UNIVERSIDAD MIGUEL HERNÁNDEZ

FACULTAD DE CIENCIAS SOCIALES Y JURÍDICAS DE ELCHE

GRADO EN ESTADÍSTICA EMPRESARIAL

TRABAJO DE FIN DE GRADO

CURSO ACADÉMICO 2020-2021

METODOLOGÍA DE ALERTA TEMPRANA PARA EL ABANDONO Y EL RENDIMIENTO ACADÉMICO EN GRADOS Y MÁSTERS UNIVERSITARIOS

Alumno/a: *Esther Sobrino Poveda*

Tutores: *Alejandro Rabasa Dolado*
y Miriam Esteve Campello

ÍNDICE

1. RESUMEN

Son muchos los motivos que conducen a que los estudiantes universitarios acaben abandonando sus carreras. La casuística que conduce al abandono es diferente en función de las ramas académicas, las facultades o escuelas y las circunstancias personales de los universitarios. También es diferente el momento en que esto ocurre. En un contexto de continuos cambios y de exceso de información, las universidades se plantean la posibilidad de extraer de sus bases de datos patrones de alerta temprana que sirvan para actuar a tiempo en aquellos casos donde la probabilidad de abandono empiece a ser relativamente alta. Este estudio propone una metodología en dos etapas para extraer tales patrones de abandono. En primer lugar, y dependiendo de cada facultad o escuela universitaria, el modelo extraerá los factores más influyentes sobre el abandono y posteriormente se generarán árboles de clasificación utilizando tales variables explicativas. La metodología propuesta ha sido testada con datos reales de la Universidad Miguel Hernández de Elche, con datos de Grados y Másteres del año 2010 hasta el 2020 (más de 24.894 registros), proporcionados directamente por el Vicerrectorado de Estudiantes y Coordinación, desde donde se lideró este proyecto. Con la experiencia computacional que acompaña a la metodología descrita, se obtienen modelos predictivos capaces de modelar el abandono con unas precisiones medias próximas al 80%. Ello no solo ofrece una fotografía real de la situación de cada facultad, sino que se sustenta en una confianza más que suficiente para que el Vicerrectorado pueda diseñar planes de contingencia para los escenarios donde el abandono se presenta como una alternativa muy probable, en caso de que no se intervenga de manera temprana.

Palabras clave: Machine Learning, Clasificación, Selección de Atributos, Alerta Temprana, Abandono de Estudiantes.

ABSTRACT

There are many reasons that lead college students to end up abandoning their careers. The casuistry that leads to dropout is different depending on the academic branches, the faculties or schools and the personal circumstances of the university students. The timing of this is also different. In a context of continuous changes and an excess of information, Universities consider the possibility of extracting early warning patterns from their databases that help to act in time in those cases where the probability of abandoning begins to be relatively high. This study proposes a two-stage methodology to extract such dropout patterns. In the first place, and depending on each faculty or university school, the model will extract the most influential factors on dropout and later, classification trees will be generated using these explanatory variables. The proposed methodology has been tested with real data from the University Miguel Hernández of Elche, with data from Degrees from the year 2010 to the present (24,894 records), provided directly by the Vice-Rector's Office for Students affairs and coordination, that leads this project. With the computational experience that accompanies the described methodology,

predictive models are capable of modelling dropout with mean accuracies close to 80%. This not only offers a real photograph of the situation of each faculty, but is supported by more than enough confidence so that the Vice-Rector's Office can design contingency plans for scenarios where dropout is presented as a very likely alternative, in case of do not intervene early.

Keywords: Machine Learning, Classification, Feature Selection, Early Warning, Students Dropping.

2. INTRODUCCIÓN Y OBJETIVOS

2.1 INTRODUCCIÓN AL TFG

En la actualidad, algunas instituciones educativas se enfrentan al reto de mejorar la calidad en la enseñanza. Uno de los factores que más influyen en esta calidad es el abandono escolar. El análisis de este abandono ha constituido desde hace décadas un tema de gran interés en todos los niveles educativos y en la actualidad son numerosos los trabajos de investigación a nivel internacional que centran sus estudios en este ámbito.

La investigación llevada a cabo en los últimos años para averiguar las causas del abandono escolar temprano nos lleva a definirlo como un proceso complejo y multidimensional en el que interfieren circunstancias de la escuela y del entorno escolar con implicaciones sociales y culturales [3]. La investigación confirma además que los alumnos que abandonan la escuela o la universidad lo hacen siendo conscientes que al no aumentar su formación, disminuirán las oportunidades de conseguir un mejor empleo y un mayor salario en un futuro.

Las técnicas de Minería de Datos, aplicadas a la información obtenida a través de las instituciones educativas, permiten establecer unos modelos predictivos que constituyen una herramienta de gran eficacia para predecir el abandono escolar de los estudiantes y para poder identificar y evaluar los factores que más influyen en este proceso proporcionando, de esta forma, una información sólida y fiable que permita desarrollar acciones encaminadas a prevenir el abandono escolar.

En la actualidad se ha generado una nueva comunidad de investigación en educación denominada Minería de Datos Educativa, que aplica las técnicas de Minería de Datos para analizar y evaluar los datos obtenidos en los entornos educativos y transformarlos en información útil [9], con la finalidad de comprender mejor los factores que propician el abandono escolar y poder servir de apoyo a la toma de decisiones por parte de las instituciones con el fin de mejorar la calidad en el proceso de enseñanza y aprendizaje de los alumnos para minimizar todo lo posible este factor [4].

2.2 OBJETIVOS DEL TFG

Este trabajo tiene dos objetivos principales. El primero es aplicar una metodología de Machine Learning, adaptada para tareas de clasificación en el marco de los estudios universitarios. El segundo es obtener modelos precisos capaces de predecir el abandono de estudios con las características más relevantes de los estudiantes y los Grados y/o Másteres.

2.3 OBJETIVOS PERSONALES

Los objetivos personales que se persiguen con este trabajo son:

- Aplicar los conocimientos adquiridos durante la formación en el Grado de Estadística Empresarial a un caso práctico real, relacionado con la educación.

- Aplicar lo aprendido durante las actividades de investigación llevadas a cabo en el marco de Prácticas Internas en el Centro de Investigación Operativa de la Universidad Miguel Hernández en el marco de un proyecto del Vicerrectorado de Estudiantes y Coordinación.

3. ESTADO DE LA CUESTIÓN Y MARCO TEÓRICO

3.1 TAREAS DESCRIPTIVAS Y DE PREPROCESAMIENTO TÍPICAS

Para realizar el preprocesamiento primero se ha realizado un estudio para la detección de valores anómalos en las variables numéricas (outliers).

Uno de los gráficos utilizados para realizar el preprocesamiento de los datos en las variables numéricas es el gráfico de cajas porque permite representar los cuantiles y los valores anómalos (outliers) en el caso de que los haya.

También se ha realizado un preprocesamiento enfocado en los valores nulos, que en este conjunto de datos solo existen en una variable y no se han eliminado debido a que eran de ayuda para imputar otra variable a partir de esta.

3.2 MÉTODOS DE SELECCIÓN DE ATRIBUTOS

En los casos en el que se dispone de un número aceptable de variables se suelen tener en cuenta todas ellas. Pero en los casos en los que se dispone de un gran número de variables, es necesario recurrir a mecanismos de selección de atributos, que ordenan a los mismos por importancia respecto a la variable objetivo para así poder elegir cuáles se deben utilizar para la construcción de los modelos de clasificación [8].

La selección de atributos afecta de forma muy importante al rendimiento y a la precisión del modelo. Una buena selección de atributos puede ayudar a interpretar el modelo más fácilmente, reduce el sobreajuste y reduce el tiempo de entrenamiento.

En este caso se ha utilizado el algoritmo GainRatio [15], que consiste en una métrica resultante de la Teoría de la Información, concretamente la que tiene que ver con la proporción de ganancia, cuyo valor sigue una distribución normal entre 0 y 1 (cuanto más alto el valor, mejor). GainRatio es la versión normalizada del algoritmo InfoGain. Esto implica que se reduce la parcialidad del algoritmo dividiendo la entropía del atributo dado. Ambos métodos son de tipo ranker, debido a que devuelven los atributos dados en forma descendente con respecto a la ganancia de la información. Una opción para asignar un punto de corte sería determinar el número de atributos que se quiere obtener.

Como ventaja, estos métodos analizan la interacción de los atributos con la variable objetivo, sin embargo, no tienen en cuenta la interacción de pares de atributos con la variable objetivo y no excluyen a los atributos redundantes porque asume que las variables son independientes entre sí.

3.3 MÉTODOS DE CLASIFICACIÓN

En el ámbito de los métodos de aprendizaje, los basados en árboles de decisión son los más amigables, tanto en utilidad como en comprensión.

La definición general de árbol de decisión es un conjunto de condiciones organizadas jerárquicamente, de forma que la decisión final se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta sus hojas. Se llevan utilizando desde hace años y son muy útiles en ámbitos educativos, médicos, legales, etc. debido a su fácil interpretación.

Los árboles de decisión son modelos jerárquicos para el aprendizaje supervisado aplicable tanto a la regresión como a la clasificación. Además, son modelos no paramétricos que se usan para crear modelos predictivos automatizados que se emplean para el aprendizaje automático, la minería de datos y las estadísticas. Este aprendizaje que está basado en árboles de decisión considera las observaciones sobre un elemento para predecir su valor.

Estos árboles están constituidos por un nodo raíz que se encuentra en la parte superior, un conjunto de nodos (cada uno de ellos asociados a una variable predictora) cuyas ramas representan decisiones de los valores de una variable y un conjunto de nodos terminales, con algún valor de la clase de la variable dependiente u objetivo.

El modelo para un árbol de clasificación se presenta como una jerarquía que establece las relaciones entre las variables predictoras y la variable dependiente. Cada ramificación contiene una serie de reglas de clasificación asociadas a una etiqueta de clase específica que se encuentra al final de la ramificación.

El árbol de decisión óptimo es el que representa la mayor cantidad de datos con el menor número de niveles posible y con una matriz de confusión aceptable. Algunos algoritmos diseñados para crear árboles de decisión incluyen CART [1], ID3 y C4.5 [7]. Cada algoritmo establece cuál es la forma óptima de dividir los datos en cada nivel.

Como en todos los métodos, la aplicación de árboles de clasificación tiene sus ventajas y sus desventajas. Una de las muchas ventajas es que las opciones posibles que se forman a partir de una condición son excluyentes y, gracias a esto, si se interpreta correctamente el árbol se puede llegar a una sola decisión a tomar. Otra ventaja es que la fiabilidad de un árbol de clasificación puede cuantificarse y, por tanto, se puede mejorar hasta obtener una alta fiabilidad. Una desventaja de este tipo de modelos es que la información obtenida tiende a favor de los atributos con más cantidad de niveles.

3.4 ÁMBITO EDUCATIVO: SITUACIÓN ACTUAL Y HERRAMIENTAS UTILIZADAS

3.4.1 Situación actual

En los últimos años, las universidades europeas tienen como uno de sus claros objetivos la mejora de la enseñanza invirtiendo en innovación educativa y en el uso de nuevas tecnologías en las aulas [6]. El objetivo es modernizar la universidad dejando atrás los métodos anticuados del siglo XX basados en la clase magistral y el estudio de conceptos mayoritariamente teóricos, alcanzando técnicas basadas en la práctica del conocimiento y el uso de nuevas tecnologías. Por lo tanto, el análisis y la monitorización del abandono estudiantil y los patrones que se siguen en estos casos tienen una crucial importancia dentro del proyecto de transformación educativa en el cual están sumergidas las universidades.

Esta tarea recae en gran parte en el Vicerrectorado de Estudiantes y Coordinación de la Universidad Miguel Hernández, siendo el responsable del uso de la información disponible (base de datos del desempeño de los estudiantes, itinerario, evolución, etc.) para planificar los cambios necesarios y las líneas principales en la nueva enseñanza.

3.4.2 Herramientas utilizadas

Las herramientas de software que se han utilizado en esta investigación (y que a su vez han sido estudiadas durante el Grado de Estadística Empresarial) son las siguientes:

- **RStudio**: es un entorno de desarrollo integrado (IDE) para el lenguaje de programación R, dedicado a la computación estadística y gráficos. Proporciona al usuario un entorno amigable para el análisis estadístico de los datos. Concretamente ha sido utilizado en este estudio para aplicar técnicas de Minería de Datos.

- **Microsoft Excel**: es un programa de software de hojas de cálculo y una herramienta de análisis y visualización de datos. En la investigación se ha utilizado para complementar las actividades realizadas con RStudio.

3.4.3 Librerías utilizadas (en R)

El lenguaje de programación utilizado para realizar el estudio es R, como ya se ha mencionado anteriormente, y a continuación se detallan todos los paquetes utilizados para llevar a cabo este trabajo y sus funcionalidades:

- **forcats** [10]: El objetivo de la librería forcats es proporcionar un conjunto de herramientas que resuelvan problemas comunes con factores, incluido el cambio del orden de los niveles o los valores.

- **sqldf**: Esta librería permite escribir lenguaje de SQL en R. En este trabajo se ha utilizado para poder imputar ciertas variables como por ejemplo la variable ABANDONO a partir de la variable ANY_EGRES como se verá más adelante.

- **MachineLearning (Algorithms for Innovation in Tourism)**: Es una librería enfocada a aplicar las herramientas de Machine Learning al Turismo que contiene las siguientes herramientas: AssociationRules, CART, Clustering, CREA.RBS, plot.MLA, plotCART, print.MLA, sampler y VariableRanker.

- **dplyr** [11]: El paquete dplyr fue desarrollado por Hadley Wickham de RStudio y es un versión optimizada de su paquete plyr. El paquete dplyr no proporciona ninguna nueva funcionalidad a R per se, en el sentido que todo aquello que podemos hacer con dplyr lo podríamos hacer con la sintaxis básica de R. Una importante contribución del paquete dplyr es que proporciona una gramática para la manipulación y operaciones con data frames. Con esta gramática podemos comunicar mediante nuestro código que es lo que estamos haciendo en los data frames a otras personas.

- **rpart** [12]: En el campo del aprendizaje automático, hay distintas maneras de obtener árboles de decisión, la que usaremos en esta ocasión es conocida como CART: Classification And Regression Trees. Esta es una técnica de aprendizaje supervisado. Tenemos una variable objetivo (dependiente) y nuestra meta es obtener una función que nos permita predecir, a partir de variables predictoras (independientes), el valor de la variable objetivo para casos desconocidos. CART es una técnica con la que se pueden obtener árboles de clasificación y de regresión. Usamos clasificación cuando nuestra variable objetivo es discreta, mientras que usamos regresión cuando es continua. La implementación particular de CART que usaremos es conocida como Recursive Partitioning and Regression Trees o RPART. Lo que hace este algoritmo es encontrar la variable independiente que mejor separa nuestros datos en grupos, que corresponden con las categorías de la variable objetivo. Esta mejor separación es expresada con una regla. A cada regla corresponde un nodo.

- **rpart.plot**: Sirve para representar el árbol que genera la librería rpart.

- **caret** [13]: Incluye una serie de funciones que facilitan el uso de decenas de métodos complejos de clasificación y regresión. Permite utilizar un código unificado para aplicar reglas de clasificación muy distintas, implementadas en diferentes paquetes. Es más fácil poner en práctica algunos procedimientos usuales en problemas de clasificación. Por ejemplo, hay funciones específicas para dividir la muestra en datos de entrenamiento y datos de test o para ajustar parámetros mediante validación cruzada.

- **ggplot2** [14]: Esta librería sirve para representar gráficos que se construyen combinando una serie de elementos básicos y comunes a muchos tipos de gráficos distintos mediante una sintaxis sencilla.

4. HIPÓTESIS DE PARTIDA

4.1 ÁMBITO DE CLASIFICACIÓN Y NATURALEZA DE LAS VARIABLES

Para estudiar el abandono y el rendimiento de los estudiantes, han sido necesarias dos variables objetivo y ambas discretas para poder realizar así árboles de clasificación, ya que si fueran continuas los árboles serían de regresión y lo que se busca es clasificar a los individuos para encontrar patrones por facultad. Para ello ha sido necesario crear dos variables (ABANDONO y MEDIAC_dis) a partir de otras variables de la base de datos como ya se ha explicado anteriormente (Apartado 3.1).

4.2 LA FACTORIZACIÓN COMO PROCESO CRÍTICO.

Debido a que la variable objetivo elegida para el estudio es MEDIAC, ya que se quiere medir el rendimiento de los alumnos de la UMH (y es una variable numérica) es necesario realizar una factorización de la misma para facilitar así la precisión a la hora de realizar árboles de clasificación interpretables. Este es un proceso crítico ya que se debe realizar bien la factorización de la variable, para conseguir el mayor porcentaje de precisión posible.

Para ello es necesario crear una variable nueva (MEDIAC_dis), que en este caso al ser la nota media del estudiante del 0 al 10, se ha decidido dividir en 4 intervalos diferentes: '[0-5)' - SUSPENSO, '[5-7)' - APROBADO, '[7-9)' - NOTABLE, '[9-10]' - SOBRESALIENTE. Si se dividiera esta variable en cinco intervalos o más, se comenzaría a perder precisión a la hora de realizar los modelos predictivos.

4.3 MODELOS PREDICTIVOS PRECISOS Y FÁCILES DE INTERPRETAR PARA TOMAR DECISIONES DESDE VICERRECTORADO.

Con este estudio se busca emplear modelos que sean fáciles de interpretar para la toma de decisiones por parte del Vicerrectorado de Estudiantes y Coordinación, sobre todo para actuar de forma temprana en aquellos casos donde la probabilidad de abandono empiece a ser relativamente alta y poder servir de apoyo a la toma de decisiones por parte de las instituciones con el fin de mejorar la calidad en el proceso de enseñanza y aprendizaje de los alumnos para minimizar todo lo posible este factor.

También se busca encontrar patrones en estudiantes cuyo rendimiento es muy alto o muy bajo, para saber con exactitud si los propicia algún factor en especial o un conjunto de ellos y poder actuar así de forma temprana en los casos de rendimiento bajo para aumentarlo y conseguir que estos estudiantes tengan situaciones semejantes a los que tienen un alto rendimiento académico.

4.4 SOFTWARE Y HARDWARE EMPLEADO

El ordenador utilizado para realizar el estudio tiene las siguientes características: Procesador AMD 3020E, memoria RAM 4GB y disco duro 128GB SSD. Gráfica: Radeon Graphics. Windows 10.

La versión utilizada de RStudio para realizar el trabajo ha sido la 1.4.1103 y la versión de R utilizada ha sido la 3.6.1 (lanzada el 5 de julio de 2019). Además las librerías de RStudio que se han usado para llevar a cabo el estudio han sido forcats, sqldf, MachineLearning (Algorithms for Innovation in Tourism), dplyr, rpart, rpart.plot, caret y ggplot2.

5. METODOLOGÍA

5.1 LOS DATOS

El Vicerrectorado de Estudiantes y Coordinación proporciona la base de datos corporativa que recopila información sobre la matriculación actual del estudiante (grado o posgrado, facultad, campus, tipo de matrícula, tipo de estudios…); información académica sobre la evolución del estudiantes (curso, créditos superados y pendientes, nota media…) e información sociodemográfica sobre el estudiante.

Para obtener datos de la mayor calidad posible, es necesario realizar un buen preprocesamiento. Dicho preprocesamiento consiste en detectar valores anómalos (outliers) y nulos. Para ello, se realizó un análisis exploratorio de los datos en el que se concluye que no existen valores anómalos o nulos. A continuación, se muestra una tabla con los atributos (Ver Tabla 1).

Tabla 1: Atributos de la base de datos

Nombre Atributo	**Tipo**	**Descripción y valores**
Ca/Ca	numérico	último año de estudios (del 2010 al 2019)
DNI	nominal	número de identificación encriptado del individuo
CAMPUS	nominal	campus de la universidad (A - 'Altea', E - 'Elche', L - 'Facultad de Ciencias Sociales y Jurídicas de Orihuela', O - 'Escuela Politécnica de Orihuela', S - 'San Juan')
CEN	nominal	facultad (C - 'Facultad de Ciencias Sociosanitarias', K - 'Escuela Politécnica de Orihuela', M - 'Facultad de Medicina', N - 'Facultad de Bellas Artes', O - 'Facultad de Ciencias Experimentales', P - 'Facultad de Farmacia', R - 'Facultad de Ciencias Sociales y Jurídicas de Elche', S - 'Escuela Politécnica de Elche', T - 'Facultad de Ciencias Sociales y Jurídicas de Orihuela')
TIPO	binario	tipo de estudiante (A - 'antiguo estudiante' no está en el primer año de su carrera, N - 'nuevo estudiante' está en el primer año de su carrera)
TIT	nominal	código de la carrera
MODALIDAD	nominal	modalidad de la carrera (O - 'Online', P - 'Presencial', S - 'Semipresencial')

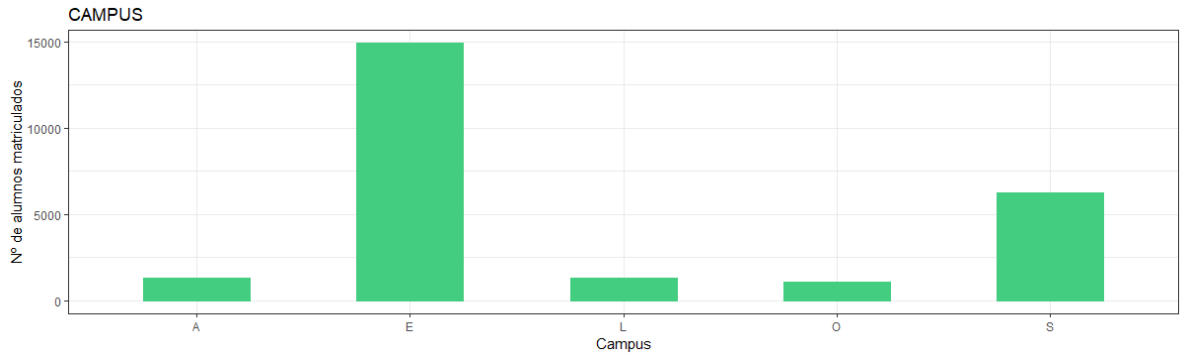
CURSO	nominal	curso en el que se encuentra el estudiante (del 1 al 6)
GRUPO	nominal	grupo al que pertenece el estudiante
ANY_EGRES	numérico	año de graduación (del 2012 al 2019, si aparece como NA es que el estudiante aún no ha acabado)
POB	nominal	ciudad del estudiante
POS	nominal	código postal
PRO	nominal	provincia del estudiante
F_NAC	nominal	fecha de nacimiento
SEGURO	binario	indica si el estudiante solicitó seguro (S - 'Sí', N - 'No')
NACIONALITAT	nominal	nacionalidad del estudiante
SEXO	binario	sexo del estudiante (H - 'masculino', M - 'femenino')
EDAD	numérico	edad del estudiante (de 17 a 81)
SOLIBEC	nominal	indica si el estudiante solicitó beca (S - 'Sí')
CONCEBEC	binario	indica si la beca fue concedida (S - 'Sí', N - 'No')
SOLIBEC_TIPO	nominal	tipo de beca (A - 'ambas', C - 'conselleria', M - 'ministerio', U - 'universidad')
CLA	nominal	tipo de matrícula (0 - 'matrícula ordinaria', A - 'estudiante visitante', B - 'estudiante de intercambio')
FAMNOM	nominal	tipo de familia numerosa (0 - 'Familia No Numerosa', 1- 'Familia Numerosa General', 2 - 'Familia Numerosa Especial', 6 - 'Familia Monoparental General', 7 - 'Familia Monoparental Especial')
PAG	nominal	tipo de pago (1 - 'Pago en efectivo en una vez', 2 - 'Pago en efectivo en dos veces', 3 - 'Cargo en cuenta en cuatro veces', 4 - 'Cargo en cuenta en una vez', 5 - 'Cargo en cuenta en dos veces', 6 - 'Cargo en cuenta en ocho veces')
DISCA	binario	discapacidad (S - 'Sí', N - 'No')

DISCA_PORCEN	numérico	porcentaje de discapacidad (de 8 a 82)
CRE_SUPTOT	numérico	créditos totales superados (de 0 a 240)
CRE_PENTOT	numérico	créditos totales pendientes (de 0 a 210)
MEDIA	numérico	nota media basada en las equivalencias: SUSPENSO 2.5, APROBADO 5.5, NOTABLE 7.5, SOBRESALIENTE 9 y MATRÍCULA DE HONOR 10 (de 0 a 10)
MEDIAB	numérico	nota media basada en las equivalencias: SUSPENSO 0, APROBADO 1, NOTABLE 2, SOBRESALIENTE 3 y MATRÍCULA DE HONOR 4 (de 0 a 4)
MEDIAC	numérico	nota media ponderada (de 0 a 10)
ABANDONO	binario	abandono de los estudios (S - 'Sí', N - 'No')
MEDIAC_dis	nominal	MEDIAC discretizada ('[0-5)' - SUSPENSO, '[5-7)' - APROBADO, '[7-9)' - NOTABLE, '[9-10]' - SOBRESALIENTE)

Considerando este conjunto de datos de entrada, se ha sometido a un estudio descriptivo detallado de las variables más relevantes para el Vicerrectorado y así conseguir una idea más exacta de las variables.

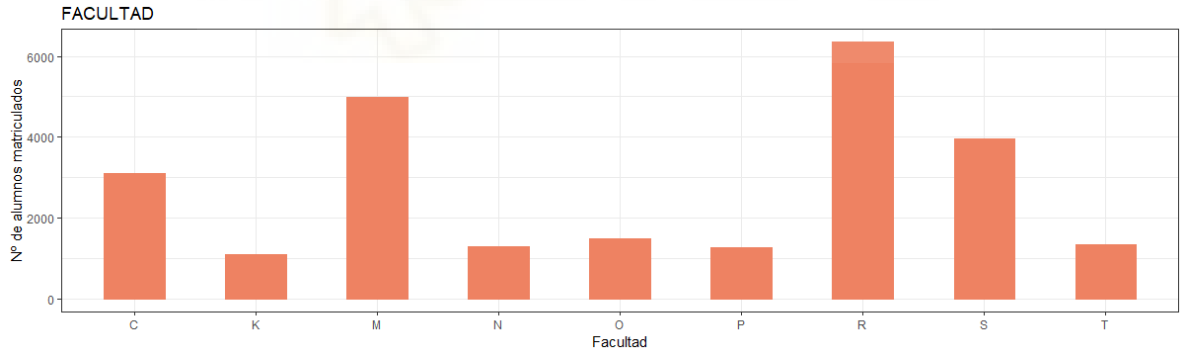
GRADOS

Figura 1: Distribución total de los estudiantes por Campus (A = Altea, E = Elche, L = Facultad de Ciencias Sociales y Jurídicas de Orihuela, O = Escuela Politécnica de Orihuela, S = San Juan).



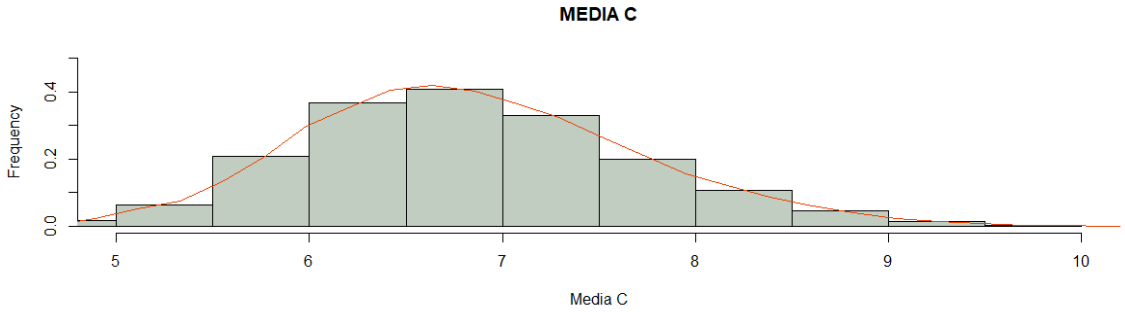
El campus de Elche es el que más alumnos matriculados en grados tiene, seguido del campus de San Juan.

Figura 2: Distribución total de los estudiantes por Facultad (C=F. Ciencias Sociosanitarias, K=Politécnica Orihuela, M=F. Medicina, N=F. Bellas Artes, O=F. Ciencias Experimentales, P=F. Farmacia, R=F.CC.SS. de Elche y Jurídicas Elche, S=Politécnica Elche, T=F.CC.SS y Jurídicas Orihuela)



El mayor número de alumnos matriculados se encuentra en la facultad de Ciencias Sociales y Jurídicas de Elche y en la facultad de Medicina.

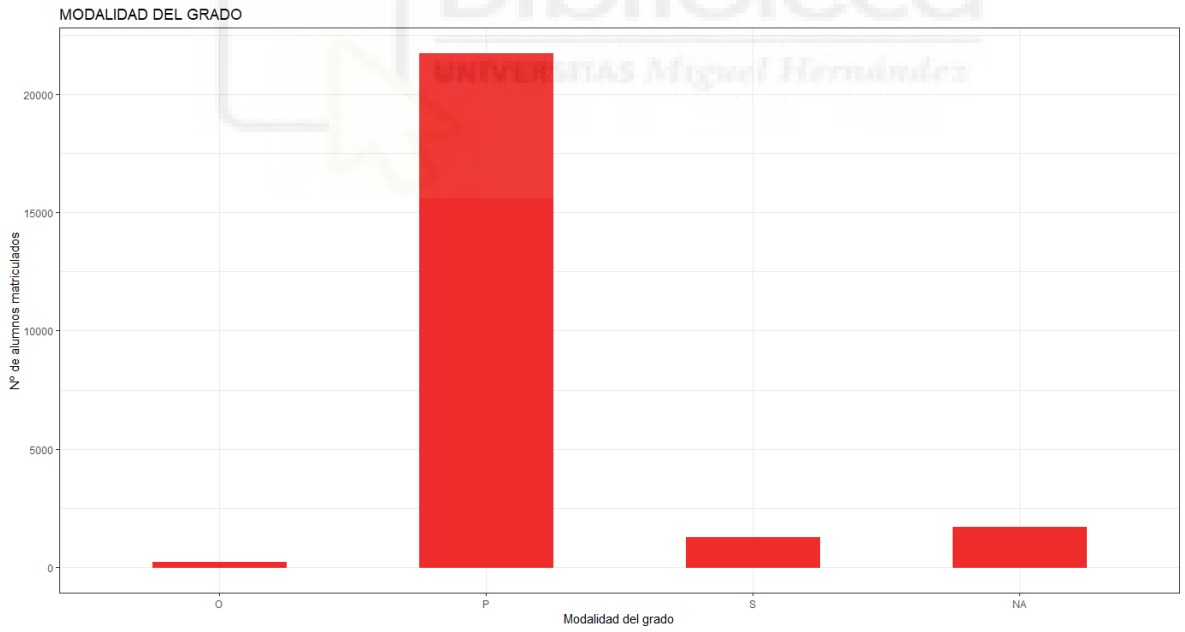
Figura 3: Distribución de las nota media (MEDIA C)



Las notas medias entre 6.5 y 7 son las más comunes en este estudio.

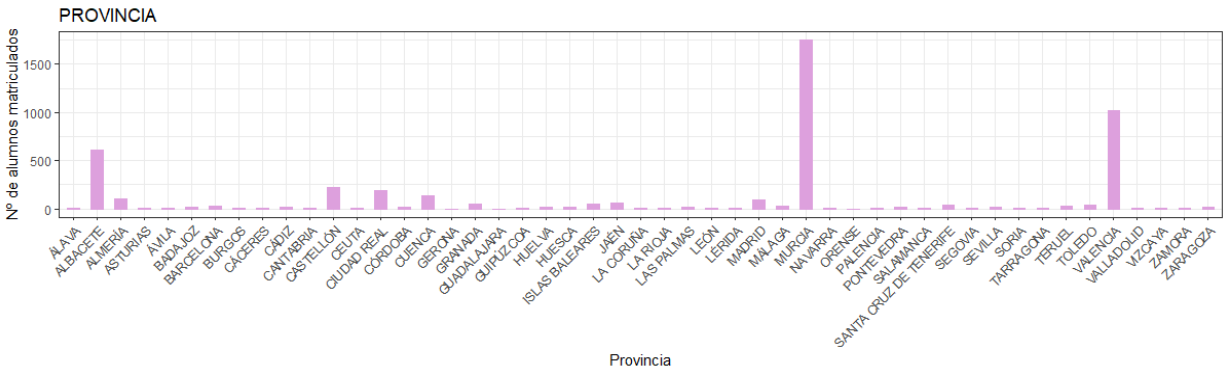
Además, también es interesante observar el análisis descriptivo del resto de las variables:

Figura 4: Distribución total de la modalidad del grado (O=Online, P=Presencial, S=Semipresencial)



Encontramos mayor número de estudiantes matriculados en la modalidad presencial, cerca del 90%.

Figura 5: Distribución total de los alumnos por provincia exceptuando Alicante

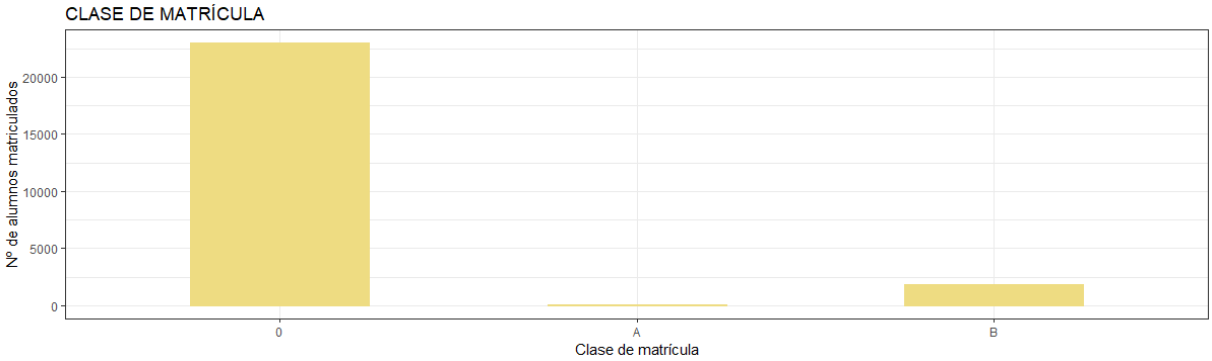


Exceptuando Alicante, la mayoría de los estudiantes matriculados en los grados de la UMH provienen de las provincias de Murcia, Valencia y Albacete, destacando Murcia con una diferencia importante.

Figura 6: Distribución total del tipo de beca que solicitan (A=Ambas, C=Conselleria, M=Ministerio, U=Universidad, NA=No solicitan beca)

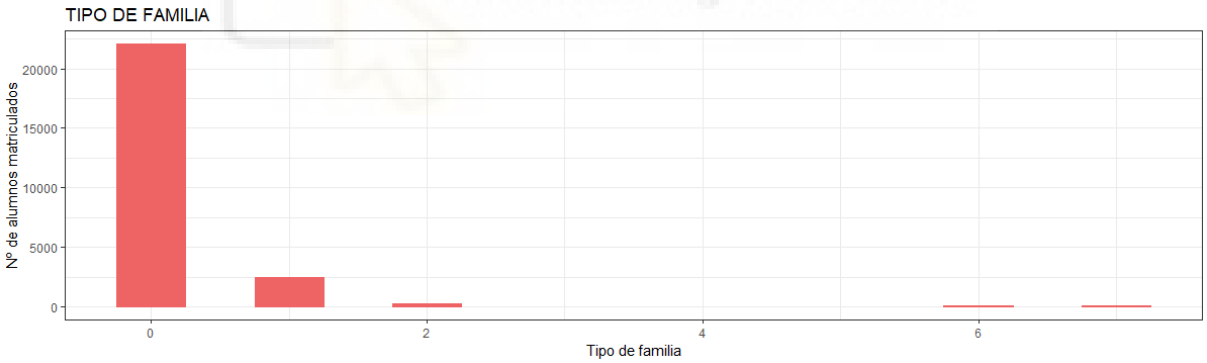
La beca más solicitada por parte de los estudiantes matriculados en grados en la UMH es la del Ministerio, pero la mayoría no solicita beca.

Figura 7: Distribución total de la clase de matrícula (0=Matrícula ordinaria, A=Alumno visitante, B=Alumno Prog. Intercambio)



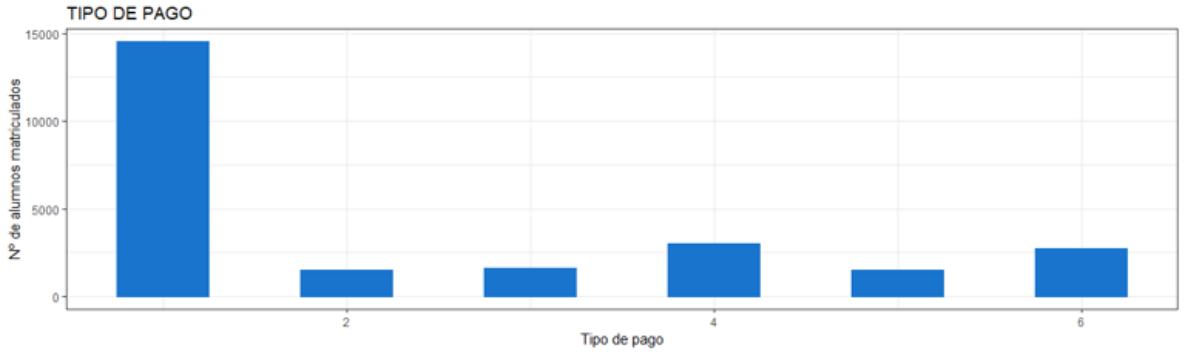
La matrícula más común en los estudiantes de grado de la UMH es la matrícula ordinaria con un 92,31 %.

Figura 8: Distribución total por tipo de familia (0=No es familia numerosa, 1=Familia Numerosa General, 2=Familia Numerosa Especial, 6=Familia Monoparental General, 7=Familia Monoparental Especial)



La mayoría de las familias no son familias numerosas, es decir son familias con uno o dos hijos (88,72%).

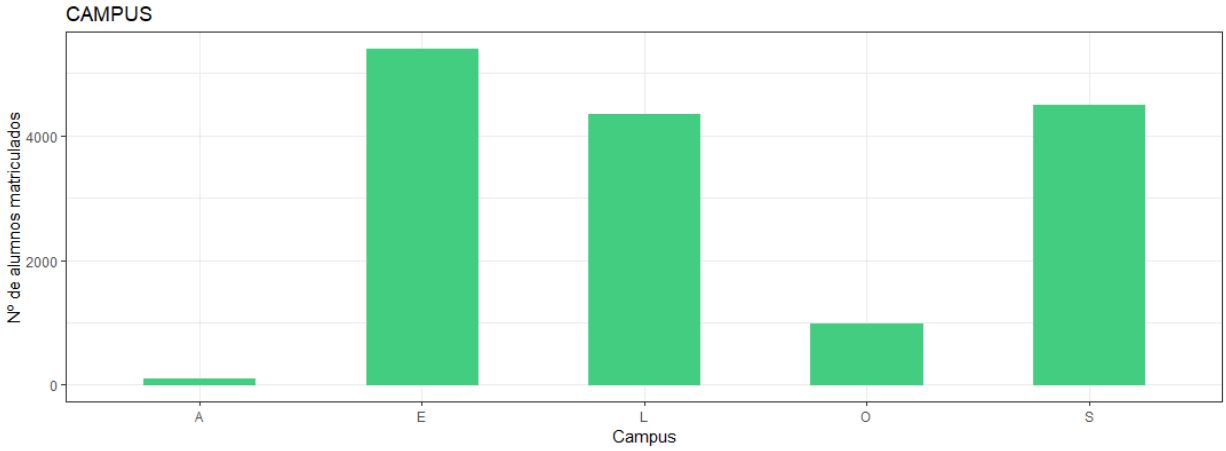
Figura 9: Distribución total según el tipo de pago (1=Pago en efectivo 1 plazo, 2=Pago en efectivo 2 plazos, 3=Cargo en cuenta 4 plazos, 4=Cargo en cuenta 1 plazo, 5=Cargo en cuenta 2 plazos, 6=Cargo en cuenta 8 plazos)



El más común entre los pagos de los estudiantes matriculados en los grados de la UMH es el pago en efectivo (1 plazo).

