

Predicción de pérdida de implicación de los participantes de un curso en línea masivo y abierto

Miguel L. Bote Lorenzo, Eduardo Gómez Sánchez

Grupo de Sistemas Inteligentes y Cooperativos / Educación, Medios, Informática y Cultura
Escuela Técnica Superior de Ingenieros de Telecomunicación. Universidad de Valladolid.
Paseo de Belén 15, 47011, Valladolid, España
migbot@tel.uva.es, edugom@tel.uva.es

Resumen—En este artículo se propone llevar a cabo la predicción de pérdida de implicación de los participantes de un curso en línea masivo y abierto a partir de dos indicadores, uno basado en el consumo de vídeos y otro en la realización de pruebas de evaluación. De esta forma resulta posible intervenir antes de que de que la implicación decaiga, de manera que los participantes continúen aprovechando el curso desde un punto de vista educativo. Los resultados muestran que es posible construir predictores con un área bajo la curva comprendida entre 0,847 y 0,909 para detectar la pérdida de implicación a partir del consumo de vídeos, y de entre 0,725 y 0,915 a partir de la realización de pruebas de evaluación.

Palabras clave—MOOC; pérdida de implicación; aprendizaje automático supervisado

I. INTRODUCCIÓN

Los Cursos en Línea Masivos y Abiertos (*Massive Open Online Courses, MOOCs*) [1] permiten obtener conocimientos sobre una determinada materia a grupos de alumnos con tamaños potencialmente muy grandes y heterogéneos tanto desde el punto de vista del origen de los participantes como del de su bagaje académico. Estos cursos se ofrecen en plataformas en línea como Coursera¹, edX² o FutureLearn³ que, generalmente, no requieren ningún pago ni establecen ningún requisito formal para la participación, como indica el término “abierto” de su denominación.

Algunos trabajos de la literatura [2]-[6] han estudiado la evolución a lo largo del curso de la implicación de los participantes de los MOOC en la realización de distintas actividades como el consumo de vídeos explicativos, la realización de pruebas de evaluación o la participación en los foros de debate. En estos trabajos se muestra la existencia en los MOOC de numerosos participantes cuya implicación disminuye a lo largo del tiempo, llegando en ocasiones al abandono completo del curso. Independientemente de que se produzca o no el abandono, esta disminución de la implicación da lugar un menor aprovechamiento del curso por parte de

esos participantes. De este modo, varios autores [3], [4] defienden la necesidad de mantener la implicación de los participantes para que los MOOC puedan tener un amplio impacto educativo.

Una posible aproximación para intentar sostener la implicación de los participantes de un MOOC consiste en detectar a aquellos alumnos cuya implicación se puede prever que disminuirá en un futuro próximo de tal forma que sea posible hacer una intervención antes de que dicha disminución se produzca. La predicción podría hacerse a partir de las trazas que deja la actividad de los participantes en el MOOC. La intervención podría consistir, por ejemplo, en la sugerencia de algún vídeo que pueda resultar interesante al alumno y aumentar así sus ganas de ver otros vídeos, o en proporcionarle pistas que le ayuden a resolver algún ejercicio que el alumno haya intentado pero no haya conseguido solucionar, de forma que pueda verse más motivado para continuar haciendo ejercicios.

En este artículo se propone llevar a cabo la predicción de pérdida de implicación de los participantes utilizando dos indicadores de implicación: el consumo de vídeos y la realización de pruebas de evaluación. También se comparan distintas alternativas para realizar dicha predicción en el contexto de un MOOC impartido en la plataforma edX.

La estructura del resto del documento es la siguiente. En primer lugar se discuten los trabajos que es posible encontrar en la literatura en relación con la propuesta que se hace en este artículo. A continuación se describen brevemente el MOOC que se utiliza como punto de partida de este trabajo así como los datos disponibles de este curso. En la siguiente sección se introduce la aproximación propuesta para llevar a cabo la predicción de pérdida de implicación en el contexto de ese MOOC. Después se presentan las características extraídas de las trazas para llevar a cabo la predicción así como los distintos métodos de reducción del número de características y de aprendizaje supervisado que han sido considerados. Seguidamente se presentan y discuten los resultados de los experimentos realizados. En la última sección se recogen las principales conclusiones y el trabajo futuro.

