J. Figueroa-Cañas and T. Sancho-Vinuesa, "Early Prediction of Dropout and Final Exam Performance in an Online Statistics Course" in IEEE-RITA, vol. 15, no. 2, pp. 86 - 94, MAY. 2020. doi: 10.1109/RITA.2020.2987727

Predicción Temprana del Abandono y Desempeño en el Examen Final en una Asignatura de Estadística en Línea

Josep Figueroa-Cañas y Teresa Sancho-Vinuesa

***Title*— Early prediction of dropout and final exam performance in an online statistics course**

***Abstract*— Higher education students who either do not complete the courses they have enrolled on or interrupt their studies indefinitely remain a major concern for practitioners and researchers. Within each course, early prediction of student dropout helps teachers to intervene in time to reduce dropout rates. Early prediction of course achievement helps teachers suggest new learning materials aimed at preventing at-risk students from failing or not completing the course. Several machine learning techniques have been used to classify or predict at-risk students, including tree-based methods, which, though not the best performers, are easy to interpret. This study presents two procedures for identifying at-risk students (dropout-prone and non-achievers) early on in an online university statistics course. These enable us to understand how classifiers work. We found that student dropout and course performance prediction was only determined by their performance in the first half of the formative quizzes. Nevertheless, other elements of participation on the virtual campus were initially considered. The classifiers will serve as a reference for intervention, despite their moderate performance metrics.**

***Index Terms*— Dropout prediction, performance prediction, decision trees, quiz completion, online university education.**

I. INTRODUCCIÓN

Este artículo es la versión extendida del presentado en el congreso LASI2019 [1]. En primer lugar, el presente trabajo añade robustez al método de clasificación, incluyendo una validación cruzada o cross validation (CV) en el procedimiento de evaluación del desempeño, mediante una aproximación de submuestreo aleatorio o *random undersampling approach* para tratar el problema del desesquilibrio de datos o *class imbalance*. En segundo lugar, se tienen en cuenta nuevos atributos de los estudiantes. Mientras que en la versión original, el único atributo considerado era la realización de pruebas evaluativas, en la actual versión se incluyen también las calificaciones de dichas pruebas. Y en tercer lugar, la nueva variable respuesta, *failure (fracaso)*, se añade a la variable *dropout (abandono)* del artículo original como variable a predecir en una fase temprana del curso.

Los estudiantes que o bien no acaban una asignatura o bien interrumpen sus estudios indefinidamente vienen causando gran preocupación desde hace mucho tiempo a profesorado e investigadores. *Dropout students* es el nombre con el que son denominados en inglés estos estudiantes. En asignaturas en línea, las elevadas tasas de abandono justifican la abundante investigación sobre este tema, como muestra la extensa revisión de [2], en la que se analizan 159 trabajos publicados entre 1999 y 2009. Más recientemente, tomando como referencia el marco europeo, se espera que para 2020 por lo menos el 40% de la población de treinta años haya finalizado con éxito estudios superiores. Ello implica que debe hacerse un esfuerzo importante en la reducción de las tasas de abandono en educación superior. Con el objetivo de mejorar nuestra práctica docente y condicionados por la estrategia del marco europeo, hemos decidido llevar a cabo una investigación sobre el no éxito en estudiantes de una asignatura de estadística en la Universitat Oberta de Catalunya.

El problema planteado en este trabajo se aborda desde la perspectiva de análisis de aprendizaje o *learning analytics* (LA). Inicialment LA fue definido como la medida, recogida, análisis y reporte de datos sobre quienes aprenden o *learners* y sus contextos, con el propósito de comprender y optimizar el aprendizaje y los entornos en los que tiene lugar [4]. Desde una perspectiva práctica, el *análisis y reporte de datos* se lleva a cabo mediante la creación de modelos predictivos. Los enfoques de LA y minería de datos educacionales o *educational data mining* (EDM) se sobreponen en varios aspectos, entre ellos la creación de modelos predictivos. No obstante, LA prioriza la interpretabilidad de los modelos sobre su optimización, de acuerdo con la definición establecida, *con el propósito de comprender el aprendizaje*, en contraste con el objectivo de

Manuscrito recibido el día de mes de año; revisado día de mes de año; aceptado día de mes de año.

English versión received September, 11-th, 2019. Revised October, 23rd, 2019. Accepted February, 12-th, 2020

Josep Figueroa-Cañas es estudiante de doctorado en la Universitat Oberta de Catalunya, España. (e-mail: jfigueroa@uoc.edu, ORCID ID: 0000-0002-6790-9142)

Teresa Sancho-Vinuesa es profesora de matemáticas en la Universitat Oberta de Catalunya, España (e-mail: tsancho@uoc.edu, ORCID ID: 0000-0002-0642-2912)

EDM [5]. Empleando un enfoque de LA, los modelos predictivos deben facilitar la intervención sobre los estudiantes con el fin de *optimizar el aprendizaje*. [6] y [7] son dos extensas revisiones de intervenciones, basadas en LA a partir de la creación de modelos predictivos, cuyo objetivo es la mejora del aprendizaje. Por citar un ejemplo, de intervención basada en LA, [8] muestra una reducción del 14% en las tasas de abandono de estudiantes mediante un plan de acción tutorial posterior a la identificación de estudiantes en riesgo de abandono. En el contexto de educación superior, encontramos dos niveles de abandono: (a) abandono a nivel de asignatura, y (b) abandono a nivel de programa. En el primer caso, el abandono tiene lugar dentro de la asignatura [9], donde los profesores pueden intervenir para evitarlo en una fase temprana, siempre que dispongan de información relevante. En el segundo caso, en el que los estudiantes abandonan sus estudios, el análisis requiere otros tipos de variables y las intervenciones dependen de otros miembros de la comunitat educativa, como por ejemplo directores de programas académicos. Los estudiantes que o bien no finalizan una asignatura o bien la finalizan aunque sin éxito, es decir, que no la aprueban, son candidatos a abandonar sus estudios.

Teniendo en cuenta que nuestro objetivo es comprender el proceso de aprendizaje y definir intervenciones a tiempo, nos proponemos identificar el mayor número posible de estudiantes en riesgo, en una asignatura de estadística, lo antes posible. Consideramos separadamente la existencia de dos tipos de estudiantes en riesgo: (a) estudiantes en riesgo de abandonar, y (b) estudiantes en riesgo de no aprobar. Los procedimientos que planteamos están basados en los de clasificación/predicción generando árboles de decisión binaria en varios instantes de tiempo a lo largo del curso. Los procedimientos tratan con datos que principalmente proceden tanto de pruebas de autoevaluación de carácter formativo (ya sea su realización o no, o las cualificaciones obtenidas), como de indicadores de compromiso o *engagement* en el aula virtual (tales como participación en el foro y número de acceso al tablón de anuncios).

En escenarios combinados o *blended* [10] y en entornos en línea [11], se ha demostrado que las pruebas de autoevaluación en línea de carácter formativo resultan ser predictores estadísticamente significativos del éxito de los estudiantes. Ambos trabajos [10, 11] tienen en cuenta todas las pruebas realizadas a lo largo del curso, de manera que el ciclo de (1) estudio de los contenidos, (2) realización de las pruebas, (3) recepción de realimentación o *feedback*, y (4) reestudio de los contenidos se va repitiendo durante todo el curso. Nuestro trabajo tiene el objetivo determinar si resulta posible predecir qué estudiantes no tendrán un buen desempeño, teniendo en cuenta únicamente los ciclos pertenecientes a la primera mitad del curso.

II. REVISIÓN DE LA LITERATURA

Un gran número de trabajos de investigación ha sido llevado a cabo en relación a modelos predictivos del éxito de los estudiantes [12]. La mayoría de los 121 autores de este trabajo definen "estudiantes sin éxito" como aquellos que o bien abandonan o suspenden sus asignaturas o cursos.

En el contexto de la educación superior, se distinguen tres niveles de abandono: (a) nivel de asignatura o curso, estudiado por diversos autores ([13] en asignaturas de informática y lingüística, [14, 15] en asignaturas de informática, y [16, 17] en asignaturas de matemáticas); (b) a nivel de grado, estudiado por [18] (en un grado de informática); y (c) a nivel institucional, objetivo de [19] (en una universidad a distancia, UDIMA). En el presente artículo, nos centramos en el abandono a nivel de asignatura.

