Expedientes del registrador universitario de minería para predecir el primer año Deserción de pregrado

Lovenoor Aulck
Universidad de Washington
laulck@uw.edu

dev nambi
Centro de Investigación del Cáncer FH
dnambi@fredhutch.org

Nishant Velagapudi
Universidad de Cal-Berkeley
nishray@berkeley.edu

Josué Blumenstock
Universidad de Cal-Berkeley
jblumenstock@berkeley.edu

jevin oeste
Universidad de Washington
jevinw@uw.edu

ABSTRACTO

Cada año, aproximadamente el 30% de los estudiantes de primer año en las instituciones de bachillerato de EE. UU. no regresan para su segundo año. y miles de millones de dólares se gastan en educar a estos estudiantes. Sin embargo, poca investigación cuantitativa ha analizado las causas y posibles remedios para la deserción estudiantil. Es más, la mayoría de los intentos previos de modelar el desgaste en los tradicionales Los campus que utilizan el aprendizaje automático se han centrado en grupos pequeños y homogéneos de estudiantes. En este trabajo, modelamos la deserción estudiantil utilizando un conjunto de datos que se compone casi exclusivamente de información recopilada de forma rutinaria para el mantenimiento de registros. en una gran universidad pública estadounidense. Al examinar la totalidad del alumnado de la universidad y no un subconjunto del mismo, usamos uno de los conjuntos de datos más grandes conocidos para examinar el trition en una universidad pública de EE. UU. (N = 66,060). Nuestros resultados muestran que la reinscripción de segundo año de los estudiantes y eventuales la graduación se puede predecir con precisión en función de un solo año de datos (AUROCs = 0.887 y 0.811, respectivamente). Encontramos que los datos demográficos (como raza, género, etc.) y datos previos a la admisión (como calificaciones académicas de la escuela secundaria, puntajes de exámenes de ingreso, etc.), según los cuales la mayoría de las admisiones los procesos están predicados - no son tan útiles como los primeros rendimiento universitario/datos de transcripciones para estas predicciones. Estos resultados resaltan el potencial de la minería de datos para impactar la retención y el éxito de los estudiantes en los campus tradicionales.

1. INTRODUCCIÓN

La deserción estudiantil ha sido durante mucho tiempo un tema de gran interés en investigación de educación superior, con informes gubernamentales sobre el desgaste que se remontan a más de 100 años [31]. Este interés se deriva del hecho de que los estudiantes que no se gradúan son una inversión perdida en muchos frentes. Para las instituciones de educación superior, limitar el desgaste es fundamental para su sostenibilidad financiera, ya que dedican recursos escasos a clases y servicios para estudiantes que no terminan [17]. En particular, se estima que el 30% de los estudiantes de primer año de los Estados Unidos (EE. UU.) no regresan para su segundo año de educación postsecundaria con Los contribuyentes de EE. UU. gastan casi $ 2 mil millones anuales solo en la educación de los estudiantes de primer año que no regresan [28]. Instituciones también están preocupados por las tasas de deserción porque son fundamentales para las estimaciones de la eficacia institucional, lo que afecta las oportunidades de financiación y el apoyo del gobierno [14].

Resaltar el impacto de la deserción a nivel institucional tampoco dice nada de su impacto en los estudiantes, que dedican tiempo, esfuerzo y las finanzas hacia actividades educativas inconclusas. Abandonar la universidad altera drásticamente las trayectorias profesionales de los estudiantes y aquellos que no tienen un título universitario se enfrentan a un crecimiento laboral en declive continuo y a un empeoramiento de las perspectivas laborales [9].

A la luz de esto, comprender las motivaciones de los estudiantes para la deserción y los posibles remedios de la misma es de gran importancia [12]. Evidencia empírica para construir la teoría de la deserción estudiantil se ha centrado tradicionalmente en la investigación basada en encuestas [30, 8]. Sin embargo, los instrumentos de encuesta suelen ser costosos de implementar, requiere mucho tiempo para la recopilación de datos y produce resultados que no siempre son generalizables entre instituciones debido a diferentes perfiles de estudiantes [34, 7, 8]. Datos institucionales que son recolectados rutinariamente en colegios y universidades (por ejemplo, estudiante solicitud y datos de transcripción) puede proporcionar una alternativa fuente de datos y una forma de complementar las medidas basadas en encuestas [8]. Aprovechar las fuentes de datos que ya existen puede agregar un medio para examinar más eficientemente la deserción de los estudiantes problema y ayudar a las instituciones a remediar el problema del desgaste.

Un campo que está preparado para aprovechar estos datos institucionales es la minería de datos educativos (EDM) y su enfoque sobre técnicas intensivas en datos en entornos educativos [26, 4].

EDM es un campo emergente con gran parte de su investigación sobre el desgaste centrada en cursos abiertos masivos en línea (MOOC) y otros entornos en línea (por ejemplo, [35, 13]). Estudiar a tiempo en MOOC y otros entornos en línea se presta a amplias oportunidades de recopilación de datos y un seguimiento detallado de los estudiantes [23]. Esto limita la medida en que este el trabajo se puede generalizar a entornos de campus más tradicionales (es decir, campus donde el aprendizaje es principalmente en el campus, en el salon de clases). Mientras tanto, el trabajo centrado en EDM sobre la predicción la deserción en los campus tradicionales ha sido escasa y, por lo general, limitada a subconjuntos pequeños y homogéneos de estudiantes en lugar de que la totalidad de una población de estudiantes universitarios. Además, el enfoque cuando se predice la deserción suele ser cómo bueno, se puede predecir y menos sobre qué tipo de datos son mejor para estas predicciones.

En este trabajo, predecimos el desgaste de un gran número de un

Lovenoor Aulck, Dev Nambi, Nishant Velagapudi, Joshua Blumenstock y Jevin West "Registrador de la Universidad Minera Registros para predecir la deserción de estudiantes universitarios de primer año" En: *The 12ª Conferencia Internacional sobre Minería de Datos Educativos,* Michel Desmarais, Collin F. Lynch, Agathe Merceron y Roger Nkambou (eds.) 2019, págs. 9 - 18

estudiantes de pregrado (N = 66,060) usando solo su primer año de datos académicos. Los estudiantes que examinamos no son de un solo departamento o especialización dentro de una universidad. Más bien, abarcan la totalidad del alumnado, por lo que comprenden un conjunto de datos con aspiraciones, antecedentes y objetivos heterogéneos. Además, dependemos casi por completo de los datos que se recopilan de forma rutinaria en las instituciones de educación superior. Con estos datos, buscamos responder dos preguntas: en qué medida se puede predecir la deserción de estudiantes de pregrado utilizando una cantidad limitada de datos de registros de registradores y qué tipos de datos de registros de registradores son más útiles para predecir la deserción. El primero de ellos se ha explorado en el pasado utilizando poblaciones de estudiantes más pequeñas y/u homogéneas; el segundo no ha sido examinado sistemáticamente en la literatura hasta donde sabemos.

Para responder a las preguntas anteriores, extraemos los registros de datos institucionales en una gran universidad pública de los EE. UU. y creamos características para las predicciones. Luego, creamos numerosos modelos de aprendizaje automático utilizando las funciones de ingeniería y comparamos el rendimiento de estos modelos entre sí. Luego, creamos modelos de aprendizaje automático separados usando solo grupos de características y no la totalidad del espacio de características para comparar el poder predictivo de diferentes subconjuntos de datos institucionales. Este trabajo es una extensión de nuestro trabajo anterior sobre el modelado de la deserción estudiantil usando una cantidad limitada de datos [3] pero donde anteriormente nos enfocamos en usar los datos del primer término para generar características para la predicción, usamos los del primer año en este trabajo. También ampliamos nuestro trabajo anterior para crear modelos adicionales de aprendizaje automático, predecir la deserción según dos definiciones diferentes (graduación general y reinscripción después del primer año de los estudiantes) y examinar los tipos de subconjuntos de características más útiles en las predicciones. Al hacerlo, presentamos dos hallazgos clave, los cuales tienen muchas implicaciones para la política administrativa en la educación superior:

- Demostramos que la graduación y reinscripción de segundo año de los estudiantes se puede predecir utilizando datos que se recopilan de forma rutinaria en las instituciones de educación superior.
- Mostramos que las características demográficas y previas al ingreso tienen menos poder predictivo que los datos académicos de los estudiantes.

