Como citar este artículo: R. Boegeholz, J. Guerra and E. Scheihing, "Exploring Risk of Delay in Academic Trajectories in Two Undergraduate Programs," in IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, vol. 17, no. 3, pp. 290-300, Aug. 2022, doi: 10.1109/RITA.2022.3191298.

Explorando el riesgo de retraso en trayectorias académicas de dos programas de pregrado

Renato Boegeholz, Julio Guerra y Eliana Scheihing

***Resumen*—El retraso en completar programas de estudio de pregrado es un problema importante en la educación superior. Este trabajo explora el retraso académico en dos programas de ingeniería usando una representación de trayectorias académicas novedosa que resume la información semestre a semestre de *riesgo de retraso* asociado a rendimiento, carga académica y dificultad de los cursos. La representación del riesgo de atraso semestre a semestre permite agrupar y contrastar trayectorias. Examinando las trayectorias en dos programas de ingeniería y comparando entre estudiantes que completaron y que abandonaron el programa, encontramos patrones de riesgo de retraso que investigadores y directores de programa pueden identificar y asociar a factores gatillantes. El modelo y análisis presentado en este trabajo contribuye a una mejor comprensión del retraso académico y puede servir para apoyar procesos de adaptación y rediseño curricular.**

***Index Terms*—time-to-degree, academic analytics, curricular analysis, learning analytics, educational data mining, educational trajectories.**

I. Introducción

El retraso en completar programas de educación superior es un problema global conocido. De acuerdo a [1], sólo 36 % de los estudiantes de educación superior de Estados Unidos completan sus estudios en el tiempo definido por estos programas. De manera similar, en países europeos como Slovenia, Austria, Holanda y Portugal, la proporción de estudiantes que se gradúan "a tiempo" está entre 23 % y 30 %; mientras que en Suiza, Suecia, Finlandia y Noruega, esta proporción está entre el 39 % al 43 %. En Chile, obtener un título universitario toma alrededor de 31 % más de lo esperado, lo que está cerca del promedio latinoamericano de 36 % [2]. Este retraso es un problema severo que tiene un costo estimado en de unos US$ 460 millones por año de para las familias y el estado [3], [4].

El retraso en completar un programa de estudios puede tener múltiples causas como el bajo rendimiento y repetición de cursos, suspensión temporal de los estudios, o inscripción de menos cursos por semestre (baja carga semestral) [5]. En el contexto de la educación universitaria en Chile, donde los programas de estudio se caracterizan por una estructura bastante rígida de cursos por semestre, el retraso académico adquiere una nueva perspectiva: los/las estudiantes se *atrasan* al repetir cursos que les impiden completar requisitos para los cursos del siguiente semestre. El retraso representa una desviación de la trayectoria académica del estudiante respeto al plan ideal. Esta distorsión es dinámica y compleja. Los estudiantes pueden acumular retraso y también disminuirlo (*ponerse al día*). El riesgo de perder beneficios como becas o la necesidad de reducir los semestres proyectados para la titulación (y sus costos) son fuertes motivadores para tratar de minimizar el retraso. Esto a menudo fomenta que los estudiantes sobrecarguen sus semestres inscribiendo muchos cursos y finalmente terminando semestres estresantes y con bajo rendimiento. La situación se agrava en programas donde los cursos iniciales tienen alta tasa de reprobación, como es el caso de los programas de ingeniería. A pesar de la importancia que el retraso académico tiene en la decisiones que los estudiantes toman, el problema ha sido poco investigado.

En este trabajo exploramos el rol del retraso académico en dos programas de ingeniería. Para capturar la dinámica y la complejidad del retraso académico, acuñamos el concepto de "riesgo de retraso" como la relación entre la información académica semestre a semestre y el retraso acumulado respecto del plan de estudios proyectado al 5to semestre de vida académica. El retraso al 5to semestre es relevante en nuestro contexto porque representa la distancia que el estudiante tiene del plan ideal de los primeros dos años que corresponden al ciclo de *Bachillerato*. El Bachillerato es también un hito común en todos los programas de ingeniería en nuestra universidad lo que permite comparar el atraso en programas similares. Bachillerato contiene todos los cursos de álgebra, cálculo, química y física, que además tienen las más alttas tasas de reprobación.

Mientras la literatura ha descrito el comportamiento académico en base a indicadores finales de rendimiento como calificaciones [6], [7], [8], nuestro trabajo incorpora otros aspectos de la información académica como la carga semestral, la repetición de cursos y la dificultad de los cursos, y de esta forma captura una fotografía mas completa de la situación académica semestre a semestre. Primero nos enfocamos en representar trayectorias académicas que sinteticen la información académica respecto del riesgo de retraso.

RQ1) ¿Cómo pueden caracterizarse las trayectoria académicas en función del riesgo de retraso y considerando carga académica, dificultad de cursos y rendimiento?

Para abordar RQ1 proponemos la construcción de trayectorias de riesgo de retraso usando métodos no supervisados para agrupar (mediante *clusters*) a los estudiantes

Renato Boegeholz completó su tesis de Magíster en Informática en la Facultad de Ciencias de la Ingeniería de la Universidad Austral de Chile, Chile. (email: renato.boegeholz@uach.cl). https://orcid.org/0000-0002-6826-1210

Julio Guerra es académico del Instituto de Informática de la Universidad Austral de Chile, Chile. (email: jguerra@inf.uach.cl). https://orcid.org/0000-0002-8296-9848

Eliana Scheihing es académica del Instituto de Informática de la Universidad Austral de Chile, Chile. (mail: escheihi@inf.uach.cl). https://orcid.org/0000-0003-1801-9167

en base a su situación académica (carga académica, dificultad de cursos y rendimiento) semestre a semestre y asociar los grupos al nivel de retraso que presentan los casos conocidos del grupo al 5to semestre. De esta forma la trayectoria de riesgo de retraso es la secuencia de la estimacion de atraso del grupo en los 4 primeros semestre de cada estudiante. Esta representación no sólo permite proyectar retraso en estudiantes nuevos, si no que también permite agrupar trayectorias académicas para descubrir patrones y contrastarlos entre estudiantes con distinto nivel de éxito académico. Debido a que el retraso puede afectar la experiencia académica y social de los/las estudiantes, es razonable explorar la relación entre riesgo de retraso y abandono. El abandono académico es un fenómeno que motiva el rediseño curricular y la innovación en los planes de estudio; alrededor de 21 % de los estudiantes chilenos abandonan sus estudios el primer año [9]; y la tasa de titulación en los programas de licenciatura en Chile sólo alcanza el 51 % [10]. Nuestra segunda pregunta de investigación es:

RQ2) ¿Qué relación existe entre las trayectorias de riesgo de retraso y el abandono académico?

Este artículo extiende nuestro trabajo previamente publicado [11] con nuevos datos y análisis, incluyendo la aplicación de las trayectorias de riesgo de retraso en dos programas de ingeniería. Estos programas tiene la misma estructura de sus primeros dos años, pero presentan indicadores académicos muy distintos que son visibles en las trayectorias de riesgo de retraso. Agregamos análisis para contrastar trayectorias de riesgo de retraso entre grupos de estudiantes: abandono vs persitente. Estos análisis ejemplifican como la representación de trayectorias propuesta puede ser usada. Hemos también extendido la discusión de resultados e incorporamos hallazgos de una sesión de validación donde analizamos la representación de trayectorias de riegos de retraso con un director de programa. El artículo se organiza de la siguiente manera. La sección 2 presenta trabajo relacionado, incluyendo trabajo que analiza información académica y abandono. Sección 3 detalla los datos usados, análisis propuestos y el método para caracterizar y representar la información académica y el riesgo de retraso. La sección 4 presenta análisis exploratorio de la variables y su capacidad predictiva respecto del retraso, y la construcción de las trayectorias de riesgo de retraso. Discusión de resultados y conclusiones son presentadas en la sección 5.

II. TRABAJO RELACIONADO

Nuestro trabajo se enmarca junto a la literatura relacionada a rendimiento académico curricular, predicción de abandono y retención, y trayectorias educacionales. Identificamos teorías relevantes, propuestas y resultados de la literatura existente, a la vez que reconocemos brechas relevantes al contexto de nuestro problema, destacando la contribución de nuestro trabajo. Investigadores han explorado, identificado y evaluado varios factores asociados al rendimiento académico con sistemas predictivos. Entre estos factores se encuentran variables socio-económicas y demográficas, calificaciones y participación que son reconocidas como las más comunes [12]. En nuestro caso, las fuentes de datos disponibles se restringen a datos académicos dentro de los programas de estudio en la universidad. Otra información acerca de los estudiantes como aspectos demográficos y socio-económicos no están disponible y a menudo no son recolectados ni tampoco disponibilizados por las instituciones. Enfocándose en datos académicos, [13] muestra que es posible predecir - con niveles de acierto razonables - rendimiento en el cuarto año, a partir de indicadores de rendimiento previos a ingresar a la universidad, más las calificaciones de los dos primeros años de estudio. En el mismo trabajo y usando árboles de decisión [14], se puedo derivar cursos que pueden servir como indicadores del nivel del programa. Respecto al progreso del rendimiento de los estudiantes, encontraron que los estudiantes generalmente tienen el mismo nivel de calificaciones (bajo, medio, alto) en todos los cursos, con patrones repetitivos a través de los años. En [15], los resultados de una auto-evaluación de entrada y el rendimiento académico del primer año se usan para agrupar a los estudiantes con k-means [16] y luego seguir sus trayectorias de rendimiento en el segundo y tercer año. Los autores encontraron que el test de auto-evaluación es un indicador importante para predecir tanto el rendimiento en el primer año como el progreso en los siguientes años. Además, los autores observan tendencias en el comportamiento académico que están fuertemente determinadas por el rendimiento en el primer año. Regresiones lineales múltiples personalizadas (Personalized Multiple Linear Regressions, PLMR) y factorización matricial (Matrix Factorization, MF, [17]), basadas en sistemas recomendadores fueron usadas en [18] para estimar con precisión el rendimiento de los estudiantes. Usando información histórica de calificaciones junto a información adicional disponible, como el número de créditos inscritos cada semestre, ellos pudieron predecir las calificaciones en el próximo semestre. Se argumenta que los métodos pueden usarse tanto con datos que provienen de clases tradicionales, como de cursos en línea masivos (Massive Open Online Courses, MOOCs).

