Exploración de datos para predicción de dropout con Zeppelin

Milagro Teruel

Programación distribuida sobre grandes volúmenes de datos

Este reporte técnico es un resumen del proceso de exploración de un conjunto de datos utilizado para la predicción de abandono (dropout) en plataformas masivas de aprendizaje online. El objetivo final es utilizar un algoritmo de aprendizaje automático que pueda clasificar cada estudiante, representado de acuerdo a sus características, en dropout y no dropout. El primer paso, que describimos a continuación, es entender qué información está disponible dentro del conjunto de datos, y qué características son potencialmente más importante para la tarea de predicción.

Luego de una breve introducción al problema, se detallan las distintas técnicas utilizadas para la extracción de métricas. Distintos aspectos han sido evaluados, incluyendo características generales, distribuciones comparativas entre cursos y etiquetas, y análisis de datos enfocados a la tarea de predicción.

En la última sección, aplicamos dos algoritmos de clasificación lineal al problema y analizamos la importancia de distintas las características en el desempeño final de cada modelo. Los datos serán procesados para utilizar también redes neuronales, pero los resultados no están incluidos en este trabajo.

1 Aprendizaje en plataformas online

La Minería de Datos Educativos (EDM por sus siglas en inglés) es un área compleja, con causas latentes desconocidas que gobiernan el comportamiento de los estudiantes y el éxito o fracaso de los cursos. En el último decenio, la cantidad de datos generados en los entornos de aprendizaje virtual ha aumentado constantemente. Estos grandes conjuntos de datos facilitan la aplicación de enfoques de aprendizaje automático y ciencia de datos para abordar tareas como la predicción y prevención de la deserción escolar, o la recomendación del siguiente ítem para mejorar el aprendizaje.

Uno de los principales desafíos en esta área es obtener abstracciones adecuadas, ya que los conjuntos de datos disponibles están compuestos por registros de bajo nivel. Es significativamente desafiante entrenar un algoritmo de aprendizaje automático que pueda modelar tareas complejas, como la predicción de deserción escolar, basándose únicamente en los registros de la interacción de los estudiantes.

Otro reto en EDM es la diversidad de los datos sobre educación. Diferentes plataformas generan datos con formatos distintos y almacenan su propio conjunto de señales. Incluso dentro de la misma plataforma, los cursos a menudo no comparten contenido, dificultando las posibilidades de transferir el aprendizaje entre modelos.

Por estas razones, la aplicación de modelos automáticos que puedan captar causas latentes profundas puede beneficiar significativamente el análisis y diagnóstico de los contenidos educativos. En el caso particular de la predicción de deserción escolar, se puede utilizar un sistema de predicción para monitorear la actividad de los estudiantes y emitir alertas para prevenir la deserción escolar a tiempo.

2 Exploración de datos

La competencia KDDCup 2015 propuso la tarea de predecir la deserción escolar (dropout) en cursos masivos o Massive Open Online Courses (MOOCs). Los datos fueron proporcionados por XuetangX, una plataforma de aprendizaje MOOC china iniciada por la Universidad de Tsinghua y socio de EdX. Para esta tarea en particular, el evento de abandono se definió como la ausencia de actividad estudiantil en los diez días siguientes, aunque no está claro si los estudiantes habían completado el curso en ese momento.

Aunque la información ya no está disponible en el sitio web del concurso, este conjunto de datos es uno de los pocos ejemplos de registro detallado en un entorno MOOC.

Los datos proporcionados incluyen información de 38 cursos diferentes. Todos los cursos tienen una duración de 29 días, comenzando en diferentes fechas desde octubre de 2013 hasta agosto de 2014. En este trabajo, sólo pudimos utilizar la parte de entrenamiento del conjunto de datos, que describimos a continuación. Los organizadores de la competición sólo publican las secuencias de interacciones de la parte de evaluación, pero no las etiquetas. Por lo tanto, no podemos usarlas para evaluar ningún clasificador.

