UNIVERSITAT D'ALACANT
UNIVERSIDAD DE ALICANTE
Facultat de Ciències Econòmiques i Empresarials
Facultad de Ciencias Económicas y Empresariales

GRADO EN ADMINISTRACIÓN Y DIRECCIÓN DE EMPRESAS

CURSO ACADÉMICO 2022 - 2023

PREDICCIÓN DEL DROP OUT DE LOS ALUMNOS EN 1º DE ADE

JOSÉ AGUILAR VAN DER HOFSTADT

PEDRO ALBARRÁN PEREZ
DEPARTAMENTO B119 FUNDAMENTOS DEL ANÁLISIS ECONÓMICO

San Vicente del Raspeig, fecha (junio 2023)

RESUMEN

El objetivo de este análisis consiste en realizar una predicción sobre el abandono de la carrera (drop out) durante el primer año de los estudiantes de ADE utilizando datos del propio estudiante y de su rendimiento académico.

El propósito de esta predicción es identificar con anticipación a aquellos estudiantes que están en riesgo de abandonar la carrera para tomar las medidas necesarias y prevenirlo o brindarles orientación hacia la mejor decisión.

Se emplearán modelos de predicción estadísticos, para llevar a cabo esta predicción y se realizará un análisis detallado de los factores de riesgo que pueden influir en el abandono de la carrera. El estudio proporcionará información valiosa para la Universidad de Alicante y para los profesores de centro, permitiendo la implementación de estrategias y políticas de apoyo adecuadas para fomentar la retención y el éxito de los estudiantes de ADE.

PALABRAS CLAVE

Drop out; Modelo de predicción; Alumnos; ADE

ÍNDICE DE ABREVIATURAS

Variable	Significado
EDM	Educational Data Mining
etc	ecétera
ADE	Administración y Dirección de Empresas
UA	Universidad de Alicante
SUE	Sistema Universitario Español
PAU	Prueba de Acceso a la Universidad
ALC	Alicante
ROC_AUC	Receiver Operating Characteristic - Area Under the Curve

ÍNDICE

1. INTRODUCCIÓN

La minería de datos educativos también conocida como Educational Data Mining (EDM) en inglés, se refiere a la aplicación de técnicas de minería de datos para descubrir patrones, tendencias y relaciones en el ámbito de la educación. Su enfoque se centra en el análisis de grandes conjuntos de datos generados en entornos educativos, como el rendimiento académico, registros de asistencia y datos personales de estudiantes y profesores, entre otros.

El objetivo de la minería de datos educativos es encontrar patrones que ayuden a comprender el comportamiento en el ámbito educativo, realizar predicciones para anticiparse a posibles eventos, así como segmentar y clasificar para ofrecer soluciones más personalizadas. A través del análisis de datos educativos, se busca obtener información valiosa que permita mejorar la enseñanza y el aprendizaje, identificar las áreas más ineficientes, adaptar estrategias educativas y tomar decisiones informadas en beneficio de los estudiantes y del sistema educativo en general.

Uno de los temas más recurrentes dentro de este campo es el análisis del drop out de los estudiantes. El drop out se refiere al abandono prematuro de los estudios, por lo que comprender las causas y los factores que han llevado a esta situación en un estudiante es de vital importancia, ya que esta decisión puede tener implicaciones muy significantes tanto para el estudiante como para la institución educativa.

Conocer estas causas ayuda a comprender las razones que puede haber detrás de esta decisión, y esto nos da la oportunidad de poder tomas medidas preventivas y diseñar estrategias para evitar el abandono, como programas de apoyo y orientación personalizado o la creación de políticas que fomenten la retención estudiantil.

Además, a través del estudio del drop out se pueden conocer posibles ineficiencias del sistema educativo, permitiendo su mejora.

En este caso, nos enfocaremos en el drop out universitario. Aunque el abandono de los estudios universitarios puede no parecer un problema de gran importancia para los estudiantes en comparación con el abandono de los estudios en etapas anteriores, prevenir este abandono tiene beneficios tanto para las instituciones educativas como para los propios estudiantes.

Desde la perspectiva de la institución educativa, evitar el drop out universitario ayuda a optimizar recursos y a mejorar la calidad de la educación. Por el otro lado, para los estudiantes, evitar el abandono universitario les permite evitar gastos económicos y ahorrar el tiempo.

Según los datos del ministerio de universidades del gobierno de España, la tasa global de abandono de los estudios en el curso 2018/19 fue del 33,9% en universidades públicas, siendo el 21,7% del abandono en el primer año de carrera, es decir, el 64,01% del abandono ocurre en el primer año de carrera. Esto supone la gran mayoría de los abandonos totales de la carrera y conseguir detectar los motivos o patrones pueden ayudar a lograr mayores retenciones estudiantiles y evitar muchos costes de deserción estudiantil para las Universidades.

En este análisis nos enfocaremos en el drop out del 1º curso de administración y dirección de empresas (ADE) de la universidad de Alicante (UA).

Según los datos del ministerio de universidades previamente mencionados, las cifras sobre la tasa de abandono en la rama de Ciencias Sociales y Jurídicas el primer año de carrera fue del 20,3%, un poco más baja que la tasa general.

En la UA estos datos en 2018/19 fueron del 18,60%, donde llama la atención que en el grado de ADE esta tasa fuese del 37,43%, que es un poco más del doble.

Atendiendo a las respuestas que da la UA sobre las tasas académicas, el coste de estas está financiado en gran parte por el estado español, siendo esta cifra entre el 70 y el 75% del coste total de la matrícula.

Contando únicamente el primer año de ADE, el cual está compuesto por 10 asignaturas de 6 créditos cada una, al precio de 12,79€ por crédito. Suponiendo que el estado financia el 70% del coste, cada crédito tiene un coste para el estado de aproximadamente 29,84€. Por lo tanto, el coste por asignatura sería de 179,06€ y, en consecuencia, el primer curso de ADE representa un coste al estado 1790,60€ por alumno.

Si este dato lo proyectamos en la tasa de abandono del primer año en el curso 2018/19, nos da un coste de abandono educativo de 250600€.

En este análisis vamos a crear un modelo que prediga que alumnos recién matriculados van a abandonar la carrera, también vamos a tratar de averiguar qué datos de los que disponemos tienen más relevancia a la hora de que el alumno tome su decisión del abandono, a su vez vamos a proponer posibles soluciones para atenuar el problema.

Se crearán los siguientes modelos: regresión logística, árbol de decisión y random Forest. Al finalizar el estudio comprobaremos cual es el modelo más preciso y por lo tanto el más conveniente para usar.

Los modelos se crearán a partir de los datos personales y académicos del estudiante. Es importante destacar, que los modelos utilizados están construidos a partir de la información disponible en la universidad. Por lo tanto, no se tienen en cuenta aspectos como las emociones, vida privada, u otros factores del estudiante que puedan ser cruciales para la decisión del abandono.

El trabajo se estructura de la siguiente manera: la revisión bibliográfica ofrece un pequeño resumen de trabajos de otras personas de características similares a este estudio. En el apartado de objetivos, se detalla en más profundidad sobre los objetivos del análisis y en las preguntas que se pretenden responder. En el apartado de la metodología, se explicará todo el proceso para la creación de los modelos: la preparación de los datos, el análisis de las variables y la creación de los modelos. En el apartado de resultados, haremos un resumen conciso sobre lo obtenido en la metodología y mencionaremos los resultados obtenidos. En el apartado de conclusiones, se responderá a las preguntas planteadas en el apartado de objetivos. Los últimos apartados del trabajo corresponden a los anexos y la bibliografía utilizada.

2. REVISIÓN BIBLIOGRÁFICA

Tratar de prevenir el drop out es un tema muy recurrente en la minería de datos educativos, por eso hay gran cantidad de trabajos al respecto en múltiples idiomas. Algunos de los trabajos que sirven para entrar en contexto y no partir de 0 en este tipo de análisis son: Gerben W. Dekker, Mykola Pechenizkiy y Jan M. Vleeshouwers (2009) que estudiaron el drop out de los estudiantes recién titulados de ingeniería eléctrica de la universidad tecnológica de Eindhoven, que se situaba en un 40%. Estos autores indican varios puntos muy interesantes. Llegaron a la conclusión de que los datos preuniversitarios eran muy poco concluyentes a la hora de desarrollar el modelo, también, hacen hincapié en tener en cuenta la matriz de confusión en la hora de predecir los datos, ya que, en esta predicción es mucho más importante detectar a los alumnos que abandonan que a los que no.

Mukesh Kumar (2017) hace una recopilación sobre las variables que se suelen incluir en este tipo de trabajos, que son: genero, edad, nacionalidad, ocupación, media del estudiante, calificaciones(internas), calificaciones(externas), educación y trabajo de los padres, renta familiar, infraestructura, entre otras. También hace una recopilación de los métodos implementados para predecir el drop out: regresión logística, árboles de decisión, JRip, Naive Bayesian Algorithm y muchos otros.

Melvin Vooren (2022) hacen la predicción del drop out con dos muestras de diferentes universidades, la Politecnico di Milano y Vrije Universiteit Amsterdam. De este trabajo podemos ver que la mayoría de los abandonos suceden en los primeros años de la carrera, sobre todo el primero. Además, demuestran que los modelos son muy precisos con la información académica del primer y segundo trimestre, menos precisos con la del primer trimestre y mucho menos con la información preuniversitaria.

Brijesh Kumar Baradwaj (2011) demuestran la efectividad de diferentes métodos que emplean árboles de decisión para identificar a aquellos alumnos en riego de abandonar o que requieran una atención especial. También se menciona la importancia del rendimiento académico como un factor de éxito.

El resultado de estos trabajos servirá para tener una base en nuestro estudio. Estas investigaciones previas han demostrado la efectividad de diversos métodos en la identificación de estudiantes en riesgo de abandono, así como su construcción y evaluación. Es importante tener en cuenta estos hallazgos en nuestro trabajo y aprovecharlos para comprender y la lograr predecir con éxito el abandono estudiantil en el grado de ADE de la UA.

3. OBJETIVOS

El objetivo principal del trabajo es predecir si un alumno de ADE en la UA va a abandonar la titulación tras cursar el primer año del grado para poder actuar en consecuencia.

Una vez resuelto el objetivo principal podremos dar respuesta a las siguientes cuestiones:

Identificar factores de riesgo ¿Qué variables son las más influyentes a la hora de tomar la decisión de abandonar la carrera?

Identificar patrones o tendencias ¿Existen patrones o tendencias comunes entre los alumnos que abandonan el grado?

Elaborar un modelo de predicción ¿Cuál es el modelo de predicción más apropiado para este objetivo?

Mejorar las tasas de retención ¿Qué medidas puede implementar la universidad para reducir el alto abandono en el grado de ADE?

Evaluar las medidas implementadas ¿Cómo podemos medir el éxito de las medidas implementas para reducir el abandono?

Para responder estas preguntas, tendremos que llevar a cabo una serie de procesos:

- Estudiar los datos y transformarlos.
- Analizar el comportamiento de las variables de manera individual y en relación con el abandono.
- Estimar diferentes modelos de predicción.
- Comparar los diferentes modelos.

4. METODOLOGÍA

Para el analizar los datos y crear los modelos de predicción utilizaré el lenguaje de programación R.

4.1. Datos

4.1.1. Extracción de los datos

Los datos comprenden desde el curso 2010/11 al 2019/20.

Hay que apuntar que los datos están anonimizados para proteger la identidad de cada individuo que aparece en los registros.

En el anexo se puede encontrar una breve descripción de algunas de las variables de los ficheros.

4.1.2. Exploración de los datos

En este apartado hablaremos de todo el proceso de la preparación de los datos.

Debemos considerar si los datos están listos para trabajar con ellos o por el contrario hay que necesitan alguna transformación o limpieza.

Lo que observamos en todos es que todas las variables están definidas como <chr> (caracteres), con este formato en R no se puede trabajar bien, por lo que el primer paso es transformar todas a factor.

Modificaremos las variables, que hace referencia al curso, las cuales están estructuradas de la siguiente forma: 20AA-AA, este formato nos puede dificultar trabajar más adelante con la variable, así que la transformaremos a un valor numérico siendo este 20AA, por ejemplo, el curso 2013-14, estará representado como 2013.

Debido a la naturaleza anónima de los datos, transformaremos las variables referidas a países, en un factor con dos categorías: España y extranjero. Esta transformación elimina a los diferentes países de nuestra base de datos, asumiendo así una ligera pérdida de información.

4.1.3. Integración, manipulación y limpieza de los datos

Lo primero que nos tenemos que plantear en este paso es que necesitamos para lograr nuestro objetivo de predecir el abandono (drop out).

Para obtener el drop out nos fijaremos en cual es el último año cursado del estudiante, si este coincide con el año en el que se matriculó por primera vez del grado el drop out tendrá un valor de TRUE, en caso contrario de FASLE. Por esta norma, cabe mencionar que los alumnos que comiencen a estudiar en el curso 2019-20 (último curso del que tenemos información) tendrían un 100% de abandono, por lo que para solucionar este problema eliminaremos de nuestra muestra a esta generación.

Por unos motivos muy similares también tendríamos que eliminar a los alumnos del curso 2010-11 (primer curso del que tenemos información), pero veremos más adelante el motivo de no lo hacerlo.

Como solo necesitamos los datos del primer curso de cada alumno, lo lógico es pensar que estos estarán formados solo por las asignaturas que se imparten en el primer curso de la carrera, aun así, vamos a comprobar que esta hipótesis sea cierta.

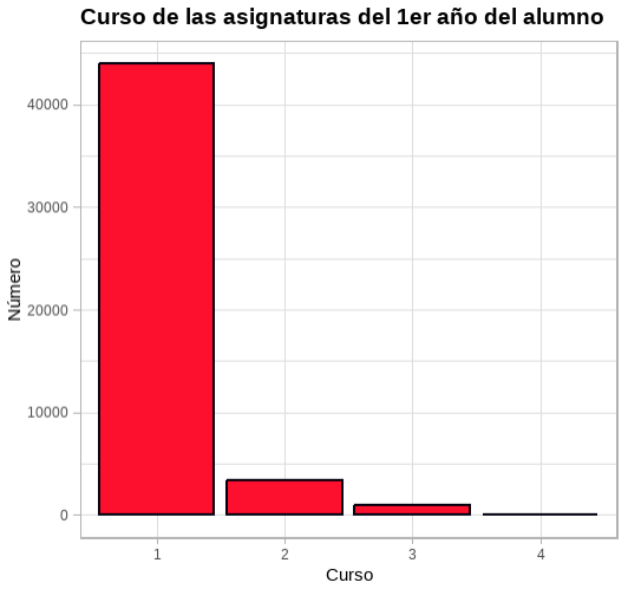


Gráfico 1. Cantidad de asignaturas en cada curso en el primer año del alumno

Como hemos comprobado, la hipótesis no iba mal encaminada pero no es del todo cierta, por lo que no podemos eliminar datos de asignaturas que no sean de 1º de carrera. Como hemos mencionado antes, debido a que el grado es de más de 1 año, puede haber estudiantes que hubiesen acabado la carrera el primer año que contabilizamos los datos, es decir, en el curso 2010-11, esto explicaría por qué hay asignaturas que se cursan "el primer año" del estudiante que no corresponden a este curso y además responderíamos a la cuestión planteada anteriormente sobre los alumnos del curso 2010-11. Para probar esto vamos a contabilizar en el primer año de cada estudiante todas las asignaturas que no son del primer curso.

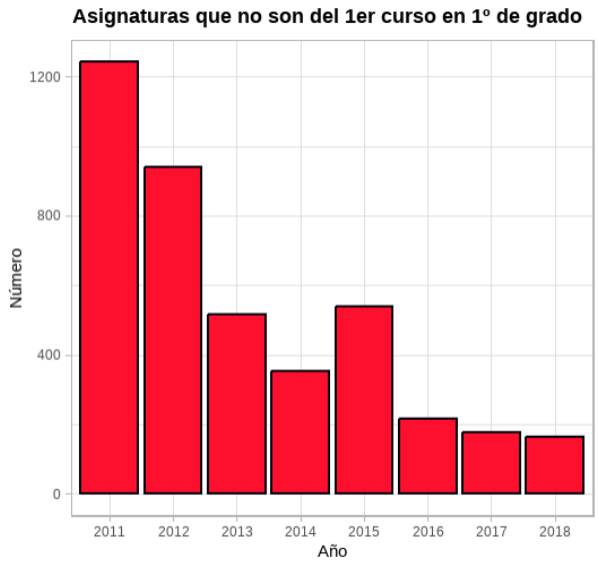


Gráfico 2. Número de asignaturas que no corresponden al primer curso de ADE y que los alumnos cursan el primer año

Como podemos observar, no hay ninguna asignatura de otro curso que no sea de primero en el curso 2010-11. Esto es incluso más extraño, ya que todos los cursos de los que se dispone información tienen alumnos que cursen asignaturas que no corresponden al primer curso de ADE, por esa razón, vamos a comprobar que no haya ningún error con los alumnos de este curso, visualizaremos el número de asignaturas totales por curso.

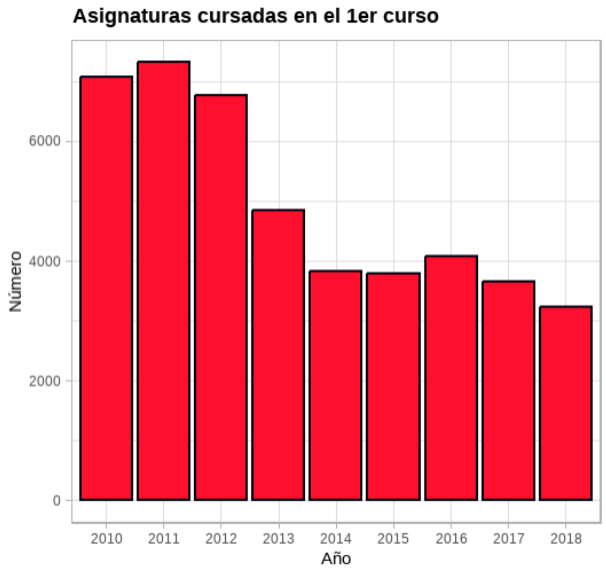


Gráfico 3. Número de asignaturas cursadas en 1º de carrera por año

Como vemos, el curso 2010-11 sale perfectamente, esto nos indica que no hay un grave problema como para excluir esta generación de la muestra, aun así, vamos a eliminar a aquellos alumnos que cursen en su primer año en la UA asignaturas correspondientes al tercer y cuarto curso, ya que, aunque no se trate de un error, estos alumnos es muy posible que correspondan al programa Erasmus+ o a un traslado de expediente de otra universidad, motivos que no son el objeto de este análisis.

En toda la literatura revisada para este trabajo, no se ha encontrado evidencia del uso de datos de los profesores para predecir el abandono o situaciones similares. Además, considerando que la inclusión de estos datos podría aumentar la complejidad de los modelos sin aportar información significativa, se ha decidido no utilizarlos en este análisis.

A priori es lógico pensar que uno de los factores más determinantes del drop out es el rendimiento del alumno. Para ello vamos a diferenciar entre el rendimiento el primer cuatrimestre y entre el segundo cuatrimestre. Mediremos el rendimiento de los alumnos en función de cuantas asignaturas aprueba, cuantas va a la recuperación, a cuantas se presenta, de cuantas se matricula y la nota media.

Crearemos las siguientes variables:

VARIABLES	DESCRIPCIÓN
APROBADO_1C	nº de asignaturas aprobadas en el primer cuatrimestre (sin recuperación)

APROBADO_2C	nº de asignaturas aprobadas en el segundo cuatrimestre (sin recuperación)
APROBADO_FINAL	nº de asignaturas aprobadas en el curso (con recuperación)
RECUPERACION_1C	nº de asignaturas a las que se presenta a la recuperación del primer cuatrimestre
RECUPERACION_2C	nº de asignaturas a las que se presenta a la recuperación del segundo cuatrimestre
PRESENTADO_1C	nº de asignaturas a las que se presenta en el primer cuatrimestre
PRESENTADO_2C	nº de asignaturas a las que se presenta en el segundo cuatrimestre
MATRICULA_1C	nº de asignaturas matriculadas en el primer cuatrimestre
MATRICULA_2C	nº de asignaturas matriculadas en el segundo cuatrimestre
NOTA_1C	nota media obtenida en el primer cuatrimestre
NOTA_2C	nota media obtenida en el segundo cuatrimestre
EDAD	edad del estudiante en el momento que entró a la carrera

Tabla 1. Variables creadas 1.

Como podemos observar, aparte de las mencionadas previamente, he añadido la variable EDAD que hace referencia a la edad del estudiante a la hora de comenzar los estudios, esta variable nos aporta mucho más que la que teníamos en referencia a su edad, que era el año de nacimiento, ya que la edad es una variable que nos permite eliminar la temporalidad de este conjunto de datos.

Tras realizar las modificaciones e integraciones necesarias damos lugar a un fichero donde estará reflejada toda la información y el rendimiento del alumno.

Para finalizar, eliminaremos del modelo las variables que puedan aportar precisión para predecir el abandono en instancias pasadas pero que no aportarían e incluso perjudicarían la predicción de eventos futuros, como el curso académico o el último año de grado del estudiante.

