

Hacia un sistema de detección temprana del riesgo de abandono en entornos en línea

M. Elena Rodríguez, David Bañeres, Ana Elena Guerrero-Roldán
Estudios de Informática, Multimedia y Telecomunicaciones
Universitat Oberta de Catalunya
Barcelona
{mrodriguezgo, dbaneres, aguerrror}@uoc.edu

Resumen

El abandono del estudiantado constituye un tema de preocupación en educación superior, y en especial en universidades en línea. En este trabajo se presenta un modelo de predicción, basado en algoritmos de aprendizaje automático, capaz de detectar de forma temprana estudiantado en riesgo de abandonar una asignatura. Para ello, el modelo analiza datos sobre el perfil del estudiantado, así como de su progresión y nivel de implicación en la asignatura, proporcionando una predicción diaria del riesgo de abandono. El modelo se integra en un sistema de alerta temprana capaz de realizar intervenciones orientadas a evitar el abandono y que ofrece diferentes cuadros de mando al profesorado. El sistema es configurable por el profesorado a nivel del contenido de las intervenciones y el número de días consecutivos que el estudiantado es marcado en riesgo por el modelo para considerarlo un abandono potencial. El sistema se ha probado en una asignatura de primer año del Grado de Ingeniería Informática en el primer semestre del curso académico 2020-2021.

Abstract

Course dropout is a challenging problem in higher education, especially in online universities. This paper presents a predictive model based on machine learning techniques to detect at-risk course dropout learners early. To this end, the model analyzes data from learners' profile, performance, and course engagement, providing a course dropout daily prediction. The model is integrated within an early warning system able to perform interventions oriented to avoid dropout. It also offers different dashboards to the teachers. The system is configurable by the teachers to define the contents of the interventions and the number of consecutive days that the learners should be detected at-risk by the model to be considered a potential dropout. The system has been tested in a first-year undergraduate course of the Degree of

Computer Science in the first semester of the 2020-2021 academic year.

Palabras clave

Riesgo de abandono, educación en línea, modelos de predicción, minería de datos educativos, sistema de detección temprana.

1. Motivación

El abandono del estudiantado constituye un elemento de preocupación en educación superior. Los grados de Ingeniería Informática tampoco son ajenos a esta problemática [4, 7]. A pesar de la amplia literatura existente, es un problema complejo que incluye múltiples factores [1] y que puede ser abordado desde diferentes perspectivas como sería el abandono de las asignaturas matriculadas, de un programa académico, o de la universidad [14].

Las universidades en línea registran mayores tasas de abandono [5, 14]. En [5] se concluye que, aunque las causas de abandono se centran principalmente en el estudiantado, el contexto institucional y de enseñanza también impactan. En el caso de nuestra universidad, se han obtenido resultados similares que cobran más relevancia en el estudiantado novel, frecuentemente sin experiencia previa en educación en línea. En [11], el estudiantado adujo factores personales y familiares como los más relevantes para el abandono. A su vez, dichos factores se relacionaban con problemas de gestión del tiempo de estudio agravado por el número de asignaturas matriculadas. Como factores pedagógicos, se incluyó la dificultad de las asignaturas y las actividades, así como la utilidad de los recursos de aprendizaje. Un estudio similar, efectuado desde la perspectiva del profesorado y disponible en [15], concluye que las razones más significativas para el abandono son los problemas relativos a la gestión del tiempo (autorregulación), así como una concepción errónea del estudiantado respecto a la exigencia de la educación en línea.

Sea como fuere, el estudiantado sigue abandonando. El profesorado necesita herramientas para su identificación temprana y disponer de mecanismos de intervención para poder ayudar al estudiantado antes de que el abandono se materialice. El desarrollo de modelos de predicción basados en algoritmos de aprendizaje automático ofrece un gran potencial para satisfacer dichos objetivos [8]. En [6] presentamos un sistema de detección capaz de detectar de forma temprana al estudiantado en riesgo de suspender una asignatura. El modelo tomaba en cuenta las calificaciones obtenidas en las actividades de evaluación continua propuestas en la asignatura. Dicho modelo evolucionó (véase [2]) añadiendo datos relativos al perfil del estudiantado, permitiendo una mejor caracterización de este. Sin embargo, el modelo de predicción presenta una limitación principal. Si bien el modelo es capaz de identificar el estudiantado que abandona, la identificación se efectúa una vez que dicho abandono se ha concretado (el estudiantado ha abandonado y por eso no entrega las actividades de evaluación). En este escenario, las medidas de intervención [12] llegan tarde y suelen ser ineficaces.

Precisamente, el objetivo de este trabajo es, en primer lugar, presentar un modelo de predicción específicamente orientado a la identificación temprana del estudiantado en riesgo de abandonar en una asignatura, el mecanismo de intervención asociado y los cuadros de mando del profesorado. Este modelo de predicción no sustituye al modelo presentado en [2] sino que lo complementa. En segundo lugar, se describe la prueba piloto realizada en una asignatura obligatoria de primer año del Grado de Ingeniería Informática de la Universitat Oberta de Catalunya (UOC), y los resultados obtenidos.