¹ <https://www.coursera.org/>

² <https://www.edx.org/>

³ <https://www.futurelearn.com/>

II. TRABAJO RELACIONADO

Los autores de este artículo no han conseguido encontrar en la literatura trabajos que traten la predicción de pérdida de implicación de los participantes de MOOC. Sí que han encontrado trabajos que se centran en la predicción de la falta de implicación en sistemas de tutores inteligentes [7], [8] o sistemas de gestión de contenidos educativos [9]. Sin embargo, ese problema de predicción es diferente al tratado en este artículo: en este último caso lo que se busca es saber si el estudiante se encontrará en el futuro en un estado de falta de implicación o en otro estado, mientras que en la predicción de pérdida de implicación se pretende determinar si bajará o no el nivel de implicación del participante respecto al que ha mantenido hasta el momento independientemente del estado en que se encuentre. Además, tanto los sistemas de tutores inteligentes como los de gestión de contenidos educativos permiten un abanico de actividades diferente al de las plataformas de MOOC, por lo que los factores que se deben considerar para determinar la implicación son distintos y las trazas de actividad disponibles para hacer la predicción también son diferentes.

En otros trabajos de la literatura [10]-[14] se estudia la posibilidad de predecir si un participante va a abandonar un MOOC con la antelación suficiente como para que sea posible intervenir de alguna forma para que dicho abandono no tenga lugar. También existen trabajos [4], [12], [15], [16] que buscan predecir si un participante se encuentra en riesgo de no superar un MOOC de forma que, de nuevo, cabe la posibilidad de intervenir para intentar evitarlo. El objetivo último de estos tipos de predicción es por tanto diferente al de la predicción de pérdida de implicación que se estudia en este artículo, la cual pretende facilitar la posibilidad de intervenir de cara a conseguir que los participantes del MOOC mantengan su nivel de implicación. Con ello se pretende no sólo que haya más alumnos que superen el curso o no lo abandonen, sino que haya un mayor número de alumnos que aprovechen el curso incluso si su objetivo no es superarlo.

III. CONTEXTO DEL ESTUDIO

Este trabajo se centra en el MOOC “6.002x *Circuits and Electronics*” impartido en la plataforma edX durante la primavera de 2013 [17]. El curso estaba estructurado en un total de 14 capítulos e incluyó un examen a mitad de curso y otro al final.

Cada capítulo contenía una secuencia de vídeos de lecciones sobre los contenidos tratados en él con una duración individual generalmente inferior a los 10 minutos. Entre los vídeos se intercalaban ejercicios de autoevaluación relacionados con el contenido del vídeo inmediatamente anterior. Los 12 primeros capítulos tenían asociados problemas y ejercicios de laboratorio, basados estos últimos en una aplicación interactiva para el análisis de circuitos. Tanto los problemas como los ejercicios debían ser entregados antes de una fecha límite dada. Algunos capítulos incluían también tutoriales con vídeos de profundización y ampliación de los contenidos. En general, cada lunes del curso se ponían a disposición de los estudiantes los materiales de un nuevo capítulo, estableciéndose el domingo de la semana siguiente

como fecha límite para el envío de problemas y ejercicios de laboratorio (es decir, entre la entrega de materiales y el envío de soluciones hay dos semanas, solapándose la segunda semana de un capítulo con la primera del siguiente). Los dos últimos capítulos tenían vídeos explicativos pero no incluían problemas ni ejercicios.

El curso ofrecía un libro de texto como material de aprendizaje complementario. También ponía a disposición de los alumnos un foro para permitir tanto a los estudiantes como a los profesores participar en debates acerca del curso. Además, existía una wiki que podía ser editada por alumnos y profesores con el objetivo de compartir conocimiento.

La nota final de los alumnos se calculaba ponderando las notas obtenidas en los problemas (15%), los ejercicios de laboratorio (15%), el examen de mitad de curso (30%) y el examen final (40%). Los alumnos que alcanzaron al menos el 50% de la nota máxima obtuvieron un certificado de superación del curso.

En este curso se encontraban registrados 26.947 participantes en la fecha límite de entrega del examen final. De ellos 6.687, sin contar a los profesores ni a sus asistentes, siguieron total o parcialmente el curso de acuerdo con el calendario establecido. Estos participantes, vieron algún vídeo o hicieron alguna entrega antes de la fecha límite de uno más de los capítulos 1 a 12 o vieron algún vídeo antes de la publicación del examen final de los capítulos 13 y 14. De esos participantes hubo un total de 1099 que consiguieron obtener la nota final necesaria para superar el curso.

La Tabla I desglosa por capítulos el número de participantes que vieron vídeos o hicieron entregas de acuerdo con el calendario del curso. En ella es posible observar cómo cayó la participación a medida que avanzó el curso a causa, principalmente, del abandono del curso. En el caso de los capítulos 13 y 14 también influyó el hecho de que no hubiera ejercicios entregables, por lo que muchos participantes decidieron no trabajar en ellos a pesar de que después hicieron el examen final. En la tabla llama también la atención el hecho de que haya bastantes participantes que sólo veían vídeos o sólo hacían envíos. Esto se debe a la existencia de distintas formas de aprovechamiento del curso por parte de los participantes.