4.2. Análisis exploratorio

4.2.1. Información general

El conjunto de datos no tiene datos nulos a primera vista, también sabemos que las variables son o factores o numéricas.

4.2.2. Análisis de variación

En este apartado vamos a estudiar la variabilidad de las variables que componen el fichero, de esta forma las podremos comprender mejor y formular hipótesis al respecto. Para ellos nos apoyaremos de una gráfica en cada variable para que sea más visual e intuitivo.

Primero de todo vamos a estudiar la variable dependiente. El drop out

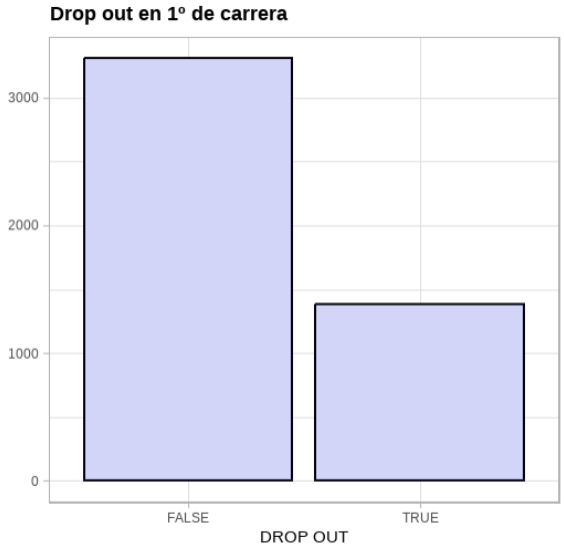


Gráfico 4. Distribución del drop out en 1º de carrera

Podemos observar que el drop out es relativamente alto. Este es exactamente del 29,48%, este dato puede parecer sorprendente, ya que en la introducción hemos visto que el porcentaje era del 37,43% en el curso 2018-19, lo que parece que la cifra obtenida por nosotros es muy buena. Pero vamos a estudiar la evolución del drop out con el paso de los años.

Gráfico 5. Evolución del drop out con los años

Como podemos apreciar, el drop out en primero va aumentando a medida que pasa el tiempo, esto amplifica la magnitud del problema inicial, que, aunque la cifra del 37,43%, siga siendo la misma, todo indica que va a seguir creciendo.

Un dato por puntualizar es que en el gráfico 5. La proporción de drop out en el curso 2018-19 es del 38,94% frente al dato oficial de la UA que es del 37,43%, esto se debe a los pequeños errores que se van asumiendo a medida que se prepara el modelo.

A continuación, iremos analizando las variables independientes y sacando conclusiones que puedan ayudar más adelante.

1. Pais de nacionalidad.

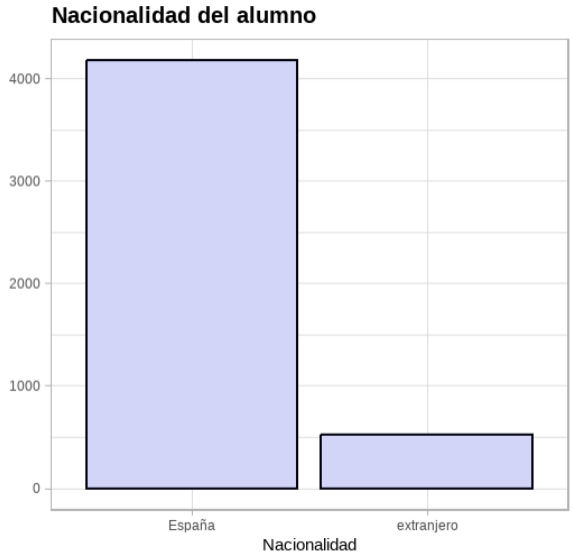


Gráfico 6. Variación de la nacionalidad del alumno

Como podíamos intuir, la gran mayoría de los alumnos son españoles, el 88,80%, igualmente me ha llamado la atención que más del 10% sean extranjeros, podría ser que el hecho de ser extranjero suponga una dificultad más a la hora de cursar el grado.

2. Sexo.

A algunas variables les iré cambiando el nombre de sus categorías para que se entiendan mejor, en este caso H es hombre y M es mujer.

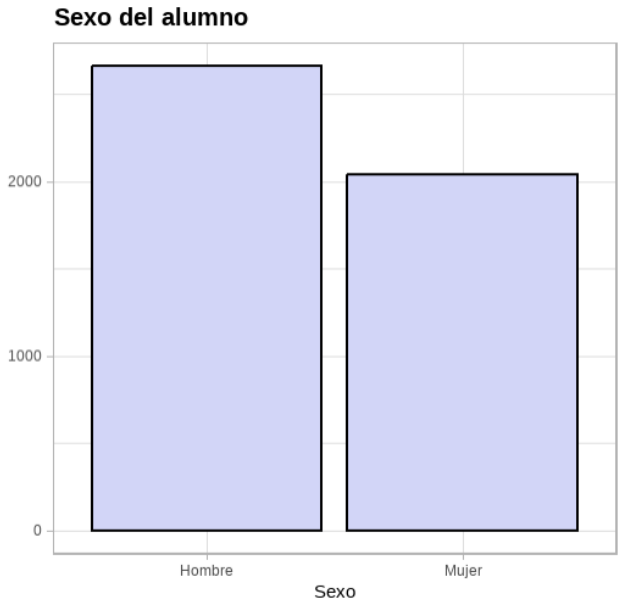


Gráfico 7. Sexo del alumno

El 56,60% de los que estudian ADE son hombres, estudiaremos si el sexo del estudiante es un factor de riesgo, aunque a priori, lo dudo.

3. Dedicación del estudio.

Esta variable contiene 3 valores nulos, para tratar con ellos los introducimos en la categoría mayoritaria (a tiempo completo) asumiendo el error correspondiente.

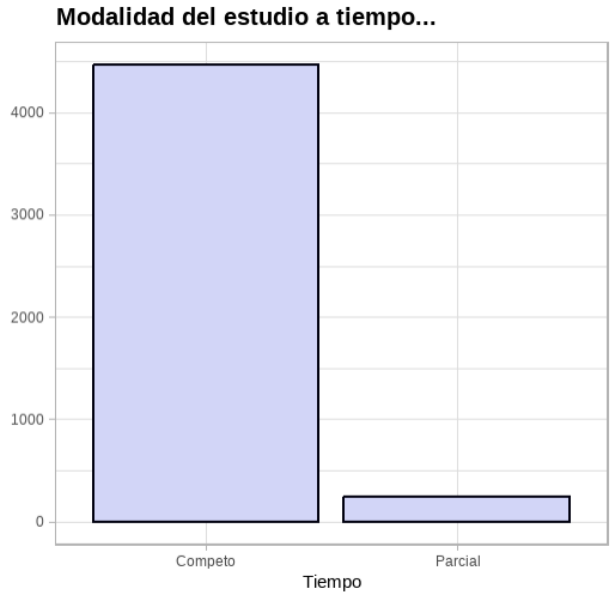


Gráfico 8. Variación de la modalidad del estudio

Observamos que la gran mayoría realiza sus estudios en una modalidad de tiempo completo, lo cual no supone una sorpresa, en este caso analizaremos si estudiar a tiempo parcial es un factor de riesgo para el drop out, estudiar a tiempo parcial implica que probablemente no se disponga de mucho tiempo y/o que se tengan otras cosas a lo largo del día, como trabajo, cuidar de otras personas…

4. Ocupación del estudiante.

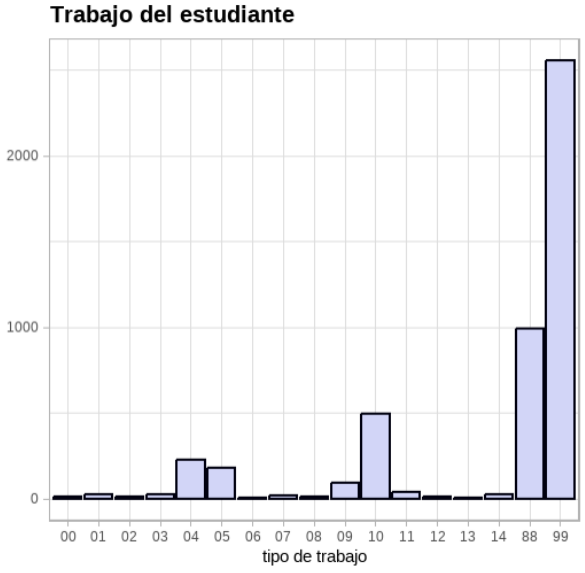


Gráfico 9. Variación de la ocupación del estudiante

El dato '99' significa que es un dato nulo, y como podemos ver, es la categoría mayoritaria, por lo que vamos a eliminar esta variable de nuestro análisis.

5. Familia numerosa.

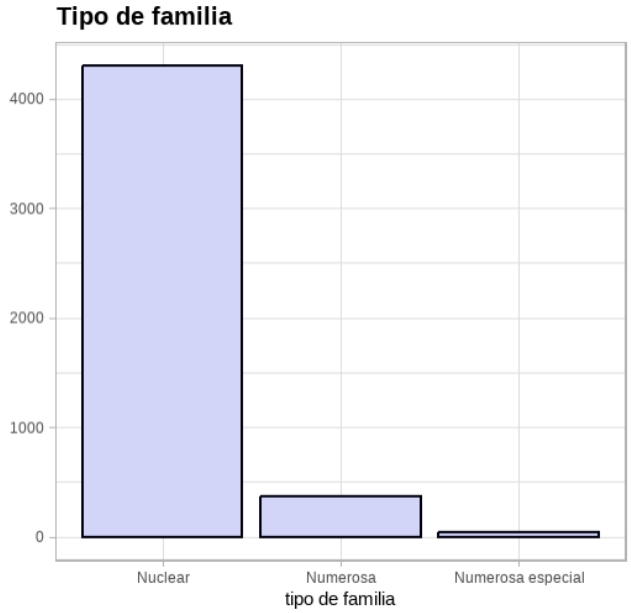


Gráfico 10. Variación del tipo de familia

Podemos ver que la categoría más numerosa es la familia nuclear, por lo que vamos a estudiar si pertenecer a una familia numerosa o a una numerosa espacial son factores de riesgo, a priori podrían ser, ya que, este tipo de familias tienen descuentos en a la hora de la matriculación, por lo que el peso de rentabilizar la inversión disminuiría y podrían sentirse menos "obligados" a rentabilizarla.

6. Nivel de estudios de los padres.

En este apartado, nos hemos encontrado con el problema de que en el nivel de estudio del padre hay 310 datos nulos y en el nivel de estudios de la madre hay 236 datos nulos. Para tratarlos, vamos a transformarlos proporcionalmente en las demás categorías asumiendo el nuevo error.

En estas variables también hemos juntado dos categorías la '1' y la '2', que significan respectivamente, analfabeto y sin estudios, en una categoría llamada ninguno que hace referencia a que no se tiene ningún estudio.

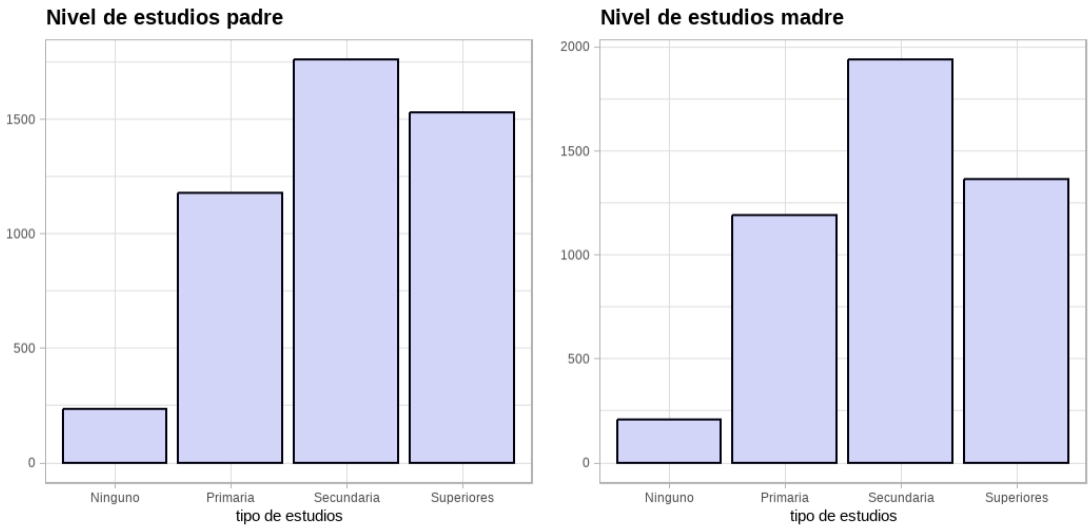


Gráfico 11. Variaciones del nivel de estudios del padre y de la madre

Podemos observar que las categorías de ningún estudio y de primaria que son prácticamente iguales. Se puede apreciar que hay más madres con estudios secundarios (178) que padres y hay más padres con estudios universitarios (166) que madres. Suponiendo que si alguien tiene estudios universitarios tiene los secundarios, aproximadamente el 70% de los padres y madres tienen estudios secundarios, mientras que aproximadamente el 30% tiene estudios universitarios.

Para tener una mejor referencia de esta variable, vamos a crear otra variable que explique cuales son los estudios más altos que hay en cada familia (padre y madre), de esta forma sabemos cuáles son los estudios más altos de los que se puede influenciar el alumno.

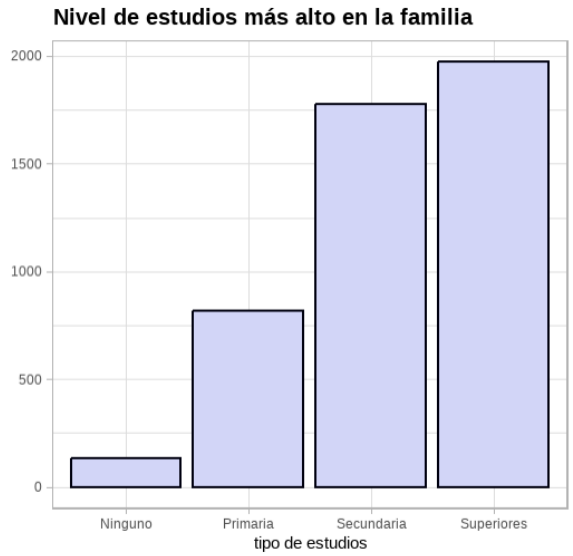


Gráfico 12. Variación del máximo nivel de estudios de los padres

Como podemos observar hay un gran cambio, los dos estudios más altos, secundaria y superiores, han aumentado frente a los estudios más bajos. Otro cambio importante, es que antes los estudios secundarios eran los más comunes, y ahora si medimos por "referente paternal" son los superiores.

El 41% de los estudiantes tienen algún padre con estudios superiores, y el 79,78% tiene como mínimo un padre que ha estudiado secundaria.

El análisis aquí será ver si no tener padres con estudios afecta a la decisión del drop out.

7. Ocupación de los padres.

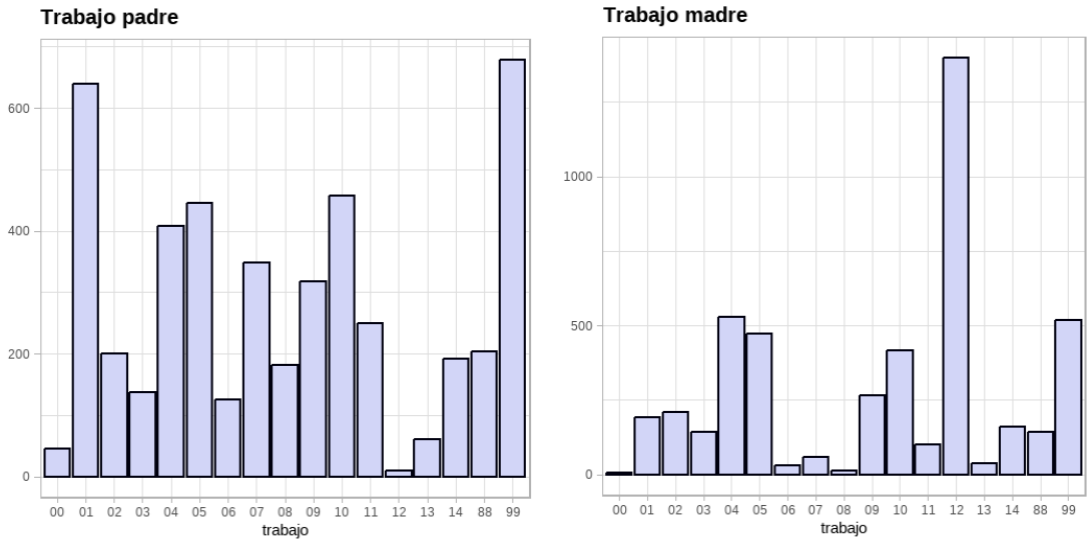


Gráfico 13. Variación de la ocupación de los padres

Podemos ver que no como en los estudios, aquí sí que hay grandes diferencias entre los padres y las madres.

Lo que más llama la atención son las categorías 01 y 99 en la ocupación de los padres, que son respectivamente 'Directores y gerentes' y 'No consta' y la categoría 12 en la ocupación de las madres 'Ama de casa', son datos que merecería la pena analizar, pero al haber tantos datos nulos ('No consta'), hace que tengamos que prescindir de esta variable.

Para ver que significa cada número en las categorías del gráfico, se puede consultar en el anexo.

8. Forma de admisión.

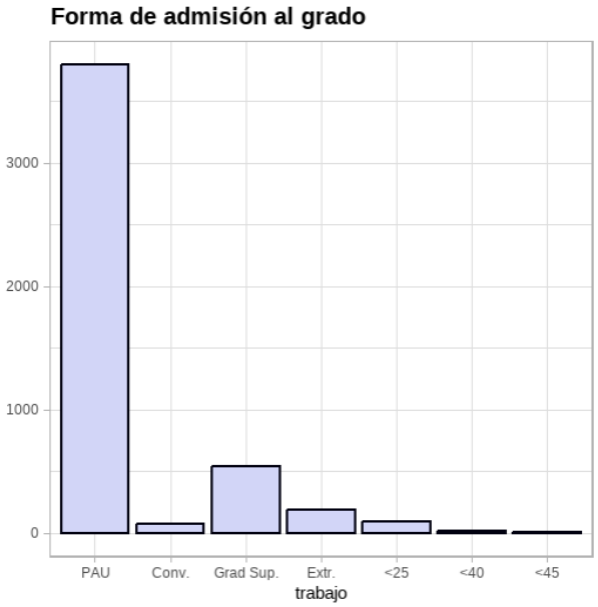


Gráfico 14. Variación de la forma de admisión al grado

Como podíamos intuir la forma de admisión más común es a través de la Prueba de acceso a la universidad (PAU). Estudiaremos sobre todo si acceder de alguna otra forma tiene alguna relevancia a la hora de decidir el drop out.

Los significados a las categorías los podréis encontrar en el anexo.

9. Estudio de acceso al grado.

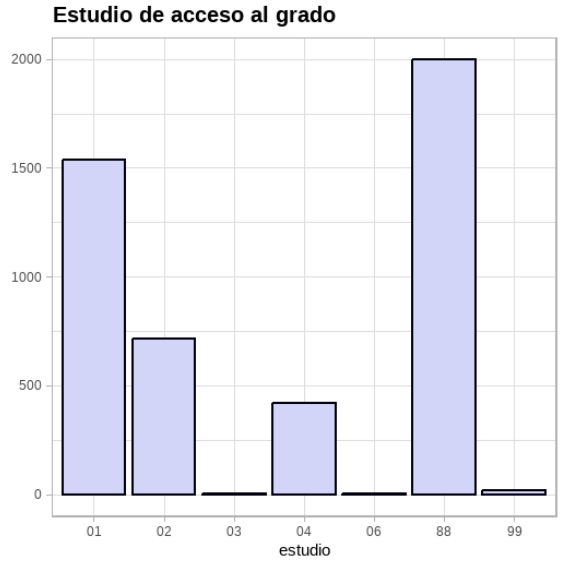


Gráfico 15. Variación del estudio de acceso al grado

Sabiendo que la categoría '88' es valor nulo, vamos a desestimar esta variable para el modelo.

10. Nuevo en el sistema universitario español.

Gráfico 16. variación de si es nuevo en el Sistema Universitario Español (SUE)

Esta gráfica puede sorprender ya que el 33,29% de los alumnos no son nuevos en el sistema universitario español, vamos a estudiar si esta variable puede tener peso en el abandono.

11. Municipio centro secundaria.

En esta variable se nos representa el municipio al que pertenecía el centro donde estudió el alumno. Este está representado por su código postal, debido a que hay muchísimas categorías, vamos a agrupar las variables por provincias. Para ello tenemos que saber cómo funciona el sistema de código postal en España.

El código postal está formado por 5 dígitos, los 2 primeros hacen referencia a la provincia a la que pertenece, por lo que, podemos diferenciar a qué provincia pertenece cada alumno y estudiar si tener que mudarte a Alicante para estudiar es un factor de riesgo para el abandono. Sabemos que en nuestros datos la gente de fuera de España está en la categoría 88888 o 99999.

