Sistema de Alerta Temprana de Estudiantes Universitarios con Riesgo de Abandono

Algoritmos de Machine Learning y Datos Abiertos

Proyecto presentado al concurso "II Datathon UniversiDATA"

FASE FINAL 3 FEBRERO 2025

No citar este borrador sin permiso del autor

Fernando A. López Hernández
Catedrático de Universidad
Universidad Politécnica de Cartagena

Abstract

Uno de los principales problemas del sistema educativo español es el alto porcentaje de estudiantes que abandonan la educación superior. Según el informe "Datos y Cifras del Sistema Universitario Español (2022-2023)", uno de cada tres estudiantes universitarios deja sus estudios o cambia de titulación. Implementar un sistema de alerta temprana basado en el análisis de datos masivos puede ser una herramienta eficaz para reducir esta tasa de abandono. Utilizando una base de datos de más de 200.000 estudiantes de de primer curso pertenecientes a cinco universidades españolas de seis cursos académicos se aplican dos algoritmos de aprendizaje automático, Multivariate Adaptive Regression Splines (MARS) y Bosques Aletorios, para asignar a cada estudiante una probabilidad de abandono, permitiendo identificar a aquellos estudiantes con mayor riesgo de abandono. Con base en estas probabilidades, se pueden diseñar estrategias específicas para tratar a los estudiantes según sus características individuales. Esta metodología no solo ayuda a disminuir el abandono, sino que también optimiza los recursos educativos y mejora la experiencia de los estudiantes y de sus familias.

Word count: 8833

1 Introducción

El abandono universitario representa uno de los problemas más desafiantes que enfrenta la educación superior a nivel global, convirtiéndose en una prioridad estratégica para los

sistemas universitarios en todo el mundo (Constante-Amores et al. 2021; Kehm, Larsen, and Sommersel 2019; Jia and Maloney 2015; Delogu et al. 2024; Núñez-Naranjo 2024). La proporción de estudiantes que no completan un programa de grado universitario dentro de la duración teórica varía desde menos del 20% en el Reino Unido, Israel, Suiza e Irlanda, hasta más del 40% en Brasil, Eslovenia, Chile, Bélgica (comunidad francófona), Suecia, Italia, Austria y Estonia (Aina et al. 2022). Este fenómeno, que tiene fuertes implicaciones tanto para los estudiantes y sus familias como para las instituciones y la sociedad en general se ha mantenido de manera persistente sin que se hayan identificado soluciones claras para reducir su impacto. En un mundo donde el capital humano es esencial para el desarrollo económico y social, la retención estudiantil debe ser una prioridad para todos los sistemas educativos, incluida la educación superior.

La búsqueda de soluciones a un problema que, de manera persistente, se repite año tras año en las universidades ha generado una extensa producción de literatura científica. Estos estudios han abordado el fenómeno desde múltiples perspectivas, presentando tanto análisis generales (Lorenzo-Quiles, Galdón-López, and Lendínez-Turón 2023; Aina et al. 2022) como casos específicos en diferentes universidades y titulaciones (Seo et al. 2024; Espinoza et al. 2024). Aunque existe una gran heterogeneidad en las investigaciones, en general, todas convergen en la identificación de un problema que es multifactorial y que está influenciado por una combinación de factores personales, sociales, económicos e institucionales (Constante-Amores et al. 2021).

Hay un elevado número de estudios identifican elementos individuales como el género (Espinoza et al. 2024), la edad (Andreas Behr et al. 2020), el origen geográfico (Delogu et al. 2024; Constante-Amores et al. 2021), la capacidad académica (Serrano 2013), la motivación (De la Cruz-Campos et al. 2023), la salud mental (Kawada 2014), el rendimiento académico previo (Bernardo et al. 2021), la vocación o las expectativas profesionales (Bernardo et al. 2021; Belloc, Maruotti, and Petrella 2010). Todos estos factores, en mayor o menor medida, son determinantes del abandono de los estudios universitarios.

El estatus socioeconómico y el entorno familiar también tiene un papel crucial. Palacio Sprockel, Vargas Babilonia, and Monroy Toro (2020) realiza un estudio bibliométrico centrado impacto de las variables socioeconómicas y sociodemográficas como causantes del bajo rendimiento y del consecuente abandono. En general los estudios coinciden en que los estudiantes de familias con menores ingresos enfrentan mayores barreras financieras y suelen tener menor acceso a recursos de apoyo. Además, los costos de matrícula y manutención pueden ser insostenibles para muchos, especialmente en países con sistemas de educación superior menos accesibles. En este grupo de factores Walsh and Robinson Kurpius (2016) señala la importancia de los estudios realizados por los padres como factor vinculado al menor nivel de abandono de los estudios universitarios. En la misma línea, Aina (2013) analiza el caso italiano. Contini, Cugnata, and Scagni (2018) explora una muestra amplia de estudiantes de las universidades de Italia y destaca la importancia que tiene el nivel educativo de los progenitores en el abandono. En el caso español el informe de 2024 del Ministerio de Ciencia Innovación y Universidades también traza un completo análisis socioeconómico del estudiantado (Ministerio de Ciencia, Innovación y Universidades 2024).

El diseño del sistema educativo también impacta el abandono. Los sistemas con menor

flexibilidad curricular, altos requisitos académicos y pocos programas de apoyo suelen tener tasas más altas de deserción. Aspectos institucionales y estructurales, como la calidad del apoyo estudiantil y las políticas de retención implementadas por las universidades (A. Behr et al. 2022; Galve-González, Bernardo, and Castro-López 2024). La integración social y académica también es determinante (Tinto 1975) cuando los estudiantes no logran establecer conexiones con el entorno universitario. Por su parte, A. Behr et al. (2022) destacan el papel de las políticas institucionales, incluyendo la orientación y el apoyo estudiantil, como elementos clave para mejorar la retención.

También todas estas investigaciones coinciden en que es en el primer año de estudios universitarios donde principalmente se produce el abandono de los estudios (Rodriguez-Muñiz et al. 2019; Goller, Diem, and Wolter 2023). Aunque es un fenómeno que también tiene lugar en los siguientes años, su incidencia es menor. Las diferencias entre ramas de conocimiento son indudables, encabezadas por las llamadas titulaciones STEM (Ciencia, Tecnología, Ingeniería y Matemáticas, por sus siglas en inglés), que por otra parte son aquellas en las que se necesita mayor capital humano para adaptarse al imparable proceso de digitalización de la economía y de todas las actividades diarias.

Un enfoque metodológico que se ha prevalecido en los estudios que buscan soluciones al abandono universitario ha sido el uso de métodos estadísticos para identificar y cuantificar los factores determinantes. Sin embargo, en los últimos años, el creciente volumen de datos y el avance de los algoritmos de aprendizaje automático (ML, por sus siglas en inglés) han otorgado un nuevo impulso a estas investigaciones. Nuestra propuesta de investigación va en esta línea en busca de contribuir al conocimiento del fenómeno en el caso español. Aprovechamos la iniciativa **UniversiDATA**, que recopila microdatos detallados de estudiantes de 5 universidades a lo largo de seis cursos académicos, para aplicar algoritmos de Machine Learning que identifiquen patrones clave y profundicen en la comprensión del abandono en el contexto español.