TABLA I. NÚMERO DE PARTICIPANTES QUE VIERON VÍDEOS O HICIERON ENTREGAS EN CADA CAPÍTULO DE ACUERDO CON EL CALENDARIO DEL CURSO.

Capítulo	Participantes			
	Videos y envíos	Sólo videos	Sólo envíos	Total
1	3059	1926	818	5803
2	1932	679	787	3398
3	1624	540	539	2703
4	1247	399	548	2194
5	1044	304	524	1872
6	821	185	565	1571
7	694	167	586	1447
8	724	130	432	1286
9	566	83	490	1139
10	528	74	509	1111
11	481	103	423	1007
12	421	133	383	937
13	0	340	0	340
14	0	166	0	166

IV. APROXIMACIÓN PROPUESTA

Los factores utilizados en la literatura para determinar la implicación de los participantes en los MOOC varían de unas propuestas a otras. Entre estos factores se encuentran el consumo de vídeos explicativos [2]-[6], [18], la descarga de esos vídeos para su posterior visionado [18], la realización de pruebas de evaluación [2]-[6], [18] y la participación en los foros de debate [3]-[6]. En este artículo se considera que la implicación de los participantes puede determinarse a partir de su consumo de vídeos y la cantidad de pruebas de evaluación que realizan por ser estos dos factores los únicos que se incluyen en todos los trabajos que se han revisado.

Concretamente, en este trabajo se establece que el nivel de implicación del participante de un MOOC una vez superada la fecha límite de entregas correspondiente al capítulo n puede obtenerse mediante dos indicadores. En primer lugar, un indicador de implicación basado en el consumo de vídeos que se obtiene promediando el porcentaje de vídeos de lecciones que el alumno ha visto en cada uno de los primeros n capítulos antes de la fecha límite correspondiente. En segundo lugar, un indicador de implicación basado en la realización de pruebas de evaluación que se calcula como el promedio del porcentaje de problemas y ejercicios de laboratorio que el alumno ha enviado en cada uno de los primeros n capítulos antes de la fecha límite correspondiente. Lógicamente, el valor de estos indicadores puede aumentar o disminuir tras sobrepasar la fecha límite de entregas del capítulo $n+1$ en función del porcentaje de vídeos vistos y del porcentaje de ejercicios y problemas entregados de ese capítulo.

De este modo, se propone predecir en la fecha límite de entregas de cada capítulo n si la implicación de cada participante definida a partir de los indicadores basados en el consumo de vídeos o la realización de pruebas de evaluación habrá disminuido o no en la fecha límite de entregas del capítulo $n+1$. Aquellos alumnos para los que se prevea la disminución de uno o los dos indicadores deberían ser objeto de una intervención encaminada a intentar evitar la pérdida de implicación predicha.

De acuerdo con el calendario establecido para el curso, la fecha límite de entregas del capítulo n tiene lugar aproximadamente una semana antes de la fecha límite del capítulo $n+1$ y, también aproximadamente, una semana después de la publicación de los materiales de este último capítulo. Esto implica que en el momento de la predicción puede haber participantes que ya hayan visto los suficientes vídeos o realizado las entregas suficientes del capítulo $n+1$ como para saber que el indicador correspondiente no disminuirá. En ese caso, lógicamente, no se lleva a cabo la predicción del indicador. Tampoco se hace la predicción de un indicador de un participante cuando en el momento de realizarla dicho indicador tiene valor 0, ya que el valor nunca podrá disminuir. Este caso se da en alumnos que participan en el MOOC viendo vídeos pero que no hacen ningún envío de ejercicios o problemas o, al contrario, que envían problemas pero no ven ningún vídeo. También se excluye de la predicción a aquellos participantes que no han visto ningún vídeo ni han realizado ningún envío en los 3 últimos capítulos puesto que se considera que probablemente hayan abandonado

el curso. Si eventualmente estos participantes vuelven a ver vídeos o a hacer entregas, se generan nuevamente predicciones para ellos.

V. CARACTERÍSTICAS Y ALGORITMOS DE PREDICCIÓN

La Tabla II muestra de forma resumida las características que fueron extraídas para cada uno de los 11 primeros capítulos. La extracción no se llevó a cabo en más capítulos dado que los capítulos 13 y 14 no tenían entregas asociadas, motivo por el cual no fueron considerados en los experimentos de predicción de indicadores de motivación. Nótese que las características del capítulo $n+1$ corresponden a la parte de ese capítulo que ocurre antes de la entrega del capítulo n .