Según [2], no existe un consenso en la definición de abandono a nivel de asignatura. Así, nos encontramos con varias definiciones en función de los autores. [20] directamente asocia abandono con fracaso: los estudiantes que abandonan o *dropout students* son aquellos que no obtienen calificaciones de A, B o C, es decir, aquellos que suspenden la asignatura o *failed students*. [21] define estudiantes que abandonan como aquellos que no finalizan la asignatura y cuyo importe de la matrícula no ha sido retornado. Y [22] considera que los estudiantes que abandonan son los que no finalizan, en sentido amplio. Así, en cierto sentido, los términos "dropout" y "failed student" pueden ser intercambiables a nivel de asignatura o curso. La definición de abandono desde la perspectiva de fracaso o *failure* [20] aparece en los trabajos de [9], [23] y [24]. La definición que hace [21] es mencionada explícitamente por [25], que añade un requisito adicional: un estudiante será considerado que abandona siempre que haya accedido a la plataforma de aprendizaje en línea al menos una vez durante el curso. Ello supone que para que se considere el abandono del estudiante, este debe haber dejado una traza en el sistema de información con anterioridad al momento de dejar la asignatura. Para [8] y [26], los estudiantes que no se presentan al examen final son aquellos a los que considera que han abandonado. [27] no indica ninguna definición precisa de estudiante que no finaliza o *non-completer student*.

En educación superior, los modelos predictivos se aplican en (a) entornos combinados o *blended* [9, 14, 16, 17], (b) entornos completamente en línea [8, 15, 23, 24, 28, 29] y también en (c) cursos masivos abiertos en línea o *massive online open courses* (MOOCS) [30, 31, 32, 33]. En entornos combinados, los instructores recopilan su propia información a través de su actividad presencial, hecho que contribuye a la mejora de los modelos. Por contra, en entornos en línea y en MOOCs, los instructores se encuentran con serias dificultades para incorporar nueva información a los modelos.

El objetivo de los modelos es predecir la retención o éxito, operacionalizado mediante variables diversas. [17] y [29] predicen los resultados finales a través de dos clases categóricas: aprobados y suspendidos. [25, 26] predicen el abandono mediante dos clases: estudiantes que abandonan y que finalizan. [8, 14] predicen las calificaciones finales en formato binario (aprobado-suspendido) y además en dos clases: puntuación alta (por lo menos 90 sobre 100) y puntuación no alta (menos que 90 sobre 100). [13] predice según la codificación académica: aprobado, suspendido, aprobado condicional y repetidor. [34] predice las cualificaciones finales de acuerdo a tres categorías: rendimiento alto (al menos 80.5 sobre 100), rendimiento medio (entre 57.5 sobre 100 y 80.5 sobre 100), y rendimiento bajo (inferior a 57.5 sobre 100). [15] predice las notas finales mediante una variable de intervalo continua.

Por lo que refiere a los atributos o características empleadas en los modelos, las tres categorías más comunes son: demográficas, interacciones con el entorno virtual de aprendizaje (VLE) o *virtual learning environment* (por ejemplo, la participación en el foro), y desempeño en actividades de evaluación o en exámenes. La primera

categoría está formada por datos invariantes en el tiempo que están disponibles desde el inicio del curso, mientras que las otras dos categorías incluyen datos que son recogidos incrementalmente a lo largo del curso. Datos demográficos tales como género e información profesional aparecen en los trabajos de [23, 24, 25, 27]. La participación en el foro es una característica estudiada por [9, 33, 35]. Por último, las calificaciones obtenidas en actividades de evaluación continuada son analizadas en [8, 15, 23, 24, 25, 26, 27, 28, 31, 32].

En cuanto a técnicas de aprendizaje de máquina o *machine learning techniques*, [15] utiliza una regresión de árbol de decisión dado que su variable respuesta es la calificación expresada como valor numérico. En casos en que los valores de la variable respuesta están divididos en clases, los metódos de clasificación empleados son: (a) algoritmos basados en árboles de decisión [13, 23, 24, 27, 29, 30, 35]; (b) algoritmos basados en redes neuronales [8, 23, 24, 25, 26, 27, 32, 35]; (c) máquinas de soporte vectorial o *support-vector machines* [8, 23, 24, 25, 26, 27, 35]; (d) naive Bayes [13, 23, 24, 27, 30, 35]; y (e) regresión logística [8, 26, 27, 31, 35].

Dentro del ámbito de la generación de modelos predictivos, el desequilibrio de datos es un problema ampliamente conocido. En general, los modelos están sesgados hacia la clase mayoritaria. Algunos autores, como [28] y [14, 23], específicamente tratan este problema empleando métodos de muestreo (enfoque a nivel de datos). [28] añade, además, la poderación de clases o *class weights* a los algoritmos de máquinas de aprendizaje (aproximación a nivel de algoritmo). Otros autores, en cambio, no tratan el problema. [30] ignora el tratamiento del problema aunque lo reconoce como limitación del trabajo, mientras que [13] y [29] simplemente no lo mencionan explícitamente.

El instante del tiempo en que se realiza la predicción difiere según los investigadores. [14] la realiza al final del curso, lo que supone que la predicción no conduce a ninguna intervención útil sobre los estudiantes. Las predicciones deben ser realizadas antes de alcanzar la mitad del curso para que resulten de utilidad [27]. En caso contrario, el profesorado no estará a tiempo de intervenir con éxito sobre los estudiantes. [36] sí lleva a cabo la predicción antes de la mitad del curso, pero utiliza atributos estáticos y, por lo tanto, no dispone de información ejecutable traducible a una intervención sobres los estudiantes. Una predicción en un único instante de tiempo es la opción escogida por [24] y [27], mientras que en múltiples instantes de tiempo, aunque no en los mismos, es la propuesta de [9], [8], [23], [25] y [26].

Respecto a las métricas de desempeño de los modelos, se encuentra una gran variedad de medidas de evaluación: precisión [5, 8, 9, 13, 25, 29, 30]; sensibilidad o *recall* [5, 8, 13, 25, 29, 30]; exactitud o *accuracy* [5, 8, 13 17, 24, 25, 30, 35, 36]; y medida-F o *F-measure* [13, 17, 23, 25, 29, 30, 31, 35]. AUC se utiliza en [13, 30, 31] y el error del cuadrado de la media o *Mean Square Error* (MSE) por [15]. Una curba ROC es estudiada por [16, 19]; una curva de precisión-sensibilidad por [37]; y una curba precisión-sensibilidad AUC por [28]. [36, 37] reportan el coeficiente Kappa de Cohen.

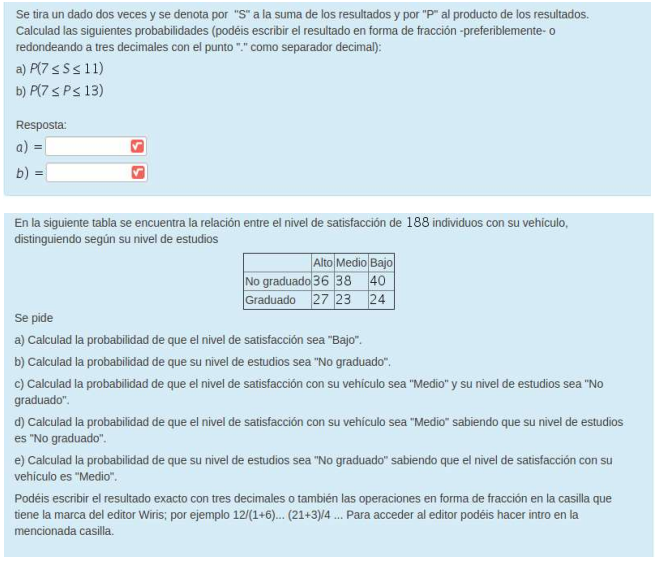
El presente artículo construye y analiza modelos predictivos tempranos para identificar estudiantes en riesgo en una asignatura de estadística en una universidad en línea. Si bien hemos considerado un amplio conjunto de atributos, nuestro modelo se basa en árboles de decisión que dependen únicamente de la media de las calificaciones de pruebas de evaluación, no obligatorias y de carácter formativo. Ello demuestra la eficacia de realizar regularmente ese tipo de pruebas, así como de obtener buenos resultados en las mismas, todo lo cual resulta ser una fortaleza de nuestra investigación.

