MASTERPROF UMH
UNIVERSITAS Miguel Hernández

MÁSTER UNIVERSITARIO EN FORMACIÓN DEL PROFESORADO
ESO Y BACHILLERATO, FP Y ENSEÑANZAS DE IDIOMAS

TRABAJO FIN DE MÁSTER

Modelos y técnicas para predecir el abandono y el fracaso escolar

Estudiante: Rubén Marco Navarro
Especialidad: Informática
Tutor: Alejandro Rabasa Dolado
Co-tutora: Kristina Polotskaya
Curso académico: 2023-24

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. Resumen y palabras clave

Resumen:

El abandono y el fracaso escolar son problemas con un impacto muy negativo sobre la sociedad y sobre el alumno. Tener la capacidad de predecirlos con la suficiente antelación es importante para poder tomar medidas orientadas hacia su prevención, y en ese aspecto los modelos de aprendizaje automático han demostrado ser de gran utilidad. En este trabajo se revisan propuestas recientes encontradas en la literatura científica con el objetivo de arrojar luz sobre qué modelos resultan más apropiados para la predicción del fracaso y el abandono escolar y qué variables poseen un mayor poder predictivo. Los resultados sugieren que los modelos basados en redes neuronales y *random forest* obtienen la mejor precisión, aunque la variabilidad en el uso de los algoritmos y en las condiciones de entrenamiento impiden afirmar esta tesis con la rotundidad deseada. Se extrae también de la revisión que el balanceo de los datos durante el preprocesamiento incide positiva y significativamente en el rendimiento de los algoritmos y que las variables con mayor poder predictivo son aquellas relacionadas con el ámbito académico.

Palabras clave**:** abandono escolar, fracaso académico, predicción abandono, aprendizaje automático, revisión.

Abstract:

School dropout and failure are problems with a highly negative impact on society and the students. The ability to predict them is important in order to take measures aimed at their prevention, and in this regard, machine learning models have proven to be very useful. This paper reviews recent proposals found in the scientific literature with the aim of shedding light on which models are most appropriate for predicting school failure and dropout and which variables have the greatest predictive power. The results suggest that models based on neural networks and random forests achieve the best accuracy, although the variability in the use of algorithms and training conditions prevents this thesis from being stated with the desired firmness. The review also reveals that data balancing during preprocessing positively and significantly affects the performance of the algorithms and that the variables with the greatest predictive power are those related to the academic domain.

Keywords: school dropout, academic failure, dropout prediction, machine learning, review.

2. Introducción

El abandono escolar es uno de los problemas más perjudiciales y comunes en el ámbito académico puesto que, además de influir en el alumno, impacta negativamente sobre la sociedad en su conjunto, su economía y su estructura. Diversos estudios han mostrado que los alumnos que abandonan sus estudios prematuramente (antes de acabar la educación secundaria) están más frecuentemente asociados con la dependencia de asistencia pública, el desempleo a largo plazo y la exclusión social que el resto de estudiantes (De Witte et al., 2013). Podría pensarse que este problema sólo atañe a los centros de educación primaria y secundaria, pero no es así: en Estados Unidos, las universidades sólo consiguen retener entre un 45 y un 50% de su alumnado inicial y únicamente un 25% de los estudiantes que comienzan el itinerario escolar acaba completando una etapa posterior a la secundaria (Franklin y Jay, 2014). En un entorno cada vez más intelectualizado y demandante de formación, es posible imaginar las implicaciones de esta problemática tanto para el alumno como para la sociedad.

La prevención del abandono y del fracaso escolar comienza con su previsión y predicción. Si esta predicción es lo suficientemente precoz, las instituciones educativas pueden implementar estrategias eficaces para ayudar, motivar y, en definitiva, retener a los estudiantes en riesgo de abandono o fracaso. No obstante, igual que ocurre en el ámbito de la medicina y adaptando una analogía a la que recurre Maquiavelo en su famosa obra *El Príncipe* (1513), los problemas son, al principio, difíciles de diagnosticar y fáciles de resolver, mientras que con el tiempo se vuelven difíciles de resolver y fáciles de diagnosticar. La comunidad científica cuenta por fortuna con aliados poderosos en esta difícil tarea: los modelos de aprendizaje automático.

El aprendizaje automático o *machine learning* puede definirse como el conjunto de técnicas informáticas que permiten a un ordenador observar un conjunto de datos, construir un modelo basado en esos datos (encontrar relaciones entre ellos) y utilizar este modelo como herramienta predictiva o de resolución de problemas (Flach, 2012). Dentro del aprendizaje automático es posible distinguir entre dos paradigmas principales en función del etiquetado de los datos: el aprendizaje supervisado y el aprendizaje no supervisado. En el aprendizaje supervisado, los modelos se "entrenan" con datos etiquetados (esto es, con resultado conocido) para encontrar relaciones entre los datos de entrada y su etiqueta asociada. De esta forma, el modelo entrenado puede inferir correctamente las etiquetas de nuevos ejemplos no vistos durante el entrenamiento (Mehryar et al., 2019). Este tipo de modelos son muy útiles en tareas de clasificación (en las que las etiquetas son categorías) y regresión (en las que las etiquetas son valores continuos). En el caso del abandono escolar, la etiqueta de cada entrada correspondería al abandono o permanencia del alumno en el curso (es, por lo tanto, un problema de clasificación binaria).

Por el contrario, los modelos de aprendizaje no supervisado se ocupan de datos que no están etiquetados y tienen el objetivo de encontrar las formas de agrupamiento más probables o eficientes para estos datos, sin referencia a variables de salida (Mehryar et al., 2019). Los problemas abordados con estos modelos se denominan de

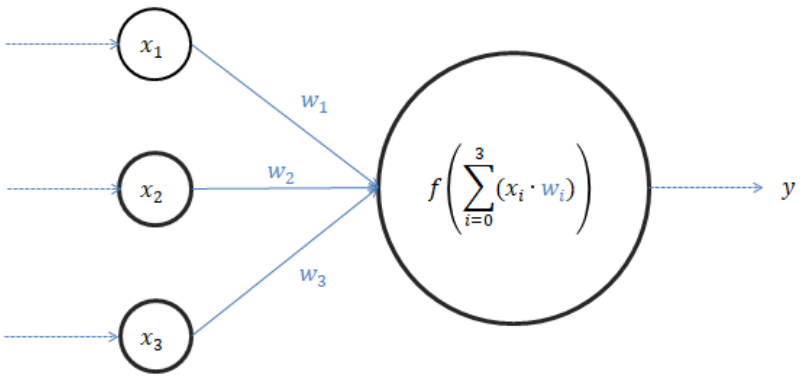
agrupamiento o *clustering* y ejemplos de ello pueden ser la creación automática de árboles filogenéticos en función del parecido entre especies o la detección del contorno de objetos en una imagen.

La naturaleza del problema sobre el que trata este trabajo hace de los modelos de aprendizaje supervisado, a priori, los más apropiados para resolverlo. A continuación se describen algunos de los modelos más utilizados en la literatura.

