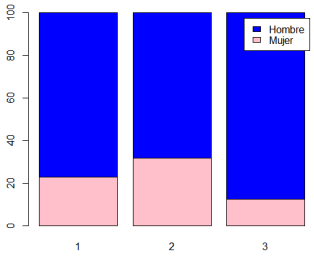
(h) Curso de abandono (**rojo:** Primero, **azul:** Segundo, **verde:** Tercero)



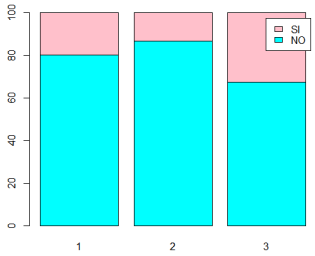
(i) Vía de acceso (**amarillo:** Selectividad, **verde:** Formación Profesional, **azul:** Titulados Universitarios, **Rojo:** Resto)



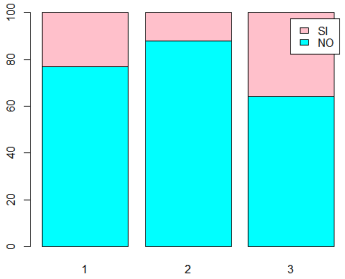
(j) Orden de preferencia (**azul claro:** Primera, **azul oscuro:** Segunda, **morado:** Resto)



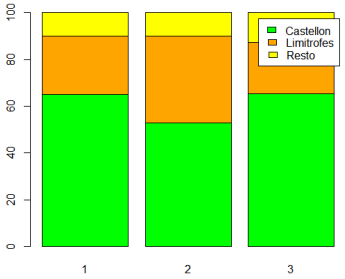
(k) Sexo (**rosa:** Mujer, **azul:** Hombre)



(l) Trabajo primer curso



(m) Trabajo último curso



(n) Provincia de residencia familiar (**verde:** Castellón, **naranja:** Provincias limítrofes a Castellón, **amarillo:** Resto)

Figura 4.16: Descripción de los grupos obtenidos en la Escuela Superior de Tecnología y Ciencias Experimentales.

Por tanto, se han encontrado tres tipologías de alumnos que abandonan los estudios.

- *Estudiantes de tipo 1:* Son en mayor parte **hombres** cuya edad está comprendida entre la edad de los alumnos del segundo grupo y la de los del tercer grupo. Se **presentan a examen de pocos créditos en primero** (sobre unos 35) **y en el último curso de matrícula a algunos menos** (sobre unos 30). **No superan prácticamente créditos en exámenes** (unos 10 por curso) y la mayor parte **no trabaja**. Suelen **abandonar en primer curso** aunque también hay bastantes que abandonan en segundo.

- *Estudiantes de tipo 2:* Son **los más jóvenes** de todos. Se presentan de bastantes créditos tanto en primer curso como en el último curso de matrícula (entre 50 y 60). Superan aproximadamente 25 créditos por año. La mayor parte **no trabaja** y abandonan tanto en primer como en segundo y tercer curso. Una gran parte de ellos pertenece a **provincias limítrofes a Castellón** en comparación con el resto de grupos luego ésta podría ser la causa de su abandono.

- *Estudiantes de tipo 3:* Suelen ser **mayores** que los alumnos clasificados en los grupos anteriores. En este grupo, la proporción de **Titulados Universitarios** es mayor que en el resto de grupos. No se presentan a prácticamente ningún crédito y no superan ninguno. Muchos alumnos **trabajan** en comparación con el resto de grupos. Suelen **abandonar en primer curso**.

★ Facultad de Ciencias Humanas y Sociales

Se ha comprobado qué variables influyen en la clasificación realizando contrastes de ANOVA y de la *Chi-cuadrado*. Se muestran los *p-valores* en la Tabla 4.3.

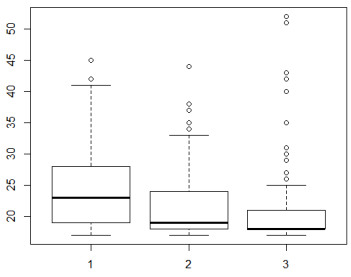
edad	nota.de.acceso	cred.pres.pri	cred.pres.ultimo	cred.honor.pri	cred.sup.exam.media
1.683e-06	1.965e-04	2.2e-16	2.2e-16	1.375e-03	2.2e-16

num.asi.rep.ult	curso.abandono	via.acc	orden.pref	sexo	trabPri	trabUltimo	prov
0.54649	4.205e-12	1.43e-03	0.8601	0.091	9.25e-03	6.733e-04	0.02312

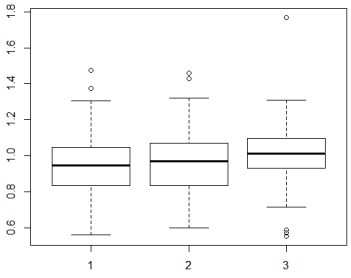
Tabla 4.3: *p-valores* de los contrastes ANOVA y *Chi-cuadrado*.

Tomando un nivel de significación de 0.05, las variables que no influyen en la clasificación son el número de asignaturas repetidas en el último curso de matrícula, el orden de preferencia y el sexo.

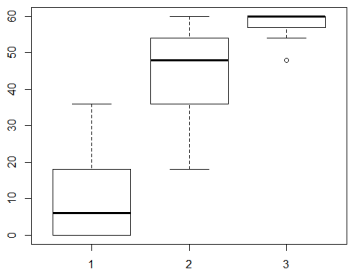
Se puede observar la descripción de los grupos en la Figura 4.17.



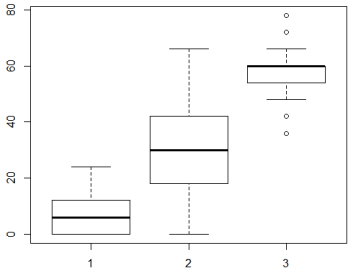
(a) Edad



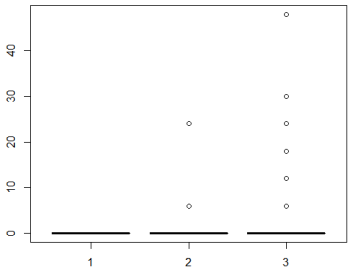
(b) Nota de acceso



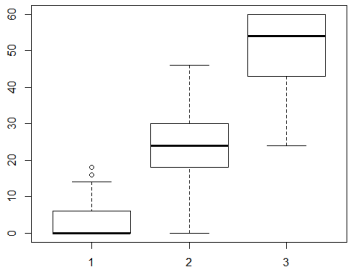
(c) Créditos presentados primer curso



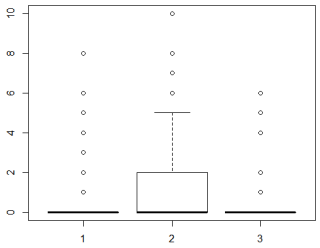
(d) Créditos presentados último curso



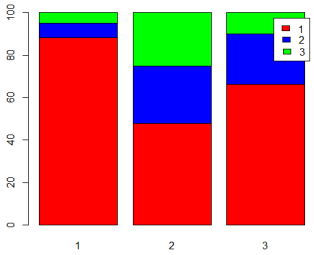
(e) Créditos de honor primer curso



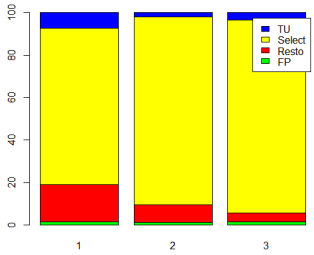
(f) Promedio de créditos superados



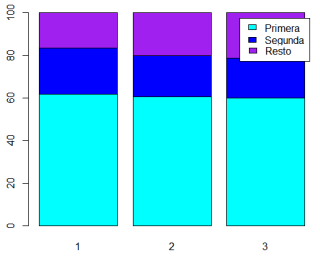
(g) Número de asignaturas repetidas último curso de matrícula



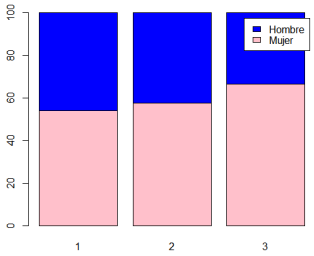
(h) Curso de abandono (**rojo:** Primero, **azul:** Segundo, **verde:** Tercero)



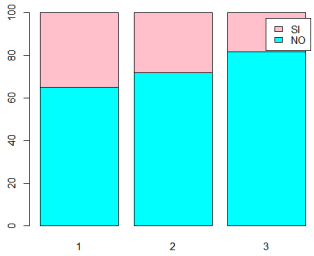
(i) Vía de acceso (**amarillo:** Selectividad, **verde:** Formación Profesional, **azul:** Titulados Universitarios, **Rojo:** Resto)



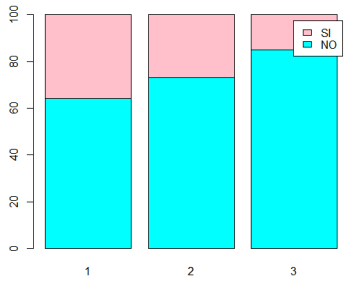
(j) Orden de preferencia (**azul claro:** Primera, **azul oscuro:** Segunda, **morado:** Resto)



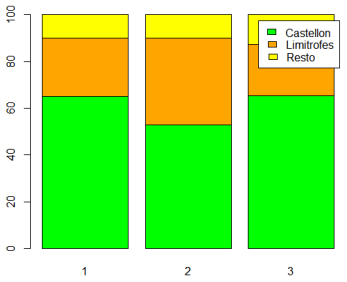
(k) Sexo (**rosa:** Mujer, **azul:** Hombre)



(l) Trabajo primer curso



(m) Trabajo último curso



(n) Provincia de residencia familiar (**verde:** Castellón, **naranja:** Provincias limítrofes a Castellón, **amarillo:** Resto)

Figura 4.17: Descripción de los grupos obtenidos en la Facultad de Ciencias Humanas y Sociales.

Por tanto, se han encontrado tres tipologías de alumnos que abandonan los estudios:

- *Estudiantes de tipo 1*: Suelen ser **mayores** que los alumnos clasificados en el resto de grupos. Muchos de ellos **trabajan** tanto en el primer curso de matrícula como en el curso anterior al abandono. Además, **se presentan a examen de muy pocos créditos y no superan prácticamente créditos en exámenes**. Aunque la mayor parte entra a través de Selectividad, **muchos de ellos entran por otras vías de acceso que no son Selectividad**. La mayor parte **abandona en primer curso**.
- *Estudiantes de tipo 2:* Son, en general, más jóvenes que los anteriores. La mayor parte entra a través de **Selectividad**. Se presentan de **bastantes créditos en primer curso** (sobre unos 50), pero **en el último curso se presentan a menos** (sobre unos 30). Abandonan tanto en primer curso como en segundo y tercer curso.
- *Estudiantes de tipo 3:* **Se presentan y superan bastantes más créditos que en el resto de grupos**. **No suelen trabajar**. Aunque la mayor parte abandona en primer curso, también hay muchos alumnos de este grupo que abandonan en segundo o tercer curso. La mayor parte pertenece a **provincias limítrofes a Castellón**, luego ésta podría ser la causa de su abandono.

★ Facultad de Ciencias Jurídicas y Económicas

Se ha comprobado qué variables influyen en la clasificación realizando contrastes de ANOVA y de la *Chi-cuadrado*. Se muestran los *p-valores* en la Tabla 4.4.

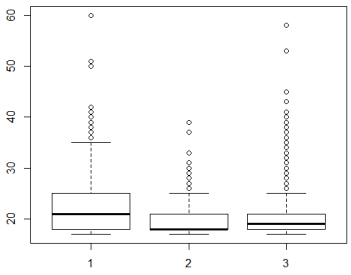
edad	nota.de.acceso	cred.pres.pri	cred.pres.ultimo	cred.honor.pri	cred.sup.exam.media
6.759e-04	0.6294	2.2e-16	2.2e-16	0.9393	6.044e-08

num.asi.rep.ult	curso.abandono	via.acc	orden.pref	sexo	trabPri	trabUltimo	prov
7.662e-16	2.2e-16	0.02406	0.1468	2.88e-03	3.14e-04	2.282e-04	0.0499

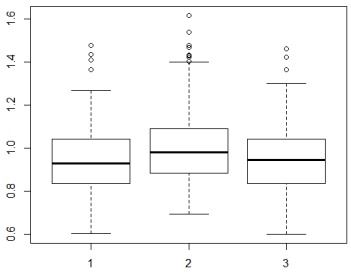
Tabla 4.4: *p-valores* de los contrastes ANOVA y *Chi-cuadrado*.

Tomando un nivel de significación de 0.05, las variables que no influyen en la clasificación son la nota de acceso, el número de créditos de honor obtenidos en primer curso y el orden de preferencia de la titulación.

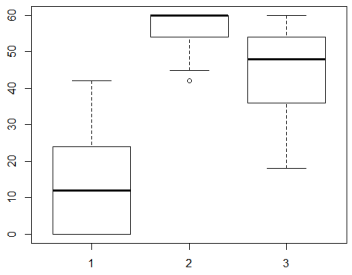
Se puede observar la descripción de los grupos en la Figura 4.18.



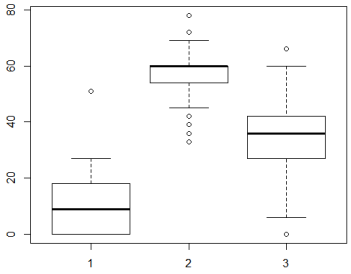
(a) Edad



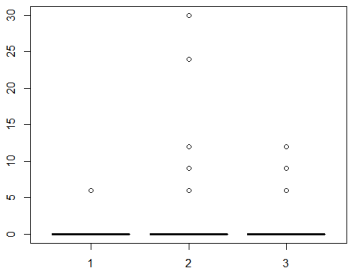
(b) Nota de acceso



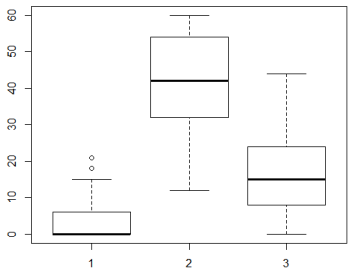
(c) Créditos presentados primer curso



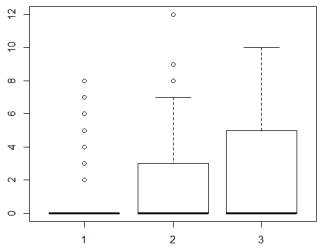
(d) Créditos presentados último curso



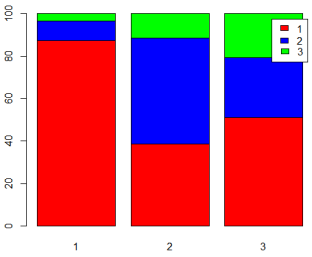
(e) Créditos de honor primer curso



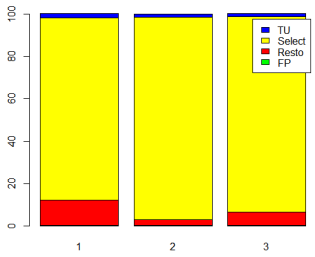
(f) Promedio de créditos superados



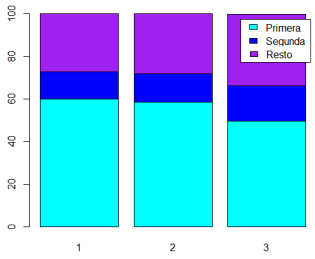
(g) Número de asignaturas repetidas último curso de matrícula



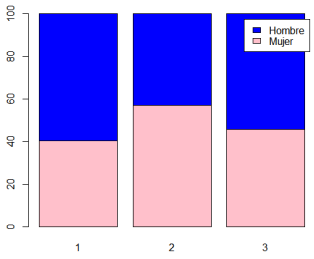
(h) Curso de abandono (**rojo:** Primero, **azul:** Segundo, **verde:** Tercero)



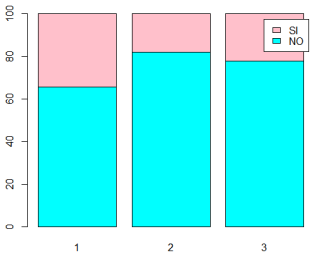
(i) Vía de acceso (**amarillo:** Selectividad, **verde:** Formación Profesional, **azul:** Titulados Universitarios, **Rojo:** Resto)



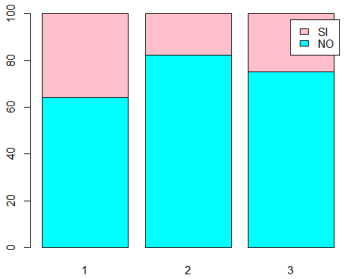
(j) Orden de preferencia (**azul claro:** Primera, **azul oscuro:** Segunda, **morado:** Resto)



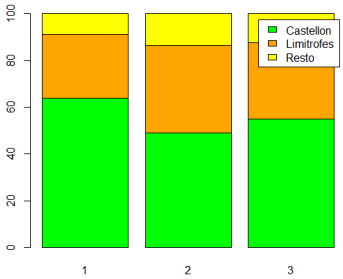
(k) Sexo (**rosa:** Mujer, **azul:** Hombre)



(l) Trabajo primer curso



(m) Trabajo último curso



(n) Provincia de residencia familiar (**verde:** Castellón, **naranja:** Provincias limítrofes a Castellón, **amarillo:** Resto)

Figura 4.18: Descripción de los grupos obtenidos en la Facultad de Ciencias Humanas y Sociales.

Por tanto, se han encontrado tres tipologías de alumnos que abandonan los estudios:

- *Estudiantes tipo 1:* Son **mayores** que los estudiantes clasificados en el resto de grupos. **Se presentan a pocos créditos** (sobre unos 10) y **no superan prácticamente ningún crédito**. Muchos de ellos **trabajan** y suelen **abandonar en primer curso**.
- *Estudiantes tipo 2:* Son **los más jóvenes**. **Se presentan a bastantes créditos** tanto en primer como en último curso y **superan bastantes créditos en exámenes**. **No suelen trabajar** y **suelen abandonar en primer o en segundo curso**. Una gran parte de ellos, en comparación con el resto de grupos, pertenece a **provincias limítrofes a Castellón** luego ésta es posiblemente la causa de su abandono.
- *Estudiantes tipo 3:* **Se presentan a una media de 40 créditos en el primer y en último curso de matrícula y superan alrededor de unos 15 créditos de media cada año. Abandonan en todos los cursos.**

★ Facultad de Ciencias de la Salud

Se ha comprobado qué variables influyen en la clasificación realizando contrastes de ANOVA y de la *Chi-cuadrado*. Se muestran los *p-valores* en la Tabla 4.5.

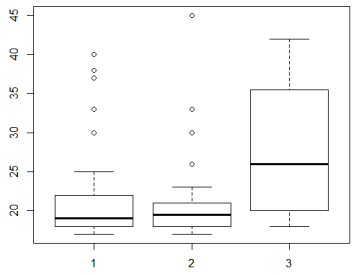
edad	nota.de.acceso	cred.pres.pri	cred.pres.ultimo	cred.honor.pri	cred.sup.exam.media
5.444e-03	0.4283	6.249e-07	1.071e-03	0.6771	0.3721

num.asi.rep.ult	curso.abandono	via.acc	orden.pref	sexo	trabPri	trabUltimo	prov
0.3796	0.1352	0.4876	0.6995	0.3154	2.113e-03	2.316e-03	0.04101

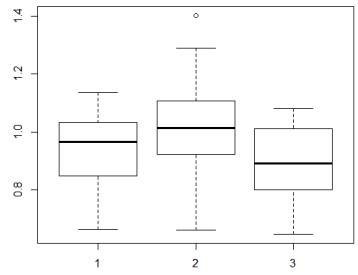
Tabla 4.5: *p-valores* de los contrastes ANOVA y *Chi-cuadrado*.

En este caso, tomando un nivel de significación de 0.05, las únicas variables en las que es posible afirmar que hay diferencias significativas entre los grupos son la edad, el número de créditos presentados en primer y último curso de matrícula, si el alumno tenía trabajo en primer y último curso de matrícula y la provincia de residencia familiar. Lo que diferencia a esta facultad de las anteriores es que el número de personas que han abandonado los estudios es mucho menor y por tanto los contrastes son más conservativos (necesitan mayor evidencia para poder rechazar la independencia).

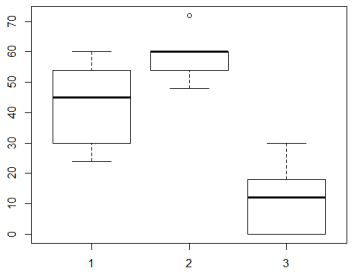
Se puede observar la descripción de los grupos en la Figura 4.19.