2.1 Características generales

El conjunto de datos comprende 120.542 registros etiquetados, lo que corresponde a 79.186 usuarios únicos. El total de 8.157.278 interacciones incluyen información como el acceso a contenidos de vídeo, la resolución de un problema, el acceso al wiki del curso, etc. Los eventos tienen fecha y hora y se identifican con el estudiante y el curso correspondiente (inscripción). Además de las actividades de los alumnos, el concurso proporcionó información sobre la organización jerárquica de los elementos del curso.

Los cursos no son del mismo tamaño, variando de 12004 a 645 (curso 0 y 38 respectivamente), con la distribución mostrada en la Figura 1. El número de interacciones únicas no corresponde directamente al número de inscripciones, con un máximo de 907.118 para el curso 1 y un mínimo de 21.216 para el curso 25. La distribución detallada se presenta en la Figura 2.

En todo el conjunto de datos, el 80% de las inscripciones corresponde a deserción escolar, aunque esta distribución cambia entre cursos. Como podemos observar en las figuras

Figure 1: Distribución de las inscripciones únicas por curso

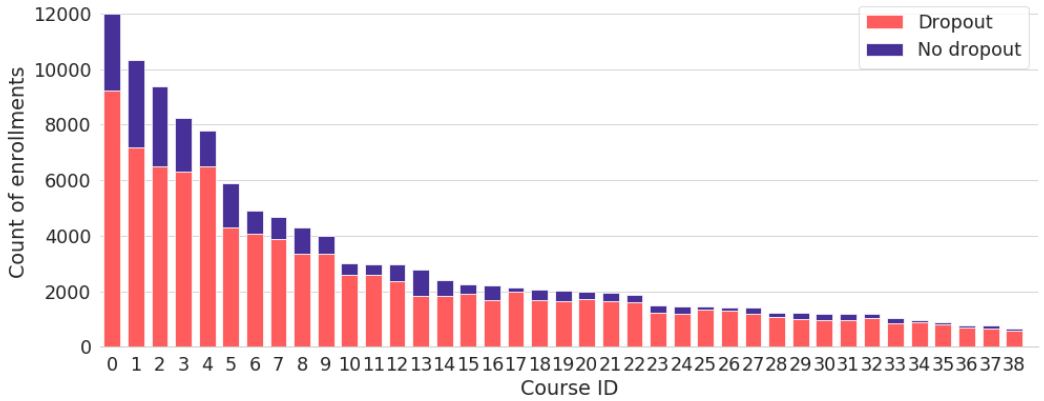


Figure 2: Distribución de interacciones únicas por curso

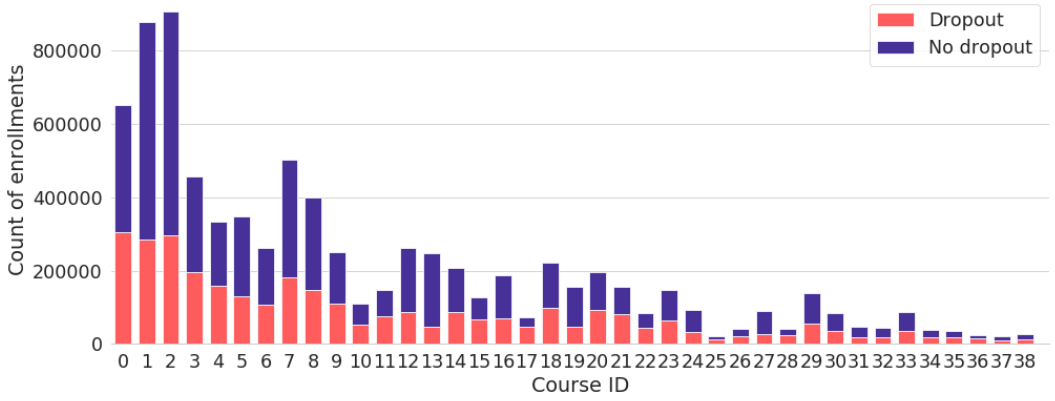
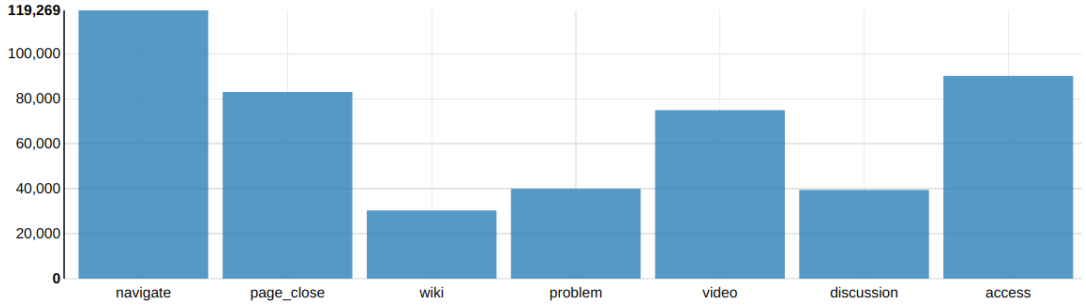