El trabajo se estructura de la siguiente manera: la Sección 2 analiza el trabajo relacionado, mientras que la Sección 3 describe el sistema de detección de estudiantado en riesgo. La Sección 4 presenta la prueba piloto y la discusión se realiza en la Sección 5. Finalmente, las conclusiones y líneas de trabajo futuro se incluyen en la Sección 6.

2. Trabajo relacionado

La mayoría de los trabajos (véase, por ejemplo, [9]) que proponen modelos de predicción específicamente orientados a la detección de abandono se centran en MOOC (*Massive Open Online Courses*) debido a la baja retención. Muchos de estos modelos tienen una precisión cercana al 90% [13] desde la mitad del curso, porque disponen de conjuntos de datos detallados dado que todo el proceso de enseñanza-aprendizaje (incluyendo la evaluación) tiene lugar exclusivamente en la propia plataforma. En general, estos modelos realizan una predicción semanal para identificar el estudiantado en riesgo de

abandono [3, 8, 10, 13]. Esta elección se basa en que la mayoría de las actividades de evaluación se proponen sobre una base semanal. Sin embargo, al examinar los conjuntos de datos, se puede observar que las clases están claramente desbalanceadas hacia el estudiantado que abandona, siendo más fácil su detección. También causan una cantidad no despreciable de falsos positivos (estudiantado identificado como abandono, y que realmente completa el curso).

En entornos de enseñanza universitaria en línea e híbridos el volumen y detalle de datos disponibles suele ser menor en comparación a los MOOC. El estudiantado puede estudiar al margen del sistema gestor de aprendizaje (o campus virtual), acceder a contenidos sin pasar por el aula virtual, comunicarse a través diferentes canales, e incluso realizar sus actividades con un acceso mínimo al campus y aula virtual. Además, cada asignatura tiene su propia propuesta de actividades, de diferente complejidad y planificación. Es decir, no necesariamente se entregan actividades semanalmente. Finalmente, la entrega de las actividades que conforman la propuesta de evaluación continua de la asignatura es buen indicador para la superación de esta [11, 15].

Precisamente, en este trabajo definimos el estudiantado en riesgo de abandonar como “aquel que no entregará la actividad en curso”, ya que el estudiantado que no entrega las actividades tiende a abandonar la asignatura. Además, proponemos un modelo de predicción capaz de hacer una predicción diaria que se combina con una ventana de identificación (medida en días) para realizar intervenciones orientadas a mitigar el abandono. Esta ventana se define como “el número de días consecutivos en que el estudiantado es marcado por el modelo en riesgo para considerarlo un abandono potencial”. Es especificada por el profesorado y, al ser variable, es capaz de adaptarse a la complejidad y duración de cada actividad, con el objetivo de intentar maximizar la detección de abandono y minimizar los casos incorrectos.

3. Descripción del sistema

3.1. Modelo de predicción

El modelo de predicción de riesgo de abandono, denominado modelo PDAR (*Profiled Dropout At-Risk model*), consta de una serie de submodelos. En concreto, hay un submodelo para cada día del semestre. El modelo es propio de cada asignatura ya que se entrena únicamente con datos de esta, y dado que las actividades de evaluación de la asignatura tienen una planificación temporal, el modelo es capaz de adaptarse a ellas. En consecuencia, el resultado de la predicción de un submodelo, para un día y estudiante concretos, es si este entregará o no la actividad en

Característica	Conjunto de datos	Tipo
Nuevo en la universidad	Perfil estudiante (Institucional)	Binario
Nota promedio de expediente (GPA)	Perfil estudiante (Institucional)	Real
Número de asignaturas matriculadas en el semestre	Matrícula (Semestre)	Entero
Número de repeticiones de la asignatura objetivo	Matrícula (Semestre)	Entero
Actividad enviada	Rendimiento (Actividad)	Binario
Nota de la actividad	Rendimiento (Actividad)	Entero
Número de días entre el último acceso al campus virtual y la fecha de entrega de la actividad	Clickstream (Diario)	Entero
Número de días entre el último acceso al aula virtual y la fecha de entrega de la actividad	Clickstream (Diario)	Entero
Número promedio de mensajes leídos por el estudiante del tablón del profesor respecto el total de mensajes leídos por los estudiantes matriculados	Clickstream (Diario)	Real
Número promedio de mensajes leídos por el estudiante del foro respecto el total de mensajes leídos por los estudiantes matriculados	Clickstream (Diario)	Real
Suma de días en que se accede a los materiales didácticos	Clickstream (Diario)	Entero
Suma de días en que se accede a otras herramientas de aprendizaje	Clickstream (Diario)	Entero

Cuadro 1. Características usadas en el modelo PDAR

curso. Es decir, el resultado es binario: no entregará (1) o entregará (0).