Redes Neuronales Artificiales: Las redes neuronales artificiales o *Artificial Neural Networks* (ANN) son algoritmos de aprendizaje automático que imitan, de manera muy simplificada, el funcionamiento del cerebro (se enmarcarían dentro de los modelos de IA bio-inspirados). Cada uno de los módulos de estos algoritmos imita el funcionamiento de una neurona, integrando la información que recibe y propagándola a los siguientes módulos en función de pesos que se van modificando durante el entrenamiento (de manera similar a como ocurre en la sinapsis biológica). El modelo más sencillo de red neuronal artificial se denomina perceptrón, diseñado por Frank Rosenblatt en 1957 (Van Der Malsburg, 1986). Este algoritmo cuenta con una serie de entradas conectadas a una sola neurona que las suma y pondera, aplica una función sobre esta combinación y ofrece el resultado como *output*. Las redes neuronales que se utilizan actualmente se basan en esta metodología pero combinan multitud de neuronas agrupadas en diferentes capas (el paradigma se suele denominar aprendizaje profundo o *deep learning* cuando el número de capas es elevado), lo cual les permite tratar con grandes volúmenes de datos y encontrar relaciones no lineales y complejas entre las variables de entrada y salida (Aggarwal, 2023).

Figura 1

Estructura de un perceptrón de tres entradas, donde f representa la función de activación y w cada peso sináptico



Nota. En la imagen, la letra *f* representa la función de activación, la letra *w* cada peso sináptico y las letras *x* e *y* las entradas y salidas del modelo, respectivamente.

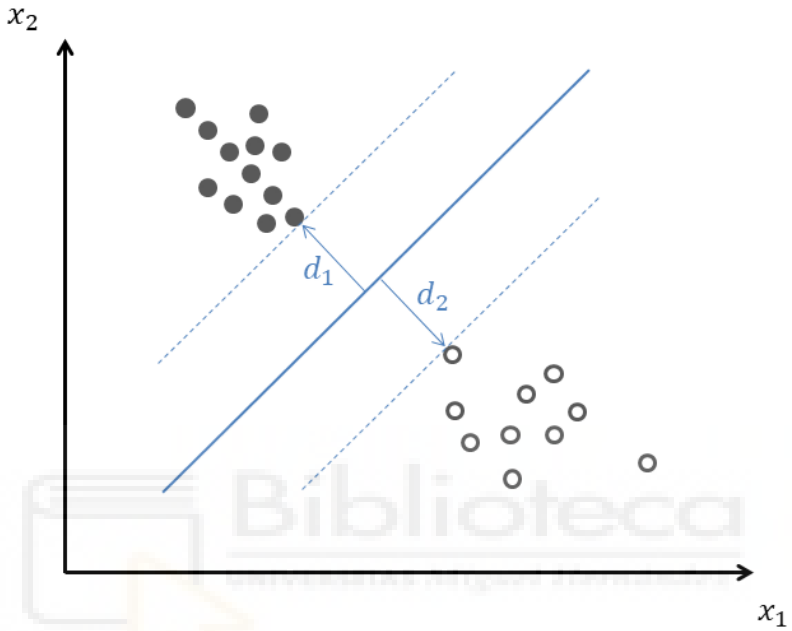
Random Forest: Los *random forest* o bosques aleatorios (aunque no se suele utilizar el término en castellano) son modelos que operan construyendo una multitud de árboles de decisión durante su entrenamiento y estableciendo posteriormente como predicción aquella obtenida en cada caso por la mayoría de los árboles generados. Fueron propuestos por L. Breiman en 2001, quien añadió este componente aleatorio a métodos similares anteriores y comprobó que esto mejoraba los resultados obtenidos (los árboles se construyen de manera diferente, dividiendo cada nodo en base a la variable más influyente de un subconjunto elegido aleatoriamente). Estos algoritmos forman parte de los denominados modelos de ensamble o *ensamble learning*, caracterizados por generar muchos clasificadores y agregar los resultados (Liaw y Wiener, 2002).

Support Vector Machines: Las *Support Vector Machines* (SVM) o máquinas de vector de soporte (aunque, de nuevo, es más común utilizar el término en inglés incluso en textos escritos en castellano) son modelos que operan mapeando los datos de entrada en un determinado espacio (generalmente de alta dimensionalidad) y encontrando el hiperplano que divide estos datos según sus categorías de la mejor manera en dicho espacio. Las SVM fueron propuestas formalmente por Corinna Cortes y Vladimir Vapnik en los años 90 (Cortes y Vapnik, 1995), aunque los conceptos fundamentales para el funcionamiento de estos modelos se fueron gestando desde los años 70 (Burges, 1998). Aunque actualmente otros algoritmos de aprendizaje automático han cobrado mayor relevancia (como las redes neuronales), el uso del algoritmos de SVM sigue siendo bastante común, especialmente en problemas de clasificación binaria (Vanneschi y Silva, 2023).

XGBoost: Este modelo optimiza los árboles de decisión recurriendo a una técnica conocida como *boosting*, que consiste en la generación progresivamente mejor de los sub-modelos (los árboles). En cada iteración, el modelo construye un árbol distinto analizando los errores o mejorías posibles del árbol generado en la iteración anterior, incrementando así el rendimiento del árbol original con el transcurso de las iteraciones. Este algoritmo fue propuesto por Tianqui Chen y Carlos Guestrin, de la Universidad de Washington, en 2016, y resulta muy útil y recurrido en la actualidad debido a su gran eficiencia y su alta escalabilidad (Chen y Guestrin, 2016).

Figura 2

Ejemplo básico de funcionamiento de un algoritmo SVM de dos entradas



Nota. Los datos se representan en un espacio que permite la división entre las clases y se encuentra un plano de dimensión *n-1* (siendo *n* el número de dimensiones de dicho espacio) que las separa de la mejor manera posible (maximizando las distancias d_1 y d_2 en este caso). Una nueva entrada se clasificará como perteneciente a una clase u otra en función de su posición respecto a este plano.

2.1. Objetivos

El principal objetivo de este trabajo es el de esclarecer qué modelos o técnicas de aprendizaje automático han demostrado ser más adecuados para la predicción del abandono y fracaso escolar a través de la revisión de estudios recientes encontrados en la literatura científica. De manera más específica, los objetivos de este trabajo pueden desglosarse en los siguientes:

- Revisar la literatura existente en busca de qué modelos de *machine learning* se utilizan más y ofrecen los mejores resultados en la predicción del abandono y el fracaso académico.
- Comparar los resultados obtenidos entre distintos modelos.
- Identificar las variables más comúnmente utilizadas para entrenar los modelos y realizar las clasificaciones, así como aquellas con un mayor poder predictivo.
- Analizar otras técnicas de relevancia (tales como las relativas al preprocesamiento de los datos) utilizadas en la predicción del abandono y fracaso académico.

Se espera con ello encontrar modelos concretos de aprendizaje automático que destaquen por su frecuencia de uso y por su rendimiento en la predicción del abandono y el fracaso académico, así como variables académicas, sociales o económicas que ejerzan una especial influencia en el resultado académico de los alumnos y, por lo tanto, convenga incluir como entrada en los clasificadores.