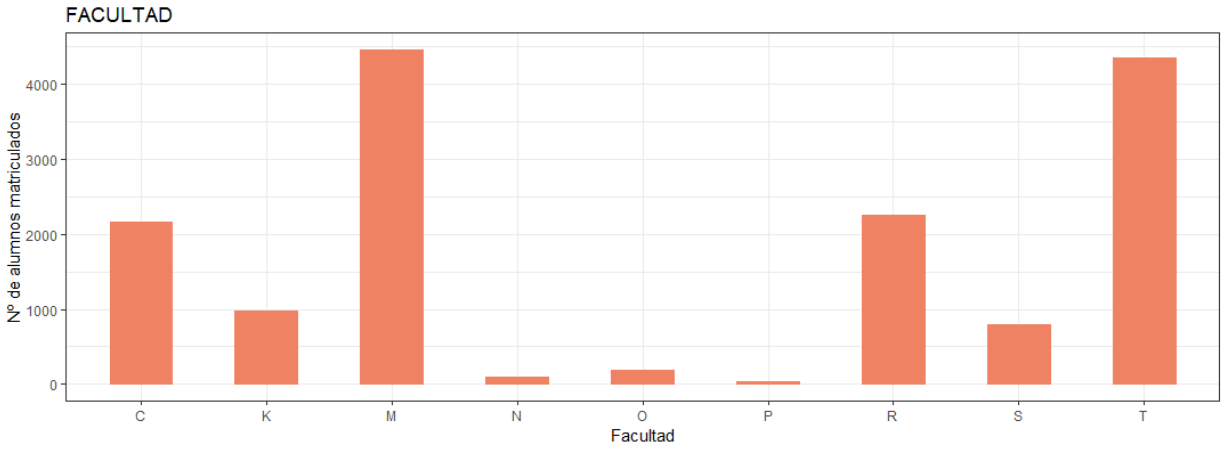
MÁSTERES

Figura 10: Distribución total de los estudiantes por Campus (A = Altea, E = Elche, L = Facultad de Ciencias Sociales y Jurídicas de Orihuela, O = Escuela Politécnica de Orihuela, S = San Juan).



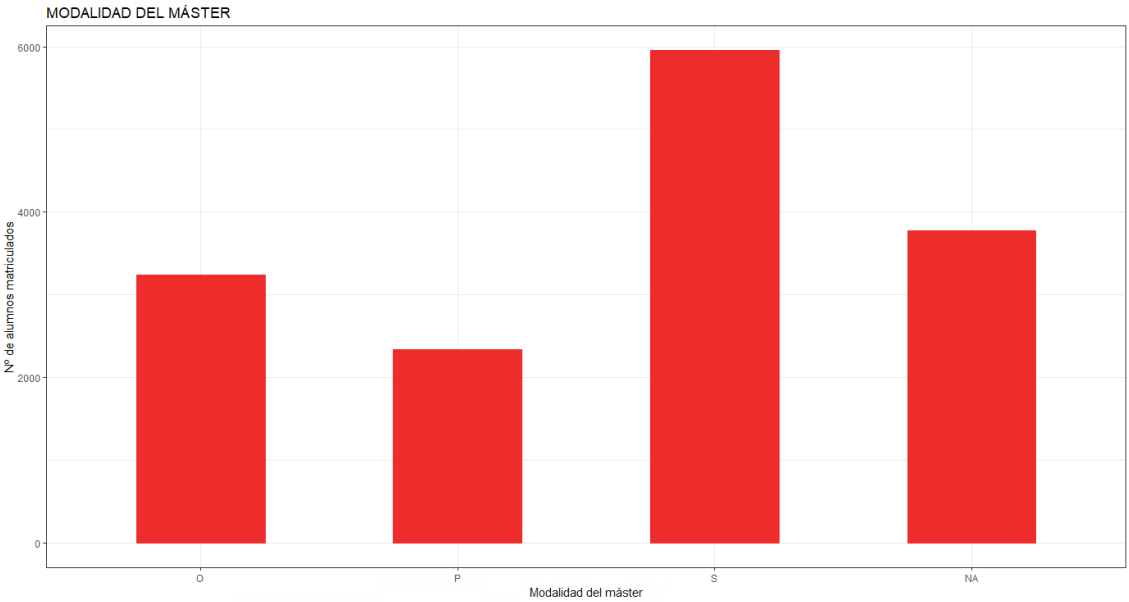
El campus de Elche es el que más alumnos matriculados tiene, le siguen el campus de San Juan y el de Facultad de Ciencias Sociales y Jurídicas de Orihuela.

Figura 11: Distribución total de los estudiantes por Facultad (C=F. Ciencias Sociosanitarias, K=Politécnica Orihuela, M=F. Medicina, N=F. Bellas Artes, O=F. Ciencias Experimentales, P=F. Farmacia, R=F.CC.SS. de Elche y Jurídicas Elche, S=Politécnica Elche, T=F.CC.SS y Jurídicas Orihuela)



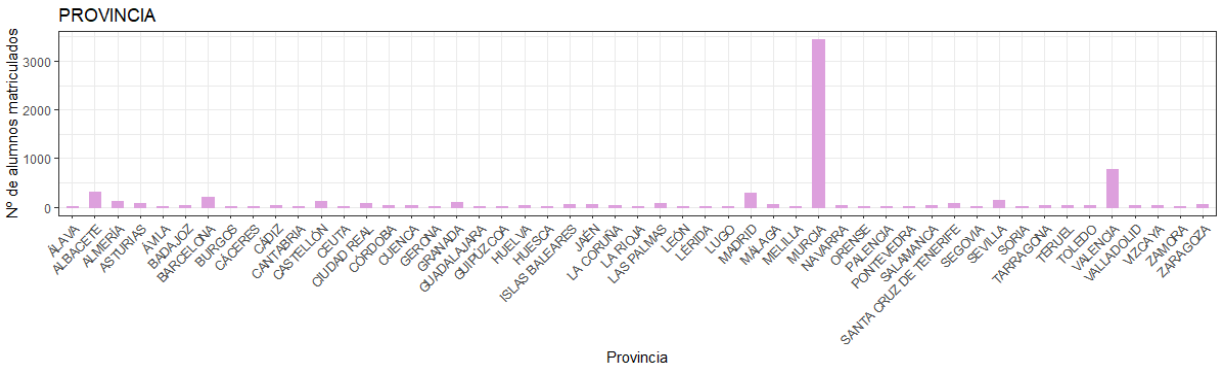
El mayor número de alumnos matriculados se encuentra en la facultad de Medicina y en la facultad de Ciencias Sociales y Jurídicas de Orihuela.

Figura 12: Distribución total de la modalidad del máster (O=Online, P=Presencial, S=Semipresencial)



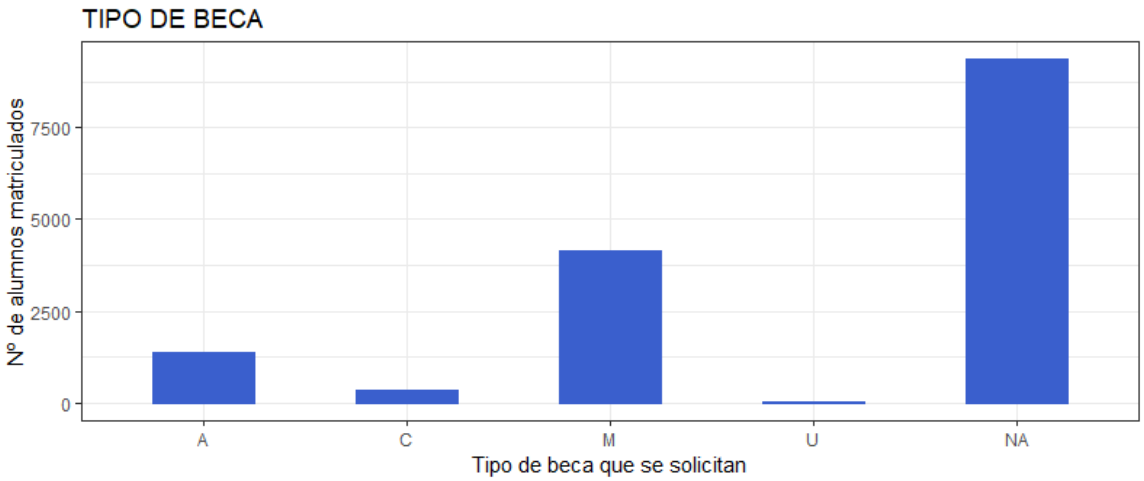
Encontramos mayor número de estudiantes matriculados en la modalidad semi-presencial (38,88%), a diferencia de los grados que los estudiantes el número de matriculados es mayor en la modalidad presencial.

Figura 13: Distribución total de los alumnos por provincia exceptuando Alicante



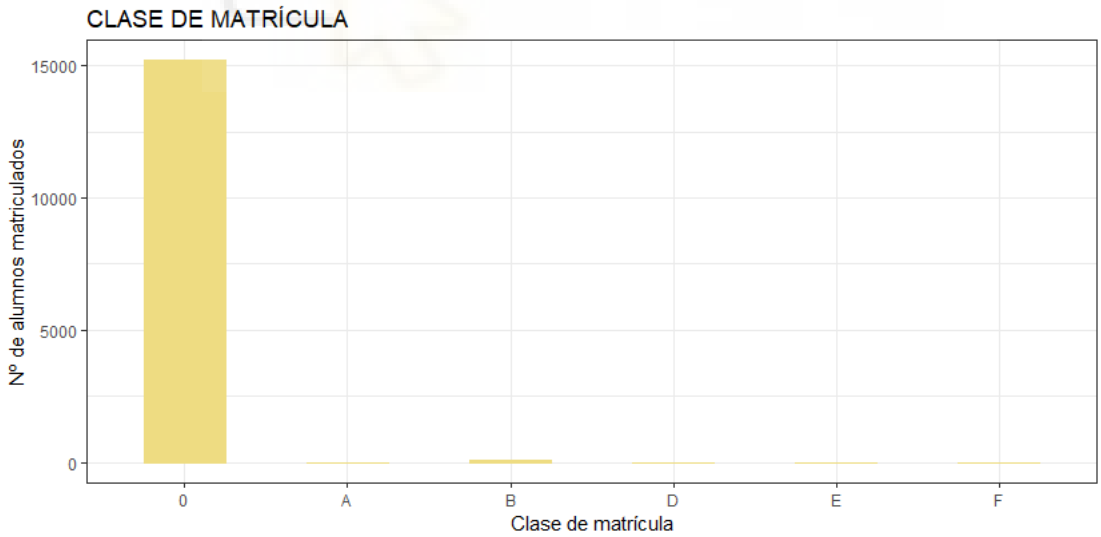
Exceptuando Alicante, la mayoría de los estudiantes matriculados en los másteres de la UMH provienen de las provincias de Murcia y Valencia, destacando Murcia con una diferencia importante.

Figura 14: Distribución total del tipo de beca que solicitan (A=Ambas, C=Conselleria, M=Ministerio, U=Universidad, NA=No solicitan beca)



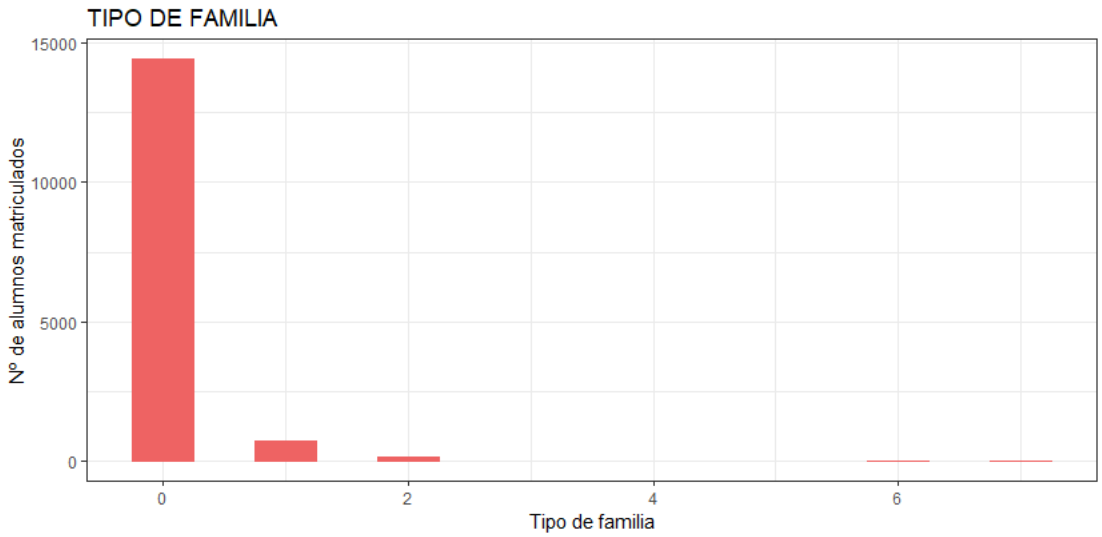
La beca más solicitada por parte de los estudiantes matriculados en másteres en la UMH es la del Ministerio, pero la mayoría no solicitan beca.

Figura 15: Distribución total de la clase de matrícula (0=Matrícula ordinaria, A=Alumno visitante, B=Alumno Prog. Intercambio)



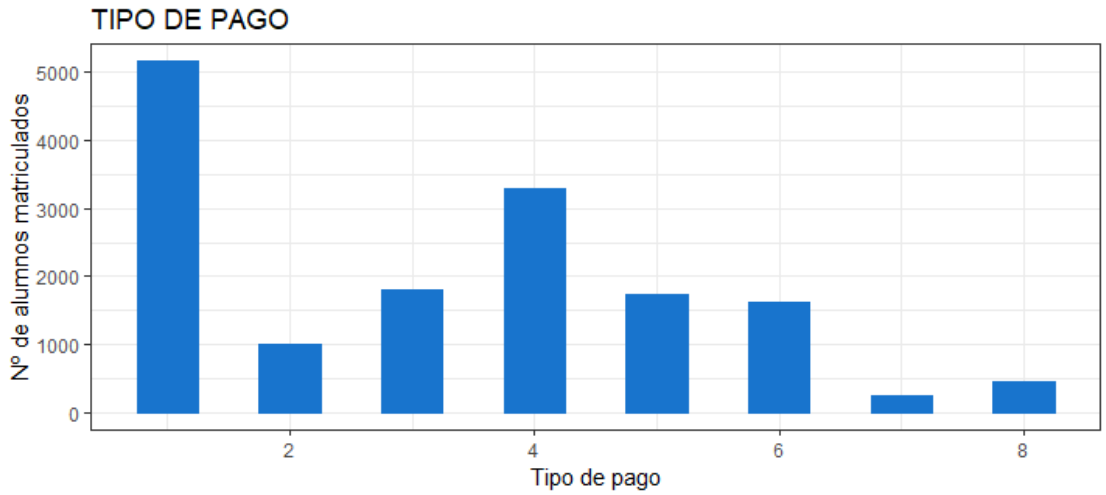
La matrícula más común en los estudiantes de máster de la UMH es la matrícula ordinaria con un 99,25%.

Figura 16: Distribución total por tipo de familia (0=No es familia numerosa, 1=Familia Numerosa General, 2=Familia Numerosa Especial, 6=Familia Monoparental General, 7=Familia Monoparental Especial)



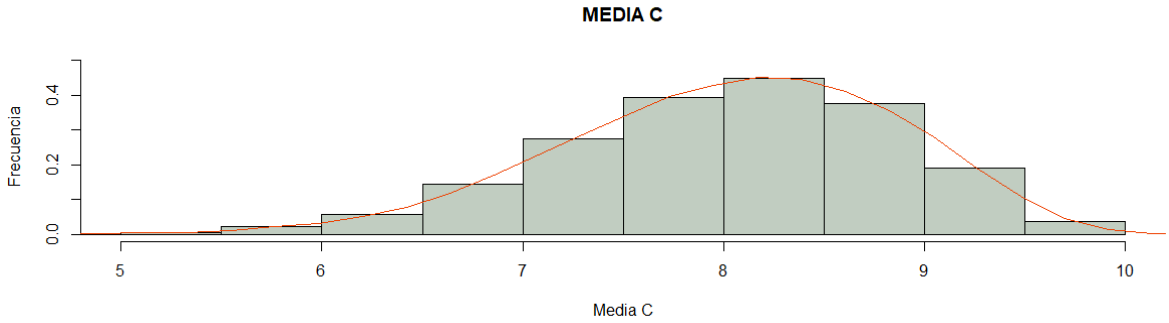
La mayoría de las familias no son familias numerosas, es decir son familias con uno o dos hijos (94,21%).

Figura 17: Distribución total según el tipo de pago (1=Pago en efectivo 1 plazo, 2=Pago en efectivo 2 plazos, 3=Cargo en cuenta 4 plazos, 4=Cargo en cuenta 1 plazo, 5=Cargo en cuenta 2 plazos, 6=Cargo en cuenta 8 plazos)



El más común entre los pagos de los estudiantes matriculados en los másteres de la UMH es el pago en efectivo en 1 plazo, en segundo lugar, le sigue el cargo en cuenta en 1 plazo.

Figura 18: Distribución de las nota media (MEDIA C)

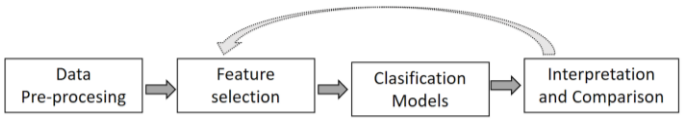


Las notas medias entre 8 y 8.5 son las más comunes en este estudio.

5.2 METODOLOGÍA DE ANÁLISIS

La metodología de la ciencia de datos se basa normalmente en el modelo CRISP-DM [2], que incluye las siguientes fases: pre-procesamiento (para la limpieza y la idoneidad de la base de datos); selección de atributos de los atributos más relevantes (seleccionando los más importantes para la construcción del modelo) y la generación de modelos de clasificación (con sus correspondientes detalles). Después de la evaluación de los modelos, se permite regresar a la fase de selección de atributos. Desde el año 2000, este esquema suele estar aceptado por muchos científicos de datos y se aplica ampliamente en diferentes marcos con ligeras variaciones [5].

Figura 19: Metodología, esquema general



A continuación, se muestra cómo estas fases han sido implementadas:

- **Pre-procesamiento de los datos**: La base de datos provista ha sido filtrada para extraer exclusivamente registros relacionados con estudiantes matriculados en grados y másteres (no se tiene en cuenta a los doctorandos o doctorados). Las tuplas que contienen errores y valores atípicos (valores fuera de rango se han eliminado. Las variables objetivo ABANDONO y MEDIAC_dis se han imputado como ya se ha explicado antes.

- **Selección de atributos**: El algoritmo VariableRanker de la librería Machine Learning de R ha sido aplicado con el fin de obtener un ranking de atributos, ordenados por relevancia. VariableRanker es una selección de atributos automática y un procedimiento de clasificación según importancia, basado sobre el concepto original de obtención de la información.

- **Modelos de clasificación**: En cada caso, el algoritmo de clasificación y regresión RPart ha sido parametrizado (RPart es una nueva implementación en R del conocido CART [1]) y alimentado con los correspondientes atributos más relevantes, con el fin de predecir las variables objetivos (ABANDONO y MEDIAC_dis). Como las variables objetivos son discretas, el algoritmo actúa como un clasificador.

- **Interpretación**: Finalmente, para las distintas facultades se han alcanzado precisiones diferentes con distintas matrices de confusión con los modelos predictivos obtenidos en la tercera fase.

6. SEGMENTACIÓN

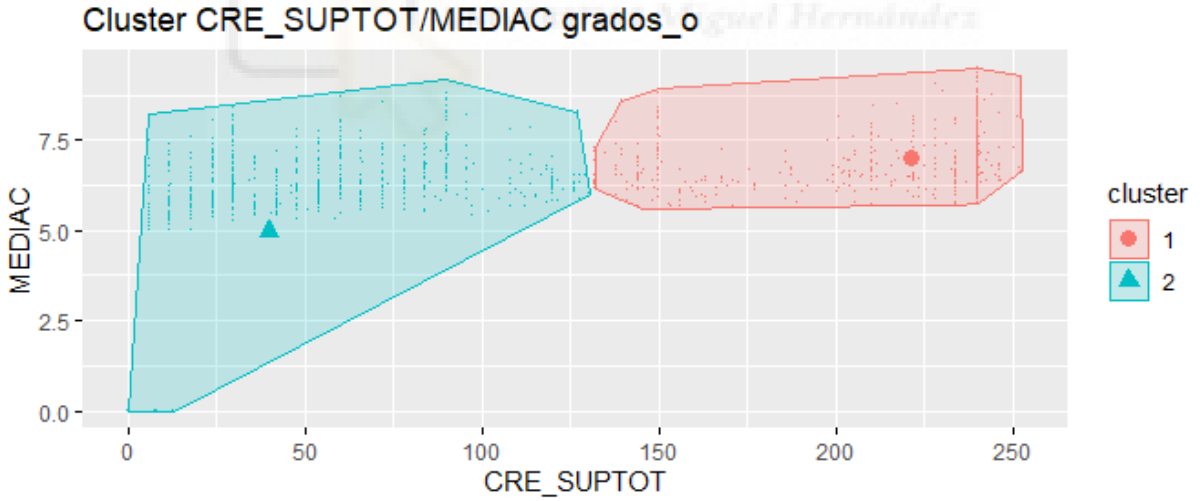
A continuación, se muestran los clusters o segmentaciones realizadas para cada facultad o escuela, generando 2, 3 y 4 grupos, siempre que en el grupo haya suficiente representatividad de individuos (superior al 5% de la muestra).

Estos gráficos representan, en diferentes colores y con su centroide o elemento más representativo, las nubes de puntos que aglutina en un mismo grupo a los estudiantes con mayores similitudes entre sí respecto a su calificación media (eje Y) y el total de créditos superados (eje X).

A continuación, se ilustra el ejemplo de la Facultad de Ciencias Experimentales, que incluyen el Grado en Biotecnología y el Grado en Ciencias Ambientales y el Máster en Biotecnología y Bioingeniería y el Máster Europeo en Cosmética Traslacional y Ciencias Dermatológicas (Erasmus Mundus). En el Anexo (apartado 8), se encuentran la totalidad de las facultades que pertenecen a la Universidad Miguel Hernández.

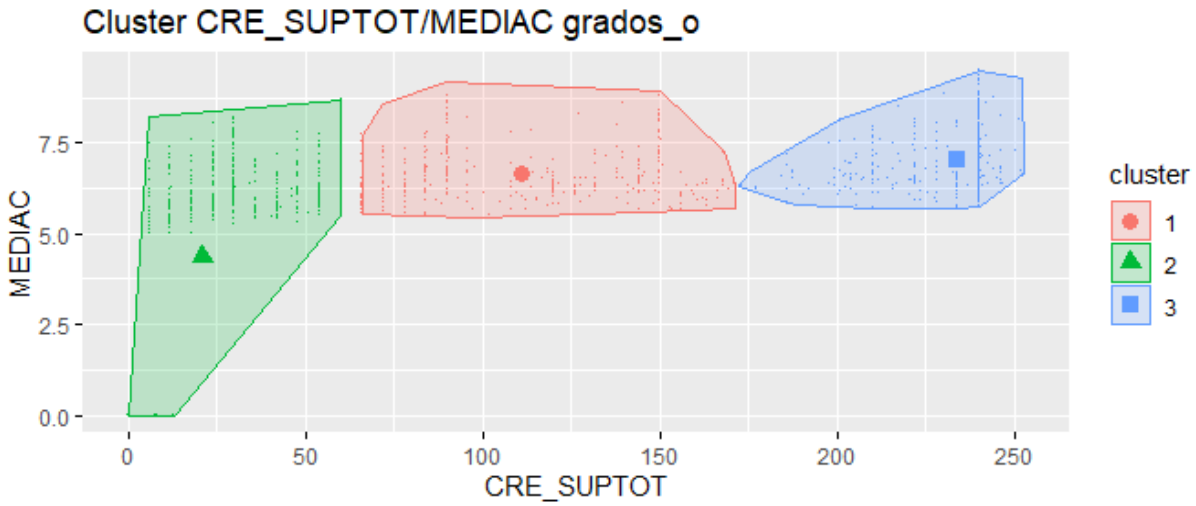
En total, en los grados, se está estudiando a 1.481 individuos y en el caso de los másteres a 186 individuos.

Figura 20: Cluster de los grados de la Facultad de Ciencias Experimentales (2 grupos)



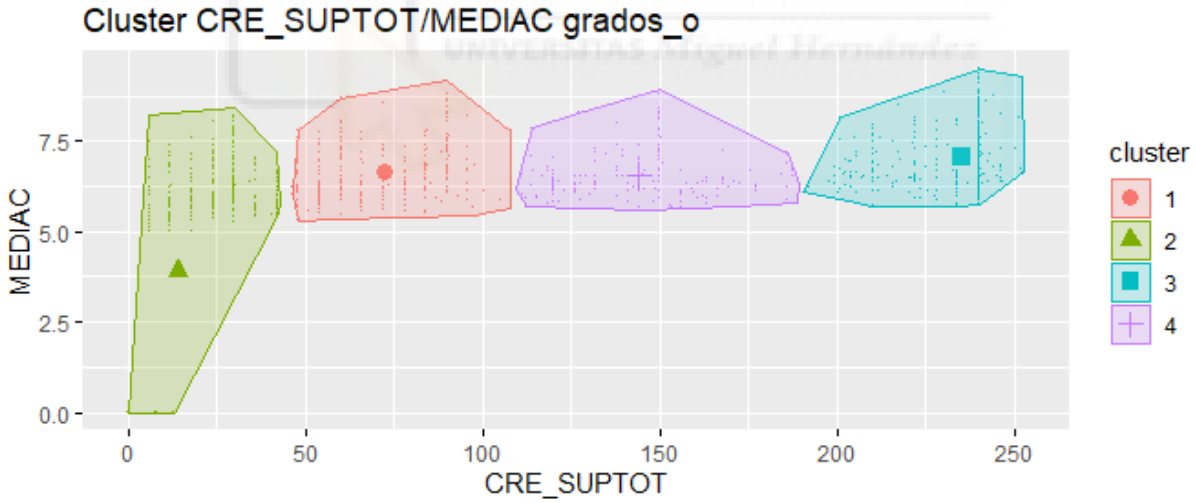
Tamaño grupos: 713, 768

Figura 21: Cluster de los grados de la Facultad de Ciencias Experimentales (3 grupos)



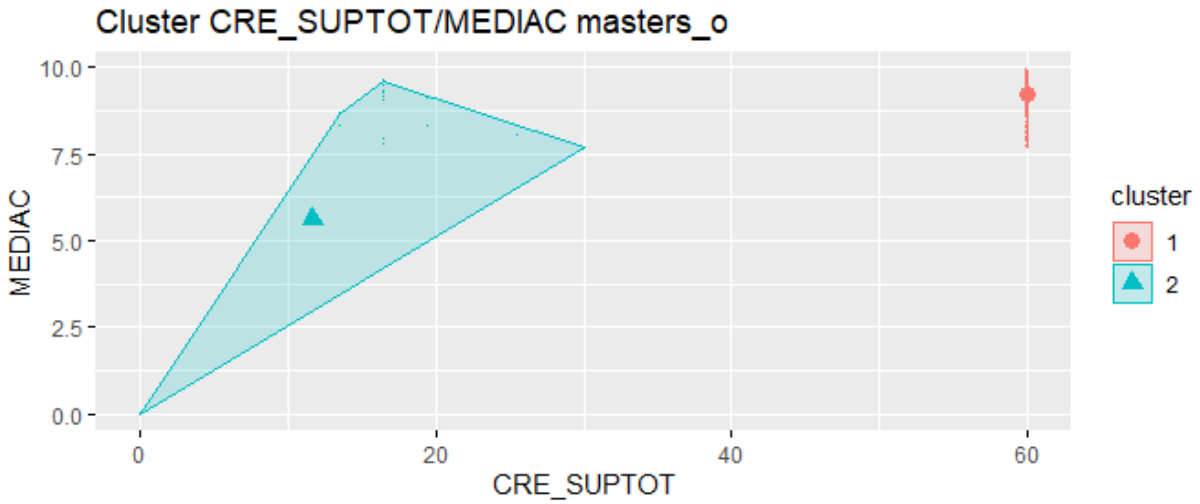
Tamaño grupos: 310, 562, 609

Figura 22: Cluster de los grados de la Facultad de Ciencias Experimentales (4 grupos)



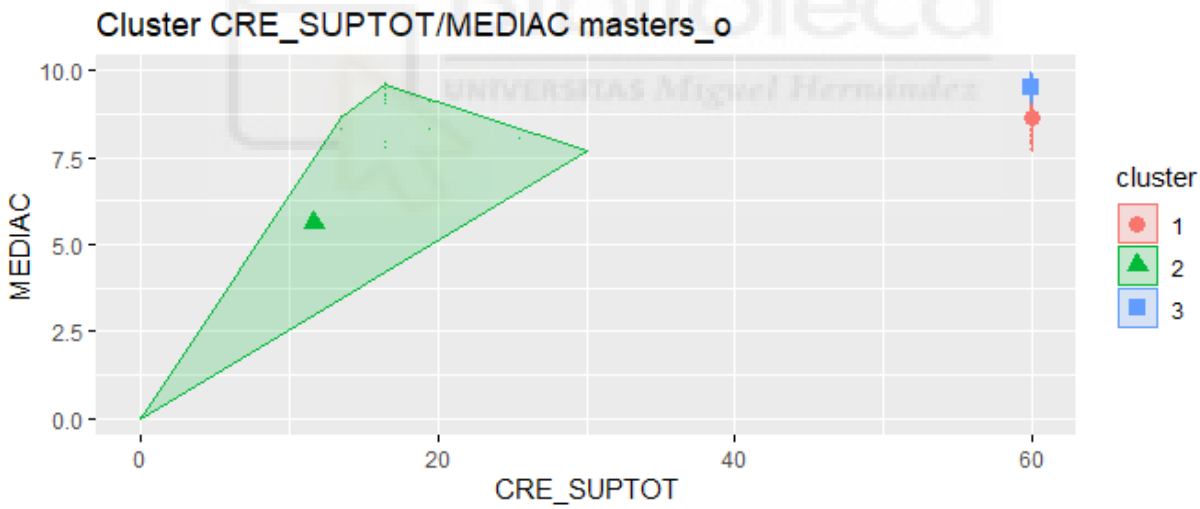
Tamaño grupos: 253, 468, 593, 167

Figura 23: Cluster de los másteres de la Facultad de Ciencias Experimentales (2 grupos)



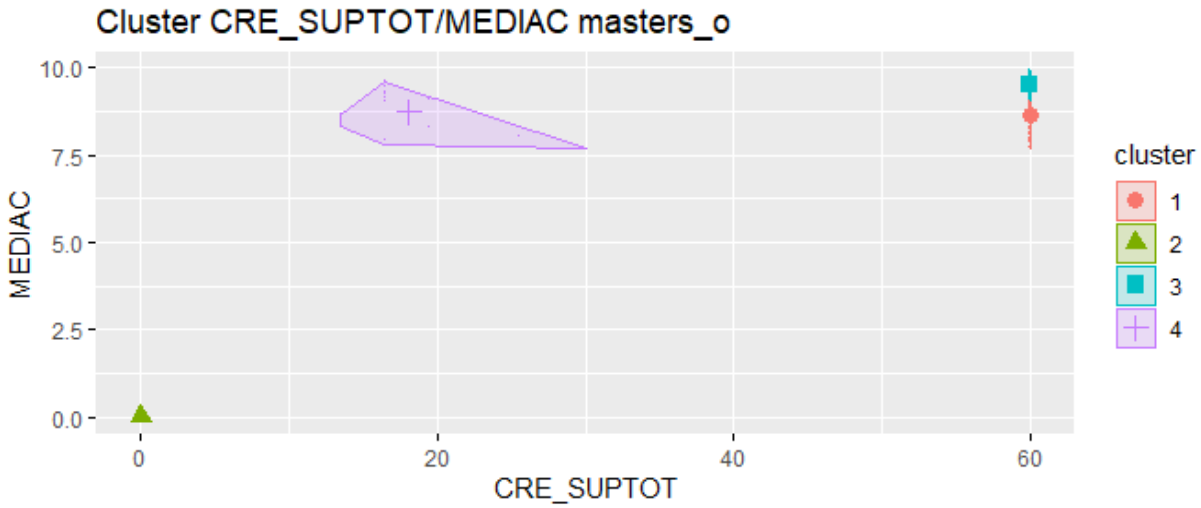
Tamaño grupos: 158, 28

Figura 24: Cluster de los másteres de la Facultad de Ciencias Experimentales (3 grupos)



Tamaño grupos: 47, 28, 111

Figura 25: Cluster de los másteres de la Facultad de Ciencias Experimentales (4 grupos)



Tamaño grupos: 47, 10, 111, 18

7. ANÁLISIS PREDICTIVOS

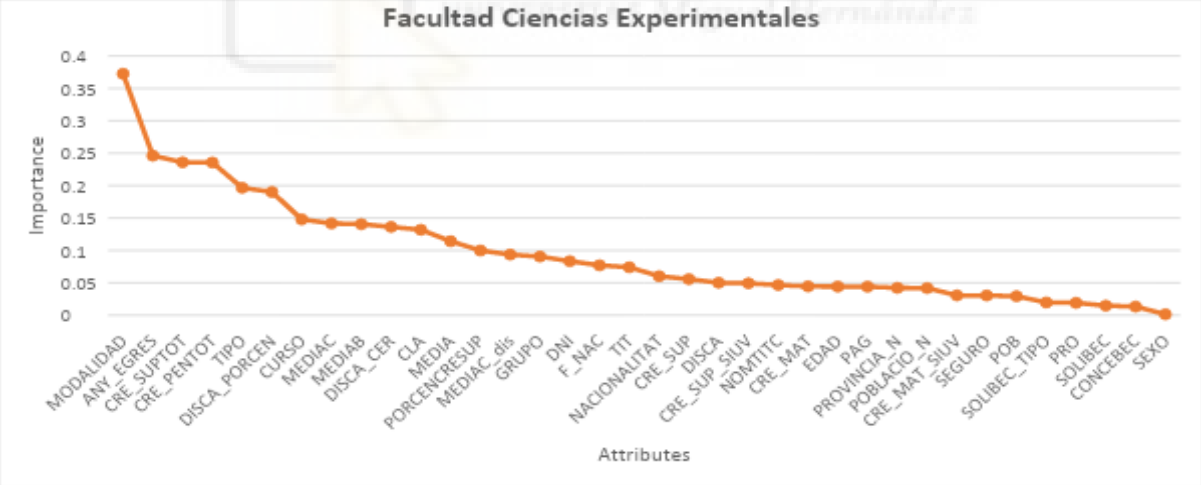
El objetivo de este estudio es crear un sistema de alerta temprana para evitar el abandono de los estudiantes cuando se cumplan una serie de patrones, a la vez que se revisa el rendimiento del estudiante mediante la nota media. Debido a esta necesidad es necesario estudiar cuáles son las variables que más influyen en estos ámbitos para poder realizar árboles de clasificación que ayuden a detectar los casos más preocupantes.

A continuación, se ilustra el ejemplo de la Facultad de Ciencias Experimentales, que incluyen el Grado en Biotecnología y el Grado en Ciencias Ambientales y el Máster en Biotecnología y Bioingeniería y el Máster Europeo en Cosmética Traslacional y Ciencias Dermatológicas (Erasmus Mundus). En el Anexo (apartado 8), se encuentran la totalidad de las facultades que pertenecen a la Universidad Miguel Hernández.

En total, en los grados, se está estudiando a 1.481 individuos y en el caso de los másteres a 186 individuos.

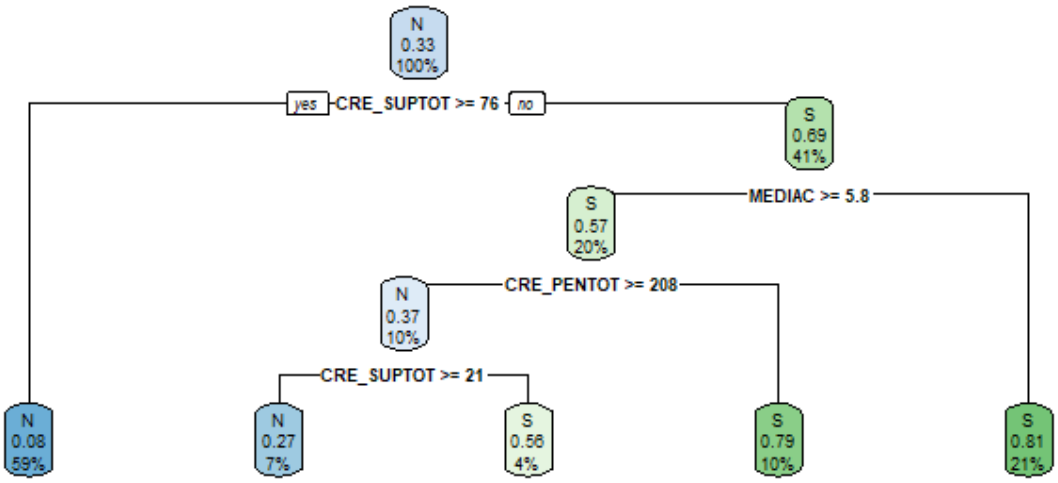
7.1 ABANDONO

Figura 26: Peso de las variables más importantes en la Facultad de Ciencias Experimentales (GRADOS)



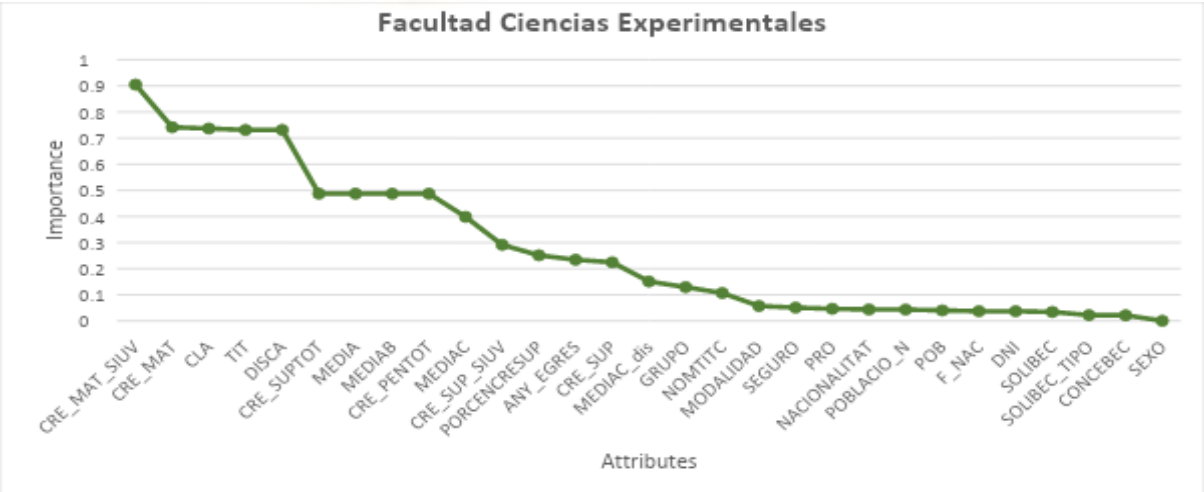
El nivel de presencialidad del grado, el año de finalización de los estudios y los créditos superados y pendientes totales son algunas de las variables que más influyen en el abandono de los grados por parte de los estudiantes de la Facultad de Ciencias Experimentales.

Figura 27: Árbol de clasificación de la Facultad de Ciencias Experimentales (GRADOS)



Como resultado del modelo con Accuracy 0.859459, por la rama más a la derecha, podemos concluir que los estudiantes que tienen menos de 76 créditos superados totales y que además su media ponderada es menor de 5.8, tienen un 81% de probabilidad de abandonar el grado.