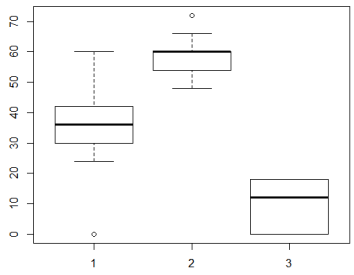
(a) Edad



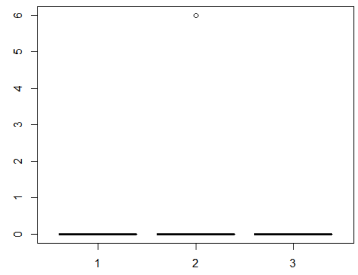
(b) Nota de acceso



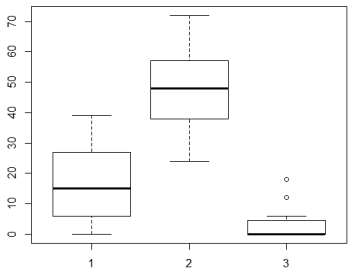
(c) Créditos presentados primer curso



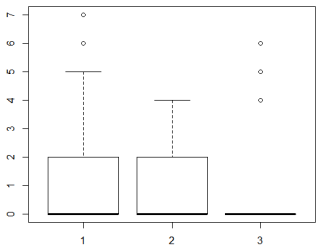
(d) Créditos presentados último curso



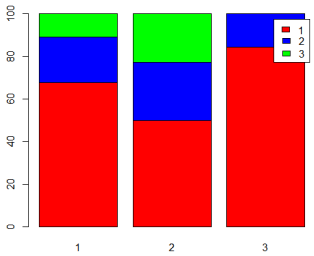
(e) Créditos de honor primer curso



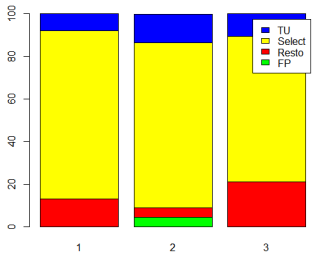
(f) Promedio de créditos superados



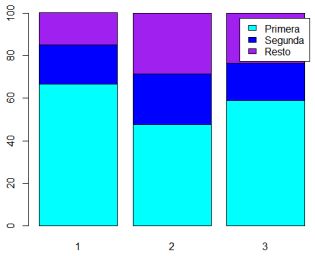
(g) Número de asignaturas repetidas último curso de matrícula



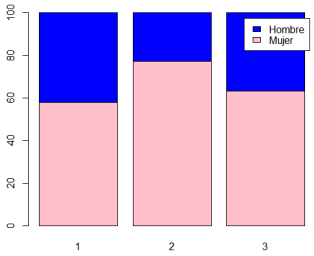
(h) Curso de abandono (**rojo:** Primero, **azul:** Segundo, **verde:** Tercero)



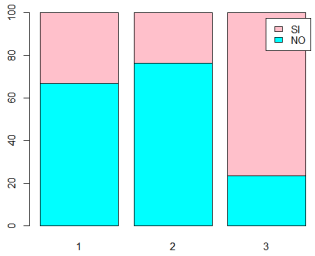
(i) Vía de acceso (**amarillo:** Selectividad, **verde:** Formación Profesional, **azul:** Titulados Universitarios, **Rojo:** Resto)



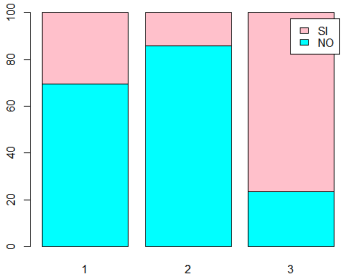
(j) Orden de preferencia (**azul claro:** Primera, **azul oscuro:** Segunda, **morado:** Resto)



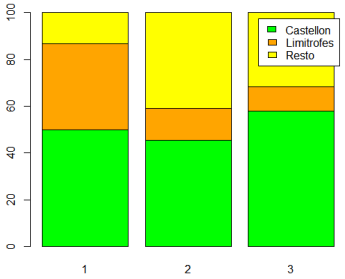
(k) Sexo (**rosa:** Mujer, **azul:** Hombre)



(l) Trabajo primer curso



(m) Trabajo último curso



(n) Provincia de residencia familiar (**verde:** Castellón, **naranja:** Provincias limítrofes a Castellón, **amarillo:** Resto)

Figura 4.19: Descripción de los grupos obtenidos en la Facultad de Ciencias de la Salud.

Por tanto, se han encontrado tres tipologías de alumnos que abandonan los estudios.

- *Estudiantes de tipo 1:* **Se presentan a bastantes exámenes** tanto en el primer como en el último curso de matrícula. En primer curso suelen presentarse a unos 45 créditos, y en el último curso a unos 35 créditos. No obstante, **no superan casi créditos** (sobre unos 15 por curso); **la mayor parte no trabaja** y muchos de ellos pertenecen a **provincias limítrofes a Castellón**, luego ésta es posiblemente la causa de su abandono.
- *Estudiantes de tipo 2:* **Se presentan a casi todos los exámenes** tanto en el primer como en último curso de matrícula. **Superan una media de 50 créditos por curso**, **no trabajan** y **abandonan en todos los cursos**. Muchos de ellos, en comparación con los otros grupos, pertenecen a **provincias que no son ni la provincia de Castellón ni ninguna de sus provincias limítrofes**.
- *Estudiantes de tipo 3.* Son **mayores** que los alumnos pertenecientes al resto de grupos. **Se presentan a muy pocos créditos** en exámenes y **no superan prácticamente ningún crédito**. Una gran parte de estos alumnos accede a través de vías de acceso como son **Titulados Universitarios o el resto de vías de acceso** sin incluir Selectividad ni Formación Profesional. **Suelen trabajar** y **abandonan en primer curso o como mucho en segundo curso**.

En todos los centros, la bondad del análisis clúster ha sido muy cercana a 0.4 por lo que se puede considerar bastante aceptable.

4.2. Análisis realizados sobre la inserción laboral

Para analizar la inserción laboral se ha hecho uso de la regresión logística y de reglas de asociación.

En esta sección se omiten los resultados de la regresión logística puesto que no han aportado prácticamente información sobre la inserción laboral y se explican los resultados obtenidos utilizando las reglas de asociación para el caso de la Escuela Superior de Tecnología, ya que las reglas obtenidas para el resto de facultades no son satisfactorias pues tienen un nivel de soporte muy bajo (únicamente un porcentaje muy pequeño de la muestra verifica cada una de las reglas).

En primer lugar, como la muestra contiene variables numéricas continuas es necesario trocearlas en intervalos antes de realizar el análisis y además, las variables numéricas discretas deben ser convertidas a variables categóricas. Una vez realizadas estas trasformaciones se debe transformar la hoja de datos que contiene la muestra a matriz de transacciones y, a continuación, se deben buscar reglas que impliquen encontrar un trabajo que requiere el título obtenido. Se han ordenado estas reglas por nivel de confianza.

Primeras reglas obtenidas en la Escuela Superior de Tecnología y Ciencias Experimentales:

```
     lhs                                rhs                       supp   conf  lift
1  {erasmus.estudios=0,
    practicas.
    extracurriculares=1} => {trabajo.titulacion=1}    0.15   0.9   2.25

2  {erasmus.estudios=0,
    erasmus.practicas=0,
    practicas.
    extracurriculares=1} => {trabajo.titulacion=1}    0.15   0.9   2.25

3  {sexo=1,
    erasmus.estudios=0,
    practicas.
    extracurriculares=1} => {trabajo.titulacion=1}    0.13   0.88  2.22

4  {sexo=1,
    erasmus.estudios=0,
    erasmus.practicas=0,
    practicas.
    extracurriculares=1} => {trabajo.titulacion=1}    0.13   0.88  2.22
5  {cursos.cuatro.o.menos=1,
    practicas.
    extracurriculares=1} => {trabajo.titulacion=1}    0.12   0.87  2.19

6  {cursos.cuatro.o.menos=1,
    erasmus.estudios=0,
    practicas.
    extracurriculares=1} => {trabajo.titulacion=1}    0.12   0.87  2.19

7  {cursos.cuatro.o.menos=1,
    erasmus.practicas=0,
    practicas.
    extracurriculares=1} => {trabajo.titulacion=1}    0.12   0.87  2.19

8  {cursos.cuatro.o.menos=1,
    erasmus.estudios=0,
    erasmus.practicas=0,
    practicas.
    extracurriculares=1} => {trabajo.titulacion=1}    0.12   0.87  2.19

9  {sexo=1,
    cursos.cuatro.o.menos=1,
    practicas.
    extracurriculares=1} => {trabajo.titulacion=1}    0.10   0.86  2.14

10 {cred.honor=0,
    erasmus.estudios=0,
    practicas.
    extracurriculares=1} => {trabajo.titulacion=1}    0.10   0.86  2.14

11 {sexo=1,
    cursos.cuatro.o.menos=1,
    erasmus.estudios=0,
    practicas.
    extracurriculares=1} => {trabajo.titulacion=1}    0.10   0.86  2.14
```

```
12 {sexo=1,
    cursos.cuatro.o.menos=1,
    erasmus.practicas=0,
    practicas.
    extracurriculares=1} => {trabajo.titulacion=1}     0.10   0.86   2.14

13 {cred.honor=0,
    erasmus.estudios=0,
    erasmus.practicas=0,
    practicas.
    extracurriculares=1} => {trabajo.titulacion=1}     0.10   0.86   2.14

14 {sexo=1,
    cursos.cuatro.o.menos=1,
    erasmus.estudios=0,
    erasmus.practicas=0,
    practicas.
    extracurriculares=1} => {trabajo.titulacion=1}     0.10   0.86   2.14
```

Por un lado, la columna *supp* indica el soporte de la regla, es decir, el porcentaje de alumnos de la muestra que están involucrados en esa regla. Por ejemplo, en la primera regla, el soporte determina que el 15 % de los alumnos de la muestra no han participado en un programa de Erasmus por motivo de estudios, pero sí han realizado prácticas extracurriculares.

Por otro lado, la columna *conf* muestra el porcentaje de alumnos que, verificando la primera parte de la regla, también cumplen la segunda. Volviendo al ejemplo anterior, de ese 15 % de alumnos de la muestra que no han participado en un programa de Erasmus por motivo de estudios y que han realizado prácticas extracurriculares, el 90 % ha conseguido un empleo que requiere el título obtenido para su desempeño.

Observando todas las reglas en conjunto, se puede apreciar que en prácticamente todas las reglas aparece haber hecho **prácticas extracurriculares** y **tardar como máximo cuatro años en obtener el título**, luego parecen ser dos requisitos a seguir para los alumnos de la Escuela Superior de Tecnología y Ciencias Experimentales si quieren encontrar un empleo.

4.3. Conclusiones

Las principales conclusiones extraídas en el análisis del abandono de los estudios son las siguientes:

- En todos los centros los hombres abandonan en mayor proporción que las mujeres.
- Una de las vías de acceso que menos abandono tiene, prácticamente en todos los centros, excepto en la Facultad de Ciencias de la Salud, es Formación Profesional.
- Una de las vías de acceso que más abandono tiene, en todos los centros, es Titulado Universitario y, aunque pueda resultar extraño _porque estas personas ya han conseguido un título universitario_, se ha conseguido explicar este hecho mediante el análisis clúster, ya

que en prácticamente todos los centros se ha obtenido un grupo donde se clasifican la mayor parte de los Titulados Universitarios, y una característica de ese grupo es que los estudiantes clasificados tienen trabajo en mayor proporción que los estudiantes pertenecientes al resto de grupos. Es decir, los Titulados Universitarios que intentan obtener otra titulación la acaban abandonando porque encuentran trabajo.

- Los estudiantes que no son de la provincia de Castellón abandonan en mayor proporción que los estudiantes que son de Castellón, principalmente los que pertenecen a provincias limítrofes a Castellón.
- Entrar en una titulación cuya opción de preferencia al realizar la preinscripción es superior a la segunda opción hace que los estudiantes abandonen en mayor proporción.
- Encontrar un trabajo influye notablemente en el abandono.

Las conclusiones obtenidas en el análisis de la inserción laboral resultan bastante escasas. Atendiendo a la Escuela Superior de Tecnología y Ciencias Experimentales, se han encontrado dos requisitos a cumplir por parte de los estudiantes para poder encontrar un empleo que requiere la titulación cursada. Estos requisitos son realizar prácticas extracurriculares y obtener el título en un máximo de cuatro años.

Capítulo 5

Conclusiones generales

El proyecto desarrollado ha sido llevado a cabo en el *Gabinete de Planificación y Prospectiva Tecnológica* de la Universidat Jaume I con el fin de poder determinar las causas que explican el abandono de los estudios y la inserción laboral.

Para lograr este objetivo, se ha hecho un estudio previo de aprendizaje sobre distintas técnicas de minería de datos, tanto a nivel teórico como a nivel práctico, para después seleccionar aquellas técnicas más adecuadas para aplicar a los datos a analizar, en función de la información que se requiriera obtener de cada una de las muestras. Las técnicas finalmente utilizadas han sido la regresión logística, el análisis clúster, las redes neuronales, los árboles de clasificación, las reglas de asociación, etc., cuyo desarrollo teórico se ha mostrado en el Capítulo 3.

Por un lado, los objetivos planteados con respecto al abandono de los estudios han sido: determinar qué variables influyen en el abandono, buscar tipologías de alumnos que abandonan y crear algoritmos de predicción para poder determinar si un alumno abandonará, o no, los estudios.

Estos objetivos han sido llevados a cabo con éxito aunque cabe destacar que si se desea profundizar sobre estos temas sería conveniente realizar de nuevo los análisis con datos más recientes e incluyendo algún tipo de variable económica, como por ejemplo, si el alumno tenía, o no, beca. Este tipo de variable no se ha podido incluir en el análisis realizado porque al tratarse de datos antiguos, se carecía de dicha información en la base de datos oficial de la UJI. No obstante, se considera que es muy probable que esta variable no fuese relevante en el estudio realizado porque todavía no se había aplicado la ley de subida de tasas que fue llevada a cabo en el año 2012.

También deberían incluirse variables sociales, de motivación y de satisfacción sobre el grado realizado. Para ello, algunos autores como Aguirre, Valdovinos, Velazquez, Eleuterio y Romero (2015) [26] elaboran encuestas con preguntas de este tipo llegando a la conclusión de que estas variables también influyen en el abandono universitario. Otras variables que se podrían añadir serían los estudios alcanzados por el padre o por la madre como lo realizan La Red, Karanik, Giovannini y Pinto (2015) [27] en una universidad de Argentina logrando comprobar que son variables que influyen en el abandono de los estudios.

Además, por lo que respecta a la Facultad de Ciencias de la Salud, también se aconsejaría repetir los análisis por cada una de sus titulaciones, debido a que heterogeneidad entre los distintos grados no ha permitido obtener conclusiones satisfactorias sobre esta facultad.

Por otro lado, los objetivos propuestos para analizar la inserción laboral han sido determinar qué variables influyen a la hora de encontrar empleo y qué patrones tienen en común los alumnos que se insertan en el mundo laboral.

En este caso, los objetivos también han sido alcanzados pero sin poder lograr grandes conclusiones al respecto como consecuencia de la poca información de la que se dispone, por lo que resultaría conveniente volver a realizar este tipo de análisis dentro de unos años, cuando la universidad disponga de un mayor número de egresados.

Bibliografía

[1] Porcel, Eduardo; Dapozo, Glayds; López María V. (2009), *Modelos predictivos y técnicas de minería de datos para la identificación de factores asociados al rendimiento académico de los alumnos universitarios.* Universidad Nacional del Nordeste, Corrientes, Argentina.

[2] Bernardo, A., Cerezo, R., Núñez, J.C, Tuero, E. y Esteban, M. (2015), *Predicción del abandono universitario: variables explicativas y medidas de prevención.* Revista Fuentes, pp.63-84

[3] Castaño, E., Gallón, S., Gómez, K. y Vásquez, J. (2004), *Deserción estudiantil universitaria: una aplicación de modelos de duración.* Lecturas de economía, 60, pp.39-66

[4] Trevizán, A.L., Beltrán, C. y Cosolito, P. (2009), *Variables que condicionan la deserción y retención durante el trayecto universitario de alumnos de carrera de Ingeniería Agronómica de la Universidad Nacional de Rosario.* Revista de Epistemología y Ciencias Humanas, 1, pp.85-95

[5] Sánchez, J. (2014), *Modelos predictivos para el estudio del abandono en centros universitarios.* (Trabajo Fin de Grado, Universidad Politécnica de Madrid) `http://oa.upm.es/31205/1/PFC_JESUS_SANCHEZ_SANTAMARIA.pdf`

[6] Araque, F., Roldán, C. y Salguero, A. (2009), *Factors influencing university drop out rates.* Computers and Education, 53, pp.563-574

[7] Marín, M., Infante, E. y Troyano, Y. (2000), *El fracaso académico en la universidad: aspectos motivadores e intereses profesionales.* Revista Lationamericana de Psicología, 32 (3), pp.505-517

[8] Rogrigo, M.F., Molina, J.G., García-Ros, R. y Pérez-González, F. (2012), *Efectos de interacción en la predicción del abandono en los estudios de Psicología.* Anales de Psicología, 28 (1), pp.113-119

[9] R Core Team (2013), *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`

[10] Jain, A.K. (2009), *Data clustering: 50 years beyond K-means.* Pattern recognition Lett.

[11] Kaufman, L. and Rousseeuw, P.J. (1987), *Clustering by means of Medoids, in Statistical Data Analysis Based on the L_1–Norm and Related Methods*, pp.405–416

[12] Steinhaus, H. (1957), *Sur la division des corps matériels en parties.* Bull. Acad. Polon. Sci. (in French) 4 (12), pp.801–804

[13] E.W. Forgy (1965), *Cluster analysis of multivariate data: efficiency versus interpretability of classifications.* Biometrics 21: pp.768–769

[14] J.MacQueen (1967), *Some methods of classification and analysis of multivariate observations.*

[15] J. A. Hartigan and M. A. Wong (1979), *Algorithm AS 136: A K-means Clustering Algorithm.* Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 28, No. 1, pp.100-108

[16] Tibshirani, R., Walther, G. and Hastie, T. (2001), *Estimating the number of clusters in a dataset via the gap statistic.* Journal of the Royal Statistical Society, Series B. 32 (2), pp.411–423

[17] Alan Agresti (2011), *Categorical data analysis.*

[18] Daniel Peña (2002), *Análisis de datos multivariantes*, pp.433-440

[19] Leo Breiman, Jerome Friendman (1984), *Classification and Regression Trees.*

[20] Daniel Peña (2002), *Análisis de datos multivariantes*, pp.446-449

[21] Historia de las Reglas de Asociación: `https://es.wikipedia.org/wiki/Reglas_de_asociaci%C3%B3n#cite_note-piatetsky-2`

[22] Trevor Hastie, Robert Tibshirani and Jerome Frinedman (2001), *The elements of Statistical Learning.*

[23] T. Menzies, Y. Hu. (2006), *Data Mining For Busy People.* IEEE Computer, pp.18-25

[24] Piatetsky-Shapiro, G. (1991), *Discovery, analysis, and presentation of strong rules*, in G. Piatetsky-Shapiro, W. J. Frawley, eds, 'Knowledge Discovery in Databases'. AAAI/MIT Press, Cambridge, MA.

[25] R. Agrawal; T. Imielinski; A. Swami (1993), *Mining Association Rules Between Sets of Items in Large Databases.* SIGMOD Conference, pp.207-216

[26] José Luis Aguirre Mendiola, Rosa María Valdovinos Rosas, Juan Alberto Antonio Velazquez, Roberto Alejo Eleuterio, José Raymundo Marcial Romero (2015), *Análisis de deserción escolar con minería de datos.* Research in Computing Science 93, pp.71-82

[27] David Luis La Red Martínez, Marcelo Karanik, Mirtha Giovannini, Noelia Pinto. (2015), *Perfiles de Rendimiento Académico: Un Modelo Basado en Minería de Datos.* Campus Virtuales, Vol IV, Num 1, pp.12-30

Anexo A

Descripción de las muestras

A.1. Descripción de la muestra sobre el abandono universitario

Para analizar el problema del abandono, se ha realizado tanto un modelo global como un modelo por cada uno de los centros académicos. En este Anexo se describen únicamente las muestras de cada uno de los centros, para poder encontrar las diferencias entre ellos, y se omite la descripción de la muestra global, pues proporciona una menor información.

Los códigos que se muestran en esta sección han sido aplicados para cada una de las cuatro muestras, una por cada centro.

La proporción de alumnos que abandona en cada uno de los centros es:

```
alumnos_abandono = length(abandono[abandono == 1])
alumnos_no_abandono = length(abandono[abandono == 0])
prop_abandonos = alumnos_abandono/(alumnos_abandono + alumnos_no_abandono)
prop_abandonos
```

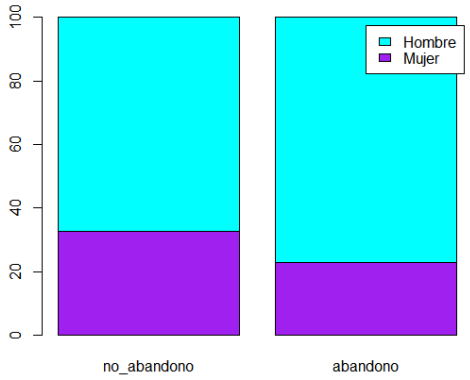
- Un 39 % en la Escuela Superior de Tecnología y Ciencias Experimentales.
- Un 19 % en la Facultad de Ciencias Humanas y Sociales.
- Un 41 % en la Facultad de Ciencias Jurídicas y Económicas.
- Un 17 % en la Facultad de Ciencias de la Salud.

Se observa una gran diferencia en la tasa de abandono entre la Facultad de Ciencias Jurídicas y Económicas y la Escuela Superior de Tecnología y Ciencias Experimentales, con respecto a las otras dos facultades.