Se estudiaron dos alternativas diferentes para reducir el número de características que fueron empleadas finalmente por los algoritmos de predicción. En una de ellas se utilizó la técnica de búsqueda hacia delante para explorar el espacio de subconjuntos de características cuya calidad fue evaluada mediante la técnica de Selección de Características basada en Correlación (CFS, *Correlation based Feature Selection*) [19]. De esta forma es posible encontrar subconjuntos de características poco correladas entre sí pero muy correlados con las categorías consideradas. En la segunda alternativa se llevó a cabo un análisis de componentes principales (PCA, *Principal Component Analysis*) para transformar los datos utilizando el número de autovectores con mayor autovalor necesario para explicar el 95% de la varianza en los datos originales.

También se evaluaron 3 algoritmos de aprendizaje automático supervisado para llevar a cabo la tarea de predicción a partir de las características obtenidas mediante las técnicas antes descritas. Estos algoritmos fueron la Regresión Logística (*Logistic Regression*, LR), las Máquinas de Vectores de Soporte (*Support Vector Machines*, SVM) [20] con un núcleo de función de base radial y las Selvas Aleatorias (*Random Forests*, RF) [21]. Los datos fueron normalizados para los experimentos realizados con el SVM, pero no para los de los otros 2 algoritmos.

VI. RESULTADOS EXPERIMENTALES

La fase de experimentación de este trabajo se ha llevado a cabo con el entorno de minería de datos Weka⁴. Concretamente, se repitió 10 veces una validación cruzada estratificada de 10 subconjuntos con cada combinación posible de las técnicas de reducción del número de características y algoritmos de aprendizaje supervisado presentados en la sección anterior para predecir la posible disminución de cada uno de los dos indicadores de implicación en cada capítulo del 2 al 12 respecto al capítulo anterior.

La Tabla III muestra el número medio de características que fueron obtenidas para cada uno de los capítulos e indicadores de implicación mediante las dos técnicas de reducción de características consideradas. En ella puede

⁴ <http://www.cs.waikato.ac.nz/ml/weka/>

TABLA II. LISTA DE CARACTERÍSTICAS CONSIDERADAS PARA LA PREDICCIÓN AGRUPADAS POR TIPO Y CON INDICACIÓN DEL PERIODO DE TIEMPO ESTABLECIDO PARA SU CÁLCULO Y LOS CAPÍTULOS EN LAS QUE FUERON EXTRAIDAS

Tipo	Características	Periodo	Capítulos
Capítulo $n+1$	C1 Porcentaje de vídeos de lecciones vistos total o parcialmente C2 Porcentaje de ejercicios de autoevaluación enviados C3 Porcentaje de problemas y ejercicios de laboratorio enviados	Inicio del capítulo n a final del capítulo n	1-11
Capítulo n	C4 Porcentaje de vídeos de lecciones vistos total o parcialmente C5 Porcentaje de otros vídeos vistos total o parcialmente C6 Porcentaje de ejercicios de autoevaluación enviados C7 Porcentaje de problemas y ejercicios de laboratorio enviados C8 Nota de los ejercicios de autoevaluación enviados C9 Nota de los problemas enviados C10 Nota de los ejercicios de laboratorio enviados C11 Suma de la nota de problemas y ejercicios de laboratorio enviados C12 Número de días consecutivos inactivo hasta la fecha de entrega C13 Número de consultas de la página de progreso C14 Número de eventos del foro de discusión generados C15 Número de eventos del libro de la asignatura generados C16 Número de eventos de la wiki generados C17 Número total de eventos generados	Inicio del capítulo n a final del capítulo n	1-11
Capítulo $n-1$	C18 Porcentaje de vídeos de lecciones vistos total o parcialmente C19 Porcentaje de ejercicios de autoevaluación enviados C20 Porcentaje de problemas y ejercicios de laboratorio enviados	Inicio del capítulo $n-1$ a final del capítulo $n-1$	2-11
Capítulo $n-2$	C21 Porcentaje de vídeos de lecciones vistos total o parcialmente C22 Porcentaje de ejercicios de autoevaluación enviados C23 Porcentaje de problemas y ejercicios de laboratorio enviados	Inicio del capítulo $n-2$ a final del capítulo $n-2$	3-11
Acumulado	C24 Media acumulada de porcentajes de vídeos de lecciones vistos total o parcialmente (indicador de implicación basado en vídeo) C25 Media acumulada de porcentajes de ejercicios de autoevaluación enviados C26 Media acumulada de porcentajes de problemas y ejercicios de laboratorio enviados (indicador de implicación basado en pruebas de evaluación) C27 Nota acumulada de los ejercicios de autoevaluación enviados C28 Nota acumulada de los problemas enviados C29 Nota acumulada de los ejercicios de laboratorio enviados C30 Suma de la nota de problemas y ejercicios de laboratorio enviados	Inicio del curso a final del capítulo n	2-11
Combinación	C31 Diferencia entre el porcentaje y la media acumulada de porcentajes de vídeos de lecciones C32 Diferencia entre el porcentaje y la media acumulada de porcentajes de ejercicios de autoevaluación C33 Diferencia entre el porcentaje y la media acumulada de porcentajes de problemas y ejercicios de laboratorio enviados	No aplicable	2-11
Examen de mitad de curso	C34 Porcentaje de ejercicios del examen de prueba mitad de curso enviados C35 Porcentaje de ejercicios del examen de mitad de curso enviados C36 Nota del examen de prueba de mitad de curso C37 Nota del examen de mitad de curso	Inicio del examen de mitad de curso a final del examen de mitad de curso	7-11