2. TRABAJO RELACIONADO Hay muchos ejemplos de predicción de la deserción en los campus tradicionales. La mayoría de estos se enfocan en subconjuntos pequeños y homogéneos de estudiantes. Moseley predijo la graduación de 528 estudiantes de enfermería usando métodos de inducción de reglas, obteniendo altas precisiones pero sin controlar el número de trimestres/semestres examinados para cada estudiante [21]. Dekker et al observaron solo las calificaciones del primer semestre de 648 estudiantes en el departamento de Ingeniería Eléctrica de la Universidad Tecnológica de Eindhoven y pudieron predecir la deserción con una precisión del 75-80% [10]. Kovaÿcí´c usó métodos basados en árboles en un conjunto de datos de tamaño similar de 453 estudiantes en el Politécnico Abierto de Nueva Zelanda, encontrando que el origen étnico y los patrones de toma de cursos de los estudiantes eran muy útiles en la predicción [18]. Bayer et al. observó a 775 estudiantes de informática aplicada en la Universidad Masaryk de la República Checa durante tres años [5]. Sin limitar la cantidad de información disponible para cada estudiante, descubrieron que incluir características relacionadas con el comportamiento social de los estudiantes puede aumentar la precisión de la predicción en más del 10 % para algunos modelos. Estos y otros estudios similares ayudan

Sin embargo, concéntrese en subgrupos relativamente pequeños (p. ej., N < 2000) de estudiantes con objetivos/enfoques académicos similares. Además, hay poca consistencia con respecto a los marcos de tiempo en los que se examinan los datos de cada estudiante. Otros enfoques para predecir la deserción en los campus tradicionales incluyen los sistemas de alerta temprana, que a menudo requieren mucha mano de obra y están mal financiados [29]. Se ha demostrado que estos sistemas de alerta benefician positivamente a los estudiantes (p. ej., [16]), pero generalmente se basan en datos recopilados en medio de un curso o un período académico (p. ej., [27, 15]), lo que no siempre es factible.

El trabajo que presentamos se relaciona más estrechamente con un subconjunto de literatura que analiza la deserción estudiantil en el contexto de la heterogeneidad de los estudiantes en todo un campus y no solo en un subconjunto del mismo. Nuestro trabajo también trata con poblaciones estudiantiles mucho más grandes que las descritas anteriormente y, en este sentido, se parece más a un cuerpo de literatura más reciente. Delen usó 8 años de datos institucionales sobre más de 25 000 estudiantes en una gran universidad pública de EE. UU. y predijo si los estudiantes regresarían para su segundo año [11]. Sin embargo, debido a los desequilibrios de clase, Delen volvió a muestrear la clase mayoritaria y, en última instancia, utilizó solo 6454 estudiantes para las predicciones. Rama et al. usó datos de aproximadamente 6500 estudiantes de primer año en una gran universidad pública de EE. UU. para predecir si los estudiantes abandonarían después de su primer semestre y, para aquellos que no lo hicieron, si abandonarían después de un período adicional [25]. Rama et al. Complementó los datos de las bases de datos institucionales con transacciones de tarjetas inteligentes de estudiantes para inferir la integración social. Más recientemente, Nagy y Molontay predijeron con cierto éxito la deserción de 15.825 estudiantes de la Universidad de Tecnología y Economía de Budapest utilizando solo su información previa al ingreso a la universidad [22].

Hay algunas formas en las que nuestro trabajo contribuye a este cuerpo de literatura. En primer lugar, utilizamos un conjunto de datos mucho más grande que el que se ha examinado previamente específicamente para la deserción (66 060 estudiantes). Examinamos la totalidad del alumnado de una gran universidad y no limitamos el grado de heterogeneidad de los estudiantes en el conjunto de datos. Además, también abordamos la cuestión de qué tipos de características son más útiles para predecir la deserción de los estudiantes. En particular, los trabajos anteriores generalmente han utilizado todas las fuentes de datos disponibles al mismo tiempo para determinar qué estudiantes abandonarán. En este trabajo, exploramos qué tipos de datos institucionales recopilados de forma rutinaria funcionan mejor al predecir el desgaste al comparar el desempeño utilizando diferentes subconjuntos de datos de forma aislada. Finalmente, comparamos al mismo tiempo las predicciones para dos definiciones diferentes de "deserción", destacando el grado en que la operacionalización del término puede afectar los resultados

3. MÉTODOS

Describimos los métodos para este trabajo detallando primero los datos utilizados en el proyecto. Luego damos definiciones operativas relevantes con respecto a cómo definimos el desgaste. Luego, discutimos los subconjuntos de datos utilizados en las predicciones y las características generadas. Por último, describimos la configuración de los experimentos de aprendizaje automático.

3.1 Descripción de los datos

Recopilamos datos anonimizados y pseudonomizados de los administradores de datos de la Universidad de Washington (la Universidad) en 2017. La Universidad es un campus tradicional donde la mayoría de la instrucción es en persona y cara a cara. No

se recopiló información de identificación personal de los estudiantes; en cambio, se hizo referencia a los estudiantes usando claves de identificación únicas. La tabla 1 muestra las tablas que se extrajeron de las bases de datos del registrador. En general, los datos incluidos en la formación sobre la demografía de los estudiantes, transcripción completa registros en la Universidad, e información de aplicaciones a la Universidad. No teníamos ninguna información sobre el estado de ayuda financiera o el estado económico de los estudiantes aparte de eso que se derivó de su código postal, como se describe a continuación. Los factores socioeconómicos pueden jugar un papel importante en el estudiante proceso de deserción [6], sin embargo, no tuvimos acceso a las finanzas de los estudiantes para usar en este trabajo. tampoco tuvimos acceso a las encuestas de salida de los estudiantes que se habían ido la Universidad o se había graduado.

Tabla 1: Datos extraídos de bases de datos de registradores

Mesa	Descripción
Datos de la solicitud	Información de las solicitudes de los estudiantes a la Universidad, incluidos cursos de secundaria
Datos del guardián	Información sobre los tutores de los estudiantes extraído de las solicitudes de los estudiantes a la Universidad
Datos demográficos	Información sobre la demografía de los estudiantes, incluida la fecha de nacimiento, raza, etnia, sexo, etc.
Datos principales	Información sobre carreras declaradas por los estudiantes trimestre a trimestre (trimestre por trimestre)
Datos de puntaje de prueba	Información sobre los resultados de las pruebas estandarizadas de los estudiantes
Datos de la transcripción	Información sobre el trabajo del estudiante y calificaciones término a término (trimestre por trimestre)

Restringimos los datos a los graduados de la escuela secundaria que se matricularon por primera vez en la Universidad como estudiantes universitarios matriculados que buscaban un título de licenciatura entre 1998 y 2010 sin haber asistido previamente a otro postsecundario institución a tiempo completo. Estos estudiantes son referidos en lo sucesivo como "estudiantes de primer año". El conjunto de datos incluía a estudiantes que estaban en un programa universitario en la escuela secundaria, pero excluyó a aquellos que asistieron a la universidad comunitaria/junior a tiempo completo después de la escuela secundaria y luego transferido a la Universidad. porque los datos eran extraído en 2017, usamos el año 2010 como punto de corte para permitir seis años completos de visibilidad sobre los estudios académicos de los estudiantes en la Universidad antes de etiquetar a un estudiante como “no terminado”, como se define en la Sección 3.2. En total, el conjunto de datos constaba de 66.060 participantes únicos de primer año. Luego limitamos aún más los datos para cada estudiante a la información a través de un año calendario desde la primera matrícula de cada alumno en la Universidad. Esta los datos se limitaron a un año calendario para todos los estudiantes, independientemente de la cantidad de cursos que tomaron/aprobaron, su calificaciones o sus antecedentes.

Después de unir tablas de interés usando los identificadores únicos de estudiantes, creamos características para los experimentos de predicción al ya sea extrayéndolos directamente de los datos sin procesar o ingeniándolos para cada estudiante. Las características se agruparon en 7 agrupaciones, que se describen en la Sección 3.3; una lista completa de características y descripciones de las mismas está disponible en solicitud, pero no se proporcionó en este escrito en interés del espacio. En total, había 1.405 características y todas las características se generaron para cada alumno sin excepción.

