Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación el 29 de noviembre de 1976.

Departamento de Electrónica, Sistemas e Informática

MAESTRÍA EN INFORMÁTICA APLICADA

MODELO PREDICTIVO DE RIESGOS EN COMPETENCIA PARA EVENTOS DE HISTORIA DE VIDA DE ALUMNOS DE UNA INGENIERÍA EN EL ITESO

Trabajo recepcional para obtener el grado de

MAESTRA EN INFORMÁTICA APLICADA

Presentan: Fuentes García, Samantha

Asesor: Mtra. Hernández Chávez, Gisel

San Pedro Tlaquepaque, Jalisco. enero de 2025

Dedicatoria

A mis padres, Alicia García y Fernando Fuentes, y a mi hermana María Fernanda Fuentes, por su amor incondicional y por estar siempre a mi lado, acompañándome en cada paso de este trayecto.

A mi pareja, Sebastián Muñoz, por acompañarme y motivarme en los momentos más desafiantes.

A mis profesores, y en especial a mi asesora la candidata a Dra. Gisel Hernández, por su dedicación, paciencia y guía. Su conocimiento y orientación fueron fundamentales para lograr realizar este trabajo.

A mis queridas mascotas, Brandy, Bongo, Bailey y Mochi, por acompañarme en tantas noches de desvelo, con su compañía y alegría, dándome ánimos en cada momento.

Y, en especial, a mí misma, por la perseverancia y la fuerza para superar los retos, y cerrar un ciclo lleno de aprendizajes (¡¡al fin!!).

Gracias a todos por su paciencia, cariño, motivación y confianza. Este trabajo es el resultado de las enseñanzas, sacrificios y dedicación de todos ustedes, quienes hicieron posible este logro.

Abstract

Este estudio compara modelos de predicción del riesgo de deserción en estudiantes de un programa de ingeniería utilizando análisis de riesgos en competencia. Se considera que los estudiantes pueden experimentar distintos eventos competitivos, como graduación, cambio de carrera o abandono sin completar el programa. Para ello, se implementa el modelo de riesgos en competencias de Fine y Gray.

El modelo se construyó a partir de variables académicas (promedio de preparatoria, preparatoria de procedencia, etc.), socioeconómicas (crédito educativo, beca académica, etc.) y demográficas (edad, género, etc.). Para la selección de variables, se analizaron tres métodos de regularización: Lasso, Ridge y Elastic Net, eligiéndose Lasso por su mejor desempeño. A partir de esto, se comparó un modelo basado en el enfoque de Fine y Gray utilizando las variables seleccionadas por Lasso con otro que emplea todas las variables disponibles.

La evaluación se realizó mediante AUC (*Área Bajo la Curva*) y Brier Score, midiendo la capacidad discriminativa y la calibración de las predicciones. Los resultados indican que el modelo con selección de variables mejora el desempeño predictivo, logrando una representación más parsimoniosa sin pérdida significativa de precisión. Sin embargo, debe considerarse que la limitación en la cantidad de variables podría inducir sobreajuste. Por ello, se recomienda que en estudios posteriores se aplique el modelo a conjuntos de datos más amplios para mejorar su robustez.

Palabras clave*:*

- Fine & Gray,
- Riegos en competencia,
- Abandono escolar,
- Deserción escolar

Tabla de contenidos

Índice de Tablas y Figuras

1. Introducción

1.1 Antecedentes

Fundado en la ciudad de Tlaquepaque, Jalisco, un 31 de julio de 1957 por la Compañía de Jesús, el Instituto Tecnológico y de Estudios Superiores de Occidente (ITESO) es una institución de educación superior privada con inspiración cristiana, comprometida con la excelencia académica y la responsabilidad social. De acuerdo su Estrategia Institucional (2022), se concibe a sí mismo como una comunidad de personas en permanente crecimiento, bajo la inspiración de la tradición educativa jesuita y el análisis constante de la realidad.

En sus primeros años, la universidad se centró en la formación de profesionistas en el área de negocios, ingeniería y arquitectura, sin embargo, también ofrecía programas en filosofía y teología. Con el paso del tiempo, el ITESO amplió su oferta académica para incluir programas en ciencias sociales, comunicación, diseño, ciencias ambientales y tecnologías de la información. Actualmente la universidad cuenta con 41 programas de licenciatura, 6 especialidades, 19 maestrías y 5 doctorados, además de programas de educación continua y, desde el 2021, preparatoria dentro del campus. (ITESO, 2021)

La universidad ha experimentado un constante crecimiento y evolución en las últimas décadas, tanto en la apertura de nuevas carreras y programas académicos como en la infraestructura del campus y especialmente en la matrícula total de estudiantes. De acuerdo con lo reportado en los Informes del Rector de 1998-1999 y 2020-2021, en la primavera de 1999 se registró una inscripción total de 6,796 alumnos, en contraste con la primavera de 2021 donde se registraron 9,960 alumnos inscritos.

En el 2021 como parte de su proceso de planeación quinquenal para el periodo 2021-2026 la universidad definió como visión a 5 años:

> *"Formamos y acompañamos personas para la reconstrucción de un mundo más justo, solidario, compasivo y sostenible, capaces de sostener diálogos interculturales y globales, que impulsen la generación de ciencia e innovación de calidad, con rigor metodológico y pertinencia social, desde nuestra identidad ignaciana, con excelencia académica y humana, de manera integral, así como con el uso de la tecnología poniendo al ser humano al centro."* (ITESO, 2021)

Abonando a esta visión, la candidata a Dra. Hernández desarrolló un modelo predictivo que permite a la universidad estimar la probabilidad de supervivencia de sus alumnos en su primera carrera, así como el nivel de riesgo del evento de abandono (Hernández-Chávez, 2024). Para realizar esta predicción se utilizan datos de los estudiantes, recabados generalmente en el proceso de inscripción, como la preparatoria de procedencia, género, edad, entre otros, así como técnicas no paramétricas utilizadas para estimar la función de supervivencia en el análisis de supervivencia.

Actualmente, la información de este análisis es generada y entregada a la Coordinación de Acompañamiento para la Excelencia Académica (CAXA), con el objetivo de dar acompañamiento y seguimiento a los estudiantes identificados como en riesgo de abandono.

1.2 Planteamiento de la problemática y justificación.

La predicción de supervivencia actualmente se realiza utilizando métodos no paramétricos como el Bosque de Supervivencia Aleatorio (Ishwaran et al., 2008) y semi paramétricos como el modelo de Cox tradicional (Cox, 1972) y regularizado(Liu et al., 2014) para la estimación de un solo riesgo, específicamente el abandono del estudiante de su primera carrera ya sea por cambio de carrera o salida definitiva de la universidad. Sin embargo, se identificó la existencia de otros riesgos que pueden influir en la ocurrencia del evento principal, tal como el egreso. , que hace que los resultados o predicciones pueden estar sesgados debido a la ocurrencia de eventos en competencia.

Este trabajo se enfoca en realizar un análisis de supervivencia utilizando métodos que permiten la estimación de riesgos competitivos y considerando los eventos: salida de la universidad, cambio de carrera y egreso. La intención es evaluar la precisión y validez de las predicciones en contextos con riesgos en competencia, proporcionando una visión más completa y precisa de los factores o variables que influyen en la ocurrencia de cada uno de estos eventos de la historia de vida estudiantil.

1.3 Objetivos

1.3.1 Objetivo general

Identificar y evaluar las variables más relevantes para la predicción de riesgos en competencia mediante un modelo de Fine & Gray con regularización. Además, comparar la precisión predictiva de este modelo con un modelo nulo, basado en estimaciones de Kaplan-Meier (para riesgos no competitivos) y de Aalen-Johansen (para riesgos en competencia), con el fin de evaluar la eficacia de las variables en mejorar las predicciones.

1.3.2 Objetivos particulares

- Identificar las variables más influyentes en la ocurrencia de eventos de riesgo en competencia mediante la aplicación de regularización en el modelo de Fine & Gray.
- Comparar la precisión predictiva de modelos de riesgos en competencia desarrollados con todas las variables, con aquellas seleccionadas a través de regularización.
- Evaluar el rendimiento del modelo de Fine & Gray en comparación con un modelo nulo basado en estimaciones de Kaplan-Meier y Aalen-Johansen, analizando su capacidad predictiva para eventos específicos en un contexto de riesgos en competencia

2. Marco teórico

2.1 Análisis de supervivencia

De acuerdo con Kleinbaum, D. G., & Klein, M (2012), en su libro *Survival Analysis*, el análisis de supervivencia es un conjunto de procedimientos estadísticos para analizar datos donde la variable de interés es el tiempo hasta que ocurre un evento, como muerte, enfermedad o recuperación. Su objetivo principal es determinar el tiempo que transcurre antes de que ocurra un evento específico, y analizar cómo diferentes variables pueden afectar el tiempo de supervivencia. Este tipo de análisis se aplica en diversas áreas, como la medicina, la biología, la ingeniería, la economía y las ciencias sociales, entre otras. Algunas de las técnicas utilizadas en el análisis de supervivencia incluyen: el estimador Kaplan Meier, el estimador Nelson Alen, riesgo relativo, censura, modelo cox-proporcional Hazard, función y curva de supervivencia, entre otros (Kleinbaum & Klein, 2012).

2.2 Censura

La censura ocurre cuando se tiene información sobre el tiempo de supervivencia de un individuo, pero no se conoce exactamente el momento del evento. Hay tres razones principales por las que puede ocurrir la censura: no se experimenta el evento antes de que termine el estudio, se pierde seguimiento durante el estudio o se retira del estudio por alguna razón. La censura afecta el tiempo de supervivencia porque solo se conoce una parte de la información (Kleinbaum & Klein, 2012).

2.3 Función de supervivencia y función de riesgos

De acuerdo con Kleinbaum y Klein (2012), la función de supervivencia es utilizada para determinar la probabilidad de que un individuo sobreviva más allá de un tiempo determinado. Por otro lado, la función de riesgos (*hazard*) se utiliza para modelar la tasa instantánea de fallas en el tiempo (tasa de riego). Amabas funciones son complementarias; la función de supervivencia se puede calcular a partir de la función de riesgos, y viceversa. Son métodos estadísticos útiles para comprender cómo cambian las tasas de “fallo” y probabilidad la supervivencia a lo largo del tiempo, en una población o grupo de individuos (Kleinbaum & Klein, 2012).

$$S(t) = exp\left[-\int_0^t h(u)du\right] \quad ..(1)$$

Función de supervivencia

$$h(t) = -\left[\frac{dS(t)/dt}{S(t)}\right] \quad ..(2)$$

Función de Hazard

2.4 Estimador Kaplan Meier

El estimador de Kaplan-Meier es una técnica no paramétrica utilizada para estimar la función de supervivencia en el análisis de supervivencia. Se basa en el cálculo de la probabilidad de supervivencia en cada punto en el tiempo, utilizando solo los datos disponibles de los individuos que aún no han experimentado el evento de interés (Kleinbaum & Klein, 2012).

$$S(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \quad ..(3)$$

Estimador de Kaplan-Meier

Donde t_i es el tiempo de cada evento, d_i es el número de eventos en t_i , y n_i es el número de individuos en riesgo justo antes de t_i .

2.5 Estimador Nelson-Aalen

El estimador de Nelson-Aalen es un método no paramétrico utilizado en el análisis de supervivencia para estimar la función acumulativa de riesgo acumulada a partir de datos de supervivencia censurados, esto de acuerdo con lo publicado por Ornulf Borgan en su artículo "*Three contributions to the Encyclopedia of Biostatistics"*. Su principal ventaja es que no requiere suposiciones sobre la distribución de los tiempos de supervivencia (Borgan, 1997).

Para una muestra de n individuos con tiempos de muerte observados $t_1, t_2, \ldots, t_k$, donde d_j es el número de individuos que mueren en t_j y r_j es el número de individuos en riesgo justo antes de t_j, el estimador de Nelson-Aalen se calcula como:

$$\hat{A}(t) = \sum_{t_j \leq t} d_j / r_j \quad ..(4)$$

Estimador de Nelson-Aalen

2.6 Estimador Aalen-Johansen

El estimador de Aalen-Johansen es un método no paramétrico utilizado en análisis de supervivencia para estimar probabilidades de transición entre estados en modelos de procesos de Markov con un número finito de estados, incluidos modelos con riesgos en competencia y modelos de enfermedad-muerte. Según lo descrito por Borgan en su contribución a la "*Encyclopedia of Biostatistics*," el estimador extiende el enfoque del estimador de Kaplan-Meier y lo generaliza a contextos donde los individuos pueden experimentar múltiples transiciones entre estados.

El estimador de Aalen-Johansen utiliza integrales de producto para calcular probabilidades de transición entre estados, tomando como base los estimadores de Nelson-Aalen para las intensidades acumuladas de transición. Este método es especialmente útil porque no requiere suposiciones paramétricas estrictas sobre los datos y permite modelar situaciones con censura y truncamiento independiente (Borgan, 1997).

2.7 Prueba Log-Rank

La prueba Log-Rank compara dos o más curvas de supervivencia. Evalúa la hipótesis nula de que no hay diferencias en la supervivencia entre los grupos. La estadística de la prueba se basa en la diferencia observada y esperada de eventos en cada grupo:

$$x^2 \approx \frac{[\sum(O_i - E_i)]^2}{\sum V_i} \quad ..(5)$$

Prueba Long-Rank

Donde $O_i$es el número observado de eventos, E_i es el número esperado, y V_i es la varianza.

2.8 Riesgos en competencia

En estudios de supervivencia, los riesgos en competencia son eventos que impiden la ocurrencia del evento de interés. Por ejemplo, en investigaciones enfocadas en la mortalidad por causas cardiovasculares, cualquier muerte por otras causas (no cardiovasculares) actúa como un riesgo en competencia. Para manejar correctamente los riesgos en competencia en el análisis de supervivencia, es necesario utilizar métodos que permitan una interpretación del impacto de las covariables y la incidencia de eventos. Los métodos como los estimadores de Kaplan-Meier pueden resultar sesgados, ya que no consideran la posibilidad de que ocurra un evento en competencia. Por esta razón, se recomienda recurrir a modelos como el desarrollador por Fine & Gray, que permite obtener estimaciones precisas en estudios que involucran múltiples tipos de eventos que pueden influir entre sí, a través de por ejemplo el cálculo de la función de incidencia acumulada (CIF) (Austin, Lee, & Fine, 2016).

2.9 Modelo Cox Proporcional Hazard (*Cox PH*)

El propósito del modelo de riesgos proporcionales de Cox es evaluar simultáneamente el efecto de múltiples factores sobre la supervivencia. Este modelo permite examinar cómo factores específicos influyen en la tasa a la que ocurre un evento particular en un momento dado, comúnmente conocida como tasa de riesgo o *Hazard Rate*. Las variables predictoras, o covariables en el análisis de supervivencia, se pueden evaluar eficazmente mediante este modelo.

La popularidad del modelo de Cox radica en su naturaleza semi-paramétrica y su robustez. Esto significa que, sin necesidad de especificar la forma de la función de riesgo, el modelo proporciona resultados que se aproximan adecuadamente a los obtenidos con modelos paramétricos, como Weibull y exponencial, siempre que

el modelo sea correcto. Este motivo vuelve al modelo de Cox PH una opción confiable y segura especialmente si se desconoce la distribución de los datos. (Kleinbaum & Klein, 2012).

$$h(t, X) = h_0(t) e^{\sum_{i=1}^{p} \beta_i X_i} \quad ..(6)$$

Función Cox PH

Donde $h_0(t)$ es la función de riesgo basal (no paramétrica), β_i es un vector de coeficientes de regresión.

2.10 Modelo de Fine & Gray

El artículo "*A Proportional Hazards Model for the subdistribution of a Competing Risk*" de Jason P. Fine y Robert J. Gray introducen un modelo semi-paramétrico de riesgos proporcionales para el análisis de riesgos competitivos. Este modelo permite evaluar el efecto de las covariables en la función de incidencia acumulativa (CIF), que es la probabilidad de falla específica en presencia de riesgos competitivos. La estimación se basa en una verosimilitud parcial adaptada para el riesgo de subdistribución y utiliza una función de puntuación ponderada para manejar datos censurados a la derecha.

$$F_1(t; \boldsymbol{Z}) = 1 - exp\left[-\int_0^t \lambda_{10}(s) exp\{\boldsymbol{Z}^T(s)\beta_0\} ds\right] \quad ..(7)$$

Función de Incidencia Acumulativa (CIF)

$$\lambda_1\{t; \boldsymbol{Z}\} = \lambda_{10}(t) exp\{\boldsymbol{Z}^T(t)\beta_0\} \quad ..(8)$$

Riesgo de Subdistribución

Donde $\lambda_{10}(t)$ es la función de riesgo basal de subdistribución, **Z** es el vector de covariables, y β coeficientes de regresión que indican el efecto de las covariables.