3. Revisión bibliográfica

3.1. Metodología de búsqueda

Para buscar los artículos revisados en este trabajo se ha recurrido, principalmente, a las plataformas *Web of Science* y *Google Scholar*. Los términos introducidos en la barra de búsqueda para encontrar estudios relacionados con el tema de esta revisión han sido "*dropout prediction*", "*school dropout*", "*academic failure prediction*", "*dropout machine learning*", "*dropout analysis*", "*school failure prediction*" y "*academic prediction model*", entre otros conceptos y combinaciones de los mencionados.

En cuanto a la criba realizada para la selección de los estudios, se han tenido en cuenta los siguientes criterios:

- El artículo es reciente, considerando como tales aquellos publicados en el año 2018 o posteriormente.
- En el artículo se aborda el problema de la predicción del abandono, fracaso o rendimiento escolar.
- En el artículo se utilizan técnicas de aprendizaje automático o *machine learning* para resolver este problema.
- El artículo no es una revisión bibliográfica, sino que en él se implementan o proponen uno o varios modelos de predicción.
- En el artículo se exponen resultados de los modelos implementados o propuestos.
- En el artículo se expone qué variables han sido utilizadas para entrenar los modelos.
- En el artículo se extraen conclusiones con respecto al rendimiento de los modelos o a la influencia de determinadas variables en la predicción.

A continuación se muestra el resultado, en un estilo narrativo, de la revisión realizada bajo estas pautas.

3.2. Revisión

En uno de los estudios más citados de los últimos años con respecto a la aplicación de técnicas de aprendizaje automático en el ámbito educativo (Yagci, 2022), el autor utiliza diferentes modelos para predecir las calificaciones finales de estudiantes universitarios de primer curso de lengua turca, comparando después los resultados. Los datos utilizados para entrenar los modelos son las calificaciones de los estudiantes de la promoción 2019-2020 en un examen parcial a mitad de curso, su facultad y su departamento (entradas) y las calificaciones en el examen final de la asignatura (salida), siendo todos ellos extraídos del *Student Information System (SIS)*, una base de datos que recopila información académica de las universidades públicas de Turquía. En el estudio se concluye que los modelos que predicen con mayor precisión los resultados de los estudiantes (y, por extensión, su fracaso o éxito en el curso) son los basados en algoritmos de *random forest,* redes neuronales y *support vector machines (SVM)*, mientras que aquellos basados en kNN obtienen los peores resultados. En cualquier caso, el autor considera que la calificación en un examen parcial, la facultad a la que pertenece el estudiante y su departamento son predictores suficientes para estimar su nota final en la asignatura, pues los modelos presentan una precisión en torno al 73%.

Cabe recalcar que en el ejemplo anterior se pretende predecir una nota y no la superación o no de la asignatura, lo cual supone, a priori, una tarea más sencilla por tener sólo dos estados de salida posibles (aunque el autor no proporciona resultados al respecto). Ahmad y Shahzadi (2018) llevan a cabo esta tarea utilizando una red neuronal multicapa y métricas relativas a los hábitos de estudio, calificaciones anteriores, ambiente doméstico, interacción durante el curso y capacidad de los estudiantes como entrada, obteniendo una precisión del 85% en la clasificación de lo que ellos denominan estudiantes en riesgo de fracaso (AR, *At Risk)* y no en riesgo de fracaso académico (NAR, *Not At Risk*). Cruz-Jesús et al. (2020) realizan un estudio con el mismo fin de categorización binaria (que ellos concretan en la determinación de si un estudiante promociona o no al siguiente curso) recurriendo a una base de datos con información (anónima) sobre 110 627 alumnos de enseñanza secundaria portugueses facilitada por la Dirección General de Estadísticas de Educación y Ciencia del Ministerio de Educación de Portugal. En este caso, los autores utilizan 17 variables de entrada entre las que se encuentran el curso del alumno, su género, su tasa de asistencia, si el alumno es o no de origen portugués, la edad del alumno, si tiene o no acceso a un ordenador y a internet en casa, el número de alumnos presentes en su clase y la presencia o no de fracasos previos en su historial académico. Con estos datos, en el estudio se entrenan diferentes algoritmos de aprendizaje automático y comparan los resultados obtenidos utilizando el área bajo la curva ROC como métrica de éxito, obteniendo los modelos basados en redes neuronales y *random forest* los mayores índices de acierto (75 y 76%, respectivamente) y los modelos basados en kNN y regresión logística, los menores (55%). Cabe destacar que, al contrario que en los anteriores ejemplos, en este estudio no se utilizan como entrada las calificaciones obtenidas por el alumno previamente en el mismo curso y, por otra parte, que la base de datos utilizada es sensiblemente más grande que en la mayoría de trabajos (según el texto, se trata de uno de los primeros trabajos en usar inteligencia artificial para

predecir el rendimiento académico a escala nacional). En base a los resultados, los autores concluyen que los modelos basados en aprendizaje automático ofrecen mejores resultados que los modelos tradicionales y que los indicadores más relevantes en la predicción del fracaso son el número de fracasos previos en el historial académico del alumno, el número de clases a las que ha asistido el alumno durante el año y su género (observan que el fracaso es menos probable en mujeres).

Directamente con el objetivo de encontrar los indicadores más relevantes para predecir el rendimiento escolar, Rebai et al. (2020) desarrollan un modelo de aprendizaje automático basado en árboles de regresión y *random forest* y lo utilizan para analizar datos de escuelas de secundaria de Túnez. En base a los resultados, los autores concluyen que los indicadores con mayor influencia en la determinación del éxito académico de los estudiantes son el tamaño de la escuela y la proporción de chicas (teniendo ambos un impacto positivo), aunque conviene matizar que en este estudio se trabaja con datos a nivel de centro y no sobre alumnos individuales. Gue et al. (2021) comparten una meta similar al analizar los indicadores más relevantes en el fracaso académico de estudiantes de ingeniería mecánica, tratando en este caso no de encontrar predictores aislados sino un conjunto de reglas a partir de estos. Los autores implementan para ello un modelo basado en redes neuronales y lógica difusa que toma como entrada tres variables: la carga de trabajo del alumno, el número de veces que no ha asistido a clase y el número de veces que ha repetido la asignatura evaluada. El modelo genera un conjunto de reglas difusas del cual los autores extraen que, contraintuitivamente, la baja carga de trabajo es el indicador individual predominante en la predicción del fracaso de un alumno y el escenario de baja asistencia a clase en conjunción con una baja carga de trabajo y cursar la asignatura por primera vez (no haber repetido antes) conduce al fracaso en todos los casos.