III. METODOLOGÍA

A. Participantes y Contexto de Aprendizaje

Los participantes en este trabajo son los 197 estudiantes matriculados en el primer semestre del curso 2018/19 en la asignatura de estadística en línea, que forma parte del grado de Ingeniería en Informática de la Universitat Oberta de Catalunya.

La asignatura de estadística incluye dos instrumentos genéricos de evaluación: (a) un examen final presencial y obligatorio, y (b) una evaluación continuada a lo largo del semestre, no obligatoria. La calificación final de la asignatura se basa principalmente en la nota del examen final, modificada según una tabla de doble entrada. A modo de ejemplo, si un estudiante obtiene una nota de A+ en la evaluación continuada y una nota en el examen final de D+, la calificación final de la asignatura es una C. De esta manera, aquellos estudiantes que suspenden el examen final, pero cuya nota está próxima al umbral del aprobado, pueden llegar a aprobar la asignatura si su nota de evaluación continuada es suficientemente alta. Esto, sin duda, aumenta el incentivo para realizar la evaluación continua. El grueso de la evaluación continua es un conjunto de seis pares diferentes de pruebas (cuestionarios y tareas con R) conocidos como pruebas de evaluación continua. Cada cuesionario está formado por preguntas de opción múltiple y preguntas de respuesta corta que son corregidas y calificadas inmediatamente, proporcionando a los estudiantes feedback automático. La Fig. 1 ilustra un ejemplo de dos preguntas de respuesta corta extaídas del segundo cuestionario. Los cuestionarios se generan mediante la suite del *Moodle quiz authoring* con el soporte del WIRIS *plugin*. Cada vez que un estudiante realiza un cuestionario se encuentra preguntas diferentes. En cada una de las seis pruebas de evaluación continua, los estudiantes disponen de dos intentos para


Se tira un dado dos veces y se denota por "S" a la suma de los resultados y por "P" al producto de los resultados. Calculad las siguientes probabilidades (podéis escribir el resultado en forma de fracción -preferiblemente- o redondeando a tres decimales con el punto "." como separador decimal):

a) $P(7 \leq S \leq 11)$

b) $P(7 \leq P \leq 13)$

Resposta:

a) =

b) =

En la siguiente tabla se encuentra la relación entre el nivel de satisfacción de 188 individuos con su vehículo, distinguiendo según su nivel de estudios

	Alto	Medio	Bajo
No graduado	36	38	40
Graduado	27	23	24

Se pide

a) Calculad la probabilidad de que el nivel de satisfacción sea "Bajo".

b) Calculad la probabilidad de que su nivel de estudios sea "No graduado".

c) Calculad la probabilidad de que el nivel de satisfacción con su vehículo sea "Medio" y su nivel de estudios sea "No graduado".

d) Calculad la probabilidad de que el nivel de satisfacción con su vehículo sea "Medio" sabiendo que su nivel de estudios es "No graduado".

e) Calculad la probabilidad de que su nivel de estudios sea "No graduado" sabiendo que el nivel de satisfacción con su vehículo es "Medio".

Podéis escribir el resultado exacto con tres decimales o también las operaciones en forma de fracción en la casilla que tiene la marca del editor Wiris; por ejemplo 12/(1+6)... (21+3)/4 ... Para acceder al editor podéis hacer intro en la mencionada casilla.

Fig.1 Captura de pantalla de dos preguntas del segundo cuestionario

resolver un cuestionario. La calificación que pasa al registro es la más alta de las dos obtenidas en ellos. Las tareas con R, en cambio, están formadas por preguntas de respuesta razonada que deben ser resueltas utilitzando el programa estadístico R. Dichas tareas requieren de la corrección manual del profesor o profesora, de manera que el feedback proporcionado conlleva un cierto retraso. Tanto los cuestionarios como las tareas con R pueden ser consideradas principalmente como instrumentos de evaluación de carácter formativo dado que su principal objetivo es proporcionar a los estudiantes información que les ayude en su proceso de aprendizaje [38]. Los profesores, además, proponen una prueba inicial durante la primera semana para determinar cuál es el conocimiento previo de estadística a nivel de educación secundaria con el que inician la asignatura. Para alentar la realización del mismo, aquellos estudiantes que voluntariamente la realizan y la envían obtienen un bonus, que también incide en la calificación final de la evaluación continua.

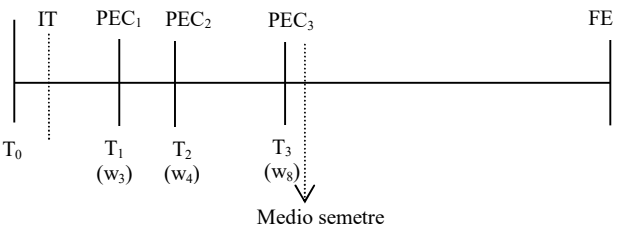
La Universitat Oberta de Catalunya pone a disposición de los estudiantes un campus virtual donde éstos realizan todas las actividades relacionadas con cada una de las asignaturas en que están matriculados. Cada aula virtual incluye acceso directo al plan docente, que contiene información específica en cuanto a metodología docente y sistema de evaluación. Además, existen tres espacios de interacción: un espacio de comunicación (el foro), un espacio de información (el tablón de anuncios) y un Moodle. Este último es utilizado por los estudiantes para realizar y enviar las pruebas de evaluación continua, y a la vez recibir feedback. El tablón de anuncios lo emplean los profesores para colgar información general de la asignatura. El foro permite a los estudiantes y profesores interactuar entre ellos, asíncronamente. Todos los accesos de lectura (al tablón, al foro y al plan docente) y de escritura (al foro) son almacenados por el sistema de información de la Universitat Oberta de Catalunya.

B. Medición y recogida de datos

La recogida de datos se ha realizado en tres instantes de tiempo (T_1, T_2, T_3), que coinciden con las fechas límite para la entrega de las tres primeras pruebas de evaluación (CAA), precisamente las de la primera mitad del curso (Fig. 2). La separación entre fechas límites es variable, oscilando entre 1 y 4 semanas. Definimos tres periodos de tiempo (*Period.1*, *Period.2*, *Period.3*) a partir de las fechas límite (T_i) de acuerdo con

$$Period.i = (T_{i-1}, T_i]$$

La selección de atributos para nuestro trabajo se basa en [8], [23] y [13]. Consideramos dos conjuntos de atributos: *Set_Engagement* y *Set_Achievement* (Tabla I). *Set_Engagement* está formado por dos atributos relacionados con la situación inicial del estudiante (*Repeating* y *EnrolledCourses*), tres atributos correspondientes al uso del aula virtual en cada uno de los tres instantes de tiempo T_i (*TeachingPlan_*T_i, *BBoard_*T_i, *ForumWr_*T_i y *ForumRe_*T_i), un atributo sobre la realización de la prueba de evaluación inicial (*T_InitialTest),* y dos atributos correspondientes a la realización de las tres pruebas de evaluación continua de los tres instantes de tiempo T_i (*N_Quizzes_*T_i y *N_RTasks_*T_i). Al margen de los atributos relacionados con la situación inicial del estudiante, el resto de ellos sirven como indicadores del compromiso o *engagement* del estudiante [12]. El conjunto de atributos *Set_Achievement* comparte los



T_i: Fecha límite para la entrega de la i-ésima Prueba de Evaluación Continuada (PEC_i). T_0 indica el inicio del semestre.
w_i: Semana i-ésima.
IT: Prueba inicial
FE: Examen Final

Fig.2 Distribución temporal de los instrumentos de evaluación.

dos atributos de la situación inicial del estudiante y los tres del uso del aula virtual con el conjunto *Set_Engagement*, mientras que sustituye los atributos de realización de pruebas de evaluación por atributos de desempeño en esas mismas pruebas (*S_InitialTest,* *A_Quizzes_*T_i y *A_RTasks_*T_i). [13] afirma que el uso de calificaciones como atributos resulta aceptable en contextos educativos donde las notas corresponden a pruebas de evaluación de carácter formativo, tal como ocurre en nuestro trabajo.