Figure 3: Cantidad de interacciones de acuerdo al tipo de evento



anteriores, la mayoría de las inscripciones corresponden a dropout, pero la mayoría de las interacciones son producidas por inscripciones sin dropout. Esto es esperable, ya que la intuición indica que las secuencias de estudiantes que abandonan son más cortas.

2.2 Detección de anomalías

Es común encontrar en los conjuntos de datos ejemplos que sobresalen y no siguen la distribución general. Este tipo de datos pueden provenir de diversas fuentes: fallas en los instrumentos de recolección de datos, errores en la entrada de datos, reporte incorrecto de un fenómeno de diferente naturaleza, entre otros. Además, también pueden ser verdaderos valores atípicos en la distribución, lo que corresponde a observaciones válidas, aunque atípicas.

Algunos modelos de aprendizaje automático, como las regresiones lineales, son susceptibles a distribuciones desequilibradas y a la presencia de ejemplos anómalos. Por esta razón, es una práctica común conservar los conjuntos de datos antes del entrenamiento, eliminando cualquier registro sospechoso.

Las redes neuronales se consideran robustas en este sentido, pero las redes neuronales recurrentes en particular pueden verse afectadas por secuencias demasiado largas. Dado que las secuencias se rellenan con la secuencia más larga del lote, si una secuencia es excesivamente larga también afectará a otros ejemplos, lo que provocará un aumento del tiempo y de los recursos necesarios para el entrenamiento del modelo. Además, i) la mayoría de las secuencias largas han sido etiquetadas como no abandonadas, ii) podrían corresponder a otros agentes, como bots o scripts automáticos, en lugar de estudiantes reales.

En este conjunto de datos nos encontramos con varias secuencias demasiado largas, que decidimos descartar por la razón mencionada anteriormente. El criterio exacto fue eliminar las inscripciones con un número de eventos mayor que el percentil 0.99, dejando 7,171,235 interacciones.

2.3 Distribución de los eventos

Hay más de 26.750 objetos descritos en el conjunto de datos, pero sólo 5.890 de ellos aparecen en los registros. Podemos ver en la Figura 3 la distribución aproximada de las interacciones en el conjunto de datos según su tipo de evento.

A través de una exploración preliminar, observamos en varios cursos que la proporción de acciones de *access* y *navigation* son más prominentes en los estudiantes con dropout. Por otro lado, los eventos de tipo *problem* son más comunes en los estudiantes que no abandonan. Para verificar que este fenómeno existe en todos los cursos, creamos una visualización más extensa.