2.11 Escalación de características (*Feature Scaling*)

La escalación de características, entre las que se encuentra la normalización y la estandarización de datos, es una técnica utilizada en el aprendizaje automático y el preprocesamiento de datos para ajustar diferentes características de un conjunto de datos a una escala similar, con el objetivo de que todas las características influyan de manera equitativa en el aprendizaje, evitando sesgos.

- Estandarización o escalamiento z-score: es un procesamiento en estadística que transforma los datos para que tengan una media de 0 y una desviación estándar de 1. Funciona restando la media de cada dato y luego dividiéndolo por la desviación estándar. Este método es especialmente útil cuando los datos están aproximadamente distribuidos de forma normal y es fundamental para estandarizar valores,

permitiendo analizar y comparar datos de diferentes escalas de manera efectiva. Se utiliza comúnmente en métodos estadísticos y pruebas de hipótesis que asumen una distribución normal. (Tay, et al., 2023)

$$Z = \frac{X_i - mean(X)}{std(X)} \quad ..(10)$$

Formula Z-score

- Escalamiento Min-Max (Normalización): transforma los datos originales a un nuevo rango, típicamente entre 0 y 1. Este proceso ajusta los valores de las características de los datos para que se encuentren dentro de este rango específico, manteniendo las relaciones relativas entre los datos originales. (Tay, et al., 2023)

$$X_{scaled} = \frac{X_i - min(X)}{max(X) - min(X)} \quad ..(11)$$

Formula Escalamiento Min-Max

- Escalamiento robusto: se basa en los cuantiles primero, segundo y tercero para escalar los datos. Para aplicar el escalado robusto a una variable, primero se deben encontrar los cuantiles de esa variable. Luego, se resta la mediana (segundo cuantil o Q2) de cada valor de los datos y se divide el resultado por el Rango Intercuartílico (IQR), que es la diferencia entre el tercer y el primer cuantil. Estos cálculos se realizan de manera simultánea para un conjunto de características. El escalado robusto es especialmente útil cuando hay valores atípicos en los datos, ya que los cuantiles son resistentes a los outliers. (Simon, Friedman, Hastie, & Tibshirani, 2011)

$$x_{robust} = \frac{x - median(x)}{IQR(x)} \quad ..(12)$$

Formula Escalamiento Robusto

2.12 Regularización

La regularización es una técnica utilizada en modelos de aprendizaje automático para prevenir el sobreajuste (*overfitting*) y mejorar la generalización del modelo. Se logra agregando una penalización al costo del modelo, basándose en la magnitud de los coeficientes de los predictores. A continuación, se describen los tipos más comunes de regularización: Lasso, Ridge y ElasticNet, junto con sus respectivas fórmulas (Simon, Friedman, Hastie, & Tibshirani, 2011).

- Lasso: Lasso es una técnica de regularización que penaliza la suma de los valores absolutos de los coeficientes. Esta técnica puede forzar a algunos coeficientes a ser exactamente cero, lo que resulta en modelos más simples y fáciles de interpretar (Simon, Friedman, Hastie, & Tibshirani, 2011).

$$min_{\beta}\left(\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right) \quad ..(5)$$

Fórmula Lasso

Donde y_i es el valor observado, x_{ij} es el valor de la característica j para la observación i, β_0 es el término de intersección, β_j son los coeficientes del modelo y λ es el parámetro de regularización que controla la fuerza de la penalización.

- Ridge: es una técnica de regularización que penaliza la suma de los cuadrados de los coeficientes. A diferencia de Lasso, Ridge no fuerza a los coeficientes a ser cero, sino que los reduce hacia cero (Simon, Friedman, Hastie, & Tibshirani, 2011).

$$min_{\beta}\left(\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2\right) \quad ..(6)$$

Fórmula Ridge

Donde y_i es el valor observado, x_{ij} es el valor de la característica j para la observación i, β_0 es el término de intersección, β_j son los coeficientes del modelo y λ es el parámetro de regularización que controla la fuerza de la penalización.

- ElasticNet: es una técnica que combina las penalizaciones de Lasso y Ridge. Esta técnica es útil cuando hay múltiples características correlacionadas, ya que puede seleccionar un grupo de características correlacionadas (Simon, Friedman, Hastie, & Tibshirani, 2011).

$$min_{\beta}\left(\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda_1\sum_{j=1}^{p}|\beta_j|+\lambda_2\sum_{j=1}^{p}\beta_j^2\right) \quad ..(7)$$

Fórmula ElasticNet

Donde y_i es el valor observado, x_{ij} es el valor de la característica j para la observación i, β_0 es el término de intersección, β_j son los coeficientes del modelo y λ_1 y λ_2 son los parámetros de regularización que controlan la fuerza de las penalizaciones Lasso y Ridge, respectivamente.

2.13 Área Bajo la Curva (AUC)

El AUC (*Area Under the Curve*) es una métrica utilizada en la evaluación de modelos de clasificación binaria. Representa el área bajo la curva ROC (*Receiver Operating Characteristic*), la cual muestra la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos para todos los umbrales posibles. Un valor de AUC más cercano a 1 indica que el modelo tiene un excelente desempeño en distinguir entre clases positivas y negativas, mientras que un valor de 0.5 refleja un rendimiento equivalente al azar. Esta métrica es especialmente útil en

conjuntos de datos con clases desbalanceadas, ya que no se ve afectada por cambios en la proporción de clases (Sadafule & Sarkar, 2022).

2.14 Brier Score

El Brier Score mide la precisión de los pronósticos probabilísticos para eventos binarios. Se calcula como el promedio del cuadrado de la diferencia entre la probabilidad pronosticada y el resultado real (0 o 1), esto de acuerdo con el artículo "*Verification of forecast expressed in terms of probability"* (Bradley, Schwartz, & Hashino, 2007). La fórmula es:

$$Brier\ Score = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2 \quad ..(14)$$

Formula Brier Score

Donde N es el número de pronósticos, f_i es la probabilidad pronosticada para el evento i, o_i es el resultado observado (1 si el evento ocurrió, 0 si no ocurrió). Este puntaje varía entre 0 (precisión perfecta) y 1 (peor precisión).

.

3. Trabajos relacionados

3.1 Fuentes de búsqueda y criterios de inclusión.

Para la búsqueda de trabajos relacionados con el tema de investigación, se realizaron consultas en las siguientes fuentes de información: IEEE Xplore, EBSCO, y Google Académico.

En esta búsqueda se consideraron los siguientes criterios para mejorar el análisis de los trabajos relacionados:

- El estudio utiliza modelos predictivos de riesgos en competencia.
- El campo de estudio es la educación y aborda el abandono de programas educativos.
- Los trabajos están en idioma inglés o español.
- No se incluirán estudios sobre programas de educación superior como certificaciones, diplomados, maestrías, especialidades o doctorados.
- No se incluirán estudios sobre educación a distancia, incluyendo modalidades en línea, MOOC, e-learning o aprendizaje remoto.
- Se analizará si se pueden considerar estudios relacionados con deserción o rotación laboral.
- Los artículos no tendrán una fecha de publicación anterior a 2014.
- Se incluirán aquellos a los que se tenga acceso a través de bases de datos públicas definidas o aquellas adquiridas por el Instituto Tecnológico y de Estudios Superiores de Occidente (ITESO).

3.2 Resultados de las búsquedas

Al realizar la búsqueda en IEEE Xplore el 25 de agosto de 2024 se obtuvieron 2 resultados, sin embargo, ninguno está relacionado con el tema de estudio; los criterios fueron:

> *(("All Metadata":competing risk) AND ("All Metadata":Predictive model OR "All Metadata":Prediction OR "All Metadata":predict OR "All Metadata":fine and gray) AND ("All Metadata":university OR "All Metadata":student OR "All Metadata":education) AND ("All Metadata":withdraw OR "All Metadata":abandon OR "All Metadata":retention OR "All Metadata":attrition OR "All Metadata":dropout)))*

También se realizó una búsqueda con la palabra 'Fine and Gray', sin resultados relevantes. Cuando se realizó la búsqueda, en la misma plataforma, de las palabras *'prediction with competing risk'* en el rango de 2014 a 2024, se encontraron 68 resultados, de los cuales ninguno estaba relacionado con el tema de investigación.

En cuanto a las búsquedas en la base de datos EBSCO, realizadas igualmente el 25 de agosto de 2024, utilizando el término principal *'competing risk'* y filtrando por el rango de fechas de 2014 a 2024, así como por las temáticas de logro académico, programas académicos, resultados educativos y educación superior. Esta búsqueda arrojo las siguientes cinco publicaciones académicas:

Tabla 1. Resultados de las búsquedas de Trabajos Relacionados

	Autores	*Título de artículo*	*Tema general del artículo*	*Método Utilizado*
1	Kemda, Lionel Establet Murray, Michael	Joint modeling of the longitudinal student mark and the competing events of degree completion and academic dropout.	Modelado del rendimiento académico y el tiempo hasta la finalización de la carrera o el abandono académico en estudiantes	Análisis longitudinal del rendimiento académico con un modelo de riesgos competitivos
2	Castro-Montoya, Bibiana Lopera-Gómez, Carlos Manrique-Hernández, Rubén Gonzalez-Gómez, Difariney	Modelo de riesgos competitivos para deserción y graduación en estudiantes universitarios de programas de pregrado de una universidad privada de Medellín (Colombia).	Análisis de factores demográficos, socioeconómicos y académicos en deserción y graduación universitaria.	Modelo de tiempo discretos y riesgos competitivos
3	Zahra, Fatima	High Hopes, Low Dropout: Gender Differences in Aspirations for Education and Marriage, and Educational Outcomes in Rural Malawi.	Relación entre aspiraciones educativas y matrimoniales con la deserción escolar en niñas de Malawi.	Modelos de tiempo discreto y de riesgos competitivos

Por último, en la plataforma Google Académico, se encontraron 10 artículos relacionados con trabajos, sin embargo ninguno de estos estaba relacionado con el tema a estudiar. Estos fueron los criterios de búsqueda:

[competing risk] [withdraw OR abandon OR retention OR attrition OR enrollment OR dropout OR spell OR "stop out"] AND [university OR college OR student] AND [predict OR survival OR fine OR gray OR "time to event" OR machine OR learning OR model OR event OR history] AND [competing risk] NOT [online OR "e learning" OR distance OR "remote learn" OR certifications OR diplomas OR courses OR master OR specializations OR doctorates OR doctoral OR phd OR teachers OR educators OR clinical OR patient OR "multiple spell"]

4. Metodología de desarrollo

El análisis se centra en la aplicación del modelo de Fine & Gray para el estudio de riesgos en competencia en un contexto educativo, específicamente para identificar las variables que más influyen en diferentes eventos de interés en presencia de eventos competidores. Se utilizó un conjunto de datos de estudiantes de la Ingeniería en Sistemas Computacionales del ITESO para analizar cómo distintos factores impactan en estos eventos.

El enfoque metodológico incluyó una selección de variables mediante regularización en el modelo de Fine & Gray, para identificar las más relevantes en la ocurrencia de eventos en competencia. Estos resultados fueron comparados con un modelo desarrollado con todas las variables y otro con variables seleccionadas mediante la regularización LASSO, evaluando la precisión de cada uno. Finalmente, se realizó una comparación de ambos modelos, con un modelo nulo basado en estimaciones de Kaplan-Meier y Aalen-Johansen.

Este trabajo se nutrió los trabajos del Mtro. Erick Román Ramos Rocha y de la candidata a Dra. Gisel Hernández Chávez, contribuyendo con la transformación de los datos, el análisis exploratorio y la comparación de modelos(Hernández-Chávez, 2024). A continuación, se detallan los pasos específicos seguidos en el estudio.

4.1 Obtención y Preparación de los Datos

Los datos fueron extraídos de registros académicos de la universidad ITESO por la Doctora Gisel Hernández Chávez, quien realizó una ingeniería de características para asegurar que las variables estuvieran preparadas para el análisis. Estos datos fueron guardados en un archivo CSV titulado *"AlumISC_afterKM_2.csv",* el cual posteriormente fue cargado en el entorno de análisis. Los pasos realizados para la preparación de los datos incluyeron:

- Carga de Datos: Los datos fueron cargados utilizando las funciones de R, con el archivo en formato CSV.
- Exploración Inicial: Se realizó una visualización preliminar de los registros confirmando que no existieran datos nulos o faltantes, así como sus dimensiones. En donde se identificaron 396 registros con 61 características.
- Identificación de eventos: Se creó una nueva variable ECR para identificar los diferentes eventos en el análisis. Esta variable categoriza a los estudiantes en función de su estado al finalizar el periodo de estudio: activos, dados de baja, cambio de plan de estudios, o graduados. La codificación de estos eventos es la siguiente:
    - 0: Censurado (estudiantes activos)
    - 1: Evento de interés (baja académica)
    - 2: Evento competidor (cambio de plan de estudios)
    - 3: Graduación

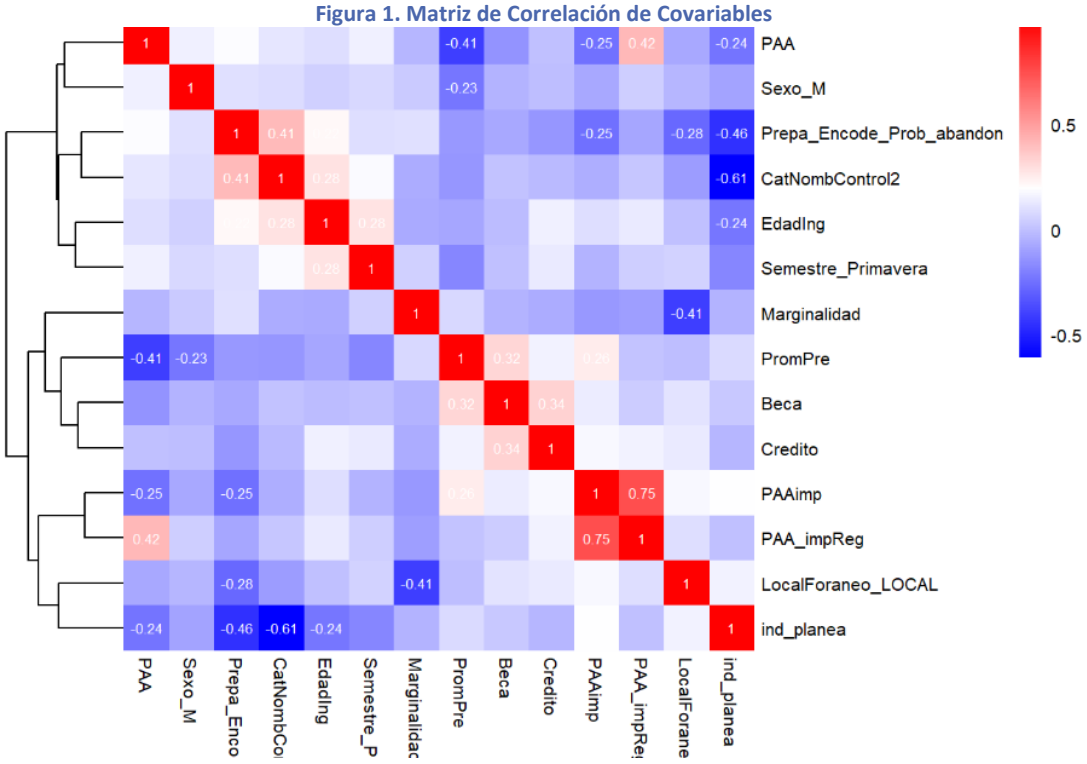
4.2 Análisis Exploratorio de Datos (EDA)

El análisis exploratorio se centró en identificar los tipos de datos asociados a cada variable, ya fueran numéricos o categóricos, y en analizar las correlaciones entre esta con el fin de preparar los datos para realizar las transformaciones necesarias antes de la aplicación del modelo. A partir de esto, se creó un nuevo conjunto de datos con el objetivo de remover las variables con alta correlación, así como las variables que representaban los eventos a analizar, como 'ECR', 'Estatus_Baja', 'Estatus_Cambio.de.plan', y 'Estatus_Egresado', así como la columna que indicaba el tiempo del evento, 'T'. Como resultado, se obtuvo un nuevo conjunto de covariables con una dimensión de 396 registros y 14 características.

El conjunto de datos presenta un rango de edades de ingreso (*EdadIng*) entre 17.28 y 34.65 años, con una mediana de 18.58. Los puntajes del examen PAA varían de 0 a 1577, con una mediana de 1208.5, y el Promedio de Preparatoria (*PromPre*) oscila entre 65 y 100, con una mediana de 84. El valor de Beca tiene un rango de 0 a 100, con una mediana de 20, mientras que el Crédito va de 0 a 45, con una mediana de 0. Las variables *PAAimp* y *PAA_impReg*, relacionadas con la imputación del puntaje *PAA*, tienen valores entre 857 y 1577. El 87.12% de los estudiantes son hombres (*Sexo_M*) y el 23.99% ingresaron en el semestre de primavera. Un 74.75% de los estudiantes son locales (*LocalForaneo_LOCAL*).