Wan et al. (2019) aplican también técnicas de aprendizaje automático para predecir el rendimiento de los estudiantes en un curso online, concretamente un modelo basado en TrAdaBoost (*Transfer Adaptive Boosting,* una extensión del algoritmo AdaBoost diseñada para escenarios de aprendizaje por transferencia, es decir, aquellos en los que se aprovecha la información de un dominio con datos etiquetados para mejorar el rendimiento del modelo en un dominio de datos no etiquetados). Mediante este algoritmo, los autores utilizan las interacciones de los alumnos con los contenidos del curso (acciones como abrir o cerrar documentos o reproducir videos), la resolución de los problemas propuestos, su participación en los foros y las respuestas a encuestas (entre otras variables) para extraer un total de 26 características relativas a los hábitos de los estudiantes. Basándose en estas características, en el estudio se compara el rendimiento de distintos algoritmos de clasificación y se concluye que el que mejores resultados ofrece es el clasificador GBDT (*Gradient Boosting Decision Tree*), por lo cual se utiliza éste para predecir las calificaciones de los alumnos durante el otoño de 2017. El año siguiente, tras cambiar la estructura del curso, los autores recurrieron al aprendizaje por transferencia para, utilizando el dataset de 2017 como entrenamiento y añadiendo los datos de 2018 (demasiado escasos como para entrenar un modelo por sí solos), entrenar un modelo basado en el algoritmo TrAdaBoost con el fin de predecir el rendimiento de los estudiantes en el nuevo curso. Los autores concluyen

que este modelo es capaz de alcanzar una alta precisión en sus predicciones aunque los nuevos datos sean reducidos, y muestran que la aplicación de una serie de intervenciones pedagógicas (como recordatorios basados en el comportamiento y recomendaciones) junto con estas predicciones tempranas repercute positiva y significativamente en el rendimiento de los estudiantes.

Los estudios comentados hasta el momento en este trabajo se centran en predecir o bien el rendimiento de los estudiantes (su calificación final en una asignatura o su grado de implicación) o bien si la superan o no (un problema de clasificación que sólo admite dos estados). Más parecido al segundo es el problema de la predicción del abandono escolar (*dropout*) que abordan otros investigadores. Baranyi et al. (2021) utilizan una serie datos de una muestra de 8319 alumnos de carreras universitarias del ámbito STEM (*Science, Technology, Engineering and Mathematics*) de la Universidad de Tecnología y Económicas de Budapest para no sólo predecir el abandono académico sino también proporcionar justificaciones que puedan ser útiles para realizar futuros diagnósticos y elaborar acciones preventivas (paradigma conocido como *Explainable* o *Interpretable AI*, tal y como se resalta en el título del artículo). Los investigadores recopilaron datos de los alumnos disponibles en el momento de su matriculación desde el año 2013 hasta el 2019 tales como su nota de entrada a la universidad (el equivalente a la nota de selectividad en España), su sexo, si es la primera vez que se matriculan en la universidad o no y si reciben ayudas del estado para pagar las tasas universitarias, entre otras variables. Sobre este conjunto de datos, en el estudio se aplican varios modelos de clasificación: XGBoost, *random forest* y redes neuronales. Tras analizar los resultados ofrecidos por los distintos modelos, los autores concluyen que las redes neuronales obtienen los mejores valores de precisión, exactitud *(accuracy*), *recall* y AUC-ROC (0.74, 0.72, 0.67 y 0.77 respectivamente). Además, gracias a las herramientas de explicabilidad incorporadas en su modelo, encuentran que los factores con mayor poder predictivo sobre la graduación de un alumno son los años transcurridos desde la finalización del instituto hasta su ingreso en la universidad, su nota de acceso a la universidad y su nota concreta en el examen de matemáticas de acceso a la universidad. Observan además que, a pesar de tratarse de carreras tecnológicas, la nota en asignaturas de humanidades durante el instituto tiene un impacto positivo sobre la probabilidad de no abandono, lo que sugiere que los estudiantes más generalistas son menos propensos a abandonar sus estudios.

A similares conclusiones llegan posteriormente dos de los autores del anterior estudio (Nagy y Molontay, 2023), quienes, gracias a su modelo interpretable basado en CatBoost, encuentran que las variables con mayor poder predictivo sobre el abandono universitario (realizando la predicción al comienzo del primer curso) son la nota media del instituto (que se puede considerar como una métrica de conocimiento general puesto que tiene en cuenta la nota de asignaturas de distinta índole) y el tiempo que transcurre entre la finalización del instituto y el comienzo de los estudios universitarios (los estudiantes son más propensos a abandonar cuanto mayor es este tiempo).

También bajo la filosofía de la "IA explicable", Krüger et al. (2023) utilizan los datos recogidos de 19 escuelas en Brasil para, además de predecir el abandono de los alumnos tempranamente, encontrar las razones o factores por los cuales un alumno es propenso a abandonar el curso. En este estudio se aplican diferentes algoritmos de clasificación tales como árboles de decisión, regresión logística, *random forest*, AdaBoost y XGBoost para determinar si un alumno dejará la escuela o finalizará el curso escolar. Utilizando las métricas "precisión" (número de verdaderos positivos entre número total de positivos) y "*recall*" (número de verdaderos positivos entre número real de positivos), los autores concluyen que el algoritmo clasificador que mejores resultados ofrece es XGBoost, alcanzando puntuaciones de hasta 0.95 y 0.93 en precisión y *recall,* respectivamente, y de 0.89 en el área bajo la curva precisión-*recall* (AUC-PR). Los autores concluyen también que es más fácil predecir si un estudiante abandonará o no el curso cuanto más tarde (más avanzado el curso) se realice la clasificación. Aunque esto pueda resultar evidente, el estudio incide en que esta mejoría en la precisión de las predicciones es especialmente notoria entre el primer y el segundo trimestre de curso, razón por la cual se argumenta que el segundo trimestre podría ser el momento más apropiado para evaluar el riesgo de abandono de los alumnos y tomar medidas al respecto. Otra de las conclusiones del estudio es que la etapa académica en la que es más complicado predecir el abandono es la de preescolar, circunstancia que los autores achacan a que esta no es una etapa obligatoria en Brasil y a que los alumnos no son evaluados con notas, por lo que es más difícil distinguir entre alumnos comprometidos y no comprometidos (o, probablemente, padres en este caso). Con respecto a la explicabilidad del clasificador, los autores extraen una serie de características especialmente influyentes en el abandono y concluyen que los indicadores más relevantes para predecirlo varían en función de la etapa académica: durante preescolar, el indicador más relevante es la tasa de matriculación; durante primaria, la suma acumulada hasta el momento de las calificaciones del alumno en las asignaturas de educación física y arte; y durante secundaria, la suma acumulada de las calificaciones del alumno en las asignaturas de portugués y geografía.