Durante el primer periodo (*Period.1*), recogemos datos de matriculación de los estudiantes, tales como el número de asignaturas matriculadas en el semestre y si el estudiante es repetidor o no. Estos datos, almacenados en el sistema de información de la Universitat Oberta de Catalunya y que nos han sido suministrados anónimamente, rellenan los atributos *Repeating* y *EnrolledCourses* de las instancias. El registro de actividad de Moodle es la fuente de información para

TABLA I
CONJUNTO DE ATRIBUTOS PARA EL PERIODO PERIOD. i (*)

Nombre	Descripción	Tipo y valores
Repeating (A,B)	Indica si el estudiante repite esta asignatura	Tipo: Booleana Valores: 1, 0
EnrolledCourses (A,B)	Indica el número total de asignaturas matriculadas en el semestre	Tipo: Entero Valores{1, …}
T_InitialTest(A)	Indica si el estudiante ha realizado la prueba inicial	Tipo: Booleana Valores: 1, 0
S_InitialTest(B)	Indica la nota obtenida en la prueba inicial	Tipo: Real Valores:[0,10]
*N_Quizzes_*T_i(A)	Indica el número de cuestionarios realizados hasta T_i	Tipo: Entero Valores:{0, …, i}
*A_Quizzes_*T_i(B)	Indica la nota media obtenida en los cuestionarios {1, …, i}	Tipo: Real Valores:[0,10]
*N_RTasks_*T_i(A)	Indica el número de tareas con R realizadas hasta T_i	Tipo: Entero Valores:{0, …, i}
*A_RTasks_*T_i(B)	Indica la nota media obtenida en las tareas con R {1, …, i}	Tipo: Real Valores:[0,10]
*TeachingPlan_*T_i (A,B)	Indica si el estudiante has visto el plan docente hasta T_i	Tipo: Booleana Valores: 1, 0
*BBoard_*T_i(A,B)	Indica el número de periodos en que el estudiante ha accedido al tablón de anuncios hasta T_i	Tipo: Entero Valores:{0, …, i}
*ForumWr_*T_i(A,B)	Indica el número de periodos en que el estudiante ha escrito mensajes en el foro hasta T_i	Tipo: Entero Valores:{0, …, i}
*ForumRe_*T_i(A,B)	Indica el número de periodos en que el estudiante ha leído mensajes en el foro hasta T_i	Tipo: Entero Valores:{0, …, i}

(*) con i = 1, 2,3 (A) Set_Engagement (B) Set_Achivement

determinar si un estudiante ha realizado y enviado la prueba inicial y la primera prueba de evaluación continua. Con estos datos, se rellenan los atributos *T_InitiaTest*, *N_Quizzes_T1* y *N_RTasks_T1* de todas las instancias. El registro de actividad del Moodle también proporciona las notas de la prueba inicial, de los cuestionarios y de las tareas con R, con lo que se calculan *S_InitiaTest*, *A_Quizzes_T1* y *A_RTasks_T1*. El registro de actividad de la aula virtual proporciona las fechas y horas de todos los accesos al foro, tablón de anuncios y plan docente que, después de ser preprocesados, rellenan los atributos *BBoard_T1*, *ForumWr_T1, ForumRe_T1* y *TeachingPlan_T1* de todas las instancias. Todos estos datos son transferidos al segundo periodo (*Period.2*), que incrementados con la información recopilada específicamente en este periodo rellena los atributos cuyo nombre termina en '_T_2' de todas las instancias. En el tercer período, se procede de manera análoga.

Para responder al propósito del presente trabajo, definimos dos variables distintas para los estudiantes en riesgo: (1) *dropout* y (2) *failure*.

1) La variable *dropout* se define a partir del abandono considerado por [21]. Un estudiante que abandona o *dropout student* es un estudiante que obtiene una calificación final en la asignatura de "No Presentado", lo que significa que el estudiante no ha realizado al examen final obligatorio (véase Fig. 3). Esta perspectiva es similar a la utilizada por [8] y [29]. La variable booleana *dropout* indica si el estudiante cumple o no con la definición previa, es decir, si el estudiante pertenece a la clase de estudiantes que abandonan (*dropout*) o a la clase de estudiantes que finalizan (*completer*). El sistema de información de la Universitat Oberta de Catalunya proporciona los datos para rellenar la variable *dropout* de las instancias. Combinando atributos, en condición de predictores, con la variable respuesta *dropout*, se crea el conjunto de datos [DS1]. Este contiene las intancias con todos los atributos del conjunto *Set_Engagement* (Tabla I) y con la variable *dropout* para cada uno de los tres períodos. De igual manera, las instancias con todos los atributos del conjunto *Set_Achievement* y con la variable respuesta *dropout* conforman la mayoría de datos del conjunto [DS2].

2) La variable *failure* se basa en la definción de [20]. Un estudiante que no aprueba es aquel que obtiene una nota en el examen final no superior o igual a 5, en una escala de 0 a 10. La variable booleana *failure* indica si un estudiante aprueba el examen final, es decir, si pertenece a la clase de estudiantes que aprueban (*passed*) o a la clase de estudiantes que no aprueban (*failed*). En la Fig.3 podemos observar como los estudiantes que abandonan (*dropout*) también son estudiantes que no aprueban (*failed*). El conjunto de datos [DS2] contiene también la variable *failure* de las instancias.

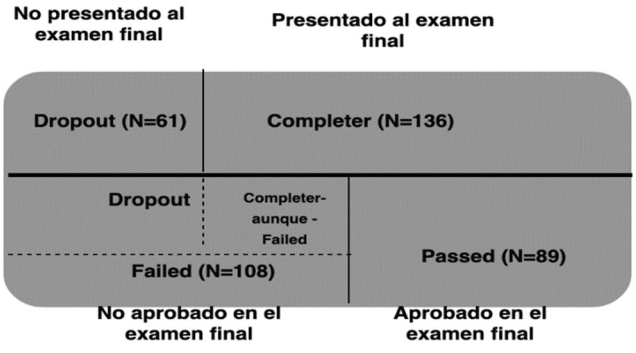


Fig.3 Distribución de los estudiantes

C. Método de clasificación

Nos hemos planteado un problema de clasificación/predicción, cuya solucion será un modelo de clasificación binaria o clasificador binario, que permita predecir si un estudiante será clasificado como estudiante en riesgo al final del semestre. Este trabajo obedece a la prioridad de interpretabilidad sobre optimización que sostiene el análisis de aprendizaje (LA) [5], dado que se pretende que el clasificador conduzca a una futura intervención docente. Hemos optado por métodos basados en árboles de decisión puesto que tal como defiende [39], estos métodos son sencillos y útiles por la claridad de su interpretación. [19] también utiliza árboles de decisión en la fase final de generación de modelos por su facilidad de interpretación. Nuestro objetivo no es tanto encontrar el clasificador de mejor desempeño, sinó uno que resulte fácilmente interpretable.

Un árbol de decisión binario es, esencialmente, un grafo orientado que se inicia en un nodo llamado raíz, que continua a través de arcos y ramas, y finaliza en nodos terminales llamados hojas. Cada nodo no terminal, incluyendo la raíz, representa un atributo, una pregunta de respueta sí/no respecto el valor de un cierto atributo. En nuestro trabajo, cada hoja representa una de las dos clases, o bien *dropout/completer* o *failed/passed*. Las ramas que proceden de un nodo representan los valores del atributo asociado al nodo, que están relacionados con la respuesta a la pregunta del propio atributo [40].