Este artículo se estructura en 5 secciones. El la segunda sección se presenta una breve revisión de la literatura centrada en recopilar trabajos sobre abandono en los que se utilicen técnicas de ML. En la sección tercera se presenta la base de datos y la metodología. En la cuarta los resultados y la quinta sección concluye.

2 Revisión de la literatura

Debido a la gran relevancia del problema del abandono en el sistema universitario, la producción científica, tanto a nivel internacional como nacional, es enorme. Varios artículos han recopilado investigaciones internacionales que abordan esta cuestión desde diferentes ángulos (Andreas Behr et al. 2020; Lorenzo-Quiles, Galdón-López, and Lendínez-Turón 2023; Aina et al. 2022; De la Cruz-Campos et al. 2023; Mayra Alban and Mauricio 2019). En el caso español, Álvarez-Ferrándiz (2021) destaca estudios sobre abandono en universidades presenciales. De la Cruz-Campos et al. (2023) realiza un análisis sistemático de estudios sobre el abandono universitario, con especial énfasis en Andalucía. Ortiz-Lozano et al. (2020) analiza cómo los datos sociodemográficos y académicos pueden predecir el abandono temprano. En Educación XX1 Constante-Amores et al. (2021) analiza los factores

asociados al abandono universitario en el caso de en la Universidad Complutense de Madrid. Algunas revistas también editaron un monográfico, dirigido por Colás Bravo (2015), que trata el problema desde múltiples perspectivas. Además de todas estas investigaciones, diversas instituciones han elaborado informes en los que se aborda el abandono universitario: el Ministerio de Universidades (Mellizo-Soto 2022), la Confederación de Rectores de las Universidades Españolas (CRUE)[1], el Ministerio de Ciencia, Innovación y Universidades (Ministerio de Ciencia, Innovación y Universidades 2024), incluso el Consejo Social de la Universidad Carlos III de Madrid elaboró dos informes en 2014 y 2019 (Madrid 2019, 2014) para analizar esta problemática.

2.1 Investigaciones que aplican algoritmos de ML

Para investigar los determinantes del abandono universitario, investigaciones previas han empleado modelos econométricos estándar, como Mínimos Cuadrados Ordinarios (OLS), Modelos Lineales Generalizados (GLM) o modelos de datos panel (Aina 2013; Belloc, Maruotti, and Petrella 2010). Sin embargo, existe ahora un consenso de que estas herramientas no son intrínsecamente predictivas, lo que ha llevado a muchos autores a sugerir el uso de algoritmos de ML para abordar el problema del abandono universitario. La literatura recoge una amplia variedad de técnicas, que van desde las más clásicas como la regresión logística (Cho, Yu, and Kim 2023; Jia and Maloney 2015) pasando por toda una variedad de algoritmos como Naïve Bayes (Kotsiantis, Pierrakeas, and Pintelas 2004), k-means (Erdogan and Timor 2005), árboles de decisión (Segura, Mello, and Hernández 2022; Sung-Hyuk and Tappert 2009; Kabra and Bichkar 2011), redes neuronales (M. Alban and Mauricio 2019), Random Forest (A. Behr et al. 2020; Urbina-Nájera, Camino-Hampshire, and Cruz Barbosa 2020), Support Vector Machine (Cho, Yu, and Kim 2023), o incluso métodos de procesamiento de lenguaje natural (Won et al. 2023). Todas ellas han mostrado ser altamente efectivas en el análisis y predicción de este fenómeno.

Son frecuentes las investigaciones que utilizan varios algoritmos de ML comparando los resultados que ofrecen cada uno de ellos. Por ejemplo, Yu et al. (2010) analiza el abandono a través de árboles de clasificación, MARS y redes neuronales. Kim et al. (2023) utiliza modelos basados Análisis de Compontenes Principales y K-means clustering para mejorar las predicciones. Kabathova and Drlik (2021) aplica seis de los algoritmos más conocidos de ML para comparar la capacidad predictiva de cada uno de ellos. Cho, Yu, and Kim (2023) también utiliza una amplia batería de algoritmos Regresión Logística, Árboles de Decisión, Random Forest, Support Vector Machine, Deep Neural Network, and LightGBM (Light Gradient Boosting Machine) siendo este último el que mejor resultados ofrece. Delogu et al. (2024) aplica 5 algoritmos (Logistic, RF, GBM, NN, LASSO) a una gran base de datos de 230.336 universitarios italianos.

En general todos los algoritmos mostraron buenas capacidades de predicción sin que a prori exista un claro modelo ganador.

[1] https://www.crue.org/wp-content/uploads/2024/06/UEC-24_Avance-04.pdf

3 Datos y metodología

3.1 Datos y estadísticas descriptivas

Esta investigación analiza la base de datos abierta más extensa que recopila y ofrece datos estadísticos sobre universidades en España. Hasta donde sabemos la más completa y extensa disponible en España y solo puede compararse en tamaño de información con la reciente publicación de Delogu et al. (2024). Cinco universidades son incluidas en el estudio: la Universidad Autónoma de Madrid, la Universidad Complutense de Madrid, la Universidad Carlos III de Madrid, la Universidad Rey Juan Carlos y la Universidad de Valladolid. Los datos analizados abarcan seis cursos académicos, desde 2017-2018 hasta 2022-2023. Los conjuntos de datos están anonimizados para proteger la privacidad de los individuos y han sido puestos en abierto con el objetivo de facilitar el desarrollo de políticas educativas basadas en la evidencia del dato. Toda la información se encuentra disponible de forma abierta en el portal **UniversiDATA**[2].

Nuestra investigación se centra en los estudiantes de primer curso que se matriculan por primera vez en la universidad ya que es en este primer contacto del estudiante con el entorno universitario cuando se produce la mayor tasa de abandono. Tras un cuidadoso proceso de ingeniería de datos se seleccionaron un total de 203.941 estudiantes que cumplían los requisitos. Asociados a estos individuos se seleccionaron de diversas bases de datos disponibles en el portal varios factores que la literatura ha asociado al abandono.

La Tabla 1 lista las variables seleccionadas para esta investigación con una breve descripción. Para caa uno de los estudiantes se calculó la tasa de rendimiento (TR) del primer año y se planteó la hipótesis de que una mayor tasa de rendimiento (TR) en el primer curso reduce la probabilidad de abandono. El rendimiento académico del estudiante en el primer año, medido como el porcentaje de créditos aprobados respecto al total matriculado (TR04), es la variable que, con diferencia, tiene más peso en el abandono. Este hallazgo confirma que la *integración académica* propuesta por Tinto (1975) tiene una relevancia crucial. Sin embargo, hay que interpretar este resultado con cautela, ya que la propia decisión de abandonar podría anteceder al bajo rendimiento. Es decir, un estudiante que ya haya decidido abandonar podría no esforzarse en los exámenes. Basándonos en esta hipótesis consideramos que un estudiante de primer grado se considera que 'abandona' los estudios si su tasa de rendimiento es inferior a 0,4. Un total de 11,9% del los estudiantes cumplen esta condición.[3].