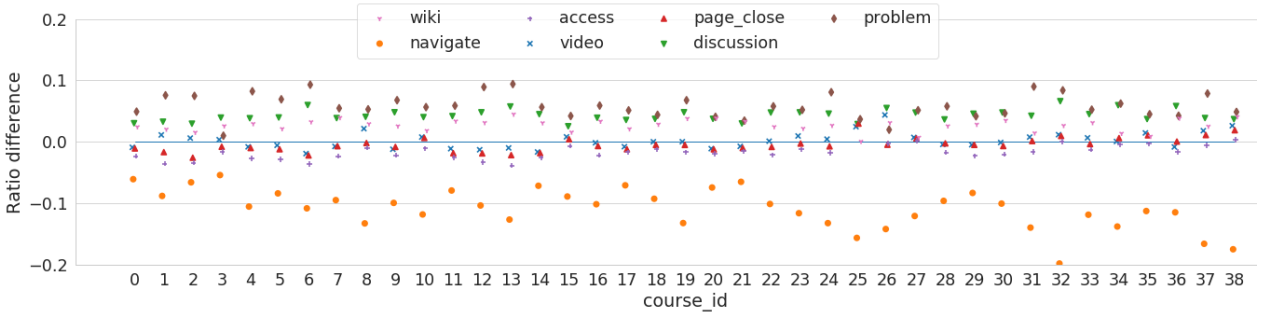


Figure 4: Diferencia en la proporción de interacciones entre estudiantes sin dropout y con dropout, para cada tipo de acción, separados por curso.

En la Figura 4 podemos observar la diferencia entre las proporciones de los tipos de eventos en los estudiantes sin dropout y con dropout. Si la diferencia es positiva, es decir, por encima de la línea, indica que los estudiantes que no abandonan proporcionalmente acceden más a ese tipo de eventos que los estudiantes que sí abandonan. Por ejemplo, en el curso 1, los estudiantes que no abandonan tienen una proporción de interacciones de acceso 9% menor que los estudiantes que abandonan. Podemos ver que, en casi todos los cursos, los tipos de eventos *problem* y discusión son tomados más por los estudiantes que no abandonan, mientras que lo opuesto es cierto para los eventos de tipo navegación y acceso. La diferencia en los tipos restantes, *video*, *page close*, y *wiki*, es cercana a cero en la mayoría de los cursos, lo que indica que pueden no estar relacionados con la etiqueta del estudiante. Esto apoya la intuición de que los estudiantes que abandonan están menos involucrados en las actividades y revisan el contenido de una manera más superficial.

3 Análisis por sesiones

Para comprender mejor los patrones presentes en las secuencias, es necesario agregar los datos de bajo nivel (interacciones) en unidades más interpretables. Elegimos para esta tarea utilizar sesiones, definidas como una secuencia de acciones donde el tiempo entre ellas no excede un límite de dos horas.

Al graficar diferentes distribuciones, notamos algunos patrones interesantes. El número de sesiones por estudiante tiene una distribución diferente entre la clase que abandona y la que no abandona, como puede verse en la Figura 5. Este patrón es consistente entre los diferentes cursos. El número de acciones y la duración de la sesión es ligeramente superior en los estudiantes que no abandonan los estudios, mientras que el promedio de minutos entre las acciones es muy similar.

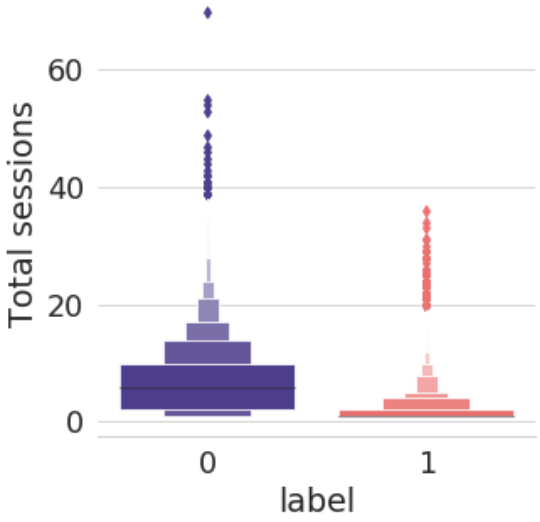


Figure 5: Distribución del número de sesiones por inscripción, de acuerdo a la etiqueta (el valor 1 corresponde a abandono)

La información codificada en las sesiones son características importantes que pueden ser indicativas de dropout, junto el número total de interacciones. Como no es posible conocer el número final de sesiones o eventos hasta que la secuencia finaliza, esta característica es menos útil en escenarios donde la predicción debe ser realizada con información parcial. Sin embargo, observamos que el tiempo transcurrido entre sesiones también es mayor en el caso de los estudiantes que abandonan, y podría utilizarse como un fuerte predictor para secuencias incompletas.