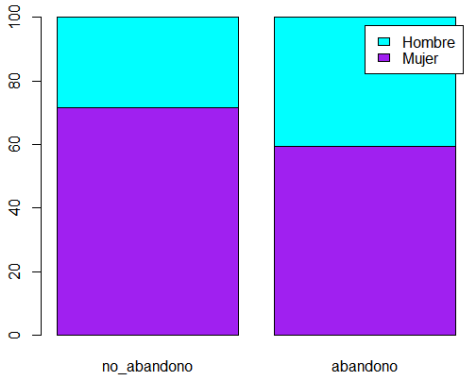
Se han realizado gráficos por cada una de las variables a estudio y se han complementado utilizando contrastes de hipótesis, para comprobar si hay diferencias significativas entre los alumnos que abandonan los estudios y los que no los abandonan. A continuación, se muestran los gráficos que se han considerado más relevantes.

Sexo:

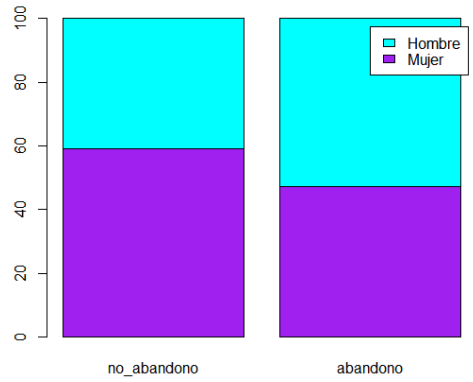
```
T = table(sexo, abandono)
Ta=colPercents(T)[1:2,]
barplot(Ta, col=c("purple","cyan"), legend.text=c("Mujer", "Hombre"), names
    .arg=c("no_abandono","abandono"))
```



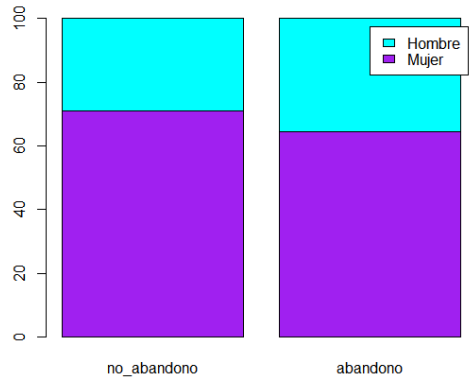
(a) Tecnología y Ciencias Experimentales



(b) Humanas y Sociales



(c) Jurídicas y Económicas



(d) Salud

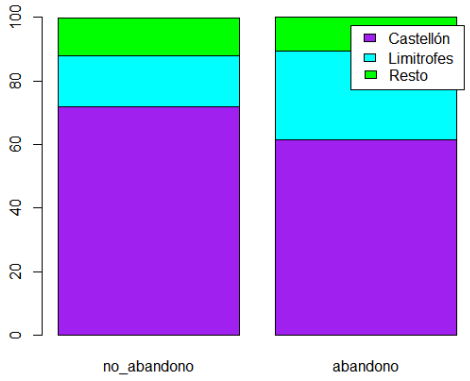
Figura A.1: Sexo, en función de abandono o no abandono de los estudios. Por un lado, se muestran los alumnos que han abandonado los estudios *(columna derecha)* y, por otro, los alumnos que no los han abandonado *(columna izquierda)*. El color **morado** se utiliza para representar a las mujeres y el color **azul** para representar a los hombres.

Para interpretar los gráficos, si se escoge por ejemplo la Escuela Superior de Tecnología y Ciencias Experimentales, del 100 % de los alumnos que no han abandonado los estudios, el 30 % son mujeres y el 70 % hombres, mientras que de los alumnos que han abandonado los estudios, el 20 % son mujeres y el 80 % hombres.

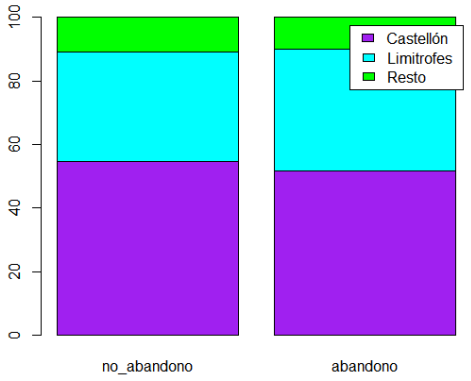
Se puede observar en todos los centros que **los hombres abandonan en mayor proporción que las mujeres**. Si se realiza un contraste de hipótesis de la *Chi-cuadrado*, esta diferencia resulta significativa en el abandono de los estudios en todos los centros, excepto en la Facultad de Ciencias de la Salud.

Provincia de residencia familiar:

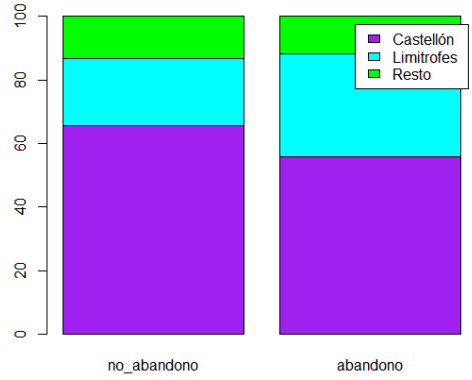
```
tabla = table(prov, abandono)
Ta = colPercents(tabla)[1:3,]
barplot(Ta, col=c("purple","cyan","green"), names.arg=c("no_abandono","
    abandono"),legend=c("Castellon", "Limitrofes", "Resto"))
```



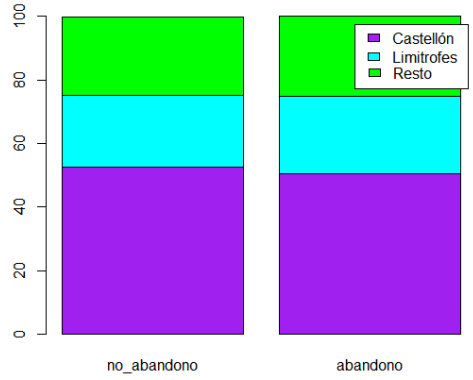
(a) Tecnología y Ciencias Experimentales



(b) Humanas y Sociales



(c) Jurídicas y Económicas



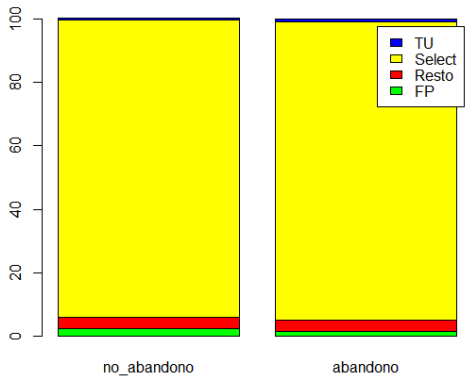
(d) Salud

Figura A.2: Provincia de residencia familiar, en función de abandono o no abandono de los estudios. Por un lado, se muestran los alumnos que han abandonado los estudios *(columna derecha)* y, por otro, los alumnos que no los han abandonado *(columna izquierda)*. El color **morado** se utiliza para representar a los alumnos de la provincia de Castellón, el color **azul** para representar a los alumnos pertenecientes a provincias limítrofes a Castellón y finalmente el color **verde** para representar al resto de provincias.

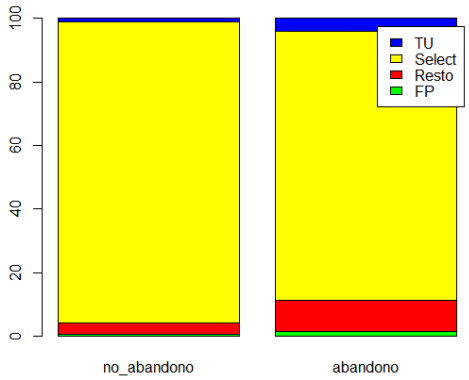
Si se acompañan los gráficos mediante un contraste de independencia de la *Chi-cuadrado*, se puede afirmar que la provincia de residencia familiar influye en el abandono en todos los centros, excepto en la Facultad de Ciencias de la Salud. **Los estudiantes que pertenecen a provincias limítrofes a Castellón son los que más abandonan**.

Vía de acceso:

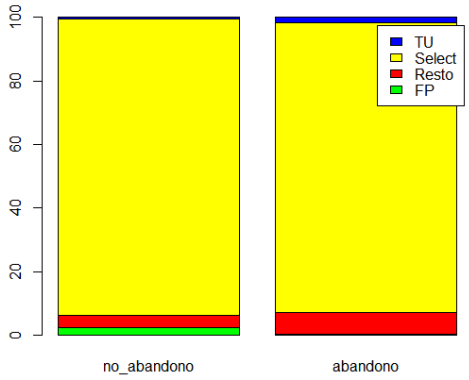
```
T = table(via.acc, abandono)
Ta=colPercents(T)[1:4,]
barplot(Ta, col=c("green", "red", "yellow", "blue"), legend.text=c("FP", "
    Resto", "Select", "TU"), names.arg=c("no_abandono","abandono"))
```



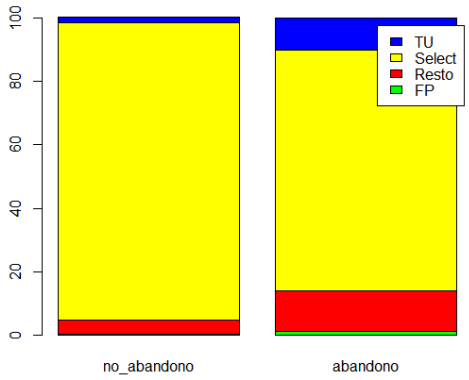
(a) Tecnología y Ciencias Experimentales



(b) Humanas y Sociales



(c) Jurídicas y Económicas



(d) Salud

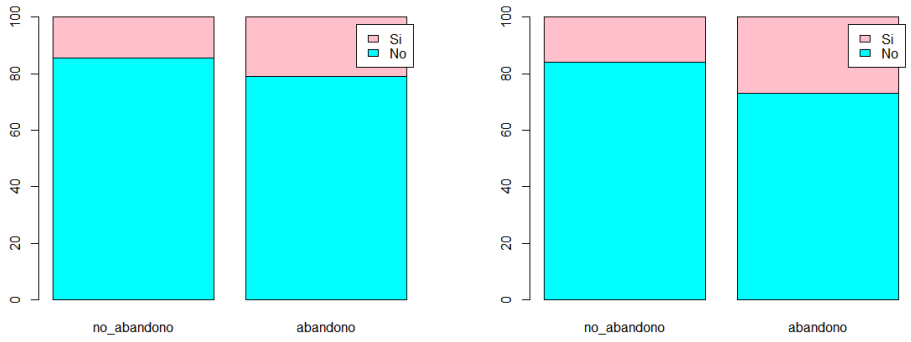
Figura A.3: Vía de acceso, en función de abandono o no abandono de los estudios. Por un lado, se muestran los alumnos que han abandonado los estudios *(columna derecha)* y, por otro, los alumnos que no los han abandonado *(columna izquierda)*. El color **amarillo** se utiliza para representar a los alumnos que han accedido a través de Selectividad, el color **verde** para los estudiantes de Formación profesional, el color **azul** para los Titulados Universitarios y el color **rojo** para el resto.

Si se realiza un contraste de independencia de la *Chi-cuadrado*, en todos los centros la vía de acceso sí influye en el abandono, excepto en la Escuela Superior de Tecnología y Ciencias Experimentales. **En las Facultades de Ciencias Humanas y Sociales, Ciencias Jurídicas y Económicas y Ciencias de la Salud, los estudiantes que más abandonan son los que entran como Titulados Universitarios o a través de otra vía de acceso que no sea Selectividad ni Formación Profesional**.

Un dato sorprendente es que en la Escuela Superior de Tecnología y Ciencias Experimentales y en la Facultad de Ciencias Jurídicas y Económicas, los estudiantes de Formación Profesional prácticamente no abandonan a diferencia de lo que suele pensar.

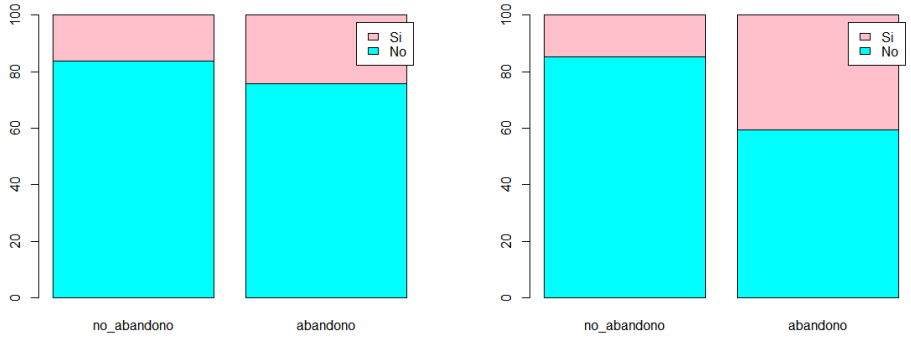
Trabajo realizado en primer curso:

```
T = table(trabPri,abandono)[2:3,]
Ta = colPercents(T)[1:2,]
barplot(Ta, col=c("cyan","pink"), legend.text=c("No","Si"), names=c("no_
    abandono", "abandono"))
```



(a) Tecnología y Ciencias Experimentales

(b) Humanas y Sociales



(c) Jurídicas y Económicas

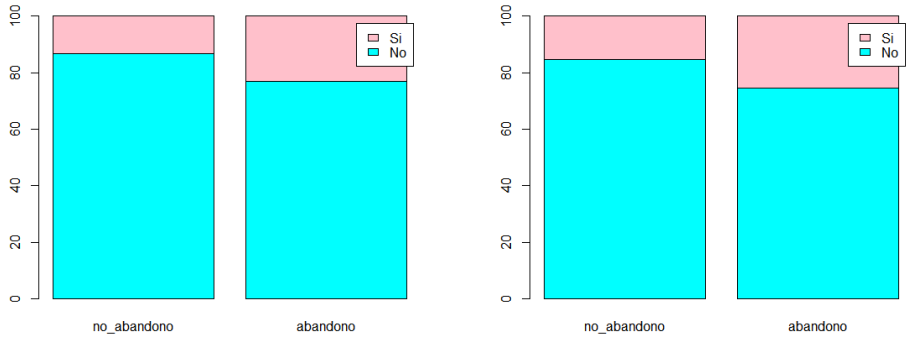
(d) Salud

Figura A.4: Trabajo realizado en primer curso, en función de abandono o no abandono de los estudios. Por un lado, se muestran los alumnos que han abandonado los estudios *(columna derecha)* y, por otro, los alumnos que no los han abandonado *(columna izquierda)*. El color **rosa** representa a los alumnos que trabajan y el color **azul** a los que no trabajan.

Si se realiza un contraste de hipótesis de la *Chi-cuadrado*, el trabajo realizado en primer curso influye en el abandono en todos los centros: **los alumnos que trabajan abandonan en mayor proporción que los alumnos que no trabajan.**

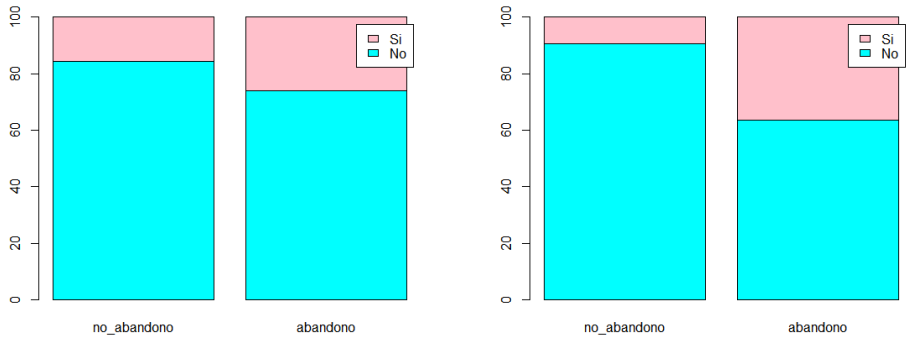
Trabajo realizado en el último curso que el alumno ha realizado matrícula:

```
T = table(trabUlt,abandono)[2:3,]
Ta = colPercents(T)[1:2,]
barplot(Ta, col=c("cyan","pink") legend.text=c("No","Si"), names=c("no_
    abandono", "abandono"))
```



(a) Tecnología y Ciencias Experimentales

(b) Humanas y Sociales



(c) Jurídicas y Económicas

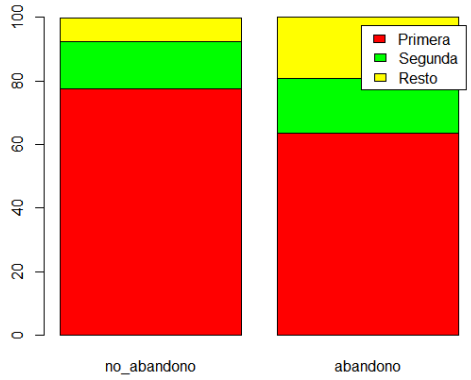
(d) Salud

Figura A.5: Trabajo realizado en último curso que el alumno realizó matrícula, en función de abandono o no abandono de los estudios. Por un lado, se muestran los alumnos que han abandonado los estudios *(columna derecha)* y, por otro, los alumnos que no los han abandonado *(columna izquierda)*. El color **rosa** representa a los alumnos que trabajan y el color **azul** a los que no trabajan.

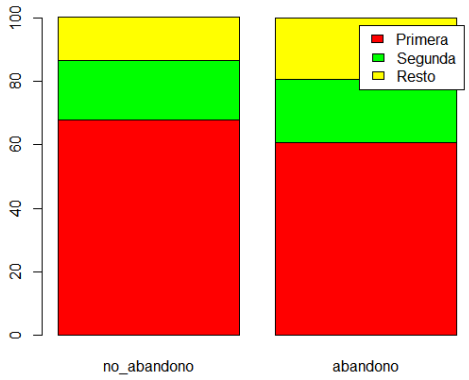
Realizando un contraste de hipótesis de la *Chi-cuadrado*, se puede afirmar que, **el hecho de que el alumno trabaje en el último curso que ha realizado matrícula, sí influye en el abandono** de los estudios en todos los centros.

Orden de preferencia de la titulación en el momento de realizar la preinscripción a la universidad:

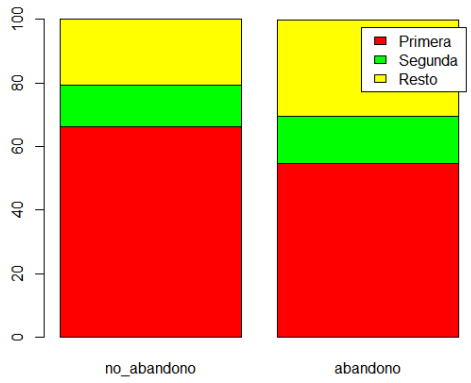
```
T = table(ordered(orden.pref, levels = c("Primera", "Segunda", "Resto")),
    abandono)
Ta = colPercents(T)[1:3,]
barplot(Ta, col = c("red", "green", "yellow"), names.arg = c("no_abandono",
    "abandono"),legend = c("Primera", "Segunda", "Resto"))
```



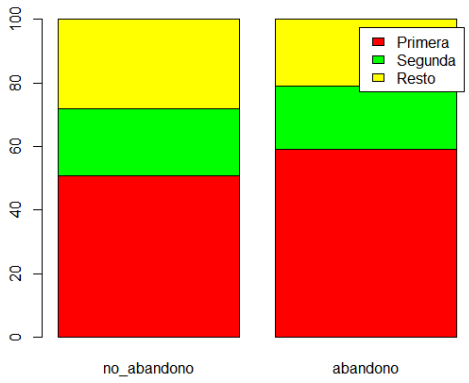
(a) Tecnología y Ciencias Experimentales



(b) Humanas y Sociales



(c) Jurídicas y Económicas



(d) Salud

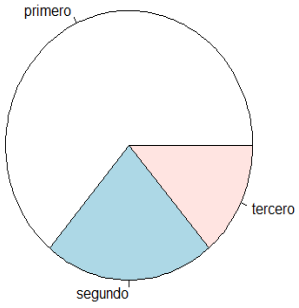
Figura A.6: Orden de preferencia de la titulación, en función de abandono o no abandono de los estudios. Por un lado, se muestran los alumnos que han abandonado los estudios *(columna derecha)* y, por otro, los alumnos que no los han abandonado *(columna izquierda)*. El color **rojo** representa la primera opción, el color **verde** la segunda opción y el color **amarillo** una opción superior.

Realizando un contraste de hipótesis de la *Chi-cuadrado*, se puede afirmar que el orden de preferencia sí influye en el abandono. En la Escuela Superior de Tecnología y Ciencias Experimentales y en las Facultades de Ciencias Humanas y Sociales y Ciencias Jurídicas y Económicas,

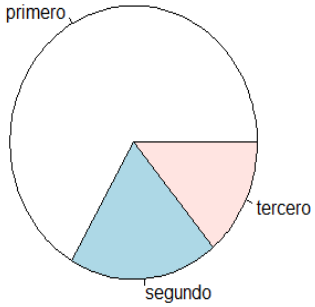
escoger la titulación en un orden superior a la segunda opción aumenta la probabilidad de abandono. En el caso de la Facultad de Ciencias de la Salud, la situación es un poco distinta, debido a que los alumnos del grado en Enfermería suelen escoger como primera opción Medicina, pero al no poder acceder por la nota de corte, entran en Enfermería y la mayor parte no abandona. Además, también hay alumnos del grado en Medicina que no pueden entrar en otras universidades que escogen como primera opción y acaban estudiando en la UJI habiendo elegido esta universidad como segunda opción o incluso una opción superior.