TABLA III. NÚMERO DE CARACTERÍSTICAS OBTENIDAS MEDIANTE CADA TÉCNICA DE REDUCCIÓN PARA CADA CAPÍTULO Y TIPO DE PREDICCIÓN CON INDICACIÓN DEL NÚMERO DE CARACTERÍSTICAS INICIAL.

Cap.	Caract.	Vídeos		Pruebas	
		CFS	PCA	CFS	PCA
1	20	8,32	11	3	12
2	30	12,53	13	8,96	13,08
3	32	6,94	13	3,03	14
4	32	12,45	14	7,3	14
5	33	9,86	15	4,99	15,72
6	32	9,74	14,18	11,26	15
7	36	16,53	16	11,13	16,02
8	36	14,36	17	13,43	17
9	37	10,28	18	13,8	18
10	37	12,12	18,91	11,4	19
11	36	11,7	17,97	12,21	17,99

llamar la atención el hecho de que el número de características extraídas por PCA no sea exactamente igual para ambos indicadores en un mismo capítulo dado que el proceso de extracción es independiente de la relación de las características con la variable predicha, a diferencia del caso de CFS. Esto se debe a que, aunque el conjunto de vectores

utilizados para la predicción de cada indicador en un mismo capítulo es muy similar, existen diferencias debidas a las posibles exclusiones de participantes de la predicción de un indicador pero no del otro de acuerdo con lo establecido en la sección IV. También se observa que el número de características seleccionadas por CFS es sensiblemente inferior al de PCA en la mayor parte de los casos, siendo superior en sólo una ocasión.

La Tabla IV y la Tabla V recogen los resultados de los experimentos de predicción de los indicadores de implicación junto con el número de participantes para los que se llevaron a cabo las predicciones en cada capítulo. Estos resultados se describen utilizando dos métricas: el área bajo la curva (*Area Under Curve*, AUC) y el porcentaje de aciertos (PA). Es importante tener en cuenta que el AUC, a diferencia de otras métricas como el PA, no se ve afectada por la presencia de sesgo en los conjuntos de datos [22]. Por este motivo, los resultados de los experimentos deben ser comparados utilizando el AUC, mientras que el PA se reporta únicamente a efectos informativos.

TABLA IV. CAPACIDAD PREDICTIVA DE CADA COMBINACIÓN DE TÉCNICAS DE REDUCCIÓN DE CARACTERÍSTICAS Y ALGORITMOS DE APRENDIZAJE SUPERVISADO Y CADA CAPÍTULO EN EL CASO DEL INDICADOR DE IMPLICACIÓN BASADO EN EL CONSUMO DE VÍDEOS.