4 Clasificadores lineales

El enfoque de aprendizaje automático más sencillo para una tarea de clasificación es, en general, la aplicación de modelos lineales. Por superficiales que sean, los modelos lineales pueden ayudar a comprender las relaciones entre sus variables ocultas y los fenómenos observados. Los modelos simples también son útiles para evaluar la complejidad de un problema, y para evaluar la necesidad de modelos más sofisticados o profundos.

Nos centraremos en dos tipos de clasificadores: Regresión Logística (RL) y Árboles de Decisión (DT, por sus siglas en inglés). Ninguno tiene la capacidad de procesar datos secuenciales y, en consecuencia, es necesario encontrar una representación adecuada de las secuencias de interacciones que pueda ser utilizada como entrada. Esta representación se basa en gran medida en características seleccionadas o construidas manualmente (*handcrafted*), y debe resaltar sólo los aspectos importantes de los datos para que un clasificador lineal alcance un rendimiento adecuado. En este sentido, se pueden utilizar clasificadores lineales

para evaluar la calidad de una representación con respecto a una tarea de clasificación. Es importante tener en cuenta que este tipo de representaciones manuales son particulares para cada conjunto de datos, y deben ser adaptadas para distintos problemas.

Con este análisis queremos responder a la pregunta: ¿son las estadísticas y las características calculadas sobre la secuencia, suficientes para predecir el dropout? A diferencia de otras tareas de EDM, hay menos investigación sobre el tipo de características que predicen el abandono. Los conjuntos de datos son más variados, e incluso no hay un consenso sobre la definición de *dropout*. Es significativo entender, para este conjunto de datos en particular, las relaciones entre las características y las etiquetas de abandono.

Para construir el vector de características que representa a cada estudiante, se utilizan las conclusiones obtenidas en el análisis anterior. Las características finales son:

- El número total de interacciones
- El número promedio de interacciones por sesión, dividido por tipo
- El número de interacciones por sesión, dividido por tipo de evento
- El número de sesiones
- La duración media de las sesiones

Esto resultó en 18 columnas de características con valores densos. El único paso de preproceso que se aplicó fue la estandarización, para evitar sobreponderar una sola característica. Los valores que faltaban se sustituyeron por ceros. Utilizamos el 80% de las filas para la formación y el 20% para las pruebas, ya que no había necesidad de un conjunto de datos de validación. Los hiperparámetros utilizados eran los predeterminados para la biblioteca Spark ML.

También evaluamos al mismo clasificador LR, entrenado sin ninguna información sobre las sesiones. Esto nos permitirá evaluar hasta qué punto esas características son esenciales para la tarea de clasificación. En este caso, sólo se utiliza el número total de interacciones y el número de interacciones por tipo.

Para tener una intuición de cuán difícil que es una tarea, es común utilizar una regla de decisión simple como punto de comparación. Por ejemplo, se pueden asignar una etiqueta aleatoria o utilizar el valor más común. En este caso, para todos los cursos asignaremos la etiqueta de abandono (1). Esto también se corresponde con el escenario de asumir que cada estudiante está en riesgo, y producir intervenciones para cada uno de ellos.