Curso de abandono:

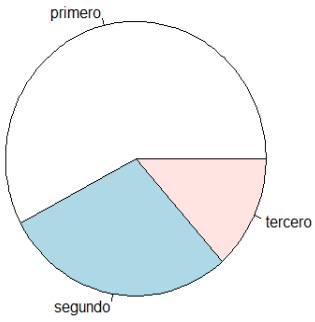
```
primero = length(curso.abandono[curso.abandono == 1 & abandono == 1])
segundo = length(curso.abandono[curso.abandono == 2 & abandono == 1])
tercero = length(curso.abandono[curso.abandono == 3 & abandono == 1])

Sector = c(primero, segundo, tercero)
names(Sector) = c("primero", "segundo", "tercero")
pie(Sector)
```



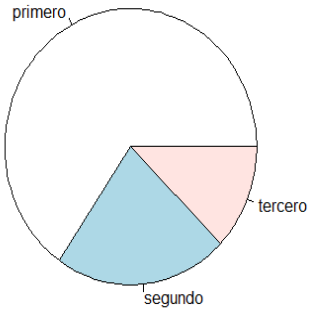
(a) Tecnología y Ciencias Experimentales



(b) Humanas y Sociales



(c) Jurídicas y Económicas



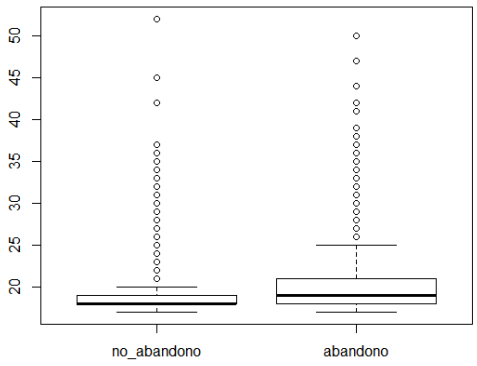
(d) Salud

Figura A.7: Curso en el que los alumnos abandonan los estudios. En **blanco** se representa la proporción de alumnos que abandona en primer curso, en **azul** la proporción de alumnos que abandona en segundo curso y en **rosa** la proporción de alumnos que abandona en tercer curso.

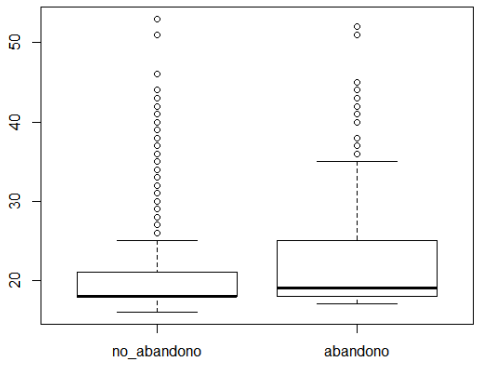
En todos los centros la mayor parte de los alumnos que abandona lo hace en primer curso, mientras que muy pocos abandonan en segundo y tercer curso.

Edad:

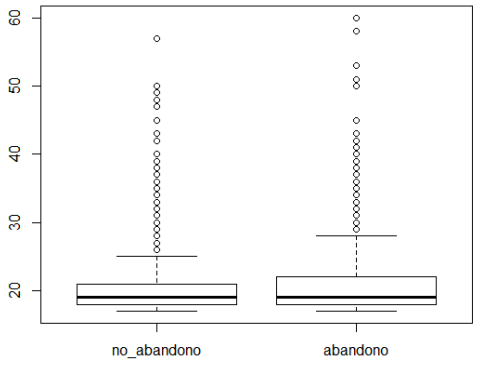
```
boxplot(edad[abandono == 0], edad[abandono == 1], names=c("no_abandono","
    abandono"))
```



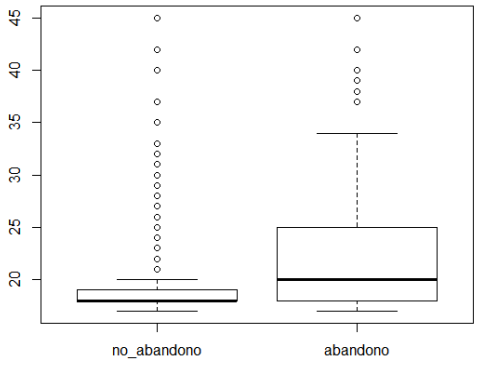
(a) Tecnología y Ciencias Experimentales



(b) Humanas y Sociales



(c) Jurídicas y Económicas



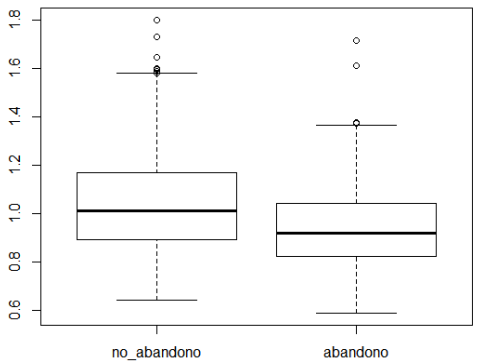
(d) Salud

Figura A.8: Edad de los estudiantes, en función de abandono o no abandono de los estudios. Por un lado, se muestran los alumnos que han abandonado los estudios *(columna derecha)* y, por otro, los alumnos que no los han abandonado *(columna izquierda).*

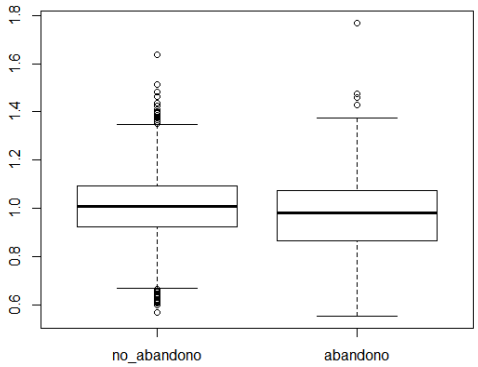
Realizando un contraste de la *t de Student*, se puede afirmar que **las personas mayores abandonan en mayor proporción** en todos los centros.

Nota de acceso:

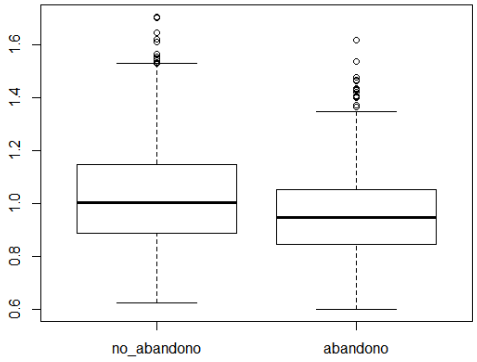
```
boxplot(nota.de.acceso[abandono == 0], nota.de.acceso[abandono == 1], names
    =c("no_abandono","abandono"))
```



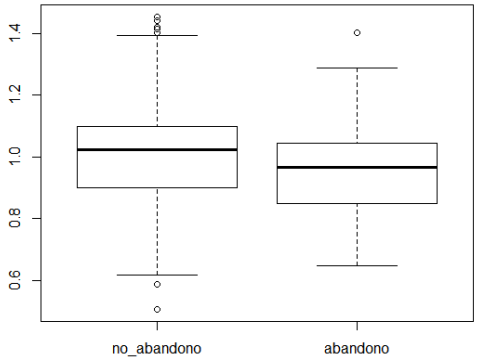
(a) Tecnología y Ciencias Experimentales



(b) Humanas y Sociales



(c) Jurídicas y Económicas



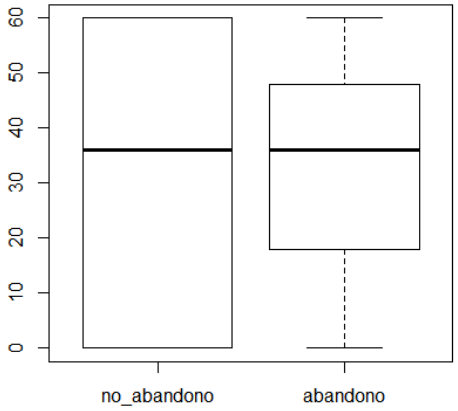
(d) Salud

Figura A.9: Nota de acceso de los estudiantes, en función de abandono o no abandono de los estudios. Por un lado, se muestran los alumnos que han abandonado los estudios *(columna derecha)* y, por otro, los alumnos que no los han abandonado *(columna izquierda).*

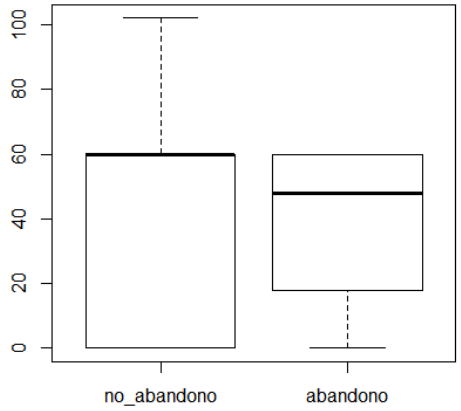
Realizando contrastes de hipótesis de la *t de Student*, la nota de acceso también influye en el abandono. **Los alumnos que abandonan tienen una nota de acceso más baja.**

Créditos presentados a examen en el primer curso de matrícula:

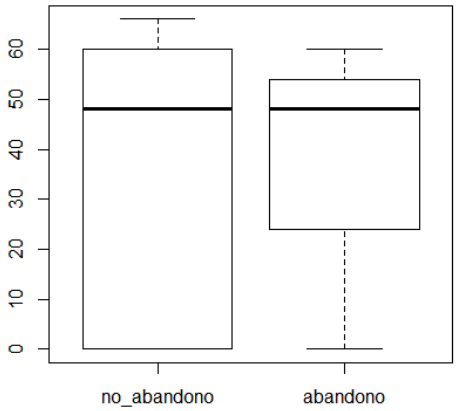
```
boxplot(cred.pres.pri[abandono == 0], cred.pres.pri[abandono == 1], names=c
    ("no_abandono","abandono"))
```



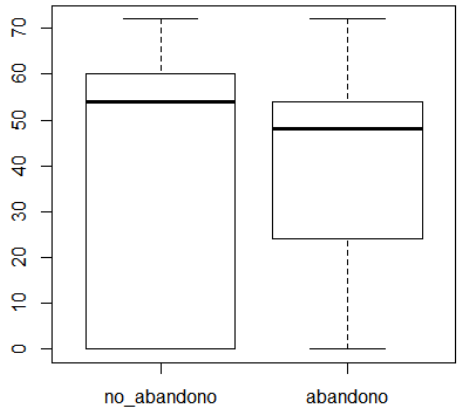
(a) Tecnología y Ciencias Experimentales



(b) Humanas y Sociales



(c) Jurídicas y Económicas



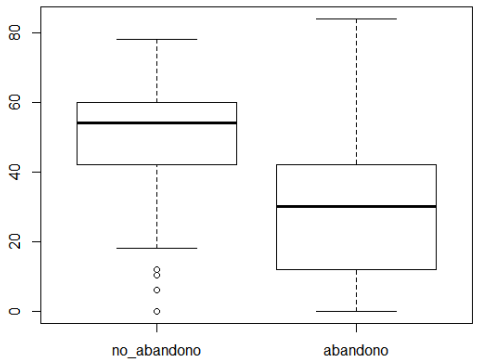
(d) Salud

Figura A.10: Número de créditos presentados a examen en el primer curso de matrícula, en función de abandono o no abandono de los estudios. Por un lado, se muestran los alumnos que han abandonado los estudios *(columna derecha)* y, por otro, los alumnos que no los han abandonado *(columna izquierda).*

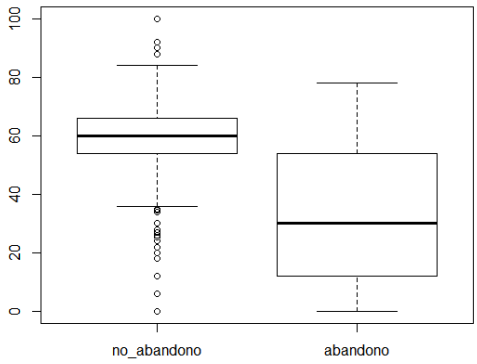
Realizando un contraste de hipótesis de la *t de Student*, se puede afirmar que el número de créditos presentados a examen en primer curso **influye en el abandono en todos los centros, excepto en la Facultad de Ciencias de la Salud**.

Créditos presentados a examen en el último curso que el alumno ha hecho matrícula:

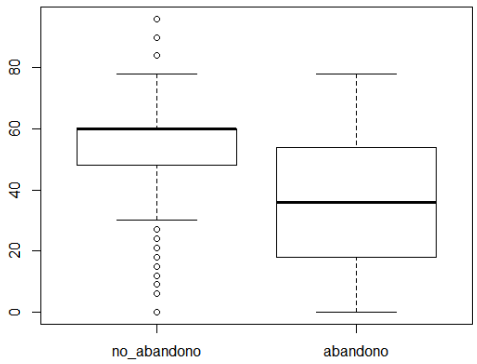
```
boxplot(cred.pres.ultimo[abandono == 0], cred.pres.ultimo[abandono == 1],
    names=c("no_abandono","abandono"))
```



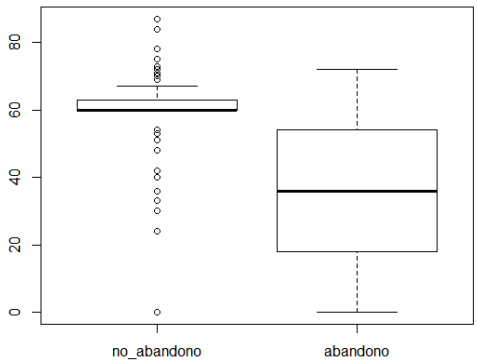
(a) Tecnología y Ciencias Experimentales



(b) Humanas y Sociales



(c) Jurídicas y Económicas



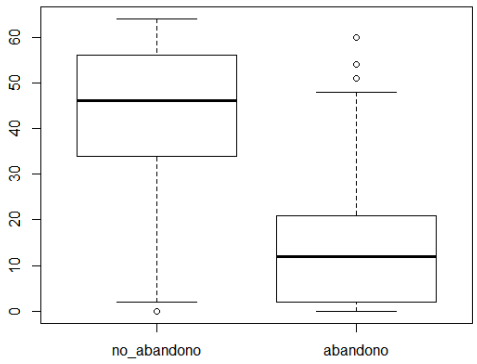
(d) Salud

Figura A.11: Créditos presentados a examen en el último curso de matrícula, en función de abandono o no abandono de los estudios. Por un lado, se muestran los alumnos que han abandonado los estudios *(columna derecha)* y, por otro, los alumnos que no los han abandonado *(columna izquierda).*

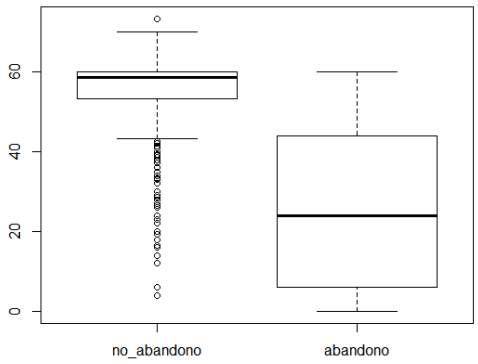
Acompañando los gráficos mediante un contraste de hipótesis de la *t de Student*, se puede afirmar que el número de créditos presentados a examen en el último curso de matrícula **influye en el abandono** en todos los centros. **Los alumnos que abandonan, se presentan a menos exámenes en el curso anterior al abandono que los alumnos que no abandonan**.

Promedio de créditos superados en los cursos en que el alumno ha hecho matrícula:

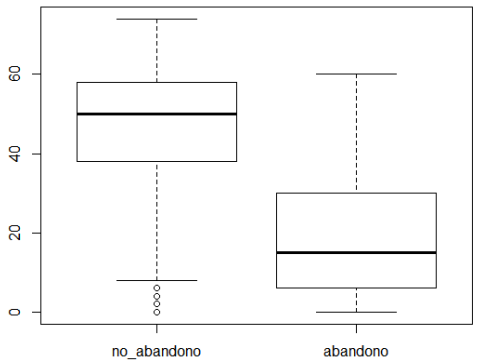
```
boxplot(cred.sup.exam.media[abandono == 0], cred.sup.exam.media[abandono ==
    1], names=c("no_abandono","abandono"))
```



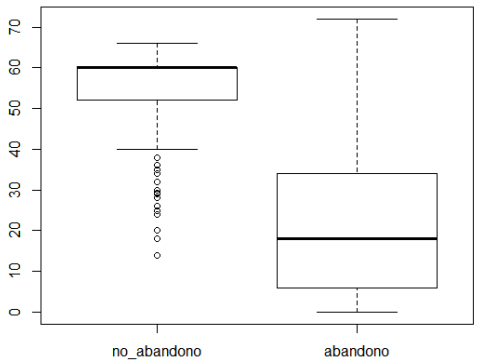
(a) Tecnología y Ciencias Experimentales



(b) Humanas y Sociales



(c) Jurídicas y Económicas



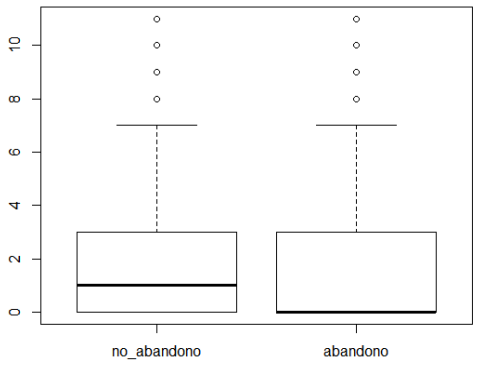
(d) Salud

Figura A.12: Promedio de créditos superados en los cursos en que el alumno ha hecho matrícula, en función de abandono o no abandono de los estudios. Por un lado, se muestran los alumnos que han abandonado los estudios *(columna derecha)* y, por otro, los alumnos que no los han abandonado *(columna izquierda)*.

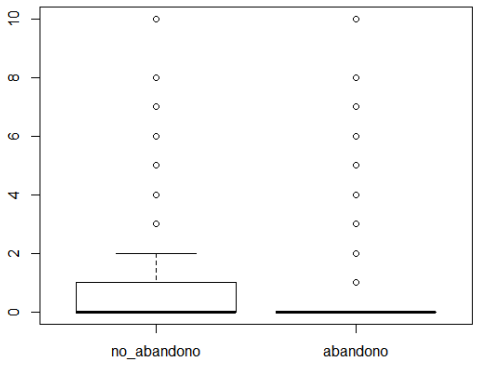
Realizando un contraste de hipótesis de la *t de Student*, se puede afirmar que el promedio de créditos superados en exámenes sí influye en el abandono en todos los centros. **Los alumnos que abandonan superan menos créditos en exámenes que los alumnos que no abandonan.**

Número de asignaturas repetidas en el último curso de matrícula:

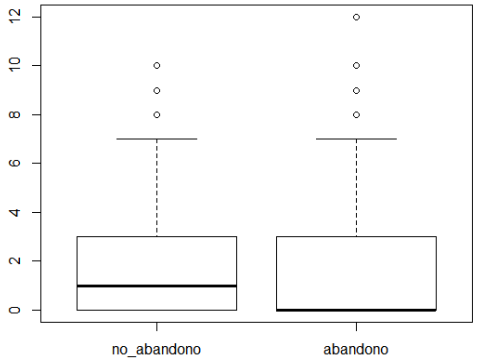
```
boxplot(num.asi.rep.ultimo[abandono == 0], num.asi.rep.ultimo[abandono ==
    1], names=c("no_abandono","abandono"))
```



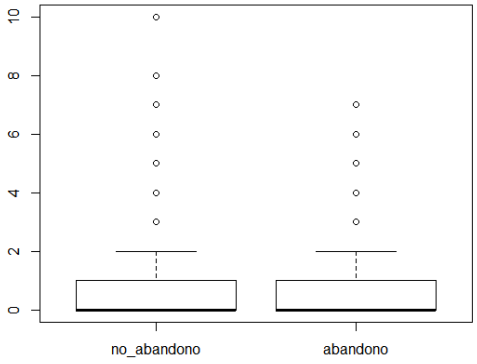
(a) Tecnología y Ciencias Experimentales



(b) Humanas y Sociales



(c) Jurídicas y Económicas



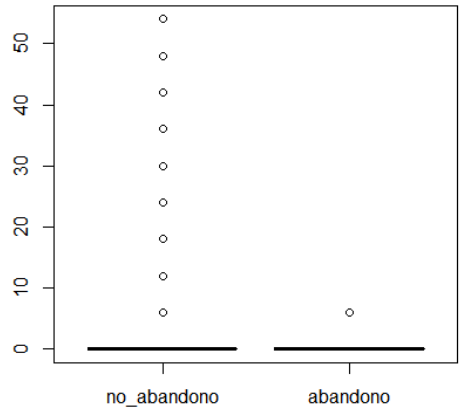
(d) Salud

Figura A.13: Número de asignaturas que el alumno ha vuelto a cursar en el último curso de matrícula, en función de abandono o no abandono de los estudios. Por un lado, se muestran los alumnos que han abandonado los estudios *(columna derecha)* y, por otro, los alumnos que no los han abandonado *(columna izquierda)*.