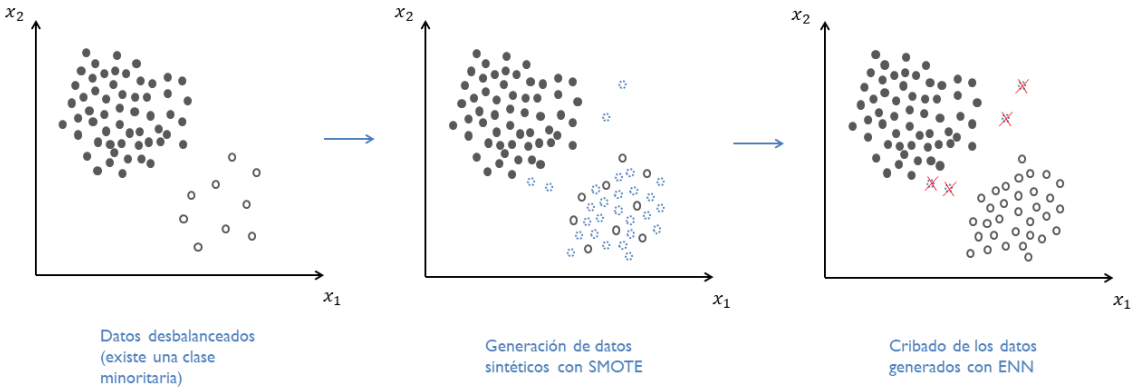
De manera similar pero en el ámbito universitario de Perú, Jiménez et al. (2023) proponen en su estudio seguir la metodología CRISP - DM (*Cross Industry Standard Process for Data Mining*) para desarrollar un modelo predictivo del abandono escolar y encontrar las variables más relevantes en esta predicción. La metodología a la que recurren los autores consta de cinco fases: entendimiento del problema, entendimiento de los datos, preparación o limpieza de los datos, modelado y evaluación. Los datos utilizados en el estudio provienen de una encuesta realizada a 385 estudiantes de universidades públicas y privadas peruanas en la que se consideran no sólo variables académicas sino también afectivas, cognitivas y familiares, entre otras (esta encuesta puede verse íntegramente en el artículo). Tras implementar y analizar el rendimiento de cuatro tipos de algoritmos de clasificación binaria (uno basado en árboles de decisión, otro en *random forests*, otro en *support vector machines* y otro en redes neuronales), los autores concluyen que el modelo que mejores resultados ofrece es el basado en el algoritmo *random forest*, obteniendo una puntuación de 0.96 en la métrica AUC-ROC al predecir el abandono universitario, y que las variables más

influyentes son la edad de los estudiantes, el método financiero utilizado por estos para pagar las tasas universitarias, el método por el cual el alumno accede a la universidad y el momento del curso en el que se recaban los datos.

Una complicación adicional que presenta el problema de la predicción del abandono escolar frente a otros problemas de clasificación binaria es la desproporción que existe entre sus categorías: puesto que, en circunstancias normales, la gran mayoría de los alumnos matriculados en un curso lo terminan, si se pretende entrenar un algoritmo de clasificación con datos etiquetados de otros años se deberá afrontar el problema de tener una cantidad de entradas mucho menor con la etiqueta "abandona" que de entradas con la etiqueta "no abandona". Para resolver este problema, Mduma (2023) evalúa la aplicación de una serie de técnicas de balanceo de datos para mejorar la precisión en la predicción de la clase minoritaria sin sacrificar la precisión general del modelo. El autor utiliza dos conjuntos de datos para realizar estas comprobaciones: el dataset "Uwezo", relativo a datos académicos de alumnos en Tanzania y recopilado en 2015, y un dataset recopilado en 2016 en la India con el objetivo de analizar el abandono escolar. Estos conjuntos de datos ejemplifican el problema descrito: en el primero, un 98.4% de las muestras corresponden a casos de no abandono y sólo el restante 1.6% a casos de abandono, mientras que en el segundo estos valores son del 95.1% y el 4.9%, respectivamente. En el estudio se utilizan diferentes métodos de preprocesamiento de datos para paliar este desbalance tales como *random oversampling, random undersampling* y SMOTE *(Synthetic Minority Over Sampling Technique)*. Comparando los resultados obtenidos al aplicar diferentes algoritmos de clasificación (regresión logística, *random forest* y perceptrón multicapa) antes y después de realizar el balanceo de los datos con estas técnicas, el autor concluye que la técnica SMOTE en combinación con *Edited Nearest Neighbor* fue la más efectiva, mejorando significativamente la precisión de las predicciones. Esta técnica consiste en la generación artificial de datos no reales de la clase minoritaria por similitud con los reales y un cierto grado de aleatoriedad (SMOTE) y un posterior cribado de los datos generados para eliminar aquellos no coherentes con la mayoría de sus vecinos (*Edited Nearest Neighbor*), como se ejemplifica en la Figura 3. Aunque no es el objetivo principal del estudio, se concluye también que el modelo de regresión logística es aquel que mejores resultados ofrece.

Figura 3

Ejemplo de funcionamiento del método SMOTE en combinación con ENN



Una técnica muy parecida para balancear los datos fue utilizada por Fernández-García et al. (2021), quienes recogieron datos de 1418 estudiantes universitarios y les aplicaron un preprocesamiento combinando el método SMOTE con un posterior cribado basado en enlaces Tomek, que funciona de manera similar al de *Edited Nearest Neighbor* mencionado anteriormente. El set de características utilizado por los autores contiene algunas relacionadas con el ingreso a la universidad de los alumnos, otras relacionadas con sus calificaciones y otras relacionadas con la financiación de sus estudios (estando así todas relacionadas con el ámbito académico y ninguna con aspectos sociodemográficos, como remarcan los investigadores). En el estudio se aplican distintos algoritmos de clasificación tales como *support vector machines, gradient boosting* y *random forest* para predecir el abandono académico en cada trimestre y esclarecer qué algoritmo ofrece los mejores resultados. Los autores concluyen que es posible predecir el abandono universitario desde el mismo principio del curso utilizando esta metodología, en su caso con un 72% de acierto utilizando *gradient boosting*. Este porcentaje de éxito mejora conforme avanza el curso, obteniendo el modelo de *support vector machine* los mejores resultados con un 91.5% en el último trimestre. También recurren al método SMOTE Barros et al. (2019) al procesar los datos de alumnos universitarios brasileños (algunos de ellos, esta vez sí, demográficos, sociales y económicos) para predecir el abandono escolar y testear el rendimiento de distintos algoritmos de aprendizaje automático al respecto. De entre los distintos modelos probados, los autores encuentran que las redes neuronales con el conjunto de datos preprocesado obtienen el mayor valor de precisión (0.991) mientras que los árboles de decisión obtienen el mayor valor de *recall* (0.977).

Vaarma y Li (2024) utilizan varios modelos de aprendizaje automático con el objetivo de predecir el abandono académico de los alumnos de una universidad finlandesa y detectar los factores que más influyen en el resultado. Concretamente, en el estudio se pretende esclarecer cómo de influyentes son las variables relativas a la utilización de

plataformas online por parte de los alumnos (Moodle) en la predicción del abandono en comparación con variables demográficas o de otra índole. Los autores implementan diez clasificadores distintos entre los que se encuentran las redes de neuronas, kNN (k *Nearest Neighbors*), XGBoost y SVM y finalmente concluyen que aquellos que mejores resultados ofrecen son las redes de neuronas, la regresión logística y el CatBoost. En estos tres modelos de clasificadores que mejores resultados ofrecen, los autores encuentran que las variables más relevantes en la predicción del abandono universitario son los créditos acumulados del alumno, el número acumulado de asignaturas suspendidas por el alumno y, como se hipotetizaba, la actividad del alumno en su cuenta de Moodle. De estos hallazgos, en el estudio se infiere que la actividad de los alumnos en plataformas online es un potente predictor del abandono escolar y que deberían utilizarse estos datos como variables de entrada en los clasificadores. Se concluye también que, utilizando los datos de plataformas online en conjunción con variables demográficas y curriculares, es posible predecir la mayoría de los abandonos universitarios mediantes algoritmos de aprendizaje automático y actuar en consecuencia.