Las características usadas por el modelo PDAR se muestran en el Cuadro 1. Se derivan de diferentes conjuntos de datos anonimizados disponibles en un *data mart* institucional que almacena datos verificados desde el curso académico 2016-2017 acerca del funcionamiento de todas las asignaturas y del propio estudiantado, y que se actualiza diariamente. Los conjuntos de datos elegidos para el modelo, en esencia, intentan caracterizar el estudiantado, y modelar su progresión y nivel de implicación en la asignatura. En primer lugar, se utilizan datos del perfil del estudiantado. En segundo lugar, el modelo toma en cuenta datos relativos a la matrícula del estudiantado en el semestre actual. En tercer lugar, se usan datos sobre el rendimiento en la asignatura donde se aplica el modelo. Finalmente, se tiene en cuenta el flujo de datos diario (*clickstream*) generado por el estudiantado en la asignatura (dentro del campus propio de la UOC y el aula virtual).

Con relación al *clickstream* se destacan diversos aspectos. Primero, se considera un intervalo acumulado que incluye los últimos siete días. Segundo, los datos de acceso se agregan dentro de dicho intervalo. En concreto, para el acceso a campus y el aula virtual, la agregación considera el período que va desde la fecha de entrega de la actividad y la última fecha de acceso del estudiantado. De forma similar, las accio-

nes del estudiantado en los espacios de comunicación del aula virtual (tablón y foro) se calculan proporcionalmente a las acciones del resto de participantes de la asignatura. Finalmente, el uso de materiales didácticos y otras herramientas (por ejemplo, tutores inteligentes) se cuenta como el número de días accedidos dentro del intervalo acumulado.

3.2. Entrenamiento

Los submodelos del modelo PDAR se entrenan usando diferentes algoritmos de clasificación con los datos históricos almacenados en el *data mart* institucional. En concreto, se usan los datos que van desde el primer semestre del curso 2016-2017 hasta el penúltimo semestre en que se quiere usar el modelo. Los algoritmos de clasificación utilizados son Naive Bayes (NB), CART Decision Tree (DT), K-Nearest Neighbors (KNN) y Support Vector Machine (SVM). Específicamente, se entrena un submodelo para cada día en que la actividad de evaluación está vigente. Cuando las actividades se solapan en el tiempo, se ordenan por fecha de entrega, y se establece como primer día de una actividad el día siguiente en que finaliza la anterior. Esta decisión permite disponer de un submodelo por día que, además, se centra en la próxima actividad a entregar.

Los submodelos tienen una precisión determinada. En nuestro caso, se usan las siguientes métricas:

$$TNR = \frac{TN}{TN+FP}$$

$$TPR = \frac{TP}{TP+FN}$$

donde TP (*true positive*) denota el número de estudiantes en riesgo correctamente identificados, TN (*true negative*) el número de estudiantes que no están en riesgo correctamente identificados, FP (*false positive*) el número de estudiantes que no están en riesgo incorrectamente identificados, y FN (*false negative*) el número de estudiantes en riesgo incorrectamente identificados, y que se utilizan para evaluar la precisión en la detección de estudiantado en riesgo (TPR, *true positive rate*), y la precisión en la detección del estudiantado que no está en riesgo (TNR, *true negative rate*). Aunque se calculan métricas adicionales como serían la precisión global del modelo, el F-score o el área bajo la curva (AUC), por su relevancia para este trabajo y por razones de espacio, únicamente presentaremos el TPR y TNR.

Los submodelos entrenados se validan con los datos del semestre anterior al semestre en que se pretende usar el modelo, y, para cada submodelo, el algoritmo de clasificación con la mejor suma entre el TPR y TNR se selecciona como el algoritmo que se utilizará para generar las predicciones. Maximizar esta suma tiende a proporcionar la mejor estrategia para detectar todo tipo de estudiantado, mientras que la maximización de una de las dos métricas tiende a penalizar la descartada y, por lo tanto, una de las dos

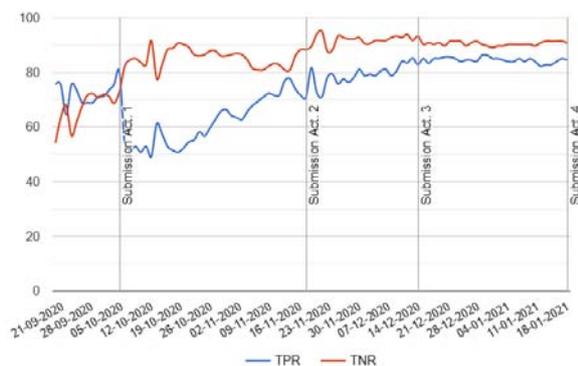


Figura 1: Precisión de los submodelos PDAR en la asignatura piloto.

posibles predicciones (no entregará, entregará). Nótese que cada submodelo tendrá una precisión (TPR) diferente, y que algún día puede tener una precisión baja. Tras diversos análisis, se ha establecido un umbral mínimo del TPR del 65% para considerar que la predicción es suficientemente precisa en la identificación del estudiantado en riesgo para ese día.