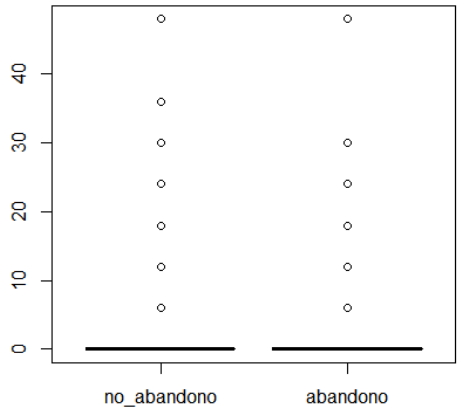
Realizando un contraste de hipótesis de la *t de Student*, se puede afirmar que el número de asignaturas repetidas **sí influye en el abandono en la Facultad de Ciencias Humanas y Sociales y en la Facultad de Ciencias Jurídicas y Económicas**.

Créditos de honor obtenidos en el primer curso de matrícula:

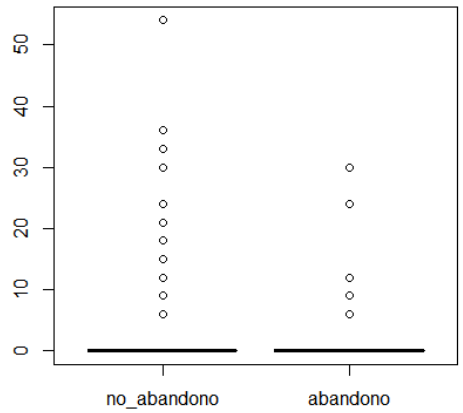
```
boxplot(cred.honor.pri[abandono == 0], cred.honor.pri[abandono == 1], names
    =c("no_abandono","abandono"))
```



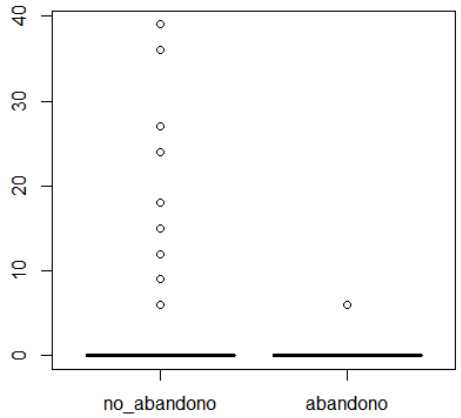
(a) Tecnología y Ciencias Experimentales



(b) Humanas y Sociales



(c) Jurídicas y Económicas



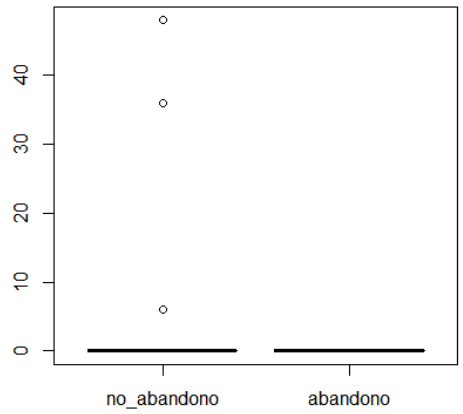
(d) Salud

Figura A.14: Créditos de honor obtenidos en primer curso, en función de abandono o no abandono de los estudios. Por un lado, se muestran los alumnos que han abandonado los estudios *(columna derecha)* y, por otro, los alumnos que no los han abandonado *(columna izquierda)*.

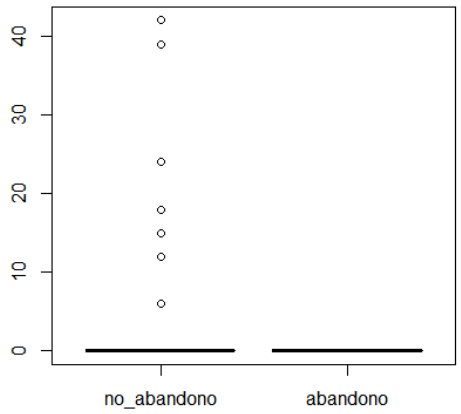
Realizando un contraste de hipótesis de la *t de Student*, se puede afirmar que el número de créditos de honor obtenidos en primer curso sí influye en el abandono en todos los centros. **Los alumnos que no abandonan obtienen más créditos de honor que los alumnos que abandonan.**

Créditos de honor obtenidos en el último curso de matrícula:

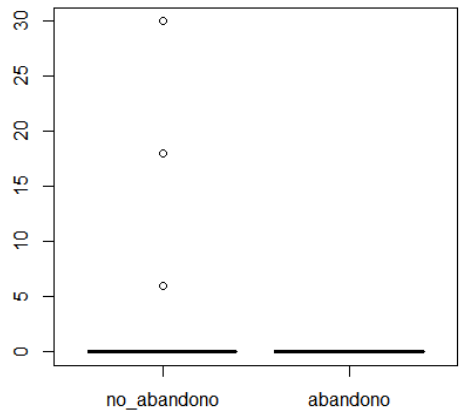
```
boxplot(cred.honor.ultimo[abandono == 0], cred.honor.ultimo[abandono == 1],
    names=c("no_abandono","abandono"), )
```



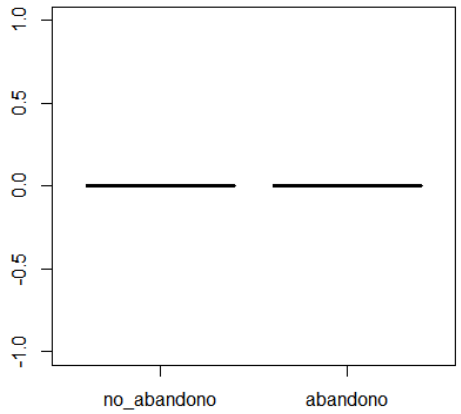
(a) Tecnología y Ciencias Experimentales



(b) Humanas y Sociales



(c) Jurídicas y Económicas



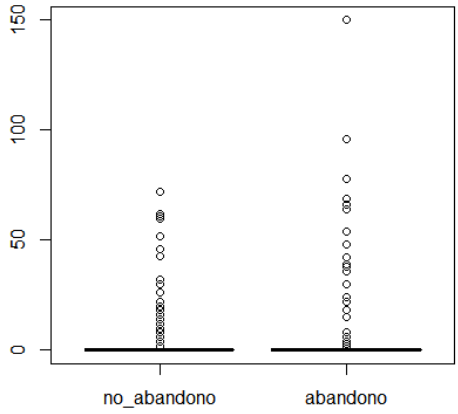
(d) Salud

Figura A.15: Créditos de honor obtenidos en el último curso de matrícula, en función de abandono o no abandono de los estudios. Por un lado, se muestran los alumnos que han abandonado los estudios *(columna derecha)* y, por otro, los alumnos que no los han abandonado *(columna izquierda)*.

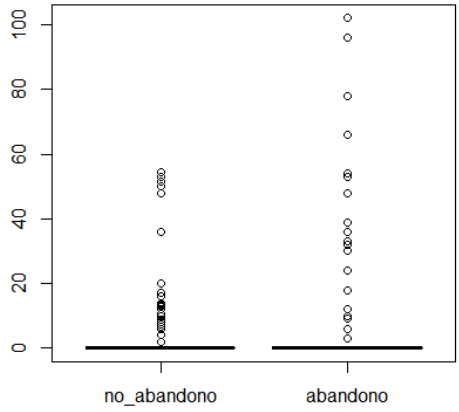
Realizando un contraste de hipótesis de la *t de Student*, se puede afirmar que el número de créditos de honor obtenidos en el último curso de matrícula influye en el abandono en la Facultad de Ciencias Humanas y Sociales. **Los alumnos que no abandonan obtienen más créditos de honor.**

Promedio de créditos reconocidos durante los cursos en los que el alumno ha hecho matrícula:

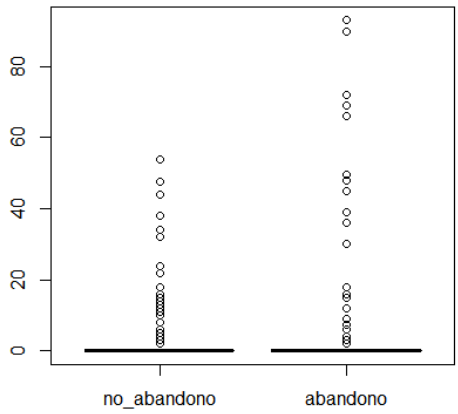
```
boxplot(cred.rec.media[abandono == 0], cred.rec.media[abandono == 1], names
    =c("no_abandono","abandono"))
```



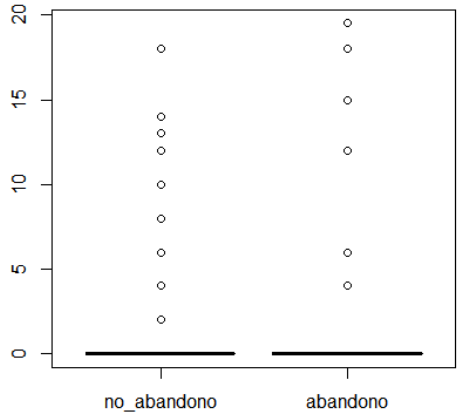
(a) Tecnología y Ciencias Experimentales



(b) Humanas



(c) Jurídicas



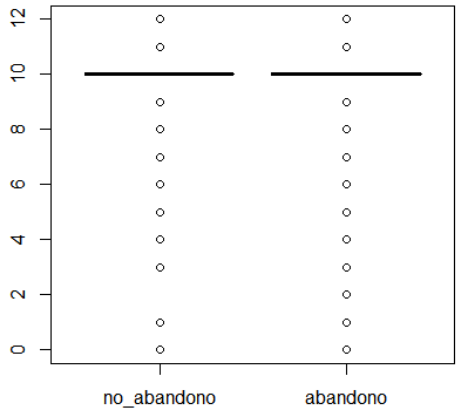
(d) Salud

Figura A.16: Promedio de créditos reconocidos durante los cursos en los que el alumno ha hecho matrícula, en función de abandono o no abandono de los estudios. Por un lado, se muestran los alumnos que han abandonado los estudios *(columna derecha)* y, por otro, los alumnos que no los han abandonado *(columna izquierda)*.

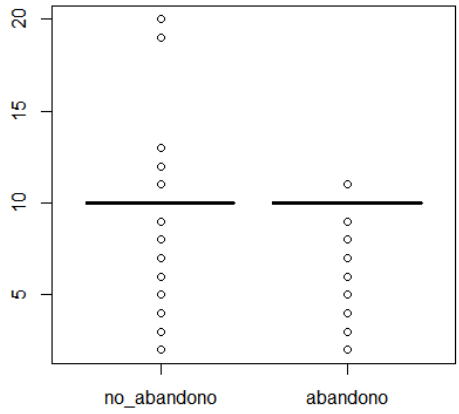
Realizando un contraste de hipótesis de la *t de Student*, se puede afirmar que el promedio de créditos reconocidos **no influye en el abandono** de los estudios en ninguno de los centros.

Número de asignaturas matriculadas en primer curso:

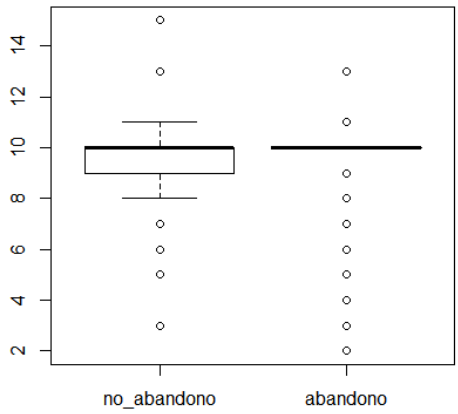
```
boxplot(num.asi.matric.pri[abandono == 0], num.asi.matric.pri[abandono ==
    1], names=c("no_abandono","abandono"))
```



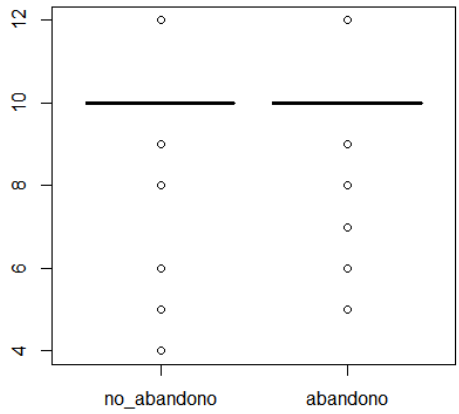
(a) Tecnología y Ciencias Experimentales



(b) Humanas y Sociales



(c) Jurídicas y Económicas



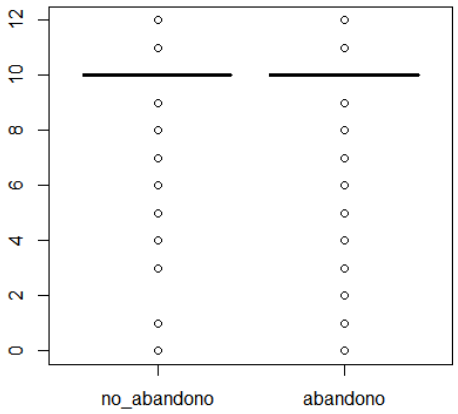
(d) Salud

Figura A.17: Número de asignaturas matriculadas en primer curso, en función de abandono o no abandono de los estudios. Por un lado, se muestran los alumnos que han abandonado los estudios *(columna derecha)* y, por otro, los alumnos que no los han abandonado *(columna izquierda)*.

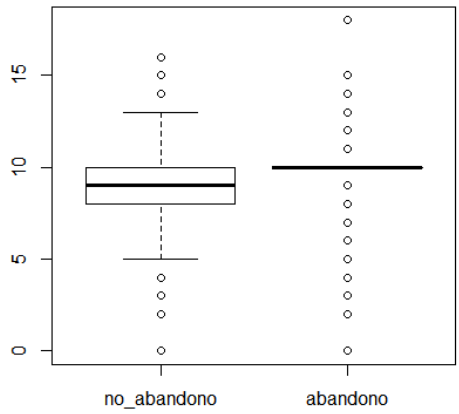
Realizando un contraste de hipótesis de la *t de Student*, se puede afirmar que el número de asignaturas matriculadas en primer curso **no influye en el abandono** de los estudios en ninguno de los centros.

Número de asignaturas matriculadas en el último curso de matrícula:

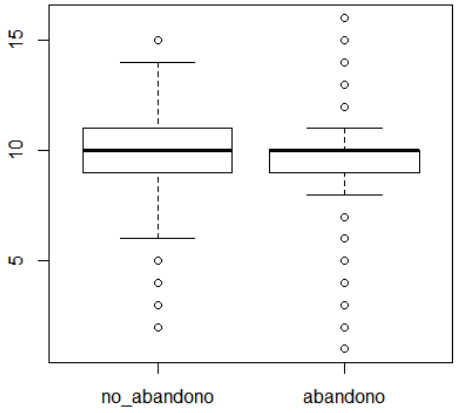
```
boxplot(num.asi.matric.ultimo[abandono == 0], num.asi.matric.ultimo[
    abandono == 1], names=c("no_abandono","abandono"))
```



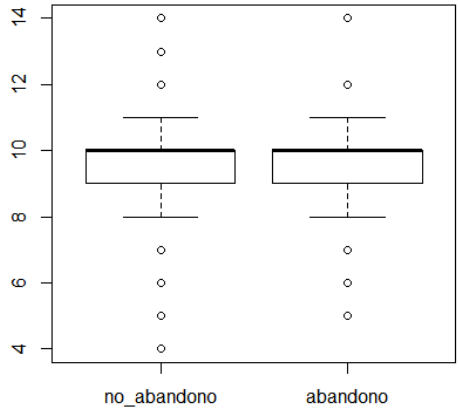
(a) Tecnología y Ciencias Experimentales



(b) Humanas y Sociales



(c) Jurídicas y Económicas



(d) Salud

Figura A.18: Número de asignaturas matriculadas en el último curso de matrícula, en función de abandono o no abandono de los estudios. Por un lado, se muestran los alumnos que han abandonado los estudios *(columna derecha)* y, por otro, los alumnos que no los han abandonado *(columna izquierda).*

Realizando un contraste de hipótesis de la *t de Student*, se puede afirmar que **el número de asignaturas matriculadas en el último curso influye en el abandono en todos los centros, excepto en la Escuela Superior de Tecnología y Ciencias Experimentales.**

A.2. Descripción de la muestra sobre la inserción

Para realizar este estudio, se han eliminado los alumnos de la muestra que no están buscando empleo, bien porque continúan estudiando o por otro motivo.

Cabe remarcar que en la Facultad de Ciencias de la Salud únicamente hay alumnos del grado en Psicología, pero no los hay ni de Enfermería ni de Medicina, pues en 2014 todavía no habían egresados en ninguna de las dos titulaciones.

Los códigos que se muestran en esta sección han sido aplicados para cada una de las cuatro muestras, una por cada centro.

Proporción de alumnos de cada centro que ha encontrado trabajo:

```
trabajoSI = length(na.omit(trabajo[trabajo == 1]))
trabajoNO = length(na.omit(trabajo[trabajo == 0]))
Sector = c(trabajoSI, trabajoNO)
names(Sector) = c("SI", "NO")
pie(Sector, col = c("purple", "cyan"))
```

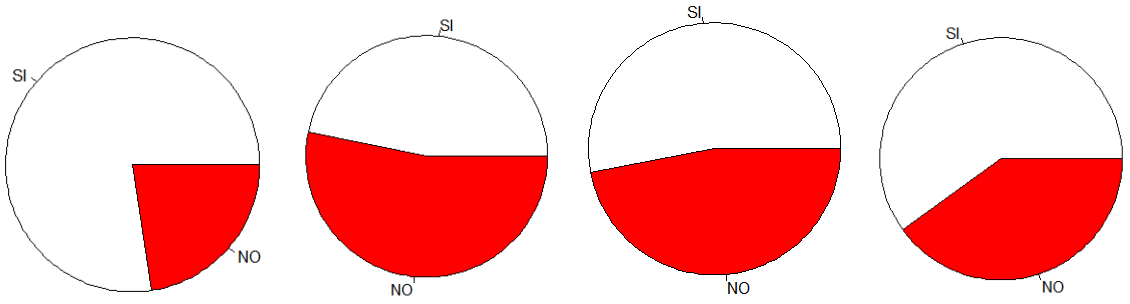
(a) Tecnología y Ciencias Experimentales (b) Humanas y Sociales (c) Jurídicas y Económicas (d) Salud

Figura A.19: Proporción de alumnos de cada centro que ha encontrado un empleo. De color **azul**, se representa la propoción de alumnos que ha conseguido un empleo y de color **morado**, la proporción de alumnos que no lo ha conseguido.

Se observa que los alumnos que encuentran empleo más fácilmente son los alumnos de la Escuela Superior de Tecnología y Ciencias Experimentales (58 %), seguidos de los alumnos de la Facultad de Ciencias Jurídicas y Económicas (50 %), a continuación, los estudiantes de la Facultad de Ciencias Humanas y Sociales (38 %) y, finalmente, los alumnos de Ciencias de la Salud (18 %).

Proporción de alumnos que necesita la titulación cursada, de entre los que han conseguido un empleo:

```
trabajo_con_titulacion = length(na.omit(trabajo[trabajo == 1 & necesita.
    titulacion == 1]))
trabajo_sin_titulacion = length(na.omit(trabajo[trabajo == 1 & necesita.
    titulacion == 0]))
Sector = c(trabajo_con_titulacion, trabajo_sin_titulacion)
names(Sector) = c("SI", "NO")
pie(Sector, col = c("white", "red"))
```



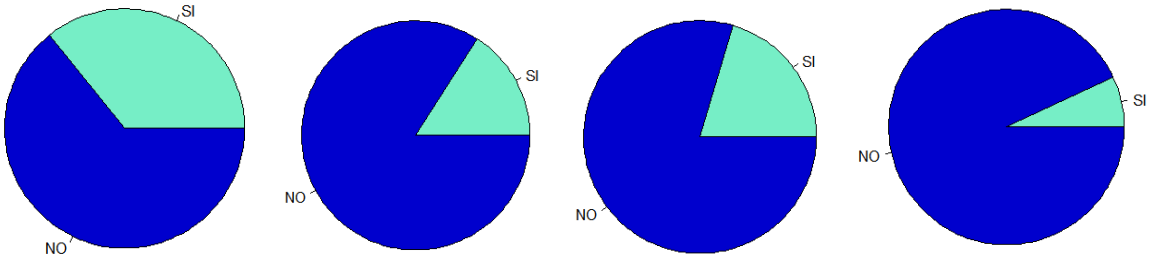
(a) Tecnología y Ciencias Experimentales (b) Humanas y Sociales (c) Jurídicas y Económicas (d) Salud

Figura A.20: Proporción de alumnos que necesita la titulación cursada, de entre los que han conseguido un empleo. De color **blanco**, se representa la proporción de alumnos que ha conseguido un empleo que requiere la titulación cursada y de color **rojo**, la proporción de alumnos que ha conseguido un empleo que no requiere la titulación cursada.

De los alumnos que trabajan, los alumnos que más fácilmente consiguen un trabajo que requiere la titulación cursada, son los de la Escuela Superior de Tecnología y Ciencias Experimentales (77 %), seguidos de los alumnos de la Facultad de Ciencias de la Salud (60 %), a continuación, los estudiantes de la Facultad de Ciencias Jurídicas y Económicas (52 %) y, finalmente, los estudiantes de la Facultad de Ciencias Humanas y Sociales (46 %).