En el ámbito de la educación exclusivamente online, Fu et al. (2021) proponen un modelo al que bautizan como CSLA para predecir el abandono de los alumnos que cursan un MOOC (*Masive Open Online Course*). En el estudio se registraron los datos de actividad de los alumnos del curso (relativos a la navegación por el contenido del curso, resoluciones de problemas, reproducciones de videos y comentarios en los foros, entre otros) durante cinco semanas consecutivas, y estos datos fueron utilizados como entradas etiquetadas para el entrenamiento del modelo, considerando que un alumno ha abandonado el curso si no se registra ninguna actividad por su parte durante diez días consecutivos y que no lo ha abandonado en caso contrario. Los autores aplican una red neuronal convolucional sobre estos datos en primera instancia para después aplicar una red neuronal de tipo LSTM (*Long Short-Term Memory*) sobre la matriz de características extraídas por la red convolucional. Tras comparar los resultados de su modelo con los obtenidos al implementar otros algoritmos de clasificación (como SVM y una red neuronal de tipo LSTM sin una etapa anterior convolucional), en el estudio se concluye que el poder predictivo y la eficiencia del primero son mayores, alcanzando valores de precisión del 87.4% frente al 68.3% que obtienen con SVM y el 78.7% con la red LSTM, por citar algunos.

Los investigadores anteriores reconocen, no obstante, que su modelo ofrece poca interpretabilidad en sus predicciones, algo que sí que abordan Benoit et al. (2024), también en el ámbito de los MOOCs, en su reciente trabajo. Aplicando una versión interpretable de modelo oculto de Márkov (HMM, *Hidden Markov Model*), los autores encuentran que la motivación de los alumnos tiende a decaer con el paso del tiempo y que esta baja motivación conduce en muchos casos al abandono, por lo que recomiendan tener en cuenta variables relacionadas con la motivación de los alumnos en los modelos predictivos y tomar medidas preventivas para evitar este decaimiento. No obstante, en el estudio se matiza que alumnos con alta motivación pueden también abandonar el curso online una vez satisfechos sus objetivos de aprendizaje, pero no consideran estos abandonos un fracaso y recomiendan diferenciarlos de los *drop-outs*

que sí lo son. Previamente, Mubarak et al. (2020) ya aplicaron modelos ocultos de Màrkov para la predicción del abandono en cursos online, en su caso en combinación con distintos modelos de clasificadores binarios (regresión logística, árboles de decisión, SVM y *random forest*). En el estudio se concluye que el modelo de regresión logística en combinación con HMM es el que mejores resultados ofrece; no obstante, en este trabajo no se explora la interpretabilidad del modelo. En otro trabajo reciente sobre la predicción del abandono en los MOOCs (Chi et al., 2023) se encuentra que los modelos basados en *random forest* son los que mejores resultados predictivos ofrecen basándose en métricas de exactitud, precisión y AUC-ROC (91.7%, 93.1% y 0.925 respectivamente) en comparación con modelos basados en regresión logística y kNN.

Kim et al. (2023) proponen un método híbrido para mejorar la eficiencia en los problemas de predicción de abandono escolar: en primer lugar, comprimen el conjunto de datos de entrada mediante la técnica PCA (*Principal Component Analysis*), que extrae de los datos "crudos" un conjunto de características y nuevas variables combinadas para mejorar la eficacia de las predicciones; en segundo lugar, aplican el método SMOTE para balancear los datos y un modelo basado en XGBoost para realizar la predicción sobre si un alumno abandonará o no el curso; y en tercer lugar aplican un algoritmo de *clustering* basado en el modelo *K-means*. El sistema propuesto por los autores no sólo es capaz de predecir con una alta precisión si un alumno abandonará el curso o no, sino que también es capaz de, gracias al algoritmo de *clustering*, predecir las razones por las cuales se realizará este abandono, que los autores clasifican en cuatro grupos: por encontrar empleo, por no matricularse correctamente, por razones personales o por ser admitido en otra universidad. Los investigadores defienden que al predecir las razones por las cuales un alumno es propenso a abandonar el curso es posible organizar una prevención personalizada, lo cual resulta más efectivo que una intervención genérica.

También desde Corea del Sur, Song et al. (2023) aplican seis algoritmos distintos de clasificación a cuatro conjuntos de datos (todos ellos con las mismas características pero recogidos en distintos momentos del curso) pertenecientes a más de 60 000 alumnos universitarios del país, conteniendo variables relativas al grado de asistencia a clases, notas, edad, ingresos y sexo, entre otras. Este trabajo y el anterior no sólo comparten país de origen y año de publicación, puesto que en este caso también se recurre al método SMOTE para balancear el conjunto de datos (ya que del total de entradas tan solo unas 7000 se corresponden con *dropouts*). Los clasificadores evaluados en el estudio son un árbol de decisión, un *random forest,* XGBoost, LightGBM, regresión logística y SVM. Los autores encuentran que no hay un modelo que domine sobre los demás en términos generales, sino que según el conjunto de datos sobre el que se aplique es un algoritmo u otro el que ofrece los mejores resultados. Se concluye también que la nota media del alumno hasta el momento (GPA), el número de faltas a clase y la diferencia entre los créditos de los que el alumno se ha matriculado y los créditos que el alumno realmente ha cursado tienen una alta correlación con el abandono académico.

4. Resultados y propuestas

El primer resultado que se extrae de la síntesis anterior es que, efectivamente, **las técnicas y modelos de aprendizaje automático son de gran utilidad en la predicción del abandono y el fracaso académico**. Los estudios analizados muestran que, implementando modelos de esta índole, es posible realizar las predicciones con niveles de precisión que varían entre el 73% y el 99% en función del modelo, las variables estudiadas, el momento en el que se realiza dicha predicción y otras condiciones relacionadas con el problema. Esto supone un éxito notable y reafirma al *machine learning* como una de las herramientas más poderosas para abordar este problema por su capacidad para tratar con grandes volúmenes de datos y realizar inferencias y generalizaciones esclarecedoras.

En cuanto a qué modelo concreto o qué tipos de técnicas de aprendizaje automático han demostrado ser más eficaces en la predicción del abandono y el fracaso escolar, se induce de la revisión realizada que **no existe un modelo que domine claramente sobre el resto** en esta tarea, puesto que la variabilidad de algoritmos utilizados en los estudios e, incluso, de ganadores resultantes en comparativas bajo condiciones similares es muy elevada. No obstante, **los modelos basados en redes neuronales y *random forest* parecen ser los más populares en este ámbito y aquellos que ofrecen los mejores resultados**, aunque las generalmente discretas diferencias con respecto a otros modelos no permiten considerarlos como las mejores opciones en términos absolutos.

Más importante que el modelo concreto a utilizar parece ser el preprocesamiento de los datos. La naturaleza del problema del abandono escolar conlleva una desproporción entre las categorías de los datos de entrada favorable al "no-abandono", puesto que, normalmente, es una minoría de alumnos la que abandona un curso académico específico. La problemática de esta circunstancia reside en la posible aparición de sesgos en el algoritmo como consecuencia de un entrenamiento no equitativo. A este respecto, se extrae de la literatura que **un preprocesamiento de los datos orientado hacia el balance entre categorías mejora notablemente los resultados**. El método SMOTE, que opera generando artificialmente nuevos ejemplos de la categoría minoritaria, en combinación con algún algoritmo de cribado supone la técnica más popular para este propósito.