Figura 28: Peso de las variables más importantes en la Facultad de Ciencias Experimentales (MÁSTERS)

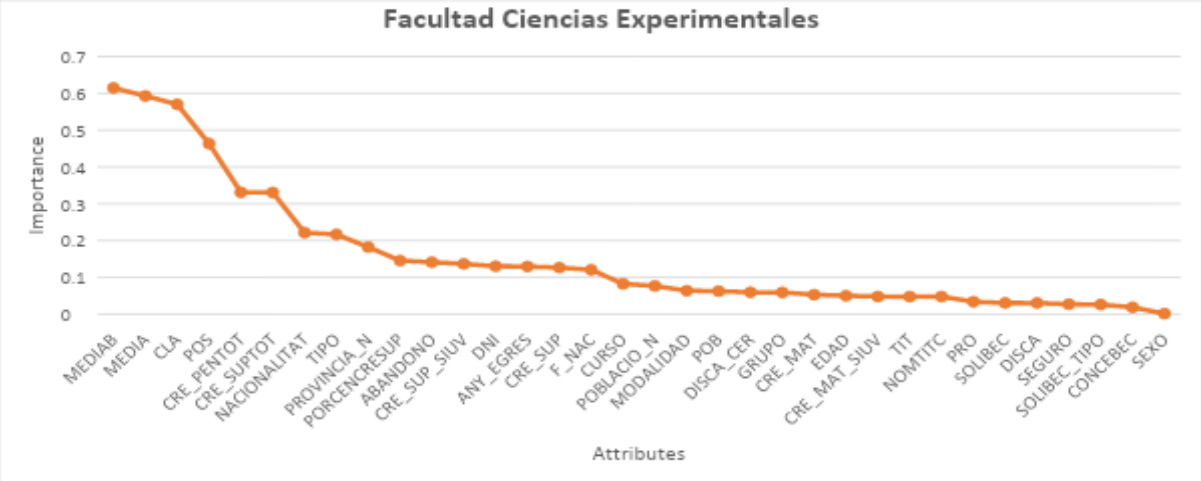


Los créditos matriculados, la clase de matrícula y si el estudiante es discapacitado o no son algunas de las variables que más influyen en el abandono de los másteres por parte de los estudiantes de la Facultad de Ciencias Experimentales.

No se puede crear un árbol de clasificación para los másteres en la Facultad de Ciencias Experimentales debido a que los datos son insuficientes.

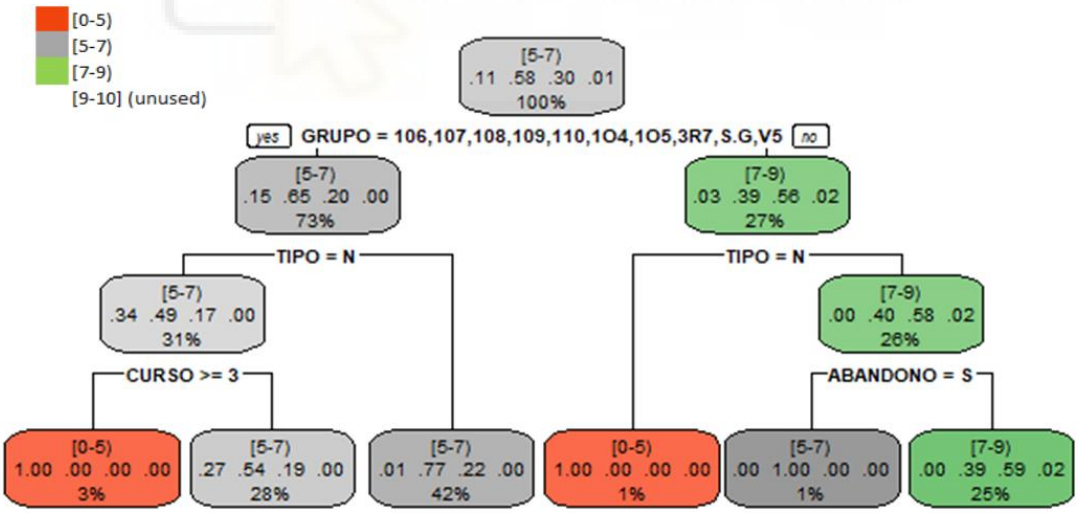
7.2 NOTA MEDIA

Figura 29: Peso de las variables más importantes en la Facultad de Ciencias Experimentales (GRADOS)



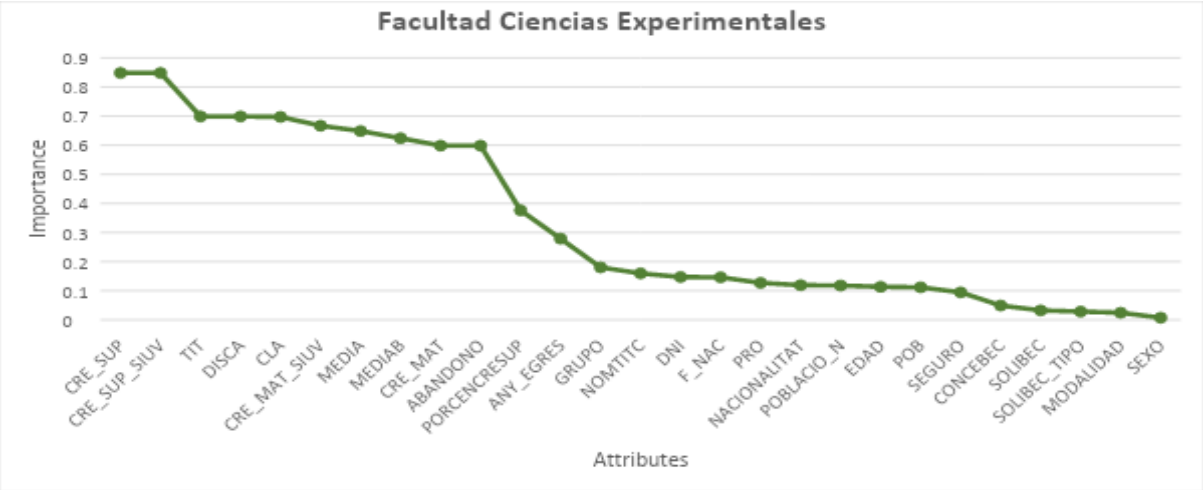
La clase de matrícula y el código postal son las variables que más influyen en la nota media obtenida por los estudiantes de la Facultad de Ciencias Experimentales.

Figura 30: Árbol de clasificación de la Facultad de Ciencias Experimentales (GRADOS)



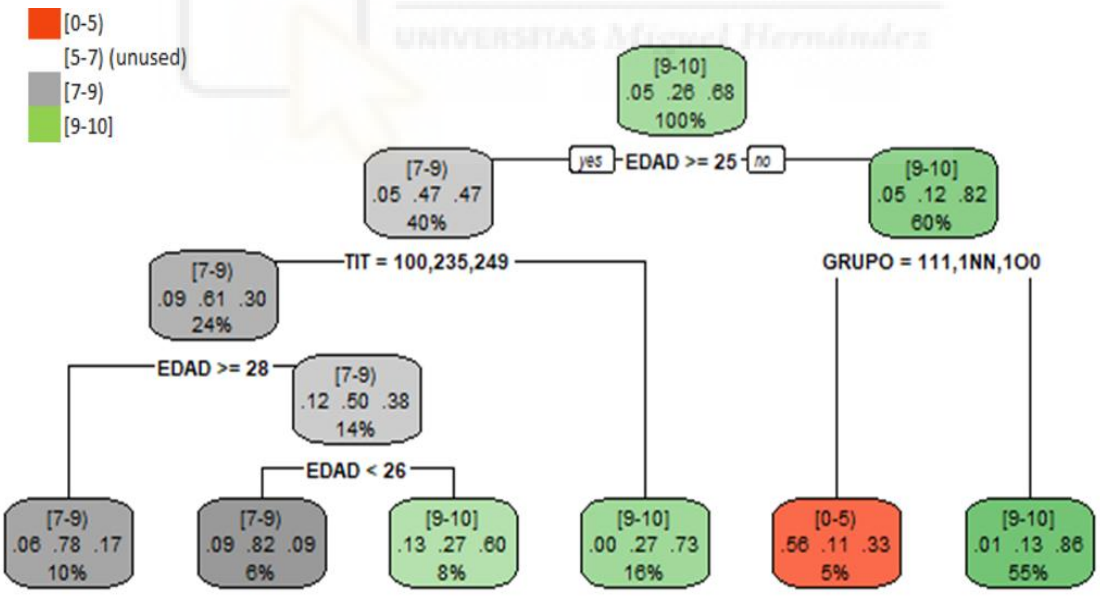
Como resultado del modelo con Accuracy 0.672973, por la rama derecha, podemos concluir que los estudiantes que pertenecen al resto de grupos que no son los indicados en la figura, que además no son estudiantes de nuevo ingreso y no abandonan el grado tienen un 59% de probabilidad de obtener una buena calificación media.

Figura 31: Peso de las variables más importantes en la Facultad de Ciencias Experimentales (MÁSTERES)



En el caso de los másteres, la variable que más influye en la nota media obtenida por los estudiantes de la Facultad de Ciencias Experimentales es los créditos superados.

Figura 32: Árbol de clasificación de la Facultad de Ciencias Experimentales (MÁSTERES)



Como resultado del modelo con Accuracy 0.673913, por la rama más a la izquierda, podemos concluir que los estudiantes con 28 años o mayores que pertenecen a las titulaciones 100, 235 y 249, tienen un 78% de probabilidad de obtener una nota media muy baja.

8. CONCLUSIONES

Para la realización de este trabajo, ha sido necesario realizar un adecuado preprocesamiento de los datos para obtener resultados lo más veraces posible. Además, también se ha cuidado la correcta discretización de las variables objetivo para optimizar lo máximo posible el accuracy de los árboles de clasificación.

Posteriormente ha sido conveniente realizar modelos de segmentación cluster, debido a que ayudan a separar a los individuos en grupos según características en común y los gráficos son de fácil interpretación.

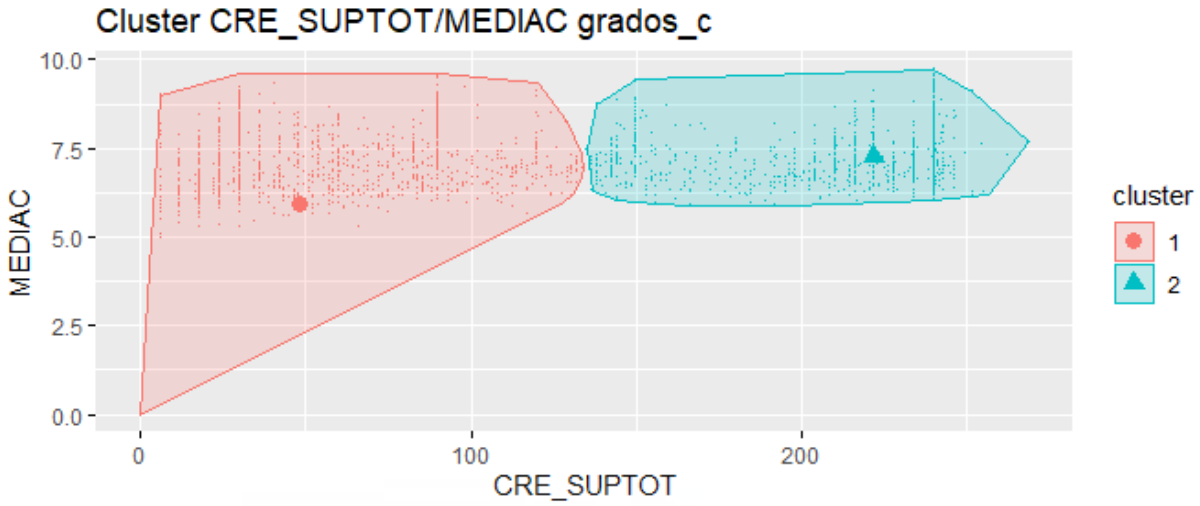
También ha sido necesario realizar los árboles de clasificación, generando previamente una selección de atributos según las variables objetivo, para la toma de decisiones organizada. La selección de atributos es un paso muy importante a realizar, ya que así el árbol realiza la clasificación directamente con las variables más significativas y supone un ahorro de tiempo.

La importancia de generar modelos basados en datos es necesario para facilitar la gestión en grandes organizaciones, como la Universidad, para poder facilitar la toma de decisiones y predecir, en este caso, el abandono estudiantil y el rendimiento académico de los alumnos para mejorar la calidad de la enseñanza y actuar de forma temprana en los casos más preocupantes de ambos escenarios.

9. ANEXO

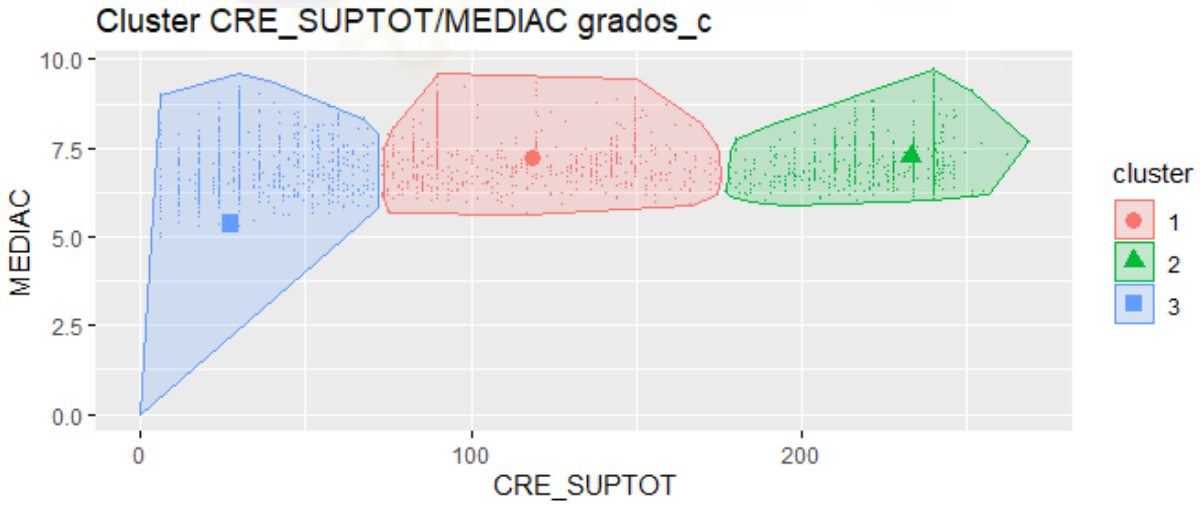
9.1 SEGMENTACIÓN

Figura 33: Cluster de los grados de la Facultad de Ciencias Sociosanitarias (2 grupos)



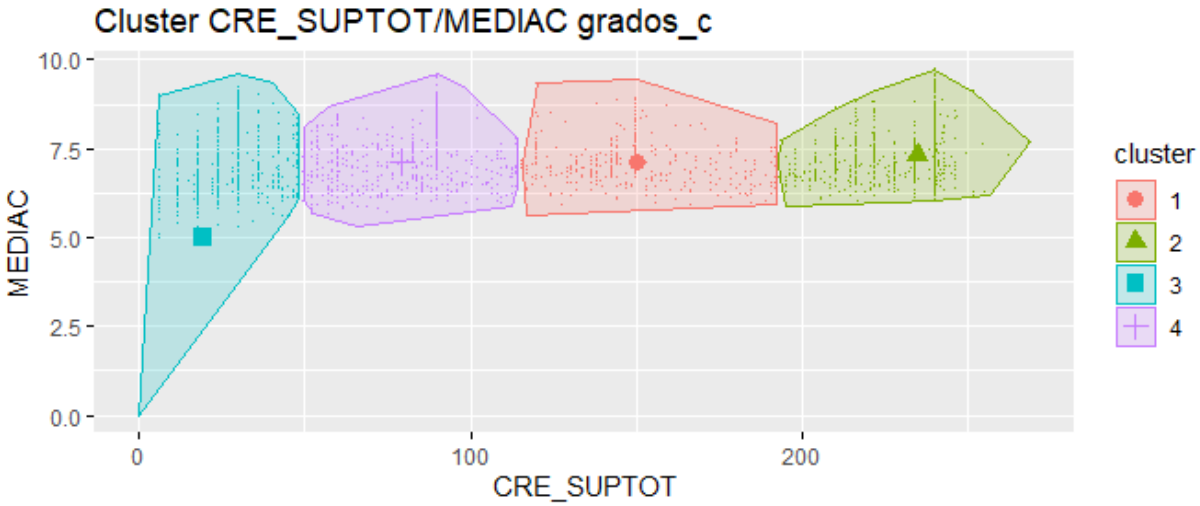
Tamaño grupos: 1327, 1781

Figura 34: Cluster de los grados de la Facultad de Ciencias Sociosanitarias (3 grupos)



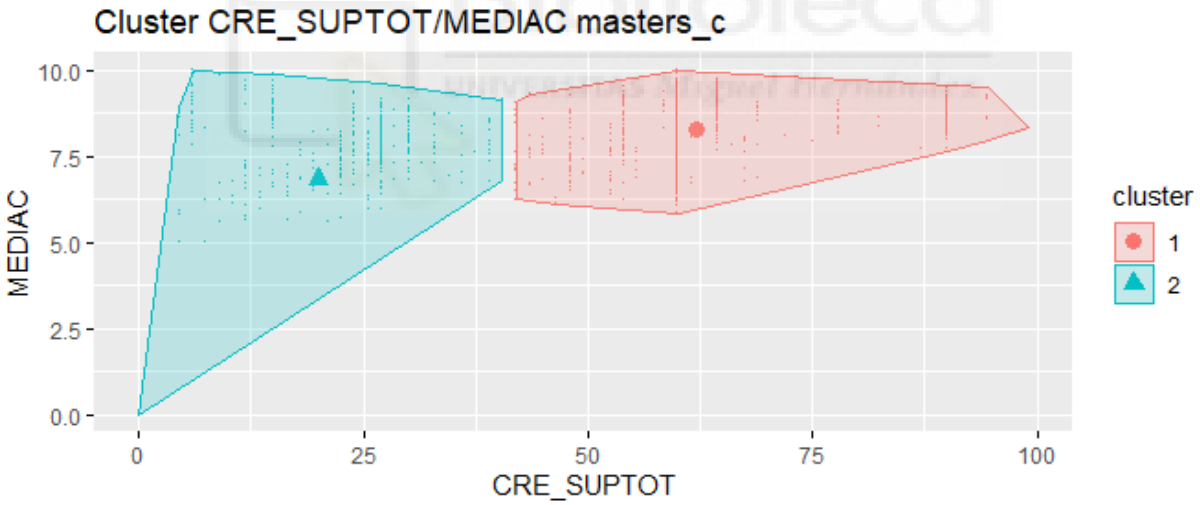
Tamaño grupos: 633, 1537, 938

Figura 35: Cluster de los grados de la Facultad de Ciencias Sociosanitarias (4 grupos)



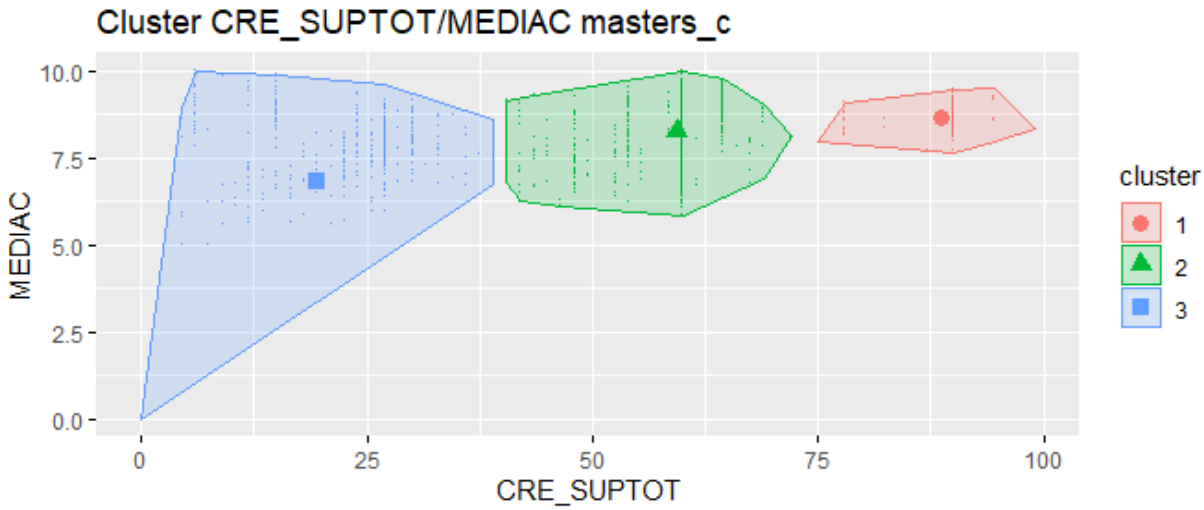
Tamaño grupos: 387, 1480, 755, 486

Figura 36: Cluster de los másteres de la Facultad de Ciencias Sociosanitarias (2 grupos)



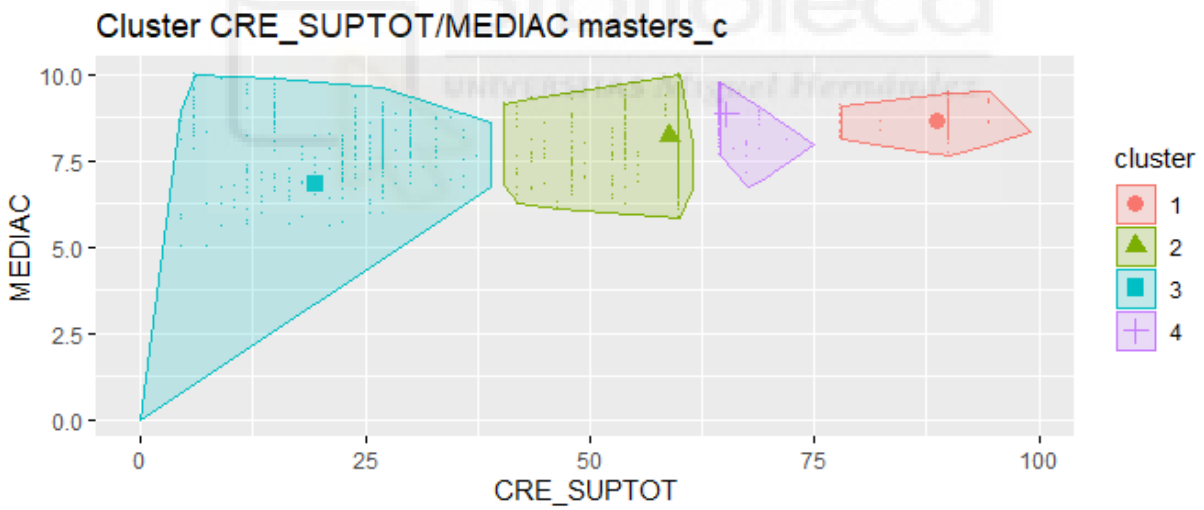
Tamaño grupos: 1725, 439

Figura 37: Cluster de los másteres de la Facultad de Ciencias Sociosanitarias (3 grupos)



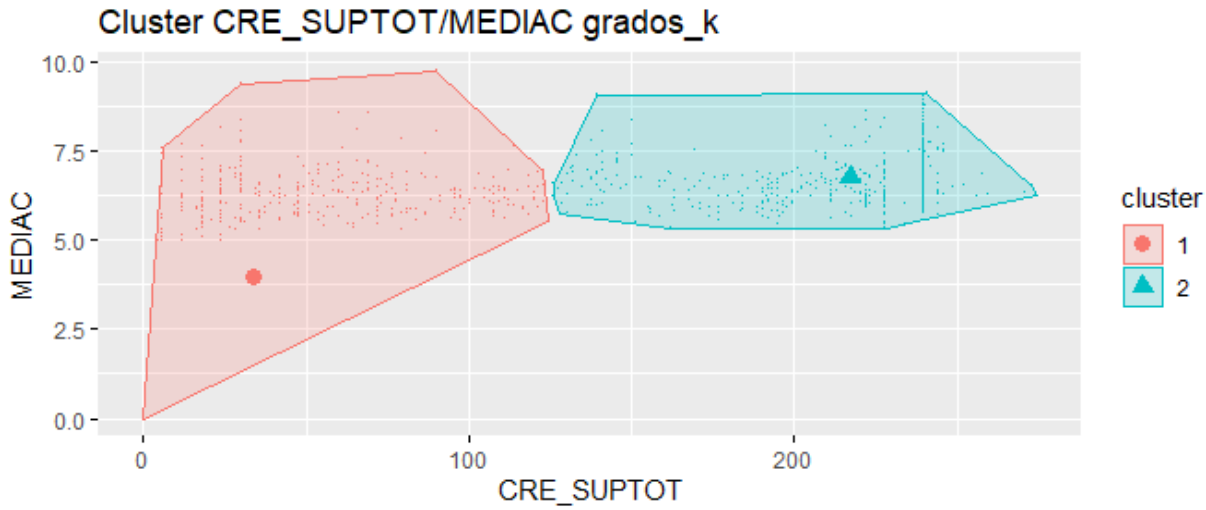
Tamaño grupos: 148, 1587, 429

Figura 38: Cluster de los másteres de la Facultad de Ciencias Sociosanitarias (4 grupos)



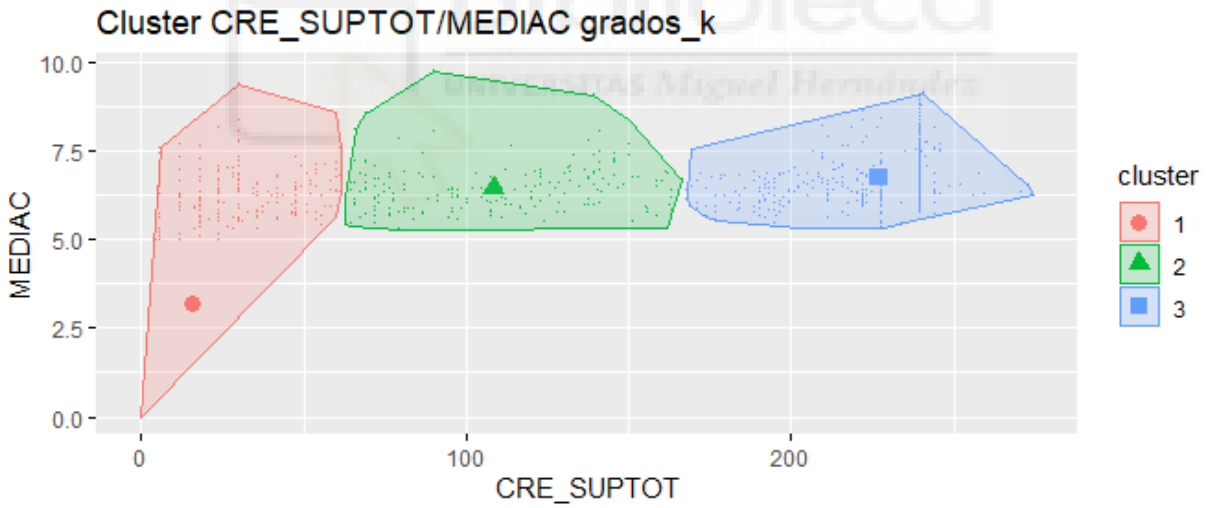
Tamaño grupos: 147, 1466, 429, 122

Figura 39: Cluster de los grados de la Escuela Politécnica de Orihuela (2 grupos)



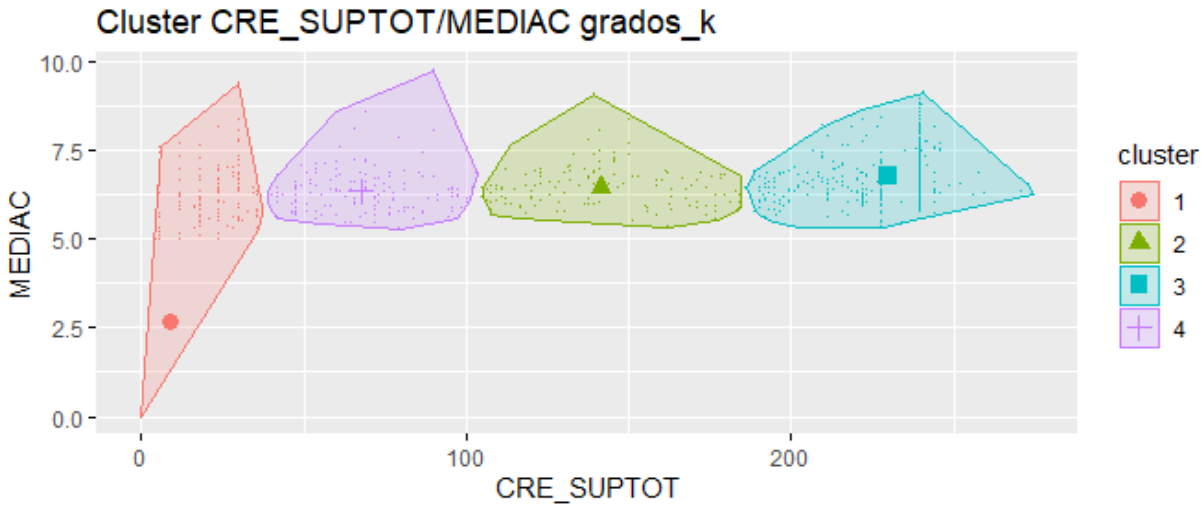
Tamaño grupos: 541, 565

Figura 40: Cluster de los grados de la Escuela Politécnica de Orihuela (3 grupos)



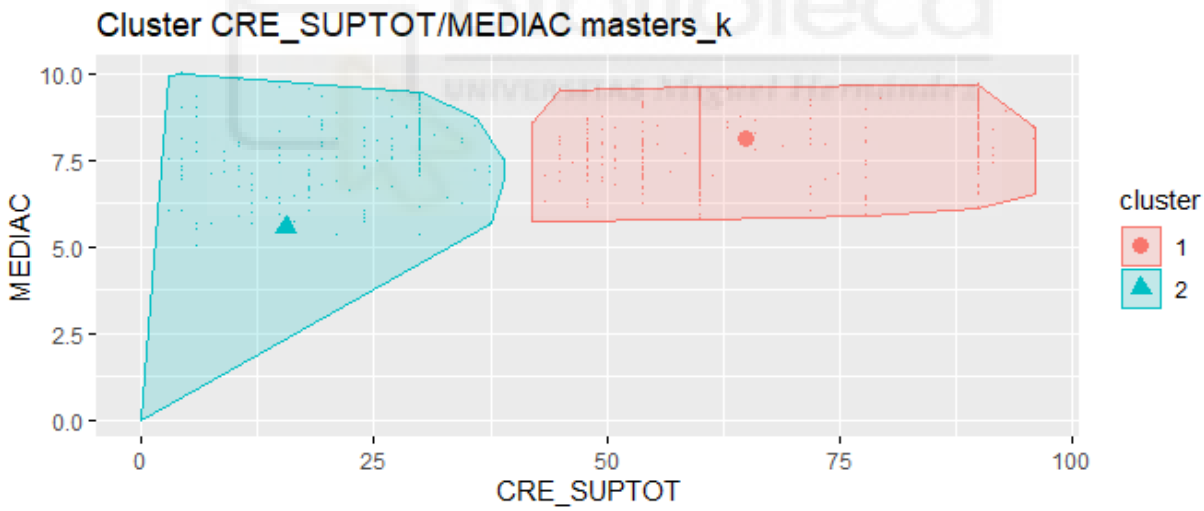
Tamaño grupos: 409, 198, 499

Figura 41: Cluster de los grados de la Escuela Politécnica de Orihuela (4 grupos)



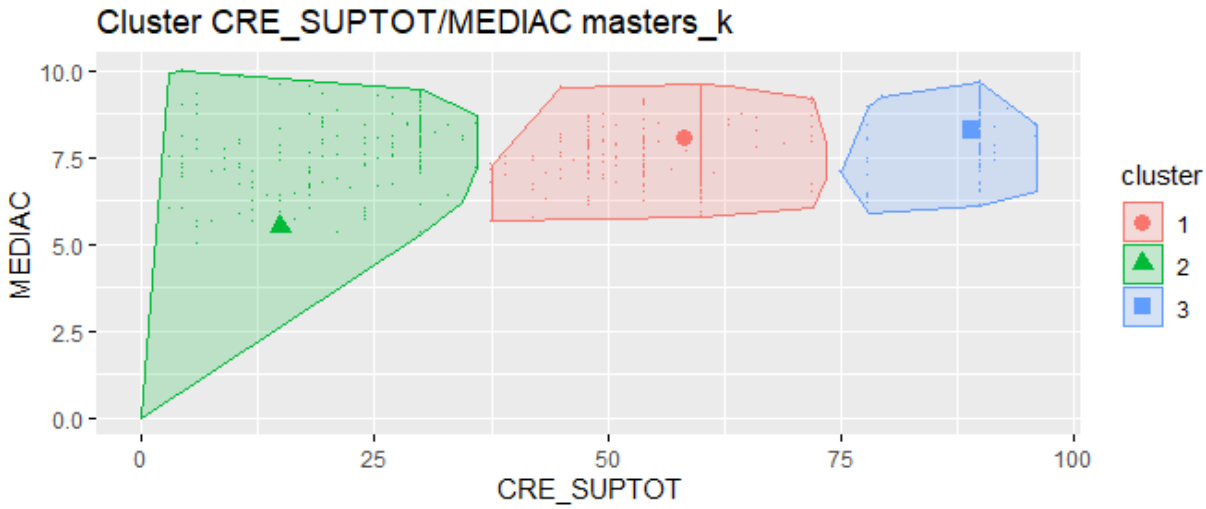
Tamaño grupos: 347, 137, 471, 151

Figura 42: Cluster de los másteres de la Escuela Politécnica de Orihuela (2 grupos)



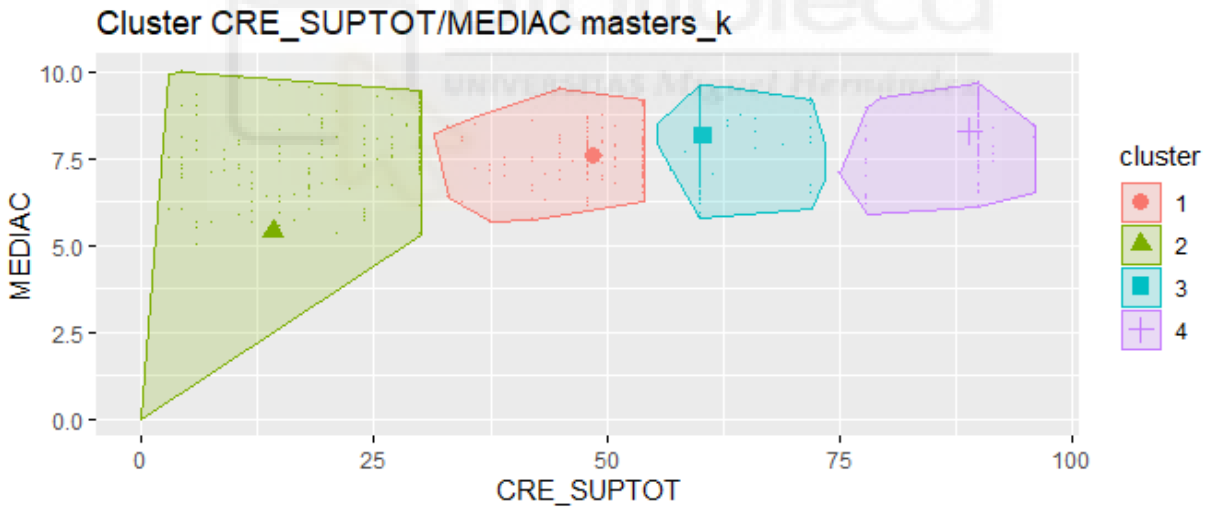
Tamaño grupos: 730, 246

Figura 43: Cluster de los másteres de la Escuela Politécnica de Orihuela (3 grupos)



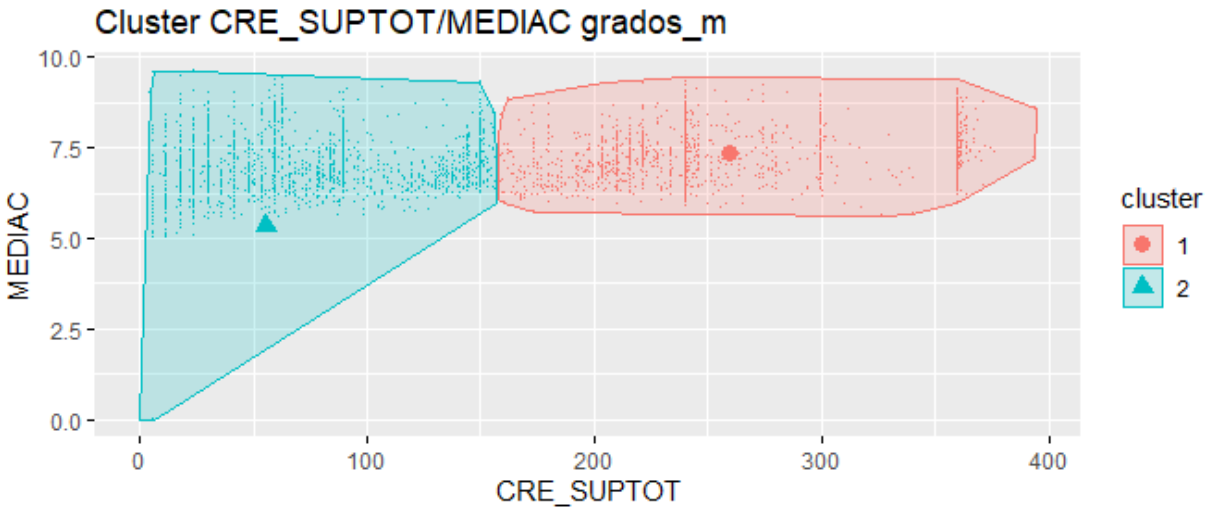
Tamaño grupos: 583, 239, 154

Figura 44: Cluster de los másteres de la Escuela Politécnica de Orihuela (4 grupos)



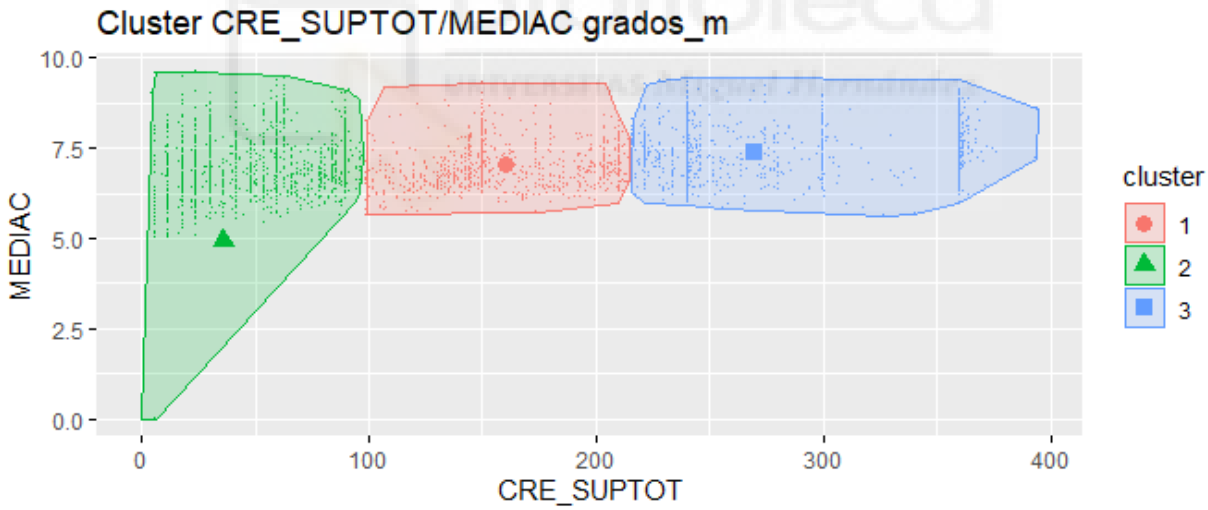
Tamaño grupos: 125, 230, 467, 154

Figura 45: Cluster de los grados de la Facultad de Medicina (2 grupos)