Proporción de alumnos que ha conseguido un trabajo que requiere el título obtenido, con respecto a todos los alumnos egresados en ese centro:

```
trabajo_titulacion = length(na.omit(trabajo.titulacion[trabajo.titulacion
    == 1]))
resto = length(na.omit(trabajo.titulacion[trabajo.titulacion == 0 ]))
Sector = c(trabajo_titulacion, resto )
names(Sector) = c("SI", "NO")
pie(Sector, col = c("aquamarine2", "blue3"))
```



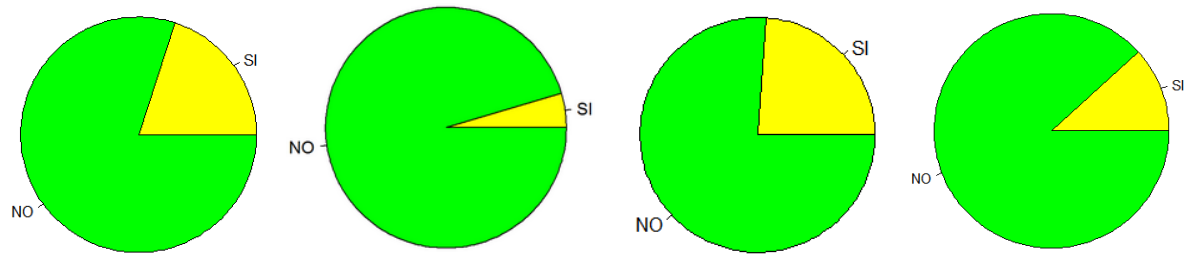
(a) Tecnología y Ciencias Experimentales (b) Humanas y Sociales (c) Jurídicas y Económicas (d) Salud

Figura A.21: Proporción de alumnos que ha conseguido un trabajo que requiere el título obtenido, con respecto a todos los alumnos egresados en ese centro. De color **azul claro**, se representa la proporción de alumnos que ha conseguido un empleo que requiere la titulación cursada y de color **azul oscuro**, la proporción de alumnos que ha conseguido un empleo que no requiere la titulación cursada o no ha conseguido ningún empleo.

Los alumnos que más fácilmente consiguen un empleo que requiere la titulación cursada, una vez finalizados sus estudios de grado, son los alumnos de la Escuela Superior de Tecnología y Ciencias Experimentales (43 %), seguidos de los alumnos de la Facultad de Ciencias Jurídicas y Económicas (20 %), a continuación, los alumnos de la Facultad de Ciencias Humanas y Sociales (17 %) y, finalmente, los alumnos de la Facultad de Ciencias de la Salud (9 %).

Proporción de alumnos, respecto de todos los que realizaron la estancia en prácticas, que se quedó trabajando en la empresa donde realizó las prácticas.

```
trabajo_empresa_practicas = length(na.omit(trabajo.empresa.practicas[
    trabajo.empresa.practicas == 1 & (practicas.obligatorias == 1 |
    practicas.extracurriculares == 1)]))
no_trabajo_empresa_practicas = length(na.omit(trabajo.empresa.practicas[
    trabajo.empresa.practicas == 0 & (practicas.obligatorias == 1 |
    practicas.extracurriculares == 1)]))
Sector = c(trabajo_empresa_practicas, no_trabajo_empresa_practicas)
names(Sector) = c("SI", "NO")
pie(Sector, col = c("yellow", "green"))
```



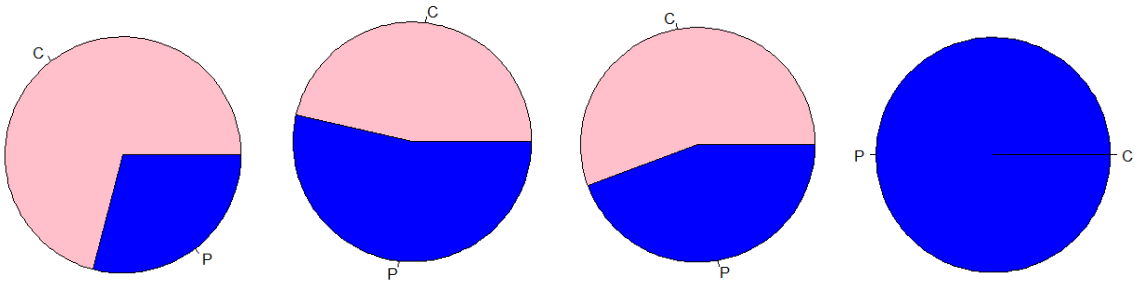
(a) Tecnología y Ciencias Experimentales (b) Humanas y Sociales (c) Jurídicas y Económicas (d) Salud

Figura A.22: De todos los alumnos que realizaron la estancia en prácticas, proporción de alumnos que permaneció trabajando una vez finalizadas las prácticas. De color **amarillo**, se muestra la proporción de alumnos que continuó trabajando y de color **verde**, la proporción de los que no lo hicieron.

Se puede observar que, los alumnos que permanecen en mayor proporción en la empresa donde realizaron la estancia en prácticas, son los estudiantes que pertenecen a la Facultad de Ciencias Jurídicas y Económicas (23 %), seguidos de los alumnos de la Escuela Superior de Tecnología y Ciencias Experimentales (20 %), a continuación, los estudiantes de la Facultad de Ciencias de la Salud (11 %) y, finalmente, los estudiantes de la Facultad de Ciencias Humanas y Sociales (11 %).

Tipo de jornada laboral que tienen los estudiantes que han conseguido un empleo:

```
trabajo_jornada_completa = length(na.omit(jornada[jornada == 1 & trabajo.
    titulacion ==1]))
trabajo_jornada_parcial = length(na.omit(jornada[jornada == 2 & trabajo.
    titulacion == 1]))
Sector = c(trabajo_jornada_completa, trabajo_jornada_parcial)
names(Sector) = c("C", "P")
pie(Sector, col = c("pink", "blue"))
```



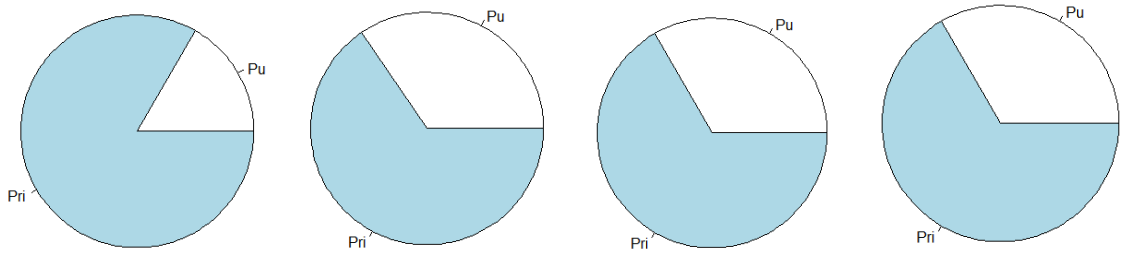
(a) Tecnología y Ciencias Experimentales (b) Humanas y Sociales (c) Jurídicas y Económicas (d) Salud

Figura A.23: Tipo de jornada laboral que tienen los estudiantes que han conseguido un empleo. De color **rosa**, se representan los estudiantes que trabajan a jornada completa y de color **azul**, los estudiantes que trabajan a jornada parcial.

Los estudiantes de la Escuela Superior de Tecnología y Ciencias Experimentales son los que más encuentran trabajo a jornada completa (70 %), seguidos de los alumnos de la Facultad de Ciencias Jurídicas y Económicas (55 %), a continuación, los de la Facultad de Ciencias Humanas y Sociales (46 %) y, finalmente, los de la Facultad de Ciencias de la Salud (0 %).

Tipo de empresa en la que los alumnos encuentran trabajo:

```
publica = length(na.omit(tipo.empresa[tipo.empresa == 1 & trabajo.
    titulacion == 1]))
privada = length(na.omit(tipo.empresa[tipo.empresa == 2 & trabajo.
    titulacion == 1]))
Sector = c(publica, privada)
names(Sector) = c("Pu", "Pri")
pie(Sector)
```



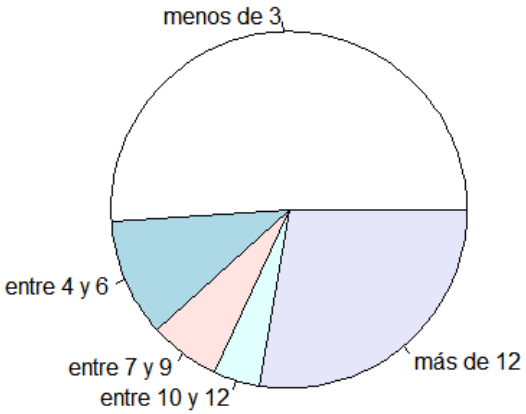
(a) Tecnología y Ciencias Experimentales (b) Humanas y Sociales (c) Jurídicas y Económicas (d) Salud

Figura A.24: Tipo de empresa en la que los alumnos encuentran trabajo. De color **blanco**, se representan los estudiantes que trabajan en empresas públicas y de color **azul**, los estudiantes que trabajan en empresas privadas.

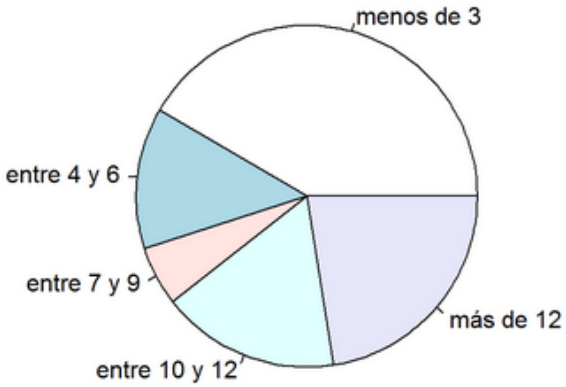
Los estudiantes que más consiguen trabajo en empresas privadas son los de la Escuela Superior de Tecnología y Ciencias Experimentales (83 %). En el resto de centros, la proporción es prácticamente la misma (sobre un 66 %).

Meses en encontrar el primer empleo:

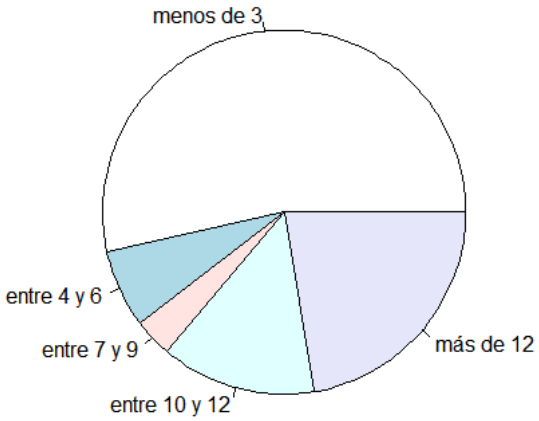
```
hasta_tres_meses = length(na.omit(tiempo.encontrar.empleo.categorica[tiempo
    .encontrar.empleo.categorica == 1 ]))
entre_cuatro_y_seis = length(na.omit(tiempo.encontrar.empleo.categorica[
    tiempo.encontrar.empleo.categorica == 2 ]))
entre_siete_y_nueve = length(na.omit(tiempo.encontrar.empleo.categorica[
    tiempo.encontrar.empleo.categorica == 3 ]))
entre_diez_y_doce = length(na.omit(tiempo.encontrar.empleo.categorica[
    tiempo.encontrar.empleo.categorica == 4 ]))
mas_de_doce = length(na.omit(tiempo.encontrar.empleo.categorica[tiempo.
    encontrar.empleo.categorica == 5 ]))
Sector = c(hasta_tres_meses, entre_cuatro_y_seis, entre_siete_y_nueve,
    entre_diez_y_doce, mas_de_doce )
names(Sector) = c("menos de 3", "entre 4 y 6", "entre 7 y 9", "entre 10 y
    12", "mas de 12" )
pie(Sector)
```



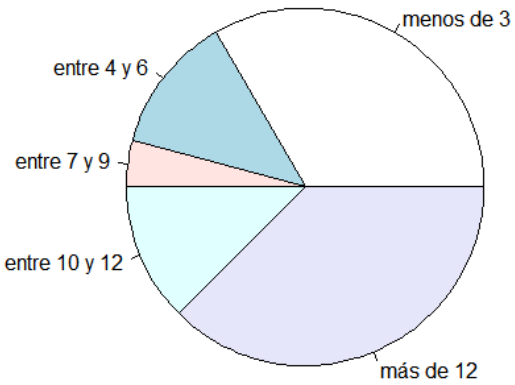
(a) Tecnología y Ciencias Experimentales



(b) Humanas y Sociales



(c) Jurídicas y Económicas



(d) Salud

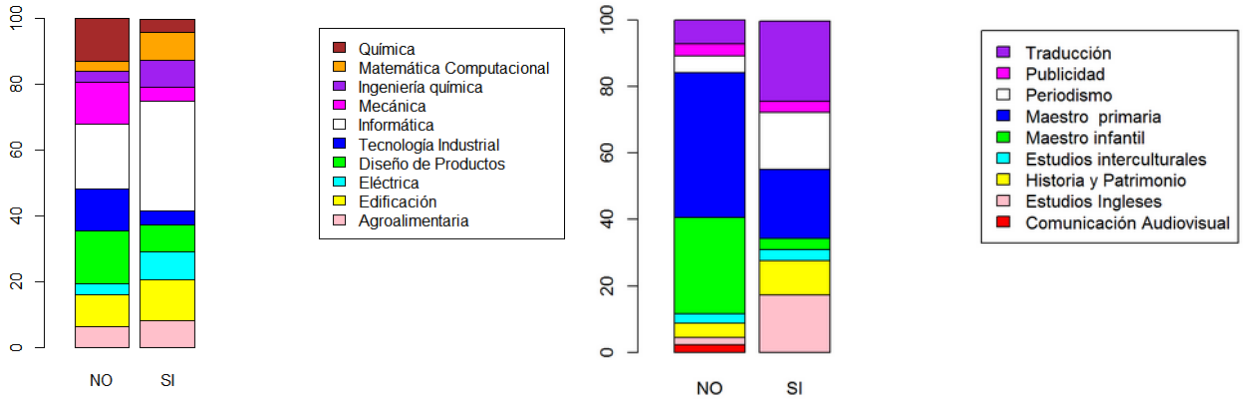
Figura A.25: Tiempo que tardan los alumnos en encontrar su primer empleo.

Los alumnos que menos tiempo tardan en encontrar su primer empleo son los estudiantes de la Facultad de Ciencias Jurídicas y Económicas, seguidos de los de la Escuela Superior de

Tecnología y Ciencias Experimentales, a continuación, los alumnos de la Facultad de Ciencias Humanas y Sociales y, por último, los alumnos de la Facultad de Ciencias de la Salud.

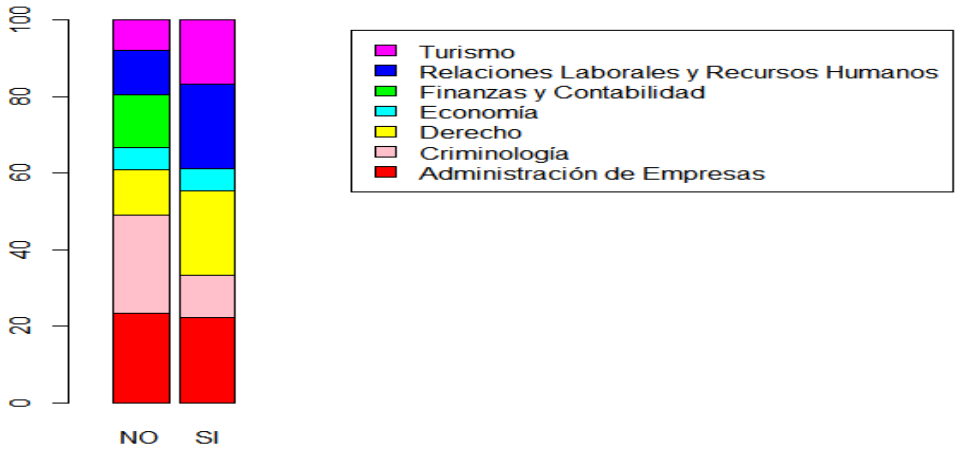
Proporción de alumnos que ha encontrado un empleo por cada una de las titulaciones. Se recuerda que no se muestra el gráfico para la Facultad de Ciencias de la Salud, pues únicamente contiene una titulación que es la de Psicología. Se muestra el código únicamente para la Escuela Superior de Tecnología y Ciencias Experimentales:

```
T = table(titulacion, trabajo.titulacion)
Ta=colPercents(T)[1:10,]
barplot(Ta, names.arg=c("NO", "SI"), xlim=c(0,9), legend=c("Agroalimentaria
    ", "Edificacion","Electrica", "Diseno de Productos", "Tecnologia
    Industrial", "Informatica", "Mecanica","Ingenieria quimica","Matematica
    Computacional", "Quimica" ), col = c("pink", "yellow","cyan", "green",
     "blue", "white", "magenta", "purple", "orange", "brown"))
```



(a) Tecnología y Ciencias Experimentales

(b) Humanas y Sociales



(c) Jurídicas y Económicas

Figura A.26: Proporción de alumnos que ha encontrado empleo por cada una de las titulaciones. La columna derecha, corresponde a los alumnos que han encontrado un empleo que requiere el título obtenido y, la columna izquierda, a los estudiantes que no han encontrado empleo o han encontrado un empleo pero no requiere la titulación cursada.

En la Escuela Superior de Tecnología y Ciencias Experimentales, las titulaciones cuyos egresados, en proporción, encuentran antes empleo son:

- Ingeniería informática.
- Ingeniería química.
- Matemática Computacional.
- Ingeniería eléctrica.

En la Facultad de Ciencias Humanas y Sociales, las titulaciones cuyos egresados, en proporción, encuentran antes empleo son:

- Traducción.
- Periodismo.
- Estudios ingleses.
- Historia y Patrimonio.

En la Facultad de Ciencias Jurídicas y Económicas, las titulaciones cuyos egresados, en proporción, encuentran antes empleo son:

- Turismo.
- Relaciones laborales y Recursos Humanos.
- Derecho.

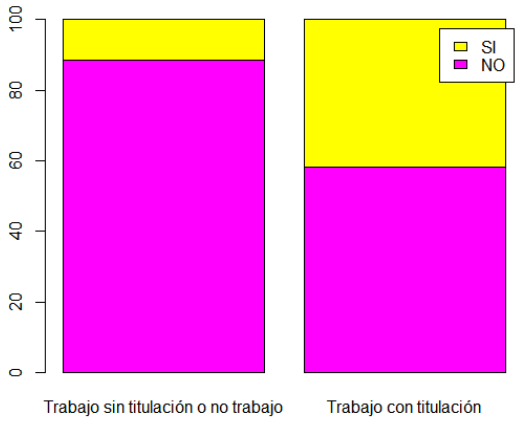
También se han realizado gráficos por cada una de las variables, para comprobar si existen diferencias entre encontrar un empleo que requiere el título obtenido y no encontrarlo.

Cabe destacar que como la Facultad de Ciencias de la Salud sólo contiene a los alumnos del grado en Psicología y muy pocos de ellos han conseguido un empleo que requiere el título obtenido, aunque en los gráficos se observen diferencias entre los alumnos que encuentran un empleo que requiere la titulación cursada y los que no, los contrastes de hipótesis no rechazarán la independencia para esa variable pues no existe suficiente evidencia.

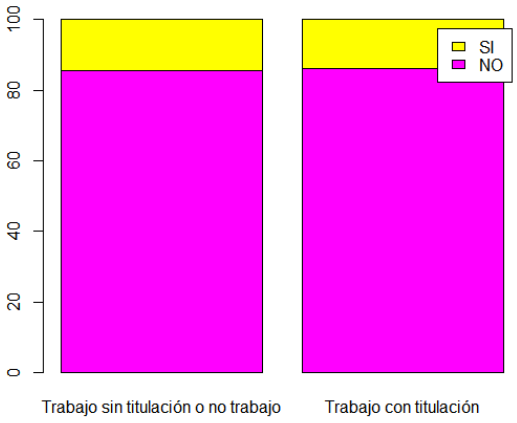
A continuación, se muestran los gráficos realizados para cada una de las variables.