La predicción de persistencia y abandono ha recibido amplia atención en la comunidad investigativa [19]. Investigadores han encontrado predictores significativos de persistencia como el promedio de calificaciones de educación secundaria (grade point average, GPA) y el promedio en primer semestre universitario. Además, revelan que los estudianets que tienen mejor formación académica anterior son más persistented comparados con quienes fueron obligados a asistir a cursos de remediación [20]. También en relación con el abandono, [8] construyó medidas de rendimiento académico como promedios semestrales para todos los semestres posteriores al 75 % de los créditos requeridos por el programa. Regresiones logit/logísticas aplicadas a estos predictores les permitieron a [21] descubrir una fuerte relación entre el *momentum* académico -que explica la finalización y no finalización de la titulación- y el abandono tardío. La investigación que mira trayectorias académicas se ha enfocadoprincipalmente en predictores de nivel individual como características socio-demográficas, educación previa, e indicadores de rendimiento académico. En contraste, factores institucionales que influyen en las trayectorias como la estructura de los programas de

estudio, han sido relativamente menos explrados [22]. Una excepción es [7], que analiza las trayectorias de desempeño de los estudiantes para informar acerca de los diseños curriculares. En particular, modelaron la dificultad de cada curso como su contribución (negativa o positiva) al GPA de los estudiantes y, luego, contrastaron esta medida con una encuesta de percepción aplicada a los estudiantes. Utilizando los mismos datos de rendimiento, también llevaron a cabo un análisis de los distintos *caminos o rutas* entre la inscripción y el abandono. La representación de trayectorias académicas no es trivial y es necesario considerar la dependencia temporal de las variables bajo estudio. Siguiendo esta idea, [23] usa minería de patrones frecuentes [24] para revelar trayectorias y entender que secuencia de cursos podrían mejorar el rendimiento de los estudiantes. Con otro enfoque, [25] propone un modelo de datos secuencial que explícitamente captura las dependencia temporal entre las características académicas y construye huellas o firmas, permitiendo diferentes interpretaciones analíticas y modelos predictivos para el riesgo de retraso académico.

Es necesario considerar que la mayor parte del trabajo previo se desarrolla en el contexto de sistemas educativos de cursos/créditos flexibles, donde los estudiantes se gradúan debido a la suma de cursos obligatorios y optativos aprobados [15], [20], [23], [26]. Este contexto difiere de nuestro sistema de currículo secuencial-no flexible, donde el flujo de cursos a tomar está predefinido, semestre a semestre, en el programa de estudios. Aquí, las trayectorias académicas que divergen del plan *ideal* implican, principalmente, que los estudiantes se retrasan en lugar de señalar alguna heterogeneidad de intereses o preferencias académicas como ocurre en situaciones de planes de estudio flexibles. A pesar de la amplia atención prestada a la deserción, el retraso académico, en este contexto, ha sido poco estudiado. Recientemente, [27] presenta un análisis de trayectorias académicas de estudiantes que busca identificar patrones de abandono tardío. Utilizando técnicas de minería de procesos, determinan que aquellos estudiantes que se retrasan por reprobar cursos con altas tasas de reprobación y que no retoman inmediatamente estas materias, tienen una mayor probabilidad de deserción tardía que los que sí lo hacen.

Tal como se presenta, varios de los modelos propuestos consideraron el desempeño de los estudiantes como variables explicativas, sin necesariamente reflejar la dependencia que existe entre el desempeño de un estudiante dentro de un mismo semestre y semestres sucesivos. Nuestro trabajo se distingue del trabajo anterior en su enfoque multivariado que considera, además del desempeño académico, aspectos como la carga académica y la dificultad de los cursos en cada semestre. Además, incorpora la dimensión temporal al modelar las trayectorias de los estudiantes en un contexto curricular estructurado secuencial, donde alrededor del 90 % de las asignaturas son obligatorias.

III. MÉTODOS

III-A. Descripción de los datos y modelado de características

Este trabajo examina datos académicos de una universidad "tradicional" chilena. En Chile, las universidades "tradicionales" son aquellas públicas y privadas sin fines de lucro creadas, en su mayoría, antes de la Ley General de Universidades introducida en 1981. Como ya se ha mencionado, es relevante que la mayoría de los programas de educación superior chilenos (incluidos los del presente estudio) implementan planes curriculares semiflexibles, con cursos semestrales que siguen una secuencia de prerrequisitos previamente definida.

Analizamos dos programas de pregrado: Ingeniería Civil en Informática (INFO) e Ingeniería Civil en Obras Civiles (OOCC). Durante el período que abarcan los datos recopilados, estos programas implementan un plan de estudio consistente que solo tiene pequeñas variaciones en los códigos de los cursos y los cursos electivos. Ambos programas tienen una duración total de once semestres, con seis cursos por semestre en promedio. El conjunto de datos comprendió registros de inscripción y calificaciones de los cursos tomados, aprobados, reprobados y eliminados por los estudiantes en cada período. Los conjuntos de datos están compuestos por 7.941 registros de actividad académica de 204 estudiantes diferentes para INFO y 12.032 registros de 309 estudiantes diferentes para OOCC. Los programas se analizaron con el mismo método (explicado más adelante, en esta sección) pero por separado ya que, aunque presentan un programa de estudios muy similar en los dos primeros años, desarrollan diferentes dinámicas.

Todos los datos utilizados fueron anonimizados y proporcionados por la universidad bajo sus políticas y estándares de privacidad y seguridad de la información.

Definimos una *trayectoria académica* como una representación multidimensional de la situación académica de un estudiante, distribuida en una secuencia semestre a semestre. Puesto que algunos estudiantes se retrasan y otros abandonan el programa, las trayectorias pueden tener diferentes longitudes. Para poder comparar las trayectorias y hacer análisis adicionales, consideramos solo los primeros cuatro semestres contados desde la primera inscripción del estudiante en el programa. Estos cuatro semestres coinciden con un hito denominado Bachillerato en todos los programas de ingeniería de la institución académica, que concentra las asignaturas fundamentales de matemática, química y física, las que a su vez presentan los mayores índices de reprobación. Es durante el Bachillerato cuando los estudiantes acumulan retrasos debido a reprobaciones repetidas en los cursos de matemáticas. Por lo tanto, nuestra variable de retraso corresponde al retraso al quinto semestre (DELAY5), que cubre todo el retraso acumulado potencial en los primeros cuatro semestres. Más precisamente, DELAY5 mide hasta qué punto el estudiante ha completado los primeros cinco semestres del plan, durante sus primeros cinco semestres de participación en el programa. Un estudiante que aprueba todos los cursos planificados hasta el quinto semestre en sus primeros cinco semestres contados desde su primera inscripción, tiene un DELAY5 igual a cero.

Por otra parte, combinamos los registros académicos individuales con datos históricos de calificaciones de los cursos y la estructura curricular de los programas para generar una serie de características académicas por cada semestre o período de la vida académica del estudiante. Estas características representan diferentes dimensiones académicas relacionadas con el desempeño, la carga del curso (carga de trabajo), la dificultad relativa de los cursos y la coherencia entre los cursos

tomados y su orden teórico en el plan de estudios.

Modelamos y seleccionamos un total de nueve características. Las descripciones y ecuaciones de las características se presentan a continuación, donde i es el i-ésimo semestre de la vida académica del estudiante dentro del programa y j, corresponde a un curso inscrito ese semestre:

1. El **promedio no acumulativo de calificaciones** (GPA) resume el desempeño del período agregando las calificaciones finales del curso y ponderando cada curso por el número de créditos asociados.

$$\text{GPA}_i = \frac{\sum_{j=1}^{\text{N-courses}_i} (\text{final-grade}_j \cdot \text{credits}_j)}{\sum_{j=1}^{\text{N-courses}_i} \text{credits}_j} \in [1, 7]$$

En nuestro contexto, los cursos se califican en una escala que va de 1,0 a 7,0. La calificación para aprobar generalmente se establece en 4,0.

2. La **tasa de aprobación del semestre** (PASS_R) es la relación entre el número de cursos aprobados y el número de cursos tomados.

$$\text{PASS\_R}_i = \frac{\text{passed}_i}{\text{passed}_i + \text{failed}_i} \in [0, 1]$$

Esta variable refleja una visión ligeramente diferente del rendimiento académico y resume el éxito del período académico en términos de cuántos cursos se aprobaron.