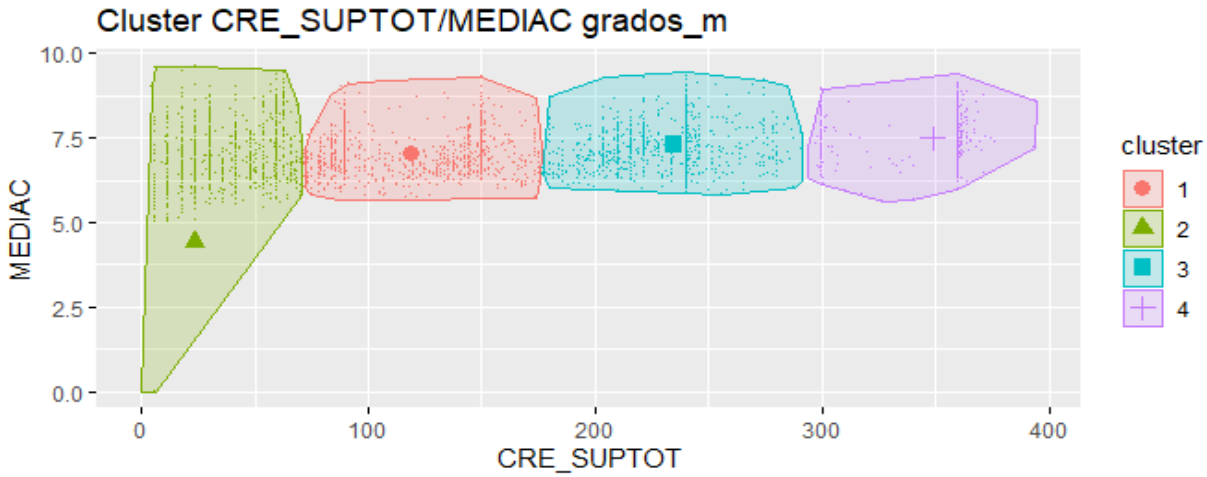
Tamaño grupos: 2600, 2389

Figura 46: Cluster de los grados de la Facultad de Medicina (3 grupos)



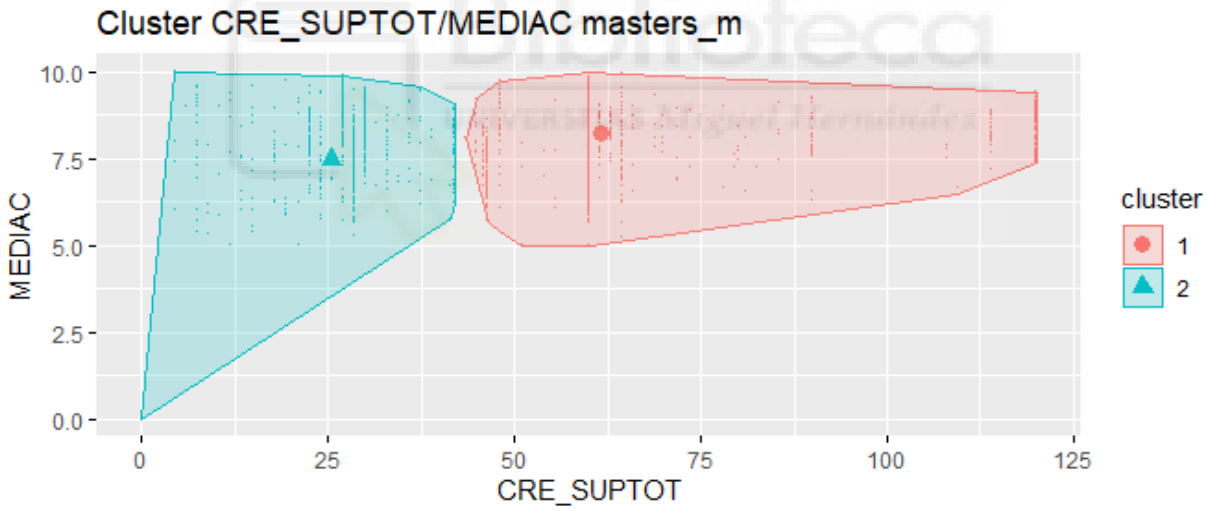
Tamaño grupos: 809, 1923, 2257

Figura 47: Cluster de los grados de la Facultad de Medicina (4 grupos)



Tamaño grupos: 927, 1541, 1914, 607

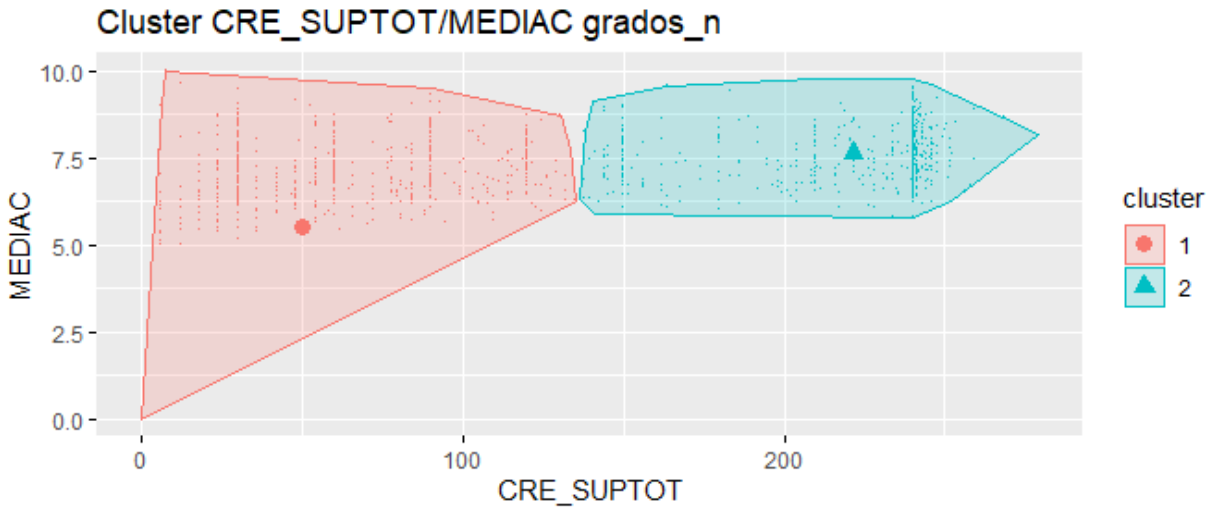
Figura 48: Cluster de los másteres de la Facultad de Medicina (2 grupos)



Tamaño grupos: 3695, 769

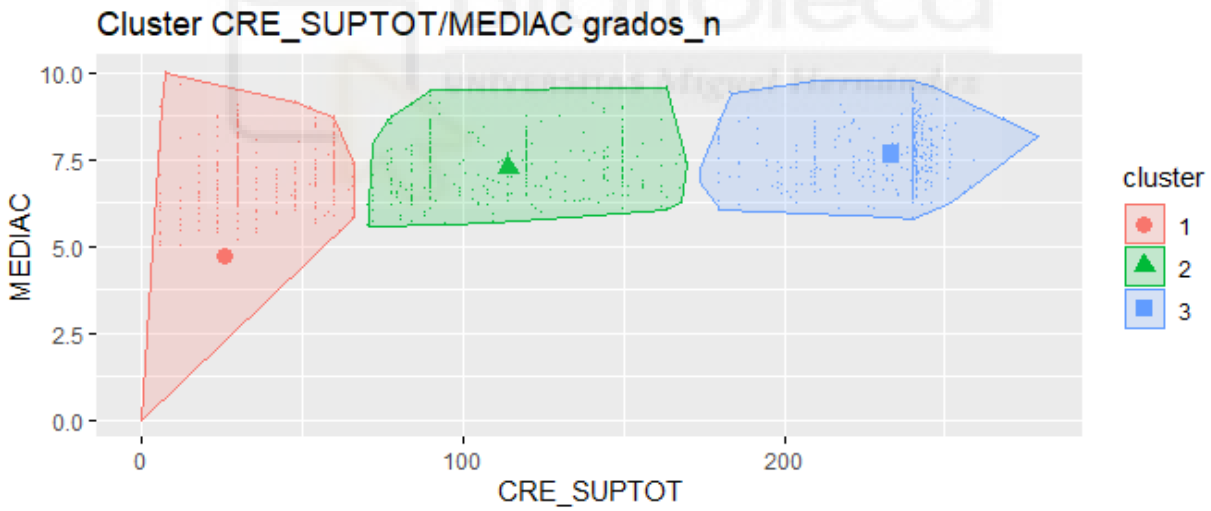
No hay clusters para 3 y 4 grupos, ya que uno de los grupos contiene menos del 5% de los individuos.

Figura 49: Cluster de los grados de la Facultad de Bellas Artes (2 grupos)



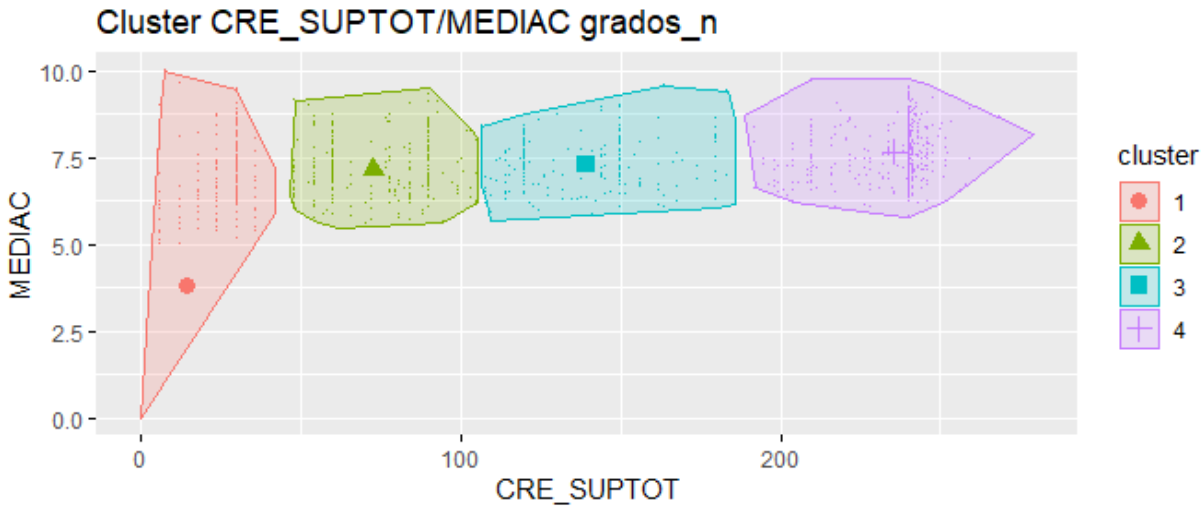
Tamaño grupos: 682, 613

Figura 50: Cluster de los grados de la Facultad de Bellas Artes (3 grupos)



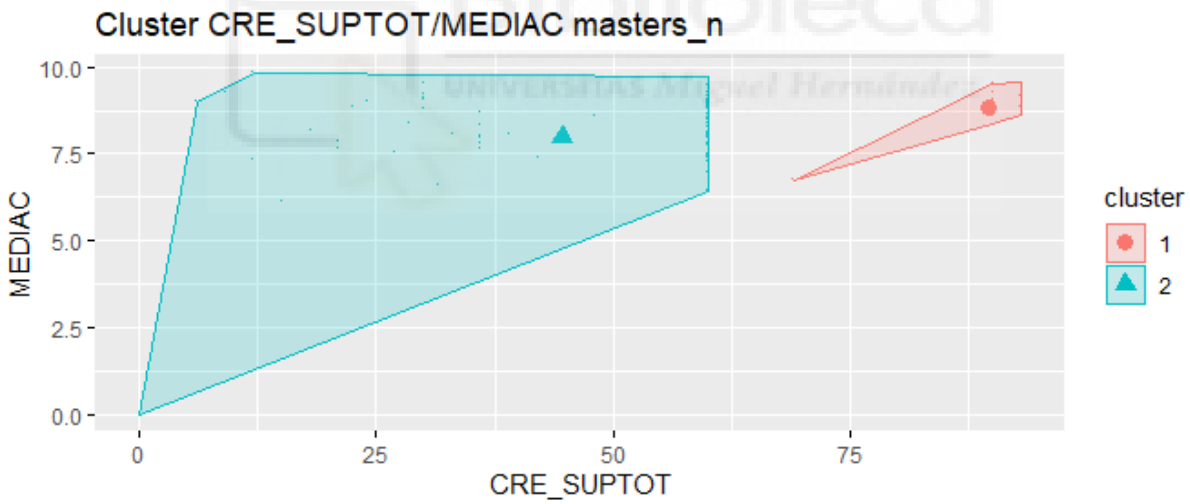
Tamaño grupos: 459, 311, 525

Figura 51: Cluster de los grados de la Facultad de Bellas Artes (4 grupos)



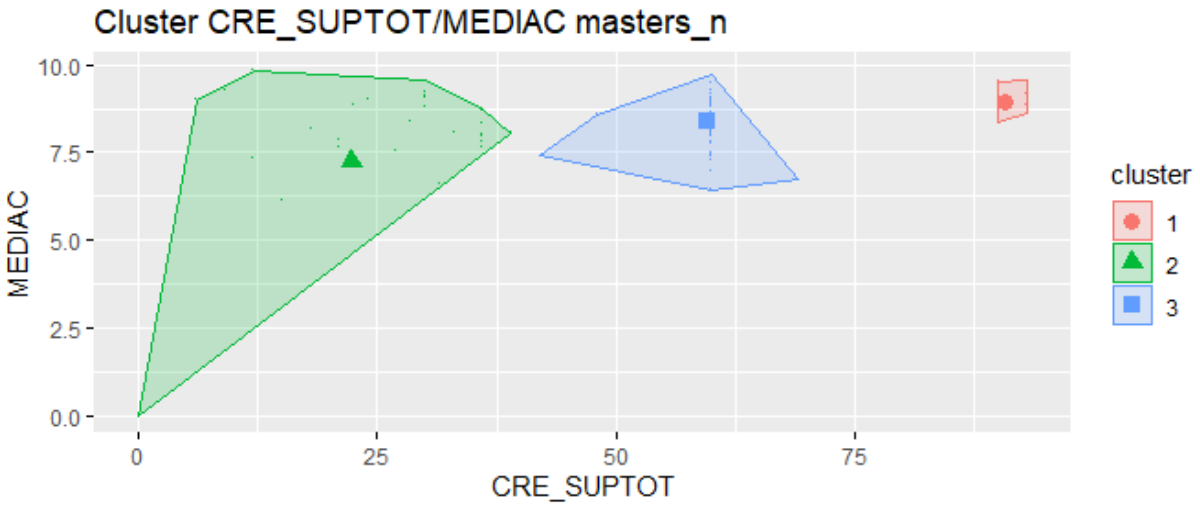
Tamaño grupos: 337, 255, 198, 505

Figura 52: Cluster de los másteres de la Facultad de Bellas Artes (2 grupos)



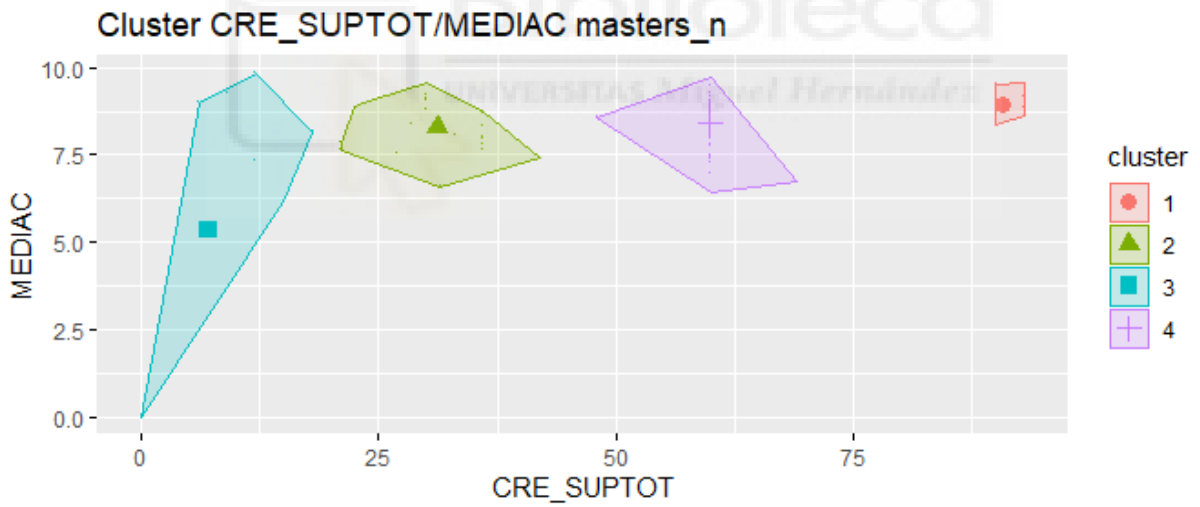
Tamaño grupos: 18, 78

Figura 53: Cluster de los másteres de la Facultad de Bellas Artes (3 grupos)



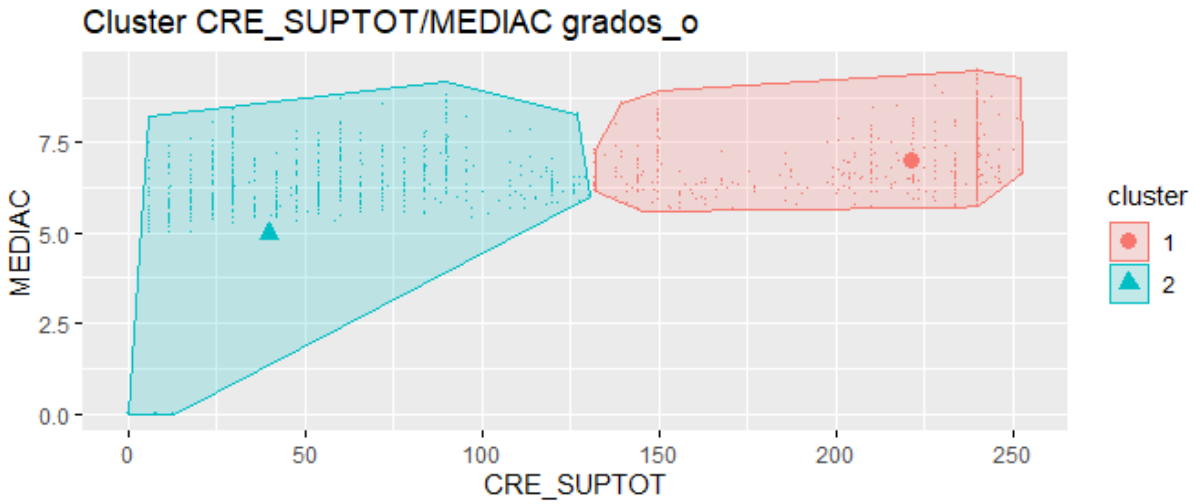
Tamaño grupos: 17, 31, 48

Figura 54: Cluster de los másteres de la Facultad de Bellas Artes (4 grupos)



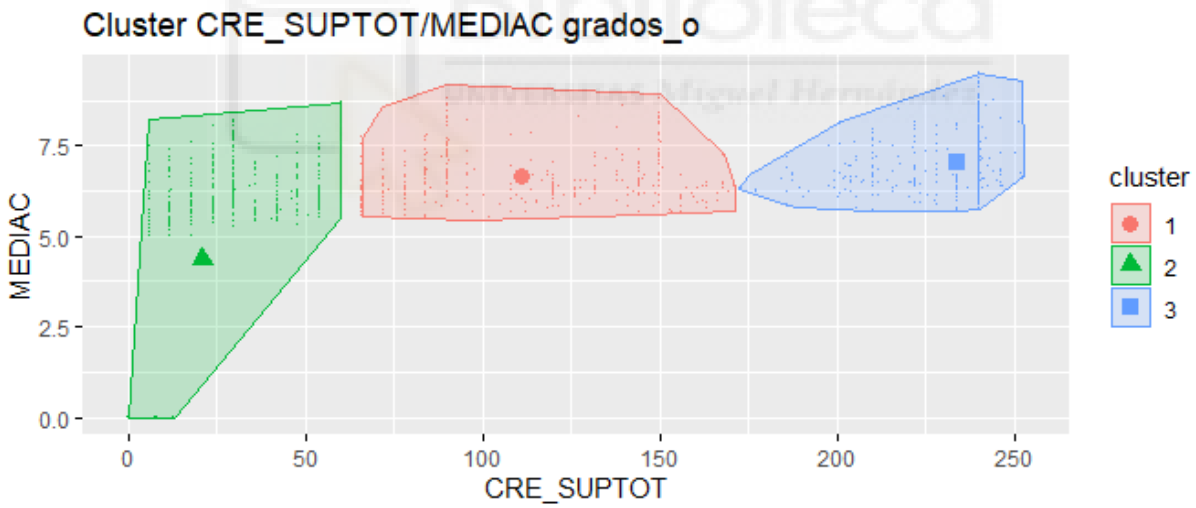
Tamaño grupos: 17, 21, 11, 47

Figura 55: Cluster de los grados de la Facultad de Ciencias Experimentales (2 grupos)



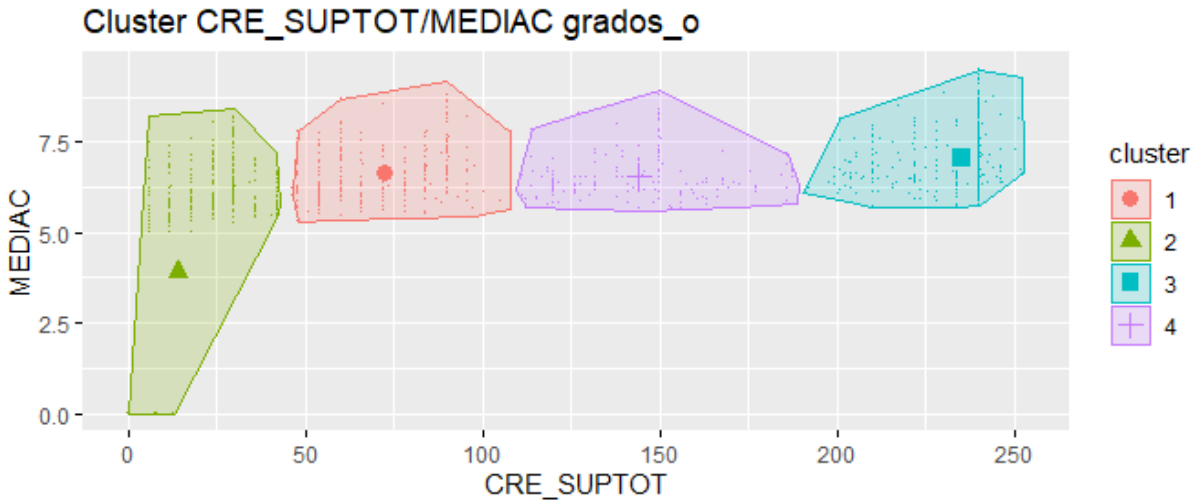
Tamaño grupos: 713, 768

Figura 56: Cluster de los grados de la Facultad de Ciencias Experimentales (3 grupos)



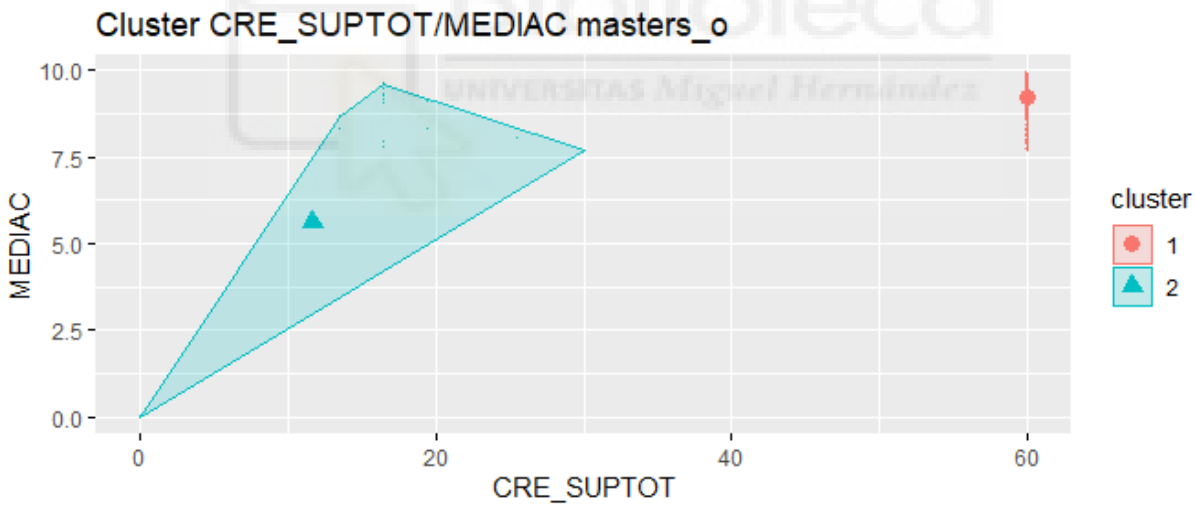
Tamaño grupos: 310, 562, 609

Figura 57: Cluster de los grados de la Facultad de Ciencias Experimentales (4 grupos)



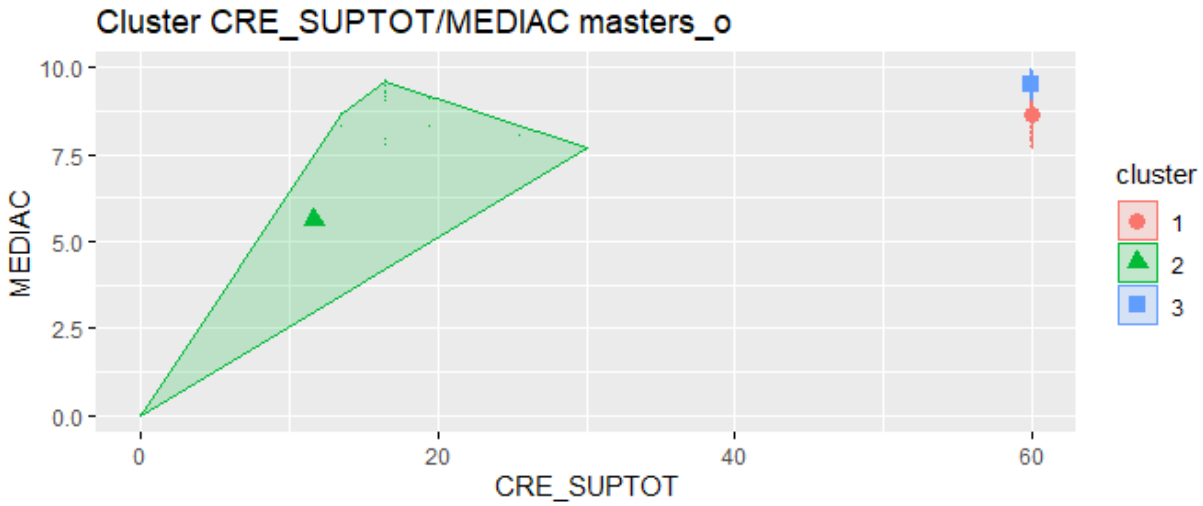
Tamaño grupos: 253, 468, 593, 167

Figura 58: Cluster de los másteres de la Facultad de Ciencias Experimentales (2 grupos)



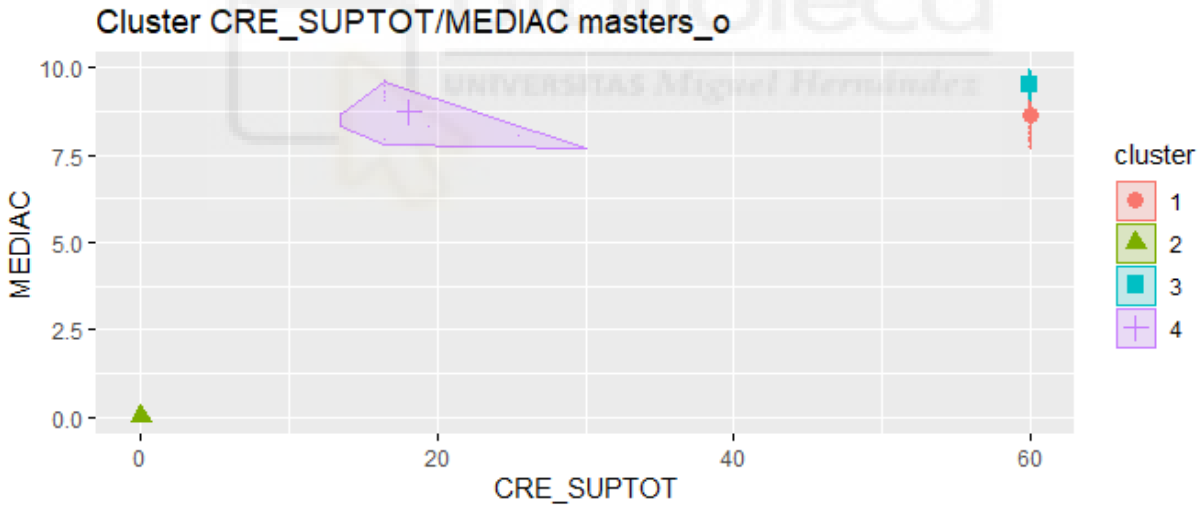
Tamaño grupos: 158, 28

Figura 59: Cluster de los másteres de la Facultad de Ciencias Experimentales (3 grupos)



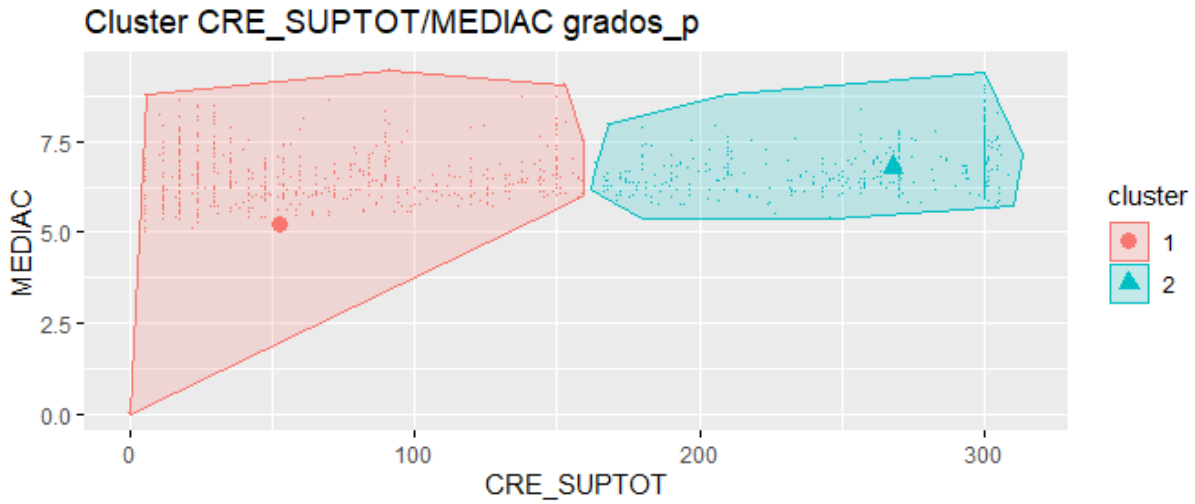
Tamaño grupos: 47, 28, 111

Figura 60: Cluster de los másteres de la Facultad de Ciencias Experimentales (4 grupos)



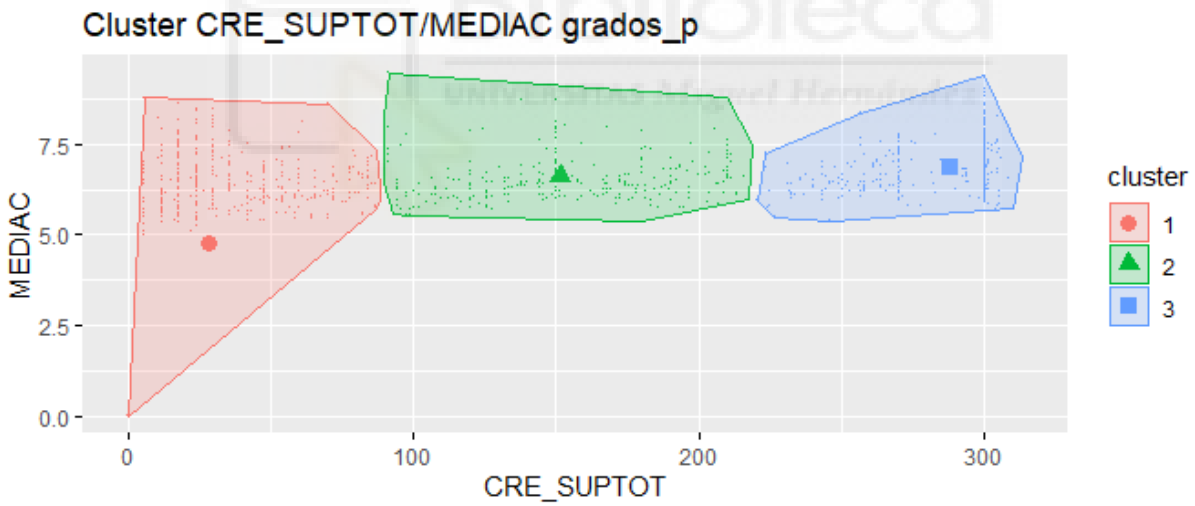
Tamaño grupos: 47, 10, 111, 18

Figura 61: Cluster de los grados de la Facultad de Farmacia (2 grupos)



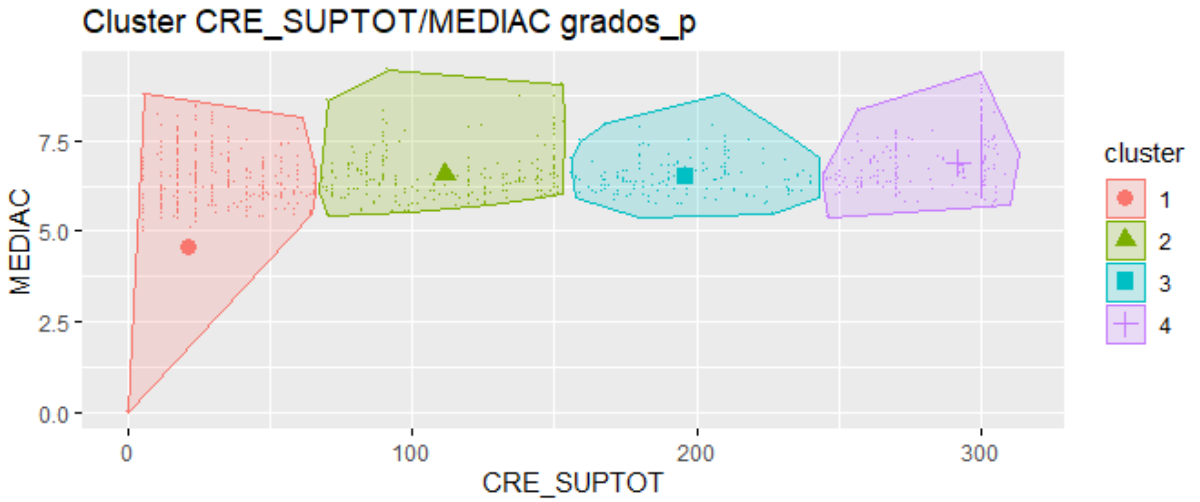
Tamaño grupos: 683, 576

Figura 62: Cluster de los grados de la Facultad de Farmacia (3 grupos)



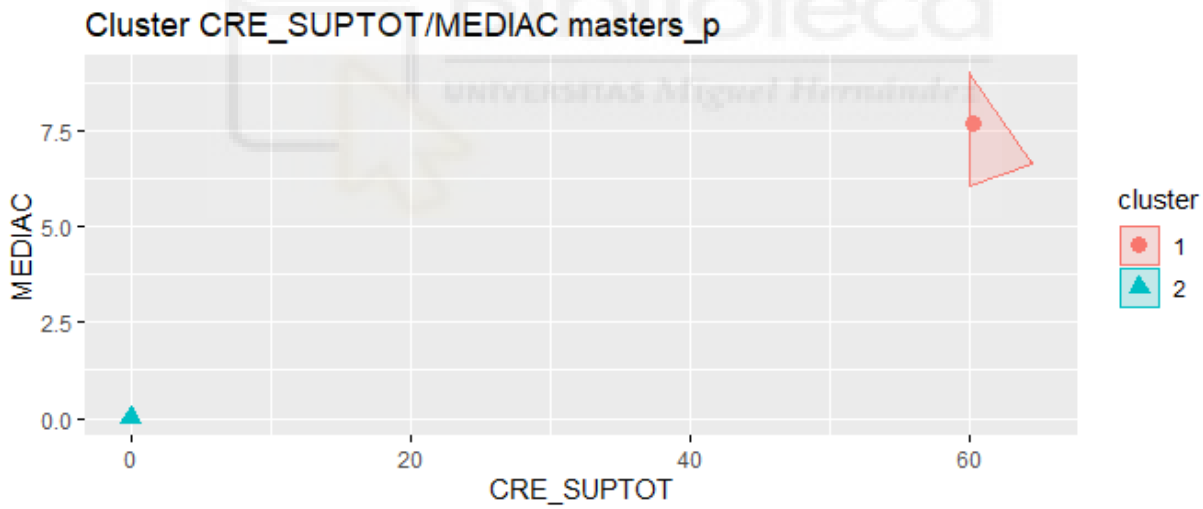
Tamaño grupos: 513, 287, 459

Figura 63: Cluster de los grados de la Facultad de Farmacia (4 grupos)



Tamaño grupos: 452, 223, 155, 429

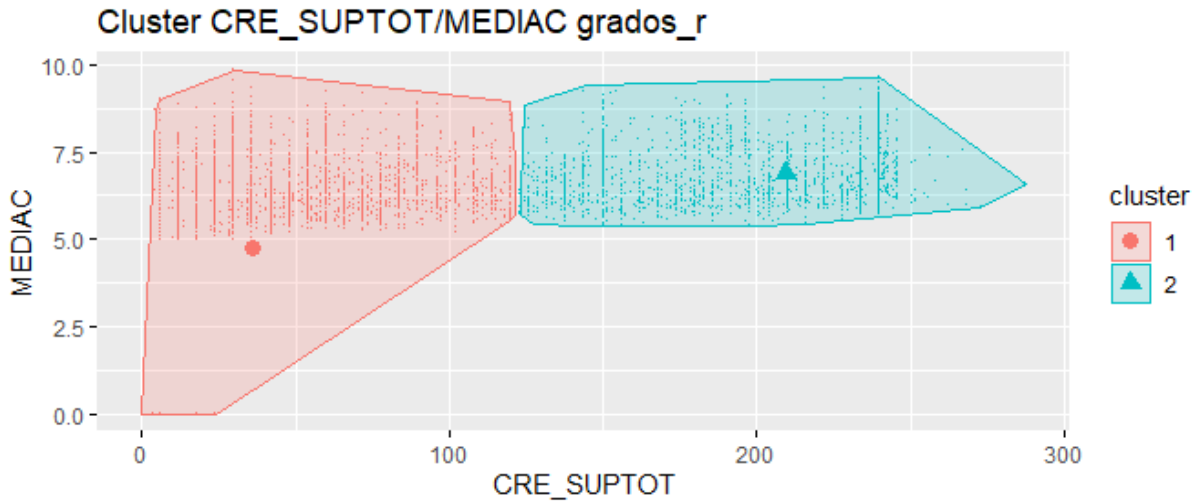
Figura 64: Cluster de los másteres de la Facultad de Farmacia (2 grupos)



Tamaño grupos: 23, 5

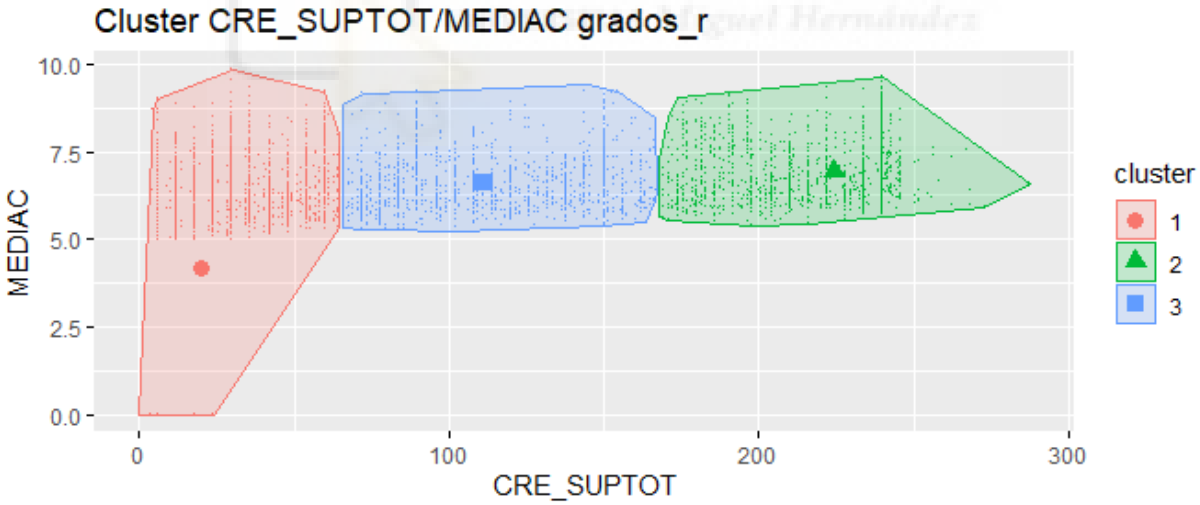
No hay clusters para 3 y 4 grupos, ya que uno de los grupos contiene menos del 5% de los individuos.