La Tabla 2 muestra la distribución porcentual de las variables categóricas (factores) del estudiante dependiendo de si la tasa de rendimiento es o no inferior a 0,4 (TR04). Los resultados observados muestran fuertes diferencias, por ejemplo, los estudiantes matriculados en centros adscritos presentan valores de TR04 inferiores a los no adscritos. El género es otro factor que genera grandes diferencias, en el conjunto de datos hay 121358 mujeres (el 59,8%) de las cuales solo un 8,85% tienen TR<0,4 mientras que el 16,75% presentan tasas de

[2] https://www.universidata.es/

[3] Advertencia: La definición de abandono es un proxy asociado a la tasa de rendimiento. Es una aproximación pobre y aunque existe alta correlación entre la tasa de rendimiento y el abandono hay varias cuestiones que pueden plantearse: La relación causal es compleja. Objetivo: visibilizar el problema del abandono

rendimiento iguales o superiores. Esta diferencias en género son similares a las identificadas en el caso italiano (Delogu et al. 2024). También la rama de conocimiento asociada a la titulación en la que el estudiante está matriculado se observan diferencias muy grandes en la variable TR04. En IA hay un 24,76% de estudiantes con TR<0,4 frente al 7,81 en CS. Finalmente, la Tabla 3 muestra la distribución de las variables asociadas a la titulación que cursa el estudiante.

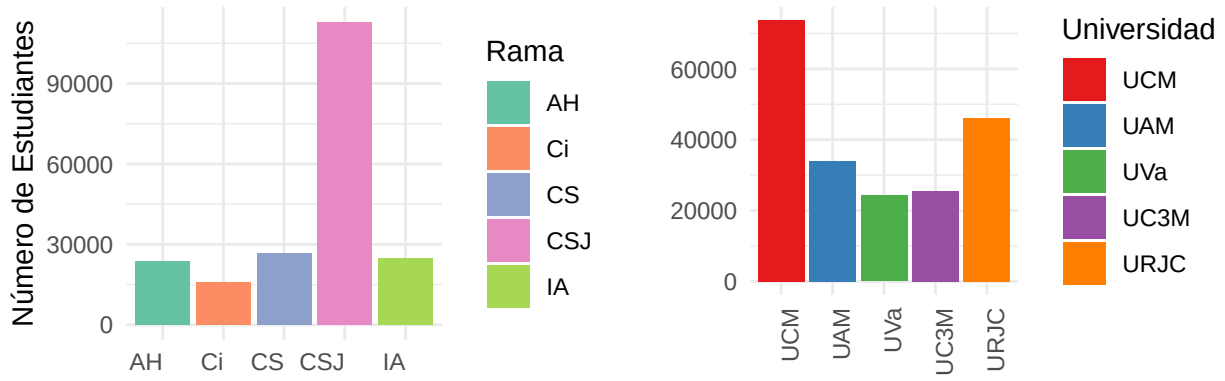


Figure 1: Histogramas número de observaciones. Rama de Conocimiento (Izda) y Universidad (Dcha)

3.2 Metodología

Los modelos econométricos quedan a la sombra del creciente boom de los modelos ensamblados y de redes profundas. El mundo los datos masivos ha traído un creciente interés por tipos de modelización que buscan la mayor precisión posible en el resultado. Esta alta precisión en busca de la mejor combinación no lineal hace que la interpretabilidad sea un punto fuerte a desarrollar (Gilpin et al. 2018). Para ello introducimos dentro del proceso de estimación una fase de deslinealización a través de la técnica de Multivariate Adaptive Regression Splines MARS (Milborrow 2011), aplicado a un modelo de regresión logística,

Table 1: Descripción de las variables (N = 203.941)

Variable	Descripción	Media(dt)
Variable Dependiente		
TR04	1 si la tasa de rendimiento del estudiante es inferior a 0.4; 0 en otro caso	0.119 ()
Variables independientes		
Género	1 si es mujer; 0 en otro caso	0.595 ()
Adscrito	1 si está matriculado en un centro adscrito; 0 en otro caso	0.078 ()
Doble	1 si el estudiante está matriculado en un doble grado; 0 en otro caso	0.119 ()
Rama	Factor con 5 categorías para cada una de las ramas de conocimiento correspondiente a la titualción en la que el estudiante está matriculado: AH Artes y Humanidades; Ci Ciencias; CS Ciencias de la Salud; CSJ Ciencias Sociales y Jurídicas; IA Ingeniería y Arquitéctura	AH 0.116; Ci 0.077; CS 0.131; CSJ 0.554; IA 0.121
Universidad	Factor con 5 categorías identificando la universidad (UCM U. Complutense de MAdrid; AUM U. Autónoma de Madrid; UC3M U. Carlos III de Madrid; URJC U. Rey Juan Carlos; UVa U. de Valladolid)	UCM 0.362; UAM 0.167; UC3M 0.125; URJC 0.227; UVa 0.119
Dato imputado al estudiante según la titulación que cursa y el curso académico en el que se matricula		
nacceso	Número medio de alumnos que acceden a cada titulación	304,9 (330,3)
EdadMedia	Edad media de los alumnos de la titulación	22,35 (1,158)
Municipio	Porcentaje de estudiantes que residen en el mismo municipio en cada titulación	0.350 (0.091)
Provincia	Porcentaje de estudiantes que residen en la misma provincia en cada titulación	0.697 (0.157)
NotaMediana	Nota mediana de acceso a la titulación que cursa el estudiante	10.299; (1.669)
NotaMin	Nota mínima de acceso a la titulación que cursa el estudiante	media=5.168
MadreUniv	Porcentaje de estudiantes cuya madre tiene estudios universitarios en la titulación en la que está matriculado el alumno	0.395 ()
PadreUniv	Porcentaje de estudiantes cuyp padre tiene estudios universitarios en la titulación en la que está matriculado el alumno	0.356 ()
Dato imputado al estudiante según la universidad en la que está matriculado y curso académico		
Ayudantes	Porcentaje de profesorado ayudante doctor sobre el total de profesorado de la universidad	0.094 (0.028)
Asociados	Porcentaje de profesorado asociado sobre el total de profesorado de la universidad	0.248 (0.089)

Valores medios y desviaciones típicas son globales para todos los cursos académcios y todas las titualciones

Tasa de rendimiento = número de créditos superados entre el número de créditos matriculados. Los datos se filtraron para solo considerar estudiantes cuyo número total de créditos matriculados estuviera en entre 30 y 90 ambos valores incluidos

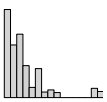
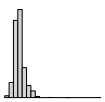
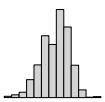
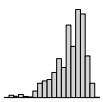
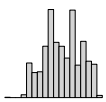
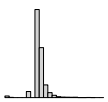
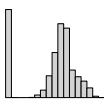
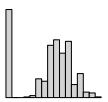
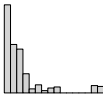
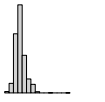
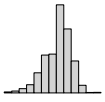
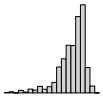
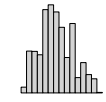
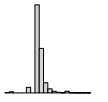
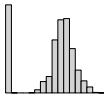
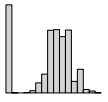
Table 2: Número estudianes (N) y distribución porcentual en base a TR04

	$N(TR \geq 0,4)$	$N(TR < 04)$	$\%(TR \geq 0,4))$	$\%TR < 0,4$
Género (mujer)				
Hombre	68751	13832	83,25	16,75
Mujer	110973	10385	91,44	8,56
Adscrito				
No	165069	23035	87,75	12,25
Si	14655	1182	92,54	7,46
Doble				
No	157033	22715	87,36	12,64
Si	22691	1502	93,79	6,21
Universidades				
UCM	64883	8944	87,89	12,11
UAM	30701	3312	90,26	9,74
UVa	19729	4545	81,28	18,72
UC3M	23419	2164	91,54	8,46
URJC	40992	5252	88,64	11,36
Rama				
AH	20506	3189	86,54	13,46
Ci	13675	2096	86,71	13,29
CS	24679	2091	92,19	7,81
CSJ	102236	10712	90,52	9,48
IA	18628	6129	75,24	24,76