Las variables creadas por Hernández incluyen *CatNombControl2*, que refleja si la escuela es privada o pública (1) o extranjera o abierta (0); está mayoritariamente en 0. *Prepa_Encode_Prob_abandon* indica la probabilidad de abandono de acuerdo con la preparatoria de procedencia del alumno, con una media de 0.2485, *ind_planea* - índice calculado por Hernández relacionado con prueba del Plan Nacional para la Evaluación de los Aprendizajes - con valores entre 0 y 194.6, y una media de 123.8, y *Marginalidad*, con valores de 0 a 3 (Hernández-Chávez, 2024).

Figura 1. Matriz de Correlación de Covariables



La **Figura 1** presenta las correlaciones entre las variables del conjunto de datos. Las correlaciones positivas más fuertes se observan entre *PAAimp* y *PAA_impReg* (0.75), ya que ambas variables provienen de un proceso de imputación y regularización realizado por Hernández a partir de la variable original *PAA*. De acuerdo con el análisis de Hernández, *PAAimp* tiene un mayor impacto en los modelos predictivos en comparación con *PAA_impReg* y *PAA*, por lo que estas dos últimas serán eliminadas del modelo. Además, se observa una correlación positiva entre *PAA* y *PromPre* (0.41), lo que sugiere que un mayor puntaje en el *PAA* está relacionado con un mejor promedio de preparatoria.

Por otro lado, las correlaciones negativas más relevantes se observan entre *CatNombControl2* y *Marginalidad* (-0.61), lo que indica que un mayor nivel de marginalidad está relacionado con un menor valor de *CatNombControl2*. También, la correlación entre *Sexo_M* y *PAA* (-0.41) muestra que los hombres tienden a tener menores puntajes en el PAA. Finalmente, algunas correlaciones son más débiles, como entre *Beca* y *Crédito*, sugiriendo relaciones menos lineales entre esas variables.

4.3 Escalamiento de características

Finalizado el análisis exploratorio se realizó el escalamiento de características utilizando la técnica Min-Max para normalizar los valores de las variables a un rango entre 0 y 1. Esto fue necesario porque las variables del conjunto de datos tienen diferentes escalas, como se observa en sus estadísticas descriptivas. Por ejemplo, *EdadIng* varía entre 17.28 y 34.65 años, mientras que *PAA* tiene un rango de 0 a 1577 y *PromPre* va de 65 a 100. Estas diferencias podrían influir en el modelo, ya que las variables con mayores rangos podrían dominar el proceso de aprendizaje del modelo.

$$X_{scaled} = \frac{X_i - min(X)}{max(X) - min(X)}$$

Formula Escalamiento Min-Max

El escalamiento Min-Max ajusta los valores de cada variable proporcionalmente dentro de un rango definido, lo que garantiza que todas las características tengan el mismo peso en el modelo. Esto mejora la estabilidad numérica y permite que el modelo aprenda de manera más equilibrada. Variables como *Beca*, *Crédito*, *PAAimp* y *PAA_impReg*, que tienen rangos muy distintos, al ser normalizadas evita que alguna influya más que las demás en el proceso de aprendizaje del modelo.

4.4 Ajuste y selección de variables mediante regularización en el modelo Fine & Gray

Se ajustaron tres modelos de Fine & Gray, uno para cada evento de interés, utilizando datos escalados. Para modelar los riesgos competitivos, se creó un objeto Crisk basado en un vector de tiempos observados (T) y un vector de estatus (ECR), en el que el estatus se definió de la siguiente manera: 0 para censurados, 1 para el evento de interés y 2 para los eventos competidores. Con el objetivo de mantener la consistencia en el análisis, los eventos competidores con valores 2 y 3 se unificaron en un único valor de 2.

Para identificar las variables más influyentes en los eventos de interés, se aplicaron las regularizaciones LASSO, Ridge y Elastic Net en cada uno de los modelos, utilizando todos los datos disponibles. Además, se definió un rango de penalización utilizando el parámetro lambda, el cual fue ajustado para seleccionar el valor óptimo que minimizará el error de predicción y maximizará la eficacia de cada modelo. Se utilizó el mismo rango de penalización que en el trabajo de Hernández (Hernández-Chávez, 2024).

$$\lambda_i = 10^{log10(\lambda max) - n - 1i - 1(log10(\lambda max) - log10(\lambda min))}$$

4.5 Aplicación y Evaluación de Modelos de Fine & Gray con Brier Score

Una vez seleccionadas las variables mediante regularización, se entrenaron los modelos ajustados utilizando la función *FGR* del paquete *riskRegression*. Se construyeron dos modelos para cada evento: uno que incluyó todas las variables disponibles y otro con las variables seleccionadas por LASSO. Esta metodología permitió comparar el impacto de la selección de variables en el desempeño de los modelos.

Para evaluar la precisión de las predicciones probabilísticas, se utilizaron el Brier Score y el AUC como métricas principales. El Brier Score mide el error cuadrático medio entre las probabilidades predichas y los resultados observados, mientras que el AUC evalúa la capacidad de los modelos para distinguir entre eventos y no eventos. Estas métricas permiten una evaluación del desempeño predictivo de los modelos a lo largo del tiempo.

Finalmente, la representación gráfica del Brier Score y el análisis del AUC permitieron evaluar la evolución del desempeño predictivo en diferentes momentos de tiempo, mostrando que los modelos seleccionados por LASSO lograron una precisión similar a los modelos completos, con la ventaja de ser más simples y parsimoniosos.

5. Resultados

5.1 Identificación de variables más influyentes

En esta sección, se analiza el impacto de cada una de las variables en los eventos de Baja Académica, Cambio de Plan de Estudios y Egreso, mediante los modelos de regularización LASSO, Ridge y Elastic Net.

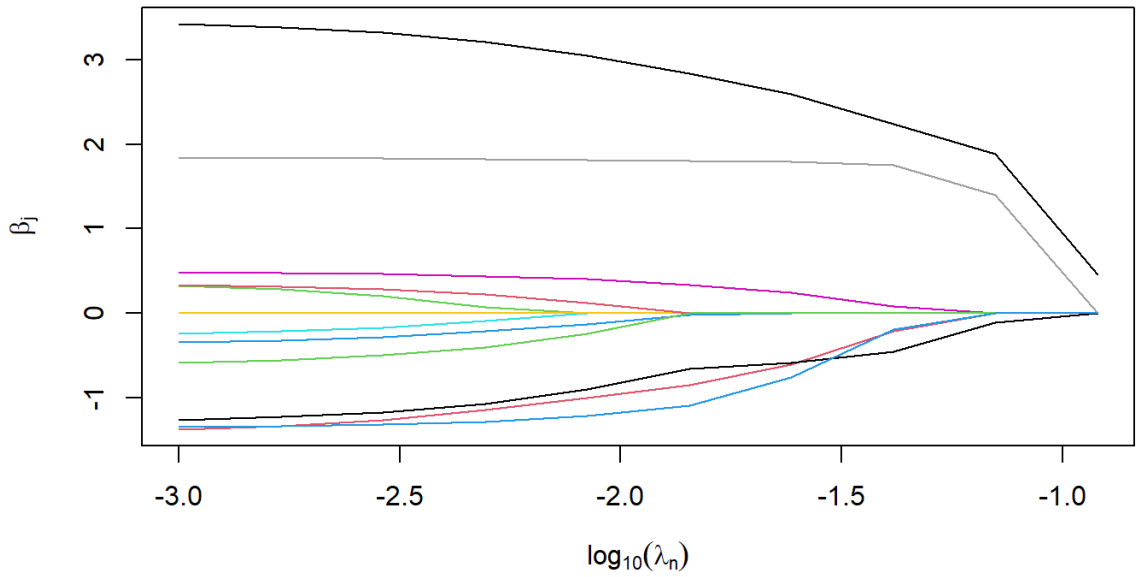
5.1.1 Evento de interés: baja académica

El primer análisis se centró en el evento de interés "Baja Académica", evaluado mediante tres métodos de regularización: LASSO, Ridge y Elastic Net (ENET). Cada modelo aplica diferentes penalizaciones para identificar la influencia de las variables predictoras en la probabilidad de abandono académico de los estudiantes. A continuación, se detallan los resultados de cada modelo y se identifican los predictores más relevantes según cada enfoque de regularización.

5.1.1.1 Regularización Lasso

El modelo LASSO penaliza ciertos coeficientes hasta llevarlos a cero, eliminando variables menos relevantes y facilitando la selección de predictores claves. El análisis de los coeficientes asociados al evento de baja académica revela cómo cada variable influye en la probabilidad de abandono, permitiendo identificar los factores más significativos, como se muestra en la **Figura 2**.

Figura 2. Evolución de Coeficientes en el Modelo para Baja Académica - LASSO



- *EdadIng*: Muestra un impacto positivo significativo, con el coeficiente creciendo desde 0 hasta 3.43 a medida que disminuye lambda. Esto sugiere que los estudiantes mayores al ingresar tienen más probabilidades de abandonar sus estudios.

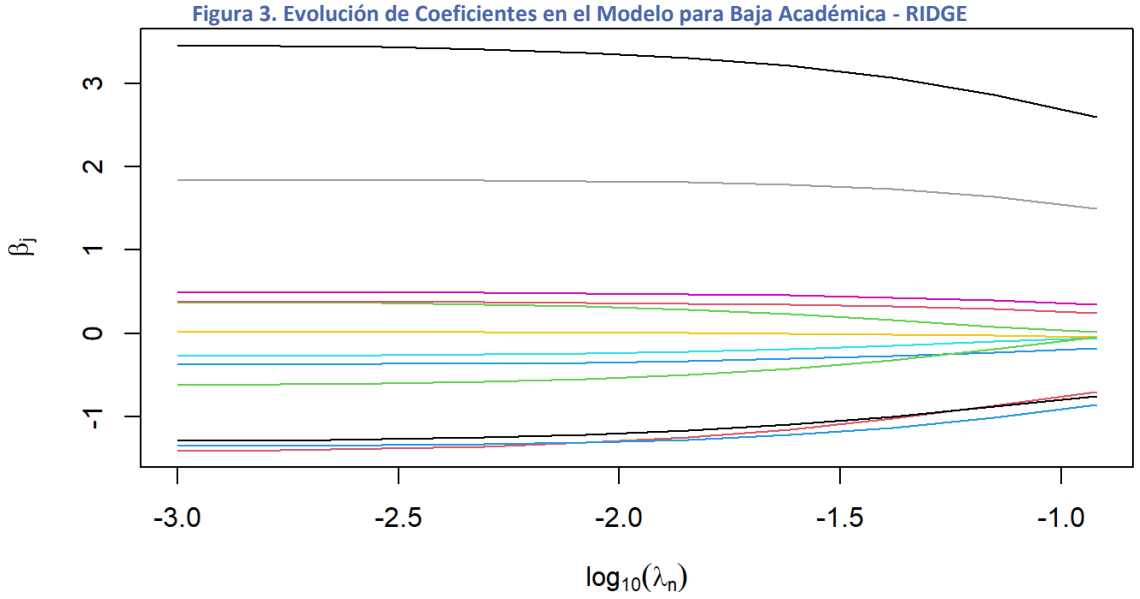
- *PromPre:* Inicialmente sin efecto, el coeficiente comienza a ser negativo y alcanza -1.37 a medida que disminuye lambda. Esto indica que los estudiantes con mejor promedio en preparatoria tienen menor riesgo de abandonar.
- *Beca*: Comienza en cero y muestra un leve impacto positivo a medida que lambda disminuye, alcanzando 0.33. Esto podría sugerir que los estudiantes con beca tienen una ligera mayor probabilidad de abandono, aunque el efecto es menor.
- *Credito:* Inicialmente cero, el coeficiente se vuelve negativo, llegando a -0.35. Esto sugiere que los estudiantes que tienen crédito educativo tienen menor probabilidad de abandono.
- *Sexo_M:* Muestra un efecto negativo constante, alcanzando -0.24. Esto sugiere que los estudiantes de sexo masculino tienen una probabilidad ligeramente menor de abandono.
- *Semestre_Primavera*: Tiene un efecto positivo moderado, alcanzando un coeficiente de 0.49, lo que sugiere que los estudiantes que inician en el semestre de primavera tienen un riesgo ligeramente mayor de abandono.
- *LocalForaneo_LOCAL:* Esta variable tiene un coeficiente constante de 0 en todos los valores de lambda, indicando que no tiene un impacto significativo en la probabilidad de abandono y puede ser irrelevante para el modelo.
- *Prepa_Encode_Prob_abandon*: Tiene un impacto positivo considerable, alcanzando un coeficiente de 1.83, lo cual sugiere que los estudiantes que vienen de preparatorias con alta probabilidad de abandono son más propensos a abandonar también.
- *ind_planea:* Su coeficiente es negativo, llegando a -1.26, lo que indica que los estudiantes con mayor valor en esta variable tienen una menor probabilidad de abandonar.
- *Marginalidad*: Comienza sin efecto, pero muestra un impacto positivo leve con un coeficiente de 0.34 en los valores más bajos de lambda, lo cual sugiere que los estudiantes de áreas con alta marginalidad pueden tener una probabilidad ligeramente mayor de abandono.
- *CatNombControl2*: Tiene un impacto negativo, alcanzando -0.59, lo cual sugiere que ciertos tipos de preparatorias están asociados con una menor probabilidad de abandono.
- *PAAimp:* Tiene un coeficiente negativo significativo, alcanzando -1.35 a medida que disminuye lambda. Esto indica que un mejor desempeño en el examen PAA reduce significativamente la probabilidad de abandono.

En conjunto, el modelo LASSO ha indica que las variables clave asociadas con el riesgo de abandono, son *EdadIng*, *PromPre*, *Prepa_Encode_Prob_abandon*, *Sexo_M*, *ind_planea* y *PAAimp* como factores significativos. Los estudiantes mayores al ingresar o con antecedentes de preparatorias con alta deserción tienen mayor riesgo de abandonar, mientras que aquellos con mejor promedio y desempeño en el examen PAA tienen menor riesgo.

5.1.1.2 Regularización Ridge

El modelo Ridge se utiliza para evaluar la importancia de las variables predictoras en el riesgo de abandono escolar, aplicando una penalización que reduce la varianza de los coeficientes sin eliminarlos. A diferencia de LASSO, el modelo Ridge mantiene todas las variables en el modelo, permitiendo una interpretación más completa de cada predictor en términos de su contribución al riesgo de abandono. La **Figura 3** muestra la evolución de los coeficientes a medida que disminuye lambda, ayudando a visualizar la influencia de cada variable.

Figura 3. Evolución de Coeficientes en el Modelo para Baja Académica - RIDGE



- *EdadIng*: El coeficiente es consistentemente positivo y aumenta ligeramente, alcanzando 3.47 a medida que disminuye lambda. Esto sugiere que los estudiantes mayores al ingresar tienen más probabilidades de abandonar sus estudios, siendo esta una de las variables más influyentes en el modelo.
- *PromPre*: Tiene un coeficiente negativo que también se amplifica al disminuir lambda, alcanzando -1.41. Esto indica que los estudiantes con mejor promedio de preparatoria tienen menor riesgo de abandono, lo cual es un factor protector en el modelo.
- *Beca*: El coeficiente es positivo y crece ligeramente, alcanzando 0.38 a medida que disminuye lambda. Esto sugiere que los estudiantes con beca podrían tener una ligera tendencia a abandonar, aunque el efecto es menor.
- *Credito*: Muestra un efecto negativo moderado, alcanzando -0.38. Esto indica que los estudiantes con crédito educativo tienen una menor probabilidad de abandono.
- *Sexo_M:* Tiene un coeficiente negativo constante que alcanza -0.27, sugiriendo que los estudiantes de sexo masculino tienen una probabilidad ligeramente menor de abandonar sus estudios.
- *Semestre_Primavera*: El coeficiente es positivo y alcanza 0.49, lo que sugiere que los estudiantes que inician en el semestre de primavera tienen un riesgo ligeramente mayor de abandono.
- *LocalForaneo_LOCAL:* Su coeficiente es cercano a cero en todos los valores de lambda, lo cual indica que esta variable no tiene un impacto significativo en el riesgo de abandono.