Figura 65: Cluster de los grados de la Facultad de Ciencias Sociales y Jurídicas de Elche (2 grupos)



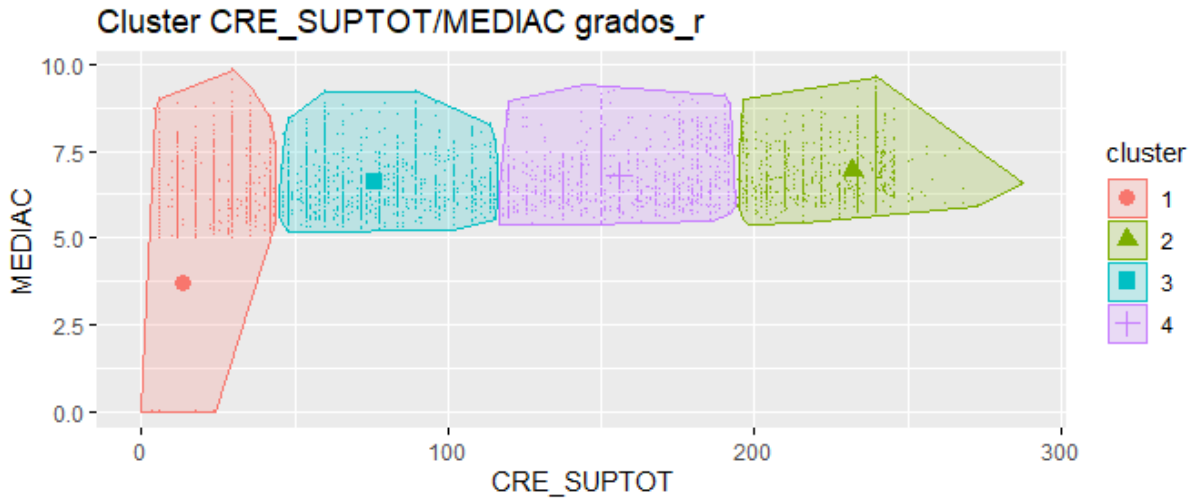
Tamaño grupos: 3729, 2625

Figura 66: Cluster de los grados de la Facultad de Ciencias Sociales y Jurídicas de Elche (3 grupos)



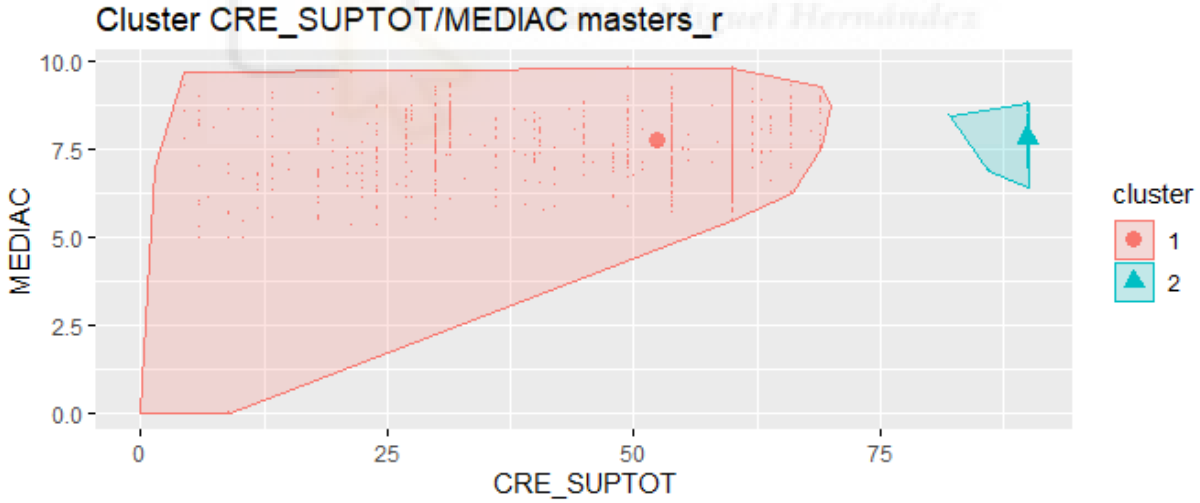
Tamaño grupos: 2882, 2133, 1339

Figura 67: Cluster de los grados de la Facultad de Ciencias Sociales y Jurídicas de Elche (4 grupos)



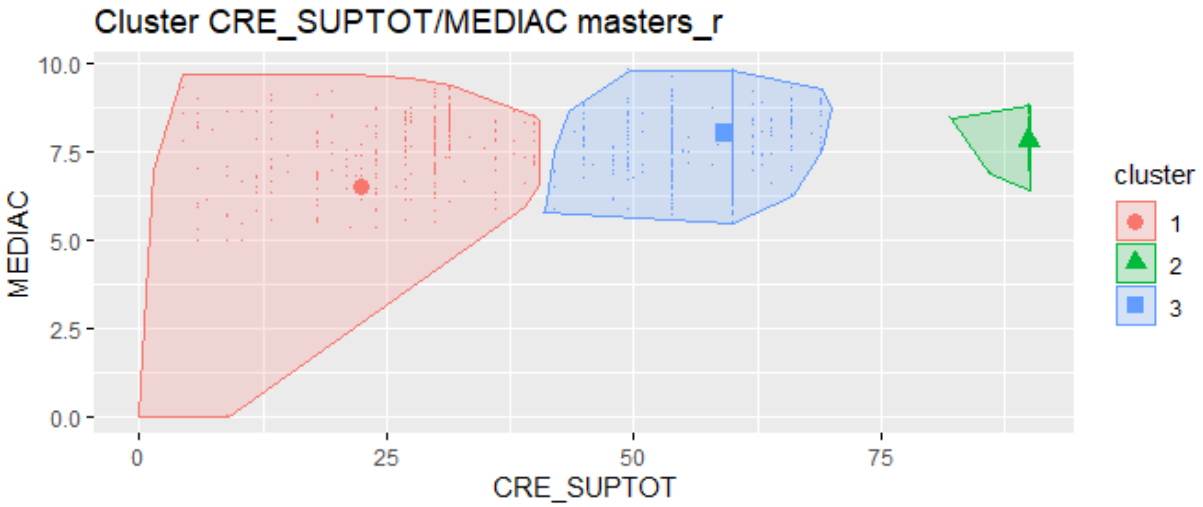
Tamaño grupos: 2416, 1814, 1243, 881

Figura 68: Cluster de los másteres de la Facultad de Ciencias Sociales y Jurídicas de Elche (2 grupos)



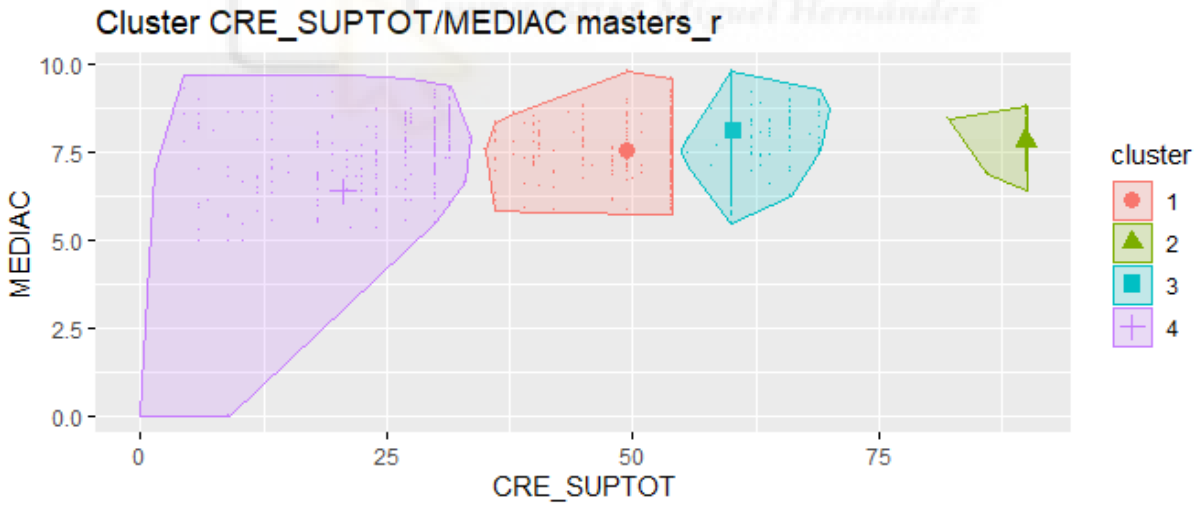
Tamaño grupos: 2052, 207

Figura 69: Cluster de los másteres de la Facultad de Ciencias Sociales y Jurídicas de Elche (3 grupos)



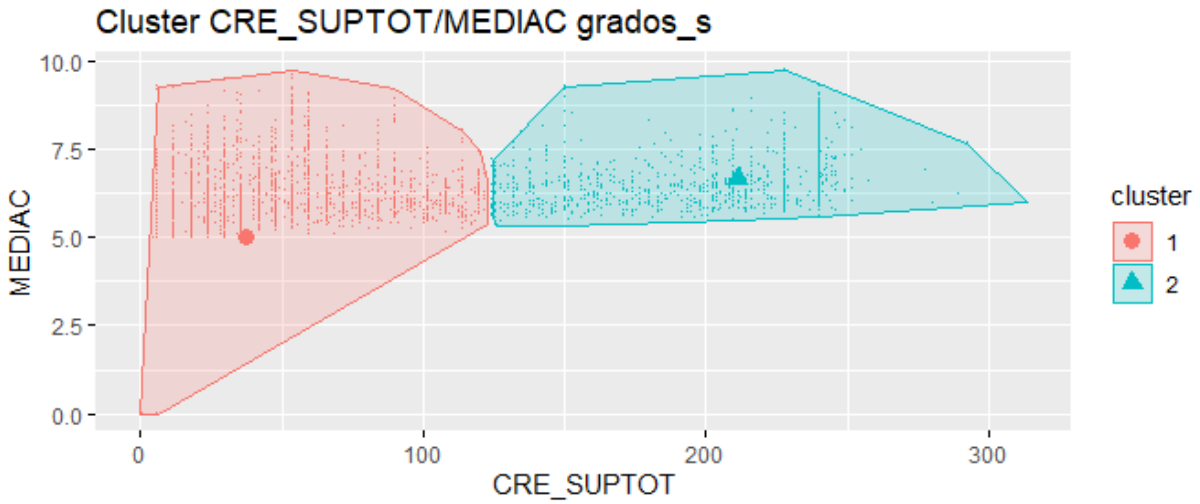
Tamaño grupos: 379, 207, 1673

Figura 70: Cluster de los másteres de la Facultad de Ciencias Sociales y Jurídicas de Elche (4 grupos)



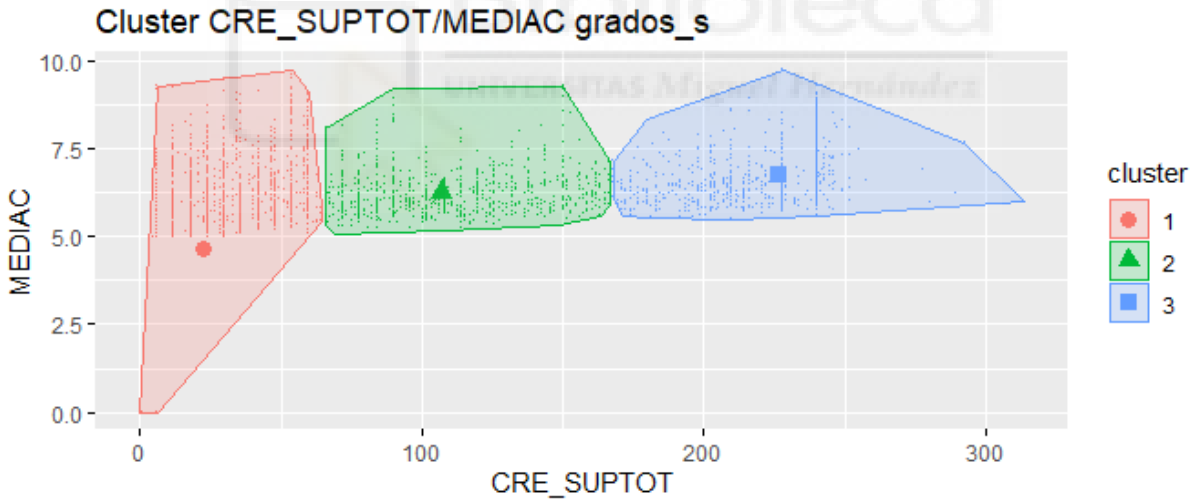
Tamaño grupos: 246, 207, 1467, 339

Figura 71: Cluster de los grados de la Escuela Politécnica de Elche (2 grupos)



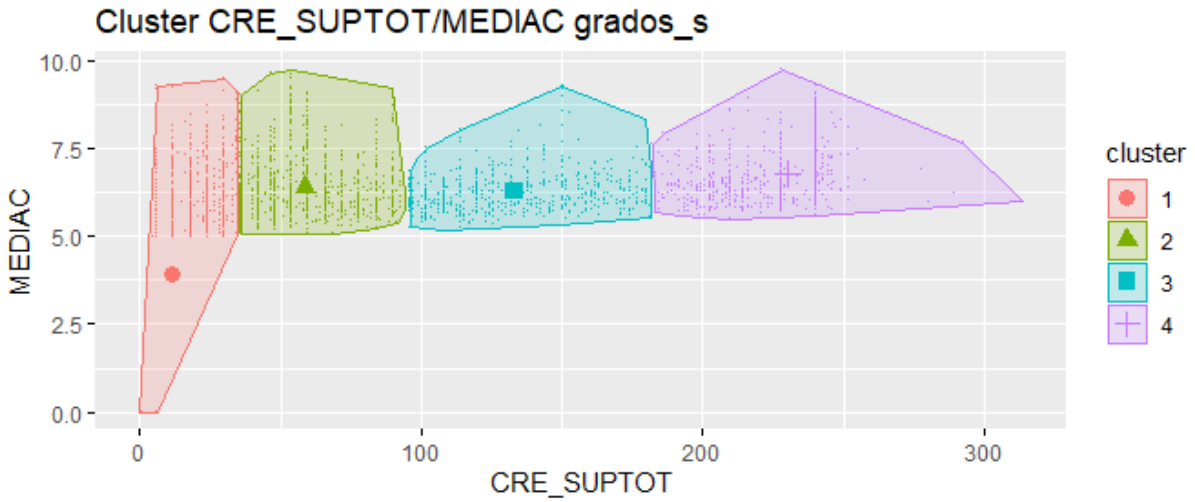
Tamaño grupos: 2537, 1431

Figura 72: Cluster de los grados de la Escuela Politécnica de Elche (3 grupos)



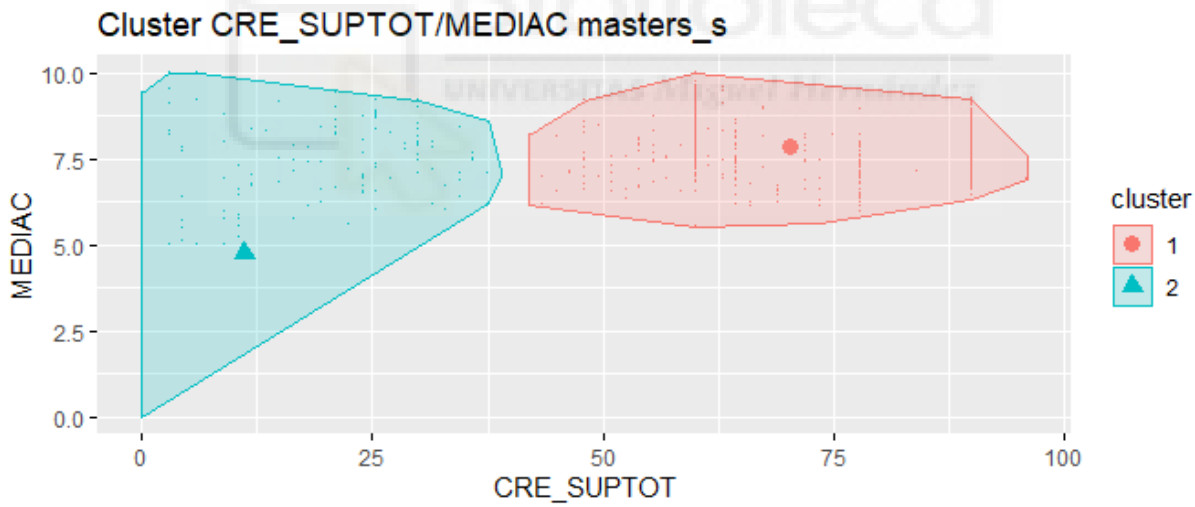
Tamaño grupos: 1976, 827, 1165

Figura 73: Cluster de los grados de la Escuela Politécnica de Elche (4 grupos)



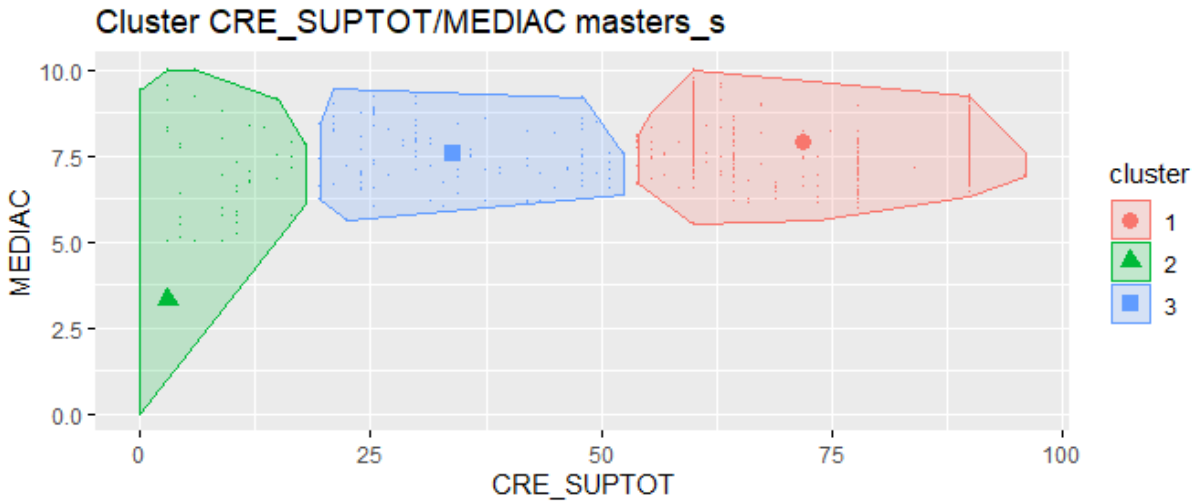
Tamaño grupos: 1387, 922, 565, 1094

Figura 74: Cluster de los másteres de la Escuela Politécnica de Elche (2 grupos)



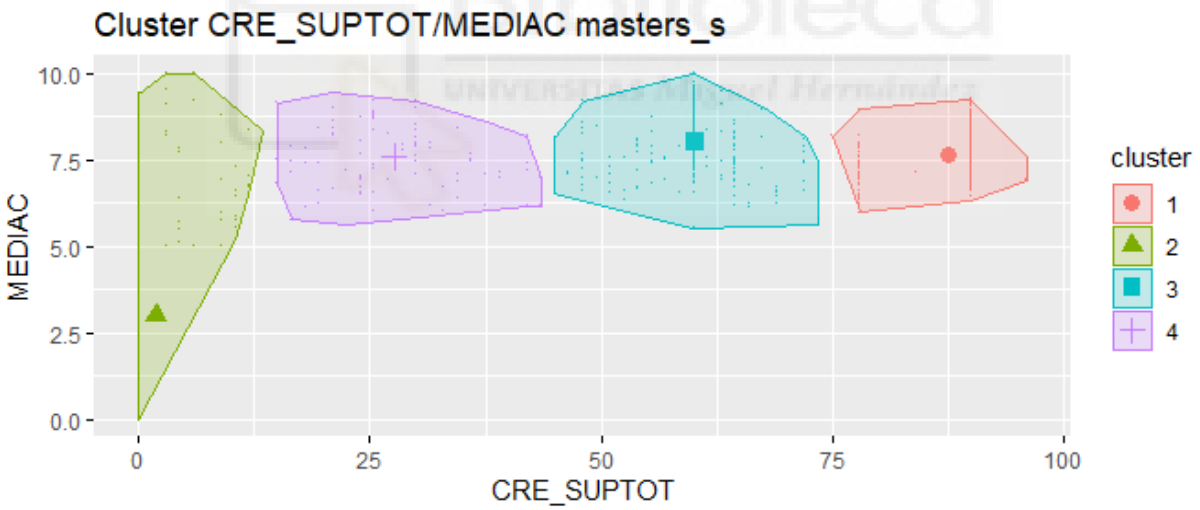
Tamaño grupos: 565, 224

Figura 75: Cluster de los másteres de la Escuela Politécnica de Elche (3 grupos)



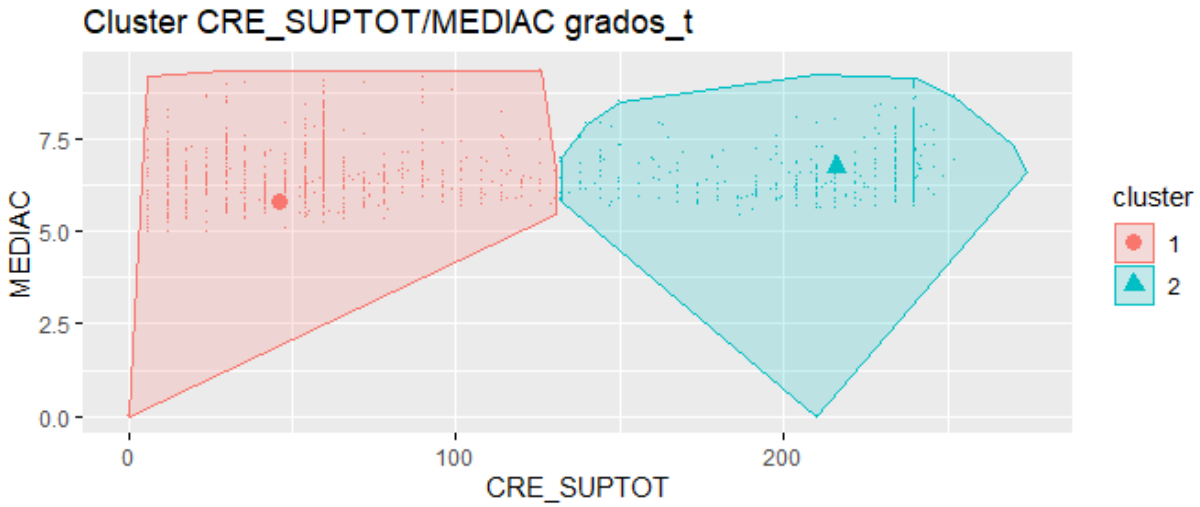
Tamaño grupos: 532, 151, 106

Figura 76: Cluster de los másteres de la Escuela Politécnica de Elche (4 grupos)



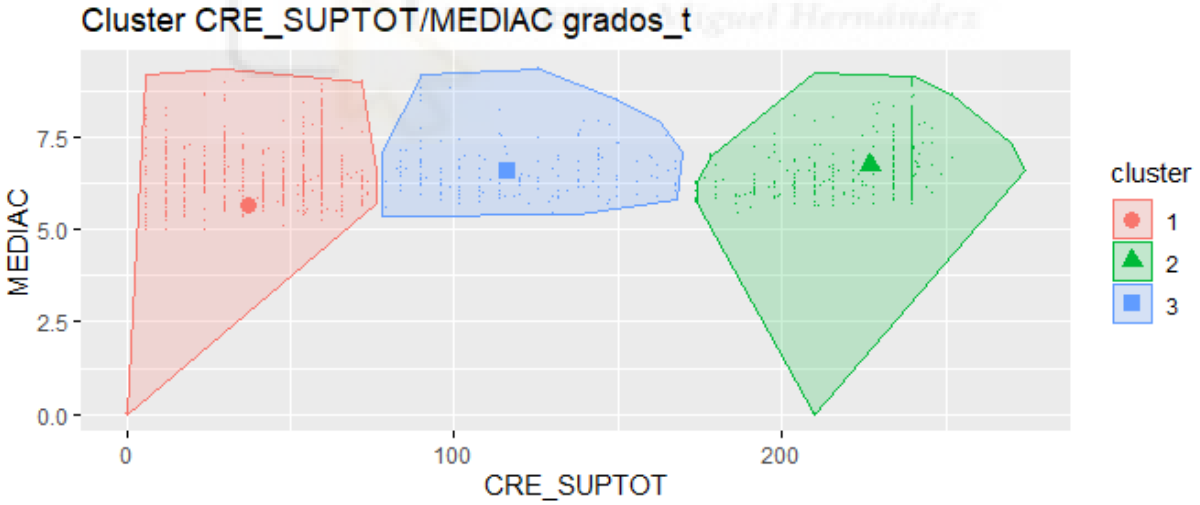
Tamaño grupos: 215, 140, 343, 91

Figura 77: Cluster de los grados de la Facultad de Ciencias Sociales y Jurídicas de Orihuela (2 grupos)



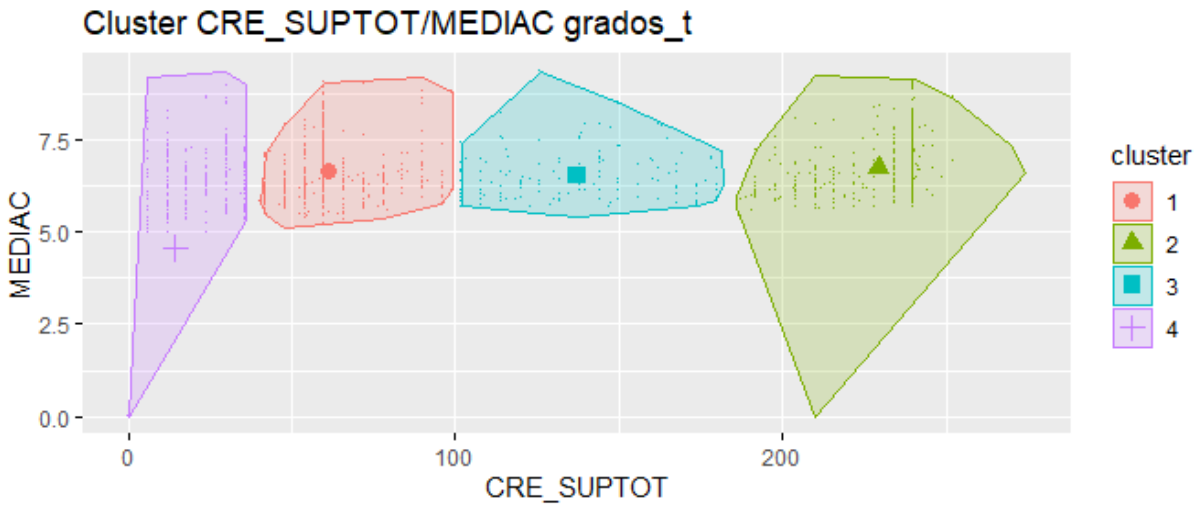
Tamaño grupos: 884, 450

Figura 78: Cluster de los grados de la Facultad de Ciencias Sociales y Jurídicas de Orihuela (3 grupos)



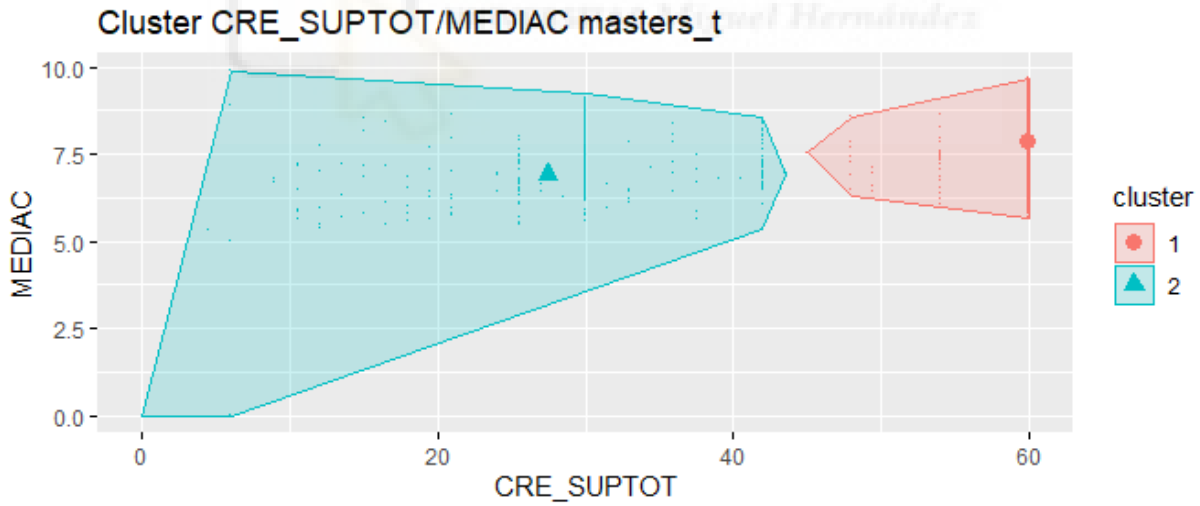
Tamaño grupos: 761, 387, 186

Figura 79: Cluster de los grados de la Facultad de Ciencias Sociales y Jurídicas de Orihuela (4 grupos)



Tamaño grupos: 467, 370, 140, 357

Figura 80: Cluster de los másteres de la Facultad de Ciencias Sociales y Jurídicas de Orihuela (2 grupos)



Tamaño grupos: 3816, 534

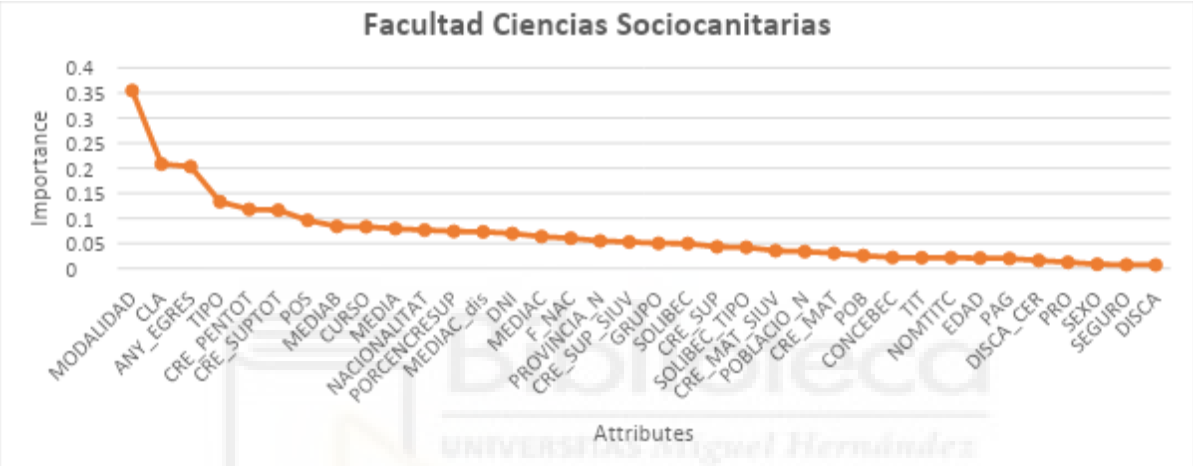
No hay clusters para 3 y 4 grupos, ya que uno de los grupos contiene menos del 5% de los individuos.

9.2 ANÁLISIS PREDICTIVOS

A continuación, se analizan como variables objetivo el abandono y después el rendimiento (nota media). Para cada facultad, primero se presentan los gráficos que indican el peso de las variables más importantes, para poder modelar la variable objetivo y a continuación se muestran los árboles de clasificación para cada una de estas variables.

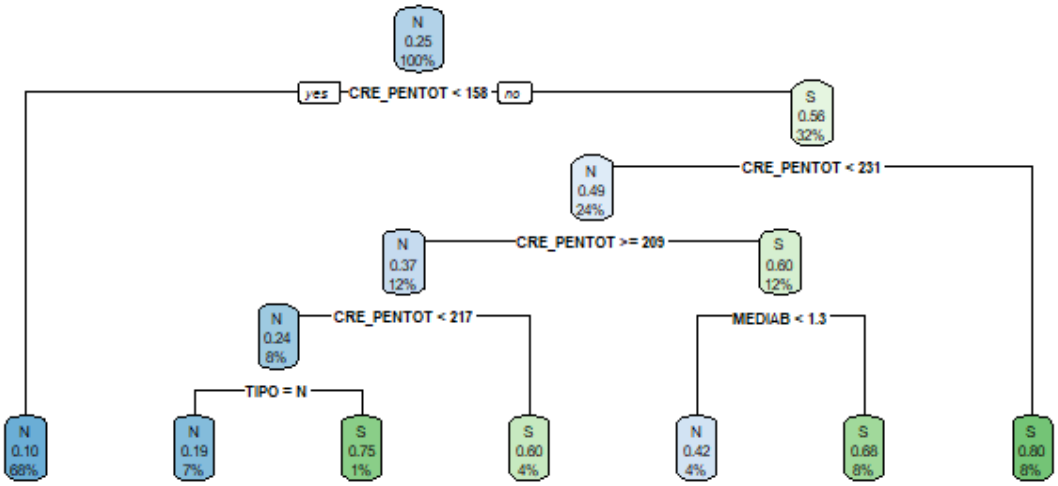
9.2.1 Abandono

Figura 81: Peso de las variables más importantes en la Facultad de Ciencias Sociosanitarias (GRADOS)



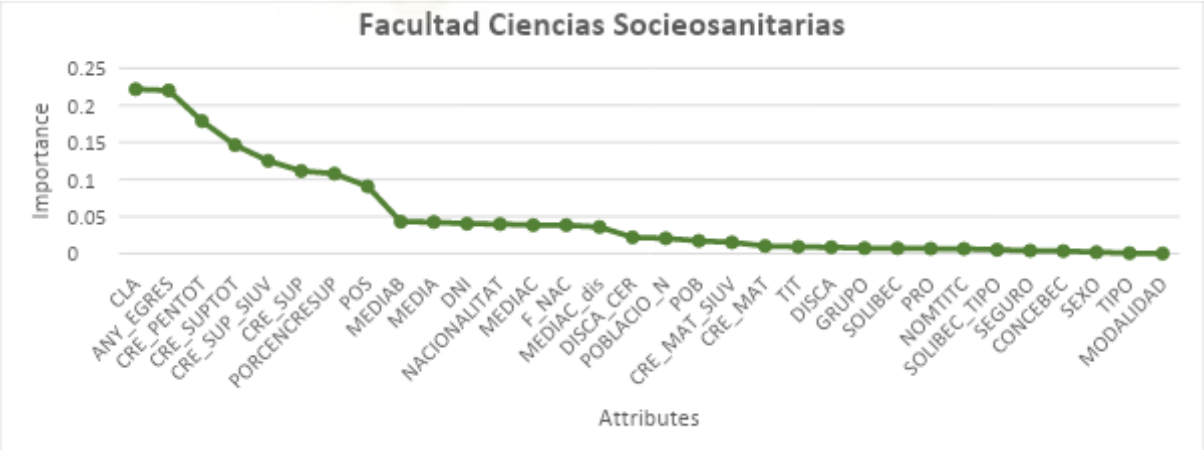
El nivel de presencialidad del grado, la clase de matrícula y el año de finalización de los estudios son las variables que más influyen en el abandono de los grados por parte de los estudiantes de esta facultad.

Figura 82: Árbol de clasificación de la Facultad de Ciencias Sociosanitarias (GRADOS)



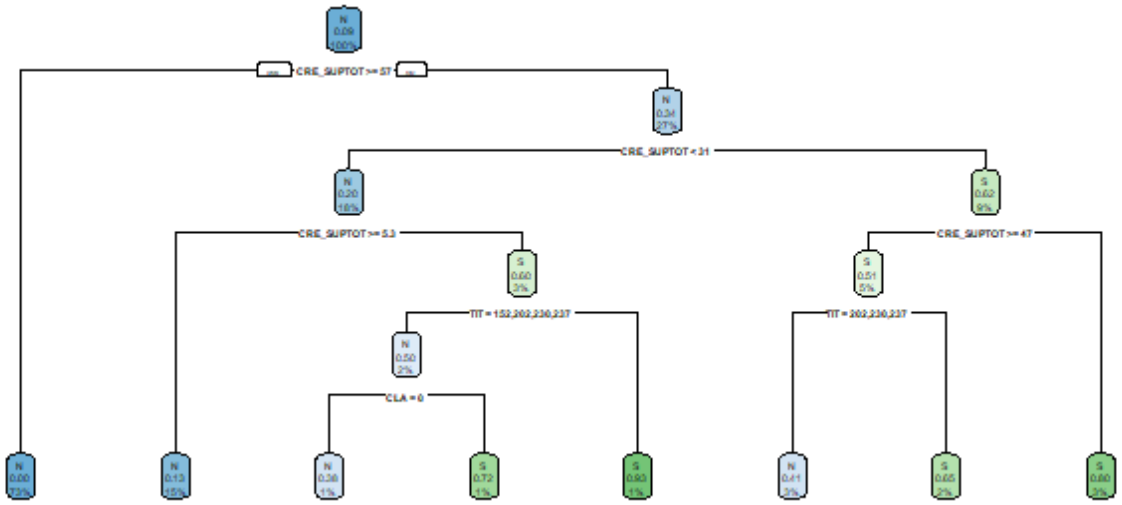
Como resultado del modelo con Accuracy=0.8326898, por la rama de la derecha, podemos concluir que los estudiantes que tienen más de 231 créditos pendientes totales tienen un 80% (0.80) de probabilidad de abandonar el grado.

Figura 83: Peso de las variables más importantes de la Facultad de Ciencias Sociosanitarias (MÁSTERES)



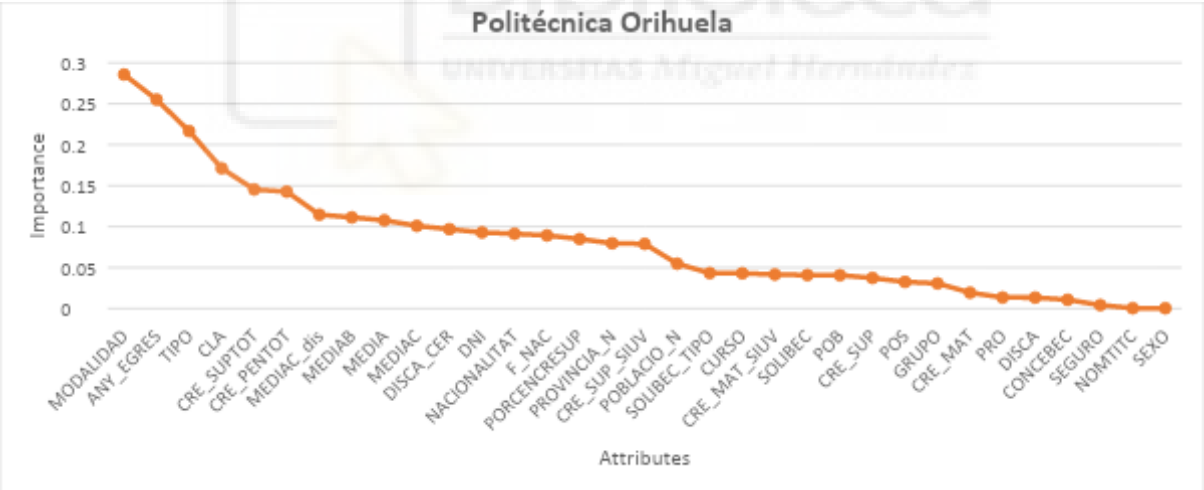
La clase de matrícula y el año de finalización son algunas de las variables que más influyen en el abandono de los másteres por parte de los estudiantes de la Facultad de Ciencias Sociosanitarias.