3.2 Definiciones

Ambigüedad con respecto a las definiciones operativas de deserción en la literatura sobre la deserción estudiantil puede dificultar la comparación de resultados entre estudios [24, 33]. Hay numerosas formas en el que la deserción se ha definido en la literatura existente, ser estudiantes que abandonan un curso en particular (p. ej. [21]), reinscribirse después de su primer período (por ejemplo, [1]), reinscribirse después su primer año (p. ej., [11]), graduarse a tiempo (p. ej., [3]), o alcanzar algún otro hito relevante (por ejemplo, [10]). En esto trabajo, definimos la deserción de dos maneras y analizamos ambas. Nosotros examinó la deserción de los estudiantes desde el primer año hasta el segundo ("reinscripción" y "no reinscripción"), así como buscar en si un estudiante se graduó a tiempo ("graduado" y “incompleto”). No examinamos la deserción término por término debido a los relativamente pocos estudiantes que dejar la Universidad después de un solo término, como se discutió en Sección 4.1. Definimos operativamente el incumplimiento y reinscripción como se describe a continuación.

3.2.1 Incumplimiento

Definimos "no culminación" como cualquier estudiante de primer año que no se graduó con un título de licenciatura de la Universidad dentro de los 6 años calendario posteriores al primer ingreso a la Universidad. Definimos un "graduado" como un estudiante de primer año que se graduó de la Universidad con un título de licenciatura dentro de los 6 años calendario posteriores a la primera inscripción. La Universidad usa un sistema de trimestres y nosotros usamos el lapso de cuatro trimestres consecutivos. trimestres académicos como una medida de un año calendario. Seis años calendario para la graduación fue, por lo tanto, el lapso de 24 trimestres académicos consecutivos. Esta definición de incumplimiento sólo se contabilizó el primer grado de bachillerato de los estudiantes y no tuvo en cuenta las dobles carreras o las dobles titulaciones. Por ejemplo, si un estudiante estaba cursando simultáneamente dos títulos de bachillerato pero solo se graduó con uno de cada cinco años, sería un graduado; alternativamente, si el estudiante se hubiera graduado con ambos grados pero durante el séptimo año, se consideraría como no culminado. Debido a que nos enfocamos en los registros del registrador de una sola institución, definir la falta de finalización de esta manera no toma en cuenta la progresión académica de los estudiantes después de dejar el Universidad. Esto se debe a que solo teníamos acceso al registrador. registros de una sola institución y no rastrear a los estudiantes a través de múltiples instituciones, podrían haberse transferido muy bien de la Universidad y haberse graduado con buena reputación.

Tomamos en cuenta a los estudiantes que participaron en una universidad en la escuela secundaria programa escolar convirtiendo el total de créditos transferidos a un conteo de trimestres académicos completados asumiendo matrícula típica de tiempo completo en la Universidad. Por ejemplo, si un estudiante completó 30 créditos en una universidad en la escuela secundaria programa, convertimos este total de crédito en un recuento de términos completados en la Universidad (en este caso, 2, ya que los estudiantes típicamente toman 15 créditos por término). Redondeamos el resultado de esta conversión en su caso. Luego deducimos esto número al determinar si el estudiante se había graduado dentro de un período de tiempo apropiado.

3.2.2 Reinscripción

Definimos "reinscripción" como un estudiante que completó al menos un curso adicional dentro de un año calendario desde el final de su primer año calendario en la Universidad (es decir, dentro de los 4 trimestres académicos desde el final de su primer año). Los “no reinscritos” eran estudiantes que no eran reinscritos.

En este trabajo, las definiciones de graduación y reinscripción se trataron mutuamente excluyentes en el sentido de que todos los graduados no eran necesariamente reinscritos. Cabe señalar que la Universidad requiere que los estudiantes que no se matriculen por dos períodos consecutivos sin permiso justificado sean readmitidos.
a criterio de la Universidad.

3.3 Agrupaciones de características

Para cada estudiante, diseñamos los subconjuntos de características que se describen a continuación. Para todas las calificaciones de los estudiantes, calculamos un percentil de calificación y una puntuación z comparando las calificaciones de cada estudiante con las calificaciones de todos los estudiantes de pregrado que habían tomado el mismo curso al mismo tiempo. Las referencias a las calificaciones incluyen el GPA del estudiante (en una escala de 4.0), su puntaje percentil (de 0 a 100) y su puntaje z para los cursos (que representa el número de desviaciones estándar de la media, suponiendo una distribución de calificaciones normal). Las referencias al "desempeño" para las agrupaciones de funciones incluyen calificaciones y créditos obtenidos, como mínimo. En algunos casos, las referencias al desempeño también pueden incluir la cantidad de créditos calificados obtenidos (en comparación con los cursos aprobados o reprobados) y la cantidad de créditos intentados. En la Tabla 2 se proporciona una breve descripción de cada uno de los subconjuntos de funciones.

Tabla 2: Subconjuntos de datos utilizados en las predicciones

Subconjunto	Descripción
Datos básicos	Año y trimestre de ingreso a la Universidad (incluido con todos los demás subconjuntos de datos)
Demográfico Datos	Datos no académicos previos al ingreso a la Universidad, incluida la demografía
Datos a nivel de departamento	Medidas de desempeño agregadas por departamento del curso
Datos resumidos del primer año	Medidas agregadas de rendimiento académico durante el primer año
Curso agrupado Datos	Medidas de desempeño agregadas por número de curso y guardianes STEM
Datos principales	Recuentos de mayores declarados término a término
Datos Pre-Ingreso	Datos académicos previos al ingreso a la Universidad

3.3.1 Base de datos

Los datos de la base constaban de solo tres características y se incluyeron en el espacio de características al hacer predicciones utilizando todos los demás subconjuntos de datos descritos. Los datos base incluían el año calendario de ingreso a la universidad de los estudiantes, su trimestre de ingreso a la universidad (es decir, cuál de los cuatro trimestres académicos fue el primero de un estudiante; que van del 1 al 4, con 1, 2, 3 y 4 correspondientes a los trimestres académicos de invierno, primavera, verano y otoño, respectivamente), y una variable de trimestre que consistía en el año de ingreso de los estudiantes multiplicado por 4 y sumado al trimestre de ingreso para crear una variable relativa escala de tiempo Estas características se incluyeron para tener en cuenta cualquier variación relacionada con el tiempo en las tasas de graduación.

3.3.2 Datos demográficos

Los datos demográficos consisten en la información no académica del estudiante antes de ingresar a la Universidad. Esto incluía, entre otros, el sexo, la raza, el origen étnico, la edad de inscripción en la universidad, el estado de veterano y el estado de estudiante atleta de los estudiantes. También incluimos información de la aplicación de los estudiantes a la Universidad, como información sobre las escuelas secundarias de los estudiantes (excluyendo los grados de la escuela secundaria), el nivel educativo de los padres y el código postal de los estudiantes, que se extrajo de su información de la escuela secundaria o, cuando no esté disponible, de su solicitud universitaria. Unimos los códigos postales de los estudiantes con los datos del censo de EE. UU. de 20151 para encontrar el ingreso promedio y el nivel educativo en cada código postal. También incluimos la distancia desde la Universidad hasta el código postal de la casa de cada estudiante. Las características derivadas de los códigos postales fueron las únicas características de fuentes externas a las bases de datos de registro de la Universidad.

3.3.3 Datos a nivel de departamento

Los datos a nivel de departamento consisten en el desempeño de los estudiantes en las ofertas de cursos agrupados por prefijo de curso. Por ejemplo, esto incluía el rendimiento en todos los cursos de BIOL (biología) agrupados, el rendimiento en todos los cursos de HIST (historia) agrupados, etc. Excluimos los prefijos de cursos en los que al menos 10 estudiantes del conjunto de datos no tomaron un curso. En total, esto incluyó 200 prefijos de cursos únicos y 1000 características, con GPA, calificación percentil, puntaje z, créditos obtenidos y créditos calificados obtenidos calculados para cada prefijo. Usamos datos a nivel de departamento en lugar de datos de cursos individuales después de que el modelado preliminar usando cursos individuales no arrojara resultados sólidos. El amplio espacio de funciones cuando se diseñaron funciones en cursos individuales también aumentó significativamente la potencia/tiempo de cómputo necesarios para el modelado y decidimos no continuar con esto.