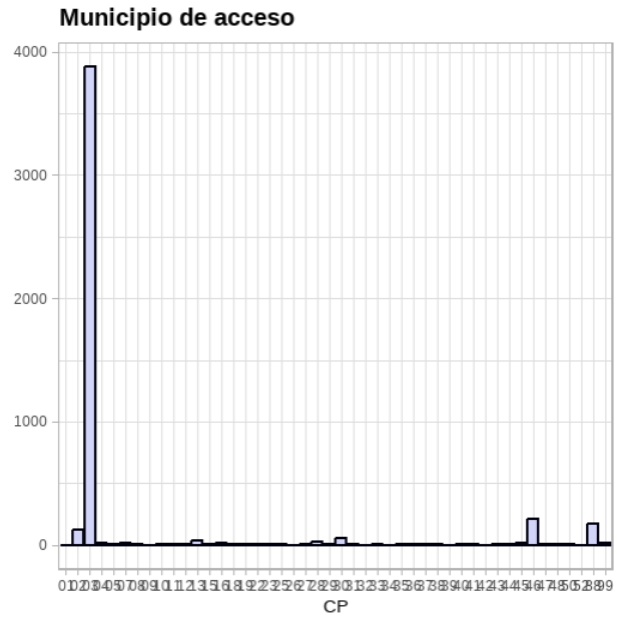


Gráfico 17. Variación del municipio donde se realizó el acceso

Podemos ver lo que podíamos intuir, que la gran mayoría de los que estudian, son de Alicante (código postal 03), aun así, estos datos no son muy claros al haber tantas provincias, por lo que vamos a diferenciar entre ser de alicante o no.

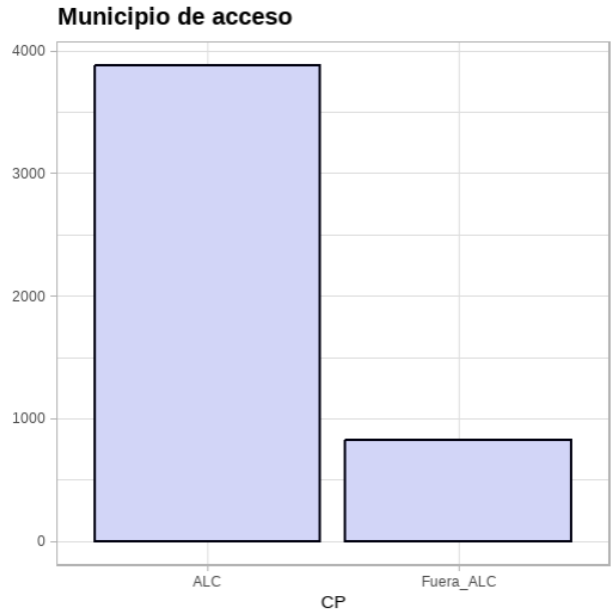


Gráfico 18. Variación de si municipio donde se realizó el acceso es Alicante

Ahora se ve mucho más claro, vamos a estudiar si tener que mudarte a una ciudad nueva es un factor importante en el drop out.

12. Naturaleza del centro donde se realizó secundaria.

En esta variable había 20 datos nulos, para tratar con ellos les asigne el valor de la categoría mayoritaria asumiendo el posible error.

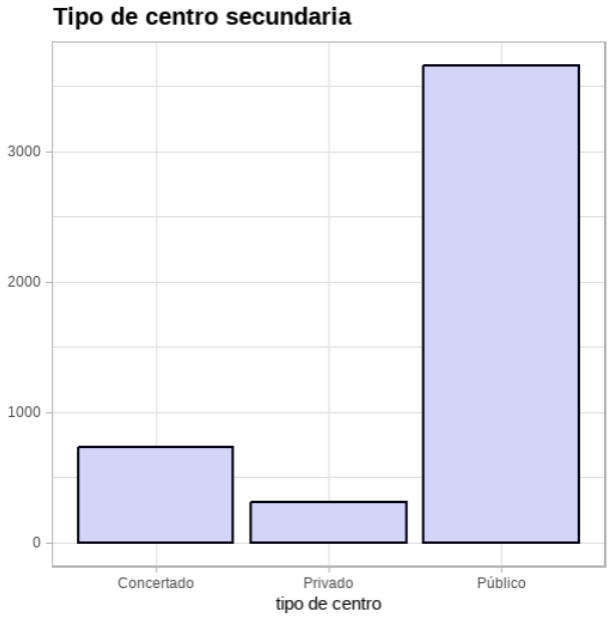


Gráfico 19. variación de la naturaleza del centro donde se estudió secundaria

Podemos observar, que la gran mayoría de los alumnos estudiaron la educación secundaria en un centro público, tendremos que estudiar si la naturaleza del centro donde estudió el alumno es significativa a la hora del drop out.

13. Nota de admisión.

En esta categoría tenemos 70 datos nulos. Para tratar con ellos, hemos generado aleatoriamente las notas en un rango que comprende desde el primer cuartil hasta el tercer cuartil.

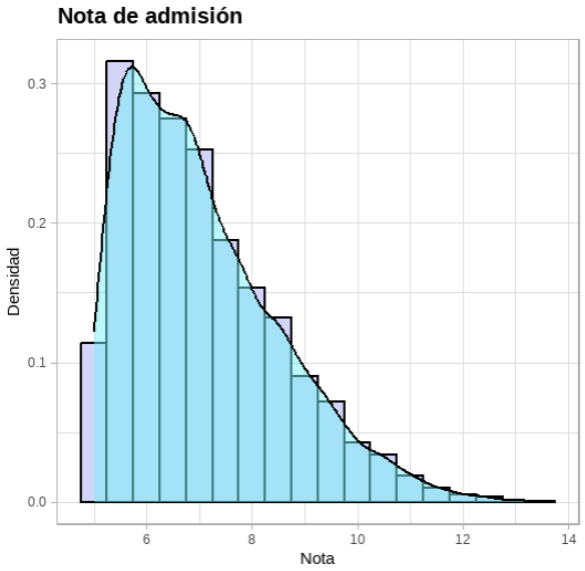


Gráfico 20. Distribución de la nota de admisión a la universidad

Como podemos ver, si tenemos en cuenta que la nota de admisión es sobre 14, es generalmente baja. El 50% central de los datos se encuentra entre 5,87 y 7,98 (primer y tercer cuartil). Vamos a estudiar si la nota de admisión es una característica importante a la hora de predecir el drop out.

14. Edad.

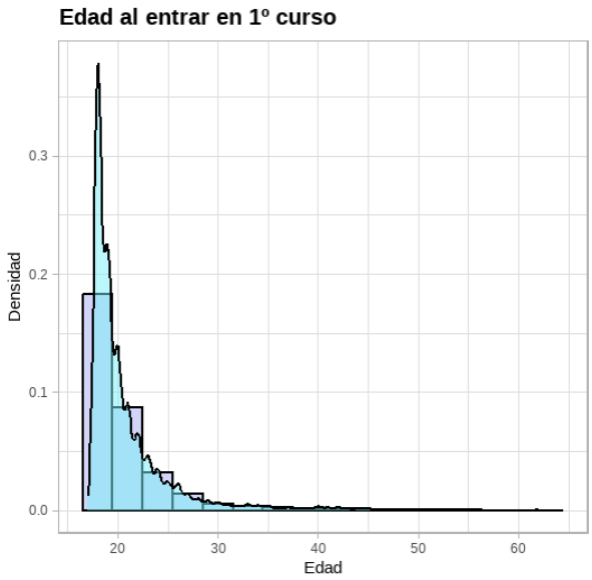


Gráfico 21. Distribución de la edad al entrar en 1º

Como era de esperar, la gran mayoría de los alumnos empiezan primero de ADE con una edad muy temprana. Vamos a ver si esta variable es un factor de riesgo, ya que suponemos que a cuanto más mayor se es, más responsabilidades se tiene, y esto puede ser una razón de peso para el drop out de la carrera.

15. Matriculación en cada cuatrimestre.

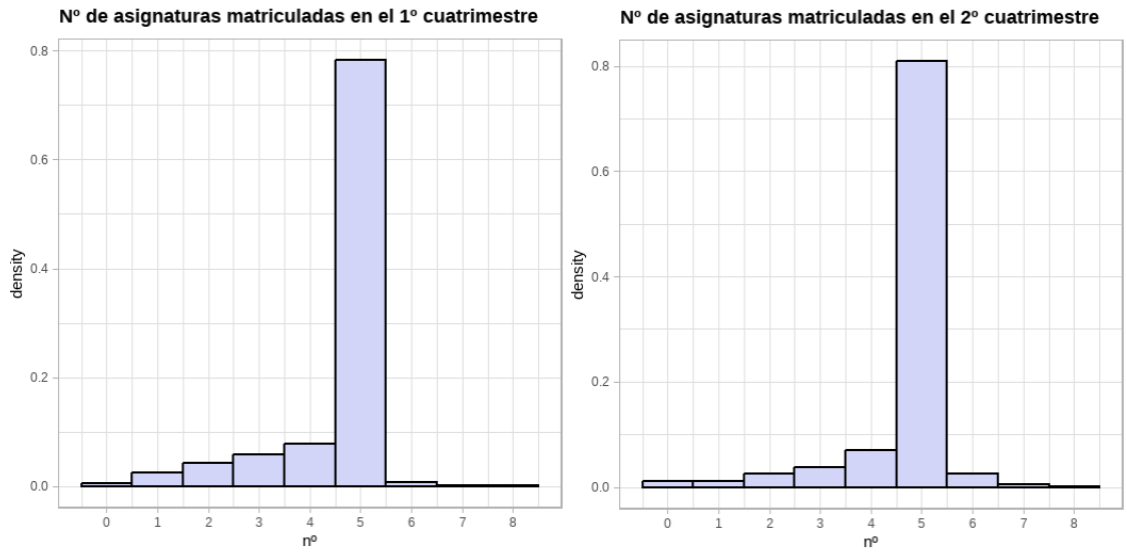


Gráfico 22. Distribución del número de asignaturas matriculadas por alumno en cada cuatrimestre

Podemos observar que no hay casi ninguna diferencia entre los cuatrimestres, y que la gran mayoría de los alumnos de matrícula de 5 asignaturas cada cuatrimestre, lo que corresponde al número de asignaturas que tiene el primer curso. La matrícula se suele hacer antes de comenzar el grado, por lo que esta variable podría mostrar las expectativas del rendimiento del alumno.

16. Asignaturas a las que el alumno se presenta.

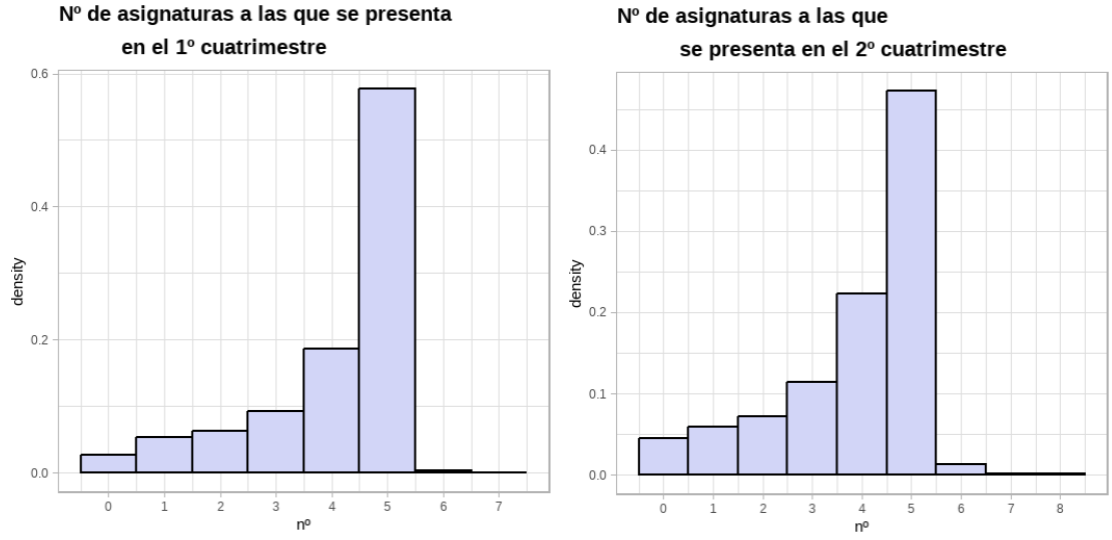


Gráfico 23. Distribución de las asignaturas a las que se presenta cada alumno por cuatrimestre

Lo primero que podemos observar es que son muy similares entre sí, pero si nos fijamos, la densidad de alumnos que se presentan a menos en 5 asignaturas crece en el segundo cuatrimestre y los que se presentan a 5 baja.

Esto puede ser debido a varias razones, entre las que se encuentran la dificultad de las asignaturas, un mal primer cuatrimestre etc.

17. Nota media del alumno en cada cuatrimestre.

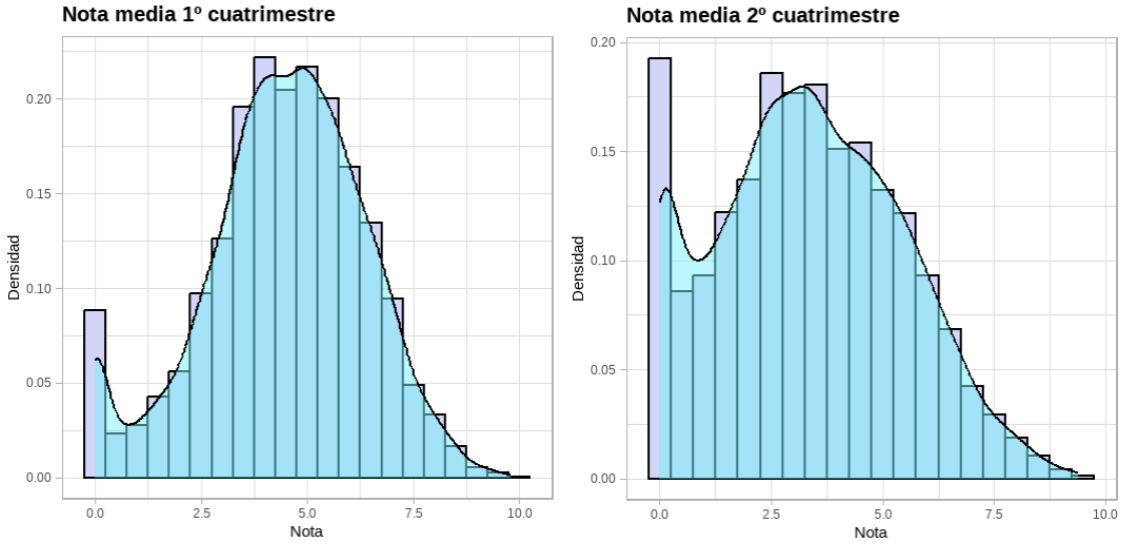


Gráfico 24. Distribución de la nota media de cada cuatrimestre

Esta comparativa ya llama mucho la atención, vemos que el rendimiento académico en el segundo cuatrimestre es mucho peor que en el primero.

En el primer cuatrimestre podemos ver que las notas siguen una distribución normal con una anomalía, ya que muchos alumnos tienen de nota media un 0. Podemos asumir que alguien que tiene una nota media de 0 es por que prácticamente ya ha abandonado el grado.

En el segundo cuatrimestre, se observa que la nota media no sigue una distribución normal. La anomalía mencionada en la distribución del primer cuatrimestre se duplica. Además, se ha registrado una disminución en la nota media, pasando de 4,45 a 3,33.

Este cambio tan brusco a notas más bajas puede deberse a que el segundo cuatrimestre es más difícil o a que muchos alumnos ya han abandonado la carrera después del primer cuatrimestre.

18. Asignaturas aprobadas por cuatrimestre sin contar la recuperación.

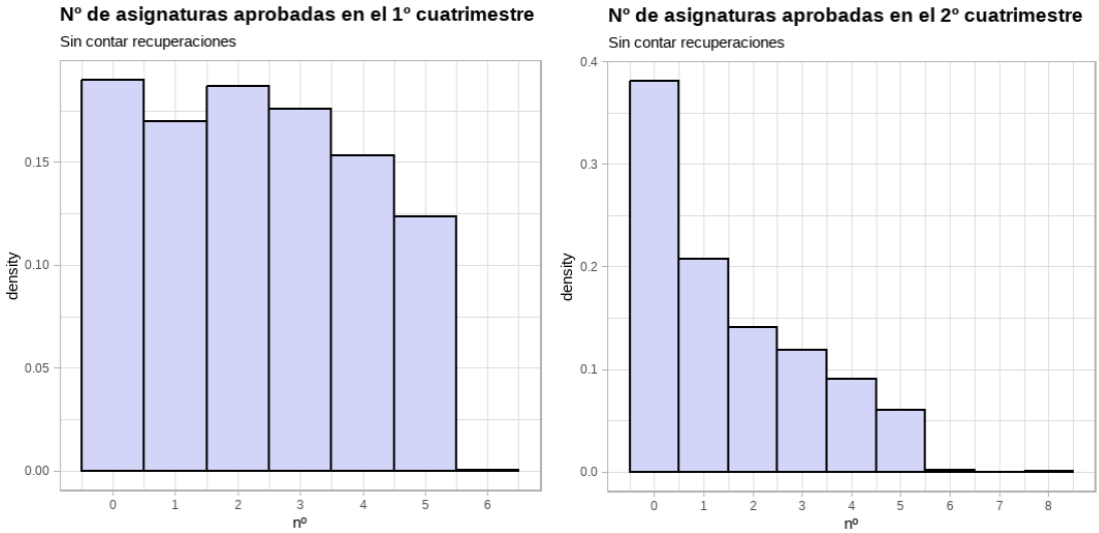


Gráfico 25. distribución de las asignaturas aprobadas en cada cuatrimestre

Podemos darnos cuenta a simple vista del gran cambio del primer al segundo cuatrimestre, en el primero vemos que la distribución es parecida entre todas las notas, apreciándose una pendiente negativa a medida que el alumno aprueba más, en cambio, en el segundo cuatrimestre, el número de personas que no aprueban ninguna asignatura es increíblemente mayor que las demás, existiendo también una pendiente negativa más acentuada a medida que se aprueban más asignaturas. Cabe destacar que estas graficas no tienen en cuenta la nota de la recuperación.

Esta gráfica nos refuerza las teorías anteriores de la dificultad del segundo cuatrimestre o de la decisión de abandonar el grado tras el primer cuatrimestre.

A continuación, les mostraré una gráfica donde podremos ver el número total de asignaturas que aprueba un alumno contando las recuperaciones.

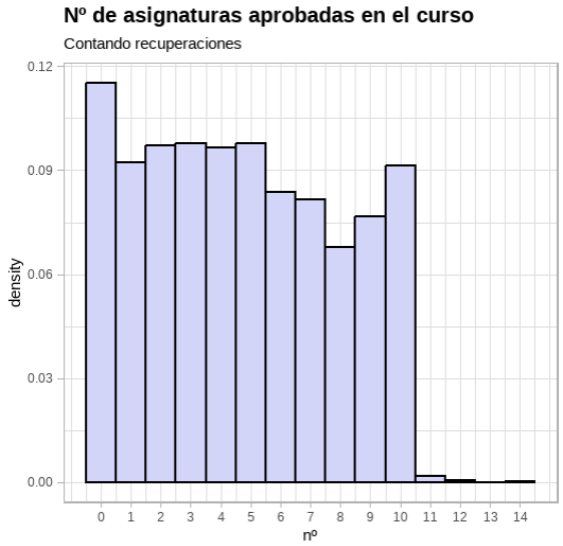


Gráfico 26. Distribución de las asignaturas aprobadas en el primer año de carrera

Podemos ver que el gráfico muestra una estructura similar a la de asignaturas aprobadas en el primer cuatrimestre, Intuimos que la gente que ha aprobado 0, 1 ó 2 asignaturas es la que finalmente se deje la carrera.

19. Recuperaciones a las que se presenta el alumno de cada cuatrimestre.

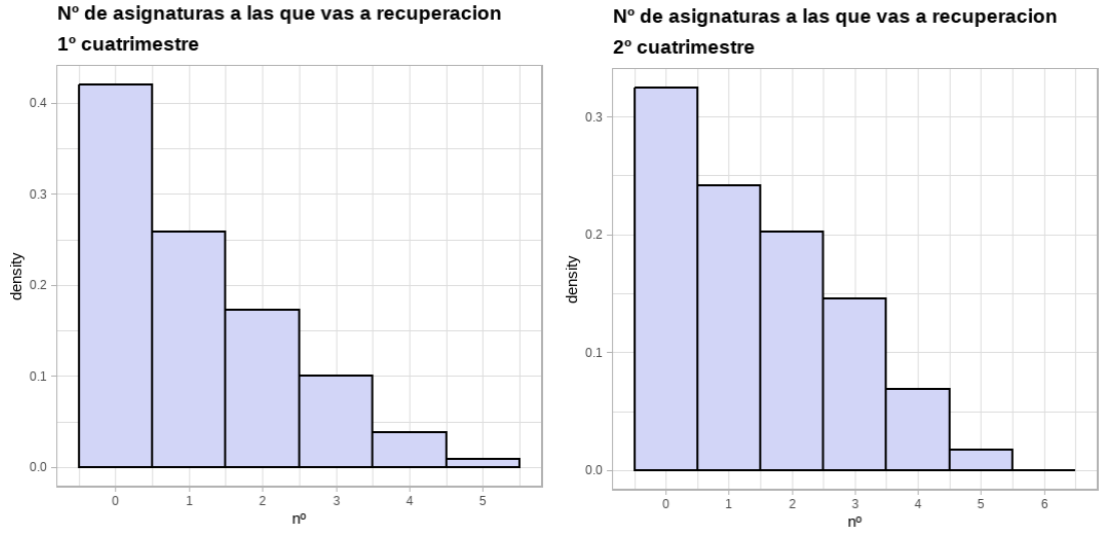


Gráfico 27. Asignaturas a las que el alumno se presenta para recuperar

Lo primero que vemos es la similitud de los gráficos. Si se presenta a recuperaciones es un indicador de que aún no ha tirado la toalla, si no se presenta, pueden ser dos opciones, que el alumno haya aprobado todas las asignaturas o que el alumno haya abandonado la carrera.