3.3. Ventana de identificación y mecanismo de intervención

El hecho de que el modelo de predicción identifique a un estudiante concreto en riesgo de no entregar la actividad en curso para un día determinado, no significa que realmente lo esté. El estudiante puede estar varios días sin acceder al aula virtual (e incluso al campus), y continuar trabajando en la asignatura según el tipo de actividad que deba realizar. Por esa razón, el profesorado (en base a su experiencia) debe indicar para cada actividad el número de días consecutivos (o ventana de identificación) en que el estudiantado se debe identificar en riesgo para ser considerado como tal, e iniciar intervenciones.

El mecanismo de intervención diseñado permite el envío automático de mensajes al estudiantado. Se tiene en cuenta para cada actividad la ventana de identificación especificada por el profesorado para ser considerado en situación de riesgo, y el envío se hace con suficiente antelación a la entrega de la actividad para enmendar la situación. El profesorado diseña los mensajes, que tienen como objetivo posibilitar un asesoramiento y un seguimiento individualizado. Solo se envía un mensaje por actividad.

Se pueden enviar tres tipos de mensajes. El primero se dirige al estudiantado que, si bien está activo en la asignatura (registra actividad y ha entregado actividades previas), se ha identificado en riesgo de abandono. El segundo tipo de mensaje se orienta al estudiantado que no ha entregado la actividad de evaluación previa, y que se identifica en riesgo de abandono. Finalmente, el tercer tipo de mensaje se dirige al estudiantado que se identifica en riesgo de abandono y no ha entregado al menos dos actividades.

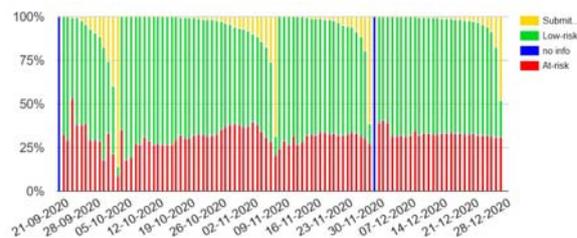


Figura 2: Agregación de las predicciones diarias del modelo PDAR para una asignatura.

El mecanismo de intervención complementa al mecanismo de intervención del modelo de predicción presentado en [2], dando apoyo tanto al estudiantado en riesgo de abandono como al estudiantado en riesgo de no superar la asignatura. La descripción de los tipos de mensajes que se pueden enviar y la evaluación de su impacto se encuentra disponible en [12].

3.4. Cuadros de mando del profesorado

El sistema de detección temprana ofrece diversos cuadros de mando para supervisar el modelo PDAR y la ventana de identificación. Respecto al modelo PDAR, ofrece un gráfico de la precisión de las métricas TPR y TNR de los submodelos de la asignatura. La precisión para la asignatura en la que se efectuó la prueba piloto (y que describiremos en la Sección 4) se puede ver en la Figura 1. Esta asignatura consta de 4 actividades de evaluación. Se puede observar que la precisión mejora a medida que el curso avanza con cierta diferencia entre la identificación entre el estudiantado en riesgo y el que no lo está.

Una vez entrenado cada submodelo e iniciado el curso, el sistema muestra la predicción de identificación del estudiantado en riesgo. Primero, el profesorado puede comprobar el porcentaje de identificación diaria del modelo PDAR. La Figura 2 muestra el resultado final para la misma asignatura en donde se puede ver el estudiantado identificado en riesgo (en rojo), sin riesgo (en verde), y el que va entregando la actividad (en dorado). En azul se marcan los días para los que no se dispone de información debido a algún problema técnico en el volcado de datos al *data mart*.

Además de la interfaz de configuración del mecanismo de intervención, la edición de los mensajes y el control de los mensajes enviados, el profesorado puede analizar la evolución individual de cada estudiante. La Figura 3 muestra un ejemplo de la identificación de los dos tipos de riesgo del sistema (no superar la asignatura y abandono). Por un lado, y en forma de color de semáforo (verde: nivel bajo, ámbar: nivel intermedio, rojo: nivel alto, y negro: estudiante que no ha presentado la actividad) se muestra el nivel de riesgo de no superar la asignatura. Por otro lado, y en forma de mini gráfico (*sparkline*) se puede revisar la identificación diaria del riesgo de abandono mediante los submodelos PDAR. Los colores utilizados,