3. **Inscripción por primera vez** como la proporción de cursos inscritos por primera vez en el semestre(FIRST_T)

$$\text{FIRST\_T}_i = \frac{\text{first-time-enrolled}_i}{\text{enrolled}_i} \in [0, 1]$$

Esta función está destinada a representar la influencia que puede tener la realización de cursos repetidos en el rendimiento general del alumno. Los valores más bajos de esta variable significan que muchos cursos tomados en el semestre son de segundo, tercer (o incluso cuarto) intento, lo que implicaría mayores niveles de estrés para el/la estudiante.

4. La característica de **progreso académico** (PROGRE) representa la contribución del número de cursos aprobados en el semestre al total de cursos requeridos para obtener el grado.

$$\text{PROGRE}_i = \frac{\text{passed}_i}{\text{prog-length}} \in [0, 1]$$

En su forma longitudinal, proporciona información sobre la tasa de progreso del estudiante.

5. **Retraso temporal académico** (DELAY) o desfase entre el semestre teórico y el semestre actual (promedio) del estudiante de acuerdo su fecha de admisión en el programa.

$$\text{DELAY}_i = \begin{cases} 0 & i \leq \text{avg-sem}_i \\ \frac{i - \text{avg-sem}_i}{\text{prog-length} - 1} & i > \text{avg-sem}_i \end{cases} \in [0, 1]$$

con avg-sem_i el semestre promedio i del estudiante, calculado como

$$\text{avg-sem}_i = \frac{\sum_{j=1}^{\text{N-courses}_i} \text{sem-in-prog}_j}{\text{N-courses}_i} \in [1, \text{prog-length}]$$

con sem-in-prog_j es el semestre en el cual se ubica al curso j dentro del programa de estudios y N-courses, la cantidad de cursos tomados por el/la estudiante ese semestre.

6. La **disparidad temporal** (DISPAR) de los cursos inscritos en el semestre, entendida como la mayor diferencia entre los semestres (preestablecidos) en que están programados los cursos en el plan de estudios. Es una medida de progreso que da información sobre el retraso relativo a la definición del plan de estudios y la brecha temporal entre los cursos inscritos.

$$\text{DISPAR}_i = \frac{\max_j(\text{sem-in-prog}_j) - \min_j(\text{sem-in-prog}_j)}{\text{prog-length} - 1} \in [0, 1]$$

con $j = 1, \cdots, \text{N-courses}_i$

7. La **carga de trabajo académico** para el semestre (WKLOAD) medida como la tasa entre los créditos inscritos y el promedio de créditos semestrales del programa de estudios.

$$\text{WKLOAD}_i = \frac{\sum_{j=1}^{\text{N-courses}_i} \text{credits}_j}{\text{avg-credits}} \in [0, \text{prog-length}]$$

La dimensión de carga de trabajo es crucial para discriminar situaciones de desempeño. Por ejemplo, dos estudiantes pueden tener el mismo GPA, pero uno puede haber tomado cinco cursos y el otro solamente uno.

8. **Dificultad del semestre como medida aditiva** (DIFFIC_A). Nombrada "Dificultad alfa" en nuestro trabajo.

$$\text{DIFFIC\_A}_i = \sum_{j=1}^{\text{N-courses}_i} \text{hist-fail}_j$$

donde hist-fail_j se refiere a la tasa histórica de reprobación en todas las ocasiones semestrales previas a i en las que se ha impartido el curso j, calculada como

$$\text{hist-fail}_j = \frac{\text{total-failed-students}_j}{\text{total-passed-students}_j + \text{total-failed-students}_j}$$

De esta forma,

$$\text{DIFFIC\_A}_i \in [0, \text{N-courses}_i \cdot max(\text{hist-fail}_j)]$$

9. **Dificultad del semestre como medida geométrica** (DIFFIC_B). Denominada "Dificultad beta" en nuestro trabajo.

$$\text{DIFFIC\_B}_i = 1 - \prod_{j=1}^{\text{N-courses}_i} (1 - \text{hist-fail}_j) \in [0, 1]$$

Utilizando la misma definición de hist-fail_j que DIFFIC_A, proporciona un valor normalizado para la característica de dificultad.

III-B. Análisis de datos y modelo secuencial

Para modelar el retraso al quinto semestre (DELAY5) en función de las otras ocho características calculadas para cada semestre, limitamos el alcance de las variables independientes (las características) a los primeros cuatro semestres. Las razones detrás de esto son: primero, la idea de predecir el retraso (y también predecir la deserción) gana relevancia si se puede hacer una predicción temprana. Por tanto, parece razonable

predecir el retraso en el quinto semestre con información de los cuatro semestres anteriores. Y segundo, los cuatro semestres iniciales (correspondientes al Bachillerato) concentran las asignaturas fundacionales de matemáticas y física. Estas asignaturas tienen las mayores tasas de reprobación y aparecen fuertemente relacionadas con el fracaso o el éxito académico. El análisis se realizó en tres pasos:

El primer paso consiste en realizar un análisis de datos exploratorio (EDA) para todas las variables. Resumimos las principales características de cada variable (media y desviación estándar) y los representamos en box plots. Realizamos análisis de componentes principales (PCA) y obtuvimos matrices de correlación para ayudar a comprender la naturaleza de las variables e identificar aquellas más significativas al explicar el retraso académico.

El segundo paso incluyó la construcción de modelos predictivos del retraso al quinto semestre, a partir de las otras características de los primeros cuatro semestres, usando dos algoritmos supervisados: regresión lineal generalizada (GLM) y máquina de vectores de soporte, SVM [28]. Estos análisis nos permiten comprender la relación entre el retraso y las características semestre a semestre que resumen la información académica.

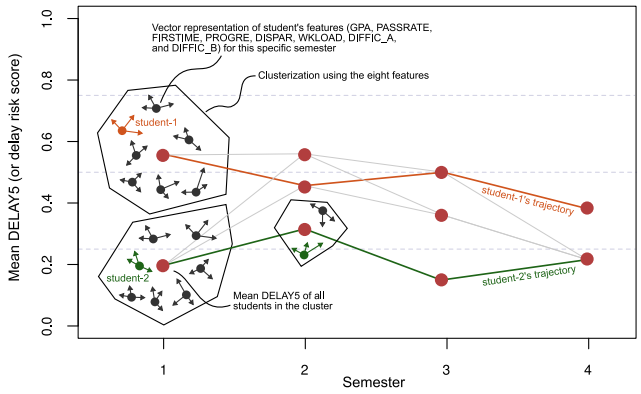

Figura 1: Resumen del proceso de cálculo de las trayectorias académicas y cómo pueden interpretarse

El tercer paso corresponde a la caracterización de las trayectorias académicas asociando el retraso con las características significativas, semestre a semestre, sugeridas por el análisis del paso previo. Implementamos una adaptación del modelo secuencial propuesto por Mahzoon et al. [25] de la siguiente manera: la trayectoria de cada estudiante se representa como una secuencia de cuatro nodos (semestres 1 a 4), donde cada nodo es un valor único que representa una puntuación de riesgo de retraso. Para calcular la puntuación de retraso del i-ésimo semestre, primero, todos los estudiantes se agrupan por las características de su i-ésimo semestre utilizando modelos de mezcla finita gaussianos (GMM) ajustados mediante el algoritmo de maximización de expectativas (EM) [29]. Así, el puntaje del i-ésimo semestre es el DELAY5 medio de todos los estudiantes del grupo. Este método genera un número finito de nodos (medias de DELAY5) cada semestre, por lo tanto, un número finito de trayectorias posibles a lo largo de los semestres. El proceso se resume en la Figura 1. Este modelo nos permite agrupar las trayectorias y calcular trayectorias de nuevos casos (casos con DELAY5 desconocido).

IV. Resultados

IV-A. Análisis Exploratorio de Datos

1) Estadísticas de las Características

La media y la desviación estándar de cada característica se presentan en las Tablas I y II. Todas las características toman valores entre 0 y 1 a excepción de DIFFIC_A. Esta variable presenta valores medios altos en el programa INFO, con tendencia a disminuir a medida que avanzan los semestres, comenzando con un valor promedio de 2,03 en el primer semestre y disminuyendo a 1,6 en el cuarto semestre. En contraste, el programa OOCC muestra un valor promedio de 0,63 el primer semestre y aumenta, en semestres sucesivos, hasta alcanzar un valor promedio de 1,61 en el cuarto semestre. De similar manera, se puede observar que las variables DELAY y DISPAR presentan valores medios bajos asociados a desviaciones estándar altas, lo que significa que, dentro de los datos estudiados, existe una alta variabilidad en estas variables.

Cuadro I: Medias y derivaciones estándar para las características de INFO.