Como métricas de evaluación, utilizaremos principalmente el área bajo la curva ROC (AUC ROC), que fue la métrica elegida para seleccionar el ganador de la competencia. Cuando la salida del clasificador es probabilística, el AUC permite comparar clasificadores en múltiples umbrales de predicción (punto de separación entre la clase positiva y la clase negativa). A modo comparativo, usaremos otras métricas adicionales. El score R2 mide la proporción de varianza de los datos capturada por el clasificador. Un puntaje R2 de 0 indica que el clasificador no explica ninguna varianza de los datos, y puede eventualmente tomar valores negativos. La raíz cuadrada del error cuadrático medio (RMSE), que estima cuán

lejos se encuentra el valor asignado por la predicción de la etiqueta real, dando mayor puntaje a clasificadores con alta certeza que polaricen los valores predichos en valores extremos 0 y 1. Finalmente, el Accuracy o exactitud es una métrica estándar de clasificación y representa la proporción de ejemplos correctamente clasificados.

4.1 Resultados

Model	AUC ROC	R2	RMSE	Accuracy
Valor más común	0.5	-0.249	0.446	0.800
DT	0.837	0.331	0.326	0.860
LR	0.840	0.344	0.323	0.865
LR + sesiones	0.848	0.357	0.320	0.866

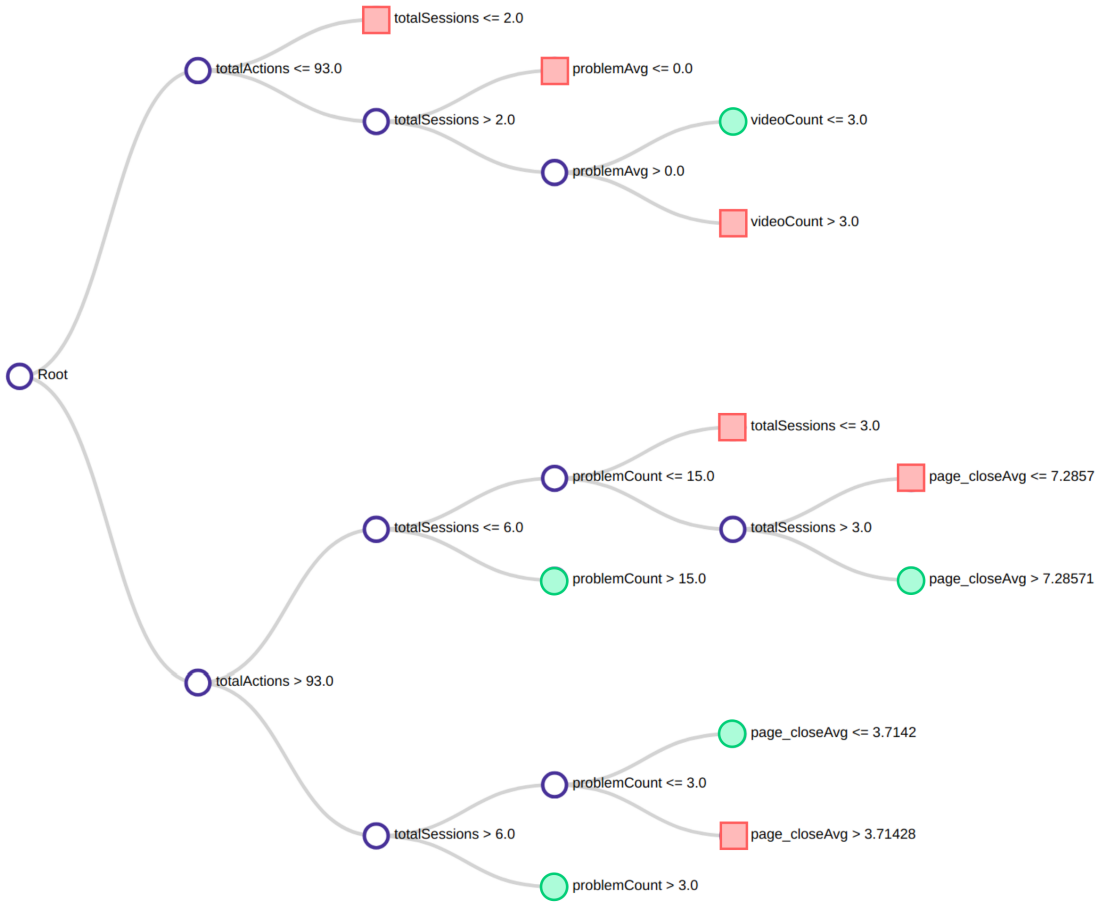
Table 1: Rendimiento de los modelos para predicción de deserción escolar, calculado en todos los cursos. LR significa Regresión Logística y DT significa Árbol de Decisión.