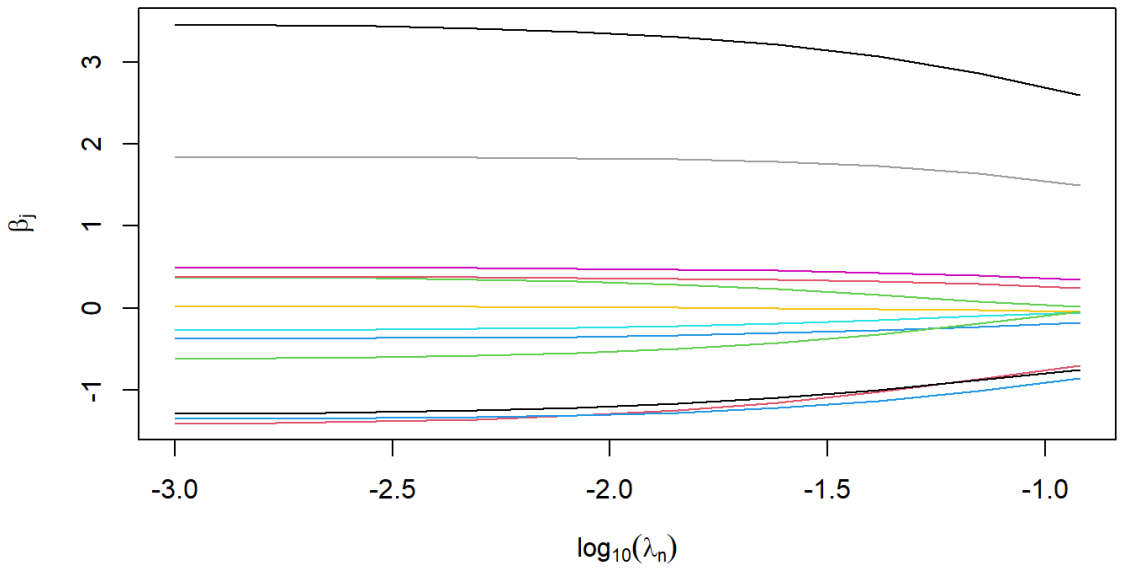
- *Prepa_Encode_Prob_abandon:* Tiene un coeficiente positivo y consistente, alcanzando 1.85, lo que sugiere que los estudiantes provenientes de preparatorias con alta probabilidad de abandono son más propensos a abandonar.
- *ind_planea:* Muestra un coeficiente negativo, alcanzando -1.30, lo que indica que los estudiantes con mayor valor en esta variable tienen una menor probabilidad de abandono.
- *Marginalidad:* Tiene un coeficiente positivo moderado, alcanzando 0.38, lo cual sugiere que los estudiantes de áreas con alta marginalidad pueden tener una probabilidad ligeramente mayor de abandono.
- *CatNombControl2:* El coeficiente es negativo, alcanzando -0.62, lo que sugiere que ciertos tipos de preparatoria están asociados con una menor probabilidad de abandono.
- *PAAimp:* Presenta un coeficiente negativo significativo, alcanzando -1.36, lo cual indica que un mejor desempeño en el examen PAA reduce significativamente la probabilidad de abandono.

De acuerdo con los resultados, variables *EdadIng*, *PromPre*, *Prepa_Encode_Prob_abandon*, *ind_planea*, y *PAAimp* son las más determinantes en la probabilidad de abandono, lo cual se observa en la **Figura 3**, donde sus coeficientes muestran una influencia significativa y consistente a medida que disminuye lambda.

5.1.1.3 Regularización Elastic Net

El modelo con regularización Elastic Net (ENET) combina las penalizaciones LASSO y Ridge, permitiendo una selección de variables y reduciendo la varianza en los coeficientes. Esto resulta útil en situaciones donde algunas variables deben ser eliminadas mientras otras se mantienen, logrando un equilibrio entre simplicidad y robustez. En la **Figura 4** se observa cómo evolucionan los coeficientes de cada predictor a medida que disminuye lambda, proporcionando una referencia visual de sus influencias en el riesgo de abandono.

Figura 4. Evolución de Coeficientes en el Modelo para Baja Académica - RIDGE



- *EdadIng*: El coeficiente es consistentemente positivo y aumenta ligeramente, alcanzando 3.47 a medida que disminuye lambda. Esto sugiere que los estudiantes mayores al ingresar tienen más probabilidades de abandonar, siendo esta una de las variables más influyentes en el modelo.
- *PromPre*: Presenta un coeficiente negativo que se amplifica al disminuir lambda, alcanzando -1.41. Esto indica que los estudiantes con mejor promedio de preparatoria tienen menor riesgo de abandono, destacándose como un factor protector.
- *Beca:* Tiene un coeficiente positivo que crece moderadamente, alcanzando 0.38. Esto sugiere una ligera mayor probabilidad de abandono para estudiantes con beca, aunque el efecto es menos pronunciado.
- *Credito:* Su coeficiente es negativo, alcanzando -0.38, indicando que los estudiantes con crédito educativo presentan menor riesgo de abandono.
- *Sexo_M:* Muestra un efecto negativo constante, alcanzando -0.27, sugiriendo que los estudiantes de sexo masculino tienen una probabilidad ligeramente menor de abandonar.
- *Semestre_Primavera:* Presenta un coeficiente positivo moderado de 0.49, lo que indica que los estudiantes que inician en el semestre de primavera tienen un riesgo ligeramente mayor de abandono.
- *LocalForaneo_LOCAL:* Su coeficiente es cercano a cero en todos los valores de lambda, lo cual indica que esta variable tiene poco impacto significativo en el riesgo de abandono.
- *Prepa_Encode_Prob_abandon:* Con un coeficiente positivo que llega a 1.85, esta variable sugiere que los estudiantes provenientes de preparatorias con alta probabilidad de abandono son más propensos a abandonar.
- *ind_planea:* Muestra un coeficiente negativo de -1.30, indicando que los estudiantes con mayor valor en esta variable tienen una menor probabilidad de abandonar.
- *Marginalidad:* Con un coeficiente positivo moderado de 0.38, sugiere que los estudiantes de áreas con alta marginalidad tienen una probabilidad ligeramente mayor de abandono.
- *CatNombControl2:* Este coeficiente es negativo y alcanza -0.62, lo que sugiere que ciertos tipos de preparatorias están asociados con una menor probabilidad de abandono.
- *PAAimp:* Presenta un coeficiente negativo significativo de -1.36, indicando que un mejor desempeño en el examen PAA reduce considerablemente el riesgo de abandono.

Este modelo resalta *EdadIng*, *PromPre*, *Prepa_Encode_Prob_abandon*, *ind_planea*, y *PAAimp* como los predictores claves, equilibrando la penalización y reteniendo la mayoría de las variables con influencia relevante.

5.1.1.4 Comparación de modelos

De acuerdo con los tres modelos, las variables que consistentemente demuestran una fuerte influencia en el riesgo de abandono son EdadIng, PromPre, Prepa_Encode_Prob_abandon, ind_planea, y PAAimp. Estas variables destacan como las mejores predictores, ya que sus coeficientes son significativos en cada modelo y mantienen su impacto incluso bajo diferentes penalizaciones.

- *EdadIng*: Asociada a un mayor riesgo de abandono para estudiantes de mayor edad al ingresar.
- *PromPre*: Protector contra el abandono, ya que un mayor promedio previo se asocia a menor riesgo.

- *Prepa_Encode_Prob_abandon*: Indica que estudiantes de preparatorias con alto riesgo de abandono también son propensos a abandonar.
- *ind_planea*: Un mayor valor en esta variable se asocia con una menor probabilidad de abandono.
- *PAAimp*: Su desempeño en el examen PAA reduce el riesgo de abandono significativamente.

Estos resultados sugieren que, independientemente de la metodología de regularización utilizada, estas variables deberían ser el enfoque principal de cualquier estrategia de intervención destinada a reducir el abandono escolar.

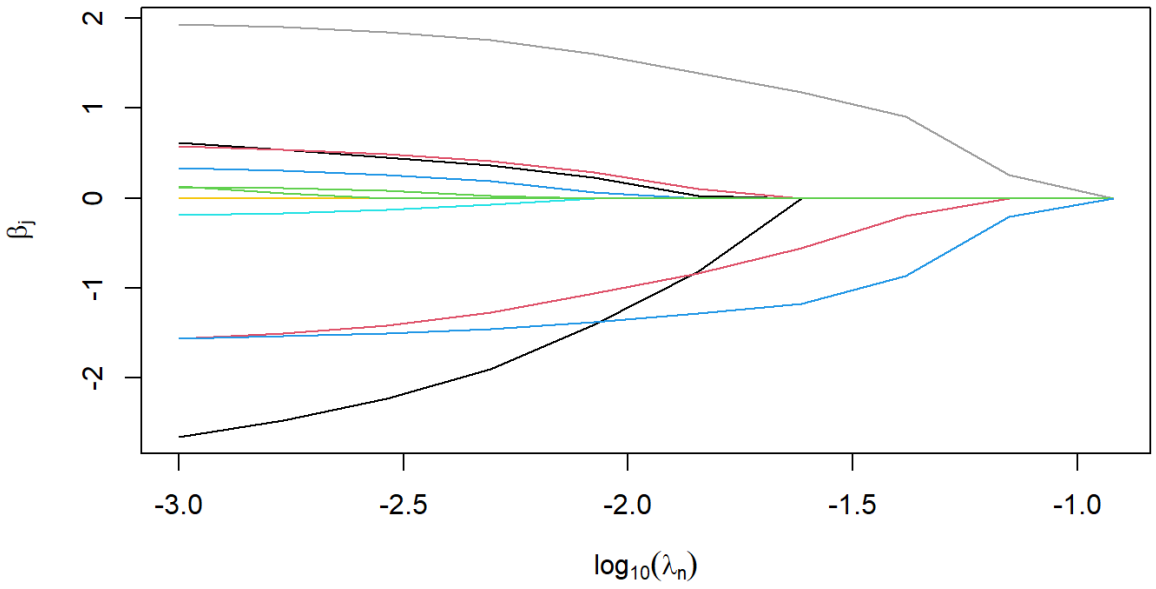
5.1.2 Evento de interés: Cambio de plan de estudios

Continuando con el análisis, el siguiente evento de interés a analizar fue "Cambio de Plan de Estudios", el cual, al igual que el anterior, fue evaluado mediante tres métodos de regularización: LASSO, Ridge y Elastic Net (ENET). A continuación, se detallan los resultados de cada modelo y se identifican los predictores más relevantes según cada enfoque de regularización.

5.1.2.1 Regularización Lasso

Iniciando con la regularización tipo LASSO, la Figura 5 muestra la evolución de los coeficientes en el modelo para el evento de "Cambio de Plan de Estudios", destacando cómo los valores de cada predictor se ajustan conforme se aplica la penalización.

Figura 5. Evolución de Coeficientes en el Modelo para Cambio de plan de estudios – Lasso



- *EdadIng*: Coeficiente negativo al disminuir lambda, alcanzando -2.66. Los estudiantes mayores tienen menos probabilidad de cambiar de plan.
- *PromPre*: Coeficiente negativo, alcanzando -1.56, lo cual indica que un mejor promedio previo reduce la probabilidad de cambio.

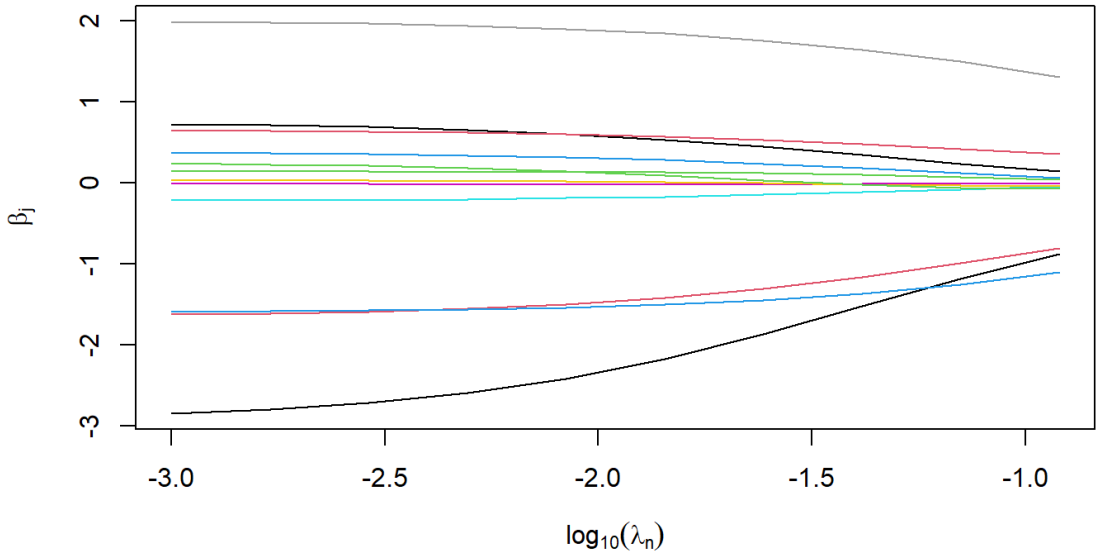
- *Beca*: El coeficiente se vuelve negativo a valores bajos de lambda (-0.19), sugiriendo que los estudiantes con beca tienen una ligera menor probabilidad de cambiar de plan.
- *Credito*: Coeficiente positivo (0.33), lo que podría indicar una relación leve con la probabilidad de cambio de plan.
- *Sexo_M*: Esta variable tiene coeficiente cercano a cero, sugiriendo poca influencia en la decisión de cambio.
- *Semestre_Primavera*: Coeficiente muy bajo (cercano a cero), lo que indica un impacto mínimo en el cambio de plan.
- *LocalForaneo_LOCAL*: Coeficiente cercano a cero, eliminando esta variable como predictor importante.
- *Prepa_Encode_Prob_abandon*: Coeficiente positivo significativo (1.93), indicando que estudiantes de preparatorias con alta tasa de abandono son más propensos a cambiar de plan.
- *ind_planea*: Efecto positivo (0.62), lo cual sugiere que un menor valor en esta variable se asocia con una mayor disposición al cambio de plan.
- *Marginalidad*: Coeficiente positivo (0.57), sugiriendo que estudiantes de áreas de alta marginalidad podrían tener mayor probabilidad de cambiar de plan.
- *CatNombControl2*: Coeficiente positivo bajo (0.13), indicando un impacto limitado en el cambio de plan.
- *PAAimp*: Coeficiente negativo (-1.56), lo cual sugiere que un buen desempeño en el examen PAA reduce la probabilidad de cambio.

En el este modelo, las variables *EdadIng*, *PromPre*, *Prepa_Encode_Prob_abandon*, *ind_planea*, *Marginalidad*, y *PAAimp* destacan como factores determinantes en el cambio de plan, mientras que las demás variables tienen menor relevancia.

5.1.2.2 Regularización Ridge

Por otro lado, la regularización Ridge, visible en la Figura 6, mostró los siguientes resultados, permitiendo observar cómo los coeficientes de cada predictor se mantienen en valores moderados sin reducirse a cero, lo que proporciona una visión más completa del impacto de cada variable en el "Cambio de Plan de Estudios".

Figura 6. Evolución de Coeficientes en el Modelo para Cambio de plan de estudios – Ridge



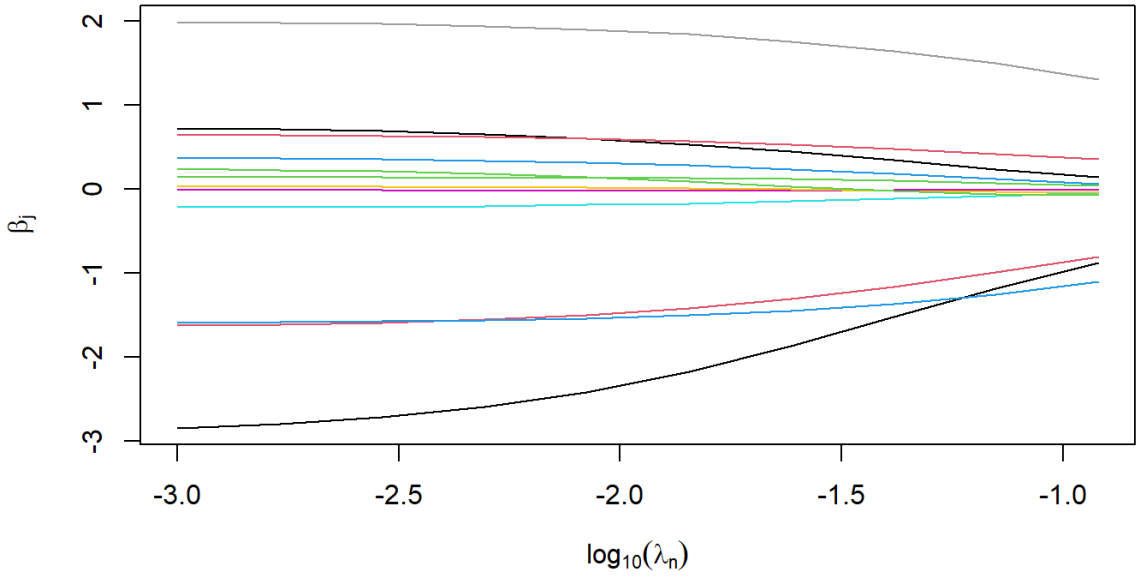
- *EdadIng*: Coeficiente negativo constante, alcanzando -2.85, indicando menor probabilidad de cambio para estudiantes de mayor edad.
- *PromPre*: Coeficiente negativo estable (-1.62), sugiriendo que un mejor promedio reduce la probabilidad de cambio.
- *Beca*: Coeficiente negativo (-0.22), indicando que los estudiantes con beca tienen menor probabilidad de cambiar de plan.
- *Credito*: Coeficiente positivo moderado (0.37), sugiriendo que estudiantes con crédito pueden estar más dispuestos al cambio.
- *Sexo_M*: Coeficiente positivo bajo (0.03), lo que indica un impacto mínimo en el cambio de plan.
- *Semestre_Primavera*: Coeficiente bajo (0.01), sugiriendo que el semestre de inicio tiene un efecto mínimo en la probabilidad de cambio.
- *LocalForaneo_LOCAL*: Coeficiente cercano a cero, indicando baja relevancia en el cambio de plan.
- *Prepa_Encode_Prob_abandon*: Coeficiente positivo y significativo (1.99), indicando mayor probabilidad de cambio para estudiantes provenientes de preparatorias con alto abandono.
- *ind_planea*: Coeficiente positivo (0.72) alumnos con un menor valor en esta variable pueden estar más dispuestos a cambiar de plan.
- *Marginalidad:* Coeficiente positivo moderado (0.72), indicando que los estudiantes de áreas de alta marginalidad tienen una mayor disposición al cambio de plan.
- *CatNombControl2*: Coeficiente positivo bajo (0.23), sugiriendo un efecto mínimo.
- *PAAimp*: Coeficiente negativo de -1.59, indicando que un buen desempeño en el examen PAA reduce la probabilidad de cambio.