	Semester 1	Semester 2	Semester 3	Semester 4
N	204	204	202	198
GPA	4.59 ± 0.83	4.21 ± 0.74	4.13 ± 0.78	4.10 ± 0.86
PASS_R	0.81 ± 0.21	0.68 ± 0.24	0.66 ± 0.29	0.61± 0.31
FIRST_T	1.00 ± 0.00	0.73 ± 0.31	0.63 ± 0.31	0.65 ± 0.31
PROGRE	0.11 ± 0.03	0.07 ± 0.03	0.06 ± 0.03	0.06 ± 0.04
DELAY	0.00 ± 0.00	0.07 ± 0.08	0.18 ± 0.13	0.27 ± 0.17
DISPAR	0.00 ± 0.00	0.11 ± 0.12	0.13 ± 0.10	0.19 ± 0.14
WKLOAD	1.05 ± 0.07	0.84 ± 0.17	0.78 ± 0.19	0.74 ± 0.24
DIFFIC_A	2.03 ± 0.26	1.96 ± 0.36	1.75 ± 0.33	1.60 ± 0.46
DIFFIC_B	0.93 ± 0.04	0.92 ± 0.07	0.89 ± 0.06	0.89 ± 0.06

Cuadro II: Medias y derivaciones estándar para las características de OOCC.

	Semester 1	Semester 2	Semester 3	Semester 4
N	309	308	307	303
GPA	5.14 ± 0.56	4.46 ± 0.61	4.35 ± 0.61	4.08 ± 0.60
PASS_R	0.94 ± 0.13	0.82 ± 0.20	0.74 ± 0.27	0.70 ± 0.28
FIRST_T	1.00 ± 0.00	0.94 ± 0.13	0.77 ± 0.28	0.74 ± 0.28
PROGRE	0.11 ± 0.02	0.11 ± 0.03	0.08 ± 0.03	0.08 ± 0.03
DELAY	0.00 ± 0.00	0.01 ± 0.03	0.07 ± 0.09	0.14 ± 0.13
DISPAR	0.00 ± 0.00	0.03 ± 0.07	0.13 ± 0.14	0.19 ± 0.15
WKLOAD	0.85 ± 0.07	1.10 ± 0.15	0.84 ± 0.17	0.90 ± 0.20
DIFFIC_A	0.68 ± 0.10	1.41 ± 0.29	1.53 ± 0.31	1.61 ± 0.39
DIFFIC_B	0.52 ± 0.06	0.81 ± 0.10	0.84 ± 0.07	0.84 ± 0.09

La distribución de la variable DELAY a explicar se muestra en la Figura 2 con diagramas box plot para el semestre 1 al 5 para los dos programas analizados (OOCC e INFO). Se observa que el retraso promedio aumenta con los semestres de forma natural, pero con mayor tasa en el programa INFO. La variable DELAY5 muestra un comportamiento de variabilidad considerable en ambos programas.

De manera similar, la Figura 3 contiene diagramas box plot para GPA (promedio no acumulativo de calificaciones) y WKLOAD (carga de trabajo), para los primeros cuatro semestres, tanto en INFO como en OOCC. Las distribuciones

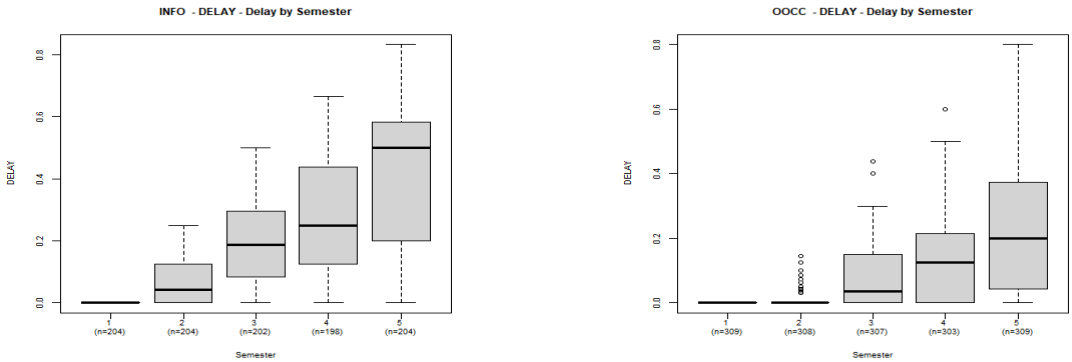

Figura 2: Box plots de la variable DELAY para los primeros cinco semestres.

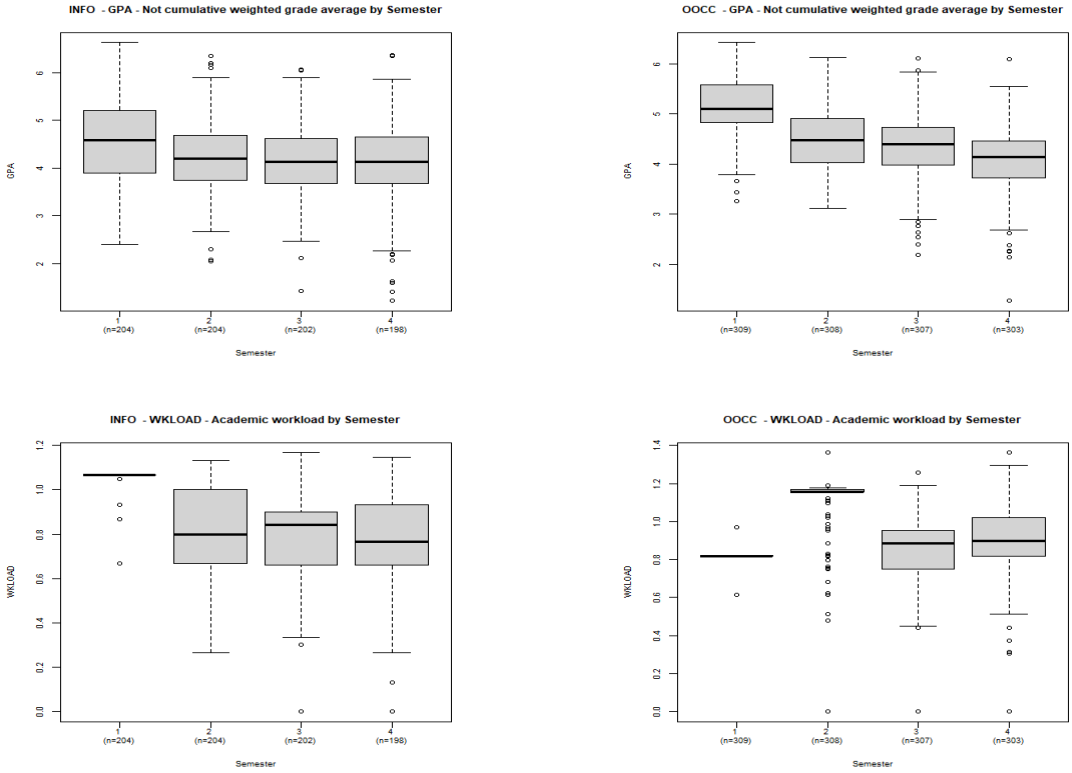

Figura 3: Box plots de las variables GPA y WKLOAD.

de las características muestran similitudes y diferencias entre programas. Por ejemplo, incluso cuando GPA sigue una tendencia similar en ambos programas, existen diferencias notorias WKLOAD, con el programa OOCC presentando un alto valor de carga de trabajo en el segundo semestre, para la mayoría de los estudiantes. La inspección de las distribuciones de las características nos permite concluir que i) el modelado de trayectorias debe realizarse en cada programa de estudios por separado, incluso cuando cada programa tiene un plan de estudios muy similar (ambos programas cumplen con el mismo plan de Bachillerato de ingeniería durante los dos primeros años); y ii) que las diferencias entre características como GPA y WKLOAD respaldan la idea de que es importante incluir toda esta información en el modelado de las trayectorias.

2) Relaciones entre Características

Construimos matrices de correlación e hicimos análisis de componentes principales (PCA) para verificar las diferentes características son informativas para análisis posteriores o si es razonable reducir el conjunto de variables. Los análisis se realizaron por semestre. La figura 5 muestra el caso del segundo semestre como ejemplo. En todos los semestres se observa que el retraso (DELAY5) se correlaciona negativamente con el rendimiento académico, el progreso curricular y la carga académica, es decir, GPA, PASS_R, PROGRE y WKLOAD, respectivamente. DELAY también muestra una