Note:

Porcentajes calculados sobre suma elementos fila

Table 3: Histogramas de las variables correspondientes a la titulación en la que el estudiante está matriculado (Si=TR04<0,4)

TR04	nacceso	EdadMedia	Municipio	Provincia	NotaMed	NotaMin	MadreUniv	PadreUniv
Si								
No								

para desarrollar un sistema de alerta temprana que identifique el riesgo de abandono de cada estudiante. A diferencia de otros algoritmos similares, MARS tiene la ventaja de seleccionar automáticamente las variables más relevantes que determinan las probabilidades de abandono, considerando también impactos no lineales de las variables independientes. Una de las principales características de este algoritmo es su fácil interpretación. Esta facilidad de interpretación es crucial, ya que permite a las universidades entender claramente los factores que influyen en el abandono y, por tanto, tomar decisiones rápidas y efectivas para reagrupar a los estudiantes en función de su riesgo.

El algoritmo MARS es una técnica de regresión flexible, no paramétrica y por partes, introducida por Friedman (1991). Esta metodología impulsada por datos es especialmente útil para identificar no linealidades en modelos de regresión sin hacer suposiciones previas sobre su forma funcional, las variables explicativas o su número. La característica principal de esta técnica es que el modelo econométrico considera diferentes pendientes de regresión en distintos intervalos para cada predictor.

A diferencia de técnicas de regresión lineal más conocidas, MARS no asume que los coeficientes sean estables a lo largo del rango de cada variable y, en su lugar, utiliza splines para ajustar funciones continuas por partes. En general, MARS construye una función lineal por partes para capturar relaciones no lineales de manera adaptativa. La principal ventaja de esta metodología, en comparación con algoritmos similares (como modelos polinómicos, LASO o GAM), es la simplicidad del modelo econométrico resultante y su interpretación intuitiva. Además, se ha reportado que los modelos MARS funcionan satisfactoriamente en términos de costo computacional, independientemente de la dimensión del conjunto de datos, destacando la escalabilidad computacional de este algoritmo para tamaños de muestra grandes en comparación con otros algoritmos.

Información técnica detallada sobre el algoritmo MARS y su aplicación en análisis de datos se puede encontar en Hastie, Tibshirani, et al. (2009) y Hoang, Chen, and Liao (2017).

3.2.1 Principios Básicos de MARS

Al igual que en cualquier modelo de regresión, el objetivo de esta metodología es construir un modelo econométrico que explique la variación de una variable dependiente $Y = (y_1, \ldots, y_n)'$ con un conjunto de posibles variables explicativas $X = (X_1, \ldots, X_p)$, donde $X_i = (x_{1i}, x_{2i}, \ldots, x_{ni})'$. Para alcanzar este objetivo clásico, MARS utiliza las llamadas funciones básicas (BF) de la forma $(x - c)_+ = \max\{0, x - c\}$ y $(c - x)_+ = \max\{0, c - x\}$, donde el subíndice "+" indica que la función toma solo el valor positivo o cero en caso de una diferencia negativa. Tales pares de funciones lineales se denominan funciones de bisagra (o funciones truncadas a ambos lados), y la constante c denota un nudo donde cambia la pendiente. La colección de todas las BF posibles, $\mathcal{C}$, es dada por:

$$\mathcal{C} = \{(x - c)_+, (c - x)_+\} \quad \text{con} \quad c \in \{x_{1i}, x_{2i}, \ldots, x_{ni}\} \quad \text{y} \quad i = 1, \ldots, p$$

Cada función es lineal a trozos con un nudo ('knot') c en cada x_{ij}, y si todos los valores de entrada son distintos, hay np funciones de bisagra o, de manera equivalente, $2np$ funciones

básicas. Usando estas BF, la estrategia de construcción del modelo es similar a una regresión paso a paso hacia adelante clásica, que utiliza las funciones del conjunto $\mathcal{C}$ y sus productos como entradas. La expresión final del modelo es:

$$Y = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X) + \epsilon$$

donde $h_m(X)$ es una BF o un producto de dos o más tales funciones, si se permiten interacciones entre variables, o quizás el predictor original, si tiene un impacto lineal en la variable dependiente. Los coeficientes β_m se estiman minimizando la suma del cuadrado de los errores (SCE), similar a un modelo de regresión lineal estándar.

3.2.2 El Paso Hacia Adelante

El proceso de entrenamiento del modelo selecciona y añade de manera iterativa algunas funciones bisagra al modelo (o al predictor original). Durante cada paso del proceso de entrenamiento, MARS selecciona nuevos términos que minimizan la suma de los cuadrados de los errores (SCE) utilizando los mínimos cuadrados ordinarios (OLS). En este paso hacia adelante, el algoritmo MARS comienza con un modelo que solo incluye el término de intersección β_0. En cada paso subsiguiente, se selecciona y añade al modelo un par reflejado de funciones bisagra y un predictor original. Este par puede entrar directamente al modelo o, alternativamente, multiplicarse por una función básica (BF) ya existente en el modelo, convirtiéndose en nuevas BFs. Este segundo caso permite modelar la interacción entre los diferentes predictores.

El proceso de paso hacia adelante continúa hasta que se cumple una de varias condiciones que pueden ser impuestas a priori, como: (i) alcanzar el número máximo de términos del modelo (elegido por el usuario) antes de la poda, o (ii) cuando añadir un término cambia (R^2) a un valor inferior al umbral seleccionado por el usuario (por ejemplo, 0,001). La búsqueda de funciones bisagra en cada paso puede realizarse mediante fuerza bruta, pero este proceso puede acelerarse utilizando una heurística que reduce el número de términos parentales a considerar Friedman (1991).

En general, al final de este proceso, obtenemos un modelo grande en la forma descrita en la ecuación (1). El modelo MARS obtenido en este paso hacia adelante es adaptable y puede mostrar un alto grado de flexibilidad, lo que puede resultar en sobreajuste si no se toman medidas correctivas. Para resolver el problema del sobreajuste y construir un modelo con mejor capacidad de generalización, se debe aplicar un procedimiento de poda.

3.2.3 El Proceso de Poda

Aunque existen otros métodos, MARS típicamente aplica un procedimiento de eliminación hacia atrás para podar el modelo. Así, en la segunda fase del algoritmo MARS, se aplica un procedimiento de eliminación "uno a la vez" en el cual se elimina repetidamente la función básica que tiene la menor contribución al modelo. Esta poda se basa en el criterio de

validación cruzada generalizada (GCV), propuesto originalmente por (**?**)} y adaptado por Friedman and Silverman (1989)}. La expresión del GCV es:

$$GCV = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \left(1 - \frac{\hat{C}(M)}{n}\right)^2$$

donde $\hat{y}_i$ son los valores ajustados y $\hat{C}(M) = C(M) + dK$, siendo $C(M)$ el número de parámetros que se están estimando (el número de funciones básicas linealmente independientes sin el término de intersección); K es el número de nudos seleccionados en el proceso hacia adelante, y d representa el costo de cada optimización de función básica. Generalmente, $d = 2$ si el modelo no involucra términos de interacción, y $d = 3$ en caso contrario Friedman (1991)}. Por lo tanto, la fórmula del GCV ajusta el SCE para tener en cuenta la flexibilidad del modelo. Valores más grandes de d resultan en menos nudos y estimaciones de funciones más suaves. La mejor aproximación de MARS es aquella con el valor más bajo de GCV. Para obtener una medida similar a $ R^2 $, se puede estandarizar el coeficiente GCV y definir un nuevo coeficiente como:

$$GRSq = 1 - \frac{GCV}{GCV_{tot}}$$

donde GCV_{tot} es el GCV de un modelo que solo contiene el término de intersección.