3.3.4 Datos resumidos del primer año

Los datos resumidos del primer año consistieron en medidas agregadas del primer año de los estudiantes en la universidad. Esto incluía, entre otras cosas, el desempeño de los estudiantes en los cursos, los créditos tomados, la cantidad de cursos reprobados, la cantidad de trimestres inscritos y la inscripción en cursos de seminario de primer año. Los datos resumidos del primer año también incluyeron medidas agregadas del desempeño de los estudiantes en su primer, segundo, tercer y cuarto trimestre, así como el desempeño de los estudiantes en el último trimestre académico en el que se inscribieron durante su primer año (independientemente de qué trimestre era). También incluimos las diferencias entre el desempeño de los estudiantes en trimestres sucesivos

3.3.5 Datos del curso agrupados

Los datos del curso agrupados consisten en el rendimiento del curso del estudiante agrupado por número de curso o por rendimiento en "guardianes STEM". Para agrupar los cursos por número de curso, agregamos el rendimiento de todos los cursos numerados por debajo de 100, del 100 al 199, del 200 al 299, del 300 al 399 y del 400 en adelante. La numeración del curso generalmente reflejaba si el curso estaba diseñado para ser tomado por hombres de clase baja o alta y, en algunos casos, también indicaba

1Del American Fact Finder de la Oficina del Censo de EE. UU.

durante qué año los estudiantes típicamente tomaron el curso. MADRE
los guardianes se refieren a cursos de introducción a la ciencia, la tecnología, la
ingeniería y las matemáticas (STEM, por sus siglas en inglés) que a menudo funcionan como
requisitos previos para carreras y títulos STEM. Estos cursos de guardianes tienden
a ser altamente competitivos y el rendimiento en estos cursos es un determinante
clave de si un
el estudiante será aceptado en cualquiera de los altamente competitivos
carreras STEM. Agrupamos el desempeño en STEM gatekeepers por departamento
del curso y tema (por ejemplo, el cálculo
serie, la serie de química general, la química orgánica
serie, etc.), así como en todos los guardianes de STEM.

3.3.6 Datos principales

Los datos principales consistieron en recuentos de las principales declaraciones de
los estudiantes durante su primer año académico. En la mayoría de los casos, los
estudiantes ingresaron a la Universidad con una designación de "pre-major"
antes de declarar su(s) especialidad(es) de interés en algún momento durante su
primer o segundo año. Estas designaciones anteriores a las principales
varió según el campo de interés (por ejemplo, preingeniería, preenfermería, presalud,
etc.). Se registraron las carreras de los estudiantes
trimestralmente por la Universidad (una vez por trimestre
registro de transcripciones) y contamos los recuentos de las principales declaraciones
de cada estudiante durante la totalidad de su primer año.
Por ejemplo, un estudiante que declaró una especialización en matemáticas en su
primeros dos cuartos solo para cambiar a geografía en su tercero
cuarto y luego agregar una doble especialización en historia en su cuarto
trimestre tendría los valores 2, 2, 1 en la especialidad de matemáticas,
características principales de geografía e historia, respectivamente.

3.3.7 Datos previos a la entrada

Los datos previos al ingreso consistieron en la información académica de los estudiantes.
antes de asistir a la Universidad. Esto incluía, entre
otras cosas, puntajes del examen de ingreso de los estudiantes, escuela secundaria
GPA, cursos de la escuela secundaria y universidad en la escuela secundaria
participación y desempeño del programa. no incluimos
cualquier información sobre los estudiantes después de su inscripción en el
Universidad en los datos de pre-entrada.

3.4 Aprendizaje automático y predicciones

Dividimos aleatoriamente a los estudiantes en entrenamiento y prueba.
conjuntos usando una división 80-20 (N en entrenamiento = 52,848; N en prueba =
13,212). Utilizamos el mismo conjunto de pruebas al evaluar el rendimiento predictivo
de cada uno de los modelos para permitir la evaluación directa.
comparaciones a realizar. Los datos estaban muy sesgados con
graduados y reingresos que comprenden el 78,5% y el 93,1%
de todos los datos, respectivamente. Graduados y Reingresos
comprendió el 78,0% y el 92,9% de los datos de prueba, respectivamente.
Aunque tratar los desequilibrios de clase es de gran interés
al examinar la deserción de los estudiantes de primer año [32], no utilizamos ninguna
técnicas de equilibrio ya que queríamos trabajar con los datos en
su forma original, inalterada. Escalamos los datos de entrenamiento por
restando la mediana de cada característica y dividiendo por la
rango intercuartílico de la característica respectiva. Posteriormente
escaló los datos de prueba usando los valores de escala para cada característica
de los datos de entrenamiento.

Utilizamos cinco modelos diferentes de aprendizaje automático para predecir
graduación y reinscripción de cada estudiante: regresión logística regularizada (LR),
K-vecinos más cercanos (KNN), aleatorio
bosques (RF), máquinas de vectores de soporte (SVM) y árboles potenciados por
gradientes (XGB). Entrenamos a cada modelo en todo el
la totalidad de los datos de entrenamiento y usó el mismo entrenamiento
instancias para entrenar cada uno de los modelos. Entrenamos a cada uno
modelo por separado para predecir la graduación y la reinscripción.
Ajustamos los hiperparámetros del modelo para cada modelo usando 5-
doblar la validación cruzada en los datos de entrenamiento, después de lo cual el
los modelos se volvieron a entrenar en la totalidad de los datos de entrenamiento
utilizando los hiperparámetros sintonizados. Informamos las métricas de error finales
y el rendimiento en el conjunto de prueba, que fue consistente
en todos los modelos, independientemente de si predicen la graduación o la
reinscripción.

Después de desarrollar modelos predictivos usando todas las características, creamos
modelos de regresión logística regularizados usando cada una de las características.
6 subconjuntos de características destacados en la Sección 3.3 de forma aislada. El
Los datos base (ver Tabla 2) se incluyeron en el espacio de características para
cada subconjunto de datos. La justificación detrás del uso de la regresión logística
regularizada para estos modelos se analiza con más detalle en la Sección 4.3.
Entendemos que un enfoque alternativo
ser probar todos los modelos enumerados anteriormente para cada uno de los datos sub
conjuntos para encontrar las mejores combinaciones de modelo/subconjunto.
Dicho esto, creemos que nuestro enfoque seguía siendo adecuado para comparar
diferentes subconjuntos de datos. Al modelar usando datos
subconjuntos, usamos las mismas observaciones que antes para entrenar
cada uno de los modelos y, como antes, desarrollamos un
modelo para predecir la graduación y la reinscripción para cada
de los subconjuntos de datos. Así, las instancias de formación fueron
lo mismo en todos los modelos, pero las características de entrenamiento diferían
dependiendo del subconjunto de funciones utilizado. Ajustamos la fuerza de
regularización para estas regresiones logísticas regularizadas.
modelos que usan validación cruzada de 5 veces en el conjunto de datos de entrenamiento
e informamos los resultados en el conjunto de prueba.

4. RESULTADOS Y DISCUSIÓN

4.1 Características de los estudiantes

Mostramos el número y la proporción de graduados y reinscritos en la Figura 1. En
total, el 78,5 % de los estudiantes fueron etiquetados como graduados, mientras que
el 93,1 % de los estudiantes fueron etiquetados como reinscritos. Estas proporciones
fueron verificadas con la oficina de análisis institucional de la Universidad. Tan
altamente sesgado
se pueden esperar datos sobre graduados y reinscripciones
en un gran entorno universitario de investigación de nivel 1 donde ha
Ha habido un esfuerzo considerable y de larga data para mejorar la tasa general de
deserción a lo largo del tiempo. Dicho esto, también hay que señalar
que en una institución con una población estudiantil tan grande,
incluso pequeñas fracciones del alumnado representan cientos
de estudiantes en forma anual. A lo largo de la línea de tiempo de la
conjunto de datos (13 cohortes), 14.196 no finalizaciones y 4.593 no reinscripciones
representan 1.092 y 351 estudiantes sobre una base anual, respectivamente.