Figura 84: Árbol de clasificación de la Facultad de Ciencias Sociosanitarias (MÁSTERES)



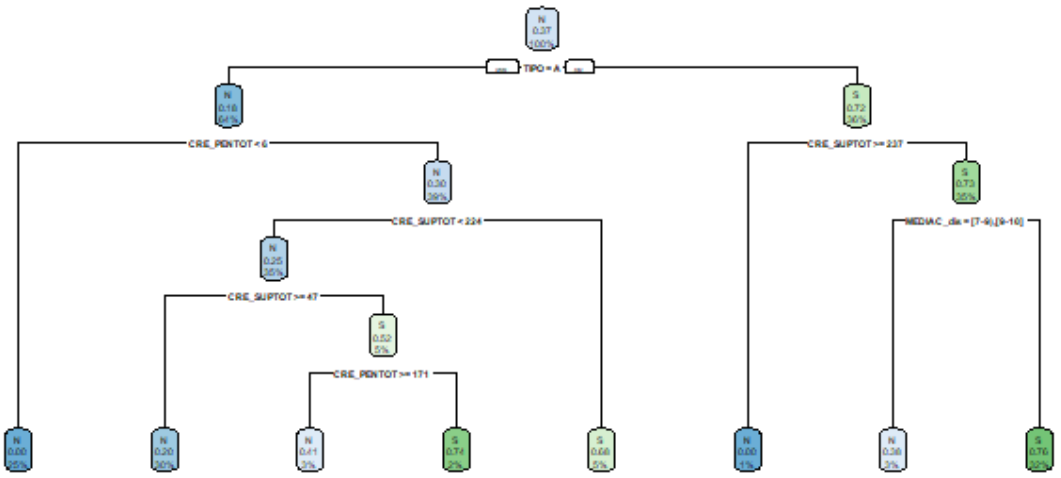
Accuracy=0.9390018

Figura 85: Peso de las variables más importantes en la Politécnica de Orihuela (GRADOS)



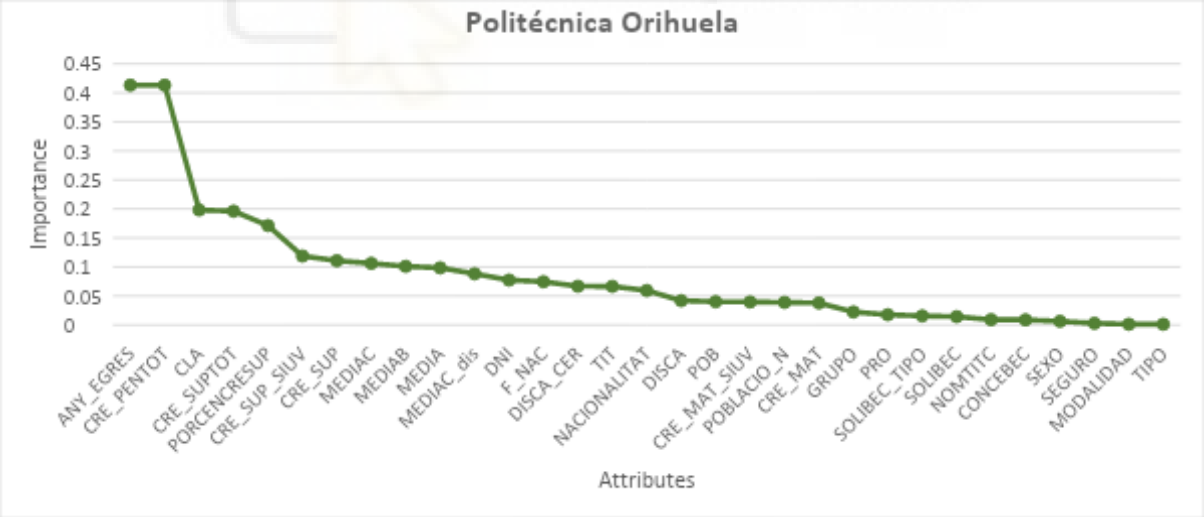
El nivel de presencialidad del grado, el año de finalización de los estudios, si es estudiante de nuevo ingreso o no y la clase de matrícula son algunas de las variables que más influyen en el abandono de los grados por parte de los estudiantes de la Universidad Politécnica de Orihuela.

Figura 86: Árbol de clasificación de la Politécnica de Orihuela (GRADOS)



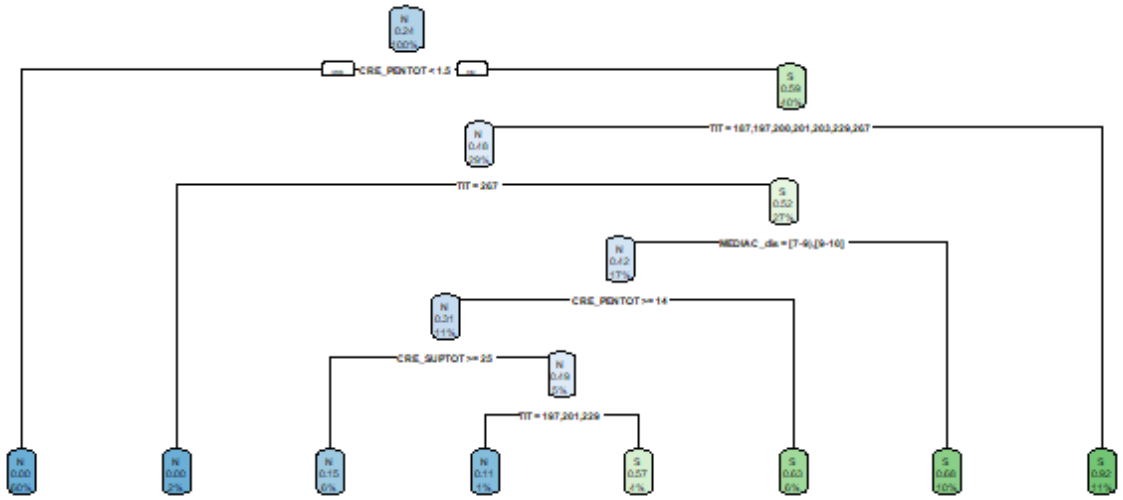
Accuracy=0.8188406

Figura 87: Peso de las variables más importantes en la Politécnica de Orihuela (MÁSTERES)



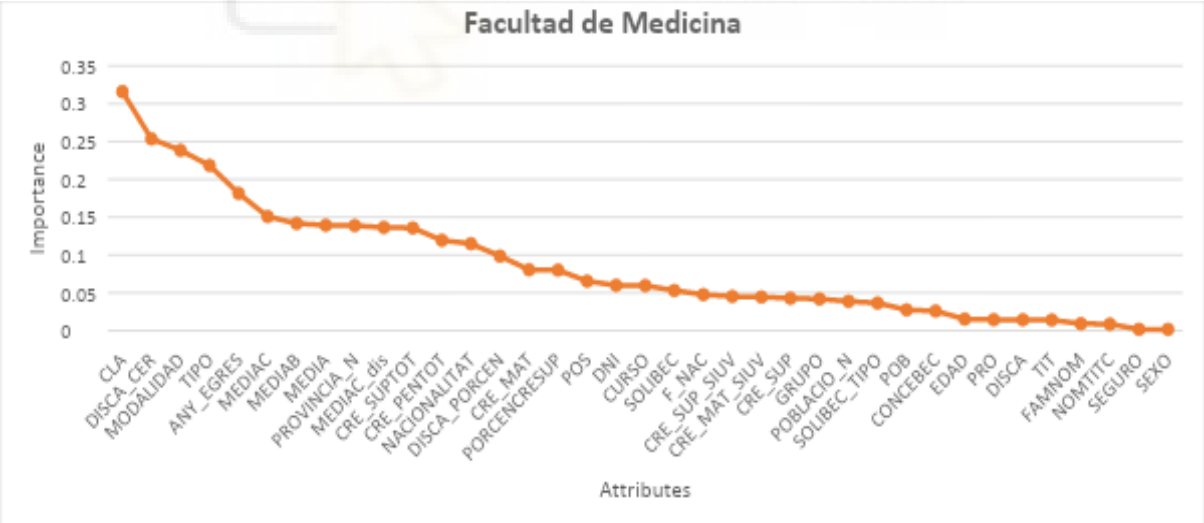
El año de finalización de los estudios y los créditos pendientes totales son las variables que más influyen en el abandono de los másteres por parte de los estudiantes de la Escuela Politécnica de Orihuela.

Figura 88: Árbol de clasificación de la Politécnica de Orihuela (MÁSTERES)



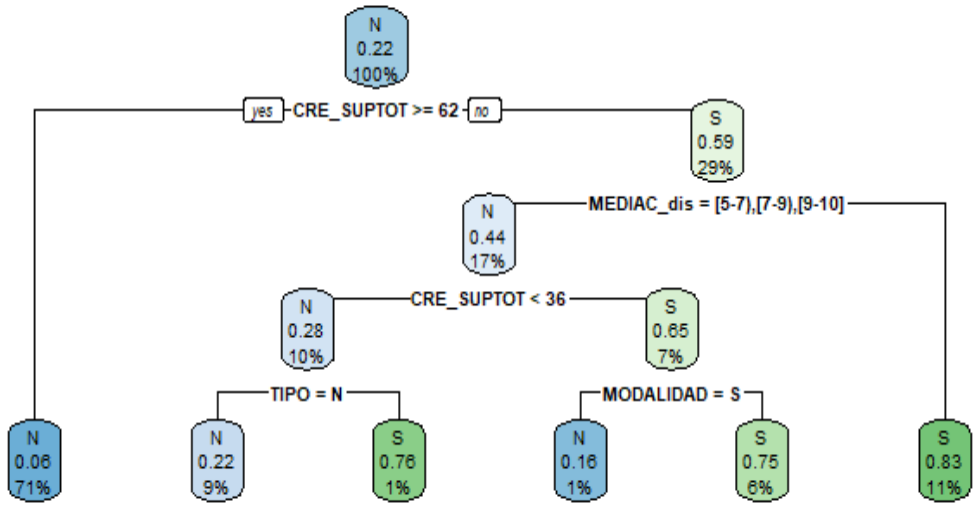
Accuracy=0.9262295

Figura 89: Peso de las variables más importantes en la Facultad de Medicina (GRADOS)



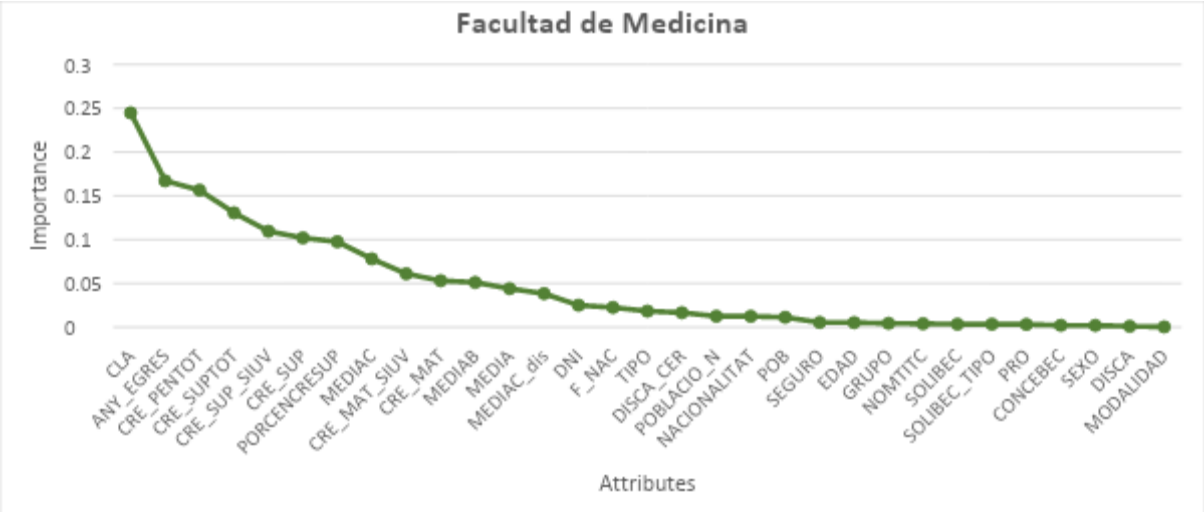
La clase de matrícula, el certificado de discapacidad, el nivel de presencialidad del grado, si es estudiante de nuevo ingreso o no y el año de finalización de los estudios son las variables que más influyen en el abandono de los grados por parte de los estudiantes de la Facultad de Medicina.

Figura 90: Árbol de clasificación de la Facultad de Medicina (GRADOS)



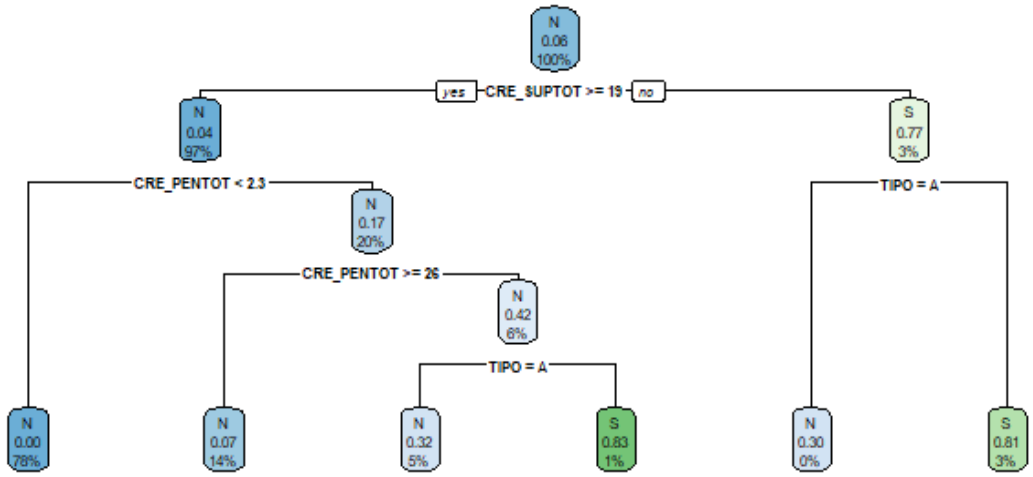
Como resultado del modelo con Accuracy 0.9093825, por la rama más a la derecha, podemos concluir que los estudiantes que tienen menos de 62 créditos superados totales y tienen una media entre 0 y 5 este último sin incluir, tienen un 83% de probabilidad de abandonar el grado.

Figura 91: Peso de las variables más importantes en la Facultad de Medicina (MÁSTERES)



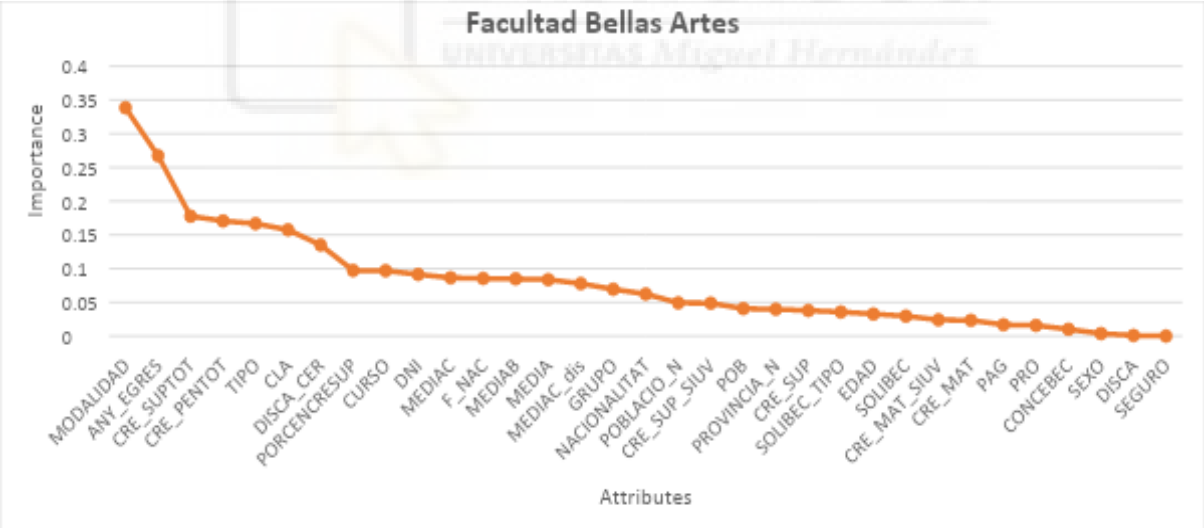
La clase de matrícula, el año de finalización de los estudios y los créditos pendientes totales son las variables que más influyen en el abandono de los másteres por parte de los estudiantes de la Facultad de Medicina, al igual que ocurre en la Politécnica de Orihuela.

Figura 92: Árbol de clasificación de la Facultad de Medicina (MÁSTERES)



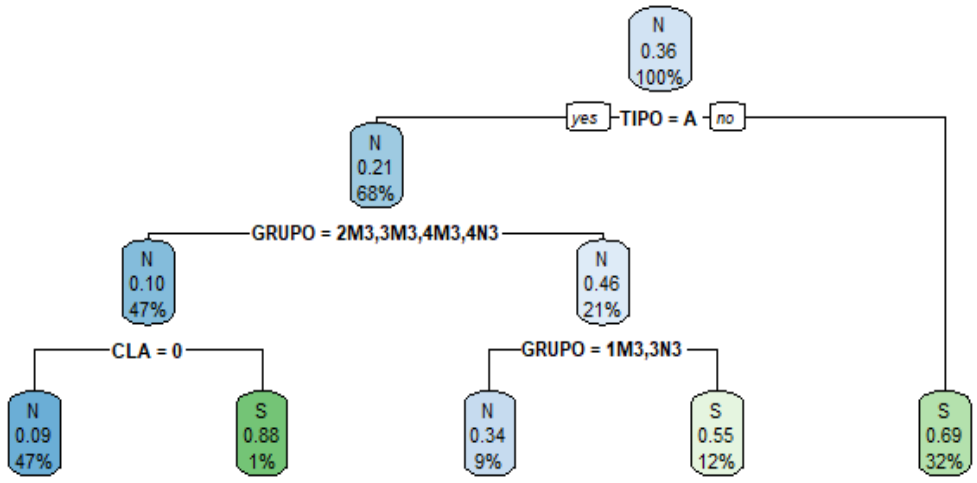
Accuracy=0.9749104

Figura 93: Peso de las variables más importantes en la Facultad de Bellas Artes (GRADOS)



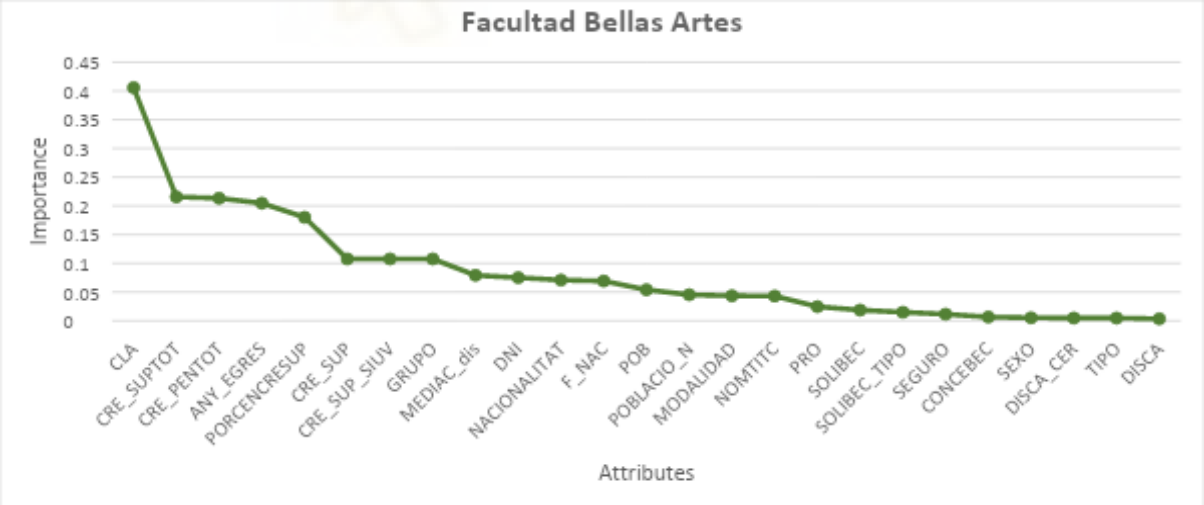
El nivel de presencialidad del grado y el año de finalización de los estudios son las variables que más influyen en el abandono de los grados por parte de los estudiantes de la Facultad de Bellas Artes.

Figura 94: Árbol de clasificación de la Facultad de Bellas Artes (GRADOS)



Como resultado del modelo con Accuracy 0.7592593, por la rama más a la derecha, podemos concluir que los estudiantes que son de nuevo ingreso tienen un 69% de probabilidad de abandonar el grado.

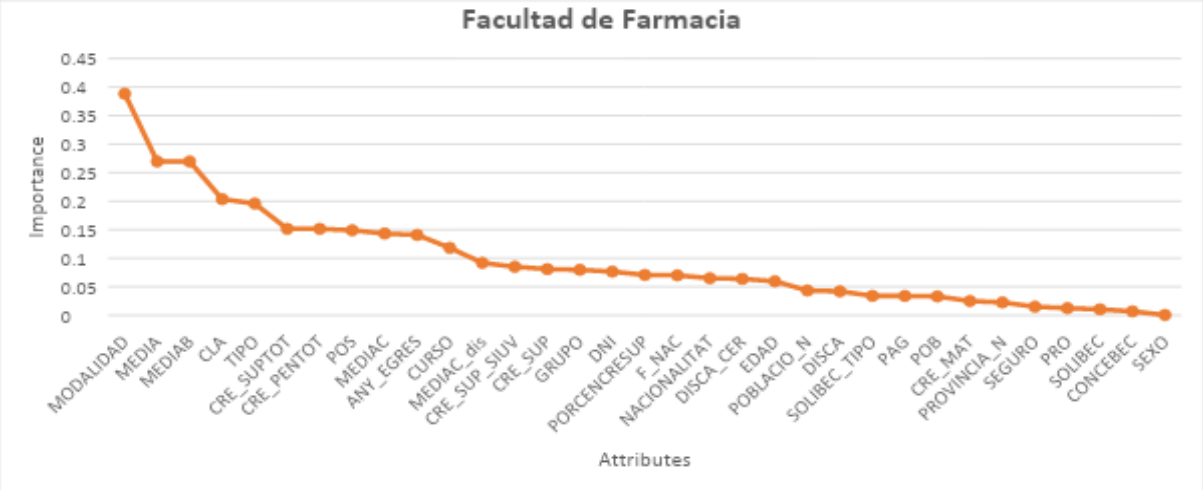
Figura 95: Peso de las variables más importantes en la Facultad de Bellas Artes (MÁSTERES)



La clase de matrícula, los créditos superados y pendientes totales y el año de finalización de los estudios son algunas de las variables que más influyen en el abandono de los másteres por parte de los estudiantes de la Facultad de Bellas Artes.

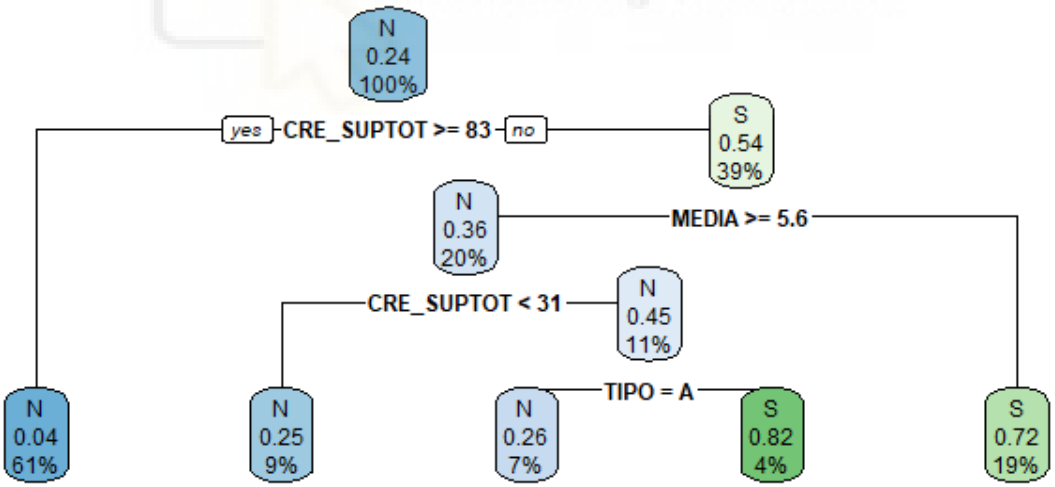
No ha sido posible realizar un árbol de clasificación de los másteres en la Facultad de Bellas Artes porque hay datos insuficientes para crear un modelo.

Figura 96: Peso de las variables más importantes en la Facultad de Farmacia (GRADOS)



El nivel de presencialidad del grado, la nota media, la clase de matrícula, y si es estudiante de nuevo ingreso o no son algunas de las variables que más influyen en el abandono de los grados por parte de los estudiantes de la Facultad de Farmacia.

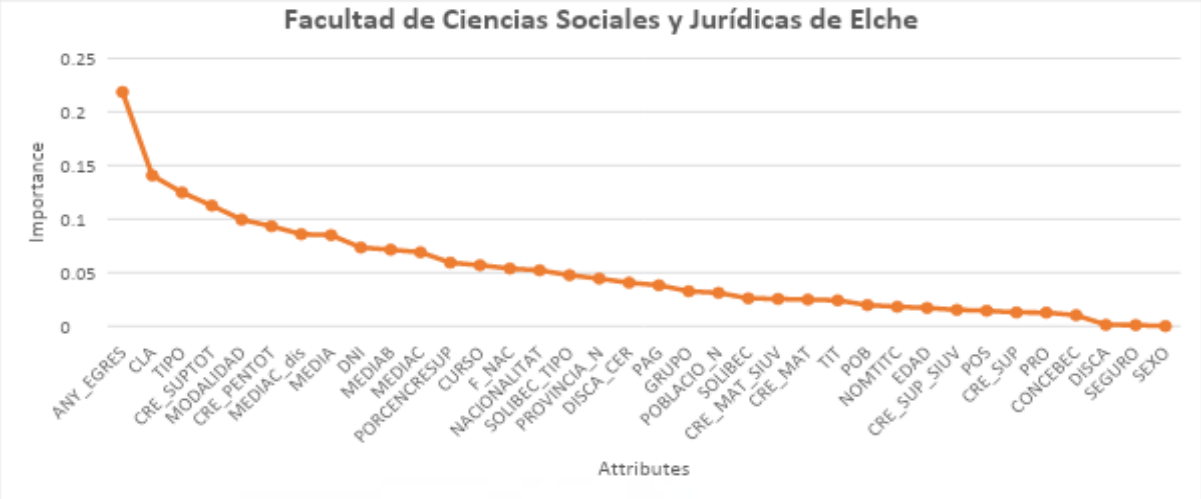
Figura 97: Árbol de clasificación de la Facultad de Farmacia (GRADOS)



Como resultado del modelo con Accuracy 0.8412698, por la rama más a la derecha, podemos concluir que los estudiantes que tienen menos de 83 créditos superados totales y que además su media, según las siguientes equivalencias SUSPENSO 2.5, APROBADO 5.5, NOTABLE 7.5, SOBRESALIENTE 9 y MATRÍCULA DE HONOR 10, es menor de 5.6, tienen un 72% de probabilidad de abandonar el grado.

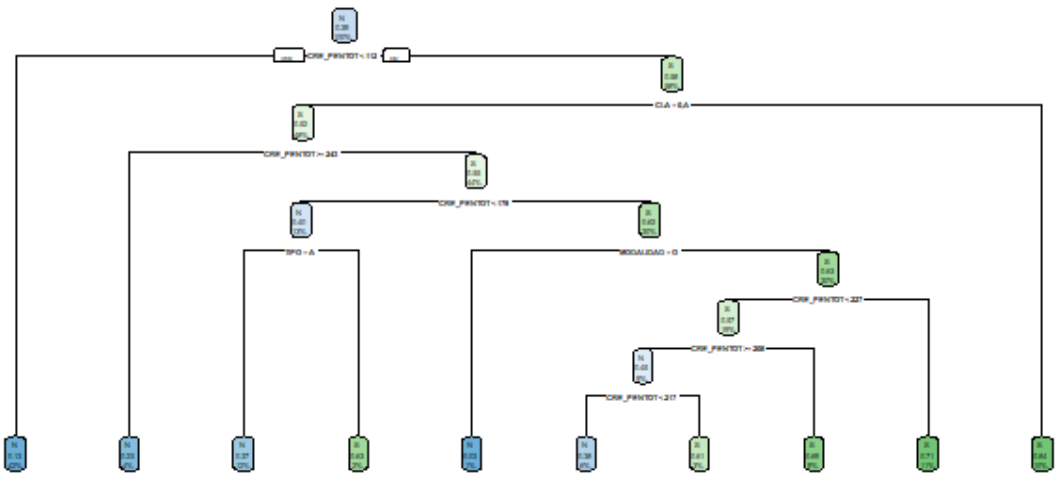
Los datos en la Facultad de Farmacia relativos a los másteres son insuficientes tanto para asignar un peso a las variables como para crear un modelo.

Figura 98: Peso de las variables más importantes en la Facultad de Ciencias Sociales y Jurídicas de Elche (GRADOS)



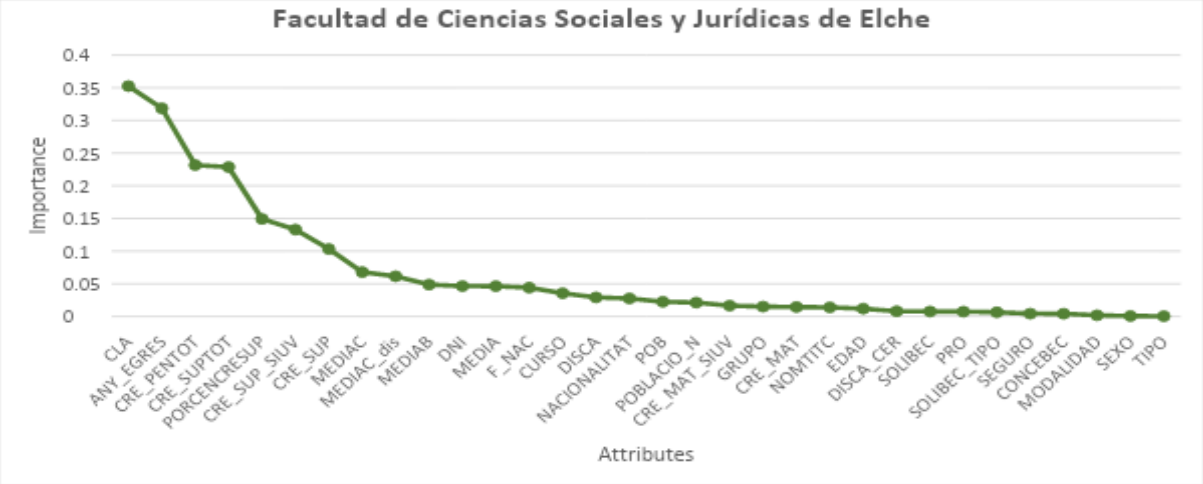
El año de finalización de estudios es la variable que más influye en el abandono de los grados por parte de los estudiantes de la Facultad de Ciencias Sociales y Jurídicas de Elche.

Figura 99: Árbol de clasificación de la Facultad de Ciencias Sociales y Jurídicas de Elche (GRADOS)



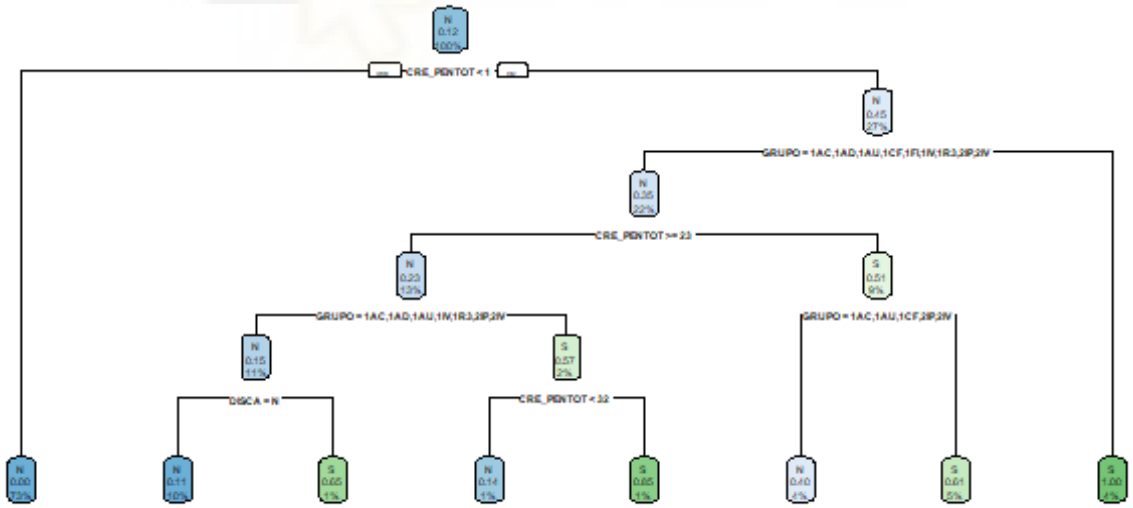
Accuracy=0.7877834

Figura 100: Peso de las variables más importantes en la Facultad de Ciencias Sociales y Jurídicas de Elche (MÁSTERES)



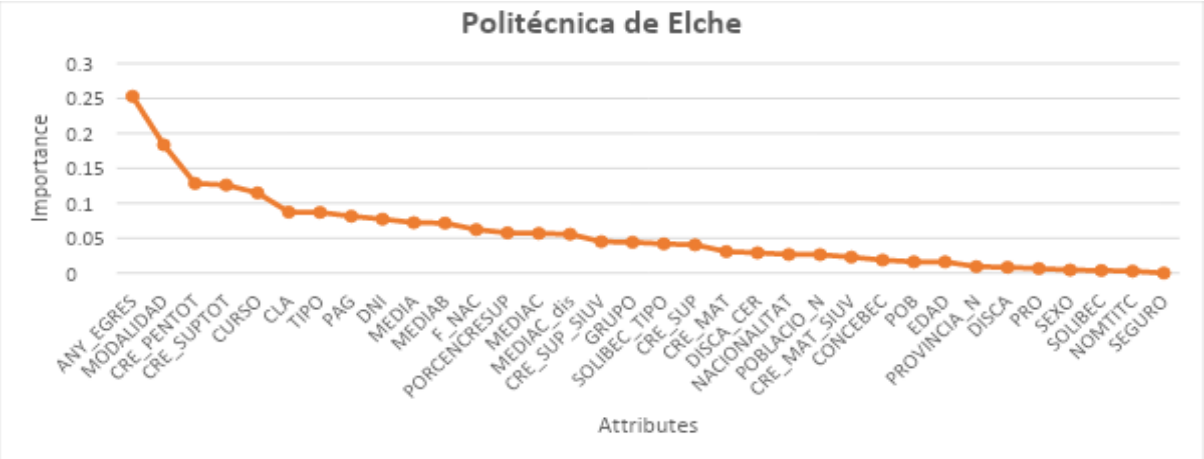
La clase de matrícula y el año de finalización de los estudios son las variables que más influyen en el abandono de másteres por parte de los estudiantes de la Facultad de Ciencias Sociales y Jurídicas de Elche.

Figura 101: Árbol de clasificación de la Facultad de Ciencias Sociales y Jurídicas de Elche (MÁSTERES)



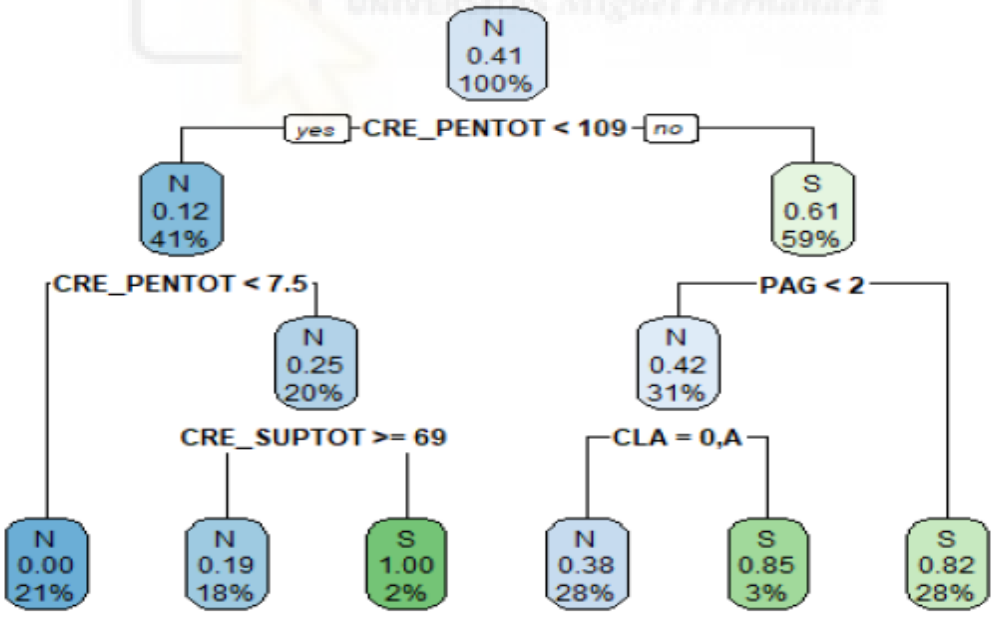
Accuracy=0.9486726

Figura 102: Peso de las variables más importantes en la Escuela Politécnica de Elche (GRADOS)



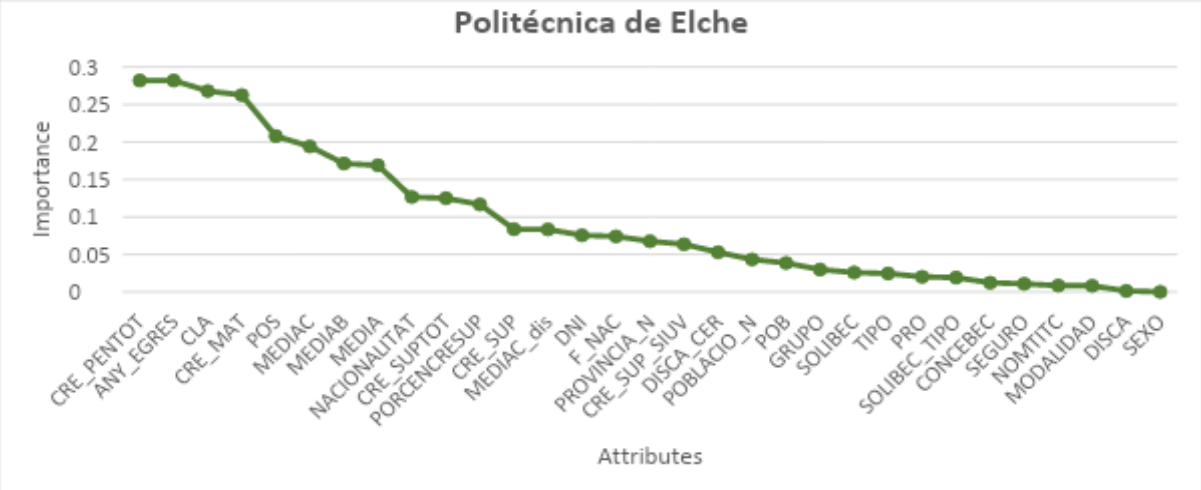
El año de finalización de los estudios, el nivel de presencialidad del grado, los créditos pendientes y superados totales y el curso en el que se está en el último año son las variables que más influyen en el abandono de los grados por parte de los estudiantes de la Escuela Politécnica de Elche.