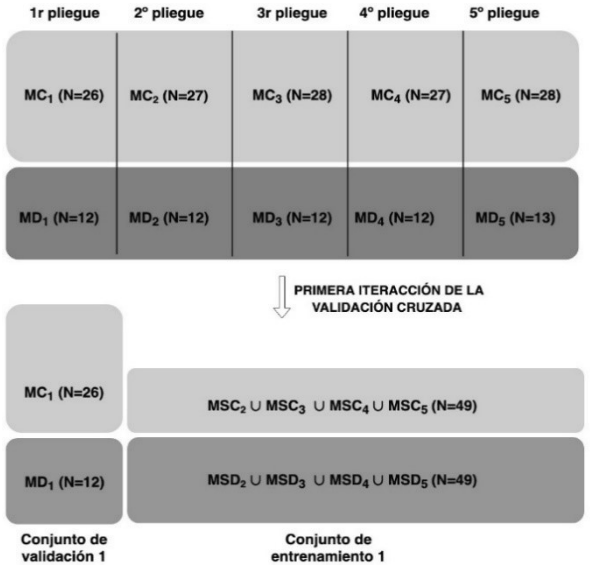
Tal como podemos observar en la Tabla I, no todos los atributos presentan el mismo número de valores posibles. En el proceso de creación de nodos, se produce un sesgo hacia atributos con mayor número de valores posibles, lo que resulta un problema detectado en trabajos que emplean modelos de árboles de decisión [41]. Hemos escogido modelos de árboles condicionales puesto que reducen ese sesgo [41]. Para la generación de los modelos de árboles condicionales hemos empleado la función *ctree()* disponible en el programa estadístico R.

Dado el conjunto de datos [DS1] y [DS2] para la variable *dropout*, el número de estudiantes que abandonan, correspondientes a la clase minoritaria *dropout*, es 61 y el número de estudiantes que finalizan, correspondiente a la clase mayoritario *completer*, es 136 (Fig. 3). Así pues, la ratio de estudiantes de la clase *dropout* respecto *completer* es 1:2.22, lo que indica que nos enfrentamos a un problema de desequilibrio de datos o *imbalanced data*. En esta situación, los clasificadores tienden a sesgar en favor de la clase mayoritaria (*completer*). Para corregirlo, se usan varias técnicas: métodos de muestro, métodos sensibles al coste, métodos kernel y métodos de aprendizaje activos [42]. La aplicación de métodos de muestreo mejora la exactitud o *accuracy* [42]. Entre estos, el submuestreo aleatorio o *random undersampling* es un método que ha sido ampliamente empleado. Consiste en eliminar aleatoriamente datos de una muestra previamente seleccionada de casos de la clase mayoritaria para igualar el tamaño de una muestra también seleccionada previamente de casos de la clase minoritaria. Nos hemos decidido por el submuestreo aleatorio para solventar nuestro problema de desequilibrio de datos. Además, para estimar el desempeño de nuestros modelos predictivos utilizamos una valización cruzada de

cinco pliegues o *five-fold cross validation* (CV). [43] afirma que el modo de combinar correctamente una CV de cinco pliegues con un submuestreo aleatorio es separar todos los datos en cinco subconjuntos y posteriormente reducir el tamaño del conjunto de entrenamiento o *traning set* a partir de un conjunto de entrenamiento original. Más detalladamente, seleccionamos aleatoriamente cinco subconjuntos o pliegues (Fig. 4), manteniendo la distribución de todo el grupo de estudiantes. Posteriormente, en la primera iteración, generamos el conjunto de validación tomando la totalidad de estudiantes del primer pliegue. El conjunto de entrenamiento lo generamos a partir de una submuestra de estudiantes que finalizan (*completer*s) y otra submuestra de estudiantes que abandonan (*dropout*) (Fig. 4), siendo la submuestra de *dropout* la unión de todas las muestras de la clase *dropout* contenidas en los pliegues del segundo al quinto. La submuestra de la clase *completer* es generada a través del submuestreo aleatorio de la muestra de la clase *completer* no contenida en el primer pliegue. No efectuando el submuestreo del conjunto de validación, nos aseguramos que todos los datos intervienen en la evaluación del desempeño del modelo. El resto de iteraciones se implementan análogamente. Se seleccionaron seis clasificadores para la variable *dropout*, uno para cada un de los dos conjuntos de datos y para cada uno de los tres periodos (Fig. 2).

Dado el conjunto de datos [DS2] para la variable *failure*, el número de estudiantes que no aprueban, correspondientes a la clase mayoritaria *failed* es 108 y el de estudiantes que sí aprueban, correspondientes a la clase mayoritaria *passer* es 89 (Fig. 3). Así pues, la ratio de estudiantes de la clase *passer* respecto *failed* es 1: 1.21, por lo que hemos considerado que no tenemos un problema de desequilibrio de datos.

De igual modo que en el caso anterior, hemos llevado a cabo una validación cruzada de cinco pliegues. Hemos separado todos los datos en cinco sunconjuntos o pliegues, manteniendo la distribución de subpoblaciones: estudiantes que aprueban (*passers*), estudiantes que abandonan (*dropout*) y estudiantes que finalizan pero suspenden o *completer-aunque-failed*. En cada iteración, se toma un subconjunto o pliegue como conjunto de validación y los cuatro restantes, como conjunto de entrenamiento. Hemos seleccionado tres clasificadores, uno para cada uno de los tres periodos (Fig. 2).



MCi: Muestra de *Completers* del pliegue i-ésimo
MD_i: Muestra de *Dropouts* del pliegue i-ésimo
MSC_i: Submuestra de *Completers* del pliegue i-ésimo

Fig.1 Validación cruzada con submuestreo aleatorio

Nuestra principal preocupación al predecir estudiantes en riesgo es reducir el número real de estudiantes que abandonan (*dropout*) o no aprueban (*failed*) que son clasificados erróneamente como estudiantes que finalizan (*completer)* o sí aprueban (*passer*), respectivamente. Teniendo en cuenta que nuestro objetivo es utilizar los modelos predictivos para implementar intervenciones sobre estudiantes en riesgo, nuestra intención es identificar tantos estudiantes como sea posible. Considerando a los estudiantes en riesgo como casos *positivos*, nuestro principal objectivo es conseguir valores bajos de *falsos positivos*. Por este motivo, hemos decidido hemos optado por medir el desempeño de nuestro modelo empleando principalmente la métrica *sensibilidad* o *recall*. Nuestra segunda preocupación es lograr valores bajos de *falsos negativos* para evitar que el modelo se dirija a estudiantes que no están realmente en riesgo. Por esto último, hemos considerado también la métrica *precisión*. Finalmente, también hemos optado por la combinación de ambas métricas, más concretamente su media armónica, la *Medida-F* o *F-measure*. [29, p. 3] utiliza las mismas métricas dado que el objetivo es primordialmente reconocer estudiantes en riesgo. Hemos empleado las definiciones siguientes, de acuerdo con [44]:

$$Precisión = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Positivos} \quad (1)$$

$$Sensibilidad = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + FalsosNegativos} \quad (2)$$

$$Medida - F = 2 * \frac{Sensibilidad * Precisión}{Sensibilidad + Precisión} \quad (3)$$

IV. RESULTADOS

A. Clasificadores de abandono y de finalización

Se obtuvieron dos grupos de clasificadores para la variable *dropout*, uno para cada conjunto de atributos: *Set_Engagement* y *Set_Achievement*.

A.1) Clasificadores de abandono (Dropout) empleando el conjunto Set_Engagement

Las métricas de desempeño de los tres modelos creados, empleando el compromiso o *engagement* en las pruebas de evaluación (a través de la realización de pruebas de evaluación), se encuentran por debajo del 70% (Tabla II), demostrando así falta de poder predictivo según [13]. Cada métrica del desempeño de la clasificación ha sido calculada mediante el promedio de las cinco métricas de cada una de las cinco iteraciones correspondientes a la validación cruzada de cinco pliegues.