Mostramos el porcentaje acumulado de estudiantes que
se graduó o dejó la Universidad a lo largo del tiempo en la Figura 2. Nosotros
usó el primer año como punto de corte para los datos porque, históricamente, una
gran cantidad de estudiantes deciden si
continuar con sus actividades de educación superior durante y
inmediatamente después de su primer año [28]. Como tal, desarrollar
modelos que pueden predecir si los estudiantes se reinscribirán para
un segundo año y si están en una trayectoria hacia
una graduación exitosa podría ayudar a los administradores y asesores académicos
a desarrollar y brindar intervenciones dirigidas a los estudiantes que necesitan
asistencia de manera más eficaz. Cuándo
examinando los datos, el 27,5% de todos los incumplimientos abandonan el
universidad antes del inicio de su 2º año, el 51,9% de los que no finalizan abandonan
la Universidad entre su 2º y 6º

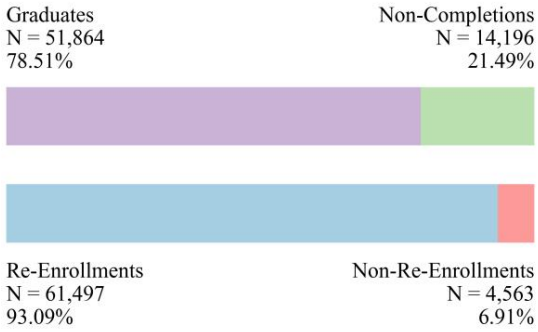

Figura 1: Conteos y porcentajes de clases en el
conjunto de datos Las definiciones se proporcionan en la Sección 3.2.

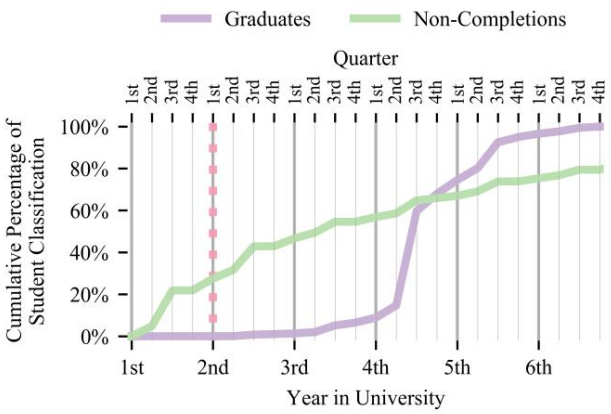

Figura 2: Curvas acumuladas de graduación y no finalización de los estudiantes. Años y trimestres
son relativos al momento de la primera inscripción. La línea punteada indica
el punto al que se limitan los datos
para cada estudiante. Solo los primeros seis años de los estudiantes son
que se muestra, según la definición de "graduado".

año, y el 20,6% continuaba matriculado en la Universidad
después de su sexto año. La diferencia en el número entre los no completados
que no regresaron para su segundo año y los no reinscritos se puede atribuir
a los no reinscritos que
más tarde regresó a la Universidad y se graduó a tiempo. Menos
del 5% de los que no completaron y menos del 15% de los que no reingresaron
abandonaron la Universidad después de solo un período, lo que lleva
no examinemos la deserción después del primer y segundo término.
En entornos donde las tasas de deserción son más altas después de que los estudiantes
primer y segundo términos, puede ser más relevante examinar
el desempeño de clasificadores después de uno o dos términos.

La Figura 2 también muestra que la mayoría de los graduados (65.6%)
completaron sus títulos durante su cuarto año en la Uni versidad. El tiempo de
finalización medio y mediano para todos los graduados fue de 16,6 y 15,0
trimestres calendario, respectivamente, de
primera inscripción. Esto es particularmente evidente debido a la
forma casi sigmoidea del gráfico acumulativo para graduados,
con un fuerte aumento durante el cuarto año de los estudiantes. también vemos
que hay una relativa falta de estudiantes que se graduaron antes
al comienzo de su tercer año. Esto pone de manifiesto la dificultad
en la predicción de la graduación basada en el primer año de los estudiantes - un
el estudiante normalmente no se gradúa hasta varios años después,
durante el cual una gran cantidad de influencias pueden dar forma a un académico
trayectoria, ya sea personal, financiera o académica.

4.2 Predicciones usando diferentes algoritmos

Tabla 3: Resultados de predicción utilizando todas las funciones de datos.
Los valores de referencia se basan en el conjunto de pruebas.

	Graduación		Reinscripción	
Modelo	Precisión	AUROC	Precisión	AUROC
Base	78,0%	0.500	92,9%	0.500
LR	83,2%	0.811	95,0%	0.882
RF	83,1%	0,806	95,3%	0.887
XGB	83,0%	0,806	95,1%	0.885
KNN	82,5%	0,798	94,8%	0.876
MVS	78,0%	0,780	92,9%	0.862

Mostramos el rendimiento de cada uno de los modelos utilizando la totalidad
del espacio de características en la Tabla 3. La medida de referencia
en la Tabla se refiere a las composiciones de clase mayoritarias en el
equipo de prueba. En términos generales, la mayoría de los modelos tuvieron
un rendimiento comparativo similar para cada tarea de predicción (es decir,
predecir la graduación o la reinscripción). Esto sugiere
un techo efectivo con respecto al poder predictivo de la
tipos de funciones que se utilizan (es decir, las extraídas del registrador)
expedientes) y que representaciones adicionales del estudiante
se debe incorporar la experiencia (ya sea académica o social). Alternativamente,
un modelo predictivo más complejo (por ejemplo,
redes neuronales profundas) también pueden obtener mejores resultados al hacer que estos
predicciones Dicho esto, dados los datos utilizados, los modelos son
capaz de predecir la eventual graduación y reinscripción
de los estudiantes con bastante éxito, como lo demuestra el relativo
mejoras sobre los valores de referencia para ambas tareas de predicción.

Para predecir la graduación, la regresión logística fue el modelo con mejor
rendimiento, seguido de los bosques aleatorios. Al predecir la reinscripción,
los bosques aleatorios se desempeñaron mejor,
seguido de árboles potenciados por gradiente y regresión logística.
Estos resultados están generalmente en línea con nuestro trabajo previo sobre
tareas similares, donde encontramos que la regresión logística tiende
funcionar bien en comparación con otros modelos para predecir la graduación
y el abandono de STEM [2]. Al examinar los modelos de peor desempeño, el
modelo SVM hizo predicciones que
consistía en su totalidad en la clase mayoritaria al predecir tanto
graduación y reinscripción, como se ve por la precisión de los modelos que es
igual a los valores de referencia. Tales resultados son
típico de clasificadores sin mucha fuerza predictiva en un
conjunto de datos que consiste en clases altamente desproporcionadas. En esto
caso específico, puede remediarse usando núcleos alternativos
para el modelo, que no exploramos en este trabajo.

Mostramos las curvas ROC para los modelos en la Figura 3.
Estas curvas ilustran aún más la falta de diferenciación
con respecto al rendimiento del modelo. Para la misma tarea de predicción,
las curvas ROC resultantes en los modelos fueron casi idénticas con poca
diferencia en la curvatura.
La diferencia más notable fue al comparar el ROC
curvas para predecir la graduación con aquellas para predecir la reinscripción,
ya que las curvas para predecir la reinscripción fueron
más prominentemente convexo en comparación con aquellos para predecir
graduación. Estas curvaturas, junto con las métricas que se muestran

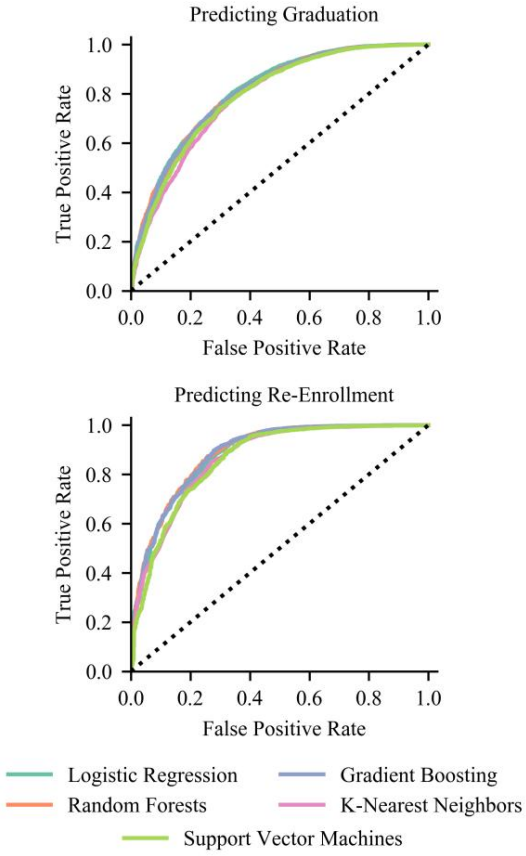

Figura 3: Curvas características de funcionamiento del receptor cuando se utilizan diferentes modelos de aprendizaje automático.

en la Tabla 3, demuestre que la predicción de los estudiantes eventuales la graduación es una tarea más difícil que predecir los estudiantes reinscripción. Esperábamos esto como el límite para los datos. utilizado en las predicciones (es decir, el primer año de los estudiantes) estaba cerca del punto en el que un estudiante es clasificado como reinscrito (después su segundo año) pero fue mucho antes que cuando un estudiante fue clasificado como no completado (después de su sexto año). Esto ayuda a resaltar el grado en que las diferentes operaciones Las definiciones de deserción pueden alterar enormemente la fuerza predictiva percibida de estos clasificadores. Para otros escenarios, las definiciones alternativas de deserción pueden ser más apropiadas y la efectividad de los esfuerzos para construir modelos predictivos será matizados por estas definiciones y contextos institucionales.