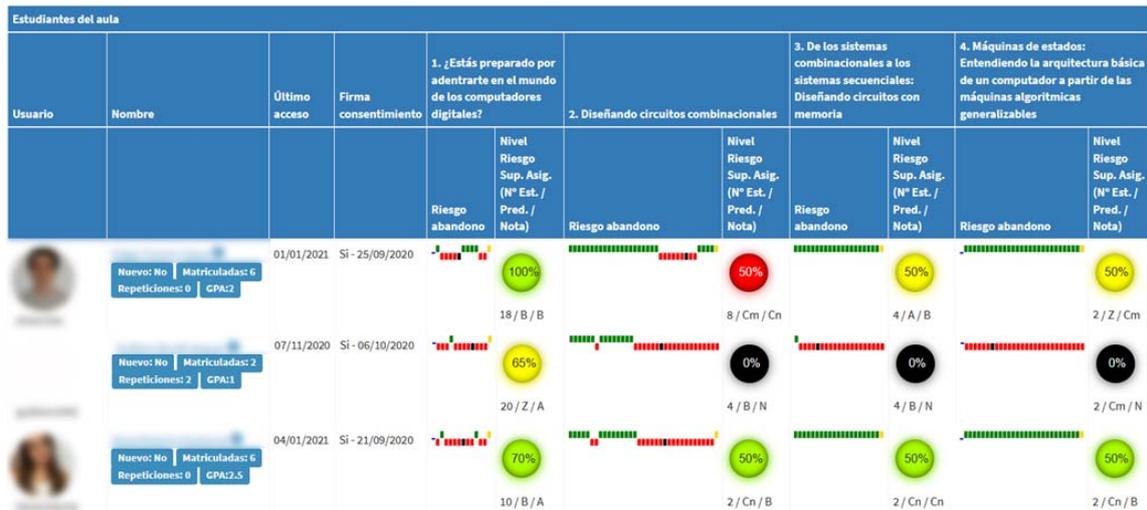


Figura 3: Evolución del nivel de riesgo individualizada por estudiante en una asignatura.

en esencia, tienen significado idéntico al otro tipo de riesgo. El verde muestra un nivel de riesgo bajo de abandono, el rojo un nivel alto y el negro identifica el día en que se realiza una intervención en forma de mensaje después de pasar los días consecutivos de la ventana de identificación especificados por el profesorado en que el estudiante es identificado en nivel de riesgo de abandono alto. Finalmente, el dorado muestra que el estudiante ha entregado la actividad. Nótese que la visualización también muestra información relativa al perfil del estudiante tal y como se ha especificado en el Cuadro 1 (si es nuevo en la universidad, número de asignaturas matriculadas, número de veces que ha repetido la asignatura, y la nota promedio de su expediente académico). Nótese que únicamente aparece la información del estudiantado que firma un consentimiento para el uso de sus datos.

Concretamente, en la Figura 3 podemos ver tres estudiantes. Con relación al nivel de riesgo de suspender, el primer estudiante pasa por diferentes niveles de riesgo, pero acaba superando la asignatura. El segundo supera la primera actividad, pero acaba suspendiendo por abandono. La tercera estudiante es un claro ejemplo de una estudiante sin nivel de riesgo. Nótese que estudiantado con nota idéntica para la actividad puede no tener el mismo nivel de riesgo debido a diferencias en su perfil. También se indica, para cada estudiante, el porcentaje de estudiantado con el mismo perfil que han superado la asignatura en el pasado (para obtener información más detallada, véase [2]). Respecto al nivel de riesgo de abandono, para el primer estudiante, se envían mensajes de intervención en la primera y segunda actividad, aunque finalmente entrega todas las actividades. El segundo estudiante constituye un claro caso de abandono después de la primera actividad de evaluación. En el caso de la tercera estudiante, vemos un ejemplo de una estudiante que durante la primera y segunda

actividad estudia mayormente desconectada del aula virtual, siendo identificada erróneamente como posible abandono.

4. Prueba piloto

4.1. Contexto

El modelo PDAR y el mecanismo de intervención han sido probados en la asignatura de *Fundamentos de computadores* (FC). Se trata de una asignatura de primer año de 6 créditos ECTS del Grado de Ingeniería Informática en donde el estudiantado aprende a analizar y sintetizar circuitos digitales, y la arquitectura subyacente. El modelo de evaluación continua incluye tres actividades de evaluación continua (PEC1, PEC2 y PEC3) y una práctica (Pr). Un examen complementa la evaluación continua al final del semestre. Es una asignatura adecuada para analizar porque tiene un rendimiento académico bajo (entre el 40% y 50% sobre el número de estudiantes matriculados), principalmente por abandono, compartiendo la problemática expuesta en [11, 15]. Aunque se puede aprobar sin realizar alguna de las PEC (la Pr y el examen son obligatorios), el profesorado sabe que el estudiantado que no las realiza tiene dificultades.

El piloto tuvo lugar en el primer semestre del curso académico 2020-2021 en donde se matricularon 380 estudiantes. El Comité de ética de la universidad aprobó el estudio. La participación fue voluntaria y los participantes (259 estudiantes, el 68% de los matriculados) firmaron un consentimiento informado.

4.2. Precisión de la ventana de identificación

Para comprobar la efectividad de la ventana de identificación, se ha entrenado el modelo PDAR con los datos históricos (anonimizados) del estudiantado

	Vent.	TP	FP	TN	FN	TPR	TNR
PEC1	5	36	117	320	16	69,23%	73,23%
PEC2	7	115	87	270	17	87,12%	75,63%
PEC3	5	136	29	295	29	82,42%	91,05%
Pr	7	156	11	279	43	78,39%	96,21%

Cuadro 2. Precisión de la ventana de identificación en FC.

de la asignatura, y disponibles en la *data mart* desde el semestre de otoño de 2017 hasta el semestre de otoño de 2019. La validación se ha efectuado con los datos del estudiantado del semestre de primavera de 2020. El objetivo es comprobar el porcentaje de acierto de la ventana en la identificación de estudiantado en riesgo de abandono (estudiantado que no entrega la actividad en curso, de acuerdo con nuestra definición de abandono incluida en la Sección 2).