A.2) Clasifiadores de abandono (Dropout) empleando el conjunto Set_Achievement

Al sustituir compromiso en las pruebas de evaluación por logros en las pruebas de evaluación (mediante las notas de las pruebas de evaluación), el modelo creado después del final de la tercera prueba de evaluación continua (*Model.Dropout.T3*) muestra métricas de desempeño por

TABLA II
MÉTRICAS DE DESEMPEÑO DE LOS CLASIFICADORES DE ABANDONO (DROPOUT) PARA EL CONJUNTO *SET_ENGAGEMENT*

Modelos	*Sensibilidad*	*Precisión*	*Medida F*
Tras T_1	55.8%	66.6%	52.8%
Tras T_2	61.0%	65.9%	61.0%
Tras T_3	65.8%	62.3%	62.7%

TABLA III
MÉTRICAS DE DESEMPEÑO DE LOS CLASIFICADORES DE ABANDONO (DROPOUT) PARA EL CONJUNTO *SET_ACHIEVEMENT*

Modelos	*Sensibilidad*	*Precisión*	*Medida F*
Tras T_1	57.2%	69.0%	61.9%
Tras T_2	60.9%	77.9%	67.5%
Tras T_3	75.1%	70.3%	72.0%

encima del 70% (Tabla III). En el resto de modelos creados, en los instantes T_2 y T_1, los valores de las métricas indican peores desempeños. Estos resultados están situados en el extremo inferior del rango de [8], que reporta valores de sensibilidad entre 69.23% y 96.73% al predecir abandono en la mitad del curso. El trabajo de [8] tiene lugar en un entorno completamente en línea, empleando algoritmos de máquinas de aprendizaje de *caja negra*, en contraposición a nuestro algoritmo de árbol de decisión de *caja blanca*. Las diferencias en la exactitud o *accuracy* en la predicción de [8], con inferior exactitud al emplear algoritmos basados en árboles de decisión, están en consonancia con [39].

El modelo resultante *Model.Dropout.T3* resulta extremadamente sencillo, si tenemos en cuenta que únicamente está compuesto por un nodo (Fig.5). El único atributo contenido que forma el modelo es el promedio de las notas de los tres primeros cuestionarios (*A_Quizzes_T3*), lo que prueba que este es el atributo más fuertemente asociado [41] con la variable respuesta *dropout*.

B. Clasificadores de desempeño

El modelo creado tras las tres primeras pruebas de evaluación continua (*Model.Failure.T3*) muestra métricas de desempeño superiores al 75% (Tabla IV). El resto de modelos generados en los instantes T_2 y T_1 muestran peores medidas de desempeño, algunas por debajo del 70%. [23], en su trabajo sobre una asignatura de informática en línea, presenta *Medidas-F* en un rango entre 77% y 82% en la predicción del resultado final (aprobado/suspendido) efectuada en la mitad del curso. En la franja baja del rango encontramos el método J48 para árboles de decisión. Nuestra *Medida-F* (76.3%) es ligeramente inferior a la de [23], que utiliza notas de tareas y uso de herramientas de aprendizaje virtual, tal como nosotros. [29] reporta *Medidas-F* inferiores a 50%, igualmente al predecir resultados finales de aprobado/suspendido a mitad de curso, empleando en su caso el método C4.5 para árboles de decisión, el único método en ese trabajo. [29] utiliza notas de tareas y datos de compromiso o *engagement* y herramientas de aprendizaje virtual en varias asignaturas de arte, de matemáticas y de negocios en línea. Nuestra M*edida-F* es ligeramente superior a la de [29].

El modelo *Model.Failure.T3* resulta también extremadamente sencillo dado que contiene un único nodo (Fig. 6). Únicamente el atributo relacionado con el logro en los cuestionarios posee poder discriminatorio (*A_Quizzes_T3)*. La diferencia para que un estudiante sea clasificado como estudiante que abandona (*dropout*) y que suspende (*failed*) es aproximadamente 1 punto en una escala de 0 a 10, observando los modelos *Model.Dropout.T3* y *Model.Failure.T3,* cantidad que no resulta demasiado relevante.

TABLA IV
MÉTRICAS DE DESEMPEÑO DE LOS CLASIFICADORES DE DESEMPEÑO EN EL EXAMEN FINAL (FAILURE)

Modelos	*Sensibilidad*	*Precisión*	*Medida F*
Tras T_1	63.0%	72.6%	66.9%
Tras T_2	66.7%	77.2%	70.6%
Tras T_3	75.9%	78.3%	76.3%

V. DISCUSIÓN

En esta sección reseñamos la interpretación y utilidad de los modelos que hemos generado.

A. La predicción del abandono y desempeño está determinada por el desempeño en las pruebas de evaluación

Tal y como se mencionó en la sección IV en cuanto a la predicción del abandono, al considerar atributos del conjunto *Set_Engagement*, las métricas de evaluación presentan valores bajos. Consecuentemente afirmamos que la realización de cuestionarios, atributos incluídos en el conjunto *Set_Engagement,* no determina la predicción del abandono. En cambio, al considerar atributos del conjunto *Set_Achievement,* tras tres pruebas de evaluación continua el modelo ciertamente posee poder discriminador. En realidad, un solo atributo presenta poder discriminador: *A_Quizzes_T3*, que está directamente relacionado con el desempeño en los cuestionarios. De todo ello también se desprende que los atributos relacionados con el uso del foro carecen de poder discriminatorio. El mismo razonamiento es válido en cuanto al desempeño de la predicción de los clasificadores de desempeño en el examen final.

El dominio de los atributos de pruebas evaluativas está en línea con [23], donde los atributos más determinantes son

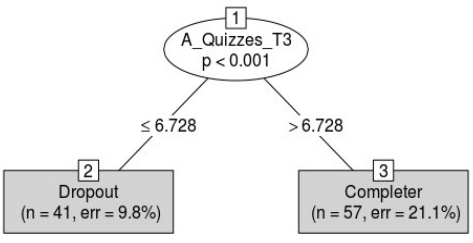


Fig. 5 Model.Dropout.T3

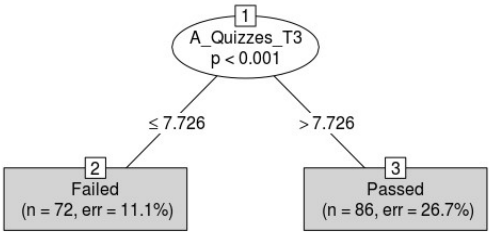


Fig. 6 Model.Failure.T3

las notas de mitad de curso.

B. *Las notas de los cuestionarios son más determinantes que las de las tareas con R*

A pesar de que las tareas con R también tienen un carácter evaluativo, no intervienen en los modelos de predicción tal y como sí hacen los cuestionarios. La principal diferencia entre cuestionarios y tareas con R radica en que estas últimas obligan a los estudiantes a redactar una respuesta extensa en que deben comunicarse matemáticamente. Aunque en principio podíamos pensar que las notas en tareas más exigentes, como las tareas con R, deberían ser más determinantes en la predicción del desempeño en el examen final, las cifras obtenidas nos dicen otra cosa. La estructura del examen final, en que el peso de las cuestiones que requieren habilidades relacionadas con R no supera el 25%, puede afectar al poder discriminatorio de las notas de las tareas con R.

C. *Los modelos de abandono y desempeño son mutuamente coherentes*

Simplificando, hemos clasificado los estudiantes que abandonan y los que suspenden según la calificación obtenida en el examen final. Los primeros se corresponden con la calificación de no presentado, mientras que los segundos, con una nota que no alcanza 5 sobre 10. Así pues, la diferencia entre los dos tipos de estudiantes queda determinada por el valor de corte establecido en la calificación del examen final, superior en el caso de aprobados. Los modelos de predicción para abandono y desempeño únicamente difieren en el promedio de la puntuación de los cuestionarios, siendo superior en el caso del desempeño. Ello es coherente con la separación de estudiantes considerado en nuestro trabajo.

D. *La simplicidad de los modelos de predicción de abandono y desempeño facilitan el procedimiento de clasificación*

La simplicidad de los modelos, tanto de abandono como de desempeño, reducen el volumen de información a, básicamente el desempeño en los cuestionarios. Hecho que a su vez conduce a dos consecuencias beneficiosas: (a) se elimina el tiempo dedicado a la recogida y procesado del resto de atributos; (b) el profesorado puede recoger los datos directamente del registro de actividades del Moodle. Aún así, el profesorado debe ser conscientes de que los modelos no son perfectos, debido a que los valores de las métricas de desempeño de los clasificadores se hallan en una zona media.