Ridge confirma la relevancia de *EdadIng*, *PromPre*, *Prepa_Encode_Prob_abandon*, *ind_planea*, *Marginalidad*, y *PAAimp* en el modelo, manteniendo una influencia moderada de todas las variables.

5.1.2.3 Regularización Elastic Net

Finalmente, la regularización Elastic Net, presentada en la Figura 7, combina las ventajas de LASSO y Ridge, equilibrando la selección de variables con la retención de coeficientes moderados. Esta técnica arroja los siguientes resultados, resaltando cómo cada predictor influye en el "Cambio de Plan de Estudios" al aplicar una penalización combinada.

Figura 7. Evolución de Coeficientes en el Modelo para Cambio de plan de estudios – Elastic Net



- *EdadIng*: Coeficiente negativo (-2.85), indicando que estudiantes mayores son menos propensos al cambio de plan.
- *PromPre*: Coeficiente negativo (-1.62), indicando que un mejor promedio previo reduce la probabilidad de cambio.
- *Beca*: Coeficiente negativo bajo (-0.22), sugiriendo que los estudiantes con beca son menos propensos al cambio.
- *Credito*: Coeficiente positivo moderado (0.37), lo que podría sugerir que los estudiantes con crédito educativo son más propensos a cambiar de plan.
- *Sexo_M*: Coeficiente bajo (0.03), indicando un efecto mínimo en el cambio de plan.
- *Semestre_Primavera*: Coeficiente cercano a cero, indicando que el semestre de inicio tiene un impacto mínimo en la probabilidad de cambio.
- *LocalForaneo_LOCAL*: Coeficiente muy bajo, sugiriendo poca relevancia.
- *Prepa_Encode_Prob_abandon*: Coeficiente positivo significativo (1.99), indicando que los estudiantes provenientes de preparatorias con alto abandono son más propensos al cambio.
- *ind_planea*: Coeficiente positivo (0.72), lo que sugiere que los estudiantes con menor valor en el índice son más propensos al cambio.
- *Marginalidad*: Coeficiente positivo moderado (0.72), sugiriendo que estudiantes de áreas de alta marginalidad podrían tener mayor probabilidad de cambio.

- *CatNombControl2*: Coeficiente positivo bajo (0.23), indicando un efecto mínimo en el cambio de plan.
- *PAAimp*: Coeficiente negativo (-1.59), indicando que un buen desempeño en el examen PAA reduce la probabilidad de cambio.

Elastic Net resalta *EdadIng*, *PromPre*, *Prepa_Encode_Prob_abandon*, *ind_planea*, *Marginalidad*, y *PAAimp* como las variables más influyentes, confirmando su importancia al equilibrar la penalización y seleccionando variables clave mientras retiene la mayoría.

5.1.2.4 Comparación de modelos

Los tres modelos coinciden en identificar las variables *EdadIng*, *PromPre*, *Prepa_Encode_Prob_abandon*, *ind_planea*, *Marginalidad*, y *PAAimp* como las más relevantes en la probabilidad de cambio de plan de estudios. Estas variables destacan por su influencia consistente en los modelos, sugiriendo que:

- *EdadIng* y *PromPre* actúan como factores predictores, disminuyendo la probabilidad de cambio de plan.
- *Prepa_Encode_Prob_abandon* e *ind_planea* aumentan la probabilidad de cambio, especialmente en estudiantes de preparatorias con alto abandono y aquellos con menor valor en el índice PLANEA.
- *Marginalidad* muestra un impacto positivo, indicando que estudiantes de áreas con mayor marginalidad tienen una mayor disposición al cambio.
- *PAAimp* tiene un efecto negativo, donde un buen desempeño en el examen reduce la probabilidad de cambio.

Esta consistencia sugiere que estas variables pueden servir como base para diseñar estrategias de retención académica y orientación en función de los factores más significativos que influyen en el cambio de plan de estudios.

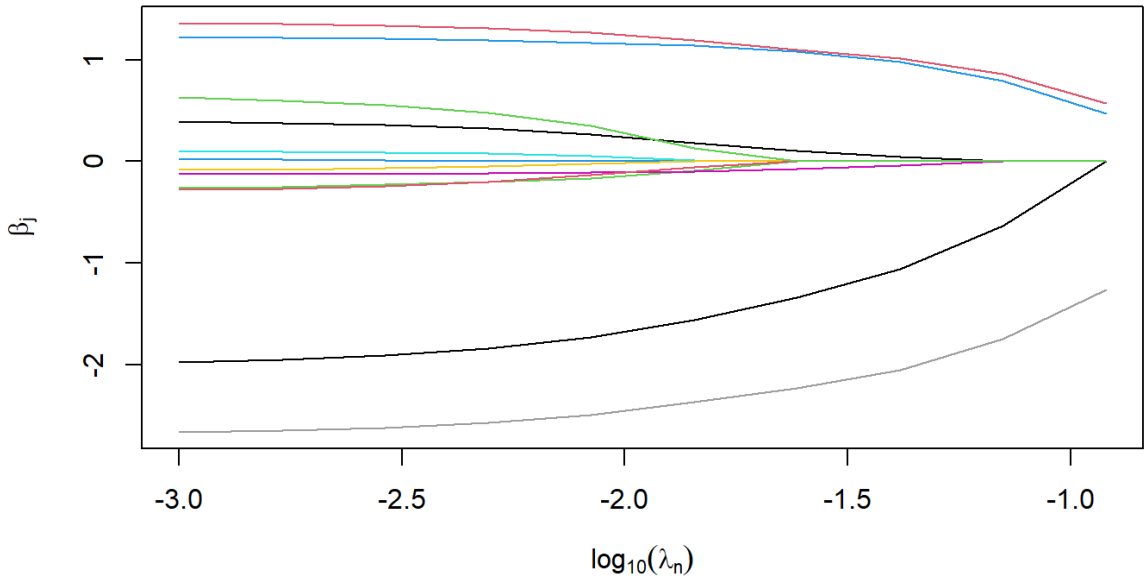
5.1.3 Evento de interés: Egresado o Graduación

Para concluir el análisis de este trabajo, se evaluó el evento de interés "Egreso o Graduación" mediante tres enfoques de regularización: LASSO, Ridge y Elastic Net. Cada método permitió examinar la influencia de diversas variables en la probabilidad de que un estudiante complete sus estudios, facilitando la identificación de los factores más determinantes en el logro académico.

5.1.3.1 Regularización Lasso

La **Figura 8** muestra los resultados de la regularización LASSO, donde se visualiza cómo los coeficientes de cada predictor se ajustan a medida que se aplica la penalización. Esta representación permite identificar las variables más relevantes para el modelo, destacando aquellos predictores que mantienen un impacto significativo en el evento de "Egreso o Graduación" mientras que otros coeficientes se reducen a cero, eliminando variables menos influyentes.

Figura 8. Evolución de Coeficientes en el Modelo para Egreso – Lasso



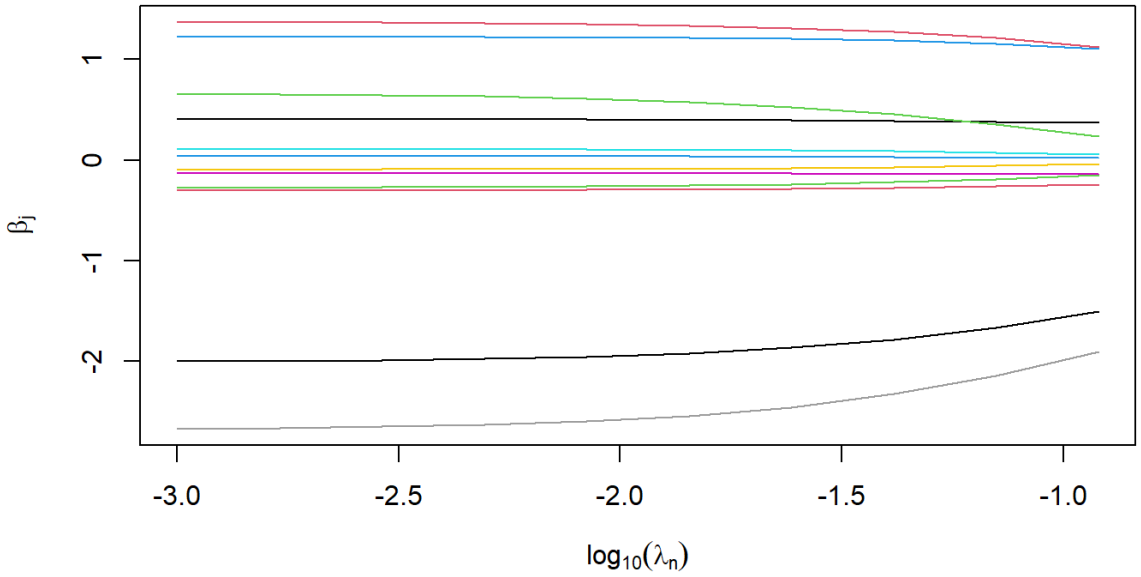
- *EdadIng*: Coeficiente negativo, alcanzando -1.97, lo que sugiere que los estudiantes mayores tienen menor probabilidad de egresar.
- *PromPre*: Coeficiente positivo (1.36), indicando que un mejor promedio previo incrementa la probabilidad de egreso.
- *Beca*: Coeficiente negativo (-0.26), lo cual sugiere que los estudiantes con beca tienen menor probabilidad de egreso.
- *Credito*: Coeficiente positivo bajo (0.03), indicando un impacto limitado.
- *Sexo_M*: Coeficiente positivo (0.10), sugiriendo una ligera mayor probabilidad de egreso para estudiantes de sexo masculino.
- *Semestre_Primavera:* Coeficiente negativo leve (-0.13), sugiriendo que los estudiantes que inician en primavera tienen una ligera menor probabilidad de egreso.
- *LocalForaneo_LOCAL*: Coeficiente negativo leve (-0.09), sugiriendo que los estudiantes locales tienen una leve menor probabilidad de egresar.
- *Prepa_Encode_Prob_abandon:* Coeficiente negativo (-2.66), indicando que los estudiantes de preparatorias con alta tasa de abandono tienen menor probabilidad de egreso.
- *ind_planea:* Coeficiente positivo (0.39), sugiriendo que estudiantes con un mayor valor en esta variable tienen una mayor probabilidad de egreso.
- *Marginalidad*: Coeficiente negativo bajo (-0.28), sugiriendo un impacto menor en el egreso.
- *CatNombControl2:* Coeficiente positivo (0.63), indicando una ligera mayor probabilidad de egreso asociada a ciertos tipos de preparatoria.
- *PAAimp:* Coeficiente positivo significativo (1.22), lo cual sugiere que un buen desempeño en el examen PAA incrementa la probabilidad de egreso.

La regularización Lasso destaca *EdadIng*, *PromPre*, *Prepa_Encode_Prob_abandon*, *ind_planea*, *CatNombControl2,* y *PAAimp* como variables clave en la probabilidad de egreso, mientras elimina o asigna bajo peso a las demás.

5.1.3.2 Regularización Ridge

Para el modelo Ridge, la **Figura 9** ilustra cómo los coeficientes de cada predictor se ajustan sin reducirse a cero, permitiendo que todas las variables mantengan algún nivel de influencia en el modelo. Esta representación ofrece una visión completa de cómo cada predictor contribuye a la probabilidad de "Egreso o Graduación".

Figura 9. Evolución de Coeficientes en el Modelo para Egreso – Ridge



- *EdadIng*: Coeficiente negativo constante, alcanzando -2.85, indicando menor probabilidad de egreso en estudiantes mayores.
- *PromPre*: Coeficiente positivo constante (1.62), sugiriendo que un mejor promedio previo incrementa la probabilidad de egreso.
- *Beca:* Coeficiente negativo leve (-0.22), indicando menor probabilidad de egreso en estudiantes con beca.
- *Credito*: Coeficiente positivo bajo (0.37), sugiriendo una leve asociación con la probabilidad de egreso.
- *Sexo_M*: Coeficiente positivo bajo (0.03), indicando un efecto mínimo en el egreso.
- *Semestre_Primavera*: Coeficiente negativo bajo (-0.13), sugiriendo que el semestre de inicio tiene un impacto menor en la probabilidad de egreso.
- *LocalForaneo_LOCAL*: Coeficiente negativo bajo (-0.09), indicando baja relevancia en la probabilidad de egreso.
- *Prepa_Encode_Prob_abandon*: Coeficiente negativo significativo (-2.66), indicando que los estudiantes de preparatorias con alta tasa de abandono tienen menor probabilidad de egreso.
- *ind_planea*: Coeficiente positivo (0.72), lo cual sugiere que los estudiantes con un mayor valor en esta variable son más propensos a egresar.

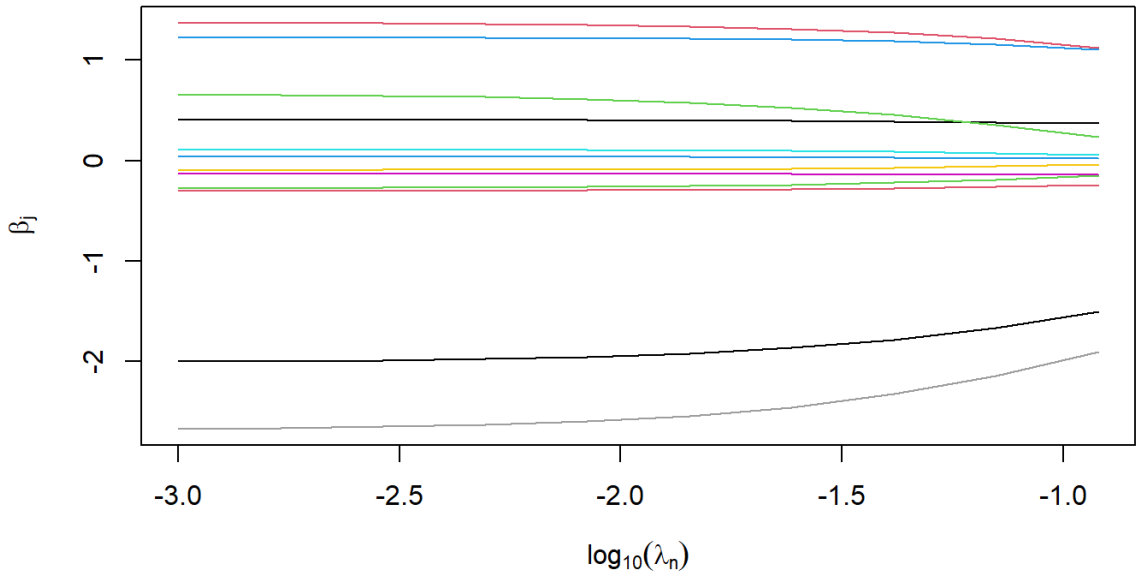
- *Marginalidad:* Coeficiente negativo bajo (-0.28), sugiriendo un efecto limitado en el egreso.
- *CatNombControl2:* Coeficiente positivo bajo (0.63), sugiriendo un efecto limitado.
- *PAAimp:* Coeficiente positivo significativo (1.22), indicando que un buen desempeño en el examen PAA incrementa la probabilidad de egreso.

Ridge confirma la importancia de *EdadIng*, *PromPre*, *Prepa_Encode_Prob_abandon*, *ind_planea*, *CatNombControl2*, y *PAAimp* en el egreso, manteniendo todas las variables en el modelo con influencia moderada.