Figura 103: Árbol de clasificación de la Escuela Politécnica de Elche (GRADOS)



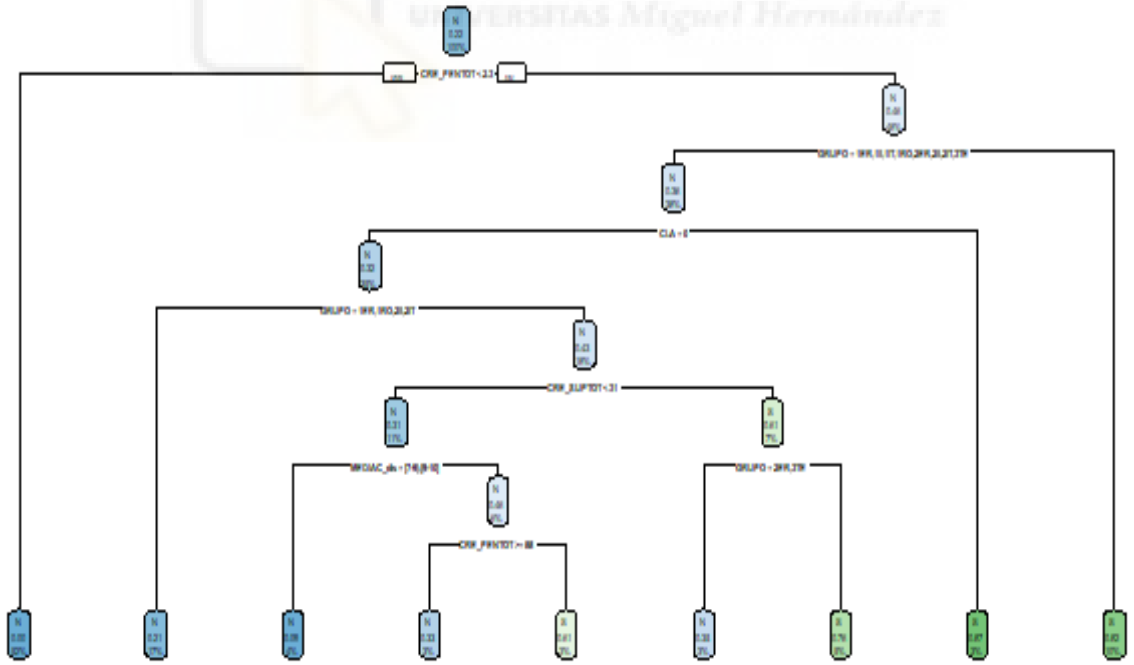
Como resultado del modelo con Accuracy 0.7832661, por la rama más a la derecha, podemos concluir que los estudiantes que tienen 109 créditos pendientes totales o más y que además el tipo de pago es 3=Cargo en cuenta (4 plazos), 4=Cargo en cuenta (1 plazo), 5=Cargo en cuenta (2 plazo)s, 6=Cargo en cuenta (8 plazos), 7=Cargo en cuenta (10 plazos) u 8=Cargo en cuenta (12 plazos), tienen un 82% de probabilidad de abandonar el grado.

Figura 104: Peso de las variables más importantes en la Escuela Politécnica de Elche (MÁSTERES)



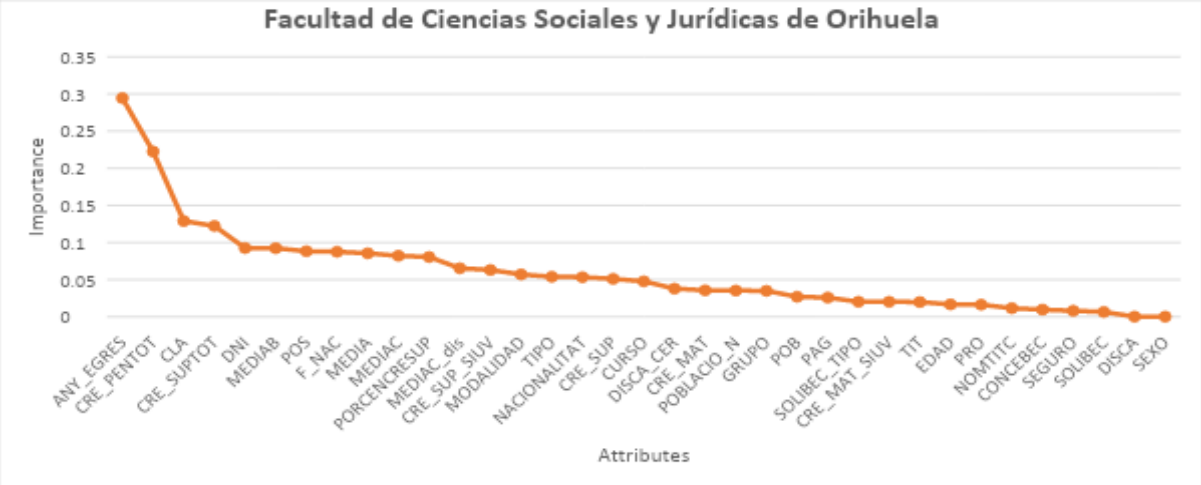
Los créditos pendientes totales, el año de finalización de los estudios, la clase de matrícula y los créditos matriculados son las variables que más influyen en el abandono de los másteres por parte de los estudiantes de la Escuela Politécnica de Elche.

Figura 105: Árbol de clasificación de la Escuela Politécnica de Elche (MÁSTERES)



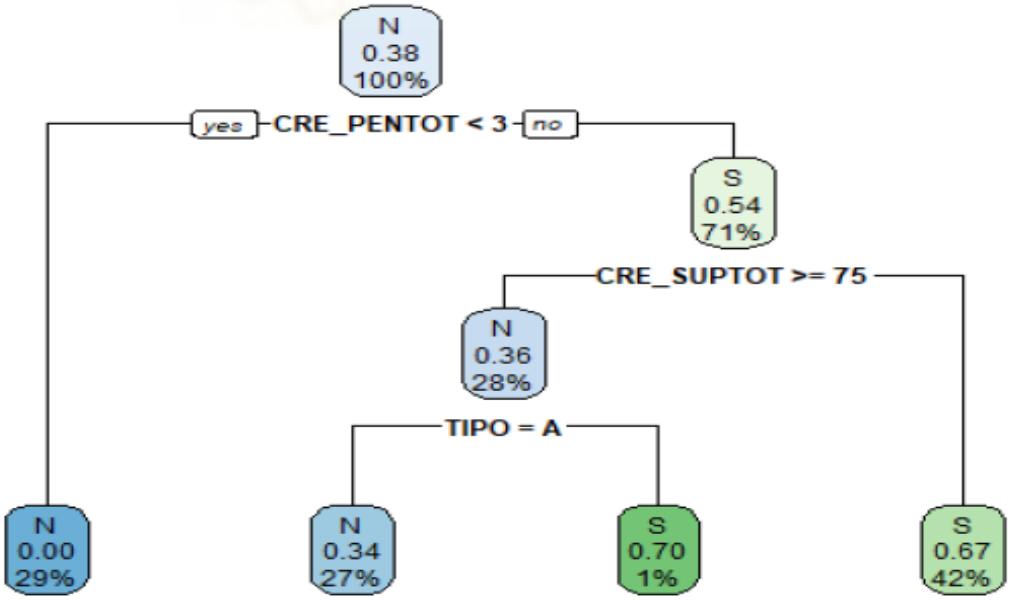
Accuracy=0.8832487

Figura 106: Peso de las variables más importantes en la Facultad de Ciencias Políticas y Jurídicas de Orihuela (GRADOS)



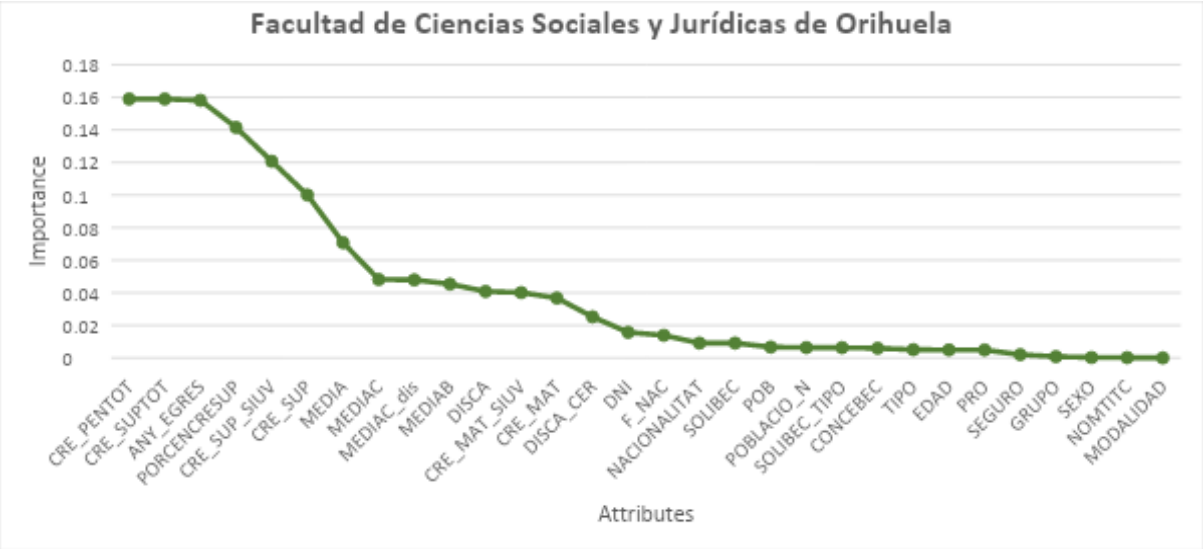
El año de finalización de los estudios y los créditos pendientes totales son las variables que más influyen en el abandono de los grados por parte de los estudiantes de la Facultad de Ciencias Sociales y Jurídicas de Orihuela.

Figura 107: Árbol de clasificación de la Facultad de Ciencias Sociales y Jurídicas de Orihuela (GRADOS)



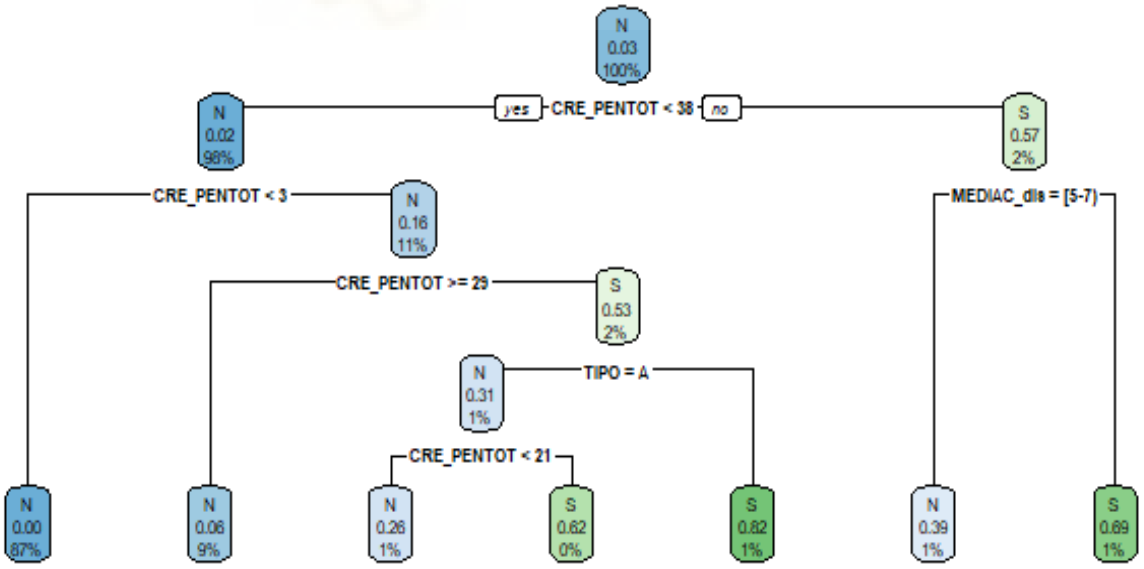
Como resultado del modelo con Accuracy 0.757485, por la rama más a la derecha, podemos concluir que los estudiantes que tienen 3 créditos pendientes totales o más y que además tienen menos de 75 créditos superados totales, tienen un 67% de probabilidad de abandonar el grado.

Figura 108: Peso de las variables más importantes en la Facultad de Ciencias Políticas y Jurídicas de Orihuela (MÁSTERES)



Los créditos pendientes y superados totales y el año de finalización de estudios son las variables que más influyen en el abandono de los másteres por parte de los estudiantes de la Facultad de Ciencias Sociales y Jurídicas de Orihuela.

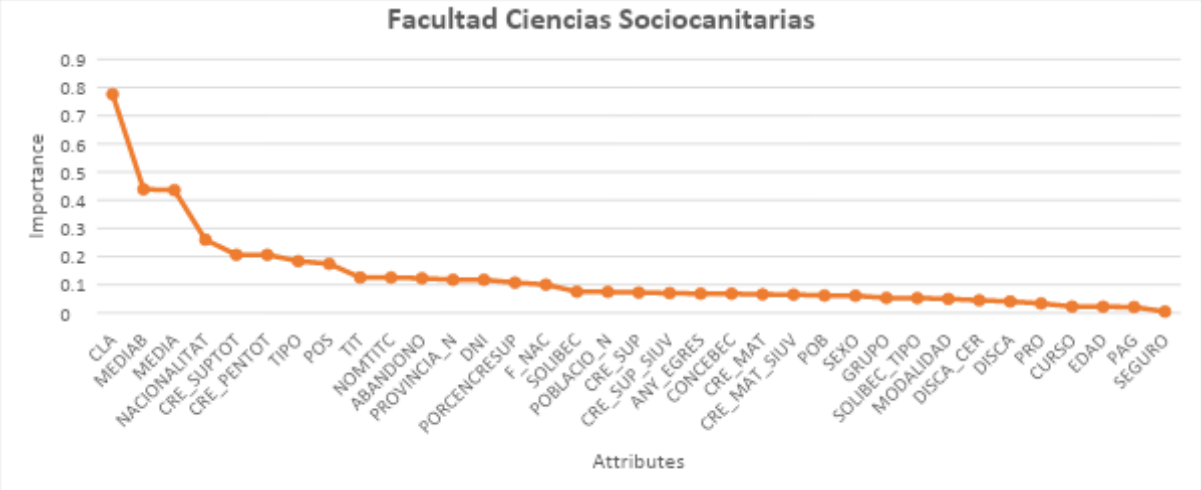
Figura 109: Árbol de clasificación de la Facultad de Ciencias Sociales y Jurídicas de Orihuela (MÁSTERES)



Como resultado del modelo con Accuracy 0.9816176, por la rama más a la derecha, podemos concluir que los estudiantes que tienen 38 créditos pendientes totales o más y cuya nota media no está entre 5 y 7 este último sin incluir, tienen un 69% de probabilidad de abandonar el grado.

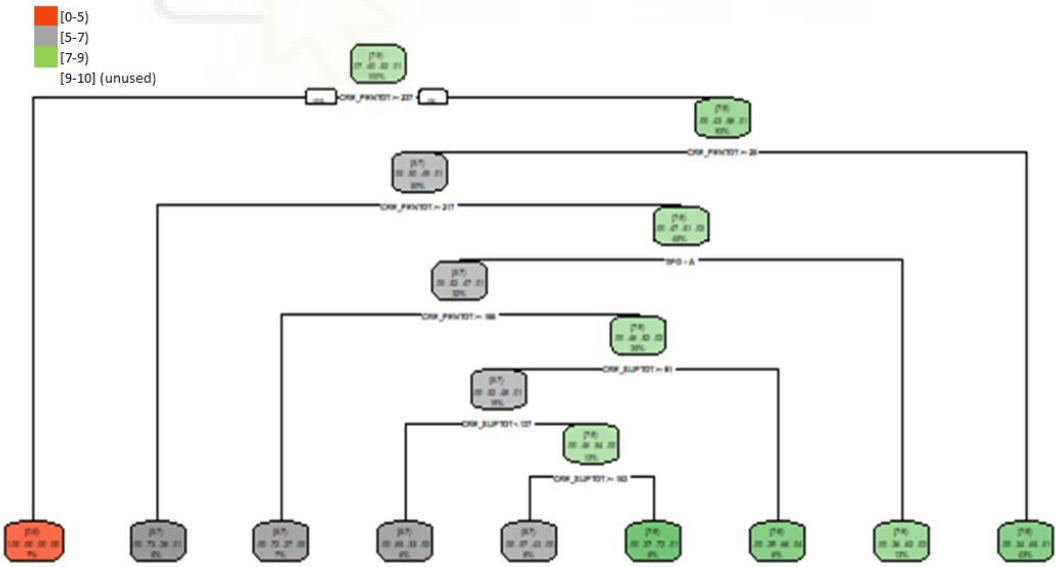
9.2.2 Nota media

Figura 110: Peso de las variables más importantes en la Facultad de Ciencias Sociosanitarias (GRADOS)



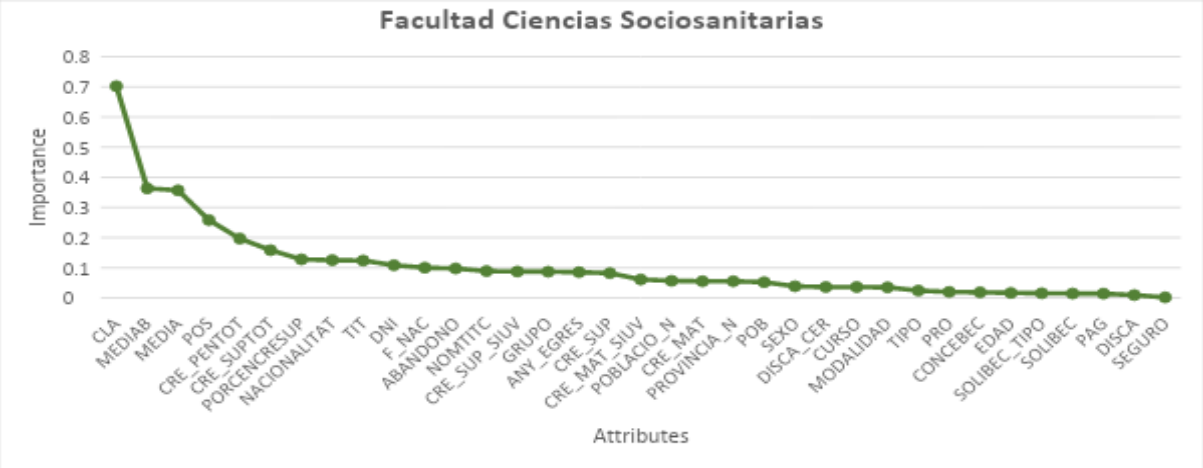
La clase de matrícula es la variable que más influye en la nota media obtenida por los estudiantes de los grados de la Facultad de Ciencias Sociosanitarias.

Figura 111: Árbol de clasificación de la Facultad de Ciencias Sociosanitarias (GRADOS)



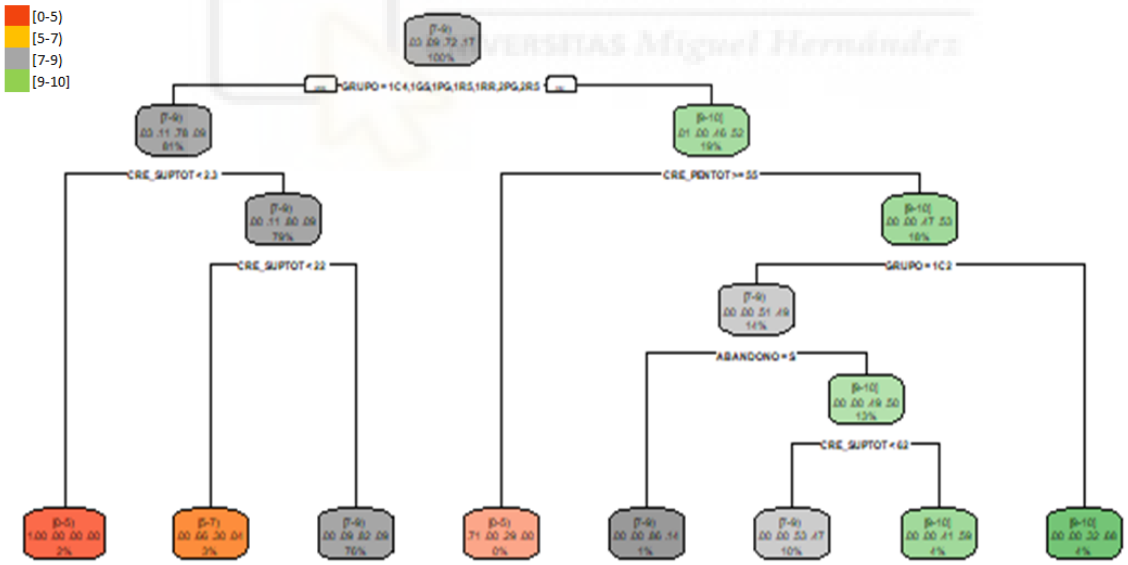
Accuracy=0.6885457

Figura 112: Peso de las variables más importantes en la Facultad de Ciencias Sociosanitarias (MÁSTERES)



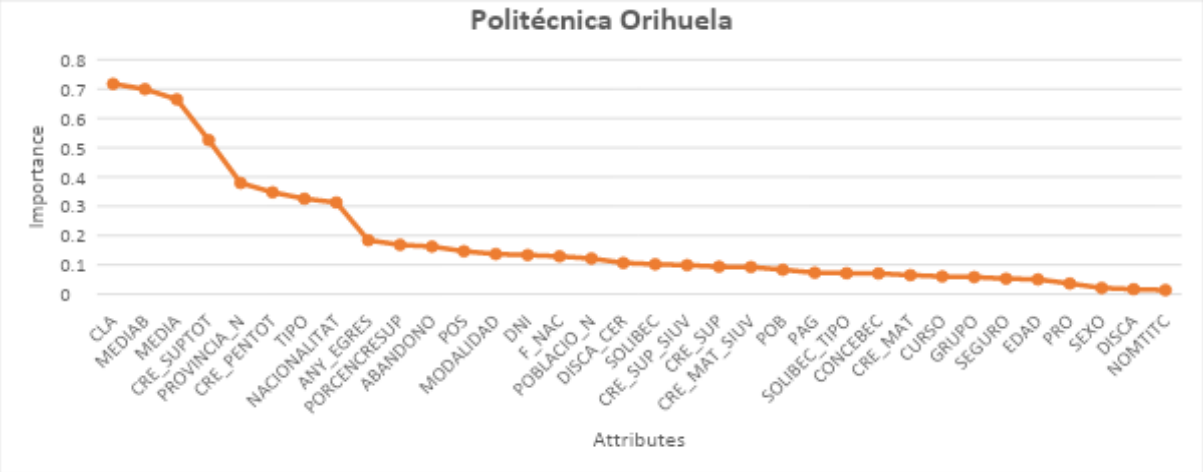
Al igual que ocurre en los grados, la clase de matrícula es la variable que más influye en la nota media obtenida por los estudiantes de los másteres de la Facultad de Ciencias Sociosanitarias.

Figura 113: Árbol de clasificación de la Facultad de Ciencias Sociosanitarias (MÁSTERES)



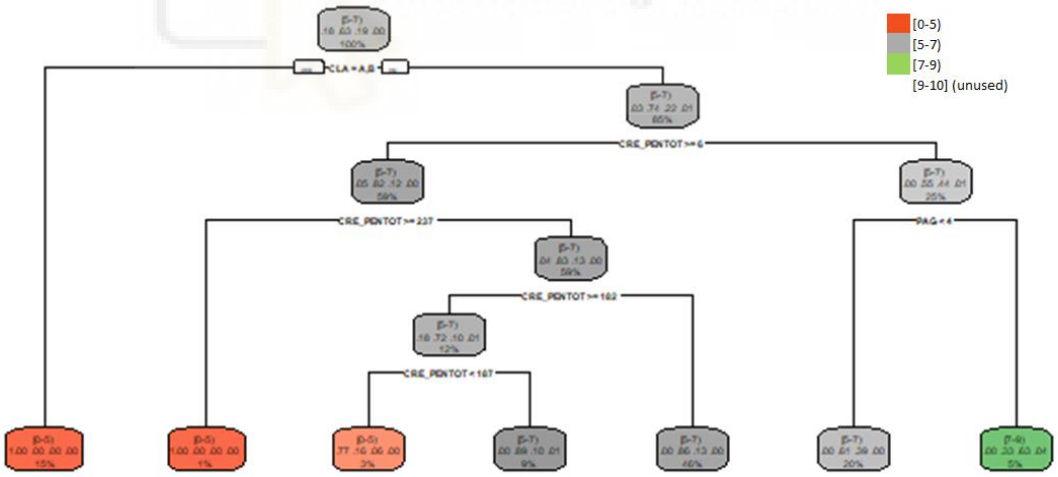
Accuracy=0.7707948

Figura 114: Peso de las variables más importantes en la Escuela Politécnica de Orihuela (GRADOS)



Al igual que ocurre en la Facultad de Ciencias Sociosanitarias, tanto en los grados como en los másteres, la clase de matrícula es la variable que más influye en la nota media obtenida por los estudiantes de la Escuela Politécnica de Orihuela.

Figura 115: Árbol de clasificación de la Escuela Politécnica de Orihuela (GRADOS)



Accuracy=0.8405797

Figura 116: Peso de las variables más importantes en la Escuela Politécnica de Orihuela (MÁSTERES)

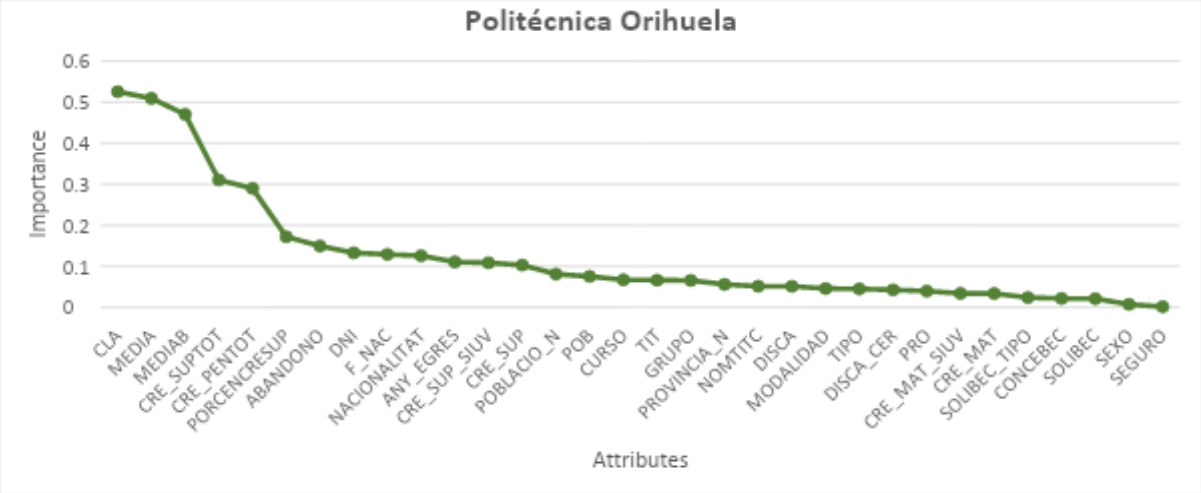
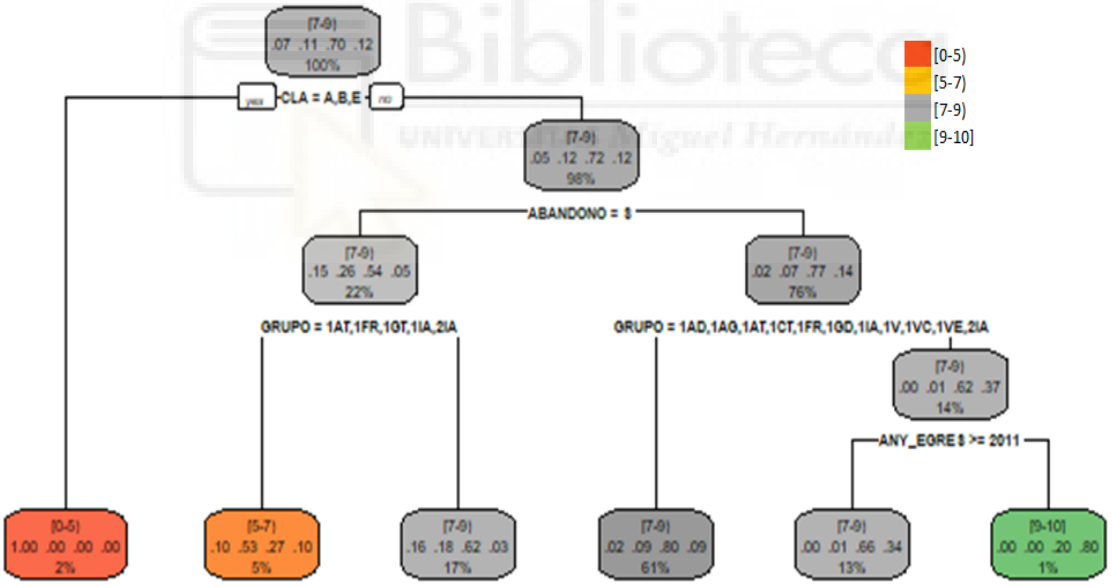
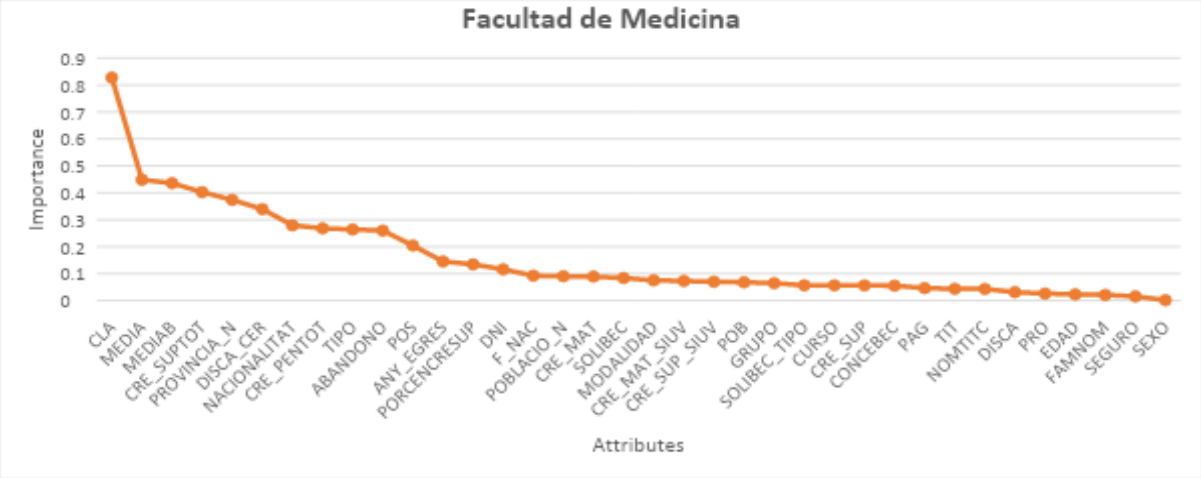


Figura 117: Árbol de clasificación de la Escuela Politécnica de Orihuela (MÁSTERES)



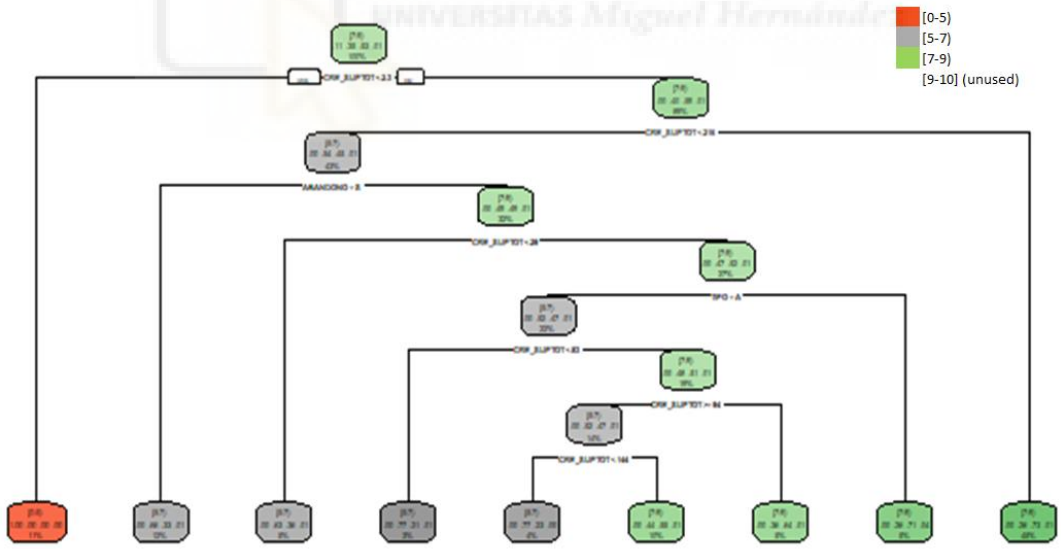
Como resultado del modelo con Accuracy 0.7868852, por la rama derecha, podemos concluir que los estudiantes que no son ni alumno visitante, ni alumno de programa de intercambio, ni tienen matrícula especial CFD, que además no abandonan el máster, que no pertenecen a los grupos de esta rama y que el último año de estudios cursado es menor que el 2011, tienen un 80% de probabilidad de obtener una calificación media muy alta.

Figura 118: Peso de las variables más importantes en la Facultad de Medicina (GRADOS)



Al igual que ocurre en la Universidad Politécnica de Orihuela y en la Facultad de Ciencias Sociosanitarias, tanto en los grados como en los másteres, la clase de matrícula es la variable que más influye en la nota media obtenida por los estudiantes de la Facultad de Medicina.

Figura 119: Árbol de clasificación de la Facultad de Medicina (GRADOS)



Accuracy=0.7297514

Figura 120: Peso de las variables más importantes en la Facultad de Medicina (MÁSTERES)

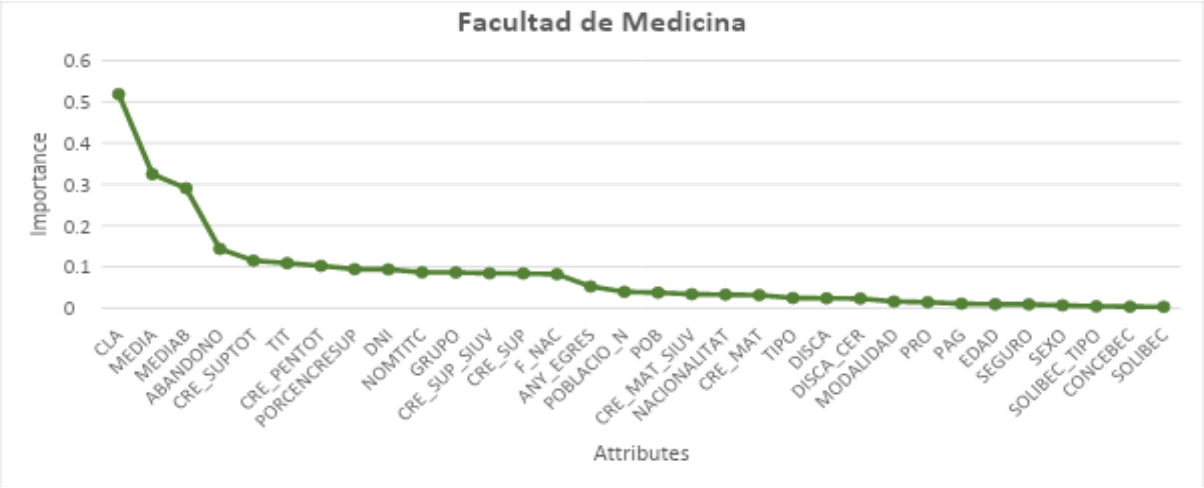
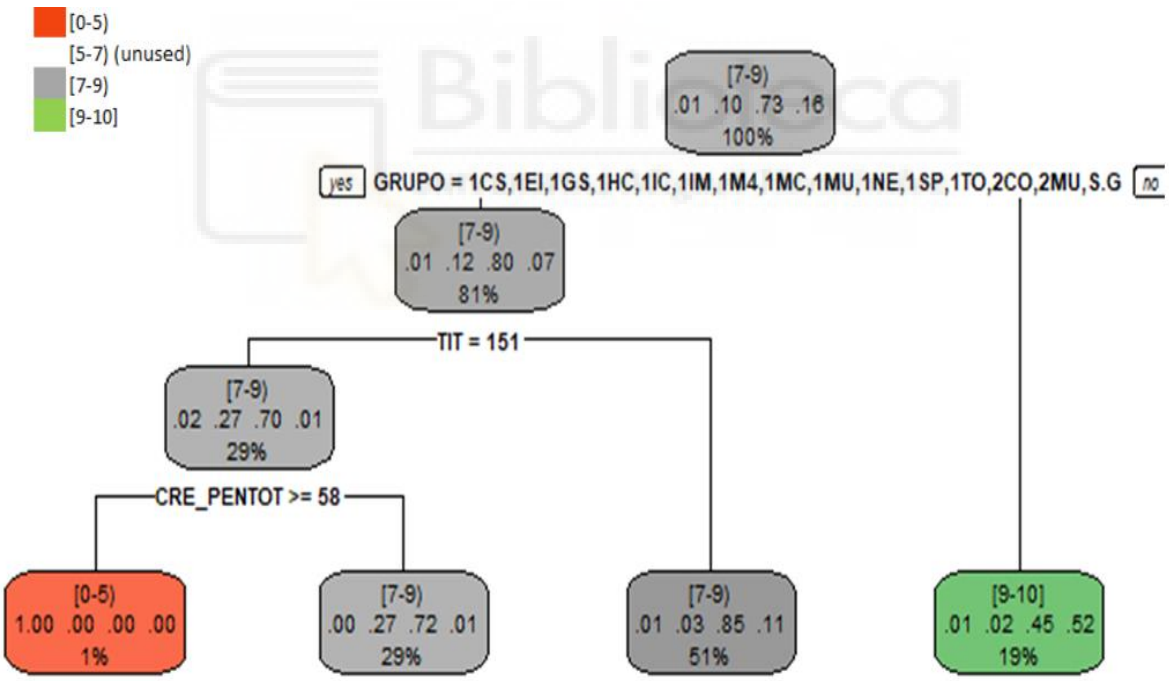
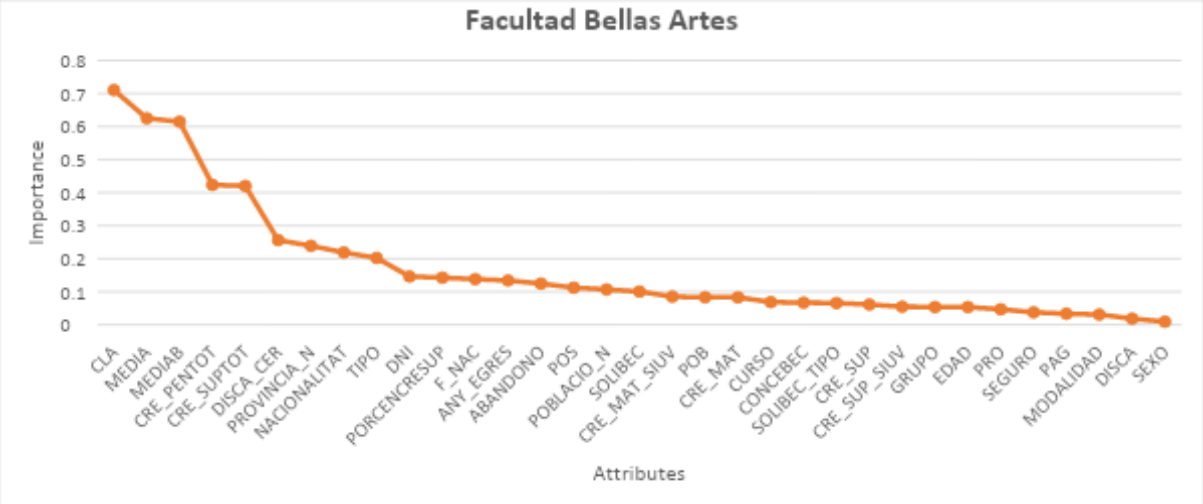


Figura 121: Árbol de clasificación de la Facultad de Medicina (MÁSTERES)



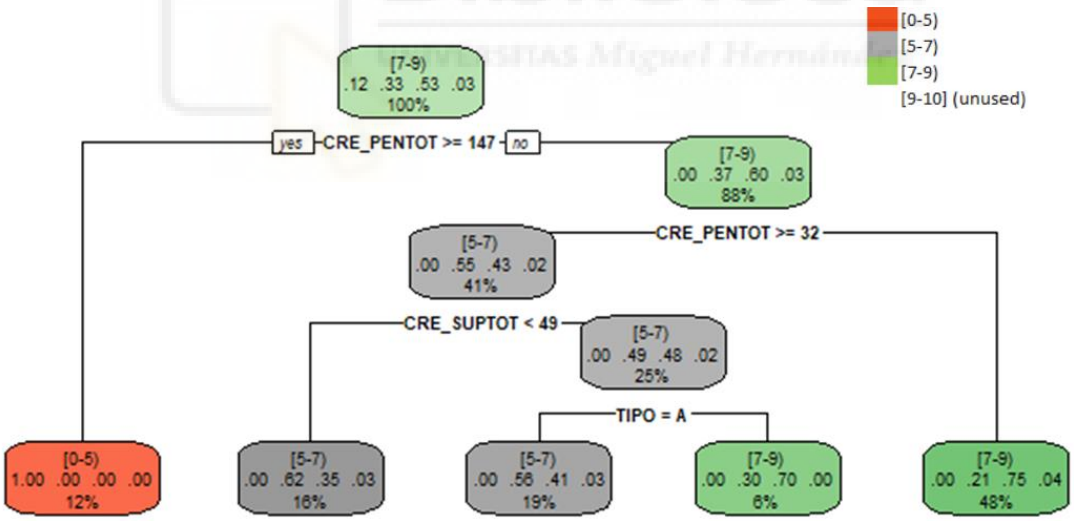
Como resultado del modelo con Accuracy 0.7491039, por la rama izquierda, podemos concluir que los estudiantes que se encuentran en los grupos indicados en la figura, cuyo código de titulación es 151 y que tienen un número mayor o igual de 58 créditos pendientes totales no tienen ninguna probabilidad de aprobar.