Cap.	Part.	Métr.	LR		SVM		RF	
			CFS	PCA	CFS	PCA	CFS	PCA
2	4.638	AUC	0,857	0,859	0,503	0,5	0,82	0,84
		PA	86,5	86,5	85,4	85,3	84,9	86
3	5.013	AUC	0,877	0,877	0,675	0,5	0,862	0,867
		PA	85,4	85,3	85,6	83,5	84,7	84,6
4	5.383	AUC	0,892	0,9	0,5	0,5	0,852	0,887
		PA	90,3	90,3	89,6	89,6	89,2	90,4
5	3.203	AUC	0,881	0,884	0,702	0,664	0,877	0,871
		PA	86,5	86,5	86,2	85,1	86,4	85,9
6	2.493	AUC	0,894	0,897	0,5	0,5	0,874	0,878
		PA	91	91,1	90,4	90,4	91,4	91,1
7	2.098	AUC	0,904	0,906	0,761	0,754	0,895	0,898
		PA	88,9	88,9	88,6	88,2	88,5	88,3
8	1.750	AUC	0,892	0,898	0,74	0,584	0,888	0,885
		PA	86,6	86,5	86,4	82,9	86,3	85,6
9	1.525	AUC	0,888	0,895	0,613	0,5	0,88	0,874
		PA	86,7	86,7	85,2	84,5	86,3	85,6
10	1.344	AUC	0,898	0,909	0,769	0,653	0,895	0,903
		PA	88,5	87,8	87,6	86	87,7	87,5
11	1.175	AUC	0,873	0,877	0,785	0,702	0,866	0,865
		PA	85,1	85,2	84,8	83,9	84,5	84,3
12	1.054	AUC	0,825	0,837	0,5	0,5	0,847	0,838
		PA	84,2	84,5	85,2	85,2	84,9	84,9

TABLA V. CAPACIDAD PREDICTIVA DE CADA COMBINACIÓN DE TÉCNICAS DE REDUCCIÓN DE CARACTERÍSTICAS Y ALGORITMOS DE APRENDIZAJE SUPERVISADO Y CADA CAPÍTULO EN EL CASO DEL INDICADOR DE IMPLICACIÓN BASADO EN LA REALIZACIÓN DE PRUEBAS DE EVALUACIÓN.

Cap.	Part.	Métr.	LR		SVM		RF	
			CFS	PCA	CFS	PCA	CFS	PCA
2	3.453	AUC	0,787	0,8	0,748	0,734	0,776	0,778
		PA	73,2	72,8	72,6	71,2	73,4	71,1
3	3.891	AUC	0,857	0,856	0,784	0,787	0,829	0,842
		PA	79,3	79,1	78,7	78,5	76,3	77,6
4	4.053	AUC	0,886	0,906	0,832	0,839	0,879	0,894
		PA	84,2	83,9	84,2	83,6	84,2	82,8
5	3.110	AUC	0,914	0,915	0,846	0,848	0,9	0,904
		PA	85,1	85,2	85,2	85	84,1	84,7
6	2.289	AUC	0,909	0,914	0,854	0,858	0,889	0,904
		PA	85,8	85,8	85,2	84,7	83,5	84,1
7	1.912	AUC	0,899	0,901	0,824	0,83	0,892	0,894
		PA	81,8	82	82,2	82,7	81,5	81,6
8	1.650	AUC	0,904	0,899	0,832	0,822	0,895	0,889
		PA	82,8	82,3	82,6	81	81,1	81,3
9	1.407	AUC	0,898	0,891	0,815	0,822	0,897	0,878
		PA	81,9	81,5	81,2	81,8	81,2	80,8
10	1.258	AUC	0,857	0,858	0,781	0,778	0,85	0,844
		PA	77,3	78,2	77,8	77,5	77,3	78,5
11	1.149	AUC	0,772	0,771	0,707	0,709	0,755	0,762
		PA	69,9	69,9	70	70,2	69	69,1
12	1.113	AUC	0,725	0,722	0,609	0,5	0,715	0,696
		PA	74,7	72,5	73,6	71,1	72,8	72,4

Los resultados muestran que es posible predecir para cada capítulo el indicador de implicación basado en vídeos con valores de AUC comprendidos entre 0,847 y 0,909 mientras que en el caso del indicador de implicación basado en la realización de pruebas de evaluación es posible hacerlo con valores de AUC que van desde 0,725 hasta 0,915. Aquí es necesario recordar que no se hacen predicciones para aquellos participantes que no han visto vídeos ni realizado pruebas de

evaluación durante los 3 últimos capítulos, aunque se vuelven a generar predicciones para ellos si en algún capítulo posterior realizan alguna de esas actividades. Esto hace que los participantes que abandonan el curso dejen de ser considerados tras 3 capítulos de inactividad, evitando así que los resultados obtenidos mejoren mediante predicciones en las que es muy fácil acertar.

En las tablas también se aprecia que los mejores resultados de predicción se consiguen en todos los casos menos en uno empleando LR como algoritmo de aprendizaje supervisado y, en la mayoría de capítulos, empleando PCA como técnica para la reducción del número de características. Sin embargo, los valores de AUC obtenidos con LR y CFS difieren en general muy poco de los conseguidos con LR y PCA, por lo que sería razonable elegir la primera combinación en caso de que se desee ahorrar la mayor carga computacional que supone la segunda opción.