Es importante notar que el SCE bruto es inadecuado para comparar modelos, ya que el SCE siempre aumenta al eliminar términos de MARS; por ende, la aplicación de este criterio en el paso hacia atrás siempre resulta en seleccionar el modelo más grande. Por lo tanto, el criterio GCV se utiliza para encontrar el mejor modelo general a partir de una secuencia de modelos ajustados. El paso hacia atrás comienza con el modelo completo devuelto por el paso hacia adelante. Luego, en cada paso, se elimina el término que resulta en el submodelo con el SCE más bajo. Este proceso iterativo continúa hasta alcanzar el modelo que solo tiene el término de intersección. Finalmente, el paso hacia atrás selecciona el modelo final: el submodelo con el GCV más bajo.

3.2.4 Hyperparámetros del Algoritmo MARS

Existen varios parámetros de ajuste asociados con el algoritmo MARS que los investigadores pueden seleccionar. Citamos aqui algunos de los mas importantes.

- *Umbral para la terminación del paso hacia adelante*: se establece un criterio para determinar cuándo debe finalizar el proceso de selección.
- *Número mínimo de observaciones entre nodos*: se especifica cuántas observaciones son necesarias antes de considerar la creación de un nuevo nodo.
- *Selección de variables*: se decide qué variables se incluirán en su forma lineal
- *Orden de interacción entre variables*: se define cuántas interacciones se permiten entre las variables.

- *Número máximo o mínimo de nodos*: se limita cuántos nodos pueden formarse durante el proceso.

Esta lista no es exhaustiva y se puede encontrar más información sobre estos parámetros en el trabajo de Friedman (1991).

3.2.5 AUC. Medida para evaluar los modelos

El AUC (Área Bajo la Curva) es un indicador de calidad para modelos logit que mide su capacidad discriminativa. Representa el área bajo la curva ROC, que evalúa la relación entre sensibilidad y especificidad. Un AUC cercano a 1 indica excelente desempeño, mientras que valores próximos a 0.5 reflejan un modelo aleatorio.

4 Resultados

Se aplicó el algoritmo MARS al conjunto de datos con la librería earth del software libre R (Trevor Hastie and Thomas Lumley's leaps wrapper. 2024). Se aplicó un procedimiento de ajuste de hiperparámetros del algoritmo, que combinó un ejercicio de validación cruzada k-fold con una búsqueda en malla (grid search) con el objetivo de maximizar el AUC. El ejercicio de validación cruzada con k=5-folks se utilizó para identificar el número máximo de términos, optiendo los mejores resultados limitando a 25 el número de términos en el modelo. Fijando el número máximo de términos en 25 se realizó una búsqueda de hiperpárametros (minspan=número mínimo de observaciones entre knots y número máximo de interacciones entre variables) para calibrar la capacidad predictiva del modelo seleccionando una submuestra de entrenamiento del 70% y una de test del 30%. Este proceso de búsqueda determinó un valor minspan=3000 y 2 términos máx de interacción. El modelo MARS-logit óptimo selecciona 22 términos de interacción entre variables y utiliza 9 predictores para analizar el abandono universitario. El AUC del modelo final resultó ser AUC-MARS=0,72.

Otros dos algoritmos se aplicaron a esta base de datos, un modelo logit y el algortitmo Random Forest. El modelo logir logró un AUC-Logit=0.70, levemente inferior al AUC-MARS. De forma paralela se lanzó en algoritmo de Random Forest con el paquete de R ranger (Wright and Ziegler 2017) al que igualmente se le aplicó una búsqueda de hiperparámetros para determinar el número óptimo de predictores. Los mejores resultados se obtuvieron con 5 predictores. En este caso el AUC-RF=0,75 fue levemente superior al que se obtuvo con MARS

La Tabla 4 muestra los resultados del modelo MARS-logit. En primer el algoritmo no considera relevantes, ni para incluirlas linealmente ni tampoco ninguna de las BF, las variables "Universidad"; NotaMinima"; "EdadMedia"; "MadreUniv"; "PAsoc"; "PAyuDoc". En segundo lugar el algoritmo seleccionó variables dicotómicas *Género*, *Adscrito* junto con *RamaIA*. Los resultados indican que el género (*Mujer* ser mujer), la rama de conocimiento en la que el estudiante se matricula (solo es relevante IA Ingeniería y Arquitectura) junto con el hecho de realizar los estudios en un centro adscrito (*Adscrito*) son factores relevantes. Mientras que ser mujer y estar matriculado en un centro adscrito disminuyen la probabilidad de abandono, el realizar los estudios en la rama de IA incrementa notablemente.

En tercer lugar el algoritmo MARS ha seleccionado algunas funciones básicas (BF) que tienen un impacto no lineal. Las variables, *nacceso*, *provincia*, *NotaMediana* y *PadreUniv* se incluyen en este grupo. El algoritmo divide la nota (mediana) el punto de corte en 8,1 de tal forma que un estudiante que entra en una titulación con nota superior a 8,1 disminuye la probabilidad $(NotaMediana - 8,1)_+$ y con nota inferior $(8,1 - NotaMediana)_+$ incrementa la probabilida de pertenecer al grupo de "abandono". Una BF es seleccionada de *nacceso* $(nacceso - 515)_+$ de tal forma que iniciar estudios en titulaciones donde acceden muchos estudiantes (>515 como media anual) incrementa la probabilidad de abandono. Solo una función básica en incluida en el modelo $(n.acceso - 515)_+$ y aparece con signo positivo, indicando que cuando el número de alumnos es inferior 515 el algoritmo no lo considera relevante y no es incluido en la modelización. Delogu et al. (2024) también identifica que esta variable es relevante cuando analiza el abandono en el caso italiano. El porcentaje de estudiantes en una titulación superior al 82% disminuye la probabilidad mientras que si el porcentaje es superior al 82% incrementa la probabilidad. Los estudiantes que están matriculados en titulaciones donde el % de padres con estudios universitarios es mayor que 0.28 incrementa la probabilidad de fracaso.

En tercer lugar varias interacciones entre dos variables mejoran el modelo. La interacción entre *Mujer* y *NotaMediana* es relavante de tal forma que si es muje y está en una titualción con nota mediana superior a 10.9 increment la ....los anteriores resultados son matizados por la incorporación de interacciones entre variables

Una cuestión relevante que se obtiene del algoritmo es la influencia de la geografía sobre la variable TR04. El porcentaje de estudiantes que en la titulación pertenecen a la misma provincia y al mismo municipio aparecen interactuando con otras variables. También Yu et al. (2010) identifica la residencia como un predictor crucial para la retención.