Prácticas extracurriculares:

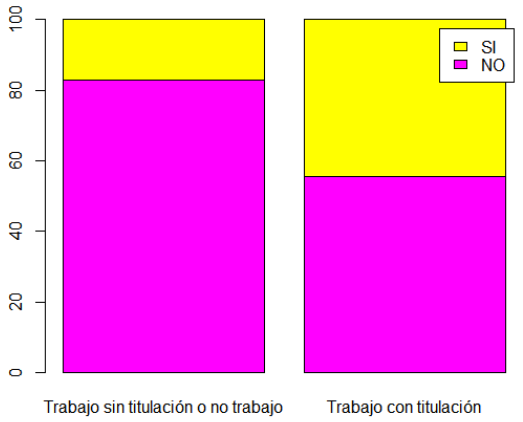
```
T = table(practicas.extracurriculares, trabajo.titulacion)
Ta=colPercents(T)[1:2,]
barplot(Ta, col=c("magenta","yellow"), names.arg=c("Trabajo sin titulacion
    o no trabajo","Trabajo con titulacion"), legend=c("NO", "SI"))
```



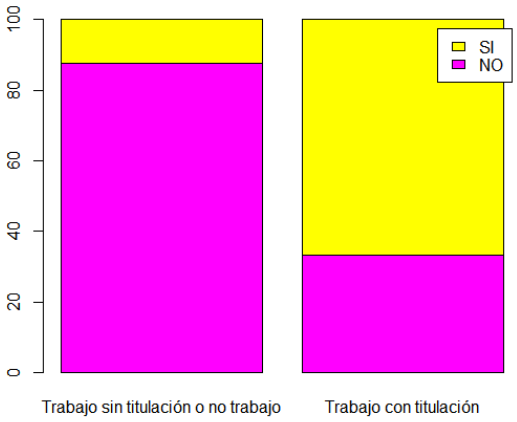
(a) Tecnología y Ciencias Experimentales



(b) Humanas y Sociales



(c) Jurídicas y Económicas



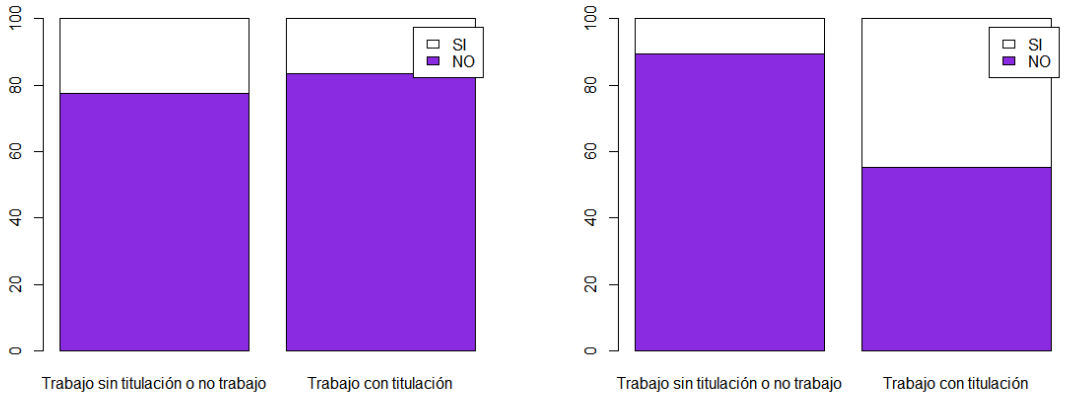
(d) Salud

Figura A.27: Prácticas extracurriculares en función de encontrar un empleo que requiere el título obtenido o no encontrarlo. Por un lado, se muestran los alumnos que han conseguido un trabajo que requiere la titulación cursada *(columna derecha)* y, por otro, los alumnos que no han conseguido un empleo o han conseguido un empleo que no requiere el título obtenido *(columna izquierda)*. El color **amarillo** se utiliza para representar a los alumnos que han realizado prácticas extracurriculares y el color **rosa** para representar a los alumnos que no las han realizado.

Si además de observar los gráficos se realiza un contraste de hipótesis de la *Chi-cuadrado*, se puede afirmar que las prácticas extracurriculares **sí facilitan la obtención de empleo en la Escuela Superior de Tecnología y Ciencias Experimentales y en la Facultad de Ciencias Jurídicas y Económicas**.

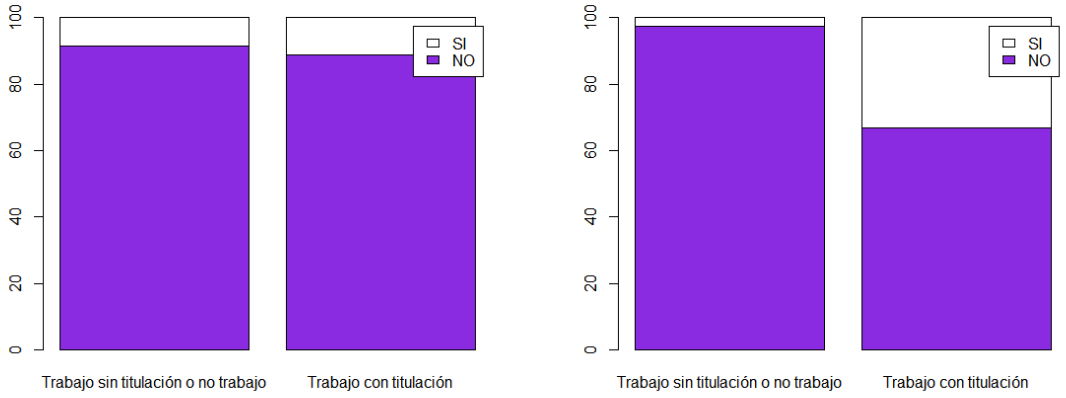
Erasmus estudios:

```
T = table(erasmus.estudios, trabajo.titulacion)
Ta=colPercents(T)[1:2,]
barplot(Ta, col=c("blueviolet","white"), names.arg=c("Trabajo sin titula-
    cion o no trabajo","Trabajo con titulacion"), legend=c("NO", "SI"))
```



(a) Tecnología y Ciencias Experimentales

(b) Humanas y Sociales



(c) Jurídicas y Económicas

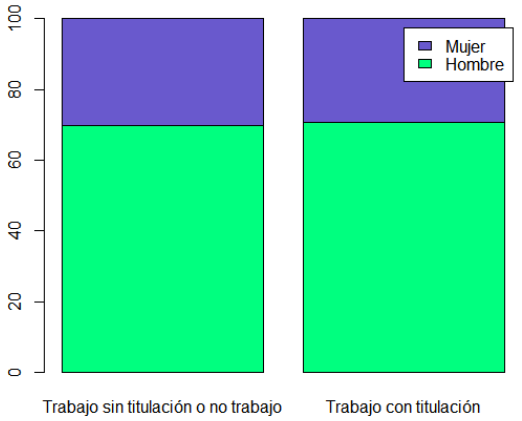
(d) Salud

Figura A.28: Erasmus por motivo de estudios, en función de encontrar un empleo que requiere el título obtenido o no encontrarlo. Por un lado, se muestran los alumnos que han conseguido un trabajo que requiere la titulación cursada *(columna derecha)* y, por otro, los alumnos que no han conseguido un empleo o han conseguido un empleo que no requiere el título obtenido *(columna izquierda)*. El color **blanco** se utiliza para representar a los alumnos que han realizado Erasmus por motivo de estudios y el color **morado** para representar a los alumnos que no lo han realizado.

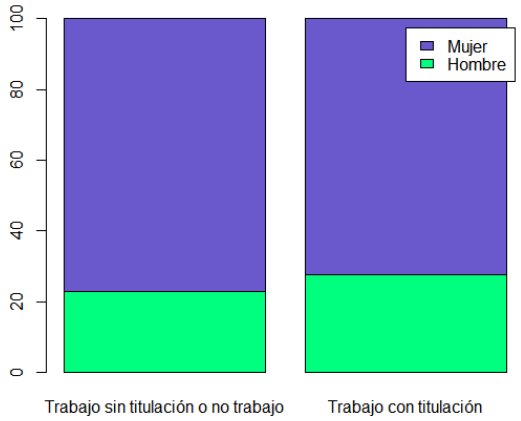
Si además de observar los gráficos se realiza un contraste de hipótesis de la *Chi-cuadrado*, se puede afirmar que **realizar Erasmus por motivo de estudios sí influye a la hora de encontrar un empleo que requiere la titulación cursada para los alumnos de la Facultad de Ciencias Humanas y Sociales**.

Sexo:

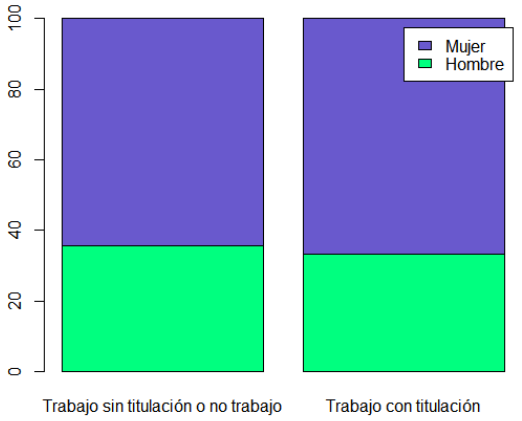
```
T = table(sexo, trabajo.titulacion)
Ta=colPercents(T)[1:2,]
barplot(Ta, col=c("springgreen","slateblue3"), names.arg=c("Trabajo sin
    titulacion o no trabajo","Trabajo con titulacion"), legend=c("Hombre",
    "Mujer"))
```



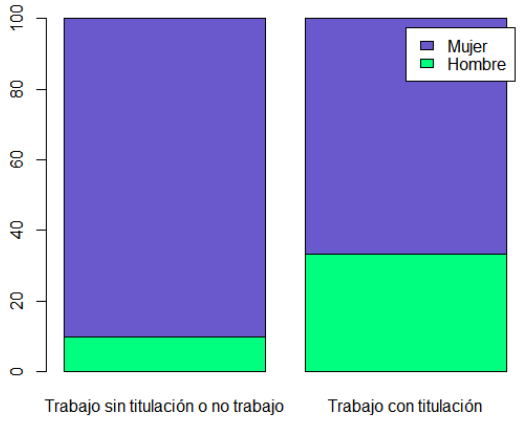
(a) Tecnología y Ciencias



(b) Humanas



(c) Jurídicas



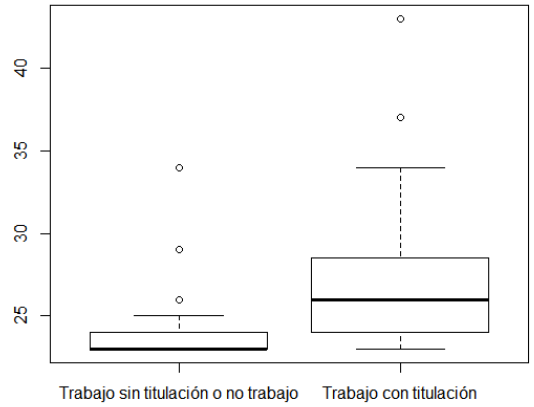
(d) Salud

Figura A.29: Sexo, en función de encontrar un empleo que requiere el título obtenido o no encontrarlo. Por un lado, se muestran los alumnos que han conseguido un trabajo que requiere la titulación cursada *(columna derecha)* y, por otro, los alumnos que no han conseguido un empleo o han conseguido un empleo que no requiere el título obtenido *(columna izquierda)*. El color **morado** se utiliza para representar a las mujeres y el color **verde** para representar a los hombres.

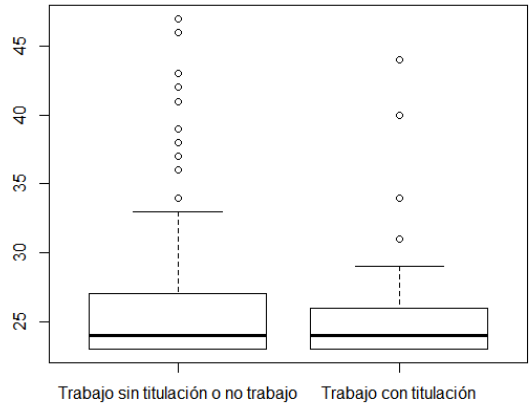
Si se realiza un contraste de hipótesis de la *Chi-cuadrado*, se puede afirmar que el sexo **no influye** en ningún centro a la hora de encontrar un empleo que requiere la titulación cursada.

Edad:

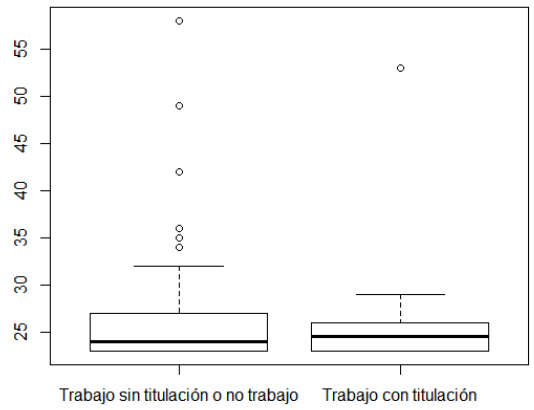
```
boxplot(edad[trabajo.titulacion == 0], edad[trabajo.titulacion == 1], names
    =c("Trabajo sin titulacion o no trabajo","Trabajo con titulacion"))
```



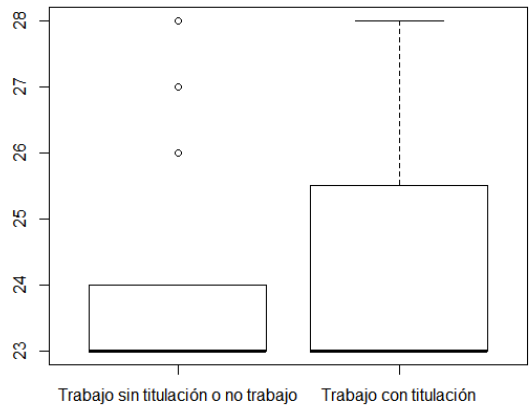
(a) Tecnología y Ciencias



(b) Humanas



(c) Jurídicas



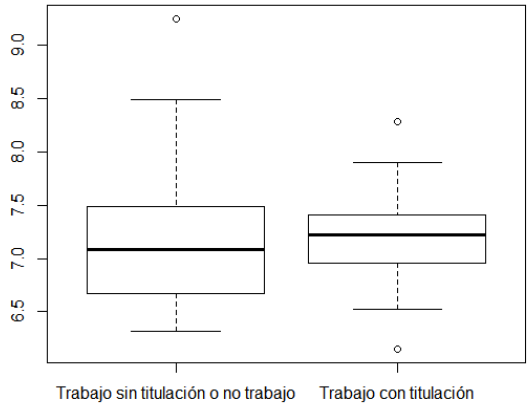
(d) Salud

Figura A.30: Edad, en función de encontrar un empleo que requiere el título obtenido o no encontrarlo. Por un lado, se muestran los alumnos que han conseguido un trabajo que requiere la titulación cursada *(columna derecha)* y, por otro, los alumnos que no han conseguido un empleo o han conseguido un empleo que no requiere el título obtenido *(columna izquierda)*.

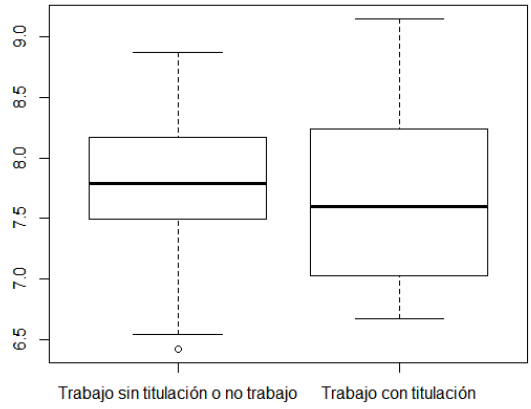
Complementando estos gráficos mediante contrastes de hipótesis de la *t de Student*, se puede afirmar que **la edad únicamente influye a la hora de encontrar un empleo que requiere el título obtenido en la Escuela Superior de Tecnología y Ciencias Experimentales. Los alumnos que encuentran un trabajo que requiere la titulación cursada son mayores que los que no lo encuentran.**

Nota del expediente:

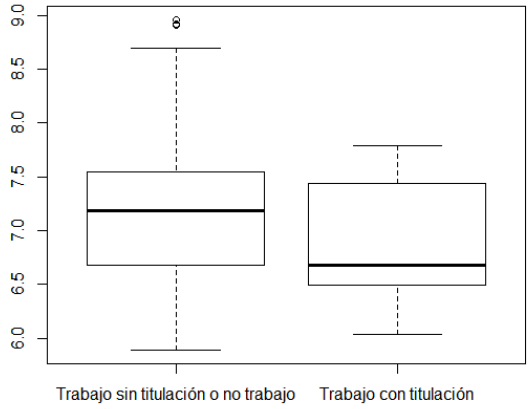
```
boxplot(nota.expediente[trabajo.titulacion == 0], nota.expediente[trabajo.
    titulacion == 1],
names=c("Trabajo sin titulacion o no trabajo","Trabajo con titulacion"))
```



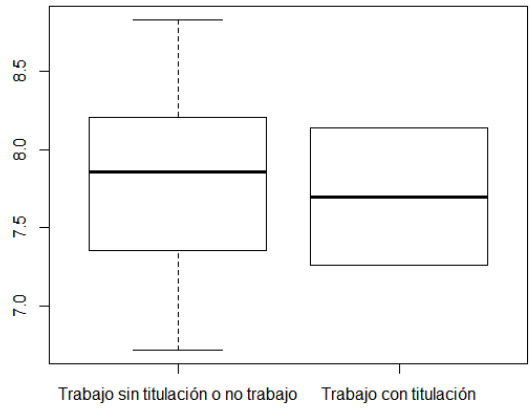
(a) Tecnología y Ciencias



(b) Humanas



(c) Jurídicas



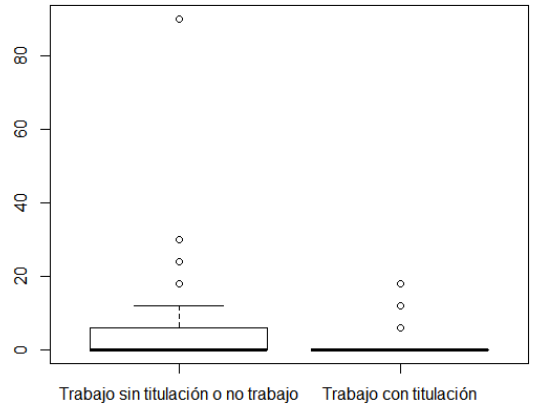
(d) Salud

Figura A.31: Nota obtenida en el expediente, en función de encontrar un empleo que requiere el título obtenido o no encontrarlo. Por un lado, se muestran los alumnos que han conseguido un trabajo que requiere la titulación cursada *(columna derecha)* y, por otro, los alumnos que no han conseguido un empleo o han conseguido un empleo que no requiere el título obtenido *(columna izquierda)*.

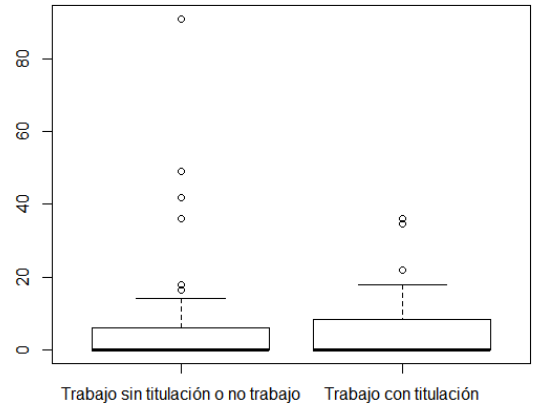
Realizando contrastes de hipótesis de la *t de Student*, se puede afirmar que **la nota de acceso no influye en la inserción laboral en ninguno de los centros**.

Créditos de honor:

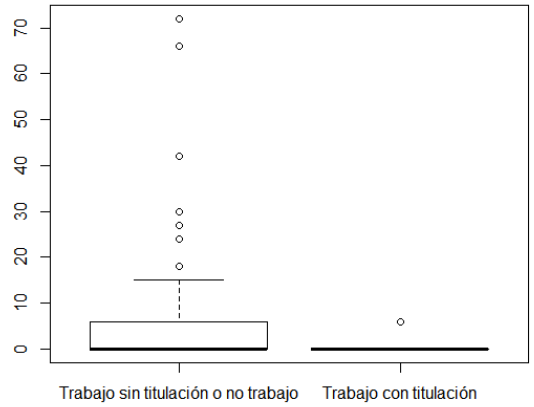
```
boxplot(cred.honor[trabajo.titulacion == 0], cred.honor[trabajo.titulacion
    == 1], names=c("Trabajo sin titulacion o no trabajo","Trabajo con
    titulacion"))
```



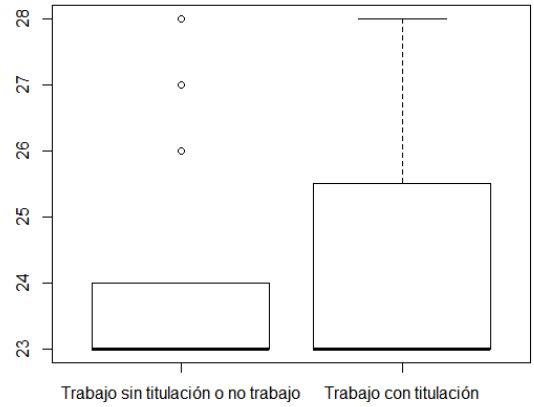
(a) Tecnología y Ciencias Experimentales



(b) Humanas y Sociales



(c) Jurídicas y Económicas



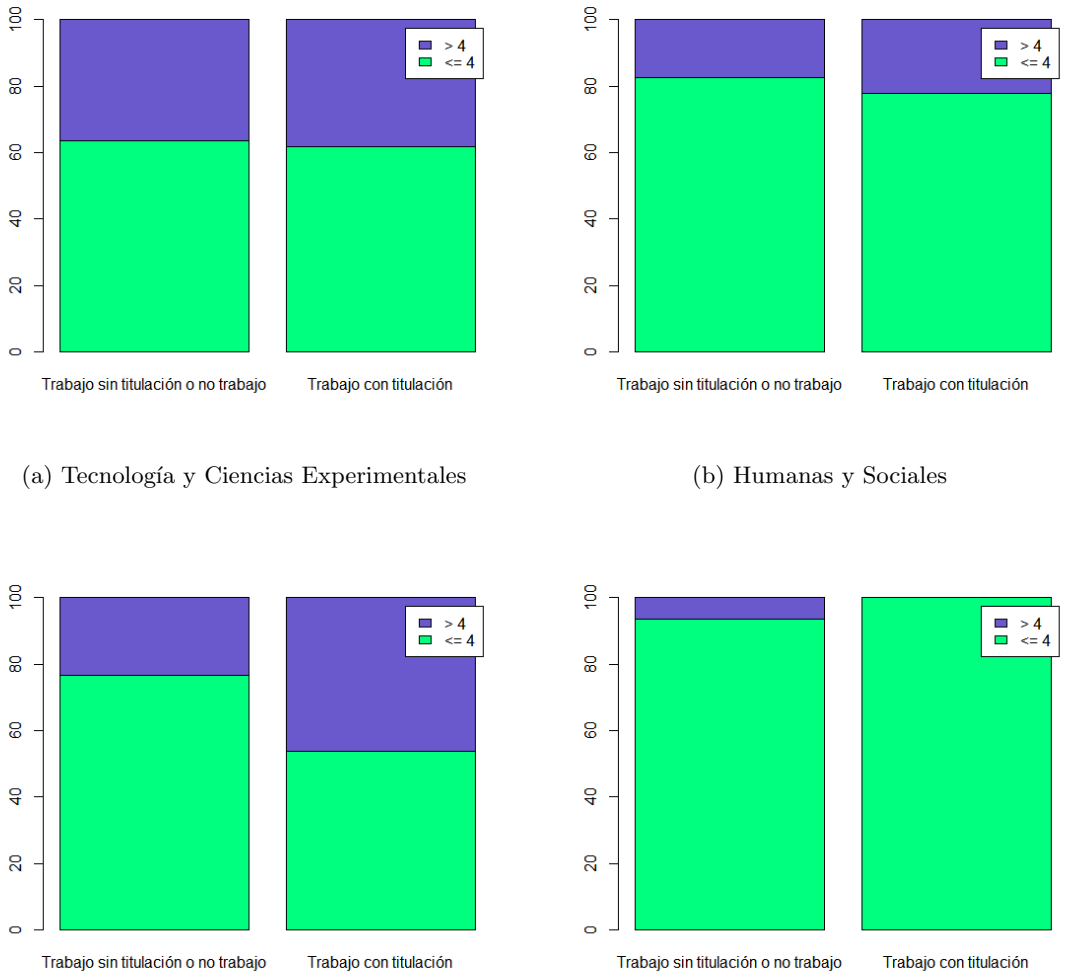
(d) Salud

Figura A.32: Créditos de honor, en función de encontrar un empleo que requiere el título obtenido o no encontrarlo. Por un lado, se muestran los alumnos que han conseguido un trabajo que requiere la titulación cursada *(columna derecha)* y, por otro, los alumnos que no han conseguido un empleo o han conseguido un empleo que no requiere el título obtenido *(columna izquierda)*.