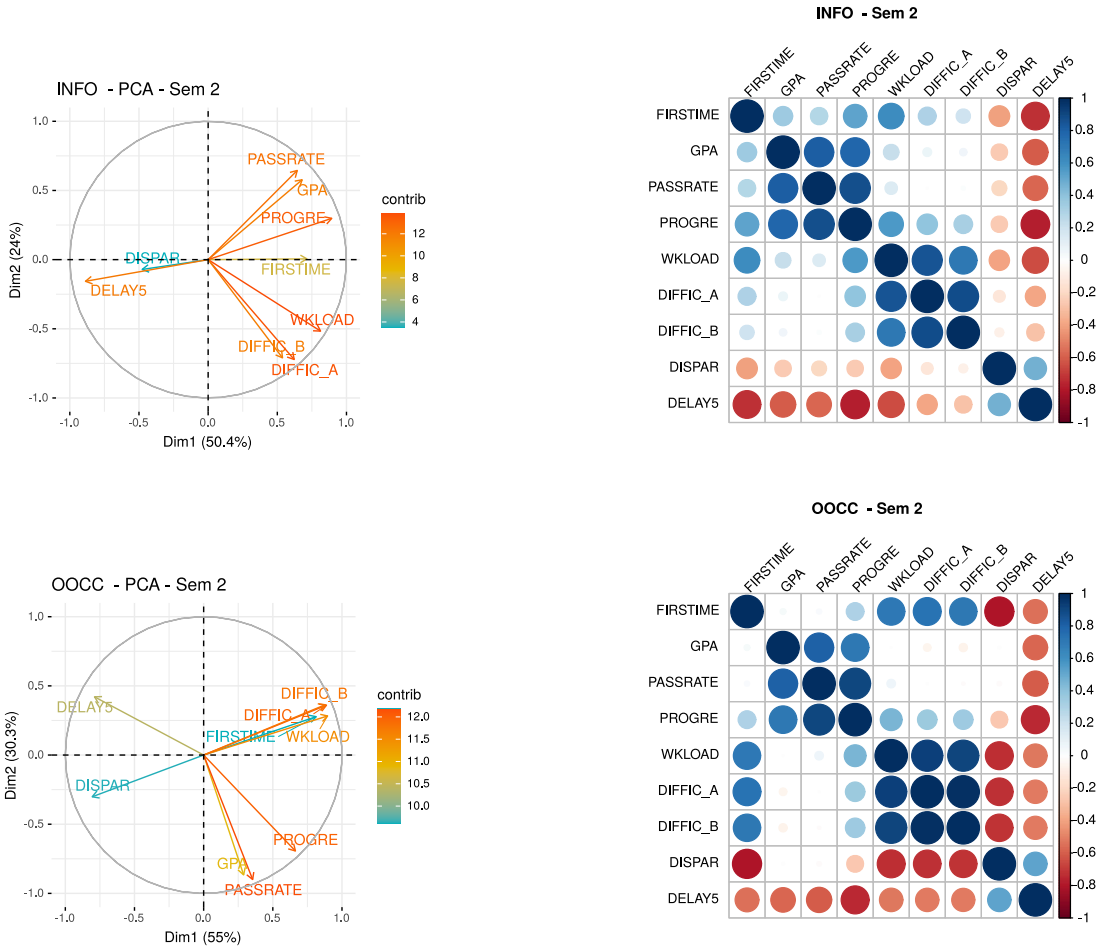

Figura 4: Biplots de los PCA y correlogramas de variables para el semestre 2.

correlación positiva débil con las medidas de dificultad de los campos (DIFFIC_A y DIFFIC_B), y una correlación negativa débil con la medida de disparidad (DISPAR). Esta observación se mantiene en ambos programas. Los dos componentes principales del PCA (ver lado izquierdo de la Figura 4) muestran dos grupos de variables: i) las relacionadas con el desempeño individual de los estudiantes: PASS_R, PROGRE y GPA; y ii) otras relacionadas con las características del programa de estudios: WKLOAD, DIFFIC_A y DIFFIC_B. Es interesante que los distintos componentes pueden ser representados por el rendimiento y la carga de trabajo. Estos análisis respaldan el valor informativo del conjunto diverso de características, incluidas la dificultad, la carga de trabajo y el rendimiento.

IV-B. Capacidad Predictiva de las Características

Para analizar la capacidad predictiva de las características sobre el retraso (DELAY5), se construyeron dos familias de modelos: SGL_i y SA_i. Los modelos SGL_i usan como predictores solamente a GPA_i y $WKLOAD_i$ en cada semestre. Estas variables fueron elegidas como representantes de los dos grupos de variables que se extraen del análisis de componentes principales: GPA asociado a las variables que caracterizan el desempeño de los estudiantes y WKLOAD, para las variables que caracterizan la carga del programa. El modelo SGL_i se formaliza como:

DELAY5 $\sim$ GPAi + WKLOADi, $i = 1, \cdots, 4$

Los modelos SA_i utilizan como predictores a todas las características de cada semestre, a excepción de $DELAY_i$. Así, SA_i se formaliza como:

DELAY5 $\sim$ GPAi + PASS_Ri + WKLOADi + DIFFIC_Ai + DIFFIC_Bi + DISPARi + FIRST_Ti + PROGREi, $i = 1, \cdots, 4$

Cuadro III: Modelos predictivos para los datos de INFO

	MSE 10-fold cross validation				
		SGL_i (dos variables)		SA_i (todas las variables)	
	N	GLM	SVM	GLM	SVM
Semestre 1	204	0.0194	0.0188	0.0185	0.0174
Semestre 2	204	0.0203	0.0174	0.0131	0.0138
Semestre 3	202	0.0184	0.0168	0.0138	0.0124
Semestre 4	198	0.0236	0.0221	0.0122	0.0151

Las tablas III y IV resumen los resultados de los modelos predictivos obtenidos para los programas de ingeniería INFO y OOCC, respectivamente. Las regresiones se construyeron utilizando un modelo lineal generalizado (GLM) con errores gaussianos e identidad como función de enlace [30]. Los modelos con mejores resultados se construyeron utilizando una

Cuadro IV: Modelos predictivos para los datos de OOCC

	MSE 10-fold cross validation				
		SGL_i (dos variables)		SA_i (todas las variables)	
	N	GLM	SVM	GLM	SVM
Semestre 1	309	0.0188	0.0183	0.0185	0.0184
Semestre 2	308	0.0146	0.0124	0.0109	0.0114
Semestre 3	307	0.0123	0.0112	0.0083	0.0076
Semestre 4	303	0.0205	0.0209	0.0122	0.0113

validación cruzada de 10 iteraciones (*10-fold cross validation*) considerando el error cuadrático medio (MSE) como medida de bondad de ajuste. Para esta tarea, se utilizó la función *glm* predefinida de *R* [31] y la función *tune* del paquete *e1071* de *R* [32], ambas con parámetros estándar. De similar manera, se utilizó el método de regresión de máquinas de vector soporte (SVM) para desarrollar la predicción con los datos disponibles, utilizando las funciones *svm* y *tune*, ambas con parámetros estándar.

En la mayoría de los casos, los modelos SVM se comportaron ligeramente mejor que los GLM, manteniendo las mismas tendencias en los diferentes casos estudiados. Un hallazgo importante es que el poder predictivo en el primer semestre es bastante similar entre el modelo con dos variables (SA_i) y el modelo con todas las variables (SGL_i) para los dos programas de estudio analizados. Por el contrario, para los semestres 2, 3 y 4, el poder predictivo del modelo completo (SA_i) es mayor, con el mayor poder predictivo del riesgo de retraso al quinto semestre (DELAY5) en el semestre 3, para ambos programas. Estos resultados sugieren que el uso de los modelos SA_i podría producir trayectorias más representativas del comportamiento del estudiante que utilizando los modelos SGL_i.

IV-C. Trayectorias de Riesgo de Retraso

Para construir las trayectorias de riesgo de retraso, realizamos un agrupamiento basado en modelos por semestre utilizando la información académica, y luego en cada grupo calculamos el DELAY5 promedio. Nos referimos a este valor como el *riesgo de retraso al quinto semestre* porque los estudiantes aosciados a un determinado grupo tiene un comportamiento académico similar ese semestre (representado por las características académicas consideradas) a otros estudiantes que luego muestran ciertos niveles de retraso. Una trayectoria de riesgo de retraso es entonces la secuencia de estos valores a lo largo de los semestres. Como se dijo anteriormente, formamos trayectorias de los primeros 4 semestres de los estudiantes en el programa. El número de conglomerados obtenidos en cada término (para los dos programas) varía de dos a seis. Dado que cada nodo de trayectoria es el DELAY5 medio de los casos (conocidos) en el grupo, hay un número limitado de trayectorias diferentes. Esta característica hace que las trayectorias asi definidas sean más adecuadas para ser agrupadas y comparadas.

Las tablas V y VI describen los grupos resultantes en términos del riesgo asociado de retraso (D5 en las tablas). Los gráficos de PCA de la Figura 5 muestran las proyecciones de los conglomerados en el primer plano de componentes principales junto con sus relaciones con las variables definidas

Cuadro V: Resumen de los resultados de la agrupación para INFO. *n* es el número de estudiantes en el grupo y DELAY5 (D5) es la media de retraso al 5°semestre de los miembros del grupo.

Semestre	1		2		3		4	
Modelo	EEV		VEV		VEV		VVE	
BIC	7305.904		5081.467		2947.037		2113.103	
N°Grupos	3		4		4		4	
Etiqueta	*n*	D5	*n*	D5	*n*	D5	*n*	D5
1	108	0.244	24	0.05	26	0.053	23	0.048
2	11	0.516	47	0.264	46	0.41	88	0.371
3	85	0.599	48	0.473	66	0.418	58	0.471
4	-	-	85	0.55	64	0.526	29	0.625

para el segundo semestre y programa de estudios analizados. En el caso de INFO, el Grupo 1 incluye alumnos con buen desempeño en general, es decir, valores más altos de las variables GPA, Tasa de Progreso y Aprobado, y valores altos de las variables de carga (Dificultad y Carga de Trabajo) al mismo tiempo. El Grupo 2 es similar al Grupo 1 excepto que sus variables de desempeño muestran valores más bajos. El Grupo 3 es el más disperso y tiene algunos individuos con una carga muy baja. El Grupo 4 tiene valores altos de la variable Disparidad y valores bajos de las variables de desempeño. Estas interpretaciones son consistentes con los valores de riesgo de retraso asociados con cada grupo, presentados en la Tabla V. En particular, los grupos 1 y 2 están representados con valores bajos de riesgo de retraso (0,05 y 0,264 respectivamente), mientras que los grupos 3 y 4 con valores más altos de riesgo de retraso (0,473 y 0,55 respectivamente). En el caso de la agrupación obtenida en el segundo semestre para el programa OOCC, se observa en el primer plano de componentes principales, que el Grupo 2 está bien concentrado con valores altos de las variables de carga y se puede interpretar como los estudiantes sin demora. Por otro lado, el Grupo 1 se asocia a valores altos de la variable Disparidad y tiene mayor variabilidad en el resto de características. Este grupo corresponde a los alumnos que se han retrasado en algún curso. De acuerdo con la interpretación anterior, se puede ver en la Tabla VI que el valor asignado para el riesgo de demora al quinto semestre es mucho más alto para el Grupo 1 (0,44) que para el Grupo 2 (0,17).