En el Cuadro 2 podemos ver el resumen de la precisión. Para cada actividad, se muestra la ventana de identificación (medida en días) definida por el profesor (Vent.) y se incluyen las dos métricas (TPR y TNR). Podemos ver que el sistema tiene dificultades para detectar abandono en la primera actividad (PEC1). Al usar únicamente datos de perfil del estudiantado y los accesos al entorno virtual (*clickstream*), el modelo tiene dificultades para predecir el posible abandono. La precisión mejora a medida que evoluciona el curso, debido a la adición de los datos de rendimiento de las actividades ya evaluadas. Además, el sistema identifica más fácilmente el estudiantado que no está en riesgo ya que tiende a comportarse de forma similar cada semestre (entrega las actividades, supera las actividades y accede regularmente al aula virtual). La precisión en todos los casos es superior al 70% (excepto la PEC1). Podemos ver un pequeño decremento de la métrica TPR para la Pr. La razón es porque la asignatura se puede aprobar únicamente superando la Pr y el examen final. Aunque el estudiantado, por lo general, no se decanta por esta opción, existen algunos casos, fallando el sistema de predicción. Teniendo en cuenta la dificultad de modelar el comportamiento de abandono del estudiantado, la precisión es más que aceptable.

4.3. Efectividad del mecanismo de intervención

Respecto al estudiantado que participó en la prueba piloto, se ha comprobado la tasa de acierto del mecanismo de intervención. Un sistema de detección temprana para ser efectivo debe enviar los mensajes al público objetivo. En el Cuadro 3 se pueden ver los resultados. En concreto se muestra, para cada actividad y nivel de riesgo, el número de estudiantes (N.), y el porcentaje de estos que no han presentado la actividad (Abandono) y el que sí (Presentado). Con relación a los mensajes, se muestra el nivel de riesgo bajo (V) que representa el estudiantado que no recibe ningún mensaje con respecto al nivel de riesgo de abandono; el nivel alto (R) que sí recibe un mensaje por estar en nivel de riesgo a pesar de haber presentado la actividad anterior; el nivel alto para el estudiantado que ya no ha presentado la actividad anterior (NP); y finalmente el estudiantado en nivel alto que no ha presentado las dos actividades anteriores (NP2).

Se puede observar que se detecta correctamente el estudiantado con nivel de riesgo bajo (V) con una precisión superior al 75%. El estudiantado con alguna actividad anterior no presentada, también se detecta correctamente. La razón es porque, en estos casos, el abandono ya se ha materializado. Solo se aprecia un caso en la Pr y el nivel de riesgo NP, donde un estudiante que no había presentado la PEC3, el sistema falla ya que acaba presentando la práctica. Como resultado positivo, no atribuible únicamente al mecanismo de intervención, es que no todo el estudiantado en nivel alto (R) se materializa como abandono en las siguientes actividades (mensajes NP y NP2).

Finalmente, el nivel de riesgo relevante para este trabajo es el de riesgo alto: el estudiantado que continúa activo (R) es detectado razonablemente bien por el sistema, con una precisión superior al 70%.

5. Discusión

Los sistemas de detección temprana pueden ser una herramienta fundamental para ayudar al profesorado en su docencia, tanto en entornos en línea como en entornos híbridos. El único requerimiento es que se disponga de datos del estudiantado en formato digital,

NR	PEC1			PEC2			PEC3			Pr		
	N.	Abandono	Presentado	N.	Abandono	Presentado	N.	Abandono	Presentado	N.	Abandono	Presentado
V	252	21,43%	78,57%	227	20,26%	79,74%	207	4,83%	95,17%	194	3,61%	96,39%
R	7	71,43%	28,57%	26	73,08%	26,92%	23	76,19%	23,81%	20	85,00%	15,00%
NP	--	--	--	6	100,00%	0,00%	23	100,00%	0,00%	18	94,44%	5,56%
NP2	--	--	--	--	--	--	6	100,00%	0,00%	27	100,00%	0,00%

Cuadro 3. Tasa de acierto del riesgo de abandono respecto presentar la actividad correspondiente en FC.

tanto del rendimiento académico como de su progreso dentro de la universidad. Se puede pensar que únicamente es factible en entornos totalmente virtuales, pero hoy en día todas las universidades disponen de plataformas de aprendizaje en línea que dan apoyo a la docencia presencial, tal y como se ha demostrado durante la situación derivada de la COVID. El desarrollo de sistemas de alerta temprana se ve facilitado si existe una política de gestión de los datos definida e impulsada por la institución.