5.1.3.3 Modelo Elastic Net

En el caso de Elastic Net, la **Figura 10** muestra los resultados combinando las características de LASSO y Ridge. En este modelo, algunos coeficientes se reducen a cero mientras otros se mantienen, logrando un balance entre la selección de variables y la retención de predictores claves. Esta gráfica permite identificar las variables más influyentes en el "Egreso o Graduación" y observar cómo se aplica la penalización combinada para optimizar la precisión del modelo.

Figura 10. Evolución de Coeficientes en el Modelo para Egreso – ENET



- *EdadIng*: Coeficiente negativo (-2.00), indicando que estudiantes mayores tienen menor probabilidad de egreso.
- *PromPre*: Coeficiente positivo (1.37), indicando que un buen promedio previo incrementa la probabilidad de egreso.
- *Beca*: Coeficiente negativo leve (-0.27), indicando que los estudiantes con beca tienen menor probabilidad de egresar.
- *Credito*: Coeficiente positivo bajo (0.03), lo cual indica un impacto limitado.
- *Sexo_M*: Coeficiente positivo (0.10), indicando que los estudiantes de sexo masculino tienen una ligera mayor probabilidad de egreso.

- *Semestre_Primavera*: Coeficiente negativo leve (-0.13), indicando que estudiantes que inician en primavera tienen una ligera menor probabilidad de egreso.
- *LocalForaneo_LOCAL*: Coeficiente negativo bajo (-0.09), sugiriendo poca relevancia.
- *Prepa_Encode_Prob_abandon:* Coeficiente negativo significativo (-2.66), indicando que estudiantes de preparatorias con alta tasa de abandono tienen menor probabilidad de egreso.
- *ind_planea:* Coeficiente positivo (0.72), lo cual sugiere que los estudiantes con mayor valor en esta variable son más propensos a egresar.
- *Marginalidad*: Coeficiente negativo bajo (-0.28), indicando un impacto menor.
- *CatNombControl2:* Coeficiente positivo bajo (0.63), indicando un efecto limitado en el egreso.
- *PAAimp:* Coeficiente positivo significativo (1.22), sugiriendo que un buen desempeño en el examen PAA incrementa la probabilidad de egreso.

Para evento de "Graduación o egreso", Elastic Net confirma la importancia de *EdadIng*, *PromPre*, *Prepa_Encode_Prob_abandon*, *ind_planea*, y *PAAimp* al mantener estas variables como las más influyentes.

5.1.3.4 Comparación de modelos

Los tres modelos coinciden en identificar las variables *EdadIng*, *PromPre*, *Prepa_Encode_Prob_abandon*, *ind_planea*, *CatNombControl2*, y *PAAimp* como las más relevantes para la probabilidad de egreso. Esto sugiere que:

- *EdadIng* actúa como un factor protector, reduciendo la probabilidad de egreso en estudiantes de mayor edad.
- *PromPre* incrementa la probabilidad de egreso en estudiantes con mejores promedios.
- *Prepa_Encode_Prob_abandon* disminuye la probabilidad de egreso para estudiantes provenientes de preparatorias con alta tasa de abandono.
- *ind_planea* muestra que un mayor valor en el índice favorece el egreso.
- *CatNombControl2* y *PAAimp* aumentan la probabilidad de egreso, especialmente en estudiantes con buen desempeño en el examen PAA.

Esta consistencia entre modelos sugiere que estas variables deberían ser el foco principal en estrategias de retención y apoyo para incrementar las tasas de egreso.

5.1.4 Resumen de resultados

Este análisis evaluó los factores asociados con tres eventos de interés: Baja Académica, Cambio de Plan de Estudios y Graduación. Para cada evento, se aplicaron tres enfoques de regularización – LASSO, Ridge y Elastic Net (ENET) – permitiendo observar las variables clave en cada caso y comprender su influencia en el rendimiento académico y las decisiones de los estudiantes.

Resultados de Variables por Evento y Regularización

- Baja Académica:
    - LASSO: Resalta las variables *EdadIng*, *PromPre*, *Prepa_Encode_Prob_abandon*, *ind_planea*, y *PAAimp*, reduciendo a cero otras variables de menor impacto.
    - Ridge: Mantiene todas las variables con influencias moderadas, destacando también EdadIng, *PromPre*, *Prepa_Encode_Prob_abandon*, *ind_planea,* y *PAAimp* como las más significativas.
    - Elastic Net: Confirma la relevancia de *EdadIng*, *PromPre, Prepa_Encode_Prob_abandon*, *ind_planea*, y *PAAimp*, logrando un balance entre la retención de variables y la eliminación de aquellas menos relevantes.
- Cambio de Plan de Estudios:
    - LASSO: Identifica como clave a *EdadIng*, *Prepa_Encode_Prob_abandon*, *ind_planea, Marginalidad*, y *PAAimp*, excluyendo otras variables de impacto menor.
    - Ridge: Conserva todas las variables, mostrando que *EdadIng*, *PromPre*, *Prepa_Encode_Prob_abandon*, *ind_planea*, y *Marginalidad* influyen más en la probabilidad de cambio de plan.
    - Elastic Net: Destaca *EdadIng*, *PromPre*, *Prepa_Encode_Prob_abandon*, *ind_planea*, y *Marginalidad* como los predictores principales, logrando un equilibrio entre simplificación y retención de variables.
- Graduación:
    - LASSO: Identifica como factores clave *EdadIng*, *PromPre*, *Prepa_Encode_Prob_abandon*, *ind_planea*, y *PAAimp* en la probabilidad de graduación, eliminando otros predictores menores.
    - Ridge: Mantiene todas las variables en el modelo, pero también destaca *EdadIng*, *PromPre*, *Prepa_Encode_Prob_abandon*, *ind_planea*, y *PAAimp* como las más influyentes.
    - Elastic Net: Refuerza la importancia de *EdadIng*, *PromPre*, *Prepa_Encode_Prob_abandon*, *ind_planea*, y *PAAimp* como variables clave, permitiendo la eliminación de variables menos significativas.

Comparación de Regularizaciones y Variables Más Influyentes

A lo largo de los tres eventos, los tres enfoques de regularización – LASSO, Ridge y Elastic Net – destacan consistentemente las siguientes variables:

- *EdadIng*: Se observa que la edad al ingreso tiene un impacto significativo en los tres eventos, con los estudiantes de mayor edad menos propensos a cambios en su trayectoria académica.
- *PromPre*: Un mejor promedio en preparatoria está asociado con una menor probabilidad de baja académica y mayor probabilidad de graduación.
- *Prepa_Encode_Prob_abandon* (tasa de abandono en la preparatoria de origen): Los estudiantes de preparatorias con alta probabilidad de abandono muestran mayor riesgo en todos los eventos.
- *ind_planea:* Este índice parece beneficiar a los estudiantes en los tres eventos, ya que un alto valor está asociado con menores probabilidades de baja y cambio de plan, y mayores probabilidades de graduación.

- PAAimp: Un buen desempeño en el examen PAA se relaciona con una mayor probabilidad de graduación y menor probabilidad de baja.

Estas cinco variables, identificadas de manera consistente en los tres eventos y en todos los métodos de regularización, son las más influyentes en las trayectorias académicas de los estudiantes.

5.2 Ajuste de modelos

En esta sección se analiza el impacto de cada una de las variables en los eventos de Baja Académica, Cambio de Plan de Estudios y Graduación. Primero, se realizaron las predicciones considerando todas las variables incluidas en el modelo para cada evento. Posteriormente, se llevaron a cabo las predicciones utilizando únicamente las variables seleccionadas mediante el método LASSO. A continuación, se presentan los resultados obtenidos.

5.2.1.1 Modelos bajas académicas.

En el primer caso, el evento de interés es el abandono o baja académica. El modelo se ajustó utilizando todas las variables predictoras disponibles. Para ello, los datos se dividieron en un 70% para el conjunto de entrenamiento, incluyendo un total de 277 casos.

El modelo presentó un Pseudo Log-likelihood de -192, lo que refleja la calidad del ajuste del modelo a los datos. Además, el Pseudo likelihood ratio test arrojó un valor de 50.5 con 12 grados de libertad, indicando que el modelo, en su conjunto, es significativo y explica de manera adecuada el evento de interés en presencia de riesgos competitivos. A continuación, la **Tabla 2** muestra los resultados de todas las variables:

Tabla 2 Resultados del Modelo Todas las Variables para la Predicción de Baja Académica

Variable	Coeficiente (β)	Exp(β)	Error estándar	Valor-Z	p-valor	IC 95% (Límite Inferior)	IC 95% (Límite Superior)
EdadIng	3.671	39.295	0.734	5.005	5.6E-07	9.332	165.46
PromPre	-0.358	0.699	0.632	-0.567	0.57	0.202	2.41
Beca	-0.327	0.721	0.61	-0.536	0.59	0.218	2.38
Credito	-0.48	0.619	0.543	-0.884	0.38	0.213	1.79
Sexo_M	-0.202	0.817	0.574	-0.352	0.73	0.265	2.52
Semestre_Primavera	0.18	1.198	0.393	0.459	0.65	0.554	2.59
LocalForaneo_LOCAL	0.508	1.663	0.502	1.013	0.31	0.622	4.45
Prepa_Encode_Prob_abandon	3.224	25.127	0.766	4.21	0.000026	5.602	112.7
ind_planea	-0.665	0.514	1.107	-0.601	0.55	0.058	4.5
Marginalidad	0.291	1.337	0.941	0.309	0.76	0.211	8.45
CatNombControl2	-0.381	0.683	0.843	-0.452	0.65	0.131	3.56
PAAimp	-1.26	0.284	0.758	-1.662	0.096	0.064	1.25

1. Variables significativas
    a. Edad de ingreso ($p < 0.001$): Por cada aumento unitario en la edad al momento de ingresar, el riesgo de abandono académico aumenta significativamente (Exp(β) = 39.295).
    b. *Prepa_Encode_Prob_abandon* ($p < 0.001$): Los estudiantes con provenientes de preparatorias con alto riesgo de abandono tienen una probabilidad 25 veces mayor de abandonar (Exp(β) = 25.127).
    c. *PAAimp* ($p = 0.096$): Aunque no significativamente, esta variable parece estar asociada con una reducción en el riesgo de abandono.
2. Variables no significativas:
    a. Variables como promedio preparatoria, beca, crédito, y género no mostraron asociaciones estadísticamente significativas con el riesgo de abandono académico ($p > 0.05$).

Posteriormente, se creó un nuevo modelo utilizando solo las variables seleccionadas por LASSO: *EdadIng*, *PromPre*, *Prepa_Encode_Prob_abandon*, *ind_planea*, y *PAAimp*. El modelo resultante fue ajustado utilizando los datos de entrenamiento con un total de 277 casos.

El modelo presentó un Pseudo Log-likelihood de -193, ligeramente menor al modelo con todas las variables. El Pseudo likelihood ratio test resultó en un valor de 47.2 con 5 grados de libertad, mostrando que el modelo es significativo. A continuación, la **Tabla 2** muestra los resultados de las variables seleccionadas por LASSO:

Tabla 3 Resultados del Modelo con Variables Seleccionadas por LASSO para la Predicción de Baja Académica

Variable	*Coeficiente (β)*	*Exp(β)*	*Error estándar*	*Valor-Z*	*p-valor*	*IC 95% (Límite Inferior)*	*IC 95% (Límite Superior)*
EdadIng	3.613	37.08	0.47	7.692	1.4E-14	14.768	93.103
PromPre	-0.62	0.538	0.574	-1.079	0.28	0.175	1.659
Prepa_Encode_Prob_abandon	2.96	19.298	0.669	4.425	9.6E-06	5.202	71.596
ind_planea	-0.328	0.72	0.55	-0.596	0.55	0.245	2.119
PAAimp	-1.588	0.204	0.748	-2.124	0.034	0.047	0.884

1. Variables significativas:
    a. Edad de ingreso ($p<0.001$): Es el predictor más relevante, con un aumento significativo en el riesgo de abandono académico por cada incremento unitario en la edad al momento de ingresar(Exp(β) =37.080).
    b. *Prepa_Encode_Prob_abandon* ($p<0.001$): Los estudiantes con provenientes de preparatorias con alto riesgo de abandono tienen una probabilidad casi 19 veces mayor de abandonar.
    c. *PAAimp* ($p=0.034$): Se asocia significativamente con una disminución en el riesgo de abandono académico (Exp(β) =0.204).
2. Variables no significativas:
    a. Promedio preparatoria e *ind_planea* no mostraron asociaciones significativas ($p>0.05$).

Comparación de Modelos y métricas de evaluación

El modelo completo incluyó 12 variables predictoras, mientras que el modelo LASSO seleccionó únicamente 5 de ellas, reduciendo la complejidad del modelo sin comprometer su desempeño general. En ambos modelos, las variables *EdadIng* y *Prepa_Encode_Prob_abandon* resultaron ser los predictores más relevantes y significativos ($p<0.001$), mientras que otras variables como *PromPre* y *ind_planea* no fueron significativas en ninguno de los dos.

El Modelo Completo mostró un Pseudo Log-likelihood de -192 y un Pseudo likelihood ratio test de 50.5 ($p<0.001$), mientras que el Modelo LASSO, con solo 5 variables, presentó un Pseudo Log-likelihood de -193 y un Pseudo likelihood ratio test de 47.2 ($p<0.001$). Esto indica que el modelo reducido mantiene un desempeño estadístico muy similar al completo, con la ventaja de mayor simplicidad y parsimonia.

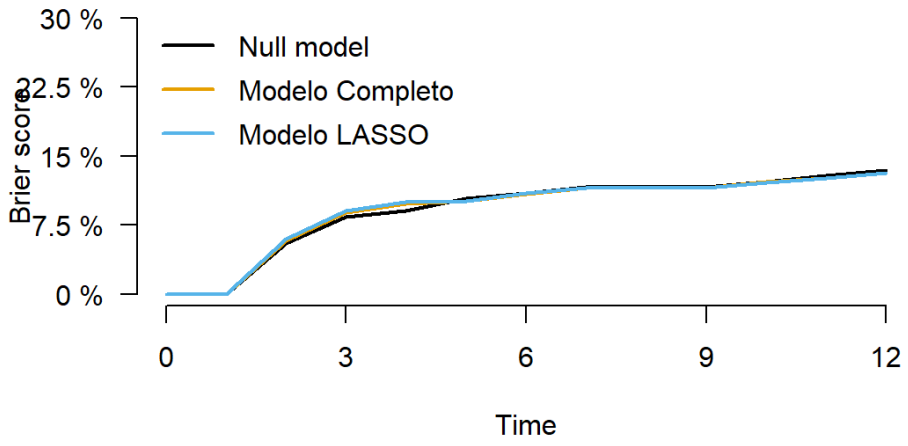
Comparación del Brier Score

El Brier Score, que evalúa la precisión de las predicciones probabilísticas, fue utilizado para comparar ambos modelos a través del tiempo. Los resultados muestran:

- En tiempos iniciales (t=0 a t=3): Ambos modelos tienen un Brier Score muy bajo, cercano a 0%, indicando alta precisión en predicciones iniciales. Las diferencias entre el Modelo Completo y el Modelo LASSO son mínimas (Δ Brier Score entre 0.0%y 0.2%).
- En tiempos intermedios y largos (t=4 a t=12): Los Brier Scores de ambos modelos se estabilizan en un rango de 15% a 25%, mostrando un desempeño similar. Las diferencias en Δ Brier Score no son estadísticamente significativas, oscilando entre -0.1% y 0.2%, con amplios intervalos de confianza ([−0.6%;0.6%]).

La **Figura 11** del comparativo de Brier Score refuerza estas observaciones, mostrando curvas prácticamente idénticas para ambos modelos a lo largo del tiempo. Esto confirma que el Modelo LASSO logra mantener una precisión similar al Modelo Completo, a pesar de la reducción en el número de variables.

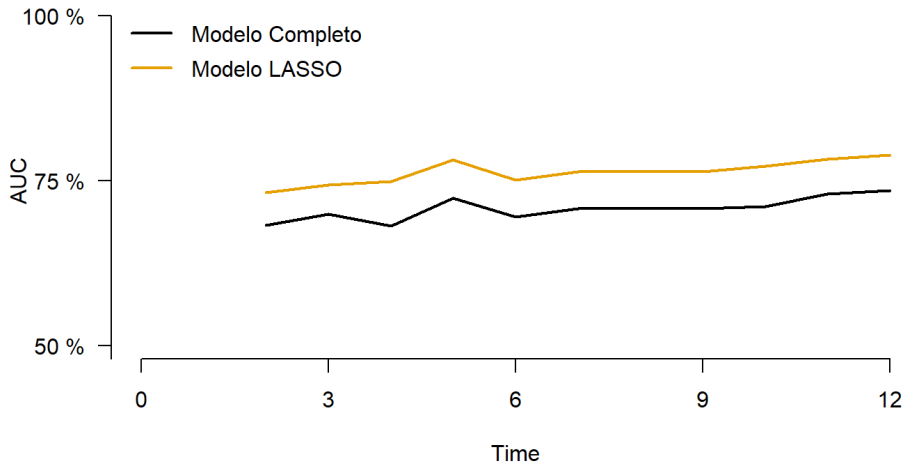
Figura 11 Comparación de Brier Score entre Modelos con todas las variables y LASSO para Baja Académica



Comparación con AUC

El desempeño de los modelos comparados se presenta en la **Figura 12**Error! Reference source not found.**,** donde se observa la evolución del AUC (%) a lo largo del tiempo para el Modelo Completo y el Modelo LASSO. El Modelo LASSO muestra un desempeño consistentemente superior al Modelo Completo en todos los intervalos evaluados, destacando su capacidad para manejar datos complejos y seleccionar variables relevantes.