Es posible resaltar la subida de la AUC que se produce en el capítulo 4 respecto a los dos capítulos anteriores en la predicción del indicador basado en el envío de pruebas de evaluación y, en menor medida, del indicador basado en el consumo de vídeos. Una parte de esta subida puede ser atribuida a que en ese momento no se dispone de tanta información de los participantes como en capítulos posteriores para hacer la predicción. Otra parte de la subida puede deberse al hecho de que hubo bastantes participantes que abandonaron el curso en los capítulos 3 ó 4 a pesar de que en los capítulos anteriores participaron activamente al igual que muchos otros participantes que no abandonaron. La pérdida de implicación de estos participantes, conocidos en la literatura como *strong starters* [5], no es fácil de predecir.

También resulta llamativa la caída en la AUC en la predicción de ambos indicadores para los capítulos 11 y 12 respecto al capítulo 10. Esta caída parece deberse en buena parte al hecho de que para la nota final del curso no contaban las dos notas más bajas de las entregas de todos los capítulos del curso. Aparentemente esto dio lugar a que bastantes participantes que habían mantenido un alto nivel de implicación y obtenido buenas notas a lo largo del curso redujeran significativamente su actividad en los dos últimos capítulos al considerar que no necesitaban las notas de sus entregas.

VII. CONCLUSIONES Y TRABAJO FUTURO

La predicción de pérdida de implicación de los participantes de un MOOC permite la intervención antes de que dicha pérdida se produzca con el objetivo de evitarla. Así es posible conseguir que los participantes hagan un mayor aprovechamiento del curso. En este artículo se ha estudiado la posibilidad de predecir la pérdida de implicación en un MOOC impartido en la plataforma edX a partir de dos indicadores, uno basado en el consumo de vídeos y otro en la realización de pruebas de evaluación. Para ello se ha llevado a cabo la extracción de una serie de características que describen la actividad de los participantes a lo largo del curso. Estos datos han permitido realizar la comparación de 2 técnicas de reducción del número de características y 3 algoritmos de aprendizaje supervisado para llevar a cabo la predicción de

ambos indicadores. Así se ha comprobado que es posible conseguir buenos resultados de predicción para ambos indicadores utilizando LR como algoritmo de aprendizaje supervisado en junto con PCA como técnica de selección de características. Esta combinación también ha dado buenos resultados en problemas relacionados como la predicción de abandono [14].

Las líneas de trabajo futuro que se abren con este trabajo son numerosas. La más inmediata pasa por analizar la importancia que tienen las distintas características en la predicción de los indicadores de implicación. También se pretende ampliar el estudio considerando nuevas características, técnicas para la reducción del número de características y algoritmos de aprendizaje supervisado en el contexto de un conjunto más amplio de MOOC. Además, se analizará la posibilidad de utilizar nuevos indicadores de implicación como la participación en foros. Otra cuestión que merece la pena tratar es la adaptación de la aproximación para la predicción de pérdida de implicación propuesta a otros formatos de MOOC diferentes al considerado en este artículo como, por ejemplo, los cursos que permiten el envío de pruebas evaluables después de la fecha límite establecida. Además, se pretende estudiar hasta qué punto es posible utilizar los datos de un curso para construir sistemas de predicción de la pérdida de implicación que se utilicen en una edición posterior de ese mismo curso. También se trabajará en el diseño de los mecanismos de intervención que se puedan utilizar exitosamente con los participantes para los que se predice la pérdida de implicación.

AGRADECIMIENTOS

El acceso a los datos utilizados en este artículo ha sido posible gracias a un acuerdo suscrito con la Oficina de Investigación Institucional del *Massachusetts Institute of Technology*. Este trabajo ha sido parcialmente financiado por el Ministerio de Economía y Competitividad del Gobierno de España (proyecto de investigación TIN2014-53199-C3-2-R) y la Consejería de Educación de la Junta de Castilla y León (proyecto de investigación VA277U14). Los autores agradecen también al resto de miembros del grupo de investigación GSIC/EMIC su apoyo y participación en el debate de ideas.