La interacción entre Provincia y nacceso es relevante. Altos porcentajes (>0,72) de estudiantes de la provincia en titulaciones con nacceso <515 reduce la probabilidad de abandono. $(515 - nacceso)_+ \mathring{u} (Provincia - 0,72)_+$

4.1 La importancia de las variables

La Figura 2 muestra la importancia relativa de las variables obtenida mediante dos algoritmos de aprendizaje: MARS y Random Forest. En el eje horizontal se enumeran las variables analizadas, mientras que el eje vertical muestra su peso en el modelo predictivo, medido en función de su contribución a la precisión. Cada barra representa la importancia asignada por cada algoritmo, diferenciada por colores. La comparación destaca cómo cada método prioriza distintos factores según su enfoque: MARS se centra en relaciones lineales y no lineales ajustadas, mientras que Random Forest evalúa interacciones complejas y jerarquías entre variables.

El gráfico muestra la importancia relativa de las variables según los algoritmos MARS (izquierda) y Random Forest (derecha). En ambos casos, las variables **"Rama"** y **"nota.mediana"** son las más relevantes, lo que indica su peso significativo en los modelos predictivos. En el caso de MARS, la distribución de las importancias es más uniforme, mientras que Random Forest asigna una mayor diferenciación entre las variables, destacando

Table 4: Resultados del modelo MARS-logit

Variable	**Coeficiente**	**OR**
Constante	-1.681	0.186
Género (Mujer)	-0.424	0.655
Rama (Ingeniería y Arquitectura)	0.986	2.681
Adscrito	-0.591	0.554
$(8.1 - NotaMediana)_+$	-0.316	0.729
$(NotaMediana - 8.1)_+$	-0.331	0.718
$(0.82 - Provincia)_+$	3.262	26.121
$(Provincia - 0.82)_+$	-14.299	6.2e-07
$(nacceso - 515)_+$	0.001	1.001
$(PadreUniv - 0.28)_+$	1.691	5.420
Mujer $\cdot$ $(NotaMediana - 10.9)_+$	0.077	1.08
Mujer $\cdot$ $(10.9 - NotaMediana)_+$	-0.091	0.91
Doble $\cdot$ $(515 - nacceso)_+$	-0.002	0.998
RamaIA $\cdot$ $(NotaMediana - 8.8)_+$	-0.211	0.81
$(12 - nacceso)_+ \cdot (0.82 - Provincia)_+$	0.587	1.80
$(nacceso - 12)_+ \cdot (0.82 - Provincia)_+$	-0.008	0.992
$(0.32 - Municipio)_+ \cdot (0.82 - Provincia)_+$	-6.690	0.0012
$(Municipio - 0.32)_+ \cdot (0.82 - Provincia)_+$	33.857	5.1e14
$(515 - nacceso)_+ \cdot (Provincia - 0.72)_+$	0.020	1.02
$(515 - nacceso)_+ \cdot (0.72 - Provincia)_+$	-0.004	0.996
Earth selected 22 of 23 terms, and 9 of 22 predictors (nprune=25)		
AIC = 134400		
AUC = 0,72		
GCV=0.097; RSS=19693.02; GRSq=0.077; RSq=0.077		

también **"n.acceso"** y **"edad.media.tit"**. Variables como **"muni.local"** y **"prov.local"** tienen un impacto moderado en ambos enfoques. Random Forest ofrece un mayor contraste en las variables menos influyentes, evidenciando diferencias metodológicas entre ambos algoritmos.

5 Conclusiones

tururu

El abandono universitario tiene efectos adversos en distintos niveles. A nivel social, implica una pérdida de capital humano que puede limitar el desarrollo económico. Además, supone un desperdicio de recursos públicos invertidos en educación. A nivel individual, se traduce en menores oportunidades laborales y salariales para aquellos estudiantes que abandonan sus estudios y también incrementa para sus familias que ven un proyecto arruinado (Aina et al. 2022). Aina et al. (2022) también presenta un resumen en la Tabla 2 sobre los efectos estimados de los predictores del abandono universitario según la evidencia empírica.

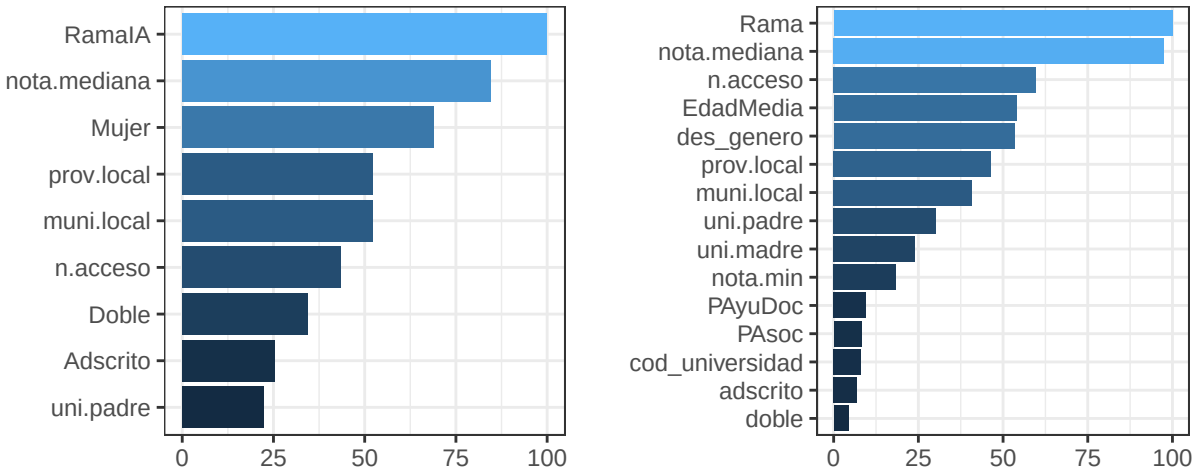


Figure 2: Importancia normalizada de las variables. A la izquierda aplicando el algoritmo MARS a la derecha con el algoritmo Random Forest. Las variables no utilizadas no aparecen en el gráfico