E. *Las notas de las pruebas de evaluación continua están asociadas con el desempeño en el examen final*

El promedio de las notas de los tres cuestionarios programados previamente a la cuarta prueba de evaluación continua (PEC_4) aparece en el modelo de predicción del desempeño *Model.Failure.T3*. Las notas de los tres primeros cuestionarios, incluídos en las notas de las PEC_1, PEC_2 y PEC_3, están consecuentemente asociados con la variable respuesta *failure*, y así también con el desempeño en el examen final de los estudiantes. Esto último está en consonancia con una asociación similar hallada entre notas de pruebas evaluativas de carácter formativo y el desempeño en el examen final en una asignatura de análisis matemático en línea [11], y en una asignatura de tecnología de la información en un entorno combinado o *blended* [10].

VI. CONCLUSIÓN Y FUTURAS INVESTIGACIONES

En una asignatura de estadística en línea en que se utilizan cuestionarios como instrumento de evaluación formativa, encontramos que dichos cuestionarios juegan un papel muy destacado en la predicción temprana de estudiantes en riesgo de abandonarla o suspenderla.

La principal contribución de este trabajo es presentar un procedimiento sencillo e interpretable para la identificación, antes de la mitad del curso, de estudiantes en riesgo de abandonar o suspender, empleando modelos de clasificación basados en árboles de decisión. El procedimiento es directo dado que los estudiantes son clasificados teniendo en cuenta un único atributo, relacionado con el desempeño en tareas evaluativas de baja puntuación o *low-stake assessment assignments*, tales como cuestionarios planificados por el profesorado, y no relacionados con el uso de elementos del entorno de aprendizaje virtual, como el foro. Además, puesto que la información que se necesita resulta fácilmente accesible al profesorado y no requiere preprocesado, es posible controlar el procedimiento de forma autónoma. Los modelos nos indican que el principal factor que contribuye al desempeño en el examen final es el aprendizaje continuado adquirido durante como mínimo la primera mitad del curso. Ello prueba que, en una asignatura de matemáticas o estadística, siempre difícil e incluso aún más si tenemos en cuenta el entorno en línea, el compromiso constante y regular en las actividades de aprendizaje resulta capital. Dejarlo todo para el examen final, no parece una buena estrategia, al menos para la mayoría de los estudiantes.

Con respecto a la simplicidad del procedimiento, en un entorno en línea similar, [8] difiere del nuestro en que presenta múltiples modelos de predicción de abandono mucho más complejos. También trabajando en entornos en línea, los modelos de [29] resultan fáciles de interpretar dado que también están basados en árboles de decisión. Sin embargo, contrariamente a nosotros, [29] no detalla explícitamente qué atributos son más significativos. Nuestras métricas de evaluación del desempeño de la predicción muestran valores superiores a los de [29], aunque en la franja baja de los reportados por [8] y [29]. A pesar de que reconocemos que nuestras medidas de evaluación del desempeño en la clasificación presentan amplio margen de mejora, las consideramos suficientes para justificar una intervención sobre estudiantes, tal como propone [37] con unos valores incluso peores (entre 65% y 70%).

La primera limitación que merece ser mencionada es el bajo valor de las métricas de evaluación del desempeño en la predicción, del cual el profesorado debe ser consciente. La segunda limitación está asociada al conjunto de validación. Según la metodología empleada, los estudiantes que forman el conjunto de entrenamiento y sobre los cuales se han ajustado los modelos de clasificación, así como los estudiantes que forman el conjunto de validación, sobre los que se han efectuado las medidas de evaluación del desempeño de la clasificación, pertenecen a la promoción de estudiantes matriculados en el mismo curso académico. Ambas limitaciones sugieren futuros trabajos de investigación. En primer lugar, más variables deben ser incluídas en el conjunto de atributos para incrementar la exactitud de la predicción. Variables relacionadas con

estrategias de aprendizaje autoregulado o *self-regulated learning*, y motivación podrían ser incorporadas, por poner algunos ejemplos. En segundo lugar, se debería considerar un conjunto de pueba (*test set*) empleando estudiantes de un curso académico distinto al de los estudiantes del conjunto de entrenamiento, tal como hacen [25] y [26]. Además, dado que nuestra investigación ha generado un modelo predictivo capaz de proporcionar suficiente información traducible en acciones, una futura investigación se concentrará en intervenciones docentes sobre estudiantes en riesgo que incluirá un análisis de la eficacia de la acción llevada a cabo, en términos de resultados académicos y motivación.

AGRADECIMIENTOS

Esta investigación ha sido parcialmente financiada por una beca de la Fundació IBADA y por el proyecto 2017SGR1619 del Gobierno de Catalunya. Queremos agradecer al Sr. Paul Garbutt y a la Dra. Laura Calvet por sus valiosos comentarios y sugerencias que han ayudado a la mejora de este trabajo.

REFERENCIAS

[1] J. Figueroa-Cañas and T. Sancho-Vinuesa, "Predicting early dropout student is a matter of checking completed quizzes: the case of an online statistics module", *CEUR Workshop Proc.*, vol. 2415, pp. 100–111, 2019.

[2] Y. Lee and J. Choi, "A review of online course dropout research: Implications for practice and future research," *Educ. Technol. Res. Dev.*, vol. 59, no. 5, pp. 593–618, 2011.

[3] H. Vossensteyn et al., Drop-Out and Completion in Higher Education in Europe - Literature Review. European Union, 2015.

[4] LAK11. Description of the 1st International Conference on Learning Analytics and Knowledge 2011 (LAK11). Banff, Alberta. Retrieved from https://tekri.athabascau.ca/analytics/

[5] W. Xing, R. Guo, E. Petakovic, and S. Goggins, "Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory," *Comput. Human Behav.*, vol. 47, pp. 168–181, 2015.

[6] A. Larrabee Sønderlund, E. Hughes, and J. Smith, "The efficacy of learning analytics interventions in higher education: A systematic review," *Br. J. Educ. Technol.*, vol. 50, no. 5, pp. 2594–2618, 2019.

[7] P. Sander and I. Services, "Using Learning Analytics to Predict Academic Outcomes of First-year Students in Higher Education," *CAPSTONE Rep. Pete Sander Manag. Inf. Serv. Oregon State Univ. Univ. Oregon Appl. Inf. Manag. Progr. Spring*, vol. 1277, no. 800, pp. 2–41, 2016.

[8] C. Burgos, M. L. Campanario, D. de la Peña, J. A. Lara, D. Lizcano, and M. A. Martínez, "Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout," Comput. Electr. Eng., vol. 66, pp. 541–556, 2018.

[9] A. Cohen, "Analysis of student activity in web-supported courses as a tool for predicting dropout," Educ. Technol. Res. Dev., vol. 65, no. 5, pp. 1285–1304, 2017.

[10] M. Ćukušić, Ž. Garača, and M. Jadrić, "Online self-assessment and students' success in higher education institutions," *Comput. Educ.*, vol. 72, pp. 100–109, 2014.

[11] J.Figueroa-Cañas and T. Sancho-Vinuesa, "Investigating the relationship between optional quizzes and final exam performance in a fully asynchronous online calculus module," Interact. Learn. Environ., 2018

[12] Y. Cui, F. Chen, A. Shiri, and Y. Fan, "Predictive analytic models of student success in higher education: A review of methodology," Inf. Learn. Sci., vol. 120, no. 3–4, pp. 208–227, 2019.

[13] C. C. Gray and D. Perkins, "Utilizing early engagement and machine learning to predict student outcomes," Comput. Educ., vol. 131, no. July 2018, pp. 22–32, 2019.

[14] X. Xu, J. Wang, H. Peng, and R. Wu, "Prediction of academic performance associated with internet usage behaviors using machine learning algorithms," *Comput. Human Behav.*, vol. 98, no. April, pp. 166–173, 2019.

[15] E. Wakelam, A. Jefferies, N. Davey, and Y. Sun, "The potential for student performance prediction in small cohorts with minimal available attributes," *Br. J. Educ. Technol.*, vol. 0, no. 0, 2019.

[16] H. Hirose, "Success/Failure Prediction for Final Examination Using the Trend of Weekly Online Testing," *Proc. - 2018 7th Int. Congr. Adv. Appl. Informatics, IIAI-AAI 2018*, no. 17, pp. 139–145, 2018.