Figura 12 Comparación de AUC entre Modelos con todas las variables y LASSO para Baja Académica



Por ejemplo, en el tiempo 5, el AUC del Modelo LASSO alcanza el 78.2%, mientras que el del Modelo Completo es de 72.3%, con una diferencia de 5.9 puntos porcentuales. Esta tendencia se mantiene a lo largo del análisis, alcanzando una ventaja de hasta 6 puntos porcentuales hacia el tiempo 12. Además, los intervalos de confianza del Modelo LASSO son más estrechos, lo que indica mayor precisión y estabilidad en sus predicciones.

5.2.1.2 Modelos Cambio de Carrera.

Para el evento de cambio de carrera se ajustó de igual forma un modelo con las 12 variables predictoras en el conjunto de datos de entrenamiento (N=277), cuyos resultados se muestran en la **Tabla 4**, y posteriormente otro modelo solo con las 5 variables más significativas de acuerdo con la regularización Lasso de este evento en particular.

El modelo presentó un Pseudo Log-likelihood de -251 y un Pseudo likelihood ratio test de 19.6 con 12 grados de libertad (p=0.074).Aunque el modelo completo es significativo, algunas variables específicas destacan por su asociación con el cambio de carrera.

Tabla 4 Resultados del Modelo Todas las Variables para la Predicción de Cambio de Carrera

Variable	*Coeficiente (β)*	*Exp(β)*	*Error estándar*	*Valor-Z*	*p-valor*	*IC 95% Inferior*	*IC 95% Superior*
EdadIng	-4.241	0.0144	2.8	-1.514	0.13	5.95E-05	3.482
PromPre	-1.777	0.1691	0.836	-2.126	0.034	0.0329	0.87
Beca	-0.385	0.6804	0.684	-0.563	0.57	0.178	2.601

Credito	0.432	1.5404	0.497	0.87	0.38	0.582	4.078
Sexo_M	0.101	1.1067	0.559	0.181	0.86	0.37	3.312
Semestre_Primavera	0.191	1.2107	0.343	0.558	0.58	0.618	2.37
LocalForaneo_LOCAL	0.29	1.3362	0.415	0.698	0.49	0.592	3.017
Prepa_Encode_Prob_abandon	1.553	4.7268	0.657	2.365	0.018	1.301	17.124
ind_planea	1.341	3.8223	0.933	1.438	0.15	0.614	23.778
Marginalidad	1.238	3.4475	0.964	1.285	0.2	0.522	22.785
CatNombControl2	0.401	1.4941	0.896	0.448	0.65	0.258	8.658
PAAimp	-1.629	0.1961	0.77	-2.116	0.034	0.0434	0.887

1. Variables significativas:
    a. Promedio preparatoria (p=0.034): Un menor promedio en preparatoria está asociado con un mayor riesgo de cambio de carrera (Exp(β) =0.169), lo que sugiere que estudiantes con bajo desempeño académico son más propensos a cambiar de programa.
    b. *Prepa_Encode_Prob_abandon* (p=0.018): Los estudiantes provenientes de preparatorias con alto riesgo de abandono tienen una probabilidad significativamente mayor de cambiar de carrera (Exp(β) =4.727).
    c. *PAAimp* (p=0.034): Esta variable está asociada con una reducción del riesgo de cambio de carrera (Exp(β) =0.196).
2. Variables no significativas:
    a. Variables como Edad de ingreso (*EdadIng*), *Beca*, *Crédito*, y otras no mostraron asociaciones estadísticamente significativas (p>0.05).

En cuanto al modelo ajustado con variables seleccionadas mediante LASSO, cuyos resultados se muestran en la **Tabla 5**, incluye cinco variables predictoras: *EdadIng*, *Prepa_Encode_Prob_abandon*, *ind_planea*, *Marginalidad,* y *PAAimp.* Este modelo busca simplificar el análisis identificando las variables más relevantes para predecir el cambio de carrera, sin perder precisión.

El modelo presentó un Pseudo Log-likelihood de -255 y un Pseudo likelihood ratio test de 11.9 con 5 grados de libertad. Esto indica que el modelo es globalmente significativo y explica de manera adecuada el cambio de carrera con las variables seleccionadas.

Tabla 5 Resultados del Modelo con Variables Seleccionadas por LASSO para la Predicción de Cambio de Carrera

Variable	*Coeficiente (β)*	*Exp(β)*	*Error estándar*	*Valor-Z*	*p-valor*	*IC 95% Inferior*	*IC 95% Superior*
EdadIng	-1.793	0.166	1.745	-1.027	0.3	0.00544	5.09
Prepa_Encode_Prob_abandon	1.451	4.266	0.674	2.153	0.031	1.139	15.98
ind_planea	0.955	2.599	0.809	1.18	0.24	0.532	12.69
Marginalidad	0.588	1.801	0.874	0.673	0.5	0.324	10
PAAimp	-1.89	0.151	0.775	-2.439	0.015	0.033	0.69

1. Variables significativas:
    a. *Prepa_Encode_Prob_abandon* (p=0.031): Los estudiantes provenientes de preparatorias con alto riesgo de abandono tienen una probabilidad significativamente mayor de cambiar de carrera (Exp(β) =4.266).
    b. *PAAimp* (p=0.015): Esta variable está asociada con una reducción significativa del riesgo de cambio de carrera (Exp(β) =0.151).
2. Variables no significativas:
    a. *EdadIng, ind_planea*, y *Marginalidad* no mostraron asociaciones estadísticamente significativas (p>0.05).

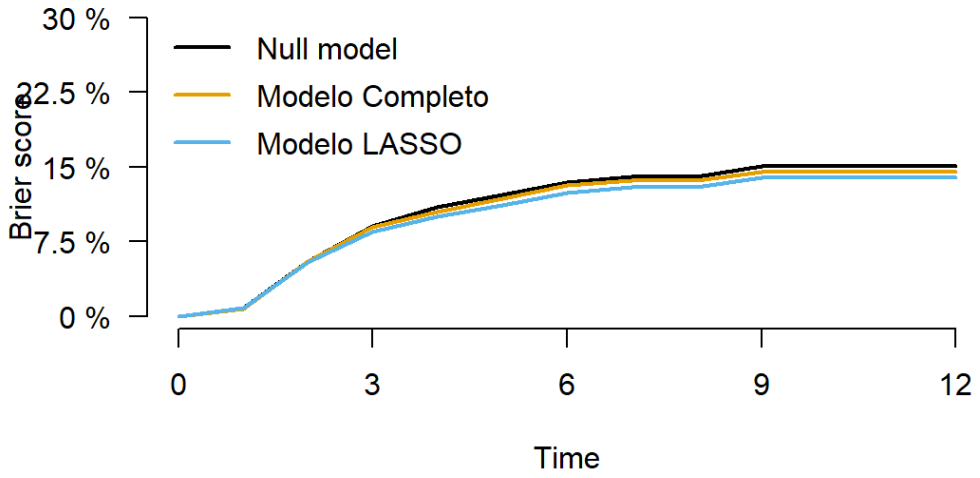
Comparación de Modelos y métricas de evaluación

La comparación entre el modelo completo y el modelo reducido con LASSO para predecir el cambio de carrera demuestra que ambos modelos tienen un desempeño comparable. El modelo completo incluye 12 variables, con un Pseudo Log-likelihood de -251 y un Pseudo likelihood ratio test marginalmente significativo (p = 0.074). Las variables clave identificadas fueron *PromPre*, *Prepa_Encode_Prob_abandon*, y *PAAimp*. Por otro lado, el modelo LASSO simplifica el análisis al incluir solo 5 variables, manteniendo un ajuste estadísticamente significativo (p = 0.036).

Comparación del Brier Score

En términos de precisión predictiva, ambos modelos muestran valores de Brier Score similares en todos los tiempos evaluados ,con ligeras ventajas para el modelo LASSO en tiempos intermedios y avanzados. La **Figura 13**Figura 13 sobre la comparación de los Brier Scores muestra que ambos modelos ajustados mejoran respecto al modelo nulo y tienen curvas casi idénticas, validando que el modelo LASSO es una alternativa eficiente y parsimoniosa sin comprometer el desempeño predictivo.

Figura 13 Comparación de Brier Score entre Modelos con todas las variables y LASSO para Baja Académica

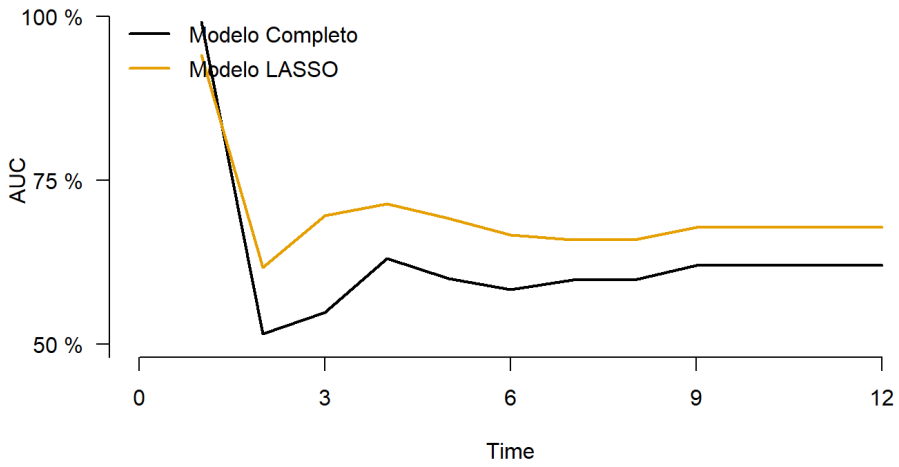


- En tiempos iniciales (t=0 a t=3): los tres modelos (nulo, completo y LASSO) presentan valores muy bajos (⩽9.1%). La diferencia entre el modelo LASSO y el completo es insignificante (Δ BrierScore=0.0%).
- En tiempos intermedios y largos (t=4 a t=12): Ambos modelos ajustados muestran un desempeño similar, con un Brier Score entre 15% y 25%. El modelo LASSO tiende a tener un Brier Score ligeramente menor, aunque las diferencias no son estadísticamente significativas (Δ BrierScore≈−0.6%).

Comparación con AUC

El desempeño de los modelos se resume en la **Figura 14**, que compara el AUC (%) del Modelo Completo y el Modelo LASSO a lo largo del tiempo. En general, el Modelo LASSO supera al Modelo Completo en términos absolutos, aunque las diferencias no son estadísticamente significativas en los intervalos evaluados.

Figura 14 Comparación de AUC entre Modelos con todas las variables y LASSO para Cambio de Carrera



En el tiempo 2, el Modelo LASSO muestra un delta AUC del 10.2% frente al Modelo Completo, con un intervalo de confianza amplio (-8.1; 28.5) y un p-valor de 0.2746. A medida que avanza el tiempo, las diferencias en el AUC entre los modelos se reducen ligeramente, siendo del 5.8% en el tiempo 12, aunque todavía no alcanzan significancia estadística (p-valor de 0.4093).

El Modelo LASSO demuestra una mejora consistente en el AUC frente al Modelo Completo, lo que sugiere su efectividad en la selección de variables. Sin embargo, los intervalos de confianza amplios indican una posible variabilidad en los resultados, lo que podría requerir un análisis adicional o un mayor tamaño de muestra para confirmar estas tendencias.

5.2.1.3 Modelos egreso o graduación.

Por último, el modelo para el evento de graduación se ajustó de igual forma considerando las 12 variables predictoras disponibles en el conjunto de datos (N=277); los resultados se muestran en la **Tabla 6.** El modelo presentó un Pseudo Log-likelihood de -916 y un Pseudo likelihood ratio test de 61.5 con 12 grados de libertad ($p<0.001$), lo que indica que el modelo es significativo en su conjunto.

Tabla 6 Resultados del Modelo Todas las Variables para la Predicción de Egreso o Graduación

Variable	*Coeficiente (β)*	*Exp(β)*	*Error estándar*	*Valor-Z*	*p-valor*	*IC 95% Inferior*	*IC 95% Superior*
EdadIng	-2.586	0.0753	1.073	-2.409	0.016	0.0092	0.617
PromPre	1.16	3.1887	0.435	2.665	0.0077	1.3589	7.482
Beca	-0.043	0.9583	0.322	-0.132	0.89	0.5095	1.802
Credito	0.047	1.0481	0.213	0.22	0.83	0.6902	1.591
Sexo_M	-0.025	0.9753	0.237	-0.105	0.92	0.6129	1.552
Semestre_Primavera	-0.067	0.9355	0.168	-0.396	0.69	0.6727	1.301
LocalForaneo_LOCAL	-0.32	0.7263	0.169	-1.895	0.058	0.5217	1.011
Prepa_Encode_Prob_abandon	-2.628	0.0722	0.691	-3.801	0.00014	0.0186	0.28
ind_planea	-0.009	0.9915	0.443	-0.019	0.98	0.416	2.363
Marginalidad	-0.384	0.6814	0.478	-0.803	0.42	0.267	1.739
CatNombControl2	0.536	1.7095	0.634	0.845	0.4	0.4929	5.928
PAAimp	1.296	3.6537	0.487	2.658	0.0078	1.4056	9.497

1. Variables significativas:
    a. Edad de ingreso (*EdadIng*) (p=0.016): Los estudiantes que ingresan a mayor edad tienen menos probabilidad de graduarse (Exp(β) =0.0753), lo que sugiere que la edad actúa como un factor de riesgo.
    b. Promedio preparatoria (*PromPre*) (p=0.0077): Los estudiantes con un promedio académico previo más alto tienen una probabilidad significativamente mayor de graduarse (Exp(β) =3.1887)
    c. *Prepa_Encode_Prob_abandon* (p=0.00014): Los estudiantes provenientes de preparatorias con alto riesgo de abandono tienen una probabilidad significativamente menor de graduarse (Exp(β) =0.0722)
    d. *PAAimp* (p=0.0078): Una puntuación más alta en esta variable está asociada con un mayor riesgo de graduación (Exp(β) =3.6537).
    e. *LocalForaneo_LOCAL* (p=0.058): Los estudiantes locales parecen tener una mayor probabilidad de graduarse (Exp(β) =0.7263), aunque no es estadísticamente significativo.
2. Variables no significativas:
    a. Variables como *Beca*, *Crédito*, *Sexo_M*, *Semestre_Primavera*, *Marginalidad*, y *CatNombControl2* no mostraron asociaciones estadísticamente significativas con la graduación (p>0.05).

Este modelo identifica que la edad de ingreso, el promedio preparatoria, la probabilidad de abandono de preparatoria de procedencia, y *PAAimp* son los factores más relevantes asociados con la graduación.

En cuanto al modelo ajustado son con LASSO, para el evento de graduación, se incluyen cinco predictores: *EdadIng*, *PromPre*, *Prepa_Encode_Prob_abandon*, *ind_planea*, y *Marginalidad*; los resultados se muestran en la **Tabla 7**. El modelo presentó un Pseudo Log-likelihood de -921 y un Pseudo likelihood ratio test de 52.8 con 5 grados de libertad (p<0.001), indicando que el modelo es estadísticamente significativo.