Figura 122: Peso de las variables más importantes en la Facultad de Bellas Artes (GRADOS)



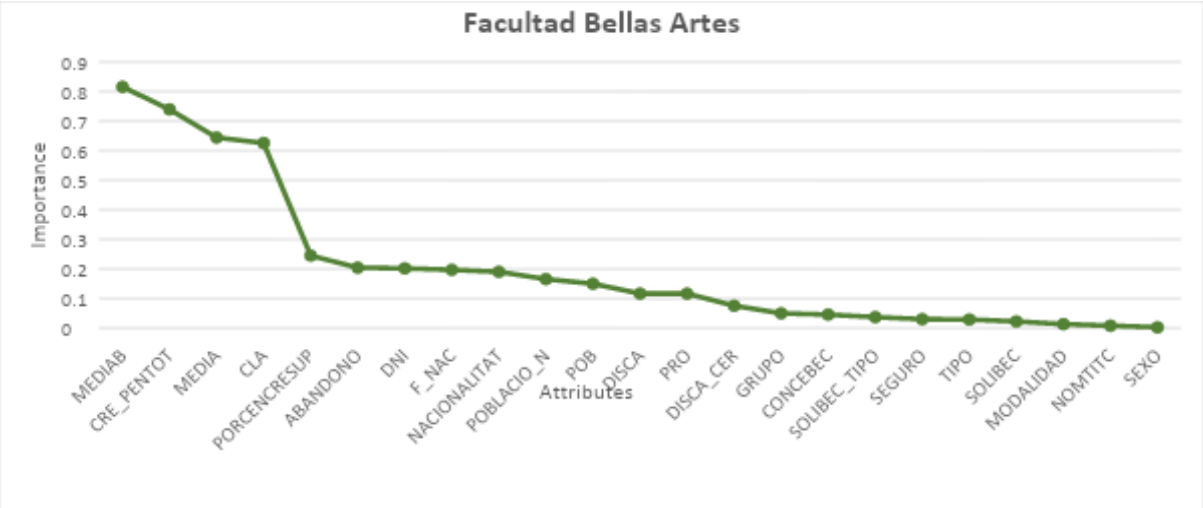
Al igual que ocurre en las facultades anteriores, la clase de matrícula es la variable que más influye en la nota media obtenida por los estudiantes de la Facultad de Bellas Artes.

Figura 123: Árbol de clasificación de la Facultad de Bellas Artes (GRADOS)



Como resultado del modelo con Accuracy 0.7222222, por la rama izquierda, podemos concluir que los estudiantes que tienen 147 créditos pendientes totales o más no tienen ninguna probabilidad de aprobar.

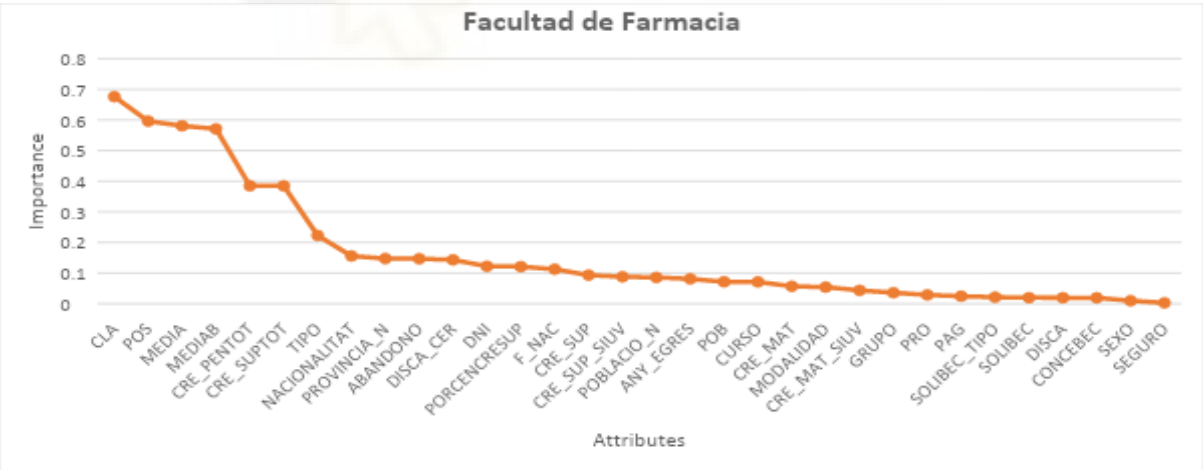
Figura 124: Peso de las variables más importantes en la Facultad de Bellas Artes (MÁSTERES)



En el caso de los másteres, también influyen los créditos pendientes totales y la clase de matrícula en la nota media obtenida por los estudiantes de la Facultad de Bellas Artes.

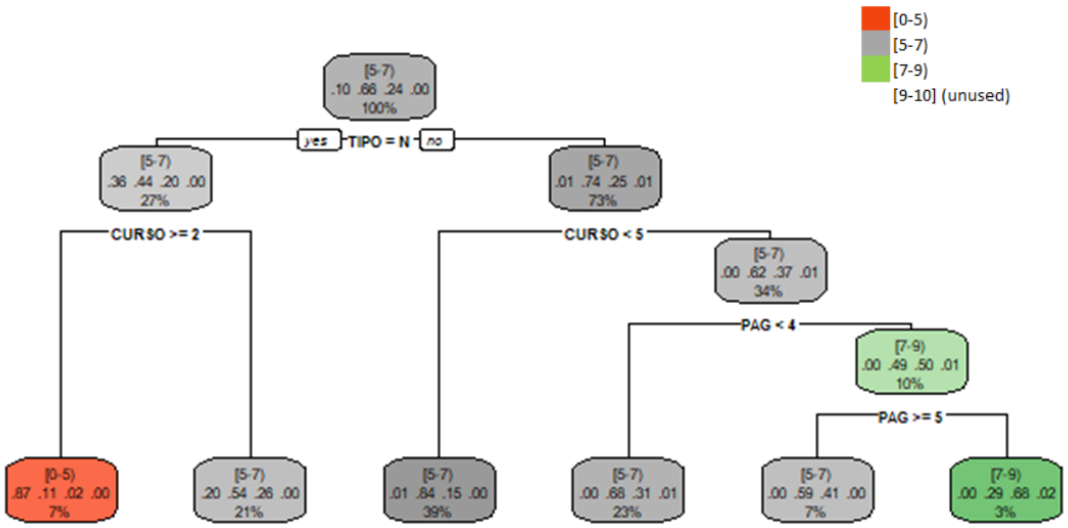
No se puede crear un modelo debido a que no hay suficientes datos de los másteres en la Facultad de Bellas Artes.

Figura 125: Peso de las variables más importantes en la Facultad de Farmacia (GRADOS)



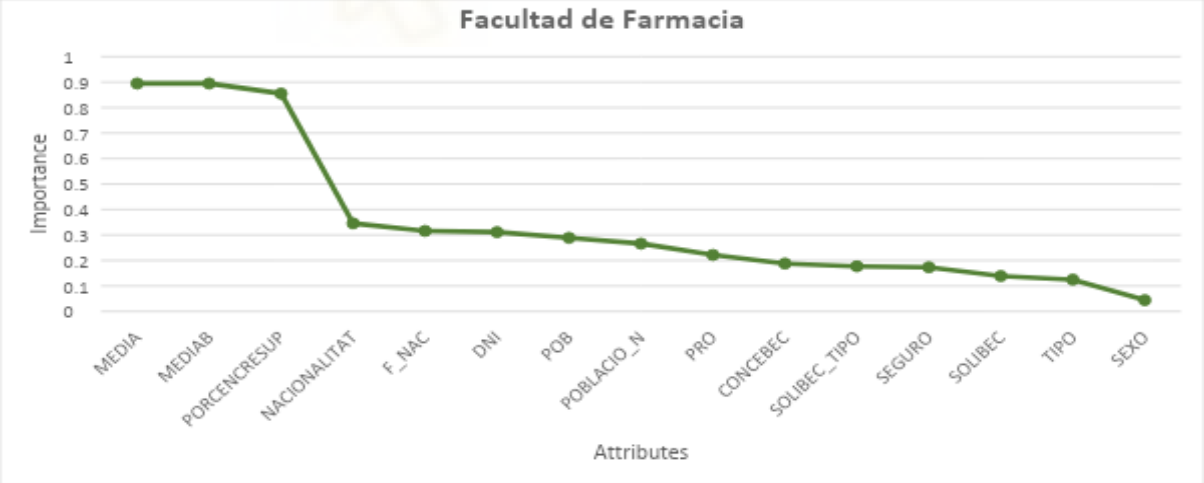
La clase de matrícula y el código postal son las variables que más influyen en la nota media obtenida por los estudiantes de la Facultad de Farmacia, al igual que ocurría en el caso de la Facultad de Ciencias Experimentales.

Figura 126: Árbol de clasificación de la Facultad de Farmacia (GRADOS)



Como resultado del modelo con Accuracy 0.6952381 podemos concluir, por la rama izquierda, que los estudiantes de nuevo ingreso que están en segundo curso o más tienen una probabilidad del 87% de suspender.

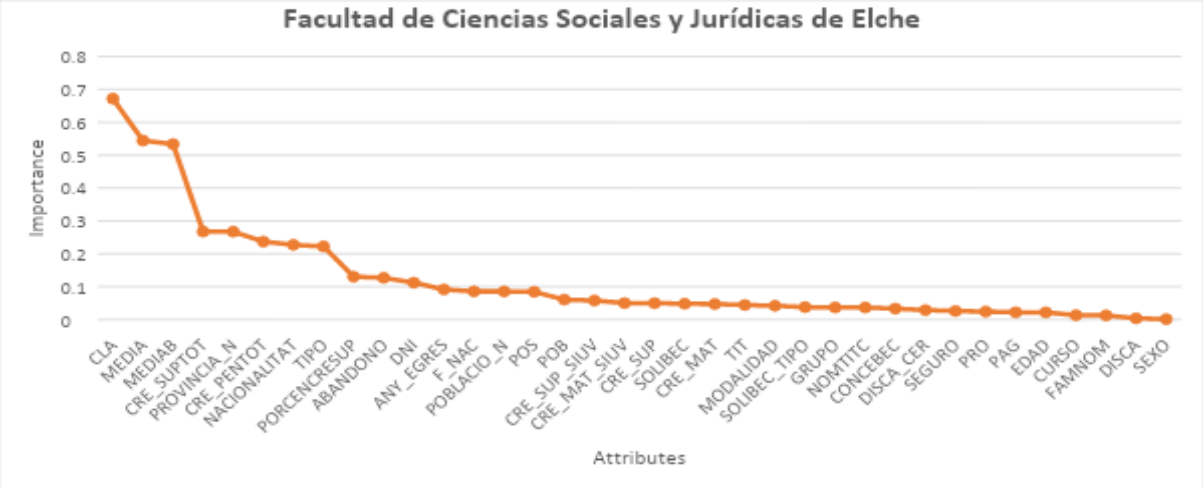
Figura 127: Peso de las variables más importantes en la Facultad de Farmacia (MÁSTERES)



En el caso de los másteres, el porcentaje de créditos superados es una de las variables que más influye en la nota media obtenida por los estudiantes de la Facultad de Farmacia.

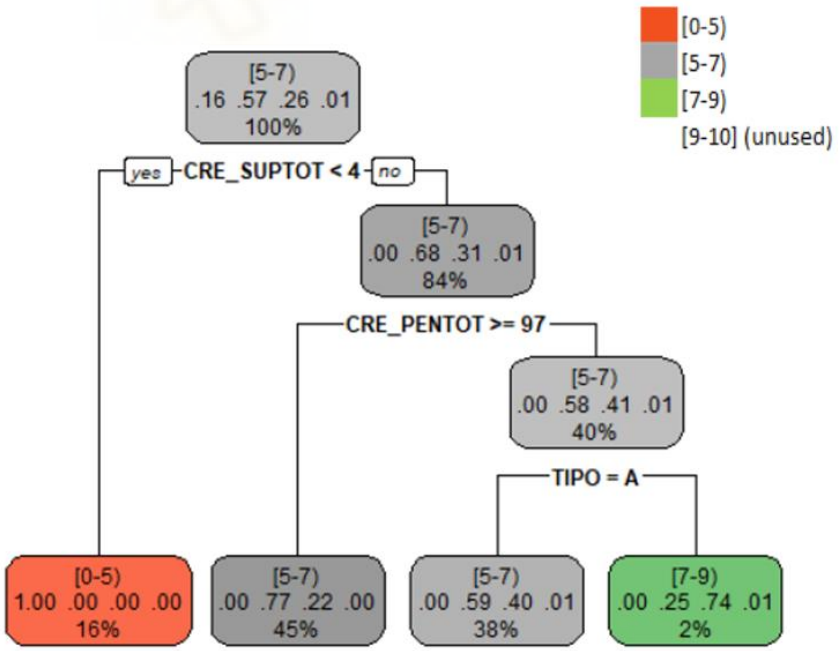
No ha sido posible crear un modelo sobre la nota media en la Facultad de Farmacia respectivo a los másteres debido a la insuficiencia de datos.

Figura 128: Peso de las variables más importantes en la Facultad de Ciencias Sociales y Jurídicas de Elche (GRADOS)



Al igual que ocurre en la Escuela Politécnica de Orihuela, en la Facultad de Ciencias Sociosanitarias y en la Facultad de Medicina, tanto en los grados como en los másteres, la clase de matrícula es la variable que más influye en la nota media obtenida por los estudiantes de la Facultad de Ciencias Sociales y Jurídicas de Elche.

Figura 129: Árbol de clasificación de la Facultad de Ciencias Sociales y Jurídicas de Elche (GRADOS)



Como resultado del modelo con Accuracy 0.7418136 podemos concluir, por la rama derecha, que los estudiantes que tienen 4 créditos superados totales o más y además tienen menos de 97

créditos pendientes y son estudiantes de nuevo ingreso tienen una probabilidad alta 74% de obtener una nota media entre 7 y 9, este último sin incluir.

Figura 130: Peso de las variables más importantes en la Facultad de Ciencias Sociales y Jurídicas de Elche (MÁSTERES)

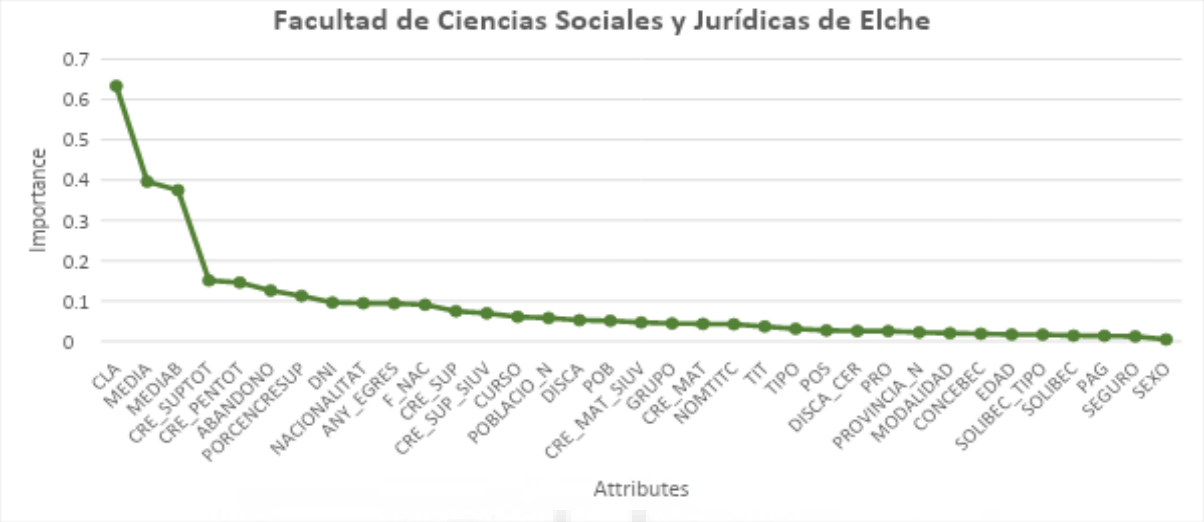
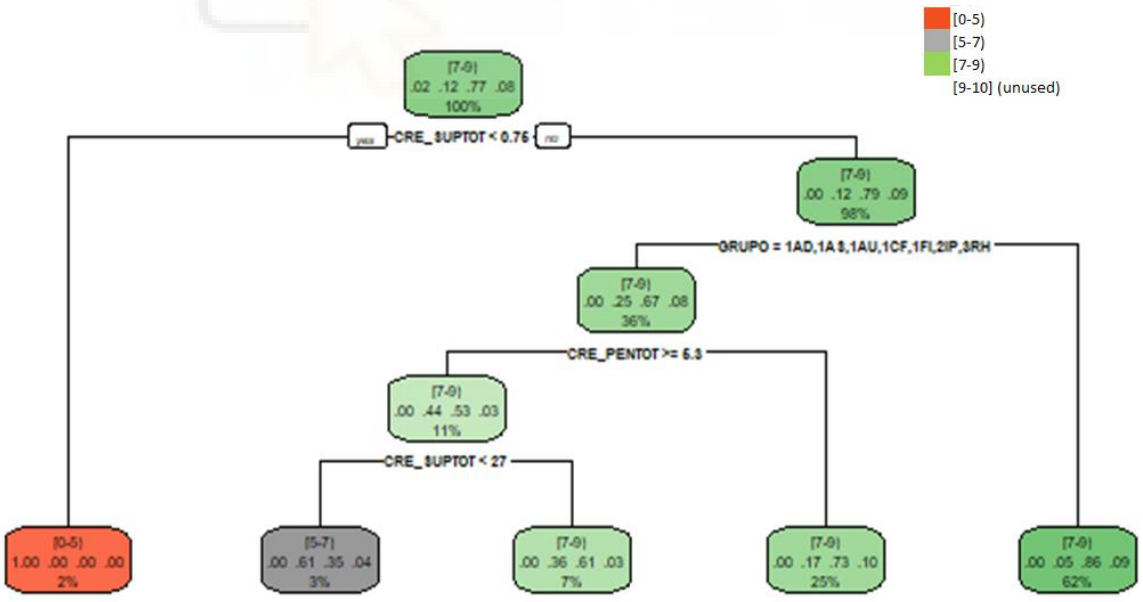
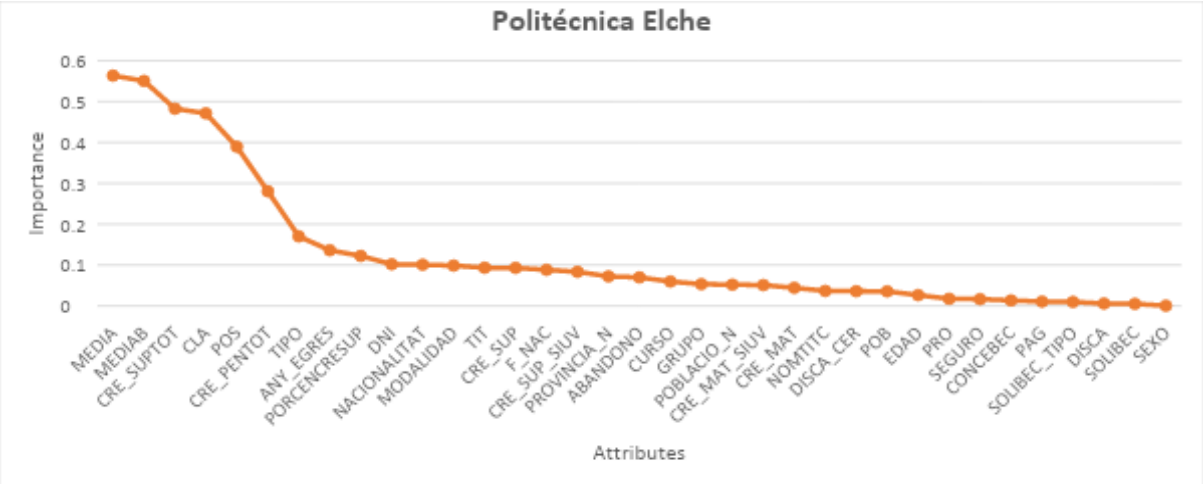


Figura 131: Árbol de clasificación de la Facultad de Ciencias Sociales y Jurídicas de Elche (MÁSTERES)



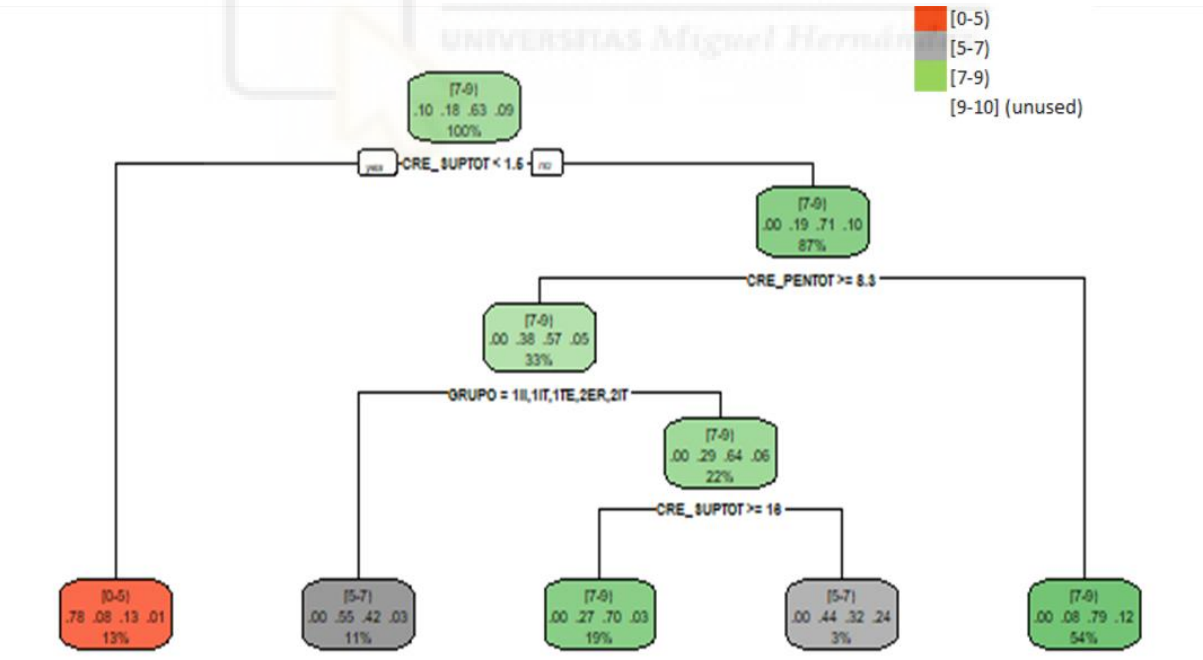
Como resultado del modelo con Accuracy 0.780531 podemos concluir, por la rama izquierda, que los estudiantes que tienen menos de 0.76 créditos superados totales no tienen ninguna probabilidad de aprobar.

Figura 132: Peso de las variables más importantes en la Escuela Politécnica de Elche (GRADOS)



Los créditos superados totales, la clase de matrícula y el código postal son algunas de las variables que más influyen en la nota media obtenida por los estudiantes de la Escuela Politécnica de Elche.

Figura 133: Árbol de clasificación de la Escuela Politécnica de Elche (GRADOS)



Como resultado del modelo con Accuracy 0.8366935 podemos concluir, por la rama derecha, que los estudiantes que tienen igual o más de 1.5 créditos superados totales y además tienen menos de 8.3 créditos pendientes tienen una probabilidad alta 79% de obtener una nota media entre 7 y 9, este último sin incluir.

Figura 134: Peso de las variables más importantes en la Escuela Politécnica de Elche (MÁSTERES)

En el caso de los másteres, son también la clase de matrícula y el código postal algunas de las variables que más influyen en la nota media obtenida por los estudiantes de la Universidad Politécnica de Elche.

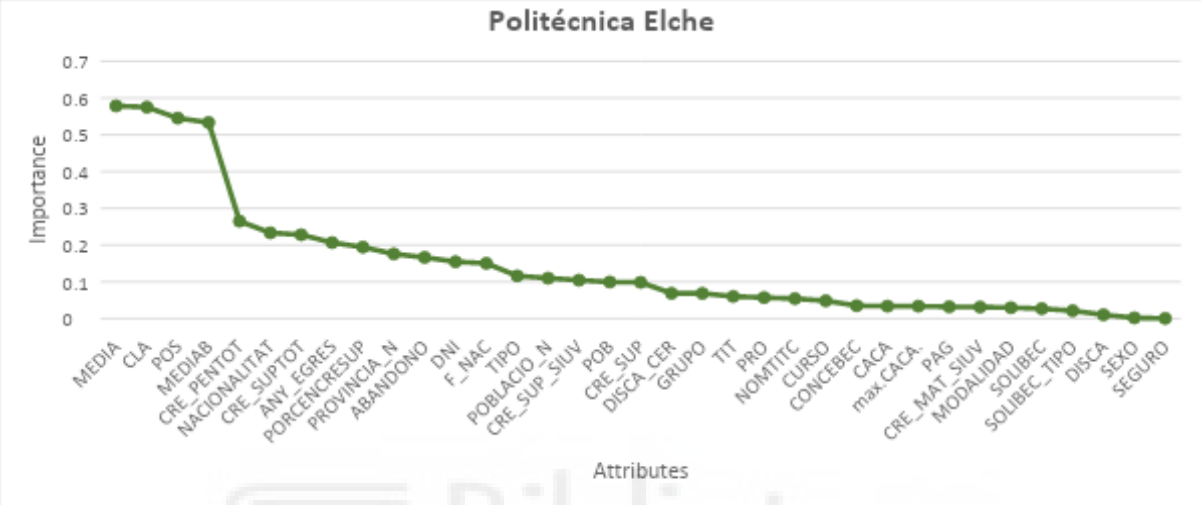
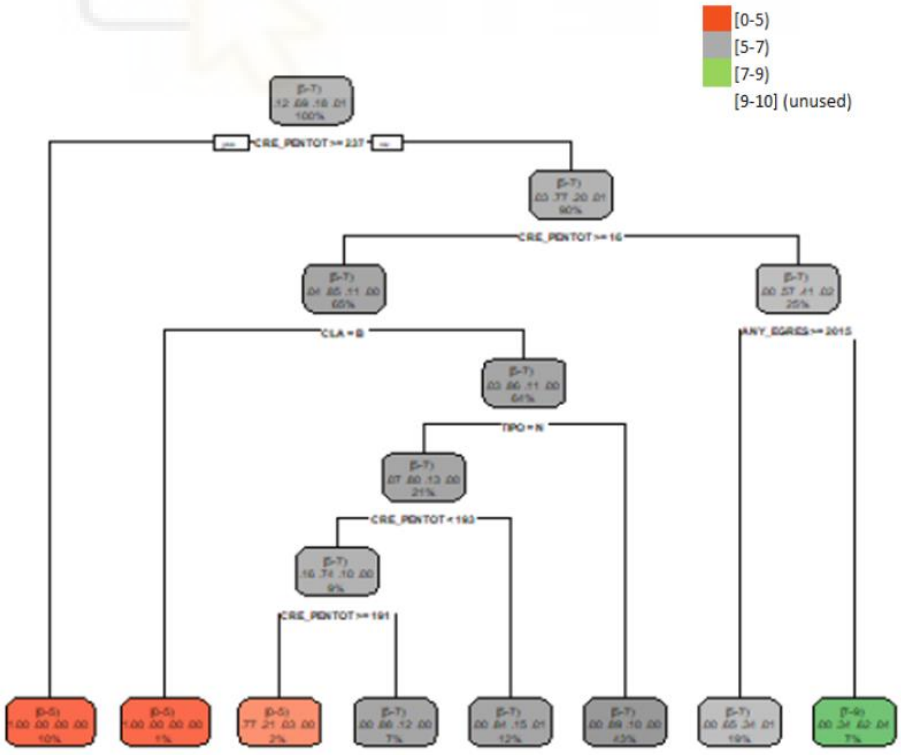
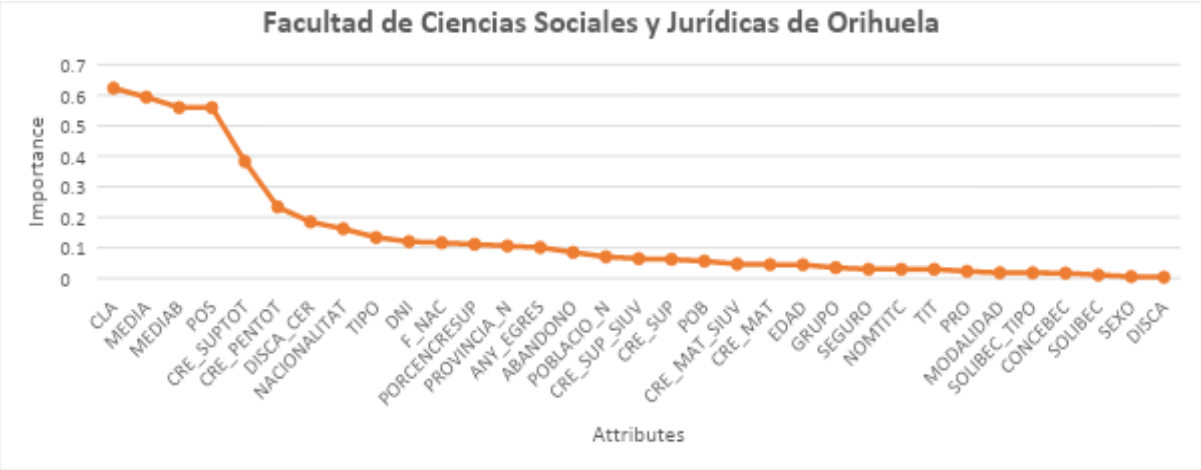


Figura 135: Árbol de clasificación de la Escuela Politécnica de Elche (MÁSTERES)



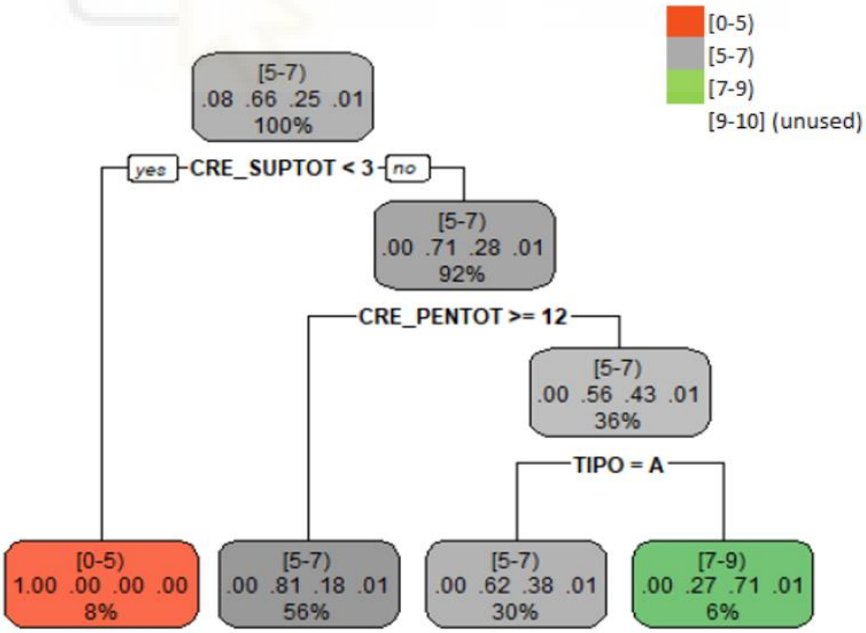
Accuracy=0.7258883

Figura 136: Peso de las variables más importantes en la Facultad de Ciencias Sociales y Jurídicas de Orihuela (GRADOS)



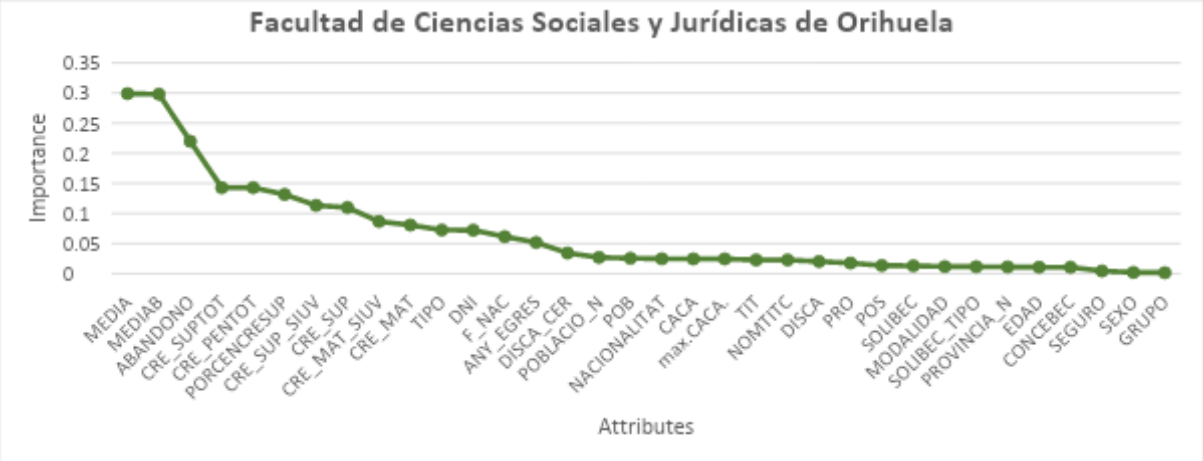
La clase de matrícula y el código postal son algunas de las variables que más influyen en la nota media obtenida por los estudiantes de la Facultad de Ciencias Sociales y Jurídicas de Orihuela.

Figura 137: Árbol de clasificación de la Facultad de Ciencias Sociales y Jurídicas de Orihuela (GRADOS)



Como resultado del modelo con Accuracy 0.760479 podemos concluir, por la rama derecha, que los estudiantes que tienen igual o más de 3 créditos superados totales y además tienen menos de 12 créditos pendientes y no son antiguos alumnos tienen una probabilidad alta 71% de obtener una nota media entre 7 y 9, este último sin incluir.

Figura 138: Peso de las variables más importantes en la Facultad de Ciencias Sociales y Jurídicas de Orihuela (MÁSTERES)



En el caso de los másteres, es el abandono de los estudios realizados una de las variables que más influye en la nota media obtenida por los estudiantes de la Facultad de Ciencias Sociales y Jurídicas de Orihuela.

No ha sido posible realizar un modelo de los másteres en la Facultad de Ciencias Sociales y Jurídicas de Orihuela debido a la falta de datos.

10. BIBLIOGRAFÍA

[1] L. Breiman, J. Friedman, C.J. Stone and R.A. Olshen. 1984. Classification and Regression Trees. Taylor & Francis.

[2] P. Chapman, J. Clinton, R. Kerber, R.T. Khabaza, T. Reinartz, C. Shearer and R. Wirth. 1999. CRISP-DM 1.0 Step by step Data Mining Guide. SPSS Internal Report.

[3] E. Crisol Moya y M.A. Romero López. 2015. Inclusive leadership as a strategy to avoid school leaving: opinion of families. Education Siglo XXI, Vol. 38, nº2, 2020, pp. 45 - 66.

[4] A. García and C. Adrogué. Fuentes. 2015. Abandono de los estudios universitarios: Dimensión, factores asociados y desafíos para la política pública. Revista Fuentes, nº16, junio 2015, pp. 85 - 106.

[5] J.B. Liu and J. Han. 2002. A practical knowledge discovery process for distributed data mining. 11th International Conference on Intelligent Systems - Emerging Technologies. Int Soc Comp & Their Applicat; Hewlett Packard Inc.

[6] C. Marcelo. 2013. Technologies for innovation and teaching practice. Brazilian Journal of Education, Vol. 18, nº52, Rio de Janeiro jan./mar. 2013.

[7] J.R. Quinlan. 1986. Inducción de árboles de decisión Machine Learning, Vol. 1.1, pp. 81 - 106.

[8] C.E. Shannon. 1948. A Mathematical Theory of Communication. The Bell System Technical Journal, Vol. 27, pp. 379 - 423, 623 - 656.

[9] O.M. Spositto, M. Etcheverry, H.L. Ryckeboer and J. Bossero. 2010. Aplicación de Técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil. Memorias de la Novena Conferencia Iberoamericana de Sistemas. Cibernética e Informática (CISCI 2010), Orlando, Florida, EE.UU.

10.1 RECURSOS WEB

En este apartado se presentan diferentes enlaces web con materiales que pueden resultar de utilidad para la materia tratada en este trabajo:

[10] Tools for Working with Categorical Variables: **https://forcats.tidyverse.org/** (consultada por última vez el 18 de julio de 2021)

[11] El paquete dplyr | Programación en R: **https://rsanchezs.gitbooks.io/rprogramming/content/chapter9/dplyr.html** (consultada por última vez el 18 de julio de 2021)

[12] Árboles de decisión con R - Clasificación: **https://rpubs.com/jboscomendoza/arboles_decision_clasificacion** (consultada por última vez el 18 de julio de 2021)

[13] Introducción al paquete Caret: **https://rpubs.com/joser/caret** (consultada por última vez el 18 de julio de 2021)

[14] R para profesionales de los datos: una introducción: **https://datanalytics.com/libro_r/elementos-de-un-grafico-en-ggplot2.html** (consultada por última vez el 18 de julio de 2021)

[15] (PDF) Métodos de selección de atributos para clasificación supervisada basados en teoría de información: **https://www.researchgate.net/publication/331155838_Metodos_de_seleccion_de_atributos_para_clasificacion_supervisada_basados_en_teoria_de_informacion** (consultada por última vez el 30 de agosto de 2021)