Mostramos las matrices de confusión para los mejores modelos para predicción de graduación y reinscripción (regresión logística y random forest, respectivamente) en la Figura 4. Estas matrices muestran una menor tasa de falsos negativos para los modelos pero una mayor tasa de falsos positivos (es decir, estudiantes clasificados incorrectamente por los modelos como graduados o reinscritos). Para comprender mejor esta tasa más alta de falsos positivos, examinamos los registros completos de la transcripción de estudiantes que fueron clasificados en consecuencia. A través de lo falso positivos, encontramos numerosos casos de incumplimientos y los no reinscritos que habían salido de la Universidad con calificaciones relativamente buenas en comparación con su graduación y

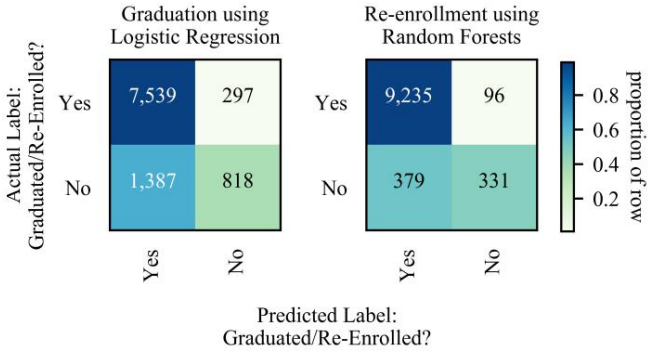

Figura 4: Matrices de confusión al examinar el algoritmos de alto rendimiento para predecir la graduación (LR, izquierda) y reinscripción (RF, derecha).

reinscripción de compañeros. Estos estudiantes también a menudo parecían estar persiguiendo carreras muy competitivas y/o parecía tener rigurosos planes de posgrado (por ejemplo, estudiantes de pre-medicina y pre-odontología). Muchos de estos estudiantes permanecieron en un estado previo a su partida antes de su partida, lo que indica que aunque tenían calificaciones relativamente buenas, probablemente no pudieron para entrar en su(s) programa(s) de grado de elección para varios razones y tuvo que dejar la Universidad para perseguir estas ambiciones como resultado. Desafortunadamente, la Universidad no tener una base de datos de aplicaciones principal centralizada para admisiones y rechazos a carreras específicas. Haberlo podría arrojar luz gran parte de la motivación detrás del deseo de estos estudiantes de dejar la Universidad y si fue, de hecho, motivado por no ingresar a carreras competitivas. Dicho esto, el hecho de que muchos de estos estudiantes eran académicamente similares a sus las contrapartes que se gradúan y reinscriben ilustran aún más por qué parece haber un techo efectivo con respecto a poder predictivo usando los datos dados, como se ve en la Tabla 3.

Desde una perspectiva práctica, cabe señalar que la los umbrales de clasificación para estos modelos no se ajustaron con respecto a la sensibilidad o la especificidad. En la práctica, al desarrollar sistemas institucionales para identificar a los estudiantes en riesgo de irse, puede ser útil elevar el umbral de clasificación al predecir si un estudiante se graduará o se reinscribirá, favoreciendo así un menor recuerdo a expensas de mayor precisión. Esto reduciría efectivamente el número de estudiantes que se prevé que se graduarán pero en realidad no (es decir, falsos positivos) a expensas de más falsos negativos, lo que podría ser más aceptable al desarrollar un Sistema de alerta para alumnos en riesgo de abandono.

4.3 Predicciones usando diferentes subconjuntos de datos

Después de examinar los resultados de la predicción de graduados y reinscripciones usando todas las características, usamos la regresión logística regularizada para predecir la graduación y la reinscripción usando subconjuntos de los datos. Usamos la regresión logística después de ver que funcionó muy bien en relación con otros modelos tanto para tareas de predicción (consulte la Sección 4.2) y porque tenía tiempos de entrenamiento relativamente rápidos debido a que tenía menos hiperparámetros para ajustar. Esto nos permitió entrenar más eficientemente a los 12 modelos diferentes que se necesitaban al examinar el desempeño de subconjuntos de datos específicos (es decir, modelar por separado la graduación y la reinscripción mientras se usan 6

Tabla 4: Resultados de predicción utilizando subconjuntos de datos específicos. Los valores de referencia se basan en el conjunto de pruebas.

	Graduación		Reinscripción	
Subconjunto	Precisión	AUROC	Precisión	AUROC
Base	78,0%	0.500	92,9%	0.500
Todo	83,2%	0,811	95,0%	0.882
FY-Suma.	83,0%	0,795	94,9%	0.855
Pre-entrada de	82,3%	0,788	94,6%	0.847
demostración	82,5%	0,781	94,6%	0.845
principal agrupada	79,9%	0,661	94,2%	0.768
por departamentos	78,0%	0,634	92,9%	0.643
	77,3%	0,630	92,9%	0.616

subconjuntos de datos de forma aislada para cada uno).

Mostramos los resultados cuando usamos subconjuntos de datos en la Tabla 4 junto con el desempeño del clasificador de regresión logística de la Sección 4.2. Las características basadas en transcripciones tienden para desempeñarse mejor que la información sobre los estudiantes antes de su matrícula en la Universidad. Más específicamente, los datos demográficos y la información previa a la entrada fueron relativamente mal para predecir tanto la graduación como la reinscripción. Intuitivamente, esto no es una sorpresa ya que el proceso de admisión en Las universidades altamente competitivas tienden a ser bastante selectivas. con énfasis en apoyar y mantener un cuerpo estudiantil exitoso pero diverso. Además, tales instituciones Es posible que ya se estén realizando esfuerzos para reducir la demografía. disparidades para el éxito de los estudiantes. Mientras tanto, al mirar subconjuntos de datos basados en transcripciones, los datos de resumen del primer año se desempeñaron mejor con un rendimiento similar al uso la totalidad de los datos. Esto es particularmente notable como los datos resumidos del primer año contenían menos características que los otros subconjuntos de datos basados en transcripciones, pero se centró en resúmenes de rendimiento a lo largo del tiempo en lugar de agregaciones entre departamentos/numeraciones del curso.

Estos hallazgos son particularmente interesantes a la luz del trabajo de otros investigadores. Por ejemplo, Nagy y Molontay encontraron que el desgaste podría predecirse con precisión usando lo que perfilar como características demográficas y previas a la entrada únicamente [22]. Sin embargo, no vemos un éxito similar aquí. Creemos esto podría deberse a entornos educativos y perfiles de estudiantes muy diferentes (p. ej., aquí, la mayoría de los estudiantes tienden a graduarse o volver a matricularse, mientras que la población estudiantil de Nagy disminuyó principalmente fuera). En un trabajo anterior, Dekker et al. encontró que las características basadas en transcripciones tienden a tener más fuerza predictiva que características previas a la entrada, pero examinó esto a través de un número bastante limitado subconjuntos de datos [10]. Nuestros resultados se hacen eco de este hallazgo. Recientemente, Manrique et al. descubrió que la deserción se podía predecir utilizando el desempeño de los estudiantes en algunos cursos clave [20]. Aquí, encontramos que los agregados durante el primer año tienden a funcionar mejor que representaciones más detalladas de la toma de cursos (por ejemplo, agrupar clases por prefijo de curso y numeración). Como discutido en la Sección 3.3.3, decidimos no usar representaciones de cursos individuales en este trabajo.