La figura 6 muestra gráficos de trayectorias agregadas. Cada línea corresponde a un alumno diferente. Se agregó un ruido de jitter a los puntos para poder percibir la densidad en las figuras. Como referencia, se han agregado marcas horizontales a los gráficos en cada valor equivalente a un semestre de retraso. Según nuestra escala, estos valores se ubican en 0,25; 0,5; 0,75 y 1. La primera fila de gráficas muestra representa todas las trayectorias de cada programa, INFO y OOCC. Las gráficas nos permiten ver las diferentes regiones de *riesgo de retraso* donde los estudiantes se concentran y cómo se mueven hacia un mayor o menor riesgo de retraso entre semestres. Para INFO, hay tres valores posibles de DELAY5 medio (tres posibles nodos de trayectoria) para el semestre 1 y cuatro valores posibles para los semestres 2, 3 y 4. Para OOCC, el análisis proporciona cuatro agrupaciones para el semestre

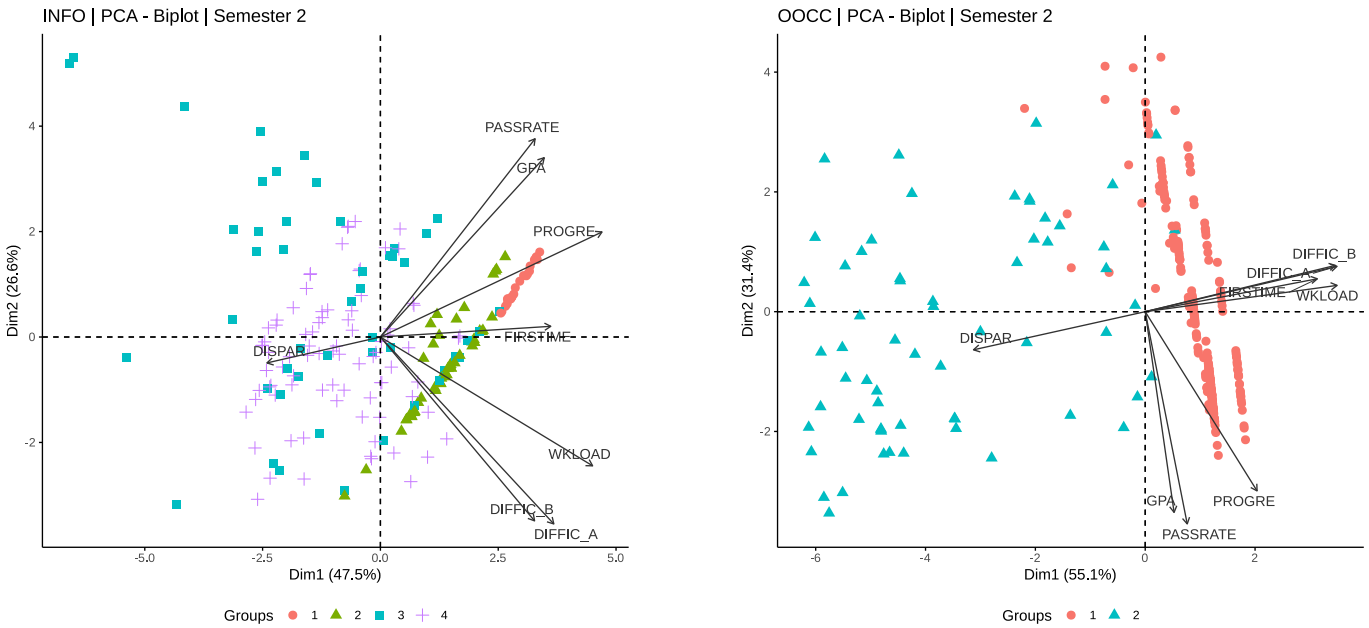

Figura 5: Agrupamiento (clustering) para el segundo semestre, para cada uno de los programas (INFO y OOCC).

Cuadro VI: Resumen de los resultados de la agrupación para OOCC. *n* es el número de estudiantes en el grupo y DELAY5 (D5) es la media de retraso al 5°semestre de los miembros del grupo.

Semestre	1		2		3		4	
Modelo	EEE		EEV		VEV		EEV	
BIC	8889.29		7418.559		10055.65		4212.788	
N°Grupos	4		2		6		4	
Etiqueta	*n*	D5	*n*	D5	*n*	D5	*n*	D5
1	73	0.093	245	0.172	23	0.051	42	0.097
2	77	0.224	63	0.44	77	0.061	134	0.202
3	48	0.284	-	-	13	0.071	113	0.283
4	111	0.296	-	-	58	0.28	14	0.316
5	-	-	-	-	66	0.291	-	-
6	-	-	-	-	70	0.386	-	-

1; dos para el semestre 2; seis para el tercer semestre; y cuatro para el último cuarto semestre. Todas las trayectorias y todos los semestres, tienen valores medios de DELAY5 por debajo de 0,5 (menos de dos semestres de retraso). Aunque hay 6 grupos para el tercer semestre (6 posibles nodos de trayectoria), varios de ellos tienen valores de retardo bastante cercanos. Este hallazgo es importante ya que podría exponer que los estudiantes con diferentes características (asignados a diferentes grupos) pueden tener riesgos de demora similares. Esta situación ilustra que los factores académicos por sí solos pueden no ser suficientes para explicar el fenómeno del retraso.

IV-D. Trayectorias de persistencia y abandono

Comparamos las trayectorias entre los estudiantes que abandonaron y los estudiantes que persisten (DROPOUT = 1 y DROPOUT = 0, respectivamente) después del quinto semestre. Es importante aclarar que esta no es la tasa bruta de deserción, sino aquella que considera solo la población de estudiantes que tienen al menos 5 semestres inscritos. Para INFO, la proporción de alumnos que abandonaron (después del 5º semestre) alcanza el 19,6 %, mientras que para OOCC es del 8,1 %. La Figura 6, segunda y tercera fila, presenta gráficas de trayectorias para los programas INFO y OOCC para estudiantes persistentes (DROPOUT = 0, en segunda fila) y que abandonan (DROPOUT = 1, en tercera fila). Estos gráficos utilizan diferentes colores para ayudar a visualizar las trayectorias más comunes dentro de los grupos de persistencia y abandono en cada programa. A partir de estos gráficos pudimos ver cómo la persistencia se asocia a mantener niveles más bajos de riesgo de retraso, y este patrón es más evidente en OOCC. Sin embargo, hay una cantidad considerable de casos en INFO con altos niveles sostenidos de riesgo de retraso que persisten a lo largo del programa y no abandonan. Ampliamos estas observaciones en la sección de discusión.

V. Discusión y Conclusiones

En este trabajo aplicamos métodos estadísticos y de minería de datos para describir trayectorias académicas relacionadas con el *retraso* acumulado, entendido como el desplazamiento entre el progreso académico del estudiante y el programa curricular secuencial. El retraso académico es un factor importante porque está directamente asociado a beneficios como las becas. El retraso académico es dinámico, puede crecer a medida que el estudiante reprueba los cursos y no puede cumplir con los requisitos para los siguientes cursos; y también puede reducirse, ya que el estudiante puede sobrecargarse al registrar muchos cursos y tener éxito en ellos (si falla, ¡el retraso es aún mayor!). La dinámica del retraso académico depende de una serie de factores académicos y nos fijamos como objetivo relacionar el retraso académico con diversa información académica en cada semestre.