Los resultados se presentan en la Tabla 1. Podemos observar que todos los clasificadores funcionan significativamente mejor que utilizar el valor más común. Esto implica que el problema es efectivamente resulto por los modelos propuestos, y que los valores obtenidos no son simplemente producto del azar o la distribución de datos. El rendimiento del modelo de Regresión Logística sin incluir las características relacionadas con la sesión no disminuye significativamente, lo que indica que esas características no están codificando ninguna información nueva que resulte útil para el clasificador. Cabe destacar que el mejor rendimiento obtenido durante la competencia fue de 0.91, pero sobre un conjunto de datos de evaluación distinto. El modelo ganador fue un ensemble de distintos tipos de clasificadores. Consideramos que, utilizando tan poca información y un modelo tan simple, el resultado obtenido es prometedor.

El uso de Árboles de Decisión tampoco supera la Regresión Logística. Sin embargo, son clasificadores interpretables, ya que se basan en reglas de clasificación. Pudimos observar, a través de la inspección manual, que en general las características más importantes son el número de sesiones, el número de interacciones totales y el recuento de interacciones por tipo. El orden entre las tres características varía de un curso a otro. Presentamos una visualización de un Árbol de Decisión para el curso 6 en la Figura 6, donde se puede observar claramente este fenómeno. Hemos colapsado los nodos que predicen un solo valor en sus sub-ramas, coloreados de acuerdo al tipo de la predicción.

Finalmente, realizamos un análisis de desempeño en distintos cursos. Notamos en general que el rendimiento del clasificador disminuye con el tamaño del campo. Para los primeros cursos, de 0 a 5, no hay variación entre los clasificadores. Sin embargo, de muchos de los otros cursos, que tienen menos secuencias, la diferencia es más significativa. Con la excepción de los cursos 25, 26 y 27, la Regresión Logística obtiene mejores resultados que el Árbol de Decisión. Para resumir estos resultados en una visualización concisa, agruparemos los cursos de acuerdo a su tamaño, en las siguientes categorías: 5 cursos grandes con más de 6000

Figure 6: Visualización de las reglas internas de un Árbol de Decisión entrenado en el curso 6. Los nodos indican la característica y el valor utilizado para la decisión. Los nodos sombreados corresponden a los puntos de decisión, con cuadrados rojos para la deserción y círculos verdes para la etiqueta de no deserción.