Tabla 7 Resultados del Modelo con Variables Seleccionadas por LASSO para la Predicción de Egreso o Graduación

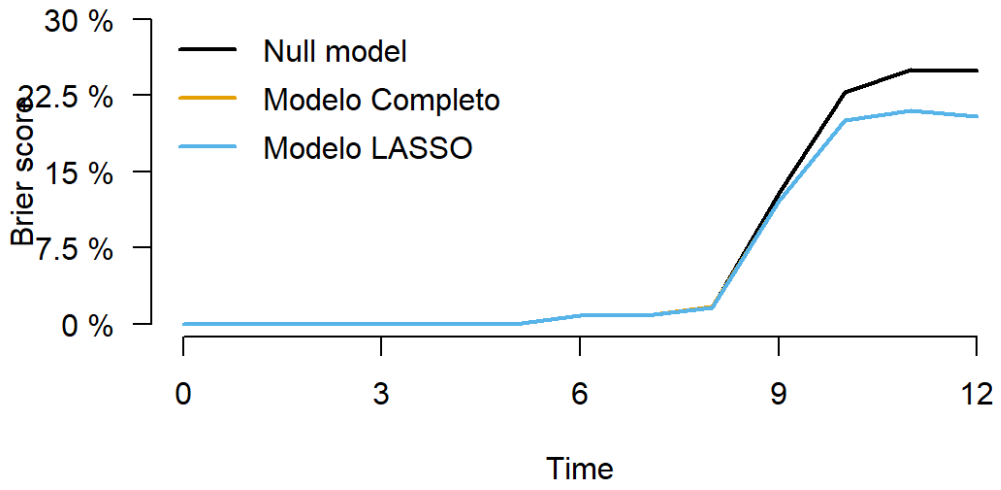
Variable	*Coeficiente (β)*	*Exp(β)*	*Error estándar*	*Valor-Z*	*p-valor*	*IC 95% Inferior*	*IC 95% Superior*
EdadIng	-2.247	0.1057	0.975	-2.304	0.021	0.0156	0.715
PromPre	1.468	4.3421	0.381	3.857	0.00011	2.059	9.157
Prepa_Encode_Prob_abandon	-2.659	0.0701	0.651	-4.082	0.000045	0.0195	0.251
ind_planea	0.171	1.1862	0.421	0.406	0.68	0.5201	2.706
Marginalidad	-0.162	0.8503	0.449	-0.361	0.72	0.3527	2.05

1. Variables significativas:
    a. Edad de ingreso (*EdadIng*) (p=0.021): Los estudiantes que ingresan a mayor edad tienen una probabilidad significativamente menor de graduarse (Exp(β) =0.1057).
    b. Promedio preparatoria (*PromPre*) (p<0.001): Un mayor promedio en preparatoria está asociado con una mayor probabilidad de graduarse (Exp(β) =4.3421).
    c. *Prepa_Encode_Prob_abandon* (p<0.001): la alta probabilidad de abandono de la prepara de procedencia reduce drásticamente la probabilidad de graduarse (Exp(β) =0.0701).
2. Variables no significativas:
    a. *ind_planea* (p=0.68) y *Marginalidad* (p=0.72) no mostraron asociaciones estadísticamente significativas con la graduación.

Comparación de Modelos y métricas de evaluación

El modelo LASSO reduce el número de variables de 12 a 5, eliminando predictores que no aportan significativamente al ajuste. Aunque el Pseudo Log-likelihood del modelo LASSO (−921) es ligeramente inferior al del modelo completo (−916), las variables seleccionadas explican adecuadamente el fenómeno. La **Figura 15** de la comparación de los valores de Brier Score muestran como ambos modelos tienen un desempeño muy similar a lo largo del tiempo.

Figura 15 Comparación de Brier Score entre Modelos con todas las variables y LASSO para Egreso o Graduación



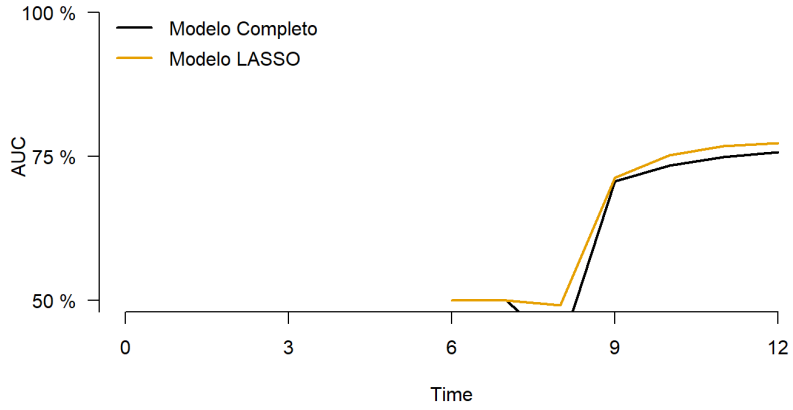
- En tiempos iniciales (t=0 a t=8): Los modelos (nulo, completo y LASSO) tienen un Brier Score de 0.0%, reflejando predicciones perfectas debido a la ausencia de eventos en estos períodos.
- En tiempos intermedios y largos (t=9a t=12): A partir de t=9, los modelos ajustados muestran un desempeño mejor que el modelo nulo. Por ejemplo:
    - En t=10, el modelo con todas las variables y el Modelo LASSO tienen un Brier Score de aproximadamente 20.0%, frente al 22.8% del modelo nulo.
    - En t=12, ambos modelos ajustados mantienen un Brier Score de 20.4%, mostrando predicciones más precisas en comparación con el modelo nulo (24.9%).

Ambos modelos ajustados superan consistentemente al modelo nulo en tiempos avanzados, indicando que las variables predictoras seleccionadas mejoran las predicciones del evento de graduación. El Modelo LASSO mantiene curvas casi idénticas al modelo con todas las variables, validando que la reducción de variables no compromete la precisión predictiva.

Comparación con AUC

El desempeño de los modelos se resume en la **Figura 16,** que compara el AUC (%) del Modelo Completo y el Modelo LASSO a lo largo del tiempo. Durante los primeros intervalos (tiempos 0 a 7), los resultados no presentan variaciones o diferencias significativas, lo que indica que ambos modelos tienen un desempeño equivalente en esos periodos

Figura 16 Comparación de AUC entre Modelos con todas las variables y LASSO para Egreso o Graduación



En el tiempo 8, el Modelo LASSO muestra una mejora destacable respecto al Modelo Completo, con un delta AUC del 7.7% y un intervalo de confianza ajustado (0.1; 15.3). Este es el único punto donde la diferencia entre ambos modelos alcanza significancia estadística (p-valor de 0.04681). A partir del tiempo 9, las diferencias vuelven a ser mínimas, con deltas AUC cercanos al 0% y p-valores que indican ausencia de significancia estadística (e.g., 0.85811 en el tiempo 9).

5.2.1.4 Análisis general de modelos

El análisis de los tres eventos principales —baja académica, cambio de carrera, y graduación— permitió identificar factores clave que influyen en el desempeño y trayectoria académica de los estudiantes. En todos los

casos, los modelos ajustados con LASSO lograron identificar las variables más relevantes, mostrando un desempeño comparable al de los modelos completos, pero con la ventaja de mayor simplicidad y parsimonia.

- Baja Académica: Los predictores más significativos fueron la edad de ingreso y la probabilidad de abandono de preparatoria de procedencia, que aumentan notablemente el riesgo de abandono. El modelo LASSO destacó estas variables como clave, eliminando otras que no aportaban significativamente.
- Cambio de Carrera: Factores como el promedio de preparatoria y la probabilidad de abandono de preparatoria de procedencia se asociaron con una mayor probabilidad de cambio de programa. El modelo LASSO simplificó el análisis al retener estas variables críticas.
- Egreso o Graduación: La edad de ingreso, el promedio de preparatoria, y la probabilidad de abandono de preparatoria de procedencia fueron determinantes para la graduación. Ambos modelos ajustados demostraron un buen desempeño predictivo, con el LASSO manteniendo resultados similares al modelo completo.

En términos de precisión predictiva, los modelos ajustados superaron al modelo nulo en todos los eventos, especialmente en tiempos avanzados, como se refleja en el Brier Score. Los resultados respaldan el uso del modelo LASSO como una herramienta eficiente para identificar variables clave y mejorar la interpretación sin comprometer la calidad de las predicciones.

6. Conclusiones

El primer paso de este análisis consistió en preparar y escalar las variables predictoras para garantizar que estuvieran en una escala comparable. Esto es especialmente relevante porque algunos algoritmos de regularización, como LASSO, son sensibles a la magnitud de las variables. Aunque este conjunto de datos tiene un número limitado de variables, este algoritmo podría utilizarse de manera más efectiva en futuros análisis, al incorporar más características y garantizar que la selección de variables se base en su importancia predictiva relativa y no en diferencias de escala.

Se aplicaron tres regularizaciones para evaluar el comportamiento de los tres eventos y determinar qué variables resultaban más relevantes. En su mayoría, las variables seleccionadas fueron consistentes entre los eventos y coincidieron con las seleccionadas por LASSO, lo que demostró ser una herramienta eficaz para simplificar los modelos, reducir la dimensionalidad y mantener un buen desempeño predictivo. Entre las variables más destacadas se encontraron:

- Edad de ingreso (EdadIng): Factor relevante para todos los eventos, indicando su impacto en la trayectoria académica.
- Promedio preparatoria (PromPre): Asociado con una menor probabilidad de abandono y cambio de carrera, pero con una mayor probabilidad de graduación.
- Probabilidad de abandono de preparatoria de procedencia (Prepa_Encode_Prob_abandon): Un predictor constante y significativo en los tres eventos.

Siguiendo la metodología propuesta por la Dra. Gisel Hernández, se optó por utilizar en el modelo de predicción las variables seleccionadas LASSO; respectivas para cada evento. Se construyeron dos modelos para cada evento utilizando el método de Fine and Gray para riesgos competitivos: uno con todas las variables y otro con las seleccionadas por LASSO. Esta comparación permitió analizar el impacto de las variables seleccionadas en la predicción del modelo.

La precisión de las predicciones se evaluó mediante el Brier Score, que mostró que ambos modelos ajustados tuvieron un desempeño significativamente mejor que el modelo nulo en todos los eventos y a lo largo del tiempo. Los modelos basados en LASSO lograron reducir la complejidad sin comprometer la precisión predictiva, destacándose como una opción simple y eficiente en contextos con similares a este, con conjuntos de datos pequeños y pocos predictores o variables.

<u>Recomendaciones</u>

1. Ampliar el conjunto de datos: El tamaño reducido del conjunto de datos limita la robustez y generalización de los modelos. Se recomienda recolectar más datos, idealmente de cohortes adicionales o instituciones similares, para mejorar la capacidad predictiva y la estabilidad de los modelos.

2. Incluir más variables predictoras: Aunque las variables seleccionadas son relevantes, los modelos podrían beneficiarse de la inclusión de factores adicionales, como características socioeconómicas más detalladas, indicadores de participación académica en preparatoria o datos cualitativos sobre el bienestar de los estudiantes antes de ingresar a la Universidad.
3. Ampliar el modelo a todas las carreras ofrecidas por la universidad: Se sugiere entrenar este modelo predictivo para todas las carreras que ofrece la institución, con el objetivo de identificar las variables específicas que influyen en cada evento (baja académica, cambio de carrera y graduación). Esto permitiría ajustar las estrategias de intervención a las características particulares de los estudiantes que ingresen a cada programa académico.
4. Desarrollar un sistema de apoyo basado en el modelo predictivo: Crear una plataforma o sistema que incorpore este modelo y proporcione herramientas para los coordinadores de cada carrera. Con este sistema se podrían:
    a. Identificar, de acuerdo con las características de los estudiantes de ingreso, la probabilidad de que suceda cada tipo de evento.
    b. Proporcionar recomendaciones automatizadas para crear estrategias personalizadas en conjunto con el área de CAXXA.
    c. Diseñar intervenciones específicas, como tutorías personalizadas para estudiantes con mayor riesgo de abandono, con el objetivo de aumentar su probabilidad de permanencia.
    d. Orientar a los estudiantes con alta probabilidad de cambiar de carrera, ayudándolos a encontrar opciones dentro de la universidad que se adapten mejor a sus intereses y habilidades, evitando así que abandonen la institución.

Estas recomendaciones no solo contribuirían a mejorar las tasas de retención y graduación de la universidad, sino también a proporcionar una experiencia académica más enriquecedora y personalizada para los estudiantes. Además, reforzarían la reputación institucional al implementar una gestión proactiva y basada en datos para el éxito estudiantil.

En conclusión, este análisis demuestra que, incluso con un conjunto de datos pequeño y un número limitado de variables, es posible construir modelos predictivos efectivos y parsimoniosos mediante el uso de regularización como LASSO. Los modelos ajustados son herramientas útiles para identificar factores clave que impactan las trayectorias académicas de los estudiantes. Sin embargo, la generalización de estos modelos requiere un esfuerzo adicional en términos de ampliación del conjunto de datos e incorporación de nuevas variables para capturar la complejidad del fenómeno en su totalidad.

Bibliografía

Ameri, S., Chinnam, R. B., Fard, M. J., & Reddy, C. K. (903-912). Survival Analysis based Framework for Early Prediction of Student Dropouts. *Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM '16)*, 2016.

Anushruthika. (2023, Aug 16). *From Raw to Rescaled: A Guide to Z-Score, Normalization, and Standardization in Data Preprocessing*. Retrieved from Medium: https://medium.com/@anushruthikae/from-raw-to-rescaled-a-guide-to-z-score-normalization-and-standardization-in-data-preprocessing-173874df077d

Anwar, A. (2021, Feb 4). *Types of Regularization in Machine Learning*. Retrieved from Towards Data Science: https://towardsdatascience.com/types-of-regularization-in-machine-learning-eb5ce5f9bf50

Austin, P. C., Lee, D. S., & Fine, J. P. (2016). *Introduction to the Analysis of Survival Data in the Presence of Competing Risks.*

Bonifro, F. D., M. G., G. L., & Zingaro, S. P. (2020). Student Dropout Prediction. *Springer*, pp. 129-140. doi: https://doi.org/10.1007/978-3-030-52237-7_11

Borgan, O. (1997). Three contributions to the Encyclop edia of Biostatistics: The Nelson- Aalen, Kaplan-Meier and Aalen-Johansen. In J. W. Ltd, *Encyclopedia of Biostatistics.* Oslo, Norway.

Bradley, A., Schwartz, S., & Hashino, T. (2007). Sampling Uncertainty and Confidence Intervals for the Brier Score and Brier Skill Score.

Eryk, L. (2020, 08 23). *Towards Data Science*. Retrieved from https://towardsdatascience.com/introduction-to-survival-analysis-the-nelson-aalen-estimator-9780c63d549d

Fine, J., & Gray, R. (1999). A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Taylor & Francis*.

Gray, J. P. (1999). A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Taylor & Francis*.

Gutierrez-Pachas, D. A., Garcia-Zanabria, G., Cuadros-Vargas, E., Camara-Chavez, G., & Gomez-Nieto, E. (2023). Supporting Decision-Making Process on Higher Education Dropout by Analyzing Academic, Socioeconomic, and Equity Factors through Machine Learning and Survival Analysis Methods in the Latin American Context. *Education Sciences*, 13(2), 154. doi:https://doi.org/10.3390/educsci13020154

ITESO. (2021). *Estrategia Institucional ITESO 2022-2026.* Tlaquepaque, Jalisco.

Kawaguchi, E. S., Shen, J. I., Li, G., & Suchard, M. A. (2021). A Fast and Scalable Implementation Method for Competing Risks Data with the R Package fastcmprsk. *National Institutes of Health Grant*.

Kleinbaum, D., & Klein, M. (2012). *Survival Analysis* (Third Edition ed.). Atlanta: Springer.

Mduma, N., Kalegele, K., & Machuve, D. (2019). A Survey of Machine Learning Approaches and Techniques for Student Dropout Prediction. *Data Science Journal*, 18(14), 1-10. https://doi.org/10.5334/dsj-2019-014.

Patacsil, F. F. (2020). Survival Analysis Approach for Early Prediction of Student Dropout Using Enrollment Student Data and Ensemble Models. *Universal Journal of Educational Research*, 4036 - 4047.

Sadafule, S., & Sarkar, S. (2022). G-AUC: An improved metric for classification model selection. *26th International Computer Science and Engineering Conference (ICSEC)*.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software,*. doi:https://doi.org/10.18637/jss.v039.i05

Tay, K., Simon, N., Friedman, J., Hastie, T., Tibshirani, R., & Narasimhan, B. (2023). Regularized Cox Regression.

Xu, W. (2019, Aug 21). *What's the difference between Linear Regression, Lasso, Ridge, and ElasticNet in sklearn?* Retrieved from Towards Data Science: https://towardsdatascience.com/whats-the-difference-between-linear-regression-lasso-ridge-and-elasticnet-8f997c60cf29

Cox, D. R. (1972). Regression Models and Life-Tables. In *Journal of the Royal Statistical Society. Series B (Methodological)* (Vol. 34, Issue 2).

Hernández-Chávez, G. (2024). *APLICACIÓN DE BOSQUES DE SUPERVIVENCIA ALEATORIOS A LA PREDICCIÓN DE ABANDONO UNIVERSITARIO*. Universidad de Guadalajara.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, *2*(3), 841–860. https://doi.org/10.1214/08-AOAS169

Liu, C., Liang, Y., Luan, X. Z., Leung, K. S., Chan, T. M., Xu, Z. Ben, & Zhang, H. (2014). The L1/2 regularization method for variable selection in the Cox model. *Applied Soft Computing*, *14*(PART C), 498–503. https://doi.org/10.1016/J.ASOC.2013.09.006