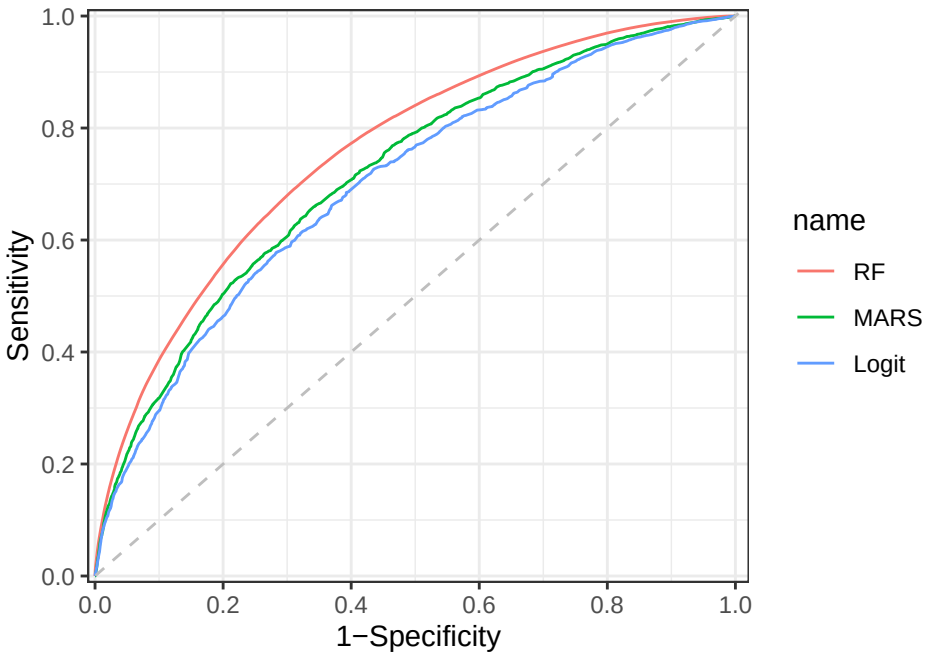


Figure 3: asdads

5.1 Propuestas para remediar el abandono

Dado el alto índice de abandono en las universidades, muchas instituciones han adoptado diversas estrategias y programas de prevención del abandono. Este tema ha sido objeto de estudio durante las últimas dos décadas, y se han propuesto numerosos planes y programas. Muchos de estos programas se desarrollan para anticiparse a las causas clave del abandono estudiantil y a la asistencia irregular. La literatura destaca la importancia de abordar estos problemas antes del inicio del curso académico o en las primeras etapas de la educación superior. Las pruebas de diagnóstico temprano han demostrado ser altamente efectivas, reduciendo significativamente las tasas de abandono estudiantil.

Además, Cholewa and Ramaswami (2015) sugiere que la personalización de la educación, junto con la orientación personalizada, puede aumentar las posibilidades de éxito académico. Por otro lado, (**?**) subraya la importancia del apoyo familiar en la toma de decisiones de los estudiantes. Una participación activa de los padres puede ser un factor determinante en la retención, mientras que su ausencia aumenta la propensión al abandono.

Existen grandes programas de prevención del abandono, pero nos centraremos en el desarrollado por la Comunidad Europea. Se ha presentado una propuesta titulada **Marco Europeo para Cursos Preparatorios de Transición Secundaria-Universidad**, que incluye diversos programas, destacando el "Lifelong Learning Programme". Este programa ofrece una base de datos que, desde 2009, ha proporcionado 118 cursos que reflejan lo que ocurre con la formación remedial en Europa.

En el caso de España, las universidades cuentan con diferentes Planes de Acción Tutorial (PAT), que orientan a los estudiantes sobre la disponibilidad y el uso de recursos para el aprendizaje, así como orientaciones académicas y profesionales, entre otras directrices (Romera, Aguayo, and Vallejo 2020). Un ejemplo destacado es el PAT de la Universidad de León, vigente desde 2002, que incluye procesos de acogida, información y orientación dirigidos a los estudiantes de nuevo ingreso, facilitando así su incorporación a una vida universitaria plena ((**?**)). Alonso García et al. (2024) demuestra que los programas de mentoría implementados en las universidades españolas son efectivos para reducir el abandono universitario y mejorar el rendimiento académico. Este enfoque no solo permite caracterizar el fenómeno, sino también ofrecer recomendaciones prácticas basadas en evidencia para reducir las tasas de abandono.

Referencias

Aina, Carmen. 2013. "Parental Background and University Dropout in Italy." *Higher Education* 65: 437–56.

Aina, Carmen, Eliana Baici, Giorgia Casalone, and Francesco Pastore. 2022. "The Determinants of University Dropout: A Review of the Socio-Economic Literature." *Socio-Economic Planning Sciences* 79: 101102.

Alban, Mayra, and David Mauricio. 2019. "Predicting University Dropout Through Data Mining: A Systematic Literature." *Indian Journal of Science and Technology* 12 (4): 1–12.

Alban, M., and D. Mauricio. 2019. "Neural Networks to Predict Dropout at the Universities." *International Journal of Machine Learning and Computing* 9 (2): 149–53.

Alonso García, M. A., A. González Ortiz de Zárate, M. D. L. Á. Gómez Flechoso, and M. Castrillón López. 2024. "Effectiveness of Peer Mentoring on University Dropout and Academic Performance."

Álvarez-Ferrándiz, D. 2021. "Análisis Del Abandono Universitario En España: Un Estudio Bibliométrico."

Behr, A., M. Giese, H. D. Teguim Kamdjou, and K. Theune. 2020. "Early Prediction of University Dropouts–a Random Forest Approach." *Jahrbücher Für Nationalökonomie Und Statistik* 240 (6): 743–89.

———. 2022. "University Dropout Problems and Solutions." *Journal of Economics* 136: 123–43. https://doi.org/10.1007/s00712-022-00814-7.

Behr, Andreas, Marco Giese, Herve D Teguim Kamdjou, and Katja Theune. 2020. "Dropping Out of University: A Literature Review." *Review of Education* 8 (2): 614–52.

Belloc, F., A. Maruotti, and L. Petrella. 2010. "University Drop-Out: An Italian Experience." *Higher Education* 60: 127–38.

Bernardo, A., A. Cervero, M. Esteban, and E. Tuero. 2021. "Student Dropout at University: A Phase-Orientated View on Quitting Studies." *European Journal of Psychology of Education* 36: 747–66. https://doi.org/10.1007/s10212-021-00557-x.

Cho, C. H., Y. W. Yu, and H. G. Kim. 2023. "A Study on Dropout Prediction for University Students Using Machine Learning." *Applied Sciences* 13 (21): 12004.

Cholewa, Blaire, and Soundaram Ramaswami. 2015. "The Effects of Counseling on the Retention and Academic Performance of Underprepared Freshmen." *Journal of College Student Retention: Research, Theory & Practice* 17 (2): 204–25.

Colás Bravo, Pilar. 2015. "El Abandono Universitario." *Revista Fuentes*, no. 16: 9–14. https://doi.org/10.12795/revistafuentes.2015.i16.

Constante-Amores, A., E. F. Martínez, E. N. Asencio, and M. Fernández-Mellizo. 2021. "Factores Asociados Al Abandono Universitario." *Educación XX1* 24 (1): 17–44.

Contini, D., F. Cugnata, and A. Scagni. 2018. "Social Selection in Higher Education. Enrolment, Dropout and Timely Degree Attainment in Italy." *Higher Education* 75 (5): 785–808.

De la Cruz-Campos, J. C., J. J. Victoria-Maldonado, J. A. Martínez-Domingo, and M. N. Campos-Soto. 2023. "Causes of Academic Dropout in Higher Education in Andalusia and Proposals for Its Prevention at University: A Systematic Review." *Frontiers in Education* 8: 1130952.

Delogu, Marco, Raffaele Lagravinese, Dimitri Paolini, and Giuliano Resce. 2024. "Predicting Dropout from Higher Education: Evidence from Italy." *Economic Modelling* 130: 106583.

Erdogan, S. Z., and M. Timor. 2005. "A Data Mining Application in a Student Database." *Journal of Aeronaut and Space Technology* 2 (2): 53–57.

Espinoza, O., L. Sandoval, L. González, K. Maldonado, Y. Larrondo, and B. Corradi. 2024. "Reasons for University Dropout in Chile: Does Student Gender Play a Role?" *Educational Review*, 1–16.

Friedman, Jerome H. 1991. "Multivariate Adaptive Regression Splines." *The Annals of Statistics* 19 (1): 1–67.