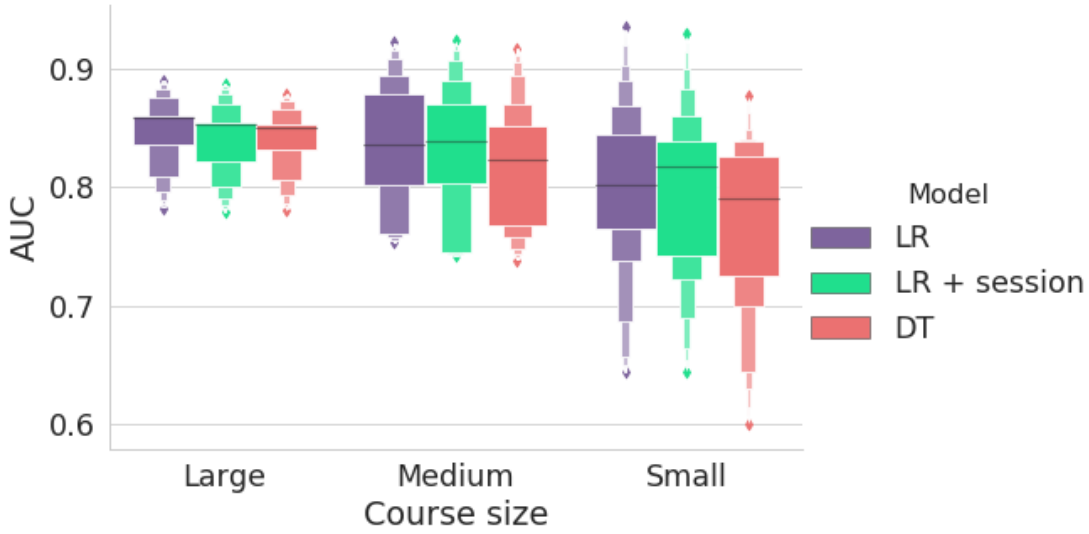


estudiantes de formación, 9 cursos medios con entre 5000 y 2000 estudiantes y 24 cursos pequeños con menos de 2000 estudiantes. En la Figura 7 la distribución obtenida por cada tipo de modelo según estas categorías.

Podemos ver que no solo la media de la distribución de resultados cae significativamente con cursos pequeños, sino también la amplitud de valores posibles. Esto quiere decir que, con pocos valores, el desempeño de este tipo de modelos podría ser tan bajo como 0.65 para LR y 0.6 para DT. Además de ello, la diferencia entre el clasificador LR con y sin información de sesión sólo es relevante cuando los cursos son pequeños, obteniendo resultados muy similares en los demás casos. Nuestra hipótesis para este comportamiento es que el análisis de sesiones incluye una mayor cantidad de información, lo cual sólo tiene impacto cuando el clasificador no tiene suficientes ejemplos para generalizar dichos patrones a partir de la representación más básica.

Asimismo, los Árboles de Decisión no están logrando buenos resultados para los cursos

Figure 7: Distribución de la métrica AUC ROC para los tipos de modelo propuestos, separadas por categorías de cursos



pequeños, donde la clasificación es más difícil y hay menos datos para construir reglas de decisión confiables.

5 Conclusiones

En este breve trabajo hemos descripto el conjunto de datos utilizado para la competencia KDDCup 2015 desde distintos puntos de vista. Podemos concluir que el fenómeno de la deserción en cursos online sigue en general principios intuitivos, estando estrechamente relacionado con el número y tipo de acciones realizadas por los estudiantes. Se han encontrado los siguientes patrones en los datos

- Los estudiantes que abandonan la escuela tienen menos interacciones. Sin embargo, esto no es suficiente para realizar una predicción adecuada, ya que hay un número significativo de ejemplos de secuencias largas con etiqueta de abandono.
- El tipo de interacciones está relacionado con el tipo de estudiantes. Los estudiantes que abandonan la escuela tienen más eventos de *access* y *navigate*, y los estudiantes que no abandonan la escuela tienen más eventos de *problem* y *discussion*.
- El número de sesiones por estudiante y los minutos pasados entre ellos tiene una distribución diferente entre la clase de deserción y la de no deserción.

Se han utilizado clasificadores lineales para resolver la tarea de predicción, obteniendo resultados mejores que un clasificador base, pero más bajos que los logrados en la competencia utilizando un modelo más complejo. A partir del análisis de desempeño de la Regresión

Logística podemos concluir que las características obtenidas con un análisis de sesiones mejoran la predicción, pero el impacto es menor. Utilizando una visualización de las características con más importancia utilizadas por los Árboles de Decisión confirmamos que los valores más importantes son el número total de acciones, el número total de sesiones y la cantidad de algunas acciones indicativas, como el acceso a problemas.

Al realizar un análisis por cursos, vemos que en general la tarea es más difícil en cursos pequeños, con menor cantidad de estudiantes. En estos casos, la Regresión Logística obtiene mejores resultados que los Árboles de Decisión, y el uso de características relacionadas a la sesión tiene más impacto en el desempeño final.