Por último, existe también una gran variabilidad entre las bases de datos utilizadas en los estudios y las variables de entrada a los modelos. Esto complica la comparativa de rendimiento entre algoritmos y, personalmente, considero que reduce el alcance de dicha comparativa a los modelos implementados en un mismo estudio y bajo las mismas condiciones. A pesar de ello, **las variables relacionadas con el número de asignaturas suspensas por un alumno, sus notas anteriores, su inactividad o faltas de asistencia y el momento de la predicción aparecen recurrentemente como indicadores de gran influencia** en la predicción del abandono y el fracaso académico. Es posible inducir de ello que, como cabría esperar, recurrir a variables de índole académica para entrenar los modelos y detectar tempranamente el abandono estudiantil resulta apropiado y aconsejable. La mayoría de estudios utiliza, sin embargo, una combinación de características académicas, sociales y económicas de los alumnos, lo cual permite un análisis más completo (encontrándose en algunos casos muy buenos predictores fuera del ámbito académico) sin conllevar un decremento en la eficiencia de los modelos gracias a algoritmos de reducción de dimensionalidad como PCA.

Tabla 1

Resumen de los trabajos revisados (parte 1)

Autores	Año	Temática	Modelo implementado / Mejor modelo	Variables utilizadas / Variables más influyentes
Ahmad y Shahzadi	2018	Predicción de fracaso académico, curso presencial	NN	Académicas y domésticas
Wan et al.	2019	Predicción de fracaso académico, curso online	TrAdaBoost	Interacción con el material, participación
Barros et al.	2019	Predicción de abandono académico, curso presencial	SMOTE + NN y SMOTE + RF	Académicas, sociales y económicas
Cruz-Jesús et al.	2020	Predicción de fracaso académico, curso presencial	NN y RF	Historial académico, asistencia a clase y género
Rebai et al.	2020	Predicción de fracaso académico, curso presencial	RF	Tamaño del centro y proporción de chicas
Mubarak et al.	2020	Predicción de abandono académico, curso online	HMM + Regresión logística	Actividad, participación
Gue et al.	2021	Predicción de fracaso académico, curso presencial	NN y Lógica difusa (sistema generador de reglas)	Carga de trabajo del alumno y asistencia a clase
Baranyi et al.	2021	Predicción de abandono académico, curso presencial	NN, IA interpretable	Tiempo entre graduación de instituto y entrada a la universidad, nota de acceso a la universidad
Fu et al.	2021	Predicción de abandono académico, curso online	NN	Actividad en la web, participación en resolución de problemas y foros
Nagy y Molontay	2023	Predicción de abandono académico, curso presencial	CatBoost	Tiempo entre graduación de instituto y entrada a la universidad, nota media del instituto
Krüger et al.	2023	Predicción de abandono académico, curso presencial	XGBoost, IA interpretable	Calificaciones en educación física y arte (primaria), calificacion en lengua y geografía (secundaria)
Jiménez et al.	2023	Predicción de abandono académico, curso presencial	RF	Edad, medio de pago de las tasas, medio de acceso a la universidad

Nota. Las siglas NN, RF y HMM corresponden, respectivamente, a *Neural Networks, Random Forest* y *Hidden Markov Model.*

Tabla 2

Resumen de los trabajos revisados (parte 2)

Autores	Año	Temática	Modelo implementado / Mejor modelo	Variables utilizadas / Variables más influyentes
Mduma	2023	Predicción de abandono académico, curso presencial	SMOTE + Regresión Logística	Académicas, familiares y económicas
Fernández-García et al.	2023	Predicción de abandono académico, curso presencial	SMOTE + SVM	Académicas
Chi et al.	2023	Predicción de abandono académico, curso online	RF	Tiempo de actividad, país de origen, reproducción de videos
Kim et al.	2023	Predicción de abandono académico, curso presencial	SMOTE + XGBoost	Académicas, personales y económicas
Song et al.	2023	Predicción de abandono académico, curso presencial	SMOTE + (RF, XGBoost, SVM, Regresión logística)	Nota media, asistencia a clase, diferencia entre créditos pagados y créditos cursados
Vaarma y Li	2024	Predicción de abandono académico, curso híbrido	Regresión Logística, CatBoost y NN	Créditos acumulados, asignaturas suspensas y actividad en Moodle
Benoit et al.	2024	Predicción de abandono académico, curso online	HMM, IA interpretable	Motivación

Nota. Las siglas NN, RF y HMM corresponden, respectivamente, a *Neural Networks, Random Forest* y *Hidden Markov Model.*

Hasta donde se ha revisado, en la literatura no se han encontrado ejemplos en los que se utilicen modelos de aprendizaje automático no sólo para la predicción del abandono y el fracaso escolar, sino también para analizar la influencia de las intervenciones motivacionales o de refuerzo sobre la probabilidad del abandono y del fracaso. Esto permitiría tener información temprana sobre qué medidas serán de ayuda en cada caso concreto antes de implementarlas y, quizás, también sobre cómo y cuándo hacerlo para maximizar su eficacia. Realizar un estudio de esta índole requeriría de la confección de una base de datos que incluyese variables relativas a medidas de intervención y del análisis de las relaciones entre la implementación o ausencia de estas medidas y su tipo, el resto de características de cada alumno y su etiqueta de salida a través del modelo. Todo ello sería laborioso pero, indudablemente, de interés y utilidad para la comunidad académica.

Por otra parte, aunque se han encontrado estudios que dan importancia a la explicabilidad de los modelos, estos son la minoría. Hacer de la transparencia o interpretabilidad de los algoritmos uno de los objetivos de investigación principales en este ámbito podría dar lugar a conjuntos de inferencias (mediante sistemas inductores de reglas, por ejemplo) que mejoren nuestra comprensión sobre qué circunstancias determinan que un alumno fracase o abandone un curso, cuándo se empiezan a gestar esas circunstancias y en qué medida es posible paliar sus efectos.

5. Conclusión

En este trabajo se ha realizado una revisión de estudios recientes encontrados en la literatura científica que abordan el problema de la predicción del abandono escolar mediante algoritmos de aprendizaje automático. Pese a no encontrar una clara mejor opción entre los modelos comúnmente utilizados al respecto, se extraen de dicha revisión algunas conclusiones interesantes tales como la importancia del balanceo previo de los datos, la especial influencia de variables académicas en la predicción y la predilección hacia el uso de modelos basados en redes neuronales y *random forest*.

Se extrae también de este estudio que mayores esfuerzos en el diseño de modelos explicativos y la inclusión de variables relativas a la aplicación de medidas destinadas a evitar el abandono y el fracaso escolar podrían conllevar un mejor entendimiento de las raíces de este problema y sobre cómo frenarlo a tiempo. Aún con estas posibles líneas de investigación por explotar, los generalmente altos índices de precisión que ofrecen los modelos revisados demuestran que las técnicas de aprendizaje automático resultan tan útiles en el ámbito académico, concretamente en la cuestión que atañe a este trabajo, como en tantos otros contextos. El impacto potencial de la aplicación de estos modelos en escuelas, institutos y universidades es enorme e invita a soñar con un futuro en el que el abandono y el fracaso académico sean anecdóticos. Fortuna mediante y gracias a la investigación en este campo, puede que algún día presenciemos un escenario en el que cada estudiante tenga la oportunidad de alcanzar su pleno desarrollo y las instituciones educativas puedan intervenir a tiempo para garantizar el éxito de sus alumnos.