Mostramos las curvas ROC para los modelos de regresión logística regularizados utilizando cada uno de los subconjuntos de datos, así como el todo el espacio de características en la Figura 5. El hecho de que la demografía y los datos previos a la entrada dieron generalmente un peor rendimiento que

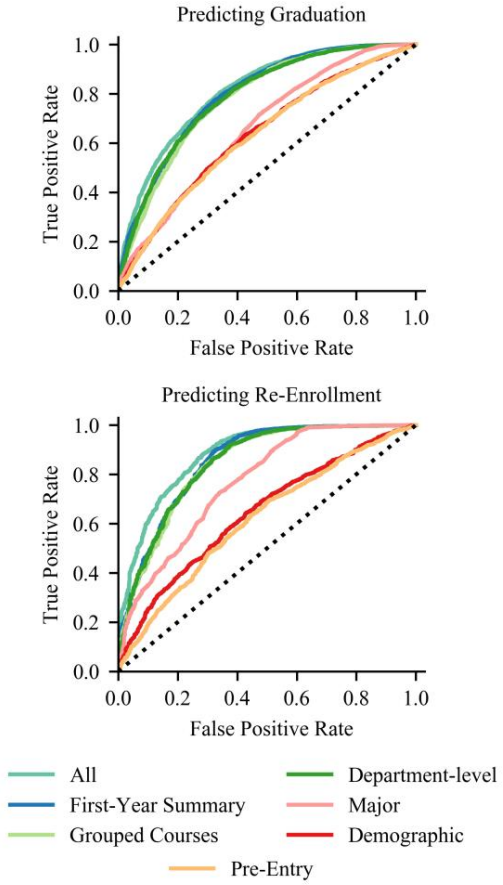

Figura 5: Curvas características de funcionamiento del receptor cuando se utilizan diferentes subconjuntos de datos.

características basadas en la transcripción es muy evidente a partir de la curvas ROC. Mientras tanto, los datos sobre carreras tendían a funcionar peor que otras funciones basadas en transcripciones, pero mejor que los datos demográficos y previos a la entrada. El hecho de usar los datos sobre las carreras no arrojaron resultados particularmente sólidos probablemente se relaciona con el hecho de que la mayoría de los estudiantes en el conjunto de datos estaban en un estado pre-principal a lo largo de su primer año y declarado formalmente su especialidad de interés más adelante en sus carreras universitarias. Como se señaló anteriormente, se creó un sistema centralizado de aplicaciones principales. no está disponible, de lo contrario podría haberse aprovechado además de datos sobre especializaciones para dibujar una imagen más clara del interés académico de los estudiantes. Mientras tanto, los otros conjuntos de datos basados en transcripciones tenían curvaturas muy similares para las curvas ROC cuando predecir tanto la graduación como la reinscripción.

Mostramos matrices de confusión a partir del uso de los mejores subconjunto de datos en la Figura 6. El subconjunto de datos de mejor rendimiento para ambas tareas de predicción fueron datos resumidos del primer año. Por Al comparar estas matrices de confusión con las que se muestran en la Figura 4, se puede ver que usando solo un subconjunto limitado de características tiende a clasificar los datos de manera similar a los modelos construidos sobre la totalidad de los datos. Esto es cierto no sólo en términos de cuán efectivos son los modelos para hacer predicciones, pero también con respecto a la tasa relativamente alta de falsos positivos visto en las cuatro matrices.

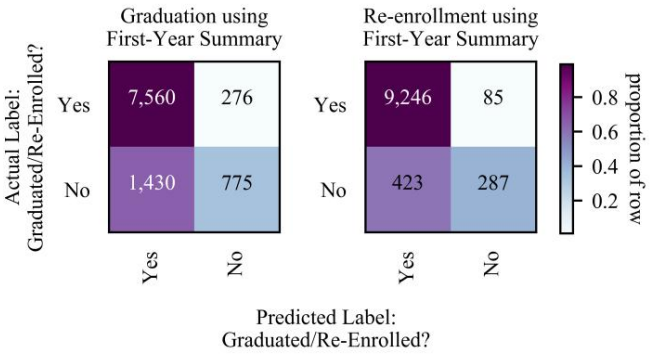

Figura 6: Matrices de confusión al examinar el subconjunto de datos de mayor rendimiento para predecir la graduación (izquierda) y la reinscripción (derecha). El subconjunto de datos de mejor rendimiento fue el mismo para ambas tareas (datos resumidos del primer año).

5. DIRECCIONES FUTURAS

Creemos que los hallazgos con respecto a los subconjuntos de datos han amplias implicaciones políticas, particularmente para identificar estudiantes en riesgo de abandonar los estudios en las grandes universidades públicas. En tales entornos, puede haber un esfuerzo de larga data para disminuir disparidades demográficas con respecto a la deserción y, como resultado, los registros de transcripciones pueden ser más viables como características en modelos predictivos que la información demográfica/previa a la entrada. Además, estos entornos también pueden tener recursos limitados con respecto al tiempo disponible para que el personal esté disponible. características del ingeniero. En tales entornos, saber qué características sería más predictivo de la deserción sin la necesidad de características de ingeniería manual en la totalidad de los datos disponibles a las instituciones podría ahorrar tiempo y esfuerzo en la construcción de modelos. Hemos tenido conversaciones con los administradores de la University por interpretar mejor nuestros resultados y mejorar los procesos para identificar a los estudiantes que necesitan asistencia.

Otra dirección de interés es comprender mejor las características utilizadas para predecir la deserción. Esto incluye no solo examinar más a fondo los determinantes individuales clave de la deserción, como hemos hecho en trabajos anteriores [3, 2], pero también encontrando el mejor combinación de características en los subconjuntos. Lo haríamos quisiera examinar este "espacio mínimo viable de características" en el contexto de los datos disponibles en las bases de datos de los registradores, así como investigar el grado en que estas características se relacionan con la teoría establecida sobre la deserción estudiantil [12].

6. CONCLUSIONES

En este trabajo, utilizamos datos de las bases de datos de registradores de un gran universidad pública de EE. UU. para predecir tanto la graduación como reinscripción usando información limitada al primer año calendario de los estudiantes en la universidad. Hacemos esto usando un conjunto de datos de estudiantes que abarca la totalidad del estudiante universitario cuerpo y, por lo tanto, es mucho más grande que los estudios previos que predicen la deserción estudiantil (N = 66,060). Al hacerlo, demostramos que tanto la graduación como la reinscripción se pueden predecir de manera efectiva utilizando características generadas a partir de datos que son rutinariamente recolectados en instituciones de educación superior. Además, también examinamos el grado en que subconjuntos específicos de los datos del registrador puede ser útil para predecir la deserción, encontrar que las funciones basadas en transcripciones tienden a superar a las funciones basado en las historias de los estudiantes antes de la universidad. Esto implica que se pueden esbozar estrategias efectivas de intervención basadas en registrar recuerdos.

Predecir la reinscripción después del primer año de los estudiantes fue un tarea mucho más manejable que predecir la graduación. Esta puede atribuirse al hecho de que predecir la graduación requiere predecir el éxito académico años en el futuro desde el punto en que los datos eran limitados mientras que la predicción la reinscripción es en un plazo mucho más corto. Teniendo en cuenta las influencias impredecibles que hacen que los estudiantes abandonen universidad antes de graduarse (por ejemplo, limitaciones financieras, dificultades personales, etc.), una tarea de predicción más confiable puede ser examinar si un estudiante regresará trimestre por trimestre. Esto podría ser particularmente útil para desarrollar alertas sistemas para identificar a los estudiantes en riesgo de deserción. Sin embargo, esto no fue explorado en este trabajo debido a los relativamente pocos estudiantes que abandonaron la Universidad después de un solo término.

Descubrimos que parece haber un límite superior para el poder predictivo de nuestro conjunto de datos. Esto demuestra las limitaciones cuando se confía únicamente en los datos del registrador y muestra la necesidad de características adicionales en la experiencia del estudiante para mejorar el poder predictivo. Algunas características potenciales de interés incluyen medidas de integración social en el campus y de ayuda financiera. También podría ser de interés comprender mejor las aspiraciones de los estudiantes más allá del simple uso de las especialidades declaradas. especialmente utilizando representaciones alternativas del comportamiento de los estudiantes para tomar un curso, como lo demostraron recientemente Luo y Pardos [19].

Por último, mostramos que las características generadas a partir de la transcripción registros, particularmente agregados y resúmenes de los académicos, funcionan mejor para las predicciones que demográficas y datos previos a la entrada. Gran parte de esto probablemente se deba a la selectividad de la Universidad y su política de admisión. No obstante, demuestra cuán útiles pueden ser los datos de transcripciones para tales tareas de predicción en contraste con la información sobre los estudiantes antes de la universidad. Demostramos que usando subconjuntos de datos de bases de datos de registradores (en este caso, agregados de primer año de los estudiantes) puede ser casi tan eficaz para las predicciones como la generación manual de una amplia franja de características de diferentes fuentes de datos institucionales.