¿Qué ventajas nos ofrecen estos sistemas de detección temprana? Sobre todo “conocer” mejor al estudiantado. Cuando tenemos un número muy alto de estudiantes en un aula (o asignatura) es muy complicado saber el conocimiento que han adquirido o sus necesidades concretas. Se podrá tener un mejor conocimiento del estudiantado que se comunica más frecuentemente con el profesorado o después de la evaluación de alguna actividad, pero el estudiantado que no interactúa o que estudia de forma más individual y autónoma es “invisible”. En el caso de las actividades, el estudiantado podrá recibir retroalimentación personalizada después de su evaluación, pero esta retroalimentación puede llegar demasiado tarde.

Un sistema de detección temprana puede identificar los problemas del estudiantado con un margen de tiempo suficiente para intervenir. Cuando se intenta detectar el riesgo de suspender (véase [2]), los beneficios son significativos, pero cuando se intenta detectar el abandono como en este trabajo, los beneficios pueden ser aún mayores. En actividades relativamente largas (dos semanas o más), el estudiantado puede tener diversos problemas que pueden impactar en su realización, de forma que el abandono se materializa en el momento de las entregas. Además, un porcentaje no despreciable del estudiantado puede percibir que no entregar una actividad significa perder la posibilidad de superar la asignatura. Por lo tanto, es importante supervisar (y predecir) la entrega de las actividades para ayudar al estudiantado.

Los factores que impactan en el abandono en el caso de este trabajo (entorno en línea) son los mismos que han sido detectados en otros trabajos previos [4, 7, 11]. El estudiantado novel suele tener problemas en el nuevo entorno de aprendizaje. Estos problemas muchas veces se combinan con el acceder a unos estudios universitarios y a un nuevo modelo de estudio y/o evaluación. Hay que destacar que en este estudio no se han podido tener en cuenta datos socioeconómicos o de estudios previos del estudiantado, ya que el Comité de ética de la universidad no permite acceder a estos datos por considerarlos sensibles.

A pesar de ello, se ha comprobado que los datos que impactan más en el abandono son el rendimiento en la asignatura y el nivel de implicación del estudiantado (accesos al entorno virtual). Por esta razón este estudio ha trabajado la intervención durante la

realización de las actividades, intentando impactar en la decisión del estudiantado de completar la actividad en curso, de continuar en la asignatura, y que el estudiantado perciba una sensibilidad y una actitud proactiva por parte del profesorado sobre los problemas potenciales que pueda tener. El hecho de realizar estas intervenciones ha podido afectar al abandono global de la asignatura que se redujo del 41,7% del semestre anterior al 37,8%.

La valoración por parte del estudiantado de los mensajes enviados por el mecanismo de intervención ha sido muy positiva, aunque no tenemos un estudio completo de los mismos. Un porcentaje muy elevado agradece la preocupación del profesorado, sobre todo en el caso de estudiantado novel. En otros casos, se aprovecha su recepción para iniciar una conversación con el profesorado sobre alguna duda relacionada con la evaluación, la actividad o los contenidos. También hemos observado que, en el caso de los mensajes relacionados con el nivel de riesgo NP o NP2 (estudiantado que ya no había presentado alguna actividad previa), el estudiantado tiende a disculparse o explicar la razón del abandono. No se ha recibido ningún mensaje negativo, seguramente debido a que la participación en el piloto era voluntaria.

Sobre la precisión del modelo PDAR, la agregación de datos de *clickstream* tiene en cuenta un período de siete días. Aunque se podría tomar un intervalo más amplio (por ejemplo, acumular todos los días desde el inicio de la actividad o incluso el semestre), diferentes análisis arrojaron como resultado que intervalos más amplios impactaron negativamente en el estudiantado que empezó más tarde el curso, o en el estudiantado que no accedió al aula virtual durante algún período, generando falsos positivos (recuérdese que el estudiantado puede realizar parte de sus actividades sin acceder al campus y aula virtual). En definitiva, se trata de datos sujetos a una alta variabilidad.

Las limitaciones de nuestro trabajo son principalmente tres. Primero, en términos de validez interna, el sesgo de autoselección. Es el estudiantado más comprometido el que tiende a participar, y suelen registrar tasas de abandono bajas. Segundo, la ventana de identificación en cada actividad es propuesta por el profesorado y, en consecuencia, puede estar sujeta a errores de apreciación. Tercero, y a pesar de la reducción del abandono en el piloto, no se ha analizado en profundidad el impacto del modelo PDAR en el abandono de la asignatura debido a la situación derivada de la COVID. La situación excepcional vivida (más tiempo de estudio y evaluación final no presencial) hace difícil extraer conclusiones fiables.

6. Conclusiones y trabajo futuro

Nuestro trabajo pone de manifiesto el potencial de la minería de datos educativos para extraer informa-

ción relevante para mejorar el proceso de enseñanza-aprendizaje. Aunque la experiencia se centra en una universidad en línea, esta puede ser de utilidad para otras universidades, dado que actualmente todas incorporan plataformas de aprendizaje en línea.