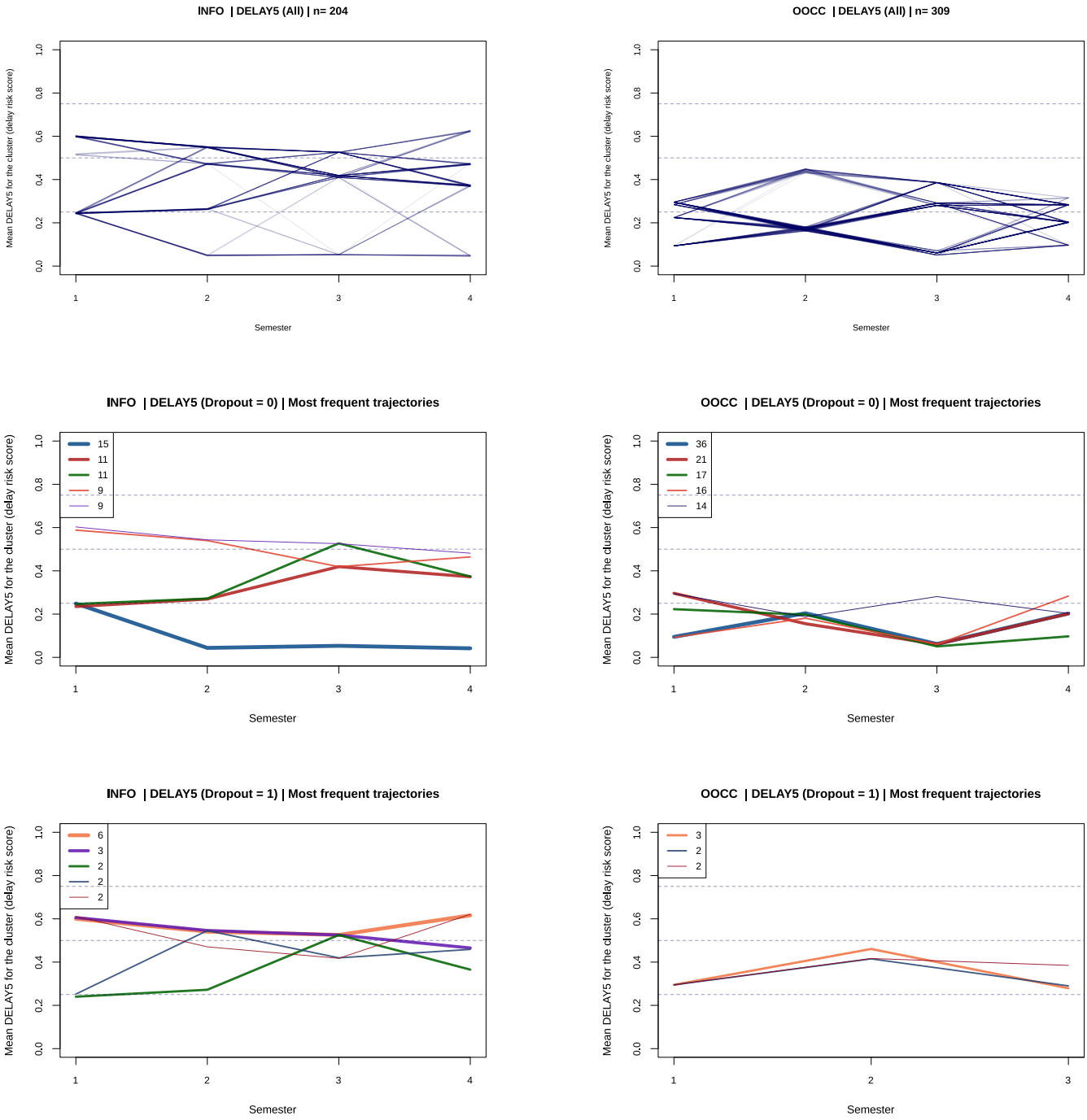

Figura 6: Trayectorias de riesgo de retraso para INFO (izquierda) y OOCC (derecha).

Para responder a la primera pregunta de investigación *(RQ1) ¿Cómo se pueden caracterizar las trayectorias académicas que involucran retraso, carga de cursos, dificultad y desempeño?*, modelamos y seleccionamos características que reflejan las diferentes dimensiones académicas de los estudiantes, como el desempeño, la carga de trabajo y dificultad del curso. El análisis exploratorio de datos de esta información en cada semestre muestra dos conjuntos distintos de variables, ortogonales dentro de las dos primeras componentes principales de un PCA: uno con todas las variables de rendimiento académico y el otro que incluye variables de carga de trabajo y dificultad. Luego, construimos un modelo del retraso acumulado en un período dado basado en las otras características académicas de períodos anteriores. Una exploración de la capacidad predictiva de las características revela que un modelo de retraso con todas las variables es más preciso que otros modelos con menos características. Este hallazgo sugiere que las trayectorias podrían ser más informativas si contienen todas las características propuestas. Luego usamos un método novedoso para relacionar la información académica semestre a semestre con el *riesgo de retraso* en el quinto semestre que corresponde al retraso potencial dado el retraso promedio en el quinto semestre de otros estudiantes con información académica similar en el semestre. Elegimos el quinto semestre porque los primeros cuatro semestres de los programas de ingeniería en estudio son el hito de Bachillerato que incluye todos los cursos básicos de matemáticas y física que registran las tasas de reprobación más altas.

Las trayectorias de riesgo de retraso generadas se pueden agrupar y trazar juntas para inspeccionar patrones. Por ejemplo, muestran que el desempeño del estudiante del primer semestre no determina estrictamente el retraso al quinto semestre.

Los que tengan un rendimiento inicial bajo pueden seguir un camino de mejora progresiva y reducir su posible retraso al finalizar el programa. Este es un fenómeno comúnmente observado en las trayectorias académicas en nuestro contexto: los estudiantes se apresuran a ponerse al día y, en ocasiones, se exponen a una mayor carga de cursos, motivados por la necesidad de reducir los costos de la educación, que aún dependen en gran medida de los recursos familiares. Una línea de investigación interesante consiste en descubrir los efectos de esta sobrecarga en la vida académica y cómo la estructura del programa podría facilitar o limitar estos fenómenos.

Luego nos enfocamos en la relación entre el retraso y el famoso indicador de éxito educativo: la deserción. Para responder a *(RQ2) ¿Qué relaciones existen entre las trayectorias de retraso y la deserción tardía?*, contrastamos las trayectorias típicas entre los estudiantes que luego abandonaron y los estudiantes que completaron los estudios, revelando patrones diferentes en dos programas de ingeniería. En los dos programas analizados, Ingeniería Civil en Informática (INFO) e Ingeniería Civil en Obras Civiles (OOCC), las trayectorias con alto y constante riesgo de retraso caracterizan a los estudiantes que desertan. Las diferencias entre los grupos de trayectorias son más evidentes en el programa INFO que en OOCC. Sin embargo, es necesario señalar que el programa OOCC presenta una menor proporción de estudiantes que finalmente abandonan.

Para obtener información sobre la interpretabilidad de las trayectorias elaboradas, nos reunimos con la directora del programa INFO para discutir las cifras que muestran el riesgo de trayectorias de retraso (ver Figura 6). Explicamos cómo se generaron las trayectorias y nos enfocamos en comparar grupos de trayectorias relacionadas con la persistencia, y aquellas relacionadas con abandono. De esta sesión surgieron dos conclusiones principales. Primero, las trayectorias son interpretables: luego de observar las figuras durante un par de minutos, la entrevistada comenzó a notar similitudes y diferencias entre las tramas y a asociarles explicaciones. Por ejemplo, relacionó las trayectorias que muestran un alto riesgo de retraso con los repetidos fracasos de los cursos de álgebra y geometría del primer semestre. También contrastó las trayectorias de alto riesgo de retraso que no están asociadas a deserción y aquellas que si se asocian con deserción, relacionando el grupo sin deserción con los estudiantes que, aunque están reprobando cursos de matemáticas repetidamente, probablemente lo estén haciendo bien y motivándose con los cursos más específicos del programa, como los cursos de programación (en el caso del programa informático). La segunda conclusión es que la representación de las trayectorias ciertamente puede mejorar, ya que no son fáciles de explicar, no son muy informativas de la información académica relacionada que podría sustentar el tipo de interpretaciones realizadas. Nuestro trabajo futuro incluye extender la representación de la trayectoria con indicadores de agrupamiento (*clustering*) de la información académica. De esta manera, las interpretaciones podrían estar mejor respaldadas.

Nuestro trabajo es novedoso al plantear el problema del retraso como factor significativo en el éxito o fracaso académico. La predicción de la persistencia y la deserción en la educación superior son campos ampliamente abordados. El fenómeno de la graduación retrasada no sigue el mismo camino. Este desequilibrio puede derivarse probablemente de la existencia de una sobrerrepresentación en la investigación de casos de regiones con sistemas de educación superior flexibles o basados en créditos, donde la dinámica de retraso ocurre mayoritariament al ingreso y no en la graduación. La literatura existente no ofrece evidencia irrefutable de cuáles son los determinantes clave del retraso y la deserción. Sin embargo, sugiere que una compleja gama de características individuales e institucionales determina los fracasos universitarios [33]. Los estudiantes enfrentan continuamente decisiones estratégicas sobre, por ejemplo, qué programa elegir o qué cursos inscribirse (o anular). Si bien inscribir tantos cursos como sea posible, podría ayudar al estudiante a recuperarse o ponerse al día y reducir su retraso académico, existe el riesgo de sobrecargar el semestre, lo que podría terminar en reprobar aún más cursos y aumentando el retraso. Debido a que la toma de decisiones académicas requiere experiencia y aprendizaje, los estudiantes a menudo muestran un desempeño cambiante durante su estadía en la universidad [33].

Los conjuntos de datos utilizados en nuestro trabajo no incluyen información de antecedentes demográficos, socioeconómicos o académicos previos de los estudiantes. Acceder a información sobre otros aspectos personales como el

emocional (excepto para un estudio específico de esa área específica) podría resultar poco práctico. Aun así, considerar solo los factores emocionales o sociales iniciales puede no ser suficiente para reflejar su influencia a lo largo de la vida académica porque estos factores son siempre dinámicos. Dado que solo utilizamos datos relacionados con la actividad académica, hemos propuesto novedosas características para abordar la multidimensionalidad y la naturaleza dinámica de la información académica. Las características que definimos también reflejan características institucionales al incluir información sobre el diseño del plan de estudios (la secuencia de cursos) y la dificultad histórica de los cursos.