7. AGRADECIMIENTOS

Los autores desean agradecer a los administradores de datos de la Universidad de Washington por su ayuda en la obtención los datos utilizados en este trabajo.

8. REFERENCIAS

[1] E. Aguiar, NV Chawla, J. Brockman, GA Ambrose y V. Goodrich. Compromiso frente a desempeño: uso de carteras electrónicas para predecir primero retención de estudiantes de ingeniería por semestre. En Procedimientos de la 4ta Conferencia Internacional sobre el Aprendizaje Análisis y conocimiento, páginas 103–112. ACM, 2014.

[2] L. Aulck, R. Aras, L. Li, C. L'Heureux, P. Lu y J. Oeste. STEM-ming the tide: Predecir STEM deserción utilizando los datos del expediente académico de los estudiantes. SIGKDD Taller de aprendizaje automático para la educación, 2017.

[3] L. Aulck, N. Velagapudi, J. Blumenstock y J. West. Predicción de la deserción estudiantil en la educación superior. Aprendizaje automático de ICML en buenas aplicaciones sociales Taller, 2016.

[4] RS Baker y PS Inventado. Datos educativos
análisis de minería y aprendizaje. En Learning Analytics, páginas 61–
75. Springer, 2014.

[5] J. Bayer, H. Bydzovsk´a, J. G´eryk, T. Obsivac y L. Popelinsky.
Predicción de la deserción a partir del comportamiento social de los
estudiantes. En Actas de la 5ª Conferencia Internacional sobre
Minería de Datos Educativos, 2012.

[6] AF Cabrera, A. Nora y MB Castañeda. El papel de las finanzas en el
proceso de persistencia: Un modelo estructural. Investigación en
Educación Superior, 33(5):571–593, 1992.

[7] AF Cabrera, A. Nora y MB Castañeda. Persistencia universitaria: prueba
de modelado de ecuaciones estructurales de un modelo integrado de
retención de estudiantes. Revista de educación superior, 64(2):123–139,
1993.

[8] ALCaison. Análisis de instituciones específicas
investigación de retención: una comparación entre encuestas y métodos
de bases de datos institucionales. Investigación en Educación Superior,
48(4):435–451, 2007.

[9] AP Carnevale, N. Smith y J. Strohl. Recuperación:
Requisitos de educación y crecimiento laboral hasta 2020. 2013.

[10] GW Dekker, M. Pechenizkiy y JM
Vleeshowers. Predicción de la deserción de los estudiantes: un
estudio de caso. Grupo de Trabajo Internacional sobre Minería de
Datos Educativos, 2009.

[11] D. Delén. Predecir la deserción estudiantil con datos
métodos de minería. Journal of College Student Retention: Research,
Theory & Practice, 13(1):17–35, 2011.

[12] C. Demetriou y A. Schmitz-Sciborski. Integración, motivación, fortalezas
y optimismo: Teorías de la retención del pasado, presente y futuro. En
Actas del 7º Simposio Nacional sobre Retención de Estudiantes,
Charleston, SC, páginas 300–312, 2011.

[13] S. Halawa, D. Greene y J. Mitchell. Abandonar
predicción en MOOC utilizando características de actividad del alumno.
Experiencias y mejores prácticas en y alrededor de los MOOC, 7, 2014.

[14] D. Hosler. Gestión de la retención de estudiantes: ¿Está el vaso medio
lleno, medio vacío o simplemente vacío? Facultad y Universidad,
81(2):11–14, 2006.

[15] WE Hudson Sr. ¿Puede una alerta temprana excesiva
sistema de advertencia de ausentismo ser eficaz en la retención de los
estudiantes de primer año? Journal of College Student Retention:
Research, Theory & Practice, 7(3):217–226, 2005.

[16] SM Jayaprakash, EW Moody, EJ Laur´ÿa, JR
Regan y JD Baron. Alerta temprana de estudiantes en riesgo
académico: una iniciativa de análisis de código abierto.
Journal of Learning Analytics, 1(1):6–47, 2014.

[17] N. Johnson. Los costos institucionales de estudiante
desgaste. Proyecto Delta Cost en American Institutes for Research,
2012.

[18] ZJ Kovaÿcí´c. Predicción temprana del éxito de los estudiantes:
extracción de datos de inscripción de estudiantes. En Actas de la
Conferencia sobre Educación en Ciencias y TI (InSITE), páginas
647–665. Citaseer, 2010.

[19] Y. Luo y ZA Pardos. Diagnosticar el dominio de la materia de los
estudiantes universitarios y predecir la finalización del título en el
espacio vectorial. En la Trigésima Segunda Conferencia AAAI sobre
Inteligencia Artificial, 2018.

[20] R. Manrique, BP Nunes, O. Marino, MA
Casanova y T. Nurmikko-Fuller. Un análisis de la representación de
los estudiantes, características representativas y algoritmos de
clasificación para predecir la deserción de grado. En Actas de la 9.ª
Conferencia Internacional sobre Análisis y Conocimiento del Aprendizaje,
páginas 401–410.
ACM, 2019.

[21] LG Moseley y DM Mead. Predecir quién abandonará los cursos de
enfermería: un ejercicio de aprendizaje automático. Educación de
enfermería hoy, 28(4):469–475, 2008.

[22] M. Nagy y R. Molontay. Predicción de la deserción en la educación
superior basada en el rendimiento de la escuela secundaria. En
2018 IEEE 22nd International Conference on Intelligent Engineering
Systems (INES), páginas 000389–000394. IEEE, 2018.

[23] D. Niemi y E. Gitin. Uso de big data para predecir
abandono estudiantil: posibilidades tecnológicas para la investigación.
En Actas de la Asociación Internacional para el Desarrollo de la
Sociedad de la Información (IADIS)
Conferencia Internacional sobre Cognición y Aprendizaje
Exploratorio en la Era Digital, 2012.

[24] TJ Pantages y CF Creedon. Estudios de abandono universitario:
1950-1975. Revisión de la investigación educativa, 48(1):49–101, 1978.

[25] S. Ram, Y. Wang, F. Currim y S. Currim. Uso de big data para predecir
la retención de estudiantes de primer año. 2015.

[26] C. Romero y S. Ventura. Minería de datos educativos:
una revisión del estado del arte. IEEE Transactions on Systems, Man,
and Cybernetics, Parte C (Aplicaciones y revisiones), 40(6):601–618,
2010.

[27] S. Sadati y NA Libre. Desarrollo de una temprana
sistema de alerta para predecir los estudiantes en riesgo de reprobar en
función de sus actividades iniciales del curso. 2017.

[28] M. Schneider. Finalización de la primera vuelta: el costo de la deserción
de estudiantes de primer año en los colegios y universidades de cuatro
años de Estados Unidos. Institutos Americanos de Investigación, 2010.

[29] JM Simons. Un estudio nacional de modelos de alerta temprana para
estudiantes en instituciones de educación superior de cuatro años.
Universidad Estatal de Arkansas, 2011.

[30] W. Spady. Abandonos de la educación superior: Hacia un modelo
empírico. Intercambio, 2(3):38–62, 1971.

[31] J. Summerskill. Abandonos de la universidad. En El Colegio
Americano. Wiley, Nueva York, 1965.

[32] D. Thammasiri, D. Delen, P. Meesad y N. Kasap.
Una evaluación crítica del problema de distribución de clases
desequilibrada: el caso de la predicción de la deserción de estudiantes
de primer año. Sistemas expertos con aplicaciones, 41(2):321–330,
2014.

[33] V. Tinto. Definición de la deserción: una cuestión de perspectiva.
Nuevas direcciones para la investigación institucional,
1982(36):3–15, 1982.

[34] V. Tinto. Dejar la universidad: repensar las causas y
curas de la deserción estudiantil. Prensa de la Universidad de Chicago,
1987.

[35] D. Yang, T. Sinha, D. Adamson y CP Ros´e. Encienda, sintonice, abandone:
Anticipando la deserción de los estudiantes en cursos masivos abiertos
en línea. En Actas del taller educativo basado en datos de NIPS de 2013,
volumen 11, página 14, 2013.