REFERENCIAS

- [1] T. R. Liyanagunawardena, A. A. Adams, and S. A. Williams, "MOOCs: a systematic study of the published literature 2008-2012," *The International Review of Research in Open and Distance Learning*, vol. 14, no. 3, pp. 202–227, 2013.
- [2] R. F. Kizilcec, C. Piech, and E. Schneider, "Deconstructing disengagement: analyzing learner subpopulations in massive open online courses," in Proceedings of the Third International Conference on Learning Analytics and Knowledge, Leuven, Belgium, 2013, pp. 170–179.
- [3] A. Ramesh, D. Goldwasser, and B. Huang, "Modeling learner engagement in MOOCs using probabilistic soft logic," in Proceedings of the Neural Information Processing Systems Conference, Workshop on Data Driven Education, Lake Tahoe, NV, USA, 2013.
- [4] A. Ramesh, D. Goldwasser, B. Huang, H. Daume, III, and L. Getoor, "Uncovering hidden engagement patterns for predicting learner performance in MOOCs," in Proceedings of the First ACM Conference on Learning at Scale, New York, New York, USA, 2014, pp. 157–158.
- [5] R. Ferguson and D. Clow, "Examining engagement," in Proceedings of the Fifth International Conference on Learning Analytics And Knowledge, Poughkeepsie, NY, USA, 2015, pp. 51–58.
- [6] R. Ferguson, D. Clow, R. Beale, A. J. Cooper, N. Morris, S. Bayne, and A. Woodgate, "Moving Through MOOCs: Pedagogy, Learning Design and Patterns of Engagement," in Proceedings of the Tenth European Conference on Technology Enhanced Learning, Toledo, Spain, 2015, pp. 70–84.
- [7] J. E. Beck, "Engagement tracing: using response times to model student disengagement," in Proceedings of the International Conference on Artificial Intelligence in Education, Pittsburgh, PA, USA, 2005, pp. 88–95.
- [8] C. Mills, N. Bosch, A. Graesser, and S. D'Mello, "To quit or not to quit: predicting future behavioral disengagement from reading patterns," in Proceedings of the International Conference on Intelligent Tutoring Systems, Honolulu, HI, USA, 2014, pp. 19–28.
- [9] M. Cocca and S. Weibelzahl, "Disengagement Detection in Online Learning: Validation Studies and Perspectives," *IEEE Trans. Learning Technol.*, vol. 4, no. 2, pp. 114–124.
- [10] D. Yang, T. Sinha, and D. Adamson, "Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses," in Proceedings of the Neural Information Processing Systems Conference, Workshop on Data Driven Education, Lake Tahoe, NV, USA, 2013.
- [11] S. Halawa, D. Greene, and J. Mitchell, "Dropout prediction in MOOCs using learner activity features," in Proceedings of the European MOOCs Stakeholders Summit, Lausanne, Switzerland, 2014, pp. 58–65.
- [12] C. Ye and G. Biswas, "Early prediction of student dropout and performance in MOOCs using higher granularity temporal information," *Journal of Learning Analytics*, vol. 1, no. 3, pp. 169–172, Dec. 2014.
- [13] W. Xing, X. Chen, J. Stein, and M. Marcinkowski, "Temporal prediction of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization," *Computers in Human Behavior*, vol. 58, pp. 119–129, May 2016.
- [14] C. Taylor, "Stopout prediction in massive open online courses," PhD dissertation, Massachusetts Institute of Technology, USA, 2014.
- [15] S. Jiang and M. Warschauer, "Predicting MOOC performance with week 1 behavior," in Proceedings of the International Conference on Educational Data Mining, London, UK, 2014, pp. 273–275.
- [16] J. He, J. Bailey, B. I. P. Rubinstein, and R. Zhang, "Identifying at-risk students in Massive Open Online Courses," in Proceedings of the Conference on Artificial Intelligence, Austin, TX, USA, 2015, pp. 1749–1755.
- [17] D. T. Seaton, J. Reich, S. O. Nesterko, T. Mullaney, J. Waldo, A. D. Ho, and I. Chuang, "6.002x Circuits and Electronics MITx on edX Course Report - 2013 Spring," 8, Jan. 2014.
- [18] A. Anderson, D. P. Huttenlocher, J. M. Kleinberg, and J. Leskovec, "Engaging with massive online courses," in Proceedings of the International Conference on World Wide Web, New York, NY, USA, 2014, pp. 687–698.
- [19] M. A. Hall, "Correlation-based feature selection for machine learning," PhD dissertation, Waikato University, Hamilton, New Zealand, 1999.
- [20] V. N. Vapnik, *The nature of statistical learning theory*, Second edition. New York, NY, USA: Springer, 2000.
- [21] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [22] L. A. Jeni, J. F. Cohn, and F. De la Torre, "Facing imbalanced data-recommendations for the use of performance metrics," in Proceedings of the Conference on Affective Computing and Intelligent Interaction, Geneva, Switzerland, 2013, pp. 245–251.