Dado que la información académica se utiliza para la agrupación no supervisada dentro del método, las trayectorias podrían formarse eventualmente con cualquier información académica disponible, como explican los autores originales del método [25].

De esa manera, este método podría aplicarse en una gama más amplia de instituciones o situaciones. Las modificaciones curriculares también agregan complejidad al análisis de las situaciones de retraso. El Proceso Europeo de Bolonia de reforma de la educación superior [El Proceso de Bolonia se refiere a una serie de reuniones y acuerdos entre países europeos que "buscan generar más coherencia a los sistemas de educación superior en toda Europa" - http://www.ehea.info] ha influido en una serie de cambios en la educación superior en sus países miembros y los que le siguieron como referencia [34]. Por ejemplo, muchas universidades en Chile han llevado a cabo un proceso de innovación curricular desde 1999 [35]. Nuestro método es apropiado en este escenario, ya que puede obtener conjuntos finitos comparables de trayectorias académicas incluso cuando los planes de estudio del programa han cambiado drásticamente.

Los métodos que utilizamos fueron adecuados para responder a nuestras preguntas de investigación. Sin embargo, también tienen limitaciones. La falta de información demográfica y de antecedentes de los estudiantes es una de ellas. La ausencia de un "semestre 0" en las trayectorias, incluyendo información previa de los estudiantes, no nos permite explorar la influencia de estos factores. El trabajo futuro tendrá como objetivo resolver las limitaciones actuales mediante la búsqueda de colaboraciones que nos permitan acceder a conjuntos de datos incluyendo más dimensiones de los estudiantes como la información preuniversitaria y demográfica inicial. También podría ser necesario modelar y probar características adicionales de factores existentes o nuevos. Y a partir de ahora, exploraremos los resultados de nuestros modelos aplicados a un conjunto más amplio de programas de estudio .

Acknowledgment

Trabajo financiado por la Universidad Austral de Chile y el proyecto LALA (Código 586120-EPP-1-2017-1-ES-EPPKA2-CBHE-JP). Este proyecto ha sido financiado con el apoyo de la Comisión Europea. Esta publicación refleja únicamente la puntos de vista de los autores, y la Comisión no se hace responsable de ningún uso que pueda hacerse de la información contenida en el mismo.

Referencias

[1] OECD, *Education at a Glance: OECD Indicators*. OECD, 2019. [Online]. Available: https://www.oecd-ilibrary.org/education/education-at-a-glance-2019_f8d7880d-en

[2] M. M. Ferreyra, C. Avitabile, J. Botero Alvarez, F. Haimovich Paz, and S. Urzúa, *At a Crossroads: Higher Education in Latin America and the Caribbean*. World Bank, Washington, DC, may 2017.

[3] SIES Servicio de Información de Educacion Superior del Ministerio de Educación de Chile, "Actual Duration and Over Duration Report of Higher Education Degrees / Informe Duración Real y Sobreduración e las carreras de Educación Superior (2014-2018)," Ministerio de Educación de Chile, Tech. Rep., 2020.

[4] Aequalis Foro de Educación Superior, "Estimation of Government and Family Spending to Finance the Over Duration of Students in their Degrees Programs: The Chilean Case / Estimación del gasto fiscal y familiar para financiar la sobreduración de los estudiantes en las carreras: caso chileno," Aequalis Foundation, https://aequalis.cl/, Tech. Rep., 2019.

[5] E. Himmel, "Modelo de análisis de la deserción estudiantil en la educación superior." *Calidad en la Educación*, no. 17, 2002.

[6] A. Abu, "Educational Data Mining & Students' Performance Prediction," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 5, 2016.

[7] G. Mendez, X. Ochoa, K. Chiluiza, and B. De Wever, "Curricular Design Analysis: A Data-Driven Perspective," *Journal of Learning Analytics*, vol. 1, no. 3, 2014.

[8] Z. Mabel and T. A. Britton, "Leaving late: Understanding the extent and predictors of college late departure," *Social Science Research*, vol. 69, 2018.

[9] SIES Servicio de Información de Educacion Superior del Ministerio de Educación de Chile, "1st Year Undergraduate Retention Report / Informe retención de 1er año de pregrado (2014-2018)," Ministerio de Educación de Chile, Tech. Rep., 2019.

[10] OECD, *Education at a Glance 2020: OECD Indicators*. Paris: OECD Publishing, 2020. [Online]. Available: https://doi.org/10.1787/69096873-en

[11] R. Boegeholz, J. Guerra, and E. Scheihing, "Modeling Trajectories to Understand the Delayed Completion of Sequential Curricula Undergraduate Programs. BT - Proceedings of the Workshop on Adoption, Adaptation and Pilots of Learning Analytics in Under-represented Regions co-located with the 15th Eur," pp. 24–47, 2020. [Online]. Available: http://ceur-ws.org/Vol-2704/paper3.pdf

[12] Z. Papamitsiou and A. A. Economides, "Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence," *Journal of Educational Technology & Society*, vol. 17, no. 4, pp. 49–64, dec 2014. [Online]. Available: http://www.jstor.org.uchile.idm.oclc.org/stable/jeductechsoci.17.4.49

[13] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," *Computers and Education*, vol. 113, 2017.

[14] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, 1986.

[15] R. Campagni, D. Merlini, and M. C. Verri, "The influence of first year behaviour in the progressions of university students," in *Communications in Computer and Information Science*, vol. 865, 2018.

[16] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in *5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.

[17] M. Kutner, C. Nachtsheim, J. Neter, and W. Li, *Applied Linear Statistical Models.*, 5th ed. New York: McGraw-Hill, Irwin, 2005.

[18] A. Elbadrawy, A. Polyzou, Z. Ren, M. Sweeney, G. Karypis, and H. Rangwala, "Predicting Student Performance Using Personalized Analytics," *Computer*, vol. 49, no. 4, 2016.

[19] M. Alban and D. Mauricio, "Predicting University Dropout trough Data Mining: A systematic Literature," *Indian Journal of Science and Technology*, vol. 12, no. 4, pp. 1–12, 2019.

[20] S. Stewart, D. H. Lim, and J. Kim, "Factors Influencing College Persistence for First-Time Students," *Journal of Developmental Education*, vol. 38, no. 3, pp. 12–20, 2015. [Online]. Available: http://www.jstor.org.uchile.idm.oclc.org/stable/24614019

[21] A. J. Martin, R. Wilson, G. A. D. Liem, and P. Ginns, "Academic Momentum at University/College: Exploring the Roles of Prior Learning, Life Experience, and Ongoing Performance in Academic Achievement across Time," *The Journal of Higher Education*, vol. 84, no. 5, 2013.

[22] C. Haas and A. Hadjar, "Students' trajectories through higher education: a review of quantitative research," 2020.

[23] O. Almatrafi, A. Johri, H. Rangwala, and J. Lester, "Identifying course trajectories of high achieving engineering students through data analytics," in *ASEE Annual Conference and Exposition, Conference Proceedings*, vol. 2016-June, 2016.

[24] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, 1993.

[25] M. J. Mahzoon, M. L. Maher, O. Eltayeby, W. Dou, and K. Grace, "A Sequence Data Model for Analyzing Temporal Patterns of Student Data," *Journal of Learning Analytics*, vol. 5, no. 1, 2018.

[26] R. Robinson, "Pathways to completion: Patterns of progression through a university degree," *Higher Education*, vol. 47, no. 1, 2004.

[27] J. Salazar-Fernandez, M. Sepulveda, J. Munoz-Gama, and M. Nussbaum, "Curricular analytics to characterize educational trajectories in high-failure rate courses that lead to late dropout," *Applied Sciences*, vol. 11, p. 1436, 02 2021.

[28] C. C. Chang and C. J. Lin, "LIBSVM: A Library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, 2011.

[29] L. Scrucca, M. Fop, T. B. Murphy, A. E. Raftery, and Tinto, "mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture ModTitle," *The R Journal*, vol. 12, no. 1, pp. 1–12, dec 2019. [Online]. Available: http://www.jstor.org.uchile.idm.oclc.org/stable/24614019 http://www.jstor.org.uchile.idm.oclc.org/stable/jeductechsoci.17.4.49

[30] J. Chambers and T. Hastie, *Statistical Models in S*. Pacific Grove, California: Wadsworth & Brooks/Cole, 1992.

[31] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018. [Online]. Available: https://www.R-project.org/

[32] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2021, r package version 1.7-7. [Online]. Available: https://CRAN.R-project.org/package=e1071

[33] C. Aina, E. Baici, G. Casalone, and F. Pastore, "The Economics of University Dropouts and Delayed Graduation: A Survey," 2018, visited March 10, 2021. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3153385

[34] Z. Y. Mngo, "Probing the progress of the external dimension of the Bologna process," *PSU Research Review*, vol. 3, no. 3, 2019.

[35] R. Pey and S. Chauriye, "Curricular Innovation in Universities of the Rectors' Council 2000-2010 / Innovación curricular en las universidades del Consejo de Rectores 2000-2010," Consejo de Rectores de las Universidades Chilenas, CRUCH, Santiago, Chile, Tech. Rep., 2011.