6. Referencias

Aggarwal, C. C. (2023). Neural Networks and Deep Learning: A Textbook. In *Neural Networks and Deep Learning: A Textbook*. https://doi.org/10.1007/978-3-031-29642-0

Ahmad, Z., & Shahzadi, E. (2018). Prediction of Students' Academic Performance using Artificial Neural Network. *Bulletin of Education and Research*, *40*(3).

Baranyi, M., Nagy, M., & Molontay, R. (2020). Interpretable Deep Learning for University Dropout Prediction. *SIGITE 2020 - Proceedings of the 21st Annual Conference on Information Technology Education*. https://doi.org/10.1145/3368308.3415382

Barros, T. M., Neto, P. A. S., Silva, I., & Guedes, L. A. (2019). Predictive models for imbalanced data: A school dropout perspective. *Education Sciences*, *9*(4). https://doi.org/10.3390/educsci9040275

Benoit, D. F., Tsang, W. K., Coussement, K., & Raes, A. (2024). High-stake student drop-out prediction using hidden Markov models in fully asynchronous subscription-based MOOCs. *Technological Forecasting and Social Change*, *198*. https://doi.org/10.1016/j.techfore.2023.123009

Breiman, L. (2001). Random forests. Machine Learning. *Kluwer Academic Publishers. Manufactured in The Netherlands.*, *45(1)*.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2). https://doi.org/10.1023/A:1009715923555

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*. https://doi.org/10.1145/2939672.2939785

Chi, Z., Zhang, S., & Shi, L. (2023). Analysis and Prediction of MOOC Learners' Dropout Behavior. *Applied Sciences (Switzerland)*, *13*(2). https://doi.org/10.3390/app13021068

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3). https://doi.org/10.1023/A:1022627411411

Cruz-Jesus, F., Castelli, M., Oliveira, T., Mendes, R., Nunes, C., Sa-Velho, M., & Rosa-Louro, A. (2020). Using artificial intelligence methods to assess academic achievement in public high schools of a European Union country. *Heliyon*, *6*(6). https://doi.org/10.1016/j.heliyon.2020.e04081

De Witte, K., Cabus, S., Thyssen, G., Groot, W., & van den Brink, H. M. (2013). A critical review of the literature on school dropout. In *Educational Research Review* (Vol. 10). https://doi.org/10.1016/j.edurev.2013.05.002

Fernandez-Garcia, A. J., Preciado, J. C., Melchor, F., Rodriguez-Echeverria, R., Conejero, J. M., & Sanchez-Figueroa, F. (2021). A real-life machine learning experience for predicting university dropout at different stages using academic data. *IEEE Access*, *9*. https://doi.org/10.1109/ACCESS.2021.3115851

Flach, P. (2012). *Machine Learning: The Art and Science of Algorithms that Make Sense of Data* (1st ed.). Cambridge University Press.

Fu, Q., Gao, Z., Zhou, J., & Zheng, Y. (2021). CLSA: A novel deep learning model for MOOC dropout prediction. *Computers and Electrical Engineering*, *94*. https://doi.org/10.1016/j.compeleceng.2021.107315

Gue, I. H. v., Sy, A. M. T., Nuñez, A. B., Loresco, P. J. M., Onia, J. G. Y., Belino, M. C., & Onia, J. G. Y. (2021). A Rule Induction Framework on the Effect of 'Negative' Attributes to Academic Performance. *International Journal of Emerging Technologies in Learning*, *16*(15). https://doi.org/10.3991/ijet.v16i15.24269

Jiménez, O., Jesús, A., & Wong, L. (2023). Model for the Prediction of Dropout in Higher Education in Peru applying Machine Learning Algorithms: Random Forest, Decision Tree, Neural Network and Support Vector Machine. *Conference of Open Innovation Association, FRUCT*, *2023-May*. https://doi.org/10.23919/FRUCT58615.2023.10143068

Kim, S., Choi, E., Jun, Y. K., & Lee, S. (2023). Student Dropout Prediction for University with High Precision and Recall. *Applied Sciences (Switzerland)*, *13*(10). https://doi.org/10.3390/app13106275

Krüger, J. G. C., Britto, A. de S., & Barddal, J. P. (2023). An explainable machine learning approach for student dropout prediction. *Expert Systems with Applications*, *233*. https://doi.org/10.1016/j.eswa.2023.120933

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3).

Maquiavelo, N. (2017). El príncipe de Nicolás Maquiavelo. *Editorial Libsa*.

Mduma, N. (2023). Data Balancing Techniques for Predicting Student Dropout Using Machine Learning. *Data*, *8*(3). https://doi.org/10.3390/data8030049

Mehryar Mohri, Afshin Rostamizadeh, & Ameet Talwalkar. (2019). Foundations of machine learning, second edition. In *Statistical Papers* (Vol. 60, Issue 5).

Mubarak, A. A., Cao, H., & Zhang, W. (2020). Prediction of students' early dropout based on their interaction logs in online learning environment. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2020.1727529

Nagy, M., & Molontay, R. (2023). Interpretable Dropout Prediction: Towards XAI-Based Personalized Intervention. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-023-00331-8

Rebai, S., ben Yahia, F., & Essid, H. (2020). A graphically based machine learning approach to predict secondary schools performance in Tunisia. *Socio-Economic Planning Sciences*, *70*. https://doi.org/10.1016/j.seps.2019.06.009

Schargel, F., & Smink, J. (2014). Strategies to help solve our school dropout problem. In *Strategies to Help Solve Our School Dropout Problem*. https://doi.org/10.4324/9781315854090

Song, Z., Sung, S. H., Park, D. M., & Park, B. K. (2023). All-Year Dropout Prediction Modeling and Analysis for University Students. *Applied Sciences (Switzerland)*, *13*(2). https://doi.org/10.3390/app13021143

Vaarma, M., & Li, H. (2024). Predicting student dropouts with machine learning: An empirical study in Finnish higher education. *Technology in Society*, *76*. https://doi.org/10.1016/j.techsoc.2024.102474

Van der Malsburg, C. (1986). Frank Rosenblatt: Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. In *Brain Theory*. https://doi.org/10.1007/978-3-642-70911-1_20

Vanneschi, L., & Silva, S. (2023). Support Vector Machines. In *Natural Computing Series*. https://doi.org/10.1007/978-3-031-17922-8_10

Wan, H., Liu, K., Yu, Q., & Gao, X. (2019). Pedagogical Intervention Practices: Improving Learning Engagement Based on Early Prediction. *IEEE Transactions on Learning Technologies*, *12*(2). https://doi.org/10.1109/TLT.2019.2911284

Yagcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1). https://doi.org/10.1186/s40561-022-00192-z