El modelo PDAR presenta dos características distintivas respecto trabajos previos. Primero, la identificación es en tiempo real al proporcionar una predicción diaria del riesgo. Segundo, la ventana de identificación especificada por el profesorado permite que el modelo sea capaz de adaptarse a las características de las actividades. El piloto muestra la efectividad del modelo PDAR y del mecanismo de intervención.

Como trabajo futuro se diseñará un mecanismo automático de cálculo de la ventana de identificación para cada actividad, a ser validada por el profesorado ya que puede estar sujeta a errores de apreciación. También se estudiará el impacto del modelo en las tasas de abandono, así como el análisis de utilidad percibida del sistema por parte del profesorado y la mejora del mecanismo de intervención.

Agradecimientos

Este trabajo ha sido financiado por el proyecto New Goals 2018NG001 “LIS: Learning Intelligent System” del eLearning Innovation Center (UOC).

Referencias

- [1] Othman Aljohani. A Comprehensive Review of the Major Studies and Theoretical Models of Student Retention in Higher Education, *High. Educ. Stud.*, 6(2):1–18, 2016.
- [2] David Baneres, M. Elena Rodríguez, Ana-Elena Guerrero-Roldán y Abdulkadir Karadeniz. An Early Warning System to Detect At-Risk Students in Online Higher Education. *Appl. Sci.*, 10(13):4427, 2020.
- [3] Mi Fei y Dit-Yan Yeung. Temporal Models for Predicting Student Dropout in Massive Open Online Courses. En *Proceedings - 15th IEEE International Conference on Data Mining Workshop*, pp. 256–263, noviembre 2015.
- [4] Alfonsa García, Ana Lías, Ángeles Mahillo y Rosa M.^a Pinero, “Abandono de primer año en la Ingeniería Informática”, En *Actas de las XX Jornadas de Enseñanza Universitaria de la Informática, JENUI 2014*, pp. 151-158, Oviedo, 2014.
- [5] Steven J. Greenland y Catherine Moore. Large qualitative sample and thematic analysis to re-define student dropout and retention strategy in open online education, *Br. J. Educ. Technol.*, 0(0):1–21, 2021.
- [6] Ana-Elena Guerrero-Roldán, M. Elena Rodríguez, David Bañeres, Cristina Pérez, Javier Panadero y Abdulkadir Karadeniz. Hacia un sistema de detección temprana de estudiantes en riesgo en entornos de enseñanza-aprendizaje en línea. En *Actas de las XXVI Jornadas de Enseñanza Universitaria de la Informática, JENUI 2020*, pp. 37–44, Valencia, julio 2020.
- [7] Carmen Lacave, Ana I. Molina, Miguel A. Redondo y Manuel Ortega. Redes bayesianas para identificar perfiles de estudiante. Aplicación al estudio del abandono de las titulaciones de informática en la Universidad de Castilla-La Mancha. En *Actas de las XXII Jornadas de Enseñanza Universitaria de Informática, JENUI 2016*, pp. 85–92, Almería, julio 2016.
- [8] Hui Luan y Chin-Chung Tsai. A Review of Using Machine Learning Approaches for Precision Education. *Educ. Technol. Soc.*, 24(1):250–266, 2021.
- [9] Pedro M. Moreno-Marcos, Carlos Alario-Hoyos, Pedro J. Muñoz-Merino y Carlos Delgado-Kloos. Prediction in MOOCs: A Review and Future Research Directions. En *IEEE Trans. Learn. Technol.*, 12(3):384–401, 2019.
- [10] Ahmed A. Mubarak, Han Cao e Ibrahim M. Hezam. Deep analytic model for student dropout prediction in massive open online courses. *Comput. Electr. Eng.*, vol. 93(C):1–14, 2021.
- [11] M. Elena Rodríguez, Ana-Elena Guerrero-Roldán, David Baneres y Abdulkadir Karadeniz. Towards an intervention mechanism for supporting learners performance in online learning. En *ICERI2019 Proceedings*, pp. 5136–5145, noviembre 2019.
- [12] M. Elena Rodríguez, Ana-Elena Guerrero-Roldán, David Baneres y Abdulkadir Karadeniz. An Intelligent Nudging System to Guide Online Learners. *Int. Rev. Res. Open Distrib. Learn.*, 23(1):41-62, 2022.
- [13] Jacob Whitehill, Kiran Mohan, Daniel Seaton, Yigal Rosen y Dustin Tingley. MOOC dropout prediction: How to measure accuracy? En *Proceedings of the 4th ACM Conference on Learning at Scale*, pp. 161–164, abril 2017.
- [14] Marlon Xavier y Julio Meneses. A Literature Review on the Definitions of Dropout in Online Higher Education. En *EDEN Conference Proceedings*, pp. 73–80, junio 2020.
- [15] Marlon Xavier y Julio Meneses. The Tensions Between Student Dropout and Flexibility in Learning Design: The Voices of Professors in Open Online Higher Education. *Int. Rev. Res. Open Distrib. Learn.*, 22(4):72–88, noviembre 2021