Si se realiza un contraste de la *t de Student*, se puede afirmar que **obtener créditos de honor durante la titulación no influye en la inserción laboral**.

Años en obtener el título:

```
T = table(cursos, trabajo.titulacion)[2:3,]
Ta=colPercents(T)[1:2,]
barplot(Ta, col=c("springgreen","slateblue3"), names.arg=c("Trabajo sin
    titulacion o no trabajo","Trabajo con titulacion"), legend=c("<=4",">4
    "))
```



(a) Tecnología y Ciencias Experimentales

(b) Humanas y Sociales

(c) Jurídicas y Económicas

(d) Salud

Figura A.33: Años en obtener el título, en función de encontrar un empleo que requiere el título obtenido o no encontrarlo. Por un lado, se muestran los alumnos que han conseguido un trabajo que requiere la titulación cursada *(columna derecha)* y, por otro, los alumnos que no han conseguido un empleo o han conseguido un empleo que no requiere el título obtenido *(columna izquierda)*. El color **verde** se utiliza para representar a los alumnos que han tardado cuatro cursos o menos en obtener el título y el color **morado** para representar a los alumnos que han tardado más de cuatro cursos en obtener el título.

Si se realiza un contraste de hipótesis de la *Chi-cuadrado*, se puede afirmar que esta variable **no influye para encontrar un empleo que requiere el título obtenido en ninguno de los centros.**

Anexo B

Listas de programas en R

Este Anexo contiene el código utilizado para la realización los diferentes análisis empleados. En el caso del abandono de los estudios, se muestra únicamente el código empleado para la realización de los análisis sobre toda la muestra. Los análisis realizados por centros se han realizado de modo similar.

Para tratar el tema de la inserción laboral, se muestra el código empleado para analizar la inserción en la Escuela Superior de Tecnología y Ciencias Experimentales, debido a que en este caso no se ha realizado un análisis global sobre toda la muestra. El resto de análisis realizados por centros han sido llevados a cabo de modo similar.

B.1. Abandono de los estudios

B.1.1. Análisis clúster

El análisis clúster se realiza únicamente sobre los alumnos que han abandonado los estudios, por tanto, se debe seleccionar únicamente a esos alumnos de la muestra.

Seleccionar las variables utilizadas en el análisis clúster:

```
x = data.frame(curso.abandono[abandono == 1], edad[abandono == 1], nota.de.
    acceso[abandono == 1], ordered(orden.pref, levels = c("Primera", "
    Segunda", "Resto")), via.acc[abandono == 1], sexo[abandono == 1], prov[
    abandono == 1], cred.pres.pri[abandono == 1], cred.honor.pri[abandono
    == 1], cred.pres.ultimo[abandono == 1], num.asi.rep.ultimo[abandono ==
    1], prov[abandono == 1], cred.sup.exam.media[abandono == 1], as.factor(
    trabPri)[abandono == 1], as.factor(trabUltimo)[abandono == 1])
```

Renombrar las columnas:

```
names(x) = c("curso.abandono","edad", "nota.de.acceso", "orden.pref","via.
    acc", "sexo", "prov", "cred.pres.pri","cred.honor.pri","cred.pres.
    ultimo", "num.asi.rep.ultimo", "prov", "cred.sup.exam.media", "trabPri"
    ,"trabUltimo")
```

Realizar un gráfico para buscar el 'codo' de la función objetivo (3.1) que permite elegir el número de grupos:

```
W<-c()
for(i in 2:8){
        res1 <- pam( x, i, metric = "gower" )
        W[i] <- res1$objective[1]
}
plot(W)
```

Lanzar el método *k-medoides* y escoger la mejor solución obtenida:

```
mejorRes = pam( x, 3, metric = "gower" )
mejorWC <- mejorRes$objective[1]
for(i in 2:10){
        res <- pam( x, 3, metric = "gower" )
        if( res$objective[1] < mejorWC ){
                mejorWC = res$objective[1];
                mejorRes = res
        }
}
res = mejorRes
```

Obtener el número de alumnos clasificados en cada grupo:

```
table(res$clustering)
```

Obtener los representantes de cada grupo:

```
res$medoids
```

Medida de bondad de ajuste del clustering:

```
res$silinfo
```

Descripción de los grupos realizando gráficos para cada una de las variables.

```
grupos = res$clustering
```

Para la variable *edad*:

```
boxplot(x$edad[grupos==1], x$edad[grupos==2], x$edad[grupos==3], main="edad
    ", names = c(1,2,3))
```

Para la variable *nota.de.acceso*:

```
boxplot(x$nota.de.acceso[grupos==1], x$nota.de.acceso[grupos==2], x$nota.de
    .acceso[grupos==3], main="nota acceso", names = c(1,2,3))
```

Para la variable *orden de preferencia*:

```
barplot(colPercents(table(x$orden.pref,grupos)[2:4,])[1:3,],
col=c("cyan","blue","purple"),
legend.text=c("Primera","Segunda", "Resto"), main="orden preferencia")
```

Para la variable *via.de.acceso*:

```
barplot(colPercents(table(x$via.acc,grupos))[1:4,], col=c("green","red","
    yellow", "blue"), legend.text=c("FP","Resto", "Select", "TU"), main="
    via acceso")
```

Para la variable *sexo*:

```
barplot(colPercents(table(x$sexo,grupos))[1:2,], col=c("pink","blue"),
    legend.text=c("Mujer", "Hombre"), main="Sexo")
```

Para la variable *prov*:

```
barplot(colPercents(table(x$prov,grupos))[1:3,],col=c("green","orange",
"yellow"),legend.text=c("Castellon","Limitrofes","Resto"),main="Provincia")
```

Para la variable *cred.pres.pri*:

```
boxplot(x$cred.pres.pri[grupos==1], x$cred.pres.pri[grupos==2],
x$cred.pres.pri[grupos==3], main="creditos presentados primer
curso", names=c(1,2,3))
```

Para la variable *cred.honor.pri*:

```
boxplot(x$cred.honor.pri[grupos==1], x$cred.honor.pri[grupos==2],
x$cred.honor.pri[grupos==3], main="creditos honor primero", names=c(1,2,3))
```

Para la variable *cred.pres.ultimo*:

```
boxplot(x$cred.pres.ultimo[grupos==1], x$cred.pres.ultimo[grupos==2],
x$cred.pres.ultimo[grupos==3], main="creditos presentados ultimo curso
matricula",names=c(1,2,3))
```

Para la variable *cred.sup.exam.media*:

```
boxplot(x$cred.sup.exam.media[grupos==1], x$cred.sup.exam.media[grupos==2]
, x$cred.sup.exam.media[grupos==3], main="promedio de creditos superados
",names=c(1,2,3))
```

Para la variable *num.asi.rep.ultimo*:

```
boxplot(x$num.asi.rep.ultimo[grupos==1], x$num.asi.rep.ultimo[grupos==2],
$num.asi.rep.ultimo[grupos==3], main="numero asignaturas repetidas en
el ultimo curso que el estudiante hizo matricula", names = c(1,2,3))
```

Para la variable *trabPri*:

```
barplot(colPercents(table(x$trabPri,grupos))[1:2,], col=c("cyan","pink"),
    legend.text=c("NO","SI"), main="Trabajo primero")
```

Para la variable *trabUltimo*:

```
barplot(colPercents(table(x$trabUltimo,grupos))[1:2,],col=c("cyan","pink"),
     legend.text=c("NO","SI"), main="Trabajo ultimo")
```

Para la variable *curso.abandono*:

```
barplot(colPercents(table(x$curso.abandono,grupos))[1:3,],col=c("red","blue
    ","green"), legend.text=c("1","2","3"), main="Curso abandono")
```

Determinar mediante contrastes de hipótesis qué variables influyen en la clasificación.

Para variables numéricas se realiza un contraste ANOVA:

```
anova(lm(variable ~ grupos))
```

Para variables categóricas se aplica el contraste *Chi-cuadrado*:

```
chisq.test(table(variable ~ grupos))
```

B.1.2. Regresión logística

Realizar el análisis de regresión logística:

```
logistica=glm(abandono~edad+nota.de.acceso+as.factor(orden.prefSegunda
)+as.factor(orden.prefResto)+via.acc+sexo+prov+cred.pres.pri+cred.hono
r.pri+cred.pres.ultimo+cred.honor.ultimo+num.asi.rep.ultimo+cred.sup.e
xam.media+as.factor(trabPri)+as.factor(trabUltimo),family=binomial,
data=alumnos)

summary(logistica)
```

Obtener la bondad de ajuste (en cada caso se deberá dividir por unos grados de libertad u otros dependiendo del modelo creado):

```
Rcuadrado = 1- ((logistica$deviance/5020)/(logistica$null.deviance/5039))
Rcuadrado
```

Realizar predicciones utilizando la misma muestra con la que el modelo ha sido realizado:

```
muestra=data.frame(edad, nota.de.acceso, orden.prefSegunda, orden.prefResto
    , via.accResto, via.accSelectividad, via.accTitulados, sexoHome,
    provLimitrofes, provResta, cred.pres.pri, cred.honor.pri, cred.pres.
    ultimo, cred.honor.ultimo, num.asi.rep.ultimo, cred.sup.exam.media,
    trabPri, trabUltimo)

pred1=predict(logistica, muestra, type="response")

grupo1=c()
for(i in 1:dim(muestra)[1]){
        if(complete.cases(pred1[i])==FALSE){
                grupo1=c(grupo1,NA)
        }else{
                if(pred1[i]>0.5){
                        grupo1 = c(grupo1,1)
                }else{
                        grupo1 = c(grupo1,0)
                }
        }
}
```

Comprobar cuántos alumnos han sido correctamente clasificados:

```
t=table(abandono,grupo1)
t

bondad=(t[1,1]+t[2,2])/sum(sum(t))
bondad
```

Realizar validación cruzada (como los primeros 1786 alumnos han abandonado los estudios y los restantes no, se toman aleatoriamente 500 alumnos que han abandonado los estudios y 1000 que no):

```
entrenar1 = sample(x=1:1786, size=500)
entrenar2 = sample(x=1787:5893, size=1000)

mitad_entrenar = muestra[c(entrenar1, entrenar2), ]

logistica=glm(abandono~edad+nota.de.acceso+as.factor(orden.prefSegunda)+as.
    factor(orden.prefResto)+as.factor(via.accResto)+as.factor(via.acc
Selectividad)+as.factor(via.accTitulados)+as.factor(sexoHome)+as.factor(
    provLimitrofes)+as.factor(provResta)+cred.pres.pri+cred.honor.pri+cred.
    pres.ultimo+cred.honor.ultimo+num.asi.rep.ultimo+cred.sup.exam.medi
a+as.factor(trabPri)+as.factor(trabUltimo),family=binomial,
data=mitad_entrenar)

mitad_predecir = muestra[-c(entrenar1, entrenar2), 1:19]
mitad_predecir_abandono = muestra[-c(entrenar1, entrenar2), 1]

pred2=predict(logistica, mitad_predecir, type="response")

grupo2=c()
for(i in 1:dim(mitad_predecir)[1]){
        if(complete.cases(pred2[i])==FALSE){
                grupo2=c(grupo2,NA)
```

```
        }else{
                if(pred2[i]>0.5){
                        grupo2 = c(grupo2,1)
                }else{
                        grupo2=c(grupo2,0)
                }
        }
}

t=table(mitad_predecir_abandono, grupo2)
t

bondad=(t[1,1]+t[2,2])/sum(sum(t))
bondad
```

B.1.3. Análisis discriminante

Realizar análisis discriminante:

```
discrimina=lda(abandono~edad+nota.de.acceso+as.factor(orden.prefSegunda)+as
    .factor(orden.prefResto)+via.acc+sexo+prov+cred.pres.pri+cred.hon
or.pri+cred.pres.ultimo+cred.honor.ultimo+num.asi.rep.ultimo+cred.sup.exam.
    media+as.factor(trabPri)+as.factor(trabUltimo), data=alumnos)
```

Observar las variables que influyen más en el abandono:

```
discrimina$scaling
```

Obtener las probabilidades de abandonar y de no abandonar los estudios:

```
discrimina$prior
```

Mostrar las medias de cada variable para el grupo abandono y el grupo no abandono:

```
discrimina$means
```

Obtener una medida de bondad del modelo:

```
discrimina2=lda(abandono~edad+nota.de.acceso+as.factor(orden.prefSegunda)+
    as.factor(orden.prefResto)+via.acc+sexo+prov+cred.pres.pri+
cred.honor.pri+cred.pres.ultimo+cred.honor.ultimo+num.asi.rep.ultimo+cred.
    sup.exam.media+as.factor(trabPri)+as.factor(trabUltimo),
data=alumnos, CV=T)

t = table(abandono[complete.cases(nota.de.acceso)==T & complete.cases(
    trabPri)==T & complete.cases(trabUltimo)==T &
complete.cases(orden.preferencia.completo) == T], discrimina2$class)
t

bondad = (t[1,1]+t[2,2])/sum(sum(t))
bondad
```

B.1.4. Redes neuronales

Separar las variables numéricas y las no numéricas para poder tipificar las variables numéricas:

```
muestra_numerica=data.frame(edad,nota.acceso,cred.pres.pri,cred.honor.pri,
    cred.pres.ultimo, cred.honor.ultimo, num.asi.rep.ultimo, cred.sup.exam.
    media)

muestra_numerica_tipificada=scale(muestra_numerica)

muestra_no_numerica=data.frame(sexo,via.acc,prov,as.factor(orden.
    prefSegunda), as.factor(orden.prefResto), as.factor(trabPri),as.factor(
    trabUltimo))

muestra = data.frame(muestra_numerica_tipificada, muestra_no_numerica)
```

Crear red neuronal con 10 neuronas en la capa oculta:

```
red = nnet(formula=abandono~.,data=muestra, size=10, entropy=T)
```

Observar cuántos elementos de la muestra se clasifican correctamente:

```
t = table(real=abandono[complete.cases(nota.acceso)==T & complete.cases(
    trabPri)==T & complete.cases(trabUltimo)==T & complete.cases(orden.
    preferencia.completo) == T],estimado=predict(red, type='class'))
t

bondad = sum(t[1,1],t[2,2])/sum(sum(t))
bondad
```

Realizar validación cruzada:

```
muestra_numerica=data.frame(edad,nota.acceso,cred.pres.pri,cred.honor.pri,
    cred.pres.ultimo, cred.honor.ultimo,num.asi.rep.ultimo, cred.sup.exam.
    media)

muestra_no_numerica=data.frame(abandono,sexo,via.acc,prov,as.factor(orden.
    prefSegunda), as.factor(orden.prefResto), as.factor(trabPri),as.factor(
    trabUltimo))

entrenar1 = sample(x=1:1786, size=500)
entrenar2 = sample(x=1787:5893, size=1000)

mitad_entrenar_numerica = muestra_numerica[c(entrenar1, entrenar2), ]
mitad_entrenar_numerica_tipificada = scale(mitad_entrenar_numerica)
mitad_entrenar_no_numerica = muestra_no_numerica[c(entrenar1,entrenar2), ]
muestra_entrenar = data.frame(mitad_entrenar_numerica_tipificada,mitad_
    entrenar_no_numerica)

mitad_predecir_numerica = muestra_numerica[-c(entrenar1, entrenar2), ]
mitad_predecir_numerica_tipificada = scale(mitad_predecir_numerica)
mitad_predecir_no_numerica = muestra_no_numerica[-c(entrenar1,entrenar2),
    2:8]
```

```
mitad_predecir_abandono = muestra_no_numerica[-c(entrenar1,entrenar2), 1]
muestra_predecir = data.frame(mitad_predecir_numerica_tipificada, mitad_
    predecir_no_numerica)

red = nnet(formula=abandono~., data=muestra_entrenar, size=10, entropy=T)

predicciones = predict(red, newdata=muestra_predecir, type='class')

t=table(mitad_predecir_abandono, predicciones)
t

bondad=sum(t[1,1],t[2,2])/sum(sum(t))
bondad
```

Script para obtener la red neuronal que mejor predice el abandono de los alumnos de la parte de la muestra utilizada para predecir:

```
neuronas = 5
mejorBondad = NULL
while( neuronas < 41 ){
        for (i in 1:50){
                red = nnet(formula=abandono ~., data=muestra_entrenar, size
                    =neuronas, entropy=T)
                predicciones = predict(red, newdata=muestra_predecir, type=
                    'class')
                t = table( mitad_predecir_abandono, predicciones )
                bondad=sum(t[1,1],t[2,2])/sum(sum(t))

                if( is.null(mejorBondad) || bondad > mejorBondad ) {
                        mejorBondad = bondad
                        mejorRed = red
                        mejorNeuronas = neuronas
                }
        }
        neuronas = neuronas + 1
}
mejorBondad
mejorNeuronas
```

B.1.5. Árboles de clasificación

Uso de árboles para clasificación:

```
x=data.frame(edad,nota.de.acceso,cred.pres.pri, cred.honor.pri,cred.pres.
    ultimo, cred.honor.ultimo, num.asi.rep.ultimo, cred.sup.exam.media,
    sexo,via.acc,prov,as.factor(orden.prefSegunda), as.factor(orden.
    prefResto), as.factor(trabPri), as.factor(trabUltimo))

arb=rpart(formula=alumnos$abandono~., data=x, method = "class")
plot(arb)
text(arb, use.n=T)
```

B.2. Inserción laboral

B.2.1. Regresión logística

```
logistica=glm(as.factor(trabajo.titulacion)~edad+as.factor(sexo)+as.factor(
    cursos.cuatro.o.menos)+cred.honor+nota.expediente+as.factor(erasmus.
    estudios)+as.factor(erasmus.pract
icas)+as.factor(practicas.extracurriculares),family = binomial, tecnologia)

summary(logistica)
```

B.2.2. Reglas de asociación

Crear una nueva hoja de datos con las variables necesarias y renombrar las columnas:

```
nuevo=data.frame(as.factor(trabajo.titulacion),as.factor(edad),as.factor(
    sexo),as.factor(cursos.cuatro.o.menos),as.factor(cred.honor),nota.
    expediente.cut,as.factor(erasmus.estudios),as.factor(erasmus.practicas)
    ,as.factor(practicas.extracurriculares))

colnames(nuevo) = c("trabajo.titulacion", "edad", "sexo", "cursos.cuatro.o.
    menos", "cred.honor", "nota.expediente", "erasmus.estudios", "erasmus.
    practicas", "practicas.extracurriculares")
```

Convertir la hoja de datos a matriz de transacciones:

```
trans = as(nuevo, "transactions")
```

Crear reglas y buscar las que impliquen encontrar un trabajo que requiere el título obtenido para su desempeño:

```
r1=apriori(trans)

inspect(subset(sort(r1, by="confidence"), subset=(rhs %in% "trabajo.
    titulacion=1")))
```