Hay varios aspectos a analizar, vemos que son muy similares los gráficos, pero en el segundo cuatrimestre hay más densidad de gente que va a recuperar alguna asignatura, esto refuerza la hipótesis de que el segundo cuatrimestre es más difícil o les cuesta más a los alumnos. La presencia de estudiantes que asisten a la recuperación sugiere que no han abandonado aún el grado, ya que su participación en este proceso indica que han intentado superar la asignatura en la convocatoria ordinaria sin éxito.

Si volvemos al gráfico de aprobados por cuatrimestre, si todos los alumnos que hubieran suspendido alguna asignatura fuesen a la recuperación, el gráfico de recuperaciones sería el inverso al de aprobados, pero como podemos apreciar, no lo es.

La similitud de los dos gráficos nos puede recordar a la similitud del número de matrículas, esto es debido al mismo motivo, las recuperaciones se hacen al final del curso, cuando la decisión está, por lo que parece, prácticamente tomada.

Debido a que en esta variable, ir a 0 recuperaciones, significa que probablemente se ha dejado la carrera, pero también significa que el rendimiento ha sido alto vamos a desestimar estas variables para el modelo final.

4.2.3. Análisis de covariación

En este apartado vamos a comprobar cómo se comportan las variables entre sí, sobre todo cómo se comportan con la variable dependiente (DROPOUT). Para ello

Esto nos servirá para encontrar relaciones entre las variables y para construir un mejor modelo de predicción.

Empezaremos analizando las variables personales de cada alumno y terminaremos con las relacionadas con el rendimiento del estudiante.

1. Nacionalidad.

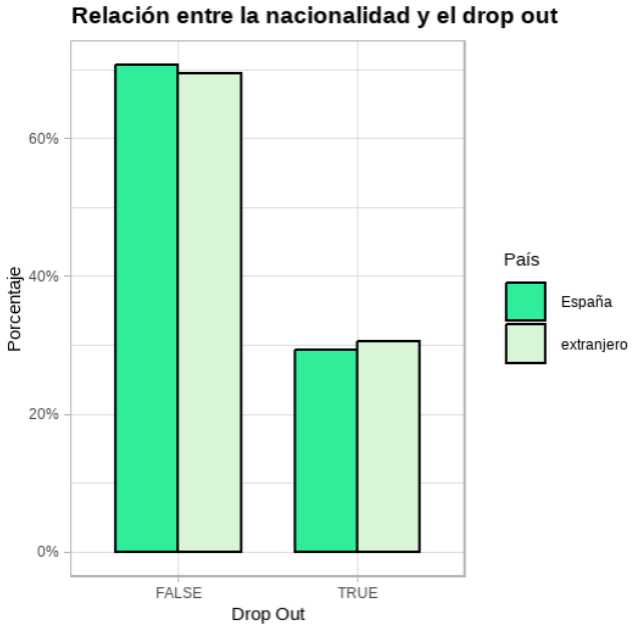


Gráfico 28. Relación entre la nacionalidad y el drop out

Podemos observar que el perfil extranjero suele abandonar el grado en primero más que el nacional, pero la diferencia es mínima, por lo que es muy poco probable que la nacionalidad sea un factor de drop out.

2. Sexo.

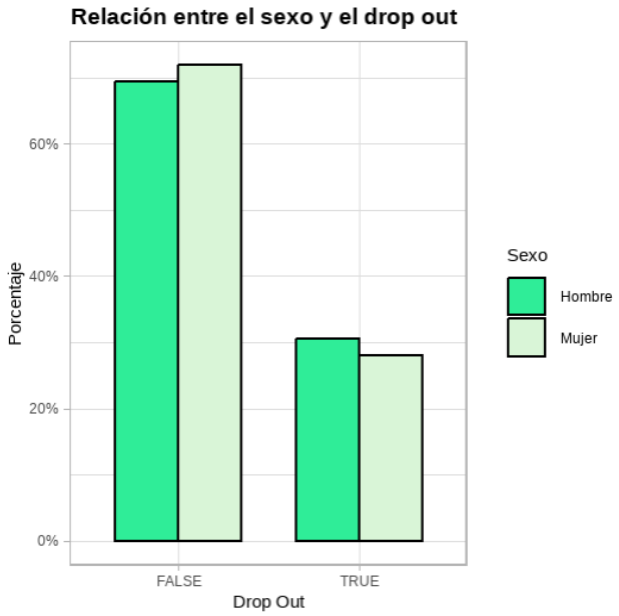


Gráfico 29. Relación entre el sexo del alumno y el drop out

Gráfica parecida a la anterior, pese a que los alumnos cuyo sexo es hombre suelen abandonar en mayor proporción que las mujeres, no parece la suficiente como para afirmar que hay relación entre el sexo del estudiante y el abandono prematuro.

3. Edad.

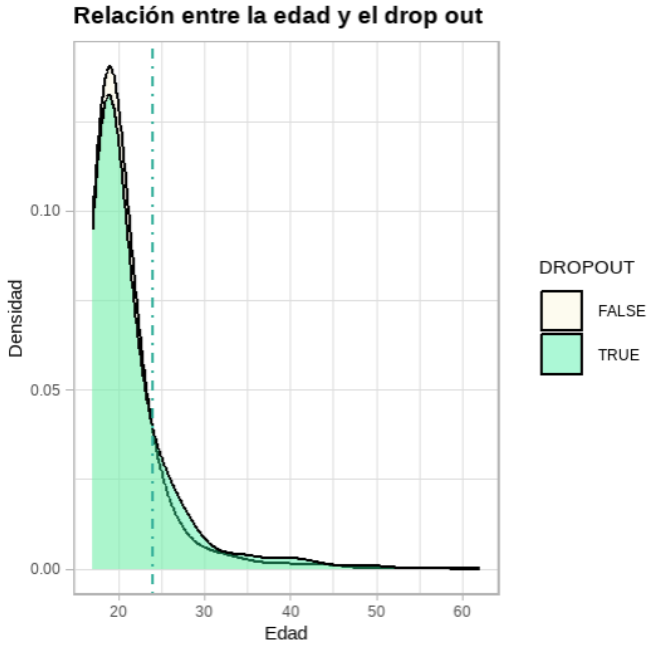


Gráfico 30. Relación entre la edad del alumno y del drop out

En esta gráfica que relaciona la edad con el drop out se puede observar que los alumnos jóvenes abandonan menos la carrera que los de edades a partir de los 24 años.

No se podría afirmar que es una variable es determinante para el drop out, pero hay que tenerla en cuenta para la realización del modelo.

4. Dedicación del estudio.

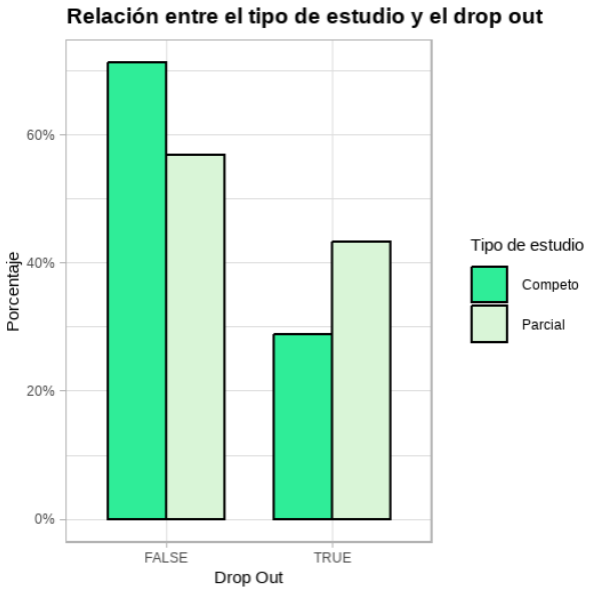


Gráfico 31. Relación entre el tipo de estudios y el drop out

Podemos ver que claramente un alumno que estudie a tiempo parcial es más probable que acabe abandonando el grado antes que uno que haga la modalidad a tiempo completo.

Esto respalda la hipótesis anterior de que estudiar a tiempo parcial puede estar asociado con tener menos tiempo disponible y/o tener responsabilidades adicionales durante el día, como trabajo o cuidado de otras personas.

5. Familia numerosa.

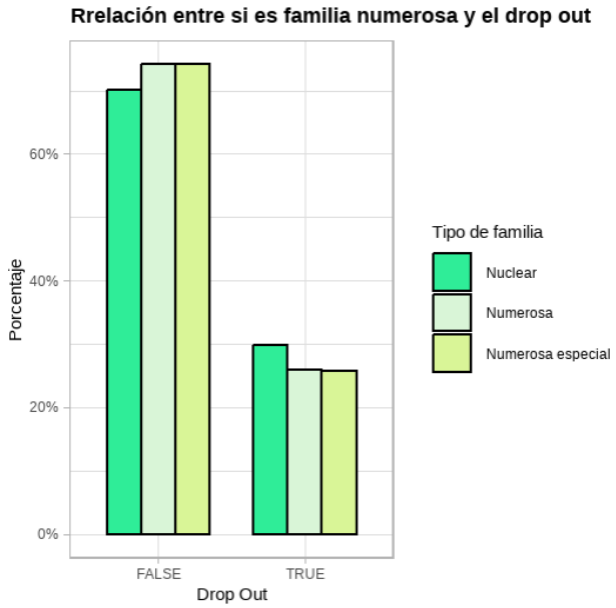


Gráfico 32. Relación entre el tipo de familia y el drop out

Lo que refleja el gráfico desmonta la hipótesis planteada anteriormente, ya que pertenecer a una familia numerosa o numerosa especial parece tener un efecto de retención. Esto puede ser debido a la misma causa de la hipótesis anterior, pero desde otro punto de vista. Se podría argumentar que un estudiante que tiene un bajo rendimiento académico y suspende algunas asignaturas, se ve obligado a volver a matricularse de nuevo de estas, al no beneficiarse del descuento, le puede suponer muy costosas las segundas matriculas, mientras que al de la familia numerosa le saldría más económico.

Pese a lo anterior, tampoco parece que la variable de pertenecer a una familia numerosa será un factor muy importante a la hora de medir el abandono.

6. Nivel de estudio de los padres.

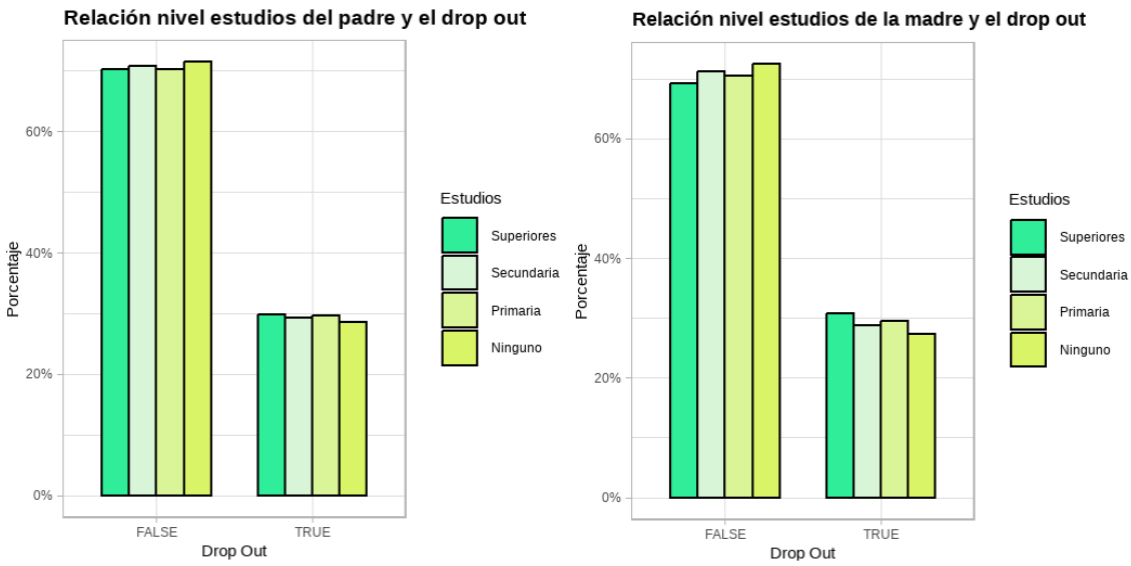


Gráfico 33. Relación entre los estudios de los padres y el drop out

Podemos observar que la relación entre los estudios de los padres y el abandono es casi, si no, prácticamente nula. Llama la atención que no cómo se podría pensar, los que tienen en casa una madre o padre con estudios superiores son los más probables de abandonar el grado en comparación con el resto.

Ahora vamos a analizar la otra variable creada anteriormente del nivel de estudios máximo en una familia.

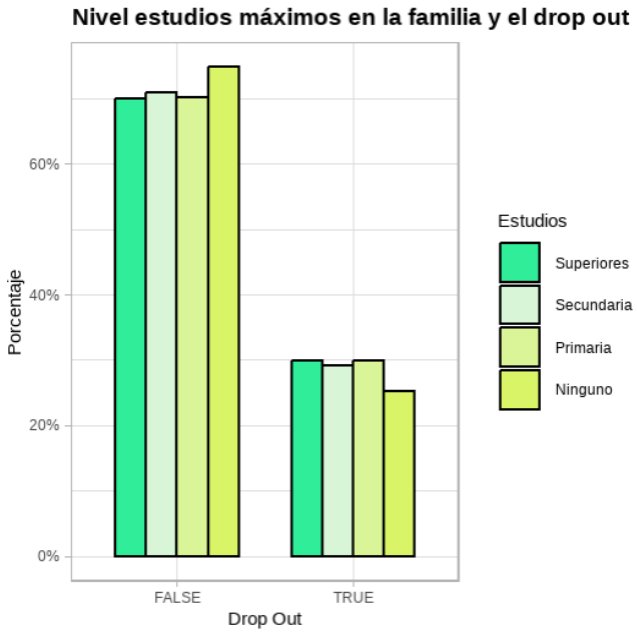


Gráfico 34. Relación entre los mayores estudios en la familia y el drop out

Ahora, podemos observar que los que menos probabilidad hay de que se dejen el grado son los alumnos cuyos padres, ninguno, tenga primaria, también hay que mencionar que esta es la categoría menos numerosa de esta variable y que se podría tratar de alguna excepción, el resto de las variables se han alineado.

7. Forma de admisión.

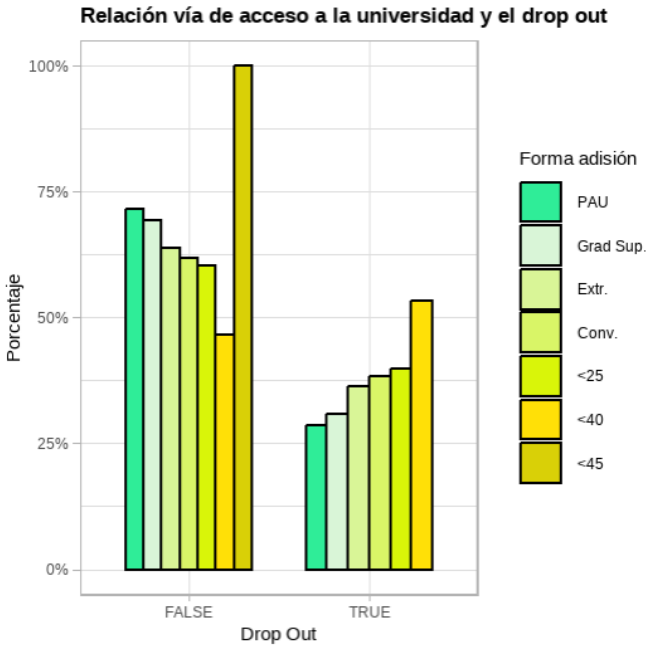


Gráfico 35. Relación entre la vía de acceso a la universidad y el drop out

Aquí parece haber una clara relación entre la vía de acceso y el drop out de la carrera, siendo cuanto menos común la vía de acceso más probable el drop out. Esto tiene una excepción en los alumnos mayores de 45 años, que, al parecer, ninguno abandonó la carrera.

Esta variable muy probablemente sea determinante en nuestro modelo.

8. Ser nuevo en el Sistema Universitario Español (SUE).

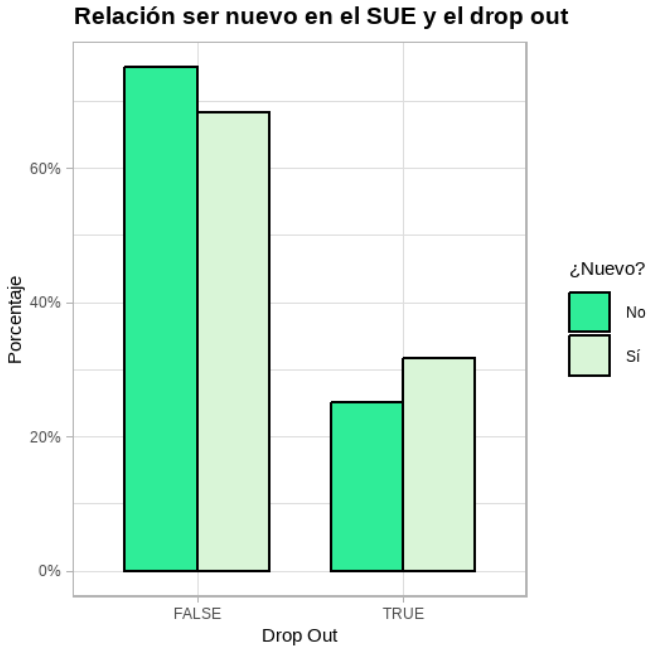


Gráfico 36. Relación entre ser nuevo en el Sistema Universitario Español (SUE) y el drop out

Como podemos apreciar, ser nuevo en el SUE es un factor de drop out, esto puede ser debido a que al ser la primera vez en este sistema no se acaben de adaptar a él y abandonen por ese motivo, por el otro lado, si no eres nuevo en el SUE puede ser que el alumno haya abandonado o terminado otro grado y decida comenzar uno nuevo para el cual ha tenido más tiempo para darse cuenta de que le gustaría estudiar.

9. Municipio de acceso.

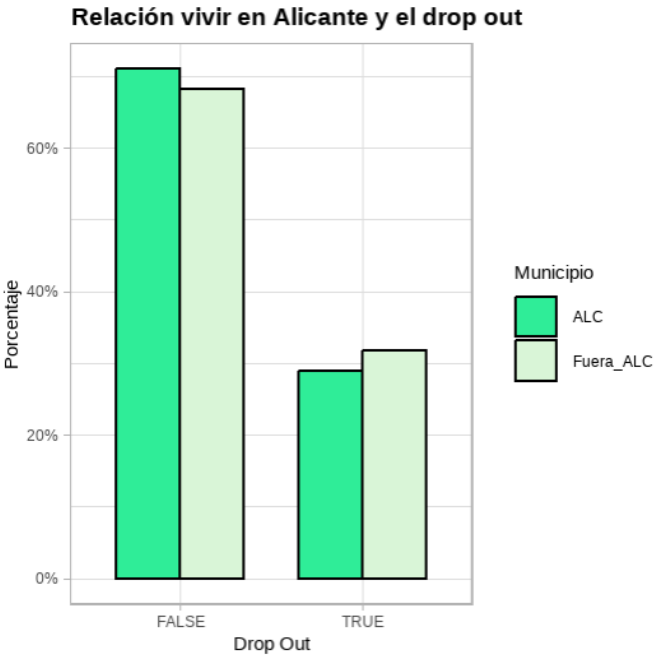


Gráfico 37. Relación entre vivir fuera de Alicante (Provincia) y el drop out

Como podíamos esperar, tener que mudarte a Alicante es un peso extra a la hora de hacer la carrera que puede derivar en el drop out de esta, aun así, no parece el suficiente como para que ser un factor de drop out.

10. Naturaleza del centro de estudios.

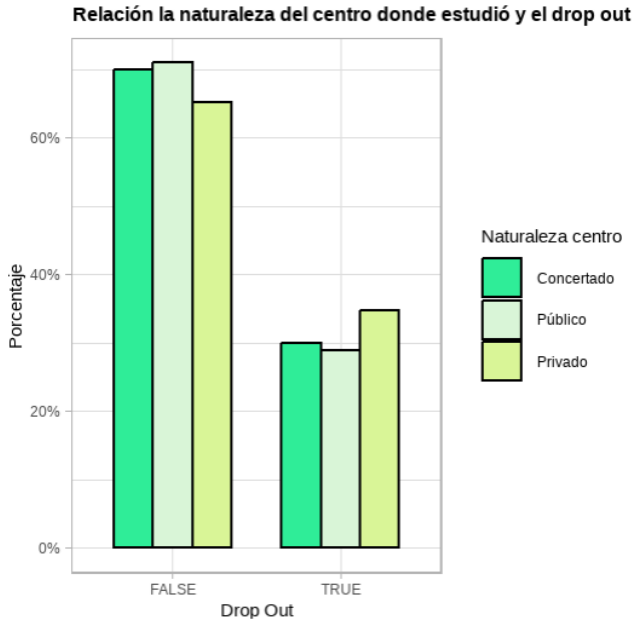


Gráfico 38. Relación entre la naturaleza del centro donde estudio el alumno y el drop out

Podemos ver que la naturaleza del centro solo es un factor de drop out si el estudiante fue a un centro privado a estudiar.

11. Nota de admisión.

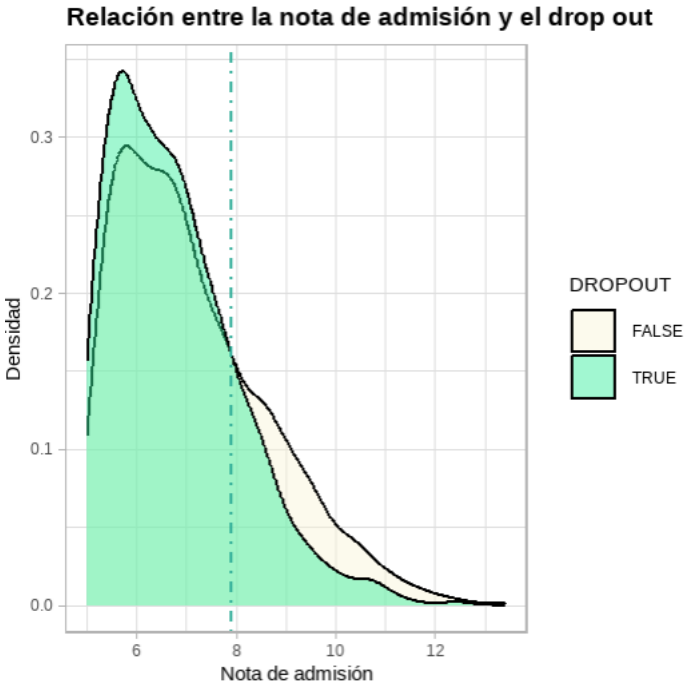


Gráfico 39. Relación entre la nota de admisión y el drop out

Podemos observar que la nota de admisión al grado sí que es un factor determinante del drop out, siendo cuanto más alta, menos probable.

Llama la atención que la tendencia comience a cambiar a partir de aproximadamente el tercer cuartil (7,97). Es decir, que, si el alumno pertenece al cuarto cuartil en la nota de admisión, es menos probable que abandone la carrera.

12. Matriculas por cuatrimestre

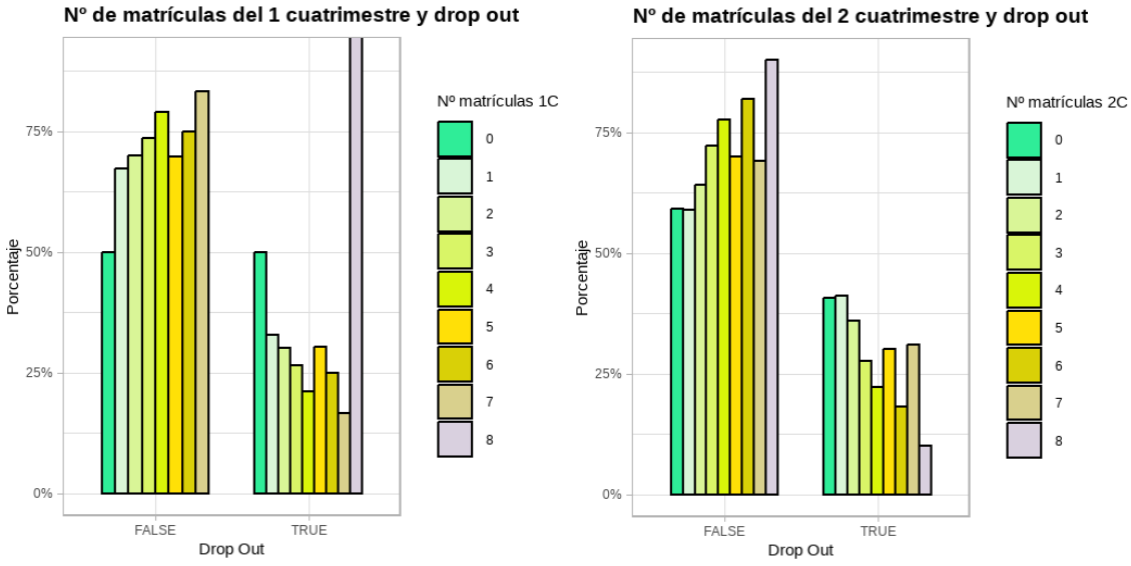


Gráfico 40. Relación entre el número de asignaturas a las que se matricula y el drop out

Podemos observar que cuantas menos matrículas más drop out, esto puede hacer referencia a lo que hemos visto antes de la modalidad a tiempo parcial. Destacar que cuando el alumno se matricula de 5 en cada cuatrimestre (la categoría más común), la probabilidad de drop out es la misma que la media.

13. Asignaturas a las que se presenta por cuatrimestre

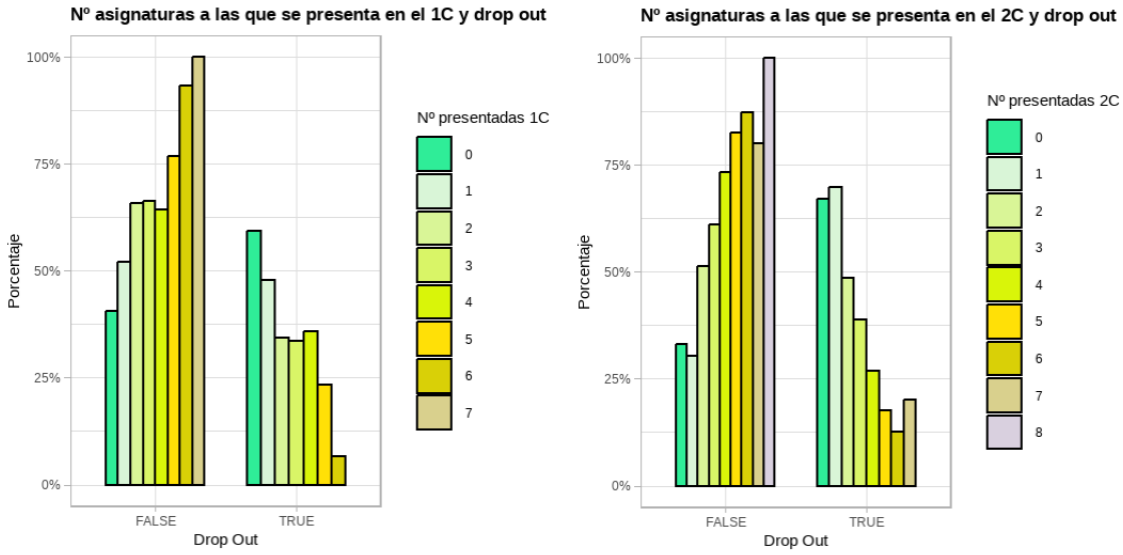


Gráfico 41. Relación entre el número de asignaturas a las que se presente el alumno y el drop out

Podemos ver que, en ambos cuatrimestres, a cuantas más asignaturas te presentes menos probabilidad de drop out, aunque en el segundo cuatrimestre está más pronunciada esta afirmación.

Esto no implica necesariamente que un alumno que se presente a todas las asignaturas de la carrera no la vaya a abandonar. Es posible que el hecho de que un alumno se presente a una asignatura indique que ha invertido tiempo y esfuerzo en prepararse para ella. Es decir, que presentarse puede indicar preparación para las asignaturas y esta puede ser un indicador de compromiso y dedicación por parte del alumno.

14. Nota media en cada cuatrimestre

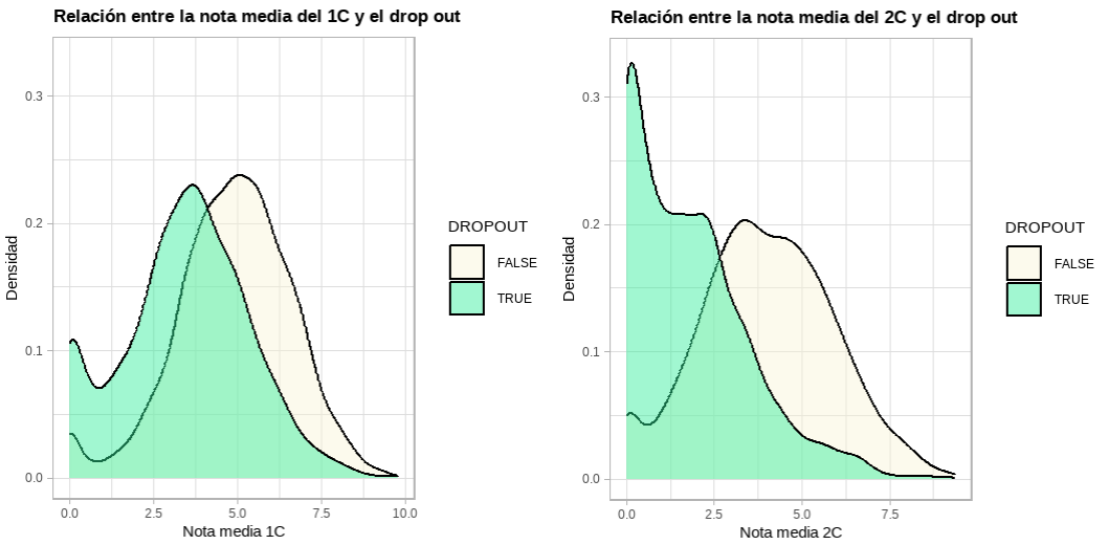


Gráfico 42. Relación entre la nota media y el drop out

Podemos ver lo que nos podríamos esperar. Las notas durante el grado son un fator muy determinante para el abandono de la carrera, lo que más llama la atención es la nota media del segundo cuatrimestre de la gente que se deja la carrera. Ver esa distribución de la nota cuando el alumno ha abandonado ratifica la hipótesis anterior de que hay gran parte del alumnado que decide abandonar la carrera pasado el primer cuatrimestre.

Para terminar de confirmar esta teoría vamos a comprobarlo con este gráfico de cajas sobre el número de asignaturas a las que se presenta el estudiante.

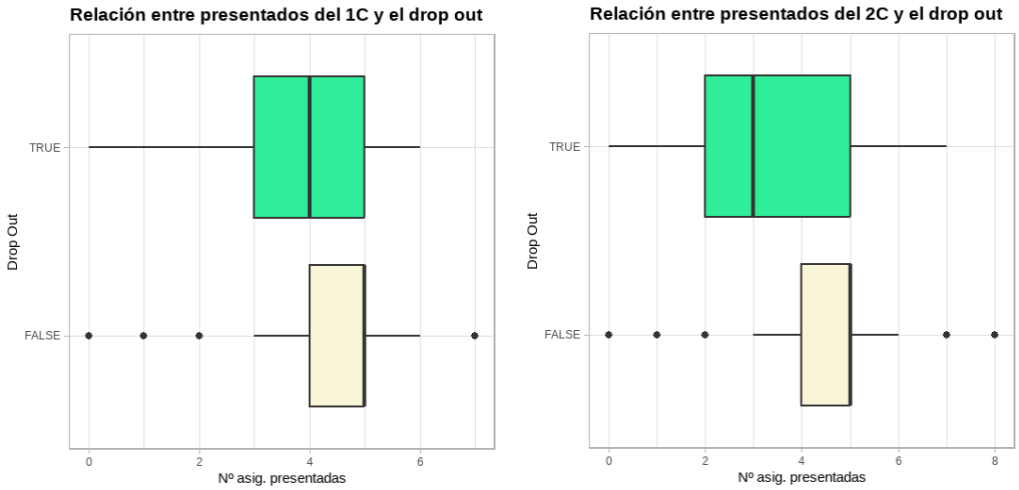


Gráfico 43. Relación entre las asignaturas a las que se presenta y el drop out

Podemos observar que en el grupo de los estudiantes que no abandonan la carrera, se presentan a un promedio de 4 o 5 asignaturas en ambos cuatrimestres. En contraste, en el grupo de los estudiantes que sí abandonan, en el primer cuatrimestre el de asignaturas presentadas, es en promedio de 3 a 5, con una media de 4, mientras que, en el segundo cuatrimestre, el rango baja de 2 a 5 asignaturas presentadas, con una media de 3.

Podemos confirmar nuestra teoría de antes, ya que si consideramos que la nota media de las asignaturas en el segundo cuatrimestre de la gente que abandona es de aproximadamente 1,25 y que el número de asignaturas a las que se presentan los estudiantes también disminuye en el segundo cuatrimestre, efectivamente podemos concluir que muchos de los estudiantes que abandonan la carrera toman esta decisión después del primer cuatrimestre.

Estas conclusiones serían muy útiles para predecir el rendimiento al final del curso, y es probable que la nota del segundo cuatrimestre sea una de las variables más determinantes. Sin embargo, dado que nuestro objetivo es detectar a los estudiantes en riesgo de abandono antes de que tomen esa decisión, con el fin de tomar medidas para prevenirlo, será mejor no utilizar variables que cobren sentido en el segundo cuatrimestre en nuestro modelo.

15. Número de asignaturas aprobadas

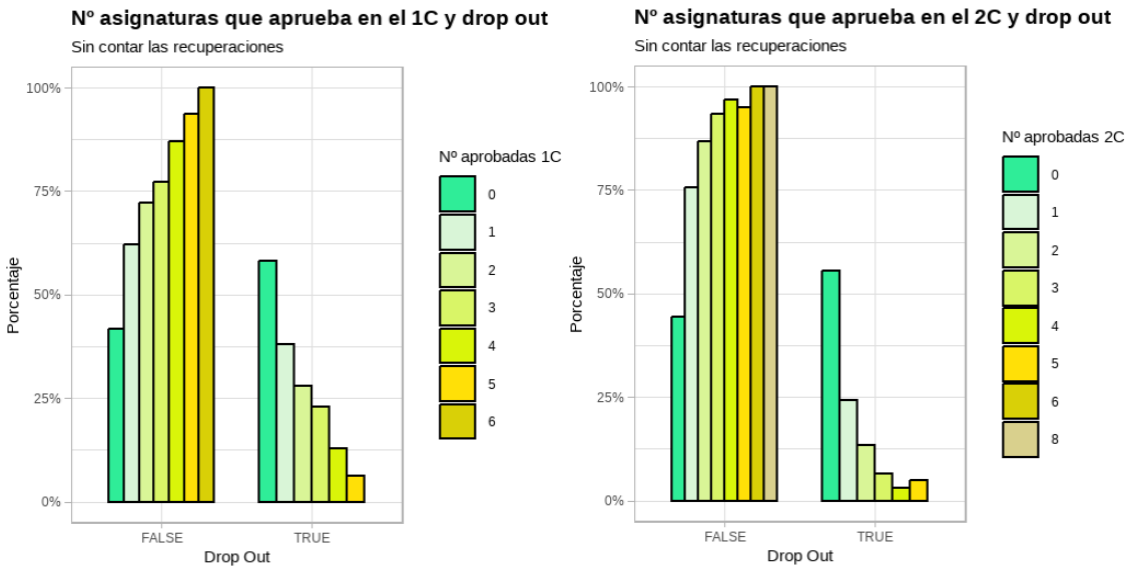


Gráfico 44. Relación entre el número de asignaturas que aprueba y el drop out por cuatrimestre

Como se podía suponer cuantas más asignaturas se aprueba menos probabilidad de drop out.

Vamos a ver ahora la relación con todas las aprobadas a final de curso.

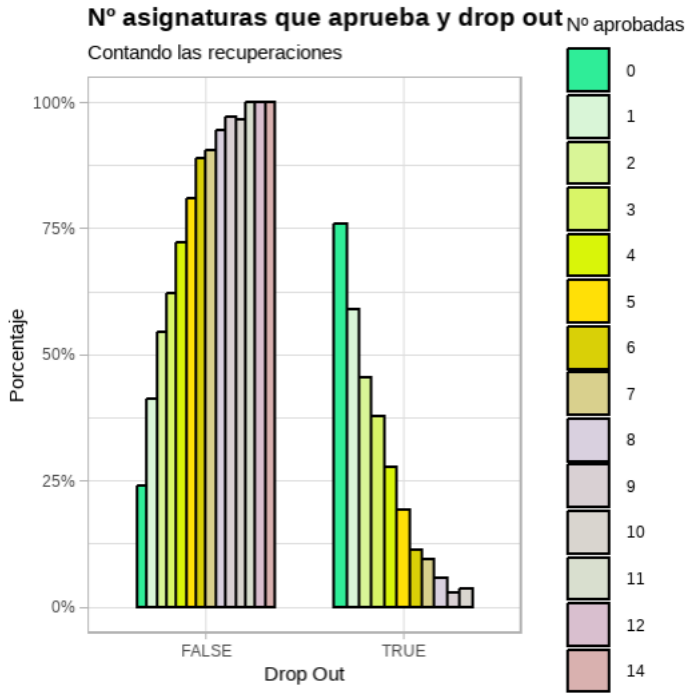
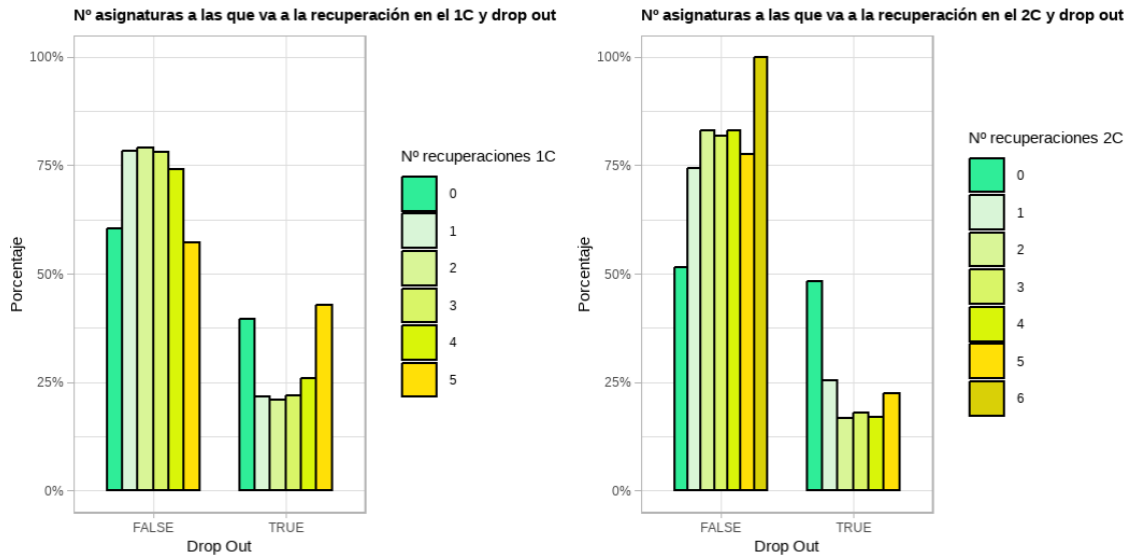


Gráfico 45. Relación entre el número de asignaturas que aprueba y el drop out por cuatrimestre

Efectivamente, cuantas más se aprueban más posibilidades hay de no abandonar la carrera. Un dato que parece curioso es que el 25% de los alumnos que han aprobado 0 siguen en la carrera.

16. Número de asignaturas a las que se presenta a recuperar



Lo que más llama la atención, sobre todo en las asignaturas del primer cuatrimestre, es que los más probables del drop out sean los que han ido a recuperar 0 y 5, los del 0, es por lo que hemos comentado anteriormente de que los estudiantes que se han dejado la carrera no van a sus respectivas recuperaciones y los de 5 puede ser debido a los resultados de estas 5 recuperaciones, que no habrán sido los esperados.

4.2.4. Conclusiones del análisis

Antes de finalizar esta sección, vamos a repasar los resultados obtenidos en el análisis.

El más importante, es la suposición de que muchos alumnos deciden abandonar la carrera tras el primer cuatrimestre. Los resultados del rendimiento del segundo cuatrimestre son probablemente las variables más importantes en la predicción del abandono prematuro, pero si nos centramos en nuestro objetivo, el cual es predecir el drop out para poder actuar en consecuencia y en la medida de lo posible poder evitarlo, y concluyendo que tras el segundo cuatrimestre muchos alumnos ya han decidido sobre su futuro, nuestro modelo deberá desestimar las variables sobre el rendimiento del segundo cuatrimestre.

Hemos llegado a la conclusión de que muy probablemente los datos más influyentes sean los del rendimiento del alumno durante el grado, aunque hay variables personales que también pueden tener peso.

Con toda esta información que hemos recogido vamos a construir los modelos predictores.

4.3. Modelos

En este apartado crearemos diferentes modelos para la predicción del drop out, los modelos serán, regresión logística, árbol de decisión y random forest.

Como hemos concluido en el apartado anterior, los modelos predecirán el abandono tras obtener los resultados del primer cuatrimestre, por lo que tenemos que eliminar todas las variables que se obtienen a partir de ese momento.

Para la evaluación de los modelos se usarán tres valores, la precisión del modelo, la precisión de verdaderos positivos y la métrica roc_auc (Receiver Operating Characteristic - Area Under the Curve), aunque por el que nos decantaremos para elegir el modelo es el roc_auc, el motivo se explica mejor en el apartado de resultados.

Antes de crear los modelos es muy importante segmentar de forma aleatoria la muestra que tenemos en dos partes, una de entrenamiento (75% de la muestra) y otra para probar el modelo (25% de la muestra).

La muestra de entrenamiento sirve como su propio nombre indica, para entrenar al modelo y la de prueba para ver cómo se comporta el modelo entrenado con una muestra diferente, de esta prueba salen los resultados de evaluación.

4.3.1. Regresión logística simple

El modelo de regresión logística es una poderosa herramienta para predecir variables categóricas binarias. Como en nuestro caso, el drop out que solo tiene dos categorías TRUE y FASLE.

Este modelo se centra en la predicción de la probabilidad del suceso en cuestión, por lo que en la categoría TRUE (que abandona la carrera) se encontrarán todos aquellos alumnos cuya probabilidad se abandono sea del 50% al 100% por defecto. Este umbral se puede modificar, pero eso lo veremos más adelante.

Una de las ventajas de usar este modelo es que nos proporciona los coeficientes de regresión de cada variable, que nos indican la dirección y magnitud que tienen sobre el drop out.

Otra ventaja es que nos permite crear interacciones entre las variables, superando así la linealidad, aunque hay que tener cuidado con no sobreajustar el modelo.

En este modelo no vamos a añadir interacciones entre variables.

Para evitar el sobreajuste y mejorar la generalización del modelo a nuevos datos vamos a regularizar el modelo.

Tenemos que ajustar el valor del hiperparámetro lambda en función del roc_auc, su valor, es el que controla la fuerza de la penalización a los coeficientes del modelo. Para obtener este valor utilizaremos la técnica de validación cruzada, esta consiste en dividir el conjunto de datos en múltiples partes y realizar varias iteraciones de entrenamiento y prueba llamadas submuestras, utilizando diferentes combinaciones de submuestras como conjunto de entrenamiento y prueba. Esto permite obtener una estimación más robusta del rendimiento del modelo y reducir el sesgo que se tendría con una única división de los datos.

El valor que nos da lambda en este modelo es una penalización de 0,00143.

A continuación, se muestran todos los coeficientes del modelo.

term	estimate	penalty
(Intercept)	1.0925	0.0014
NOTAADMISION	-0.0674	0.0014
EDAD	-0.0682	0.0014
MATRICULA_1C	-0.3476	0.0014
MATRICULA_2C	0.0203	0.0014
PRESENTADO_1C	0.2710	0.0014
NOTA_1C	0.2391	0.0014
APROBADO_1C	0.7159	0.0014
PAISNACIONALIDAD_extranjero	0.0357	0.0014
SEXO_Mujer	0.0274	0.0014
DEDICACIONESTUDIO_Parcial	-0.0544	0.0014
FAMILIANUM_Numerosa	0.0781	0.0014
FAMILIANUM_Numerosa.especial	0.0000	0.0014
NIVELESTUDIOPADRE_Secundaria	0.0426	0.0014
NIVELESTUDIOPADRE_Primaria	-0.0020	0.0014
NIVELESTUDIOPADRE_Ninguno	-0.0377	0.0014
NIVELESTUDIOMADRE_Secundaria	0.0188	0.0014
NIVELESTUDIOMADRE_Primaria	-0.0232	0.0014
NIVELESTUDIOMADRE_Ninguno	0.0000	0.0014
FORMAADMISION_Grad.Sup.	0.0014	0.0014
FORMAADMISION_Extr.	-0.0015	0.0014
FORMAADMISION_Conv.	-0.1003	0.0014
FORMAADMISION_X.25	-0.0448	0.0014
FORMAADMISION_X.40	-0.0138	0.0014
FORMAADMISION_X.45	0.0325	0.0014
MUNICIPIOCENTROSEC_Fuera_ALC	0.0191	0.0014
NATURALEZACENTROSEC_Público	0.0000	0.0014
NATURALEZACENTROSEC_Privado	-0.0236	0.0014
NUEVOSUE_Sí	-0.3356	0.0014
NIVELESTUDIOCASA_Secundaria	-0.0667	0.0014
NIVELESTUDIOCASA_Primaria	0.0000	0.0014
NIVELESTUDIOCASA_Ninguno	0.0608	0.0014

Tabla 2. Coeficientes del modelo de regresión logística simple

Y ahora podremos ver la importancia que tiene cada variable en el modelo.

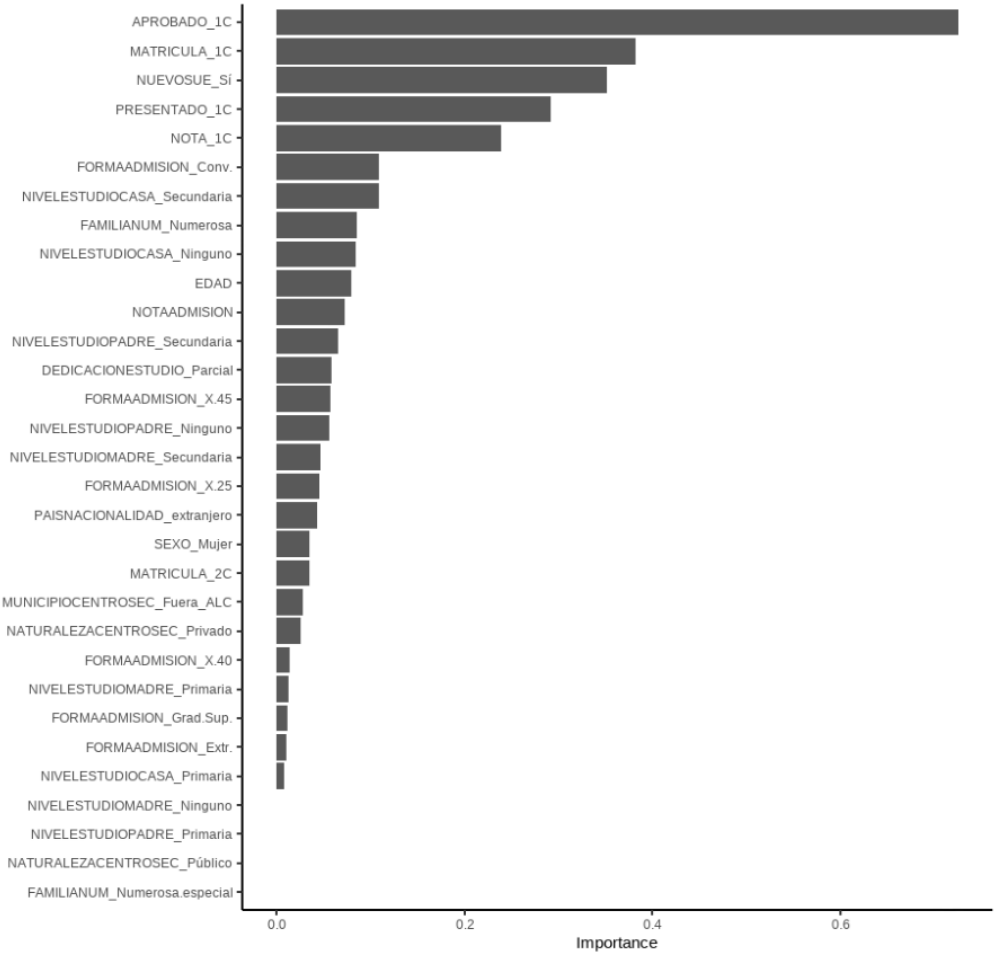


Tabla 3. Importancia de las variables del modelo de regresión logística simple

Podemos ver en orden de importancia que efectivamente, el rendimiento del estudiante es lo que más afecta en la decisión del abandono.

Llama la atención que, de las variables personales de cada estudiante, ser nuevo en el SUE sea la más significativa, bien es cierto que en análisis exploratorio hemos observado que es de las variables con más diferencia entre sus categorías.

Si observamos las variables menos importantes, todas corresponden a factores personales del alumno, entre las que destacan la naturaleza del centro de secundaria donde estudió el alumno, el nivel de estudio de los padres en general, el municipio de acceso, el sexo y un variable del rendimiento, que es las asignaturas matriculadas en el segundo cuatrimestre.

Podemos ver el error del modelo.

.metric	.estimator	.estimate	.config
accuracy	binary	0.7412766	Preprocessor1_Model1
roc_auc	binary	0.7657266	Preprocessor1_Model1

Tabla 4. Precisión y roc_auc del modelo de regresión logística simple

Vemos que la precisión del modelo es del 74,13% pero como en nuestro objetivo, es mucho más importante detectar al alumno que sí abandona el grado, vamos a ver la precisión de esta categoría en la matriz de confusión del modelo.

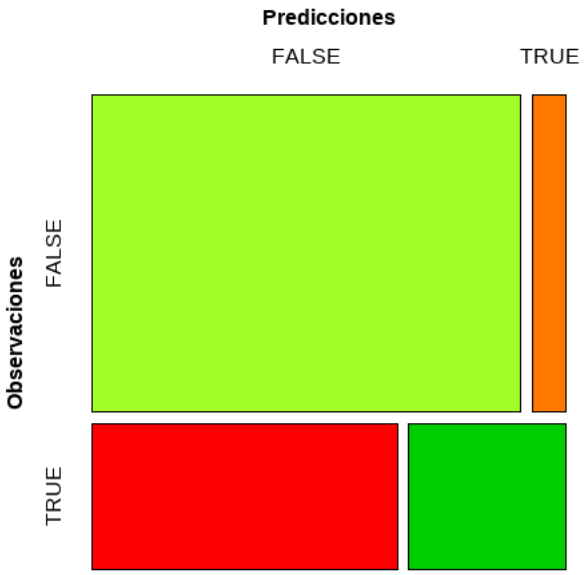


Gráfico 46. Matriz de confusión regresión logística simple (umbral = 0.5)

Como podemos ver, el falso negativo es bastante alto, podemos decir que la precisión de detectar que un alumno sí abandona la carrera es del 33,96%, una cifra bastante mejorable.

Para poder mejorar esta cifra podemos ajustar el umbral para que, pese a que el modelo pierda precisión general, gane precisión por la parte que nos interesa.

Como hemos comentado al principio de este apartado el modelo no predice si la variable es TRUE o FALSE, el modelo calcula la probabilidad de que la variable a predecir sea TRUE o FALSE. El umbral que por defecto es de 0.5, es el que se encarga de clasificar las variables según su probabilidad. Entonces, si queremos ganar precisión en el apartado de la predicción de valores TRUE, tendremos que bajar el umbral. ¿Cuánto? Esto depende de lo que estemos dispuestos a sacrificar en la predicción absoluta.

Tras probar varios umbrales, vamos a poner el umbral en 0,3.

Ahora obtendremos una nueva matriz de confusión.

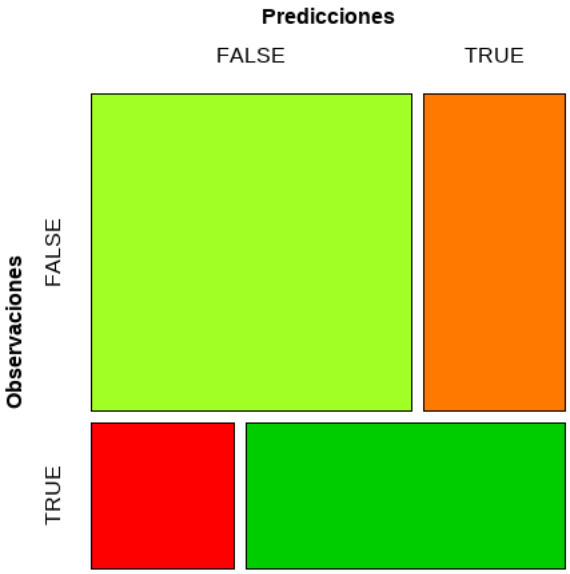


Gráfico 47. Matriz de confusión regresión logística simple (umbral = 0.3)

Como podemos comprobar, bajar el umbral, ha hecho que mejore bastante la precisión en los valores TRUE. La precisión ahora de verdaderos positivos es del 69,00%, por el contrario, hemos depreciado la precisión total, situándose esta en el 69,19%.

El valor de roc_auc no varía.

4.3.2. Regresión logística

Ahora vamos a hacer el mismo modelo de antes, pero considerando algunas otras especificaciones, creando diferentes interacciones entre las variables y eliminando las variables que eran menos importantes para el modelo anterior.

Primero vamos a transformar la nota media en un polinomio, capturando así relaciones no lineales, también vamos a discretizar la variable de la edad conforme hemos visto en el análisis de exploratorio (que a partir de los 24 cambiaba la tendencia) y la variable de la nota de admisión (que el drop out era más común en los primeros tres cuartiles (de 0 a 7,79)). Por último, vamos a crear dos interacciones entre el número de asignaturas a las que el alumno se matricula y a las que se presenta y entre el número de asignaturas a las que se presenta y las que aprueba.

El valor que nos da lambda en este modelo es una penalización de 0.

A continuación, se muestran los coeficientes del modelo.

term	estimate	penalty
(Intercept)	1.1176	0
MATRICULA_1C	-0.2299	0
PRESENTADO_1C	0.7076	0
APROBADO_1C	-0.9164	0
NOTA_1C_poly_1	0.4069	0
NOTA_1C_poly_2	-0.0662	0
MATRICULA_1C_x_PRESENTADO_1C	-0.7331	0
APROBADO_1C_x_PRESENTADO_1C	1.6535	0
PAISNACIONALIDAD_extranjero	0.0591	0
DEDICACIONESTUDIO_Parcial	-0.0552	0
FAMILIANUM_Numerosa	0.0883	0
FAMILIANUM_Numerosa.especial	0.0009	0
FORMAADMISION_Grad.Sup.	0.0156	0
FORMAADMISION_Extr.	-0.0112	0
FORMAADMISION_Conv.	-0.0861	0
FORMAADMISION_X.25	-0.0239	0
FORMAADMISION_X.40	-0.0192	0
FORMAADMISION_X.45	0.0798	0
NUEVOSUE_Sí	-0.3425	0
NOTAADMISION_X.7.97.Inf.	-0.0433	0
EDAD_X.24.Inf.	-0.1033	0

Tabla 5. Importancia de las variables del modelo de regresión logística simple

Ahora podemos ver la importancia de cada variable.

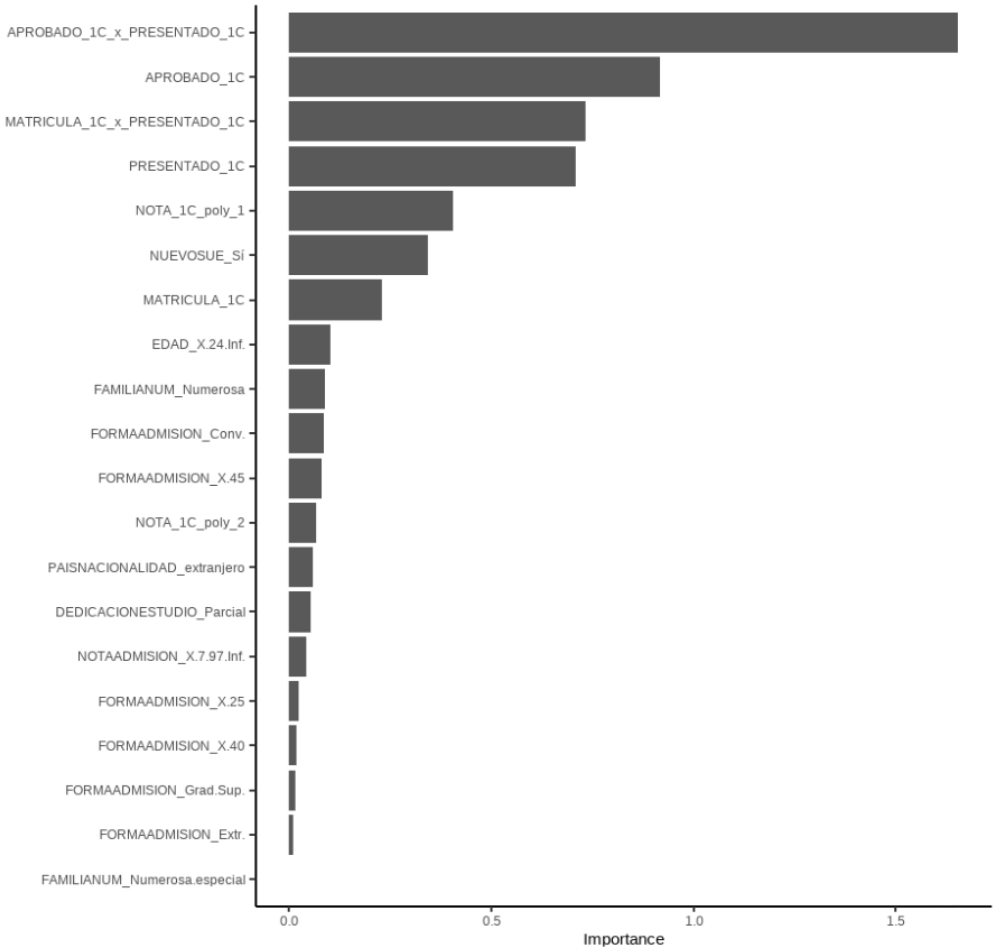


Tabla 6. Importancia de las variables del modelo de regresión logística

Como podemos ver las variables del rendimiento son las más importantes de nuevo y las variables personales las que menos.

Podemos ver el error del modelo.

.metric	.estimator	.estimate	.config
accuracy	binary	0.7361702	Preprocessor1_Model1
roc_auc	binary	0.7700664	Preprocessor1_Model1

Tabla 7. Precisión y roc_auc del modelo de regresión logística

Vemos que la precisión del modelo es del 73,61%, pero igual que en el modelo anterior nos fijaremos en la matriz de confusión.

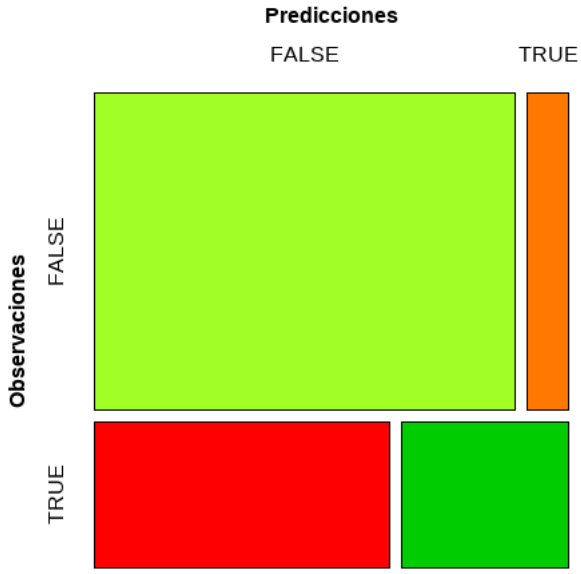


Gráfico 48. Matriz de confusión regresión logística (umbral = 0.5)

Como podemos ver, el falso negativo es bastante alto, la precisión de acertar el abandono cuando es TRUE es del 36,11%.

Vamos a hacerla de nuevo, pero con el mismo umbral de antes, 0.3.

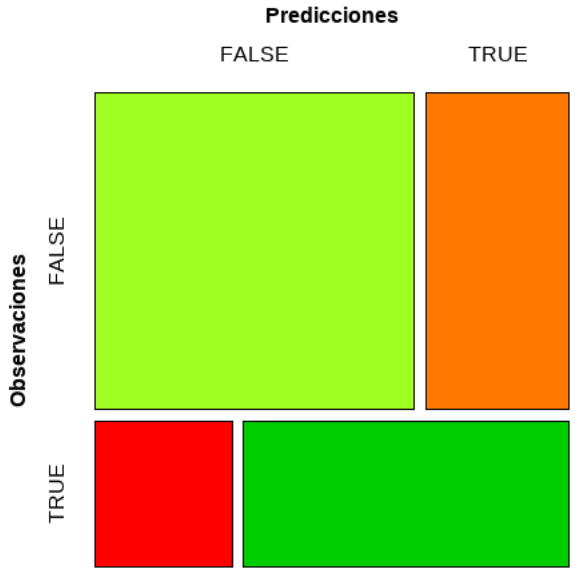


Gráfico 49. Matriz de confusión regresión logística (umbral = 0.3)

Como podemos comprobar, ha mejorado bastante la precisión en los valores TRUE. La precisión ahora es del 69,45%, por el contrario, hemos depreciado la precisión total, situándose esta en el 70,35%.

4.3.3. Árbol de decisión

El método del árbol de decisión es otra herramienta poderosa para predecir variables categóricas binarias.

Este modelo se basa en la construcción de un árbol compuesto por nodos y ramas, donde cada nodo representa una característica o variable, y las ramas representan las posibles decisiones o resultados. El árbol se divide en base a las características que mejor predicen el drop out.

Una de las ventajas del árbol de decisión es su representación visual de las reglas de decisión. Esto nos permite interpretar fácilmente las variables más importantes y su influencia en el resultado del modelo.

Además, el árbol de decisión es capaz de capturar relaciones no lineales y de interacción entre las variables, que, a diferencia de la regresión logística, no asume una relación lineal.

Sin embargo, es importante tener en cuenta que los árboles de decisión son propensos al sobreajuste (overfitting). Esto ocurre cuando el modelo se ajusta demasiado a los datos de entrenamiento y predice muy bien, pero solo en la muestra de entrenamiento. Para evitar el sobreajuste, se pueden aplicar técnicas de regularización, como la poda del árbol.

En el método de la poda (pruning), tenemos que ajustar el hiperparámetro, llamado alfa, para controlar la estructura del árbol y evitar el sobreajuste. La optimización del modelo se realiza mediante la búsqueda de la mejor combinación de alfa utilizando validación cruzada y la métrica de evaluación ROC-AUC, muy similar a lo explicado previamente en la regresión logística.

En nuestro caso el valor proporcionado por alfa es coste 0.

A continuación, muestro el árbol de decisión.

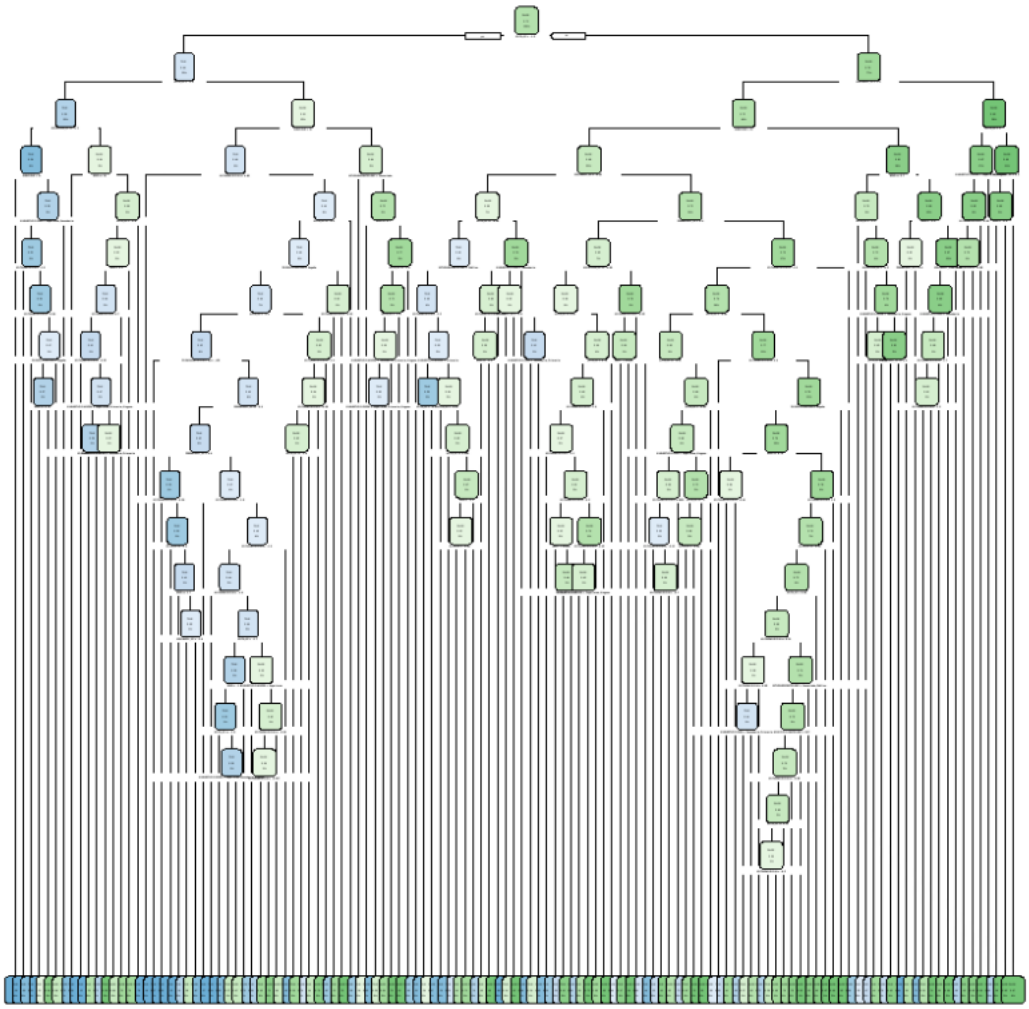


Tabla 8. Árbol de decisión

Como podemos ver, este árbol es muy complejo de entender a simple vista.

Ahora vamos a ver la importancia de las variables.

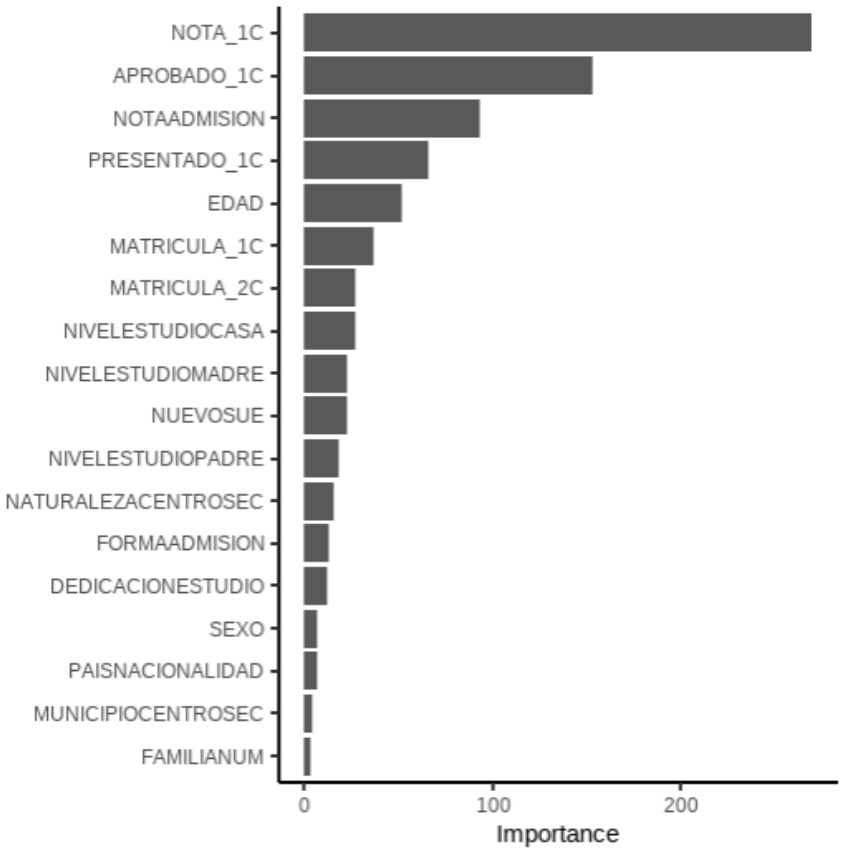


Tabla 9. Importancia de las variables del modelo del árbol de decisión.

Esta tabla nos ofrece una información muy interesante. Ya que como hemos comentado anteriormente, una de las ventajas de este método es más allá de su capacidad predictiva, es la capacidad de capturar relaciones no lineales y de interacción entre las variables.

Como podemos ver, esta vez la importancia mezcla tanto variables del rendimiento como personales, siendo se nuevo las del rendimiento más importantes. Destacar que tanto la nota media del primer cuatrimestre como la nota de admisión al grado han adquirido mucha más importancia en este modelo, siendo la nota media del primer cuatrimestre la más importante con diferencia. Mientras que ser nuevo en el SUE ha perdido importancia frente a otras variables personales.

Aquí podemos ver el error del modelo.

.metric	.estimator	.estimate	.config
accuracy	binary	0.7021277	Preprocessor1_Model1
roc_auc	binary	0.7058860	Preprocessor1_Model1

Tabla 10. Precisión y roc_auc del modelo del árbol de decisión

Vemos que la precisión del modelo es del 70,21%. Igual que en el modelo anterior nos fijaremos en la matriz de confusión.

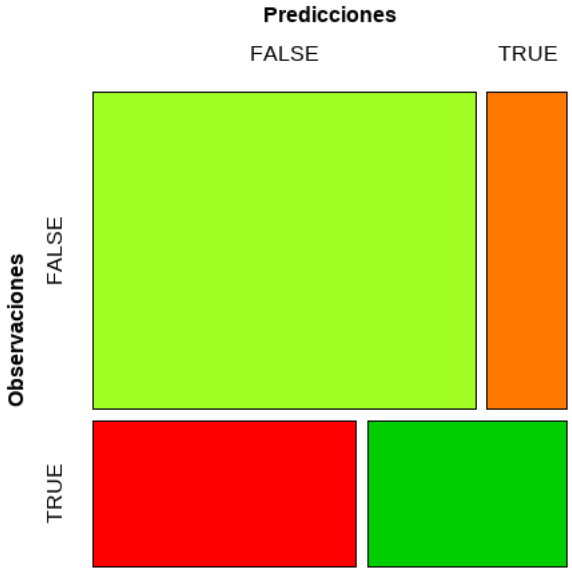


Tabla 11. Matriz de confusión del árbol de decisión (umbral = 0.5)

Como podemos ver, el falso negativo es relativamente alto, la precisión de acertar el abandono cuando es TRUE es del 43,13%.

Vamos a hacerla de nuevo, pero con el umbral utilizado previamente, 0.3.

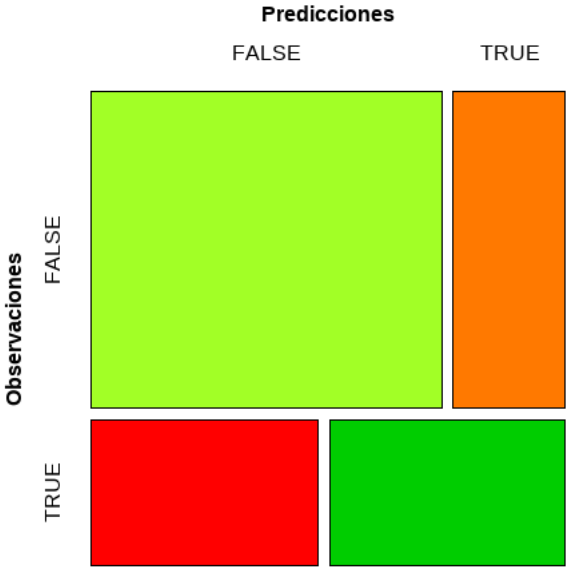


Tabla 12. Matriz de confusión del árbol de decisión (umbral = 0.3)

Como podemos comprobar, ha mejorado mucho menos la precisión en los valores TRUE en comparación con los modelos anteriores. La precisión ahora es del 50,94%, por el contrario, hemos depreciado la precisión total, situándose esta en el 68,00%. Esto es debido a que el umbral utilizado no es el óptimo (según nuestros intereses) para este modelo.

4.3.4. Random Forest

El modelo Random Forest es otro modelo para predecir variables categóricas binarias. Este modelo se basa en la construcción de múltiples árboles de decisión, donde cada árbol se ajusta a diferentes subconjuntos de datos y características. Luego, se combinan las predicciones de todos los árboles para obtener un resultado final.

Una de las ventajas del Random Forest al igual que en el árbol de decisión, es su capacidad para manejar variables no lineales y capturar interacciones complejas entre las variables predictoras.

Igual que en los árboles de decisión, el Random Forest también puede sufrir de sobreajuste si no se controla adecuadamente. La técnica de regularización utilizada en el Random Forest se basa en seleccionar un subconjunto, x, aleatorio de características en lugar de todas las características disponibles. Al no incluir todas las características consideradas en cada división, se reduce la correlación entre los árboles individuales, lo que contribuye a la regularización.

Para ello tenemos que seleccionar el número de árboles y el número, x, de variables a considerar en cada subconjunto. Para ello se puede, al igual que en los modelos anteriores, realizar una búsqueda utilizando validación cruzada y la métrica de evaluación ROC-AUC.

En nuestro caso seleccionaremos 130 árboles y el número óptimo de variables a usar (aleatoriamente) en cada nodo son 2.

Este modelo no se puede "ver", por lo que vamos a ver la importancia de las variables directamente.

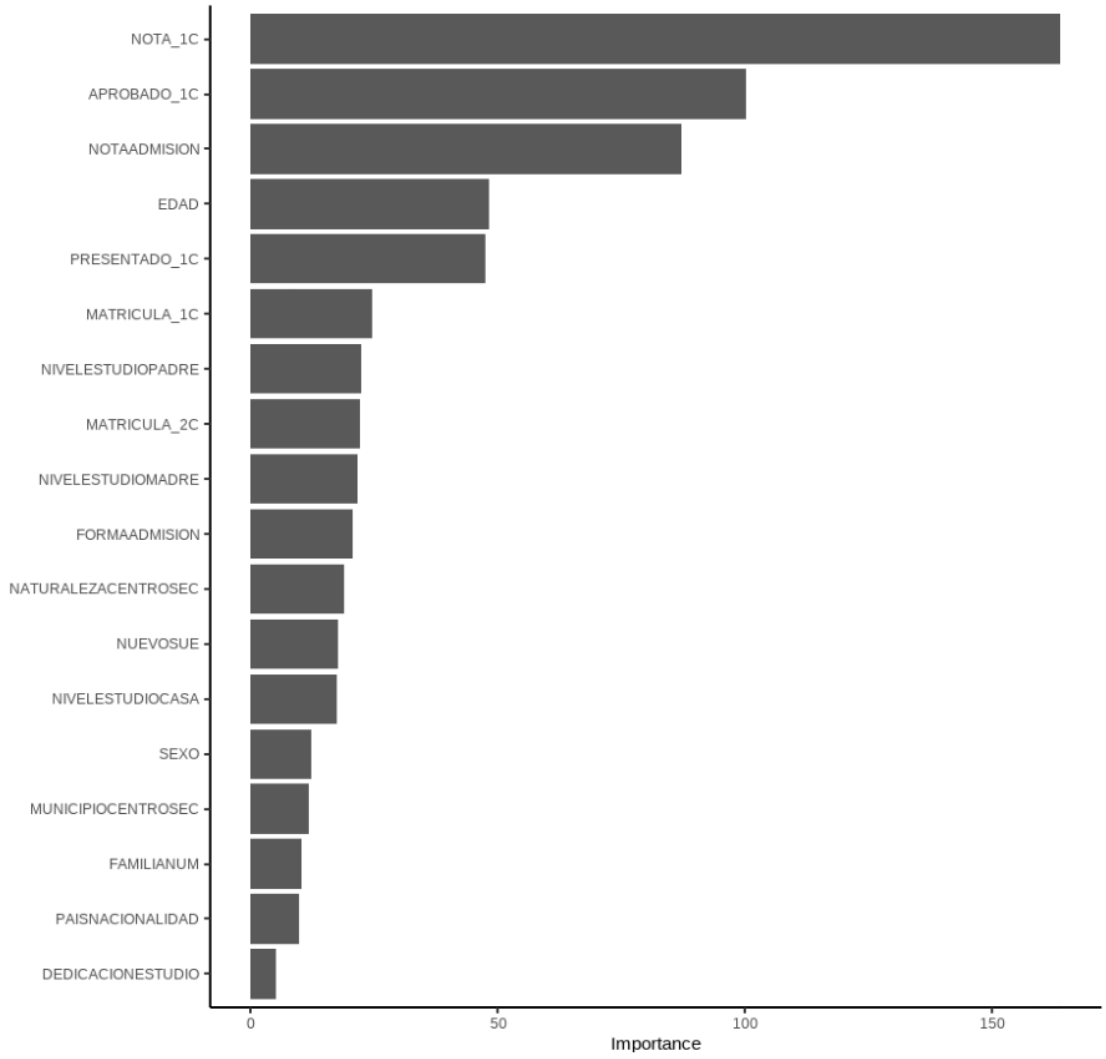


Tabla 13. Importancia de las variables del modelo Random Forest

Como podemos ver la tabla de la importancia de las variables es prácticamente igual a la del árbol de decisión, salvo por alguna pequeña variación. De nuevo, vemos que las variables más importantes son del rendimiento del alumno, y, en este caso, la más importante (con diferencia) es la nota media obtenida por el estudiante en el primer cuatrimestre.

Aquí podemos ver el error del modelo.

.metric	.estimator	.estimate	.config
accuracy	binary	0.7336170	Preprocessor1_Model1
roc_auc	binary	0.7663737	Preprocessor1_Model1

Tabla 14. Precisión y roc_auc del modelo de Random Forest

Vemos que la precisión del modelo es del 73,36%, pero igual que en el modelo anterior nos fijaremos en la matriz de confusión.

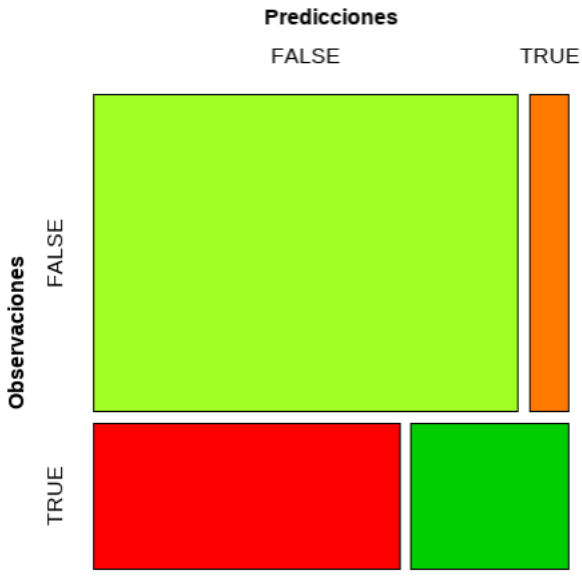


Tabla 15. Matriz de confusión Random Forest (umbral = 0.5)

Como podemos ver, el falso negativo es bastante alto, la precisión de acertar el abandono cuando es TRUE es del 33,96%.

Vamos a hacerla de nuevo, pero con el umbral utilizado previamente, 0.3.

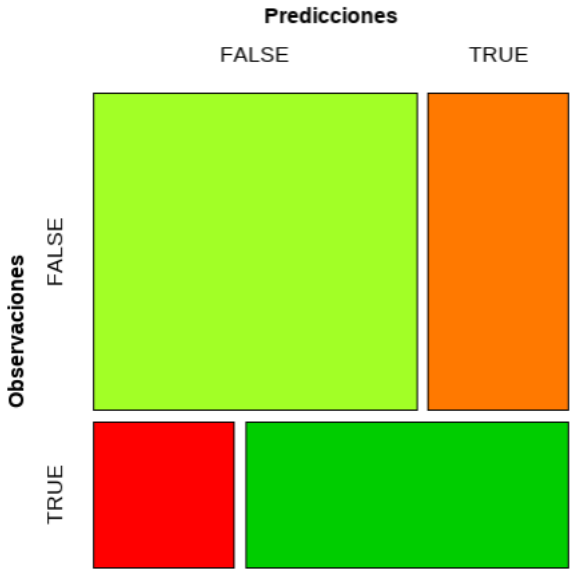


Tabla 16. Matriz de confusión Random Forest (umbral = 0.3)

Como podemos comprobar, ha mejorado mucho la precisión en los valores TRUE. La precisión ahora es del 69,54%, por el contrario, hemos depreciado la precisión total, situándose esta en el 69,70%.

4.3.5. Random Forest (2 cuatrimestre)

Como hemos concluido antes, nos interesaba calcular la predicción del abandono después del final del primer cuatrimestre. Esto era debido a que supusimos que tras el primer cuatrimestre muchos alumnos tomaban la decisión de abandonar el grado.

Tras esta decisión también concluimos que abandonábamos predictores que probablemente serían los más relevantes para la predicción del abandono (los del segundo cuatrimestre).

Para probar esta teoría vamos a construir un modelo con todos los datos que tenemos, los personales y los del rendimiento del estudiante, tanto el rendimiento del primer cuatrimestre, como el del segundo cuatrimestre.

Utilizaremos el modelo Random Forest por simplicidad, ya que, como hemos dicho, este modelo es capaz de manejar variables no lineales y capturar interacciones entre variables.

Para regularizar el modelo seleccionaremos 130 árboles y el número óptimo de variables a usar (aleatoriamente) en cada nodo son 4.

La importancia de las variables es.

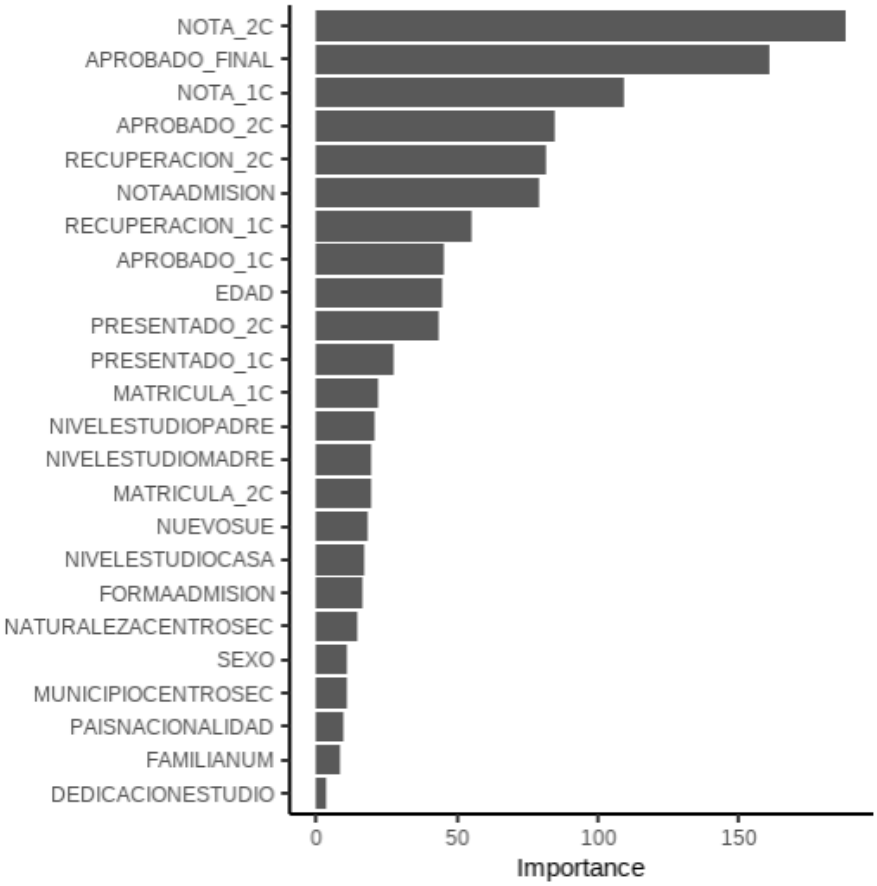


Tabla 17. Importancia de las variables del modelo Random Forest segundo cuatrimestre

Efectivamente, con muchísima diferencia, las variables más importantes son del rendimiento del alumno durante el segundo cuatrimestre, cuyo rendimiento sería nulo si había decidido abandonar el grado con anterioridad.

Ahora vamos a ver el error que comete este modelo.

.metric	.estimator	.estimate	.config
accuracy	binary	0.8263830	Preprocessor1_Model1
roc_auc	binary	0.8636702	Preprocessor1_Model1

Tabla 18. Precisión y roc_auc del modelo de Random Forest segundo cuatrimestre

Efectivamente, vemos que la precisión del modelo es del 82,26%, mucho mayor que la de los modelos anteriores.

Por curiosidad vamos a ver también las respectivas matrices de confusión de este modelo.

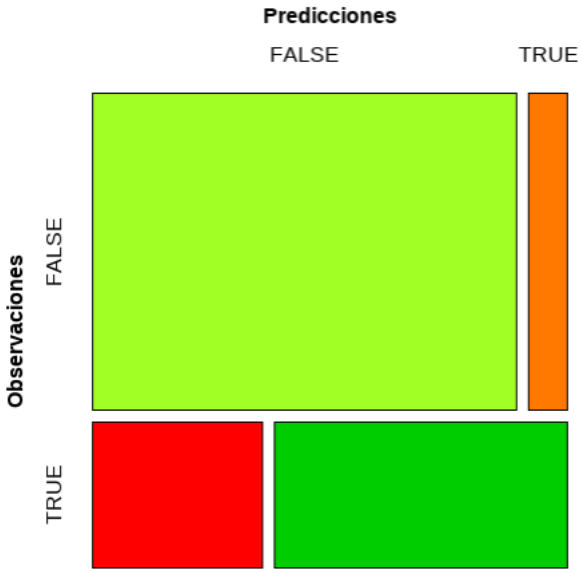


Tabla 19. Matriz de confusión Random Forest 2 cuatrimestre (umbral = 0.5)

Vemos de primeras, que la precisión de acertar el abandono ha aumentado mucho, exactamente es del 63,34%.

Para hacer una buena comparación vamos a ajustar el umbral a 0,3 de nuevo.

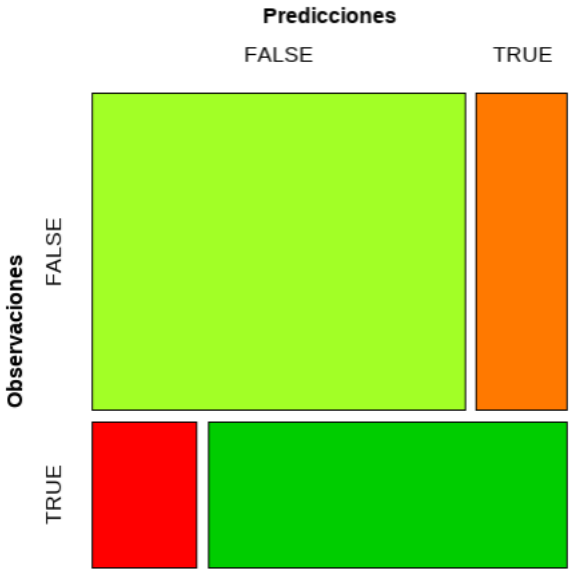


Tabla 20. Matriz de confusión Random Forest 2 cuatrimestre (umbral = 0.3)

Como podemos comprobar, la mejora ha sido relevante. La precisión ahora es del 77,36%, y la precisión total 79,49%.

5. RESULTADOS

En este apartado vamos a analizar los resultados, responder a las preguntas que nos hemos planteado y proponer posibles soluciones al problema.

5.1. Elección del modelo

Para elegir el modelo más adecuado para predecir que alumno es más probable que abandone la carrera, tenemos que comparar los resultados obtenidos previamente.

La métrica de referencia que utilizaremos para comparar los modelos es el roc_auc.

Esta métrica representa el área bajo la curva ROC, que es una representación del rendimiento de un modelo de clasificación binaria a medida que se ajusta el umbral de clasificación. La curva ROC muestra la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos del modelo en diferentes umbrales de clasificación.

El valor del roc_auc es una medida numérica del desempeño del modelo. Un valor de 0.5 indica que el modelo tiene un rendimiento aleatorio, mientras que un valor de 1 indica un modelo de predicción perfecto.

Como hemos visto, casi todos los modelos que hemos construido clasificaban mejor (según nuestro objetivo) utilizando un umbral más bajo, y puesto que a cada modelo se le ajustaría mejor un umbral diferente, la precisión bajo el umbral del 0,3 no es una métrica adecuada para la comparación de los modelos. La incluimos de forma representativa, ya que es fácil de entender a simple vista.

Los resultados obtenidos de los modelos que hemos construido son los siguientes.

	log_reg	log_reg_2	arbol	rnd_forest	2_cuatr
precisión	0.69191	0.69447	0.68	0.69872	0.79064
prec_pos	0.69003	0.7035	0.50674	0.69272	0.7655
roc_auc	0.76584	0.77007	0.70181	0.76839	0.86153

Tabla 21. Rendimiento de los modelos

En esta tabla podemos observar los cuatro modelos que hemos construido y un quinto que es el utilizado con los datos del rendimiento del segundo cuatrimestre.

Las métricas que observamos son: precisión (bajo el umbral de 0.3), precisión de los verdaderos positivos (bajo el umbral de 0.3) y roc_auc.

Como podemos observar el modelo con el roc_auc más alto, pero con muy poca diferencia, es el segundo modelo de regresión logística. Donde lo hemos contemplado con la no linealidad de las variables. Aunque los modelos de regresión logística simple y el modelo de Random Forest muestran un rendimiento muy similar.

Destacar que, si se quiere utilizar el modelo, la elección del umbral dependerá de quien lo esté utilizando y de su criterio. No hay uno ni mejor ni peor que otro.

También observamos que efectivamente, un modelo con los datos sobre el rendimiento del estudiante tras el segundo cuatrimestre predice sustancialmente mejor que los demás.

5.2. Influencia de las variables

Primero de todo hay que puntualizar un aspecto: que una variable explique el comportamiento, de en nuestro caso, el abandono, no significa que tenga que existir una causalidad.

Durante la creación de los modelos hemos ido viendo la influencia que tenía cada variable sobre el abandono (regresión logística 1, regresión logística 2, árbol de decisión, random forest y 2_cuatri), todos los modelos comparten una característica en común, las variables del rendimiento académico son las que más explican el abandono (sobre todo el número de asignaturas aprobadas y la nota media del cuatrimestre).

Dentro de las variables personales, las que más importancia tienen son: la nota de admisión, la edad del estudiante, el nivel de estudios en casa, y si es nuevo en el Sistema Universitario Español.

Un gran defecto en todos estos modelos es que solo se dispone de la información de los alumnos que tiene la universidad y esto supone que se desconocen muchos factores personales de este, como sus gustos, motivaciones, situación familiar, problemas personales etc. Factores que podrían explicar mucho mejor el abandono del grado, con la obtención de esta información se podrían obtener resultados mucho más precisos.

5.3. Retención estudiantil

El objetivo final de este trabajo era detectar a los alumnos que iban a abandonar el grado de ADE antes de que lo hiciesen y así actuar en consecuencia, ya sea con políticas de retención estudiantil, orientación personalizada o programas de apoyo.

Como hemos mencionado en el apartado anterior, el rendimiento estudiantil es la principal causa de abandono, por lo que las posibles soluciones deberían ir en sintonía con esta observación.

Al examinar el rendimiento estudiantil, vimos que la nota media del primer cuatrimestre era de 4,45 que es una nota media pobre, si tenemos en cuenta que el aprobado es sobre 5. Además, si analizamos el promedio de asignaturas aprobadas en el primer cuatrimestre vemos que son 2,3 que, de nuevo, considerando que el primer cuatrimestre de primero de carrera estar formado por 5 asignaturas, es un promedio bajo.

Sumado a que el abandono de esta carrera es bastante más elevado comparado con el resto de las carreras, nos hace llegar a dos posibles conclusiones, o que el grado es demasiado complicado o que el estudiantado no rinde como se espera.

Personalmente me decanto por la segunda opción, la carrera de ADE actualmente tiene 370 plazas, que son muchas, comparado con el resto de las carreras de la UA, además, la nota de corte no suele ser muy elevada, por lo que, el acceso es mucho más sencillo que en otro grado.

No podemos confirmar esta hipótesis con los datos disponibles, pero es posible que muchos de los estudiantes que abandonan el grado lo hagan debido a una falta de comprensión sobre qué implica o sobre lo que es ADE y lo comienzan sin tener una idea clara al respecto, ya que, como hemos mencionado, el acceso al grado no es muy complicado. Es probable que al empezar el programa y darse cuenta de que no cumple sus expectativas, decidan abandonarlo.

Para confirmar esta hipótesis, sería necesario que la universidad solicite a los estudiantes que abandonan que expliquen sus razones. De esta manera, podríamos obtener información más específica y ajustarla para desarrollar medidas más concretas destinadas a prevenir el abandono.

Si resulta que mi hipótesis anterior es cierta, se podría solucionar desde diversas soluciones. Una de ellas sería garantizar que los estudiantes estén bien informados sobre el contenido de la carrera y la conozcan en profundidad, ya sea informándoles mejor en las visitas a la universidad por parte de los institutos o mediante campañas de información. Otra posible solución es reducir las plazas disponibles para forzar que la nota de corte sea más elevada. De esta forma se dificultaría el acceso y fomentaría que aquellos realmente interesados se informen mejor de lo que es ADE. Además, supondría un ahorro de recursos que se podría redirigir a financiar estas medidas.

Otra medida posible, sería concienciar a los docentes acerca de los alumnos en riesgo de abandono, para que puedan brindarles una atención especial. Esto podría incluir la provisión de material adicional de apoyo y la adopción de medidas necesarias para ayudar al alumno a recuperar la motivación y el interés en sus estudios.

Otra medida adicional, sería ofrecer a los estudiantes en riesgo de abandono una entrevista personalizada, donde un profesional pueda brindarles consejos y orientación sobre las mejores opciones a seguir. Durante esta entrevista, se podrían discutir diferentes alternativas, como la continuidad en la carrera actual, la posibilidad de elegir otra carrera más adecuada a sus intereses o incluso considerar la opción de abandonar temporalmente los estudios para explorar otras opciones. Este enfoque individualizado permitiría comprender mejor las necesidades de cada estudiante y un asesoramiento personalizado puede ayudarles a tomar decisiones informadas que se ajusten a sus metas y aspiraciones.

Si se implementa alguna de estas medidas es importante saber medir si han funcionado, para ello, la forma más sencilla seria observar si la tendencia alcista del abandono frena, se detiene o disminuye con el paso del tiempo.

Otra forma para medir el éxito más compleja sería, recopilando los datos de a quien se le ha ayudado con que programa, hacer otro análisis de este estilo comprobando la importancia de las variables de ayuda.

6. CONCLUSIONES

El abandono en primero de carrera en el grado de Administración y Dirección de Empresas (ADE) de la Universidad de Alicante (UA) es muy superior al abandono medio en primero de carrera en España y en la UA.

Tras el análisis que hemos realizado, hemos podido crear varios modelos capaces de predecir a los alumnos que abandonan el grado de ADE en su primer año tras el primer cuatrimestre con una precisión del 70%. Siendo el modelo de regresión logística el que mejores resultados ha obtenido. También hemos concluido que la decisión del abandono suele abarcar el espacio temporal entre el inicio y el final del segundo cuatrimestre. Por este motivo es importante que la predicción y las medidas correspondientes se hagan durante este cuatrimestre.

Gracias al análisis, hemos observado que el factor que más influye en el abandono del grado es el rendimiento del estudiante. Por eso, se pueden crear programas y políticas enfocados en su mejoría. Como clases particulares a aquellos alumnos en riesgo o entrevistas personalizadas que les oriente de forma adecuada sobre su futuro.

También hemos formulado una hipótesis que sería muy interesante responder, la cual es la siguiente. Muchos de los estudiantes que abandonan el grado lo hacen debido a una falta de comprensión sobre qué implica o sobre que es ADE y lo comienzan sin tener una idea clara al respecto, sumado a que el acceso al grado no es muy complicado, es una buena opción para aquellos que no saben que estudiar. Al empezar el programa y darse cuenta de que no cumple sus expectativas, deciden abandonarlo.

Para dar respuesta a esta pregunta es necesario que la Universidad de Alicante consiga recopilar las razones del estudiante sobre su abandono.

En conclusión, con la ayuda de los modelos de predicción y la adopción de medidas preventivas y de apoyo, junto a la recopilación de información detallada sobre el abandono estudiantil, permitiría abordar esta problemática de manera más efectiva y promover una mayor retención en el grado ayudando así a reducir la tasa de abandono y a disminuir el coste de abandono, optimizando los recursos y mejorando la calidad de la educación.

7. BIBLIOGRAFÍA

Ministerio de universidades, Gobierno de España (2021-2022). *Datos y cifras del Sistema Universitario Español*. Recuperado de: https://www.universidades.gob.es/wp-content/uploads/2022/11/Datos_y_Cifras_2021_22.pdf

Unidad Técnica de Calidad, Universidad de Alicante (2023). *Tasas de rendimiento AVAP Grados y Másteres*. Recuperado de: https://utc.ua.es/es/datos/la-ua-en-cifras-apartados/tasas-rendimiento-avap-seguimiento-grados-y-masteres.html

Servicio de Información, Universidad de Alicante (2023). *Preguntas frecuentes sobre: Tasas Académicas. Importe Matrícula*. Recuperado de: https://web.ua.es/es/oia/faq/tasas-academicas-importe-matricula.html

Dekker, Gerben W.; Pechenizkiy, Mykola; Vleeshouwers, Jan M. (2009). *Predicting Students Drop Out: A Case Study*. International Working Group on Educational Data Mining.

Mukesh Kumar, A.J. Singh, Disha Handa (2017). *Literature Survey on Educational Dropout Prediction*, International Journal of Education and Management Engineering (IJEME), Vol.7, No.2, pp.8-19, 2017.DOI: 10.5815/ijeme.2017.02.02

Melvin Vooren, Chris van Klaveren, Ilja Cornelisz, Melisa Diaz Lema, Marta Cannistrà, Tommaso Agasisti (2022). *Predicting dropout in HE across borders: Exploring the common elements between the Netherlands and Italy*. AEDE 2022

Brijesh Kumar Baradwaj, Saurabh Pal (2011). *Data Mining Applications: A comparative Study for Predicting Student's performance*. (IJACSA) International Journal of Advanced Computer Science and Applications

Máté Baranyi, Marcell Nagy, Roland Molontay (2020). *Interpretable Deep Learning for University Dropout Prediction*. SIGITE '20, October 7–9, 2020, Virtual Event, USA

(2023), *Código Postal de España*, Wikipedia, Recuperado de: https://es.wikipedia.org/wiki/C%C3%B3digo_postal_de_Espa%C3%B1a

(2023), *Educación en España*, InfoEmpleo, Recuperado de: https://www.infoempleo.com/guias-informes/empleo-educacion/educacion-espana/sector-educativo-cifras.html

(2021), *Validación cruzada*, IBM, Recuperado de: https://www.ibm.com/docs/es/spss-modeler/saas?topic=settings-cross-validation

8. ANEXOS

Significado de las variables en el fichero "Matrícula"

Nombre	Descripción	Valores
ID_ALUMNO	Identificador alumno (anonimizado)	números aleatorios
ID_ANOACA	Año de matrícula	20AA-AA
ANIONACIMIENTO	Año de nacimiento	AAAA
PAISNACIONALIDAD	Pais de origen (anonimizado)	números aleatorios
SEXO	Sexo	H o M
DEDICACIONESTUDIO	Modalidad dedicación al estudio, a que tiempo	1,2,3 *
OCUPACIONESTUDIANTE	Trabajo del estudiante	00, 01, 02…*
FAMILIANUM	Tipo de familia: Nuclear, Numerosa…	1,2,3 *
NIVELESTUDIOSPADRE	Nivel de estudios del padre	1,2,3… *
NIVELESTUDIOSMADRE	Nivel de estudios de la madre	1,2,3… *
OCUPACIONPADRE	Trabajo del padre	01,02,03… *
OCUPACIONMADRE	Trabajo de la madre	01,02,03… *
FORMAADMISION	Forma de entrada al grado	01,02,03… *
ESTUDIOACCESO	Estudio de acceso 1º vez SUE	01,02,03… *
MUNICIPIOCENTROSEC	Municipio donde curso el estudio de acceso	Código Postal
ANIOFINESTUDIOACCESO	Año cuando terminó el estudio de acceso	AAAA
PAISFINESTUDIOACCESO	Pais donde estudió (anonimizado)	números aleatorios
NATURALEZACENTROSEC	Privado, concertado o público	1,2,3 *
NUEVOSUE	Es nuevo en el SUE	0,1 *
ANIOACCESOSUE	Año cuando accede al SUE	AAAA
NOTAADMISION	Nota con la que entró al grado	número
PAISRESIDENCIA	país donde reside (anonimizado)	números aleatorios

Tabla 22. Variables del fichero "Matricula"

** Cada número tiene asociado un significado, se especificará en el trabajo si requiere*

Significado de las variables en el fichero "Calificaciones"

Nombre	Descripción	Valores
ID_ALUMNO	Identificador alumno (anonimizado)	números aleatorios
ID_ANOACAD	Año en el que se cursa la asignatura	20AA-AA
ID_ASIGNATURA	Identificador asignatura (anonimizado)	números aleatorios
ID_CURSO	Curso al que pertenece la asignatura	1,2,3,4
ID_GRUPO	Grupo de la asignatura en el que se cursa	numérico
ID_SEMESTRE	Semestre al que pertenece la asignatura	1,2
NOTA_C1	Nota en la convocatoria C1	numérico
NOTA_C2	Nota en la convocatoria C2	numérico
NOTA_C3	Nota en la convocatoria C3	numérico
NOTA_C4	Nota en la convocatoria C4	numérico

Tabla 23. Variables del fichero "Calificaciones"

Significado de la ocupación de los padres

Número	Descripción
00	Ocupaciones militares
01	Directores y gerentes
02	Técnicos y profesionales científicos e intelectuales
03	Técnicos y profesionales de apoyo
04	Empleados de tipo contable y administrativo
05	Trabajadores de los servicios de restauración, personales, protección y vendedores de los comercios
06	Trabajadores cualificados en la agricultura y en la pesca
07	Artesanos y trabajadores cualificados de las industrias
08	Operadores de instalaciones y maquinaria y montadores
09	Trabajadores no cualificados
10	Parado
11	Jubilado
12	Amo de casa

13	Incapacitado para trabajar
14	Otra situación
88	No aplica
99	No consta

Tabla 24. Significado de las categorías de la variable ocupación de los padres

Significado del acceso a la universidad

Abreviatura	Descripción
PAU	Prueba de Acceso a la Universidad
Grad. Sup.	Grado Superior
<25	Mayores de 25 años
<40	Mayores de 40 años con acreditación de experiencia laboral
<45	Mayores de 45 años
Extr.	Extranjero
Conv.	Convalidación

Tabla 25. Significado de las abreviaturas en el acceso a la universidad