[17] R. Umer, T. Susnjak, A. Mathrani, and S. Suriadi, "A learning analytics approach: Using online weekly student engagement data to make predictions on student performance," *2018 Int. Conf. Comput. Electron. Electr. Eng. ICE Cube 2018*, pp. 1–5, 2019.

[18] C. Lacave, A. I. Molina, and J. A. Cruz-Lemus, "Learning Analytics to identify dropout factors of Computer Science studies through Bayesian networks," *Behav. Inf. Technol.*, vol. 37, no. 10–11, pp. 993–1007, 2018.

[19] A. Ortigosa, R. M. Carro, J. Bravo-Agapito, D. Lizcano, J. J. Alcolea, and Ó. Blanco, "From Lab to Production: Lessons Learnt and Real-Life Challenges of an Early Student-Dropout Prevention System," *IEEE Trans. Learn. Technol.*, vol. 12, no. 2, pp. 264–277, 2019.

[20] S. Liu, J. Gomez, and C.-J. Yen, "Community College Online Course Retention and Final Grade: Predictability of Social Presence," J. Interact. Online Learn., vol. 8, no. 2, pp. 165–182, 2009.

[21] Y. Levy, "Comparing dropouts and persistence in e-learning courses," Comput. Educ., vol. 48, no. 2, pp. 185–204, 2007.

[22] P. A. Dupin-Bryant, "Pre-Entry Variables Related to Retention in Online Distance Education," Am. J. Distance Educ., vol. 18, no. 4, pp. 199–206, 2011.

[23] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, and J. Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses," Comput. Human Behav., vol. 73, pp. 247–256, 2017.

[24] M. A. Santana, E. B. Costa, B. F. S. Neto, I. C. L. Silva, and J. B. A. Rego, "A predictive model for identifying students with dropout profiles in online courses," in Proceeding of the 8th international conference on educational data mining, EDM workshops, 2015.

[25] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, "Dropout prediction in e-learning courses through the combination of machine learning techniques," Comput. Educ., vol. 53, no. 3, pp. 950–965, 2009.

[26] J. A. Lara, D. Lizcano, M. A. Martínez, J. Pazos, and T. Riera, "A system for knowledge discovery in e-learning environments within the European Higher Education Area - Application to student data from Open University of Madrid, UDIMA," Comput. Educ., vol. 72, pp. 23–36, 2014.

[27] S. B. Kotsiantis, C. J. Pierrakeas, and P. E. Pintelas, "Preventing Student Dropout in Distance Learning Using Machine Learning Techniques," in Proceeding of the 7th International Conference on Knowledge-Based Intelligent Information and Engineering Systems, KES 2003., 2003, no. September 2003, pp. 267–274.

[28] M. Hlosta, Z. Zdrahal, and J. Zendulka, "Ouroboros: Early identification of at-risk students without models based on legacy data," *ACM Int. Conf. Proceeding Ser.*, pp. 6–15, 2017.

[29] A. Wolff, Z. Zdrahal, A. Nikolov, and M. Pantucek, "Improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment," in *Third Conference on Learning Analytics and Knowledge (LAK13)*, 2013, no. April 2013.

[30] A. Y. Q. Huang, O. H. T. Lu, J. C. H. Huang, C. J. Yin, and S. J. H. Yang, "Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs," *Interact. Learn. Environ.*, vol. 0, no. 0, pp. 1–25, 2019.

[31] J. A. Ruiperez-Valiente, P. J. Muñoz-Merino, A. Andújar, and C. Delgado-Kloos, "Early Prediction and Variable Importance of Certificate Accomplishment in a MOOC," in Proceedings of the European Conference on Massive Open Online Courses, 2017, no. May, pp. 263–272.

[32] K. Sharma, L. Kidzinski, P. Jermann, and P. Dillenbourg, "Towards Predicting Success in MOOCs: Programming Assignments," Proc. Eur. Stakehold. SUMMIT Exp. best Pract. around MOOCs (EMOOCS 2016), pp. 135–148, 2016.

[33] D. Yang, T. Sinha, D. Adamson, and C. Penstein Rose, "'Turn on, Tune in, Drop out': Anticipating Student Dropouts in Massive Open Online Courses," in Proceedings of the 2013 NIPS Data-driven education workshop, 2013, pp. 1–8.

[34] C. J. Villagrá-Arnedo, F. J. Gallego-Durán, F. Llorens-Largo, P. Compañ-Rosique, R. Satorre-Cuerda, and R. Molina-Carmona, "Improving the expressiveness of black-box models for predicting student performance," *Comput. Human Behav.*, vol. 72, pp. 621–631, 2017.

[35] C. Romero, M. I. López, J. M. Luna, and S. Ventura, "Predicting students' final performance from participation in on-line discussion forums," Comput. Educ., vol. 68, pp. 458–472, 2013.

[36] L. M. Abu Zohair, "Prediction of Student's performance by modelling small dataset size," *Int. J. Educ. Technol. High. Educ.*, vol. 16, no. 1, pp. 1–18, 2019.

[37] R. S. Baker, D. Lindrum, M. J. Lindrum, and D. Perkowski, "Analyzing Early At-Risk Factors in Higher Education e- Learning Courses," *Proc. 8th Int. Conf. Educ. Data Min.*, pp. 150–155, 2015.

[38] M. Yorke, "Formativeassessment in higher education: moves towards theory and the enhancement of pedagogic practice," *Higher Education*, vol. 45, nº. 4, pp. 477–501, 2003.

[39] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning, vol. 103. 2013.

[40] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: A review of classification and combining techniques," Artif. Intell. Rev., vol. 26, no. 3, pp. 159–190, 2006.

[41] T. Hothorn, K. Hornik, and A. Zeileis, "Unbiased recursive partitioning: A conditional inference framework. Research Report Series 8, Department of Statistics and Mathematics, WU Wien, 2004," J. Comput. Graph. Stat., vol. 15, no. 3, pp. 651–674, 2006.

[42] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.

[43] R. Blagus and L. Lusa, "Joint use of over-and under-sampling techniques and cross-validation for the development and assessment of prediction models," *BMC Bioinformatics*, vol. 16, no. 1, pp. 1–10, 2015.

[44] R. Pelánek, "A brief overview of metrics for evaluation of student models", *CEUR Workshop Proc.*, vol. 1183, no. 2, pp. 151–152, 2014.

[DS1] https://figshare.com/s/e6efef99ef8deb86c959

[DS2] https://figshare.com/s/da3105c22454d6a9e72f

Josep Figueroa-Cañas, nacido en Manres (España) es estudiante de doctorado del programa de e-Learning en la Universitat Oberta de Catalunya y licenciado en física por la Universitat Autònoma de Barcelona (España, 1988).

Teresa Sancho-Vinuesa, nacida en Barcelona (España), es doctora en ingeniería electrónica por la Universitat Ramon Llull (España, 1995) y licenciada en matemáticas por la Universitat of Barcelona (España, 1990).

Actualmente es profesora agregada de la Universitat Oberta de Catalunya (España) y dirige el grupo de investigación Learning Analytics for Innovation and Knowledge Application in Higher Education (LAIKA) y ha sido investigadora visitante en la Open UK. Durante el período 1990-1996, ha sido docente e investigadora en la Escuela de Ingeniería y Arquitectura, la Salle (España). Ha sido miembro del equipo pedagógico y editorial de TEXT de la Enciclopedia catalana (España). En la Universitat Oberta ha ocupado diversos cargos de responsabilidad: directora del programa de doctorado en Sociedad de la información y el conocimiento, Directora de investigación y vicerrectora de investigación e innovación. Aunque desde su ingreso en la Oberta de Catalunya ha trabajado en temas relacionados con la educación superior e Intenet. Actualmente concentra su actividad investigadora en el uso de learning analytics para la mejora de la enseñanza y aprendizaje en línea, en particular, en los procesos de evaluación y feedback. La profesora Sancho-Vinuesa ha participado en más de 10 comités técnicos y científicos y es revisora de distintas revistas académicas de reconocido prestigio en el ámbito educativo.

.