Friedman, Jerome H, and Bernard W Silverman. 1989. "Flexible Parsimonious Smoothing

and Additive Modeling." *Technometrics* 31 (1): 3–21.

Galve-González, C., A. B. Bernardo, and A. Castro-López. 2024. "Understanding the Dynamics of College Transitions Between Courses: Uncertainty Associated with the Decision to Drop Out Studies Among First and Second Year Students." *European Journal of Psychology of Education* 39: 959–78. https://doi.org/10.1007/s10212-023-00732-2.

Gilpin, Leilani H, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. "Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning." *arXiv Preprint arXiv:1806.00069*.

Goller, Daniel, Andrea Diem, and Stefan C Wolter. 2023. "Sitting Next to a Dropout: Academic Success of Students with More Educated Peers." *Economics of Education Review* 93: 102372.

Hastie, Trevor, Robert II Tibshirani, et al. 2009. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction/by Trevor Hastie, Robert Tibshirani, Jerome Frieman."

Hoang, Nhat Duc, Chun Tao Chen, and Kuo Wei Liao. 2017. "Prediction of Chloride Diffusion in Cement Mortar Using Multi-Gene Genetic Programming and Multivariate Adaptive Regression Splines." *Measurement* 112: 141–49.

Jia, Pengfei, and Tim Maloney. 2015. "Using Predictive Modelling to Identify Students at Risk of Poor University Outcomes." *Higher Education* 70: 127–49.

Kabathova, J., and M. Drlik. 2021. "Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques." *Applied Sciences* 11 (7): 3130.

Kabra, R. R., and R. S. Bichkar. 2011. "Performance Prediction of Engineering Students Using Decision Trees." *International Journal of Computer Applications* 36 (11): 8–12.

Kawada, T. 2014. "Mental Health Issues and University Student Dropouts." *Occupational Medicine* 64 (5): 392–92.

Kehm, Barbara M, Malene Rode Larsen, and Hanna Bjornoy Sommersel. 2019. "Student Dropout from Universities in Europe: A Review of Empirical Literature." *Hungarian Educational Research Journal* 9 (2): 147–64.

Kim, S., E. Choi, Y. K. Jun, and S. Lee. 2023. "Student Dropout Prediction for University with High Precision and Recall." *Applied Sciences* 13 (10): 6275.

Kotsiantis, S., C. Pierrakeas, and P. Pintelas. 2004. "Prediction of Student's Performance in Distance Learning Using Machine Learning Techniques." *Applied Artificial Intelligence* 18 (5): 411–26.

Lorenzo-Quiles, O., S. Galdón-López, and A. Lendínez-Turón. 2023. "Factors Contributing to University Dropout: A Review." *Frontiers in Education* 8: 1159864.

Madrid, Consejo Social Universidad Carlos III de. 2014. *Informe Sobre El Abandono de Los Estudios de Grado En La Universidad Carlos III de Madrid*. Consejo Social Universidad Carlos III de Madrid.

———. 2019. *Segundo Informe Sobre El Abandono de Los Estudios de Grado En La Universidad Carlos III de Madrid*. Consejo Social Universidad Carlos III de Madrid.

Mellizo-Soto, M. F. 2022. *Análisis Del Abandono de Los Estudiantes de Grado En Las Universidades Presenciales En España*. Ministerio de Universidades.

Milborrow, S. 2011. "Derived from Mda: Mars by t. Hastie and r. Tibshirani." *Earth: Multivariate Adaptive Regression Splines* 2011.

Ministerio de Ciencia, Innovación y Universidades. 2024. *El Perfil Socioeconómico Del*

Estudiantado Universitario En España. España: Ministerio de Ciencia, Innovación y Universidades.

Núñez-Naranjo, A. F. 2024. "Analysis of the Determinant Factors in University Dropout: A Case Study of Ecuador." In *Frontiers in Education*, 9:1444534. Frontiers Media SA.

Ortiz-Lozano, J M., A. Rua-Vieites, P. Bilbao-Calabuig, and M. Casadesús-Fa. 2020. "University Student Retention: Best Time and Data to Identify Undergraduate Students at Risk of Dropout." *Innovations in Education and Teaching International*.

Palacio Sprockel, L. E., J. D. Vargas Babilonia, and S. L. Monroy Toro. 2020. "Análisis Bibliométrico de Estudios Sobre Factores Socioeconómicos En Estudiantes Universitarios." *Educación y Educadores* 23 (3): 355–75.

Rodriguez-Muñiz, Luis J, Ana B Bernardo, Maria Esteban, and Irene Diaz. 2019. "Dropout and Transfer Paths: What Are the Risky Profiles When Analyzing University Persistence with Machine Learning Techniques?" *Plos One* 14 (6): e0218796.

Romera, Ana Martin, Beatriz Berrios Aguayo, and Antonio Pantoja Vallejo. 2020. "Factores y Elementos de Calidad Percibidos Por El Profesorado Participante En El Plan de Accion Tutorial de Universidades Europeas." *Educacion XX1* 23 (1): 349–71.

Segura, M., J. Mello, and A. Hernández. 2022. "Machine Learning Prediction of University Student Dropout: Does Preference Play a Key Role?" *Mathematics* 10 (18): 3359.

Seo, E. Y., J. Yang, J. E. Lee, and G. So. 2024. "Predictive Modelling of Student Dropout Risk: Practical Insights from a South Korean Distance University." *Heliyon* 10 (11).

Serrano, L. 2013. *El Abandono Educativo Temprano: Análisis Del Caso Español*. Instituto Valenciano de Investigaciones Económicas.

Sung-Hyuk, C., and C. Tappert. 2009. "Constructing Binary Decision Trees Using Genetic Algorithms." *Journal of Pattern Recognition Research* 1: 1–13.

Tinto, Vincent. 1975. "Dropout from Higher Education: A Theoretical Synthesis of Recent Research." *Review of Educational Research* 45 (1): 89–125.

Trevor Hastie, Stephen Milborrow. Derived from mda:mars by, and Rob Tibshirani. Uses Alan Miller's Fortran utilities with Thomas Lumley's leaps wrapper. 2024. *Earth: Multivariate Adaptive Regression Splines*. https://CRAN.R-project.org/package=earth.

Urbina-Nájera, A. B., J. C. Camino-Hampshire, and R. Cruz Barbosa. 2020. "University Dropout: Prevention Patterns Through the Application of Educational Data Mining." *Electronic Journal of Educational Research, Assessment and Evaluation* 26: 1–19.

Walsh, Kelsey J, and Sharon E Robinson Kurpius. 2016. "Parental, Residential, and Self-Belief Factors Influencing Academic Persistence Decisions of College Freshmen." *Journal of College Student Retention: Research, Theory & Practice* 18 (1): 49–67.

Won, Hyun-Sik, Min-Ji Kim, Dohyun Kim, Hee-Soo Kim, and Kang-Min Kim. 2023. "University Student Dropout Prediction Using Pretrained Language Models." *Applied Sciences* 13 (12): 7073.

Wright, Marvin N., and Andreas Ziegler. 2017. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." *Journal of Statistical Software* 77 (1): 1–17. https://doi.org/10.18637/jss.v077.i01.

Yu, C., S. D. Gangi, A. Jannasch-Pennell, and C. Kaprolet. 2010. "A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year." *Journal of Data